Web Data Cleaning and Analytics for Malaysia Tourism

By

Tee Hong Le

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology

(Perak Campus)

JANUARY 2019

UNIVERSITI TUNKU ABDUL RAHMAN

REPORT STATUS DECLARATION FORM

Title: _____
_____
_____

Academic Session: _____

I _____

(CAPITAL LETTER)

declare that I allow this Final Year Project Report to be kept in
Universiti Tunku Abdul Rahman Library subject to the regulations as follows:

1. The dissertation is a property of the Library.
2. The Library is allowed to make copies of this dissertation for academic purposes.

Verified by,

_____ _____

(Author's signature) (Supervisor's signature)

Address:

_____
_____ _____
_____

Supervisor's name

Date: _____ **Date**: _____

Web Data Cleaning and Analytics for Malaysia Tourism

By

Tee Hong Le

A REPORT

SUBMITTED TO

Universiti Tunku Abdul Rahman

in partial fulfilment of the requirements

for the degree of

BACHELOR OF COMPUTER SCIENCE (HONS)

Faculty of Information and Communication Technology

(Perak Campus)

JANUARY 2019

DECLARATION OF ORIGINALITY

I declare that this report entitled "**METHODOLOGY, CONCEPT AND DESIGN OF A SYSTEM MODEL FOR WEB DATA CLEANING AND ANALYTICS FOR MALAYSIA TOURISM WITH PYTHON**" is my own work except as cited in the references. The report has not been accepted for any degree and is not being submitted concurrently in candidature for any degree or other award.

Signature : _____

Name : _____

Date : _____

ACKNOWLEDGEMENTS

I would like to thank to my supervisor, Dr Liew Soung Yue who give me this bright opportunity to engage in a data analytics project. It is my first step to establish my career in data analytics field. I am assigned to do a part of this project which is Web Data Cleaning and Analytics for Malaysia Tourism. A million thanks to you and I will take this project very seriously on doing the tasks.

Finally, I must say thanks to my parents and my family for giving me mentality support when I was mentally not feeling well.

ABSTRACT

This project is about data pre-processing, cleaning and analytics in Malaysia's tourism sector. It will provide brief information about the importance of applying data analytics in tourism sector and some methodology on cleaning and pre-processing tourism dataset. To allow data analytics for some useful attributes, unwanted strings that lies in the attributes must be stripped away and only keep the numeric values which is the only value that brings the meaning of the attributes itself. New attributes are introduced which are integrated from the existing attributes such as Latitude, Longitude and text sentiment analysis scores. After that, data analytics phase is used to determine the tourism trends in Malaysia from multiple aspects such as particular tourism location, traveller's country of origin and type of travellers which have different preferences. Data analytics that on time series of tourism trends will be discussed in this project too.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

VADER	Valence Aware Dictionary and Sentiment Reasoner
NLTK	Natural Language ToolKit
TRI	Trivago Rating Index
CSV	comma separated values
GRW	gain ratio-based approach
CMC	computer mediated communication
AMT	Amazon Mechanical Turk

Chapter 1: Introduction

1.1 Problem Statement

Malaysia is a well-known tourist destination in the world for its fascinating multi-cultures, historical heritages, natural sceneries, delicious cuisine, etc. According to Malaysia Tourism Statistics in Brief (2016), Malaysia has an average of 26.42 million of tourist arrivals, and an average of RM65.7 billion of receipts made in the year from 2006-2016. This proves that tourism is one of the major income sources for Malaysia. To further increase the income from tourism in Malaysia, more advanced technology is needed. For instance, tourism data analytics in Malaysia should be launched to generally improve the positive experience of travellers. Throughout the years, it has been observed that a huge amount useful data in the Malaysia's tourism sector are there but not being analysed. This can actually be considered as a great loss in the sector the useful data are not actually used to improve the tourism. By referring to Travel Packages | Tourism Malaysia from the main website of Malaysia's tourism, there is much space for improvement. For example, there is still lack of existing customers' feedbacks although there are so many travel packages. Without collecting rating and review from customers, government or the tourism industry would not know the opinions from customers, and therefore they may have no clue how to improve. Moreover, not every travel package will include accommodation in the list. This might cause customers having difficulties to find an accommodation nearby the attractions they wish to go. Even no accommodation provided, government or industry should provide some recommendations for customers to ease their efforts in planning the visitation. In general, government has lost a lot of chances to collect and get use of the data for tourism sector. With the loss of much data and chances of collecting data, it causes a huge disadvantage for Malaysia tourism sector. In my project, I will focus on Web Data Cleaning and Analytics for Malaysia Tourism by providing useful information which transformed from raw data. This is because there are a lot of information that can be learned by analysing large dataset to generally improve tourism sector in Malaysia.

1.2 Background and motivation

It is very crucial for Malaysia government to collect and get used on every data collected from travellers. This is because every data collected can undergo data analytics to understand customers' behaviour on attractions thoroughly. By having a deep understanding of customers' behaviour, government is able to make predictions accurately based on the customers' nationality, gender, age etc. For example, government can analyse a couple from England who departed to Penang will more likely to visit beaches at Batu Ferringhi, and having an accommodation at seaside nearby. By collecting every personal data, and also reviews, ratings from customers, government is able to make promotions and recommendations to customers which have higher possibilities on visiting a particular attraction based on the data analytics made by government. With an accurate prediction on customers' behaviour, customers will feel very pleased to reduce their effort for finding suitable packages, attractions and accommodations. This will leads to positive feedback from customers, so that there are higher probability for customers to take a visit back to Malaysia. In return, Malaysia can boost up their income from tourism sector.

Since government data is not easy to request, the big picture of this project is started from developing an automated web crawler to crawl public data from travelling recommendation websites. The automated web crawler will crawl the data from these websites dynamically and store into csv file which act as our database. After that, the raw data collected undergo data pre-processing and cleaning so the attributes of data is changed to appropriate one and the unwanted noises of data can be cleared. Then, text sentiment and correlation analysis is used after the unwanted noises and changes of data attributes has been made. This project starts from collecting useful public data from a travelling recommendation website who provides every aspect of tourism sector services such as accommodation, restaurant and attraction. This system is expected to analyse the data thoroughly to find out the correlation of data set collected so that predictions can be made on customers' behaviour based on their reviews, ratings and rated date in every aspect of tourism sector. After that, data visualization is used to display some meaningful graphs to help tourism sector in Malaysia to understand their weaknesses and strengths for further improvements. There are also some meaningful graphs that are able to help customers to have better decision making on choosing

places to travel in Malaysia. For example, recommending a family or couple to particular attraction that are suitable for most of the family or couple who want to travel.

1.3 Project Objectives

This project aims to clean the raw data thoroughly to produce an accurate outcome for analysis. To be exact, this is the most important phase to identify an accurate information based on the raw data collected. This is because an inaccurate data cleaning will directly leads to an inaccurate of data analytics. This leads to wrong information obtained and affects the decision making based on the false information learned from the uncleaned raw data. According to (Press, G., 2016) survey's from existing data scientist, it shows that 60% of the time spent in the whole process is used in data cleaning. To ensure the data is cleaned thoroughly, every attributes must be checked carefully and it the time spent on data cleaning might be exponential growth based on the number of attributes.

Besides, this project aims to determine a data model of text sentiment analysis with higher accuracy that is suitable for public reviews collected from different sources. This is because producing a text sentiment with higher accuracy can lead to higher accuracy on every prediction of other attributes which correlates to text sentiment's attribute in the database. Text sentiment analysis will be completed by using one of the natural language toolkit (NLTK) in python library. This project does not involves developing a self-made library as it will consume too much time to determine the weights of every texts.

This project also aims to find out every correlation of attributes collected so that the ultimate goal which is data analytics in further of this project can be achieved easier by studying the correlation graph formed. Apart from the correlation based on attributes collected, this project aims to determine the correlation between the event launched by government for tourism such as 'Visit Malaysia 2014', or it can be focus on certain state such as ' Visit Kuala Lumpur 2018'. These meaningful correlation graphs will be visualized to prove that the usefulness of the event raised by the government and also travellers preferences in Malaysia. Peak detection algorithm is also used in this phase to allow easiness for analysing data as the dataset will be growing large when the web scrapper scraps more data into it.

1.4 Proposed approach/study

For the tools to use, Python will be the only programming language that is used to do this project. This is because Python have lots of library that is satisfied the needs of this project. For example, Numpy, Matplotlib and NLTK.

For text sentiment analysis, Valence Aware Dictionary and Sentiment Reasoner (VADER) Sentiment Analysis by the main contributors, (Hutto, C.J. & Gilbert, Eric. 2015) from Python NLTK library is used on analysing the reviews of public data collected. According to (Calderon P. 2018), VADER sentiment analysis is a rule-based sentiment and lexicon analysis tool that used in social media sentiments specifically. Unlike machine learning approaches, lexicon approaches need not to train a model using large amount of data. This is because this type of approach have a 'dictionary of sentiment' which build a lexicon by mapping words to sentiment. By using this approach, every words will mapped into the library and each words has a predefined weightage. For a sentence with many words, the weightage of each words are summed up to get a mean value to display the intensity of the whole sentence.

For correlation analysis, Matplotlib and Plotly are used to plot graphs to determine correlation between attributes. This is because Matplotlib contains various types of graph such as histogram, scatter plot and 3D plot. Scatter plot is used initially to find the pattern of the graph. For example, to prove text sentiment analysis by VADER is accurate, there must be a correlation between rating and compound of text sentiment reviews. Higher rating will have positive compound and lower rating will have negative compound in text sentiments. This can prove that the text sentiment analysis of VADER is accurate or not. Besides, for Plotly, it is an interactive graph which allows zoom in and out, focus on certain time series path and also hover function to determine the values falls on the point in the graph. It is easier to analyse the data using Plotly especially for the time series data.

1.5 Highlight of what have been achieved

In this project, a clear understanding on each attributes from the open data collected has been achieved. After understanding the data collected, data pre-processing and cleaning phase is used to greatly improve the reliability of the data as there are noises, redundancy and faulty input from the users. For example, the input from the users that states their nationality and the city they are staying. There are too many ways for them to input their nationality as the tourism recommendation website do not have a predefined input format.

Besides, there are also many different attributes with different data type. After data pre-processing, every data type are changed to the desired one. After that, few attributes are generated based on the attributes collected. For example, the helpful ratio are calculated based on the helpful votes and contributions from the reviewers. Helpful votes is voted by other users when they feel the reviews made are useful to them while the attribute 'contributions' are the total number of the reviews that the reviewers had made in the online tourism recommendation websites. The review's length of a particular review is also stored to determine the length of reviews correlates with the ratings given. There are also attributes of the latitude and longitude of the particular tourism location introduced to have a more precise tourism location on a map. These coordinates are collected based on the Names, States and Postcode of the particular tourism location. It is important so that the coordinates of the location can be tracked and so the popularity. This can also be used for data visualization by displaying a heat map that shows the popularity of the tourism locations.

In data analytics phase, there are few interesting and meaningful correlation had been studied and there are proves to justify the correlation based on the graphs visualized. There are also algorithm implemented to allow easiness of analysing the data through finding the peak count dynamically based on the query set. By sorting the most peak count graphs, it is then displayed and analysed as the higher peak counts can indicate as more undefined trends to be studied.

1.6 Report organization

This project is arranged starts from the introductions in Chapter 1, which provides some basic idea about the flow and the objectives of this project. Besides, it also describes the problem statements, background of motivation, objectives and proposed approach for this project.

For Chapter 2, there are lists of literature reviews that relates to this project. For example, data visualization, time series analysis, text sentiment analysis and data pre-processing phase. These literature reviews are studied thoroughly to find out ideas and ways from the others to apply in this project by determining the papers' strengths and weaknesses.

In Chapter 3, the project's system design will be written down. It will be more general idea which consists of data pre-processing, cleaning and analytics phase. While for the specific implementation for data pre-processing and cleaning is written in Chapter 4, and data analytics is written in Chapter 5. Flowcharts are shown in this chapter to enable easy to understand the whole process in this project. There are also implementation issues and challenges in this project written in Chapter 3 which elaborates the difficulty of data pre-processing, cleaning and analytics phase.

In Chapter 4, a more specific methodology in data pre-processing and cleaning will be elaborated. The open data collected tends to have a lot of faulty inputs and in this chapter, the objective is to clean out the noisy data and also pre-process the data so that during the data analysis process, the correctness of the data with highest precision can be kept.

In Chapter 5, specific data analytics phase is written in details. The tourism trends that obtained from the data will be listed out in this chapter. There are also extra data from different sources that will be taken into considerations such as temperature data, humidity data, Google Trend data and also number of arrivals to Malaysia from government sources. Meaningless data is then converted into useful information that will leads to useful business decision.

Chapter 6 will be the last chapter in this project which draws the conclusion in this project. Every information obtained will be summarized in this chapter.

Chapter 2: Literature Review

2.1 The Gurney Resort Hotel & Residences

In tourism, it could not be deny that a comfortable accommodation has a high impact towards the traveller's experience. According to (Sharpley.R. 2000), a tourist destination success highly depends on qualitative and quantitative characteristics of supply of the accommodation. This is because an accommodation is a basic element of the overall destination planning process. If a traveller do not satisfied to the accommodation, obviously he or she would not likely to recommend his or her friends and relatives to there, this might affect a country's image too.

As a traveller, it is important to book your hotel room before travelling to prevent inconvenience happened. With the growth of internet, most of the hotels do create their own websites to enable customers for booking hotel room in a particular date without having a phone call. By reviewing Gurney Resort Hotel website, customer can choose the date they desired few months before they travel through the hotel website. Moreover, customer can know the details of the hotel room before booking by having facilities listed on the website. To enhance the clarity of customers, Gurney Resort Hotel website do provide pictures of room, the size of room and many others information. In Gurney Resort Hotel website, they also provides location on the map and area information which provide suggestions for customers to nearby attractions for travellers. Phone number and email of the hotel are provided in the website to avoid loose contact from customers when they have questions. With all of these features, customers can book hotel room easily. Besides, without having a middleman service in a transaction, the percentage of transaction failure might decrease. It is more reliable when customers can directly pay to the hotel without a middleman.

However, there are thousands and millions of hotels in the world. Even a state in a country, do have thousands of hotels. So, customers might face difficulties to search for hotels in a particular place in Malaysia. If they want to know more details of hotels in the desired area, they need to access to each hotel's website to check and make comparison in multiple aspects of a hotel. This is absolutely not practical for customers to make a right decision on booking a hotel room. Besides, hotel websites are not transparent to customers. It is obvious that every hotel do provide customers feedback and rating to the hotel. After collecting reviews from customers, most of hotel websites

are believed to display only positive feedback from customers. They intend to hide their negative feedbacks because they want to create more positive impressions to incoming customers. This might likely caused customers to make their choice wrongly when choosing the hotels.

In general, hotel websites do allows customers who had finalised their decision to book a room from a particular hotel conveniently. It also lessen the rate of failure when booking a hotel room without a middleman service. On the other hand, for those customers who want to make comparison and search for room within their budget, it is believed to be impractical for them to open up lots of tab in browser to make comparison. The reviews from existing customers displayed are also less trustable.

2.2 Booking.com

To resolve the issue in the huge amounts of hotel websites, hotel searching engine is introduced to allow customers to search for hotels in a particular country, state and city. User can set their own requirements on a particular hotel with several of hotel searching engine websites such as booking.com, expedia.com and agoda.com. These websites bring convenience for users to book their hotel they desired by lessen customer's effort to compare for various hotels nearby the places of the area they are going to visit. The requirements can be filtered to narrow down choices of hotels to customers. For example, a customer wants to book a hotel with breakfast included, have swimming pool, budget below RM400 and free Wi-Fi. This features could be achieved by having a powerful hotel searching engine. Moreover, it is more transparent to customers as this hotel searching engine allows customers to make a review and rate the hotel that they had stayed. Every customers' review and rating on each aspect of performance on the hotel will be displayed on the hotel searching website. With the large number of reviews and rating from customers on a hotel, the hotel searching websites will display the average rating to enable customers to directly know that the hotel's performance in each aspect such as cleanliness, staff, comfort and facilities. Unfortunately, there are too many hotel searching engines which offers different price of the hotel although it is the same room. For example the hotel prices in Booking.com and Expedia.com as shown below:

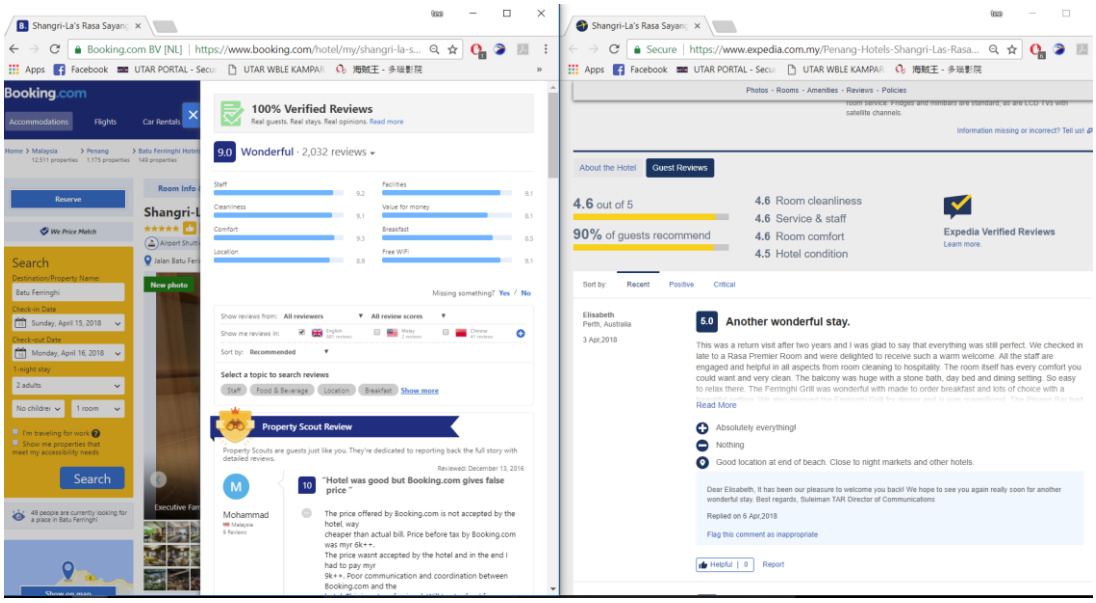


Figure 2.1: Comparison between Booking.com and Expedia.com

With different hotel searching engines, users had to do extra comparison between each hotel booking websites to check that they had chosen the right choice. Although this solves the problems of huge amount of hotel websites which leads to more effort for customers to search for hotel, the large number of hotel searching engines do have the same limitations with the hotel websites. Moreover, different hotel searching engines do have different reviews and ratings in a same hotel too. This caused the inconsistent of data of a hotel performance. With these hotel searching engines which act as a middleman service, it would have higher possibility to have failure of transaction to the particular hotel.

2.3 Trivago.com

To resolve the issues, there is a website which is Trivago.com launched. Trivago is known as a more advanced hotel searching engine which helps users to compare the prices of a particular hotel from different hotel booking websites. It claims that it only appears the list of prices of a hotel from lowest after making comparison from many of the others hotel searching engine and also the hotel websites. With Trivago, users who aims for lowest price would be very pleased to use this website.

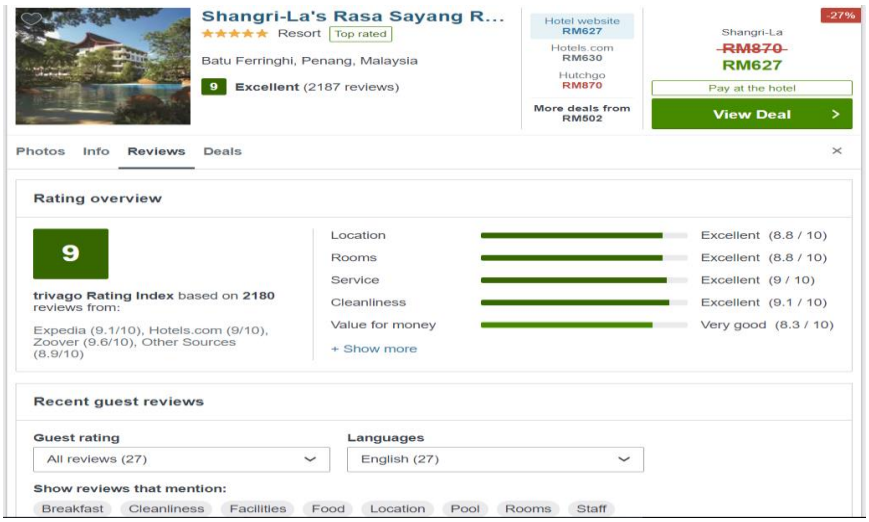


Figure 2.2: Trivago accommodation booking price comparison.

Besides, Trivago totals up the rating overview from other websites and make the average of the ratings. The ratings are classified into multiple aspects which includes location, rooms, service, food and facilities. So, users can indicate that the average performance of the particular hotel before dealing with it. Unfortunately, every hotel's performance are not consistent over time. Trivago does not have the ability to keep track of a particular hotel's performance over time with the displayed ratings in the website. So, it might not meet the expectations of the user who booked the hotel with high ratings in some aspects. For example, 'hotel A' has a rating with 10/10 in staff services category in year 2017. One year later, majority of the staffs in 'hotel A' resigned their job, 'hotel A' needs to recruit newbies to work for them. It is obvious those inexperienced staffs would not perform well as those past experienced staffs. Eventually, they would not have 10/10 in rating for staff services in year 2018.

Guest reviews are known to be the crucial part in order to let other users knows whether the hotel is suitable to them. Besides, Trivago do implement their guest reviews and named it as Trivago Rating Index (TRI). It claims that it is clear for understanding,

transparent to provide links which enable users to view ratings and reviews directly from the source. However, most of the users would not spend their time to read the reviews and see the rating of each hotel line by line. It often waste their time to make decision too. Although Trivago managed to list out the lowest price offered by each websites, but price is not the only things for customers to take into considerations. There are also ratings, reviews and value of money which is very important to increase the probability of customers to have a positive feedback after staying the particular hotel.

2.4 Microsoft Power Bi

To improve the limitations from the aspects of viewing from hotel searching engines, data visualization tools are used to display the data clearly. There are bunch of data visualization tools from the internet. One of them is the Microsoft Power Bi. (Tate,J. 2017) states that with the use of Microsoft Power Bi, it provides ease of implementation, robust access control & security, easy to learn basis, and enables data to be more accessible. Power Bi allows user to extract data directly from the cloud or hybrid systems. So, user need not to download data from cloud and export it to Power Bi to use. It also enables to read the data in comma separated values (CSV) files and also excel file format. After uploading data to Power Bi, you can choose the data you desired to display on a graph. There are also various graphs which includes pie chart, scatter chart, stacked column chart etc. By using this visualization tool, we can make our presentations using graphics so that customers can see it clearly. The graphics includes rating over time by monthly basis, so that customer is able to track the hotel's monthly rating. This is because a hotel performance might vary over time.

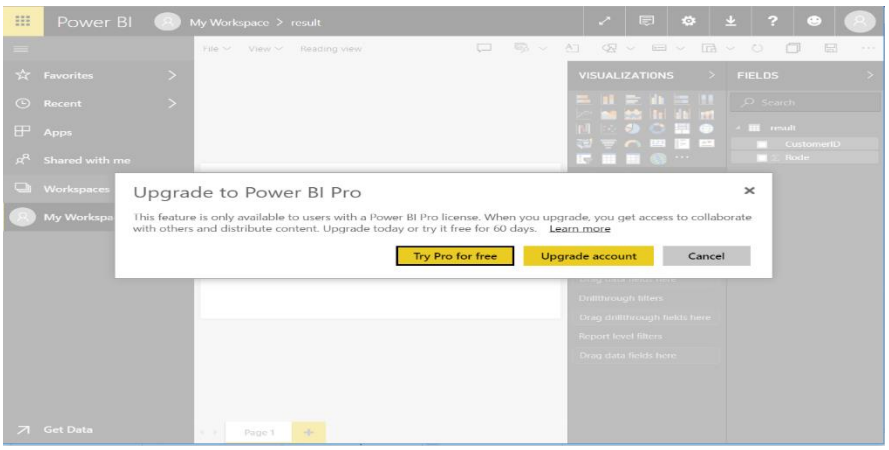


Figure 2.3: Microsoft Power Bi with limitation.

By using a powerful visualization tool, much of the works can be done easily from data visualization. However, Microsoft Power Bi requires user to pay in order for user to share the graph out for presenting. It is a good tool, but our project does not have enough budget to pay for the services. It is impossible for us to pay for the amount ourselves to complete our project. Generally, Microsoft Power Bi has the limitations that limit us on sharing the visualization data.

2.5 Zoho

To avoid spending of money for data visualization, we can use Zoho for our data visualization. It is a free software for data visualization, and it is very easy to use. Unlike Microsoft Power Bi, user can export the data as pdf to present it without paying.

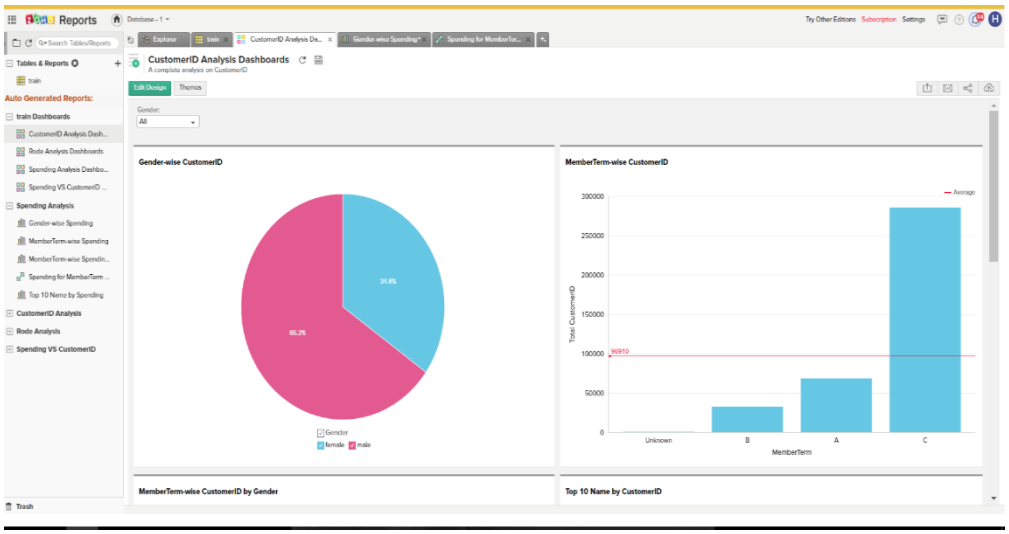


Figure 2.4: Zoho with multiple useful functions.

For example, a training set of data is insert into Zoho, after import the data into the software, it automatically appears the charts which might help user for visualization. For user who do not familiar with data visualization, it is very easy for them to adapt and learn from Zoho. Besides, Zoho do provide basic functionality of data visualization which provides pie chart and bar chart etc. Unfortunately, Zoho does not have the main function we need which is the map scatter. In our project, we want to display a 3D bar chart on the map which is the location of the hotels. We need to use it to display rating, reviews and value of money. So, Zoho is not the data visualization tool that suitable for us to develop our project.

2.6 Naïve Bayes text sentiment analysis

In this generation, there are tons and tons of comments made in social media sites every single day. Based on (Twitter Usage Statistics 2018), there are about 8000 tweets sent per second, which corresponds to roughly 700 million tweets per day. Figure is shown below:

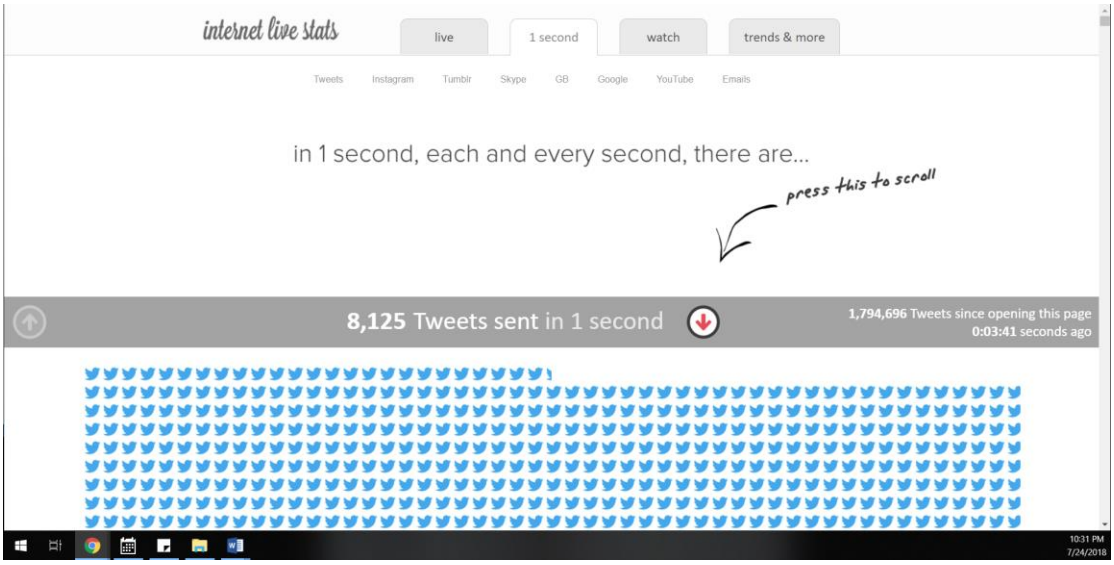


Figure 2.5: Tweets per second

It is known that every tweets sent does contains meaning. However, these huge amount of comments produce daily are impossible to let human for analysing every texts line by line to decode the meaning of the texts. So, to indicate whether the comments made are positive, neutral or negative, Natural Language Toolkit (NLTK) was implemented to allow sentiment analysis tasks to be facilitated. According to (Natural Language Toolkit n.d.), NLTK is a platform to enable human language data to be worked in Python programs. This allows sentiment analysis to work in NLTK platform.

In sentiment analysis, one of the famous machine learning approach is using Naïve Bayes techniques. (Song, J. et al. 2017) had used this kind of approach for Twitter sentiment analysis. To increase the accuracy of sentiment analysis, they proposed an approach from Multinomial Naïve Bayes to adopt two methods. First one objectives is to increase the accuracy of calculating the weights of training set which classified into positive and negative comments. While the second aims to use the average of weight differences for automatic feature selection to alter the weights of texts.

Based on their study, they did data pre-processing to remove unwanted data such as symbols, URL, email address and emoticons for a higher accuracy sentiment analysis. It is an important phase to avoid these noises from causing inaccurate of results. After that, they separate the training set into positive and negative. Each of the sets are counted separately to total up positive words and negative words. Then, they calculate the weight of the words divided in training set by using gain ratio-based approach (GRW) which assumes each attribute value is only zero or nonzero. However, there are uncountable number of attributes in Twitter data compared to fixed number of attributes of the existing feature selection approaches. Those attributes that are not in the subset are usually meaningless words and appeared more than important words. This eventually caused some meaningless words to overwhelm the important words in a single sentences and end up analysing sentiment with lower accuracy. To solve this kind of issue, they find the difference between positive weight and negative weight of all the words to adjust the weights of meaningless words. The average of differences in the training set is also calculated after that. So, some weight values are changed to zero and the meaningless words can then commendably suspended when the test set is predicted. After training and testing, their data model can achieve a maximum accuracy of 85.33% in predicting 3000 test set by having 50000 training set. The accuracy is consider great in terms of predictions. However, Naïve Bayes sentiment analysis only have the outcome of positive, or negative on the comments made. In other words, if we take negative as 0, and positive as 1. The full sentences of a single comments might also contains positive and also negative comments in the same time. For example, a user reviewing a new smartphone by comparing it with other brands. Obviously there will have some strengths and weakness of an item, and also space for improvements too. So, there should have stats to indicate a single sentence contains how much percentage of positive, negative or even neutral texts. For example, 30% of positive, 20% of negative and 50% of neutral texts in a single comments. This kind of stats will also allow data to be display in a graph to study the correlation based on dates and percentage of positivity, neutrality or negativity. There will have more variation compared to the outcome of the graph of Naïve Bayes who only have 0 or 1 as a graph display. Furthermore, even the accuracy of Naïve Bayes can be high after training, it is a machine learning approach which requires a very large data sets to train for achieving high accuracy sentiment analysis. There might also have

overfitting issues happen in machine learning approach which will leads to inaccuracy of results in other data sets.

Generally, Naïve Bayes approach is a good approach to indicate every single comments to have a positive or negative comments. However, if there are needs on analysing every texts made in a comments by percentage to plot on a graph, it is not an appropriate way to use Naïve Bayes approach.

2.7 Predicting tourists' response to an attraction using open data

The growing quantity of open data through a lot of social media platform has made the increasing of available information. For an example, based on Statista, the number of reviews in Tripadvisor is increasing steadily every year. According to (Tripadvisor, 2018), there are currently 730 millions of reviews and opinions and 480 millions of monthly average unique visitors. These open data are very helpful in analyzing tourism trends.

Based on (Open Knowledge Foundation, 2005), open data is data that can be created and pooled by everybody, everywhere for any purpose, it can be used freely for everyone. In this paper, (Anon, 2017) had taken 500 data samples from Tripadvisor to predict traveller's future choices based on their profile characteristics. Although reviews that obtained from the open data are not usually very accurate which caused by fake reviews or overrated reviewers, these reviews are more reliable than reviews that came from the particular official websites (Fotis, Buhalis, & Rossides, 2012). In their approach, to make a clearer predictions of future preferences of the travellers, they collect the rating from Tripadvisor's reviewers which consists only the scores of 0 or 5. Then, they introduced 18 features to store reviewer's interest topic such as art and architecture, eco-tourist and luxury traveller by using binary number to indicate the reviewer's profile. After that, half of the data are split into training set and the whole sample are used to determine accuracy of the classifying function. They built set of rules for training set and linked the data with projected results. Then, the prediction function is built for x times with the particular value of all the data in the sample. 2 experiments are conducted by the variation of classifying functions. Furthermore, they applied kernel trick method to effectively perform a non-linear classification which can separate the largest distance between the points.

However, the reliability of the predictions is not that desirable. In 200 classifying functions, the reliability of results is equal to 0.728. By having only 0 and 1 as an output, there are 0.5 probability for random guessing the correct output. So, the reliability of 0.728 is not reliable enough when there is only 2 output. Besides, the number of data collected is too less as we can see there are 730 millions of reviews in Tripadvisor. Thus, less amount of data will leads to overfitting to happen and can caused low accuracy in prediction when the data is unseen. In general, the approach

introduced for classification is good, however the dataset is too small for training and testing thus leads to unreliable of output results.

2.8 Forecasting tourism demand with machine learning

Tourism sector plays an important role in a country's economic. By having tourism forecasting, a country can determine the tourism trend to further improve their income via tourism sector. There are several approaches to forecast tourism demand that is used which includes neural network and time series regression.

According to (Organization of Economic Co-operation and Development, 2018) (OECD), Greece ranked the 10th tourism arrivals in 2016 in the world. This shows that the importance of tourism sector in Greece as it can bring a large income for the country. So, (Fischer, Alex et al., 2018) had study about the tourism demand forecasting in Greece which is a highly visited tourist destination in the world. Their objectives of this study is to identify the performance of having machine learning models to forecast the tourism demand of Greece by using time series between year 1996 and 2015. There are 2 approach in forecasting which consists of qualitative and quantitative approach. In their study, they used neural network with the model of feed forward network model (Bishop, 1995) for time series forecasting while logistic function is used as an activation function for this study. Using Feed forward network model is a good approach in forecasting. However, using logistic function as an activation function is not the optimum choice. This is because logistic activation is easy to saturate and it will kill gradients. When the activation saturates close to 0 or 1, the gradient will close to zero. This will affect the backpropagation process and caused the output of recurring data and kill the gradients eventually. To overcome this problem, relu activation function is a better choice at it is very light weight that allows neural network to go deeper layer effectively and lessen the computation time as the complexity of relu function is very low.

Furthermore, in neural network training process, they used 15 hidden layers to train their model. They even used Grid Search method to tune their learning rate which is [0.01, 0.02, 0.03, 0.04, 0.05]. For their validation testing, they used 10 fold cross validation due to the amount of data is not that much. They also used Support vector regression (SVR) approach to compare the outcome between neural network and SVR. They tuned 3 hyper parameters in their approach in SVR to obtain the best parameters which produces a lower test error. In their results, the difference between

neural network and SVR is very low that the outcome is nearly the same. There are not much significance difference between each approach as the dataset is not that large enough that they do not have much attributes to put into considerations. Generally, neural network can works better than SVR if the datasets is very large because neural network approach can go deeper layers to improve accuracy and the activation function needs to be change to relu activation as the complexity is lower to go through more layers.

2.9 Location recognition with geograpy3 in python

Having location information from open data, which is manually input by a user, is usually not likely to have all of the data with the same format. For example, some of the user will input the sequence of city, state, country, while for another user will only input state and country. Even some of them will only provide their country. Thus, this makes inconsistency of data. After pre-processing the user location information data, there are no guarantee that the column that store the user states and the user countries are valid or not. So, to solve the issues stated, a library in python named Geograpy by (Jonathon Morgan, 2014) is used to extract, regions, and cities from URL or text.

Geograpy is a very powerful library that consists of NLTK for entity recognition. By using NLTK, a more precise country, region and state can be identified while having a long string of texts. Another library that Geograpy used is 'newspaper' which can use for text extraction from HTML. So, by pasting an URL with articles to Geograpy, it is able to list out all the countries listed inside the article. It is very convenient for project who wants to keep track of the articles that mentioned the countries and also by the countries appeared by count too. Besides, 'pycountry' is the most important library used by Geograpy for country and region lookups. With the integrations of these libraries, Geograpy successfully create a powerful library for countries, regions and cities extraction. There are justifications that proves it can detect countries stated in articles which shown below. By inputting the URL, the list of places mentioned are in the articles are listed.

```
from geograpy3 import extraction

e = extraction.Extractor(url='http://www.bbc.com/news/world-europe-26919928')
e.find_entities()

# You can now access all of the places found by the Extractor
print (e.places)
```

```
['Media', 'Media', 'Steve Rosenberg', 'Ukraine', 'Russia', 'Ukrainian', 'Ukraine', 'Ukrainian', 'Luhansk', 'Kharkiv', 'Interi
m', 'Oleksandr Turchynov', 'Russia', 'Ukraine', 'Russia', 'Russia', 'Russian', 'Ukraine', 'Russian', 'Ukraine', 'Russia', 'Kie
v', 'Moscow', 'US', 'State', 'John Kerry', 'Russia', 'Russian', 'Sergei Lavrov', 'Ukraine', 'Russia', 'US', 'European Union',
'Russia', 'Ukraine', 'Crimean', 'Ukraine', 'Ukrainian', 'Andriy Deshchytsya', 'Russia', 'Ekho Moskvy', 'Kiev', 'Russia', 'Ukrai
ne', 'Moscow', 'Ukraine', 'BBC Moscow', 'Daniel Sandford', 'Donetsk', 'Crimea', 'Ukrainian', 'Ukraine', 'Media', 'Footage', 'Do
netsk Region People', 'Council', 'Online', 'Russian', 'People', 'Republic', 'Luhansk', 'Kharkiv', 'Ukraine', 'Ukrainian', 'Khar
kiv', 'Arseniy Yatsenyuk', 'Russia', 'Russian', 'Russian', 'Image', 'AFP Image', 'Russia', 'Kiev Image', 'AFP Image', 'Donets
k', 'Image', 'AFP Image', 'Security Service', 'Luhansk', 'Ukrainian National Security', 'Andriy Parubiy', 'Security Service',
'Valentyn Nalyvaychenko', 'Crisis', 'Viktor Yanukovych', 'EU', 'Viktor Yanukovych', 'EU', 'Kiev', 'Independence', 'Kiev', 'Inde
pendence', 'Kiev', 'Kiev', 'Yanukovych', 'Yanukovych', 'Russian', 'Crimean', 'Moscow', 'Russia', 'Eastern Ukraine', 'Arsen Avak
ov', 'Kharkiv', 'First', 'Vitaly Yarema', 'Turchynov', 'Lithuania', 'Russia', 'Ukraine', 'Luhansk', 'Kharkiv', 'Moscow', 'Ukrai
ne', 'Russia', 'Ukraine', 'European', 'Czech', 'Milos Zeman', 'Nato', 'Ukraine', 'Russia', 'Russia', 'Ukraine', 'Czech', 'Nat
o', 'Russian', 'Brussels', 'Nato', 'Moscow', 'Crimea', 'Crimea', 'Ukraine', 'Russian', 'Ukrainian', 'Kiev', 'Russian', 'Ukraini
an', 'Kiev', 'Russian', 'Ukraine', 'Ukraine', 'Russia', 'Ukraine', 'Ukraine', 'Eastern Ukraine', 'Viktor Yanukovych', 'Russia',
'Russia', 'Kiev']
```

Figure 2.6 Geograpy3 functionality

All of the places are then listed. By observing to the places shown above, we can see there are majority of places are Europe countries. However, there are also some of the strings that are not relevant to places such as media, online and people. These are the

noises that appeared by using the library. After verifying the library, it can be confirmed that for those texts that begins with a capital letter, it will be considered as places for this library as shown below.

```
import geograpy3
text_input = "Sad, Tomorrow, sad, tomorrow"
more_places = geograpy3.get_place_context(text = text_input)
more_places.countries
```

```
['Tomorrow', 'Sad']
```

Figure 2.7 Inaccuracy of Geograpy3 country determination

So, this library is not suitable for verifying the existence of the user input is a validate country as the input with capital letter will be considered as countries. This library do not help in country verifications from user input. Besides, for the users that only input their states without a country, Geograpy can be used to approximately identify the country. However, there are cities/states that have the same name in different country. For example, we all know that Brisbane is more likely to be in Australia. But there are also a city in United States that also known as Brisbane. With the list of outputs displayed, there are clearly no way to choose between the outputs.

Generally, Geograpy is powerful in approximately find the places listed in a long texts or articles. However, precise wise, it do not perform well to indicate the validity of countries on the texts inputted.

2.10 The Relation of Weather and Climate Information for Tourism

According to (Scott & Lemieux, 2010), tourism sector contributes a lot to national economics in the world. It shows that tourism sector is one of the fact that can leads to the growth of a country. To improve the field in tourism sector, they proposed the idea that weather and climate in a location will affect the tourist decision making.

Based on the study from (DRFC, 2018), they studied that according to Weather Analytics, weather affects 33% of worldwide gross domestic product (GDP). This leads to people using Big Data analytics on weather predictions as it could possibly affect the tourism sector. They also studied that weather have a strong correlations with tourism's economy. For example, to boost ice-cream sales, local industry will target on a higher temperature season so that it is more likely the customers will need it. (Susanne. B., 2010) also stated that having an understanding of potential climatic changes will helps the tourist destinations to have positive impact. On the other hand, DFRC had analysed the correlation of Crowd Analytics, temperature in weekdays and weekends by conducting the experiment at the Seoullo Bridge. The study is based on the entire winter from January to March 2018, it turns out there are more people willing to have a walk at Seoullo Bridge when the temperature is warmer which starts from 5 °C. By conducting this kind of study, a conclusion can be drawn that temperature correlates with the behaviour of the tourists. However, the temperature in Malaysia do not differs that much as Malaysia do not have four seasons. According to the data from timeanddate, the average temperature in Kuala Lumpur, Malaysia ranges between 24°C and 33°C from year 2005 to 2015. By considering lowest temperature only exists in night time, the range of temperature during the sunshine is even closer to each other. As Malaysia is a tropical country, temperature could not be seen to have much correlations between the tourist's arrivals as the temperature does not varies that much like some other countries which consists of four seasonal weather.

Generally, the fact that weather correlates with number of tourists is proved. However, this correlation analysis cannot be applied in Malaysia as the variation of temperatures are not much.

Chapter 3: System Design

3.1 Design Specifications

For the tools to use, Python will be the only programming language that is used to do this project. This is because Python have lots of library that is satisfied the needs of this project. For example, Numpy, Matplotlib, Plotly, Pandas and NLTK.

In data pre-processing and cleaning phase, the data is read from csv files using Pandas which stores the dataset inside a dataframe. Initially, there are 3 files which is Restaurant, Hotel and Attraction files respectively. Each of the dataset is labelled with their tourism type and combined into 1 csv file. Then, all sort of pre-processing and cleaning process is done on the initial attributes after understanding the meaning of attributes from the data by only using Pandas library functions. After that, new attributes are introduced by extracting data from the initial attributes. For example, the Latitude and Longitude of the tourism location and the review's length. These information are extracted to be consider for data analytics purposes.

For text sentiment analysis, to ensure the weightage of each word's is precise, (Hutto, C.J. & Gilbert, Eric. 2015) used 4 main steps for VADER quality control based on the Amazon Mechanical Turk (AMT) graders. The graders were initially required to score at least 80% of the English language tests which had a standardized college-level. Then, to make sure the rating rubric used by each independent grader is reliable, the qualified graders had to score at least 90% for matching the pre-validated mean sentiment rating of lexical items in the second phase. These items includes acronyms, emoticons, sentences and also individual words. The interface of the graders to rate is shown in (Figure 3.1 pg.16)

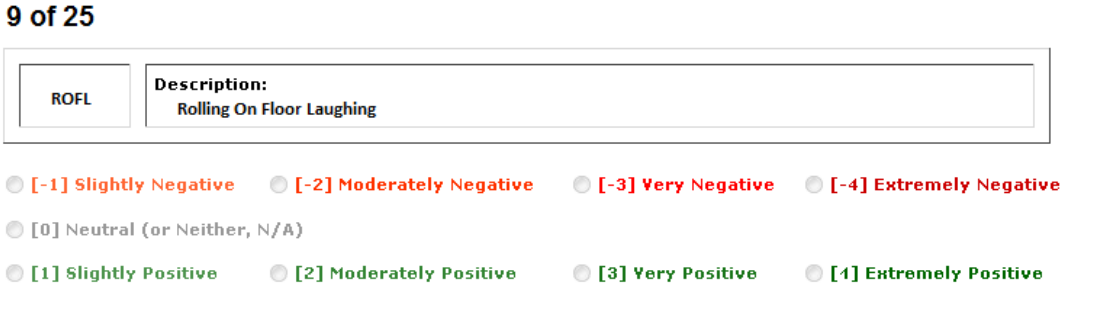


Figure 3.1: Interface implemented for obtaining valid point predicts of sentiment score for each context-free candidate feature involving the VADER sentiment lexicon.

After that, for graders that who rated have a difference of 1 standard deviation from the mean of the known distribution in more than 60% of the questions, the rating made by the graders will be disqualified to place their rating into considerations. After several filter from the graders, the final VADER program is believed to have a high precision on sentiment analysis.

Besides, VADER can categorize every sentences into negative, neutral and positive. It can provide ratio of intensity in every sentences too. For example, a sentence with 0.3 negative, 0.5 neutral, and 0.2 positive. There must be total of 1 after adding up 3 categories of text intensity. Another attribute which is known as compound is the compound of intensity of the sentences. It is ranged from -1 to 1 which indicates the intensity in compound. If the value of compound is smaller than 0, it indicates majority of negative sentences are written in particular sentences. However, if the value of compound is larger than 0, it indicates majority of positive sentences are written in particular sentences.

According to (Eru, O. and Yakin, V. 2017), the number of emoticons being used are elevating in computer mediated communication (CMC) to increase understanding and produce sense of intimacy that does not exist in the old days. Their statistics stated that there are 87% of individuals which are older than 14 years old using emoticons in CMC. This shows that emoticons are very popular nowadays. So, text sentiment analysis should include emoticons on analysis as VADER.

To verify VADER is a good social sentences text analysis tool, verification of the library is tested at (Figure 3.2 pg.18)

```
In [1]: from nltk.sentiment.vader import SentimentIntensityAnalyzer

In [2]: analyser = SentimentIntensityAnalyzer()

In [3]: def print_sentiment_scores(sentence):
            snt = analyser.polarity_scores(sentence)
            print("{:-<40} {}".format(sentence, str(snt)))

In [4]: #Start with normal sentences
        print_sentiment_scores("I just done my assignment today. When is the deadline?")

        I just done my assignment today. When is the deadline? {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

In [5]: #Add a sad emoticon behind and negative intensity
        print_sentiment_scores("I just done my assignment today. When is the deadline? :(")

        I just done my assignment today. When is the deadline? :( {'neg': 0.244, 'neu': 0.756, 'pos': 0.0, 'compound': -0.4404}

In [6]: #Using acronyms LOL which indicates as laugh out loud increase positive intensity
        print_sentiment_scores("I just done my assignment today. When is the deadline? lol")

        I just done my assignment today. When is the deadline? lol {'neg': 0.0, 'neu': 0.763, 'pos': 0.237, 'compound': 0.4215}

In [7]: print_sentiment_scores("The food is good.")

        The food is good.---------------------- {'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}

In [8]: #Whole capital letters of a positive word indicates stronger expression thus increase positive intensity.
        print_sentiment_scores("The food is GOOD.")

        The food is GOOD.---------------------- {'neg': 0.0, 'neu': 0.452, 'pos': 0.548, 'compound': 0.5622}

In [9]: #Adding ! behind a positive word express stronger expression thus increase positive intensity
        print_sentiment_scores("The food is GOOD!")

        The food is GOOD!---------------------- {'neg': 0.0, 'neu': 0.433, 'pos': 0.567, 'compound': 0.6027}

In [10]: #Adding a 'But' after a positive sentences indicates negative intensity behind. Thus, negative intensity increases.
         print_sentiment_scores("The food is really GOOD! But the service is dreadful.")

         The food is really GOOD! But the service is dreadful. {'neg': 0.192, 'neu': 0.529, 'pos': 0.279, 'compound': 0.3222}
```

Figure 3.2: Validation of VADER.

These sample of sentences proves that VADER has took emoticons, acronyms, expressions into consideration to determine intensity of sentences. So, it is suitable for text sentiment analysis on the reviews collected as it is very common for people to use acronyms and emoticons while typing.

	Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics			Ordinal Rank (by F1)		Correlation to ground truth (mean of 20 human raters)	3-class (positive, negative, neutral) Classification Accuracy Metrics		
		Overall Precision	Overall Recall	Overall F1 score				Overall Precision	Overall Recall	Overall F1 score
Social Media Text (4,200 Tweets)							**Movie Reviews (10,605 review snippets)**			
Ind. Humans	**0.888**	0.95	0.76	0.84	2	1	0.899	0.95	0.90	0.92
VADER	0.881	**0.99**	**0.94**	**0.96**	1*	2	0.451	0.70	0.55	0.61
Hu-Liu04	0.756	0.94	0.66	0.77	3	3	0.416	0.66	0.56	0.59
SCN	0.568	0.81	0.75	0.75	4	7	0.210	0.60	0.53	0.44
GI	0.580	0.84	0.58	0.69	5	5	0.343	0.66	0.50	0.55
SWN	0.488	0.75	0.62	0.67	6	4	0.251	0.60	0.55	0.57
LIWC	0.622	0.94	0.48	0.63	7	9	0.152	0.61	0.22	0.31
ANEW	0.492	0.83	0.48	0.60	8	8	0.156	0.57	0.36	0.40
WSD	0.438	0.70	0.49	0.56	9	6	0.349	0.58	0.50	0.52
Amazon.com Product Reviews (3,708 review snippets)							**NY Times Editorials (5,190 article snippets)**			
Ind. Humans	0.911	0.94	0.80	0.85	1	1	0.745	0.87	0.55	0.65
VADER	0.565	0.78	0.55	0.63	2	2	0.492	0.69	0.49	0.55
Hu-Liu04	0.571	0.74	0.56	0.62	3	3	0.487	0.70	0.45	0.52
SCN	0.316	0.64	0.60	0.51	7	7	0.252	0.62	0.47	0.38
GI	0.385	0.67	0.49	0.55	5	5	0.362	0.65	0.44	0.49
SWN	0.325	0.61	0.54	0.57	4	4	0.262	0.57	0.49	0.52
LIWC	0.313	0.73	0.29	0.36	9	9	0.220	0.66	0.17	0.21
ANEW	0.257	0.69	0.33	0.39	8	8	0.202	0.59	0.32	0.35
WSD	0.324	0.60	0.51	0.55	6	6	0.218	0.55	0.45	0.47

Table 3.1: Accuracy of VADER 3-class on classification by comparing with the rating human rates and 7 established lexicon baselines with four different domain contexts.

Based on the table shown from (Hutto, C.J. & Gilbert, Eric. 2015), the overall accuracy of VADER only varies from rank 1 to 2 with several tools and used on different platforms. It is proven to be a very high accuracy tool which only individual humans can be more accurate than it.

After all sort of data pre-processing and cleaning is done, an accurate data analytics can be made. In data analytics phase, there are few core attributes that is very important for data analytics. These are the 'Date Of Stay', 'Rating', 'TravellerType' and 'User Nation' attributes. There are many combinations based on these attributes which can produce meaningful information from it. By having time series data, the tourism trends can be determined from 'User Nation' and also 'Rating' count. We can determine which country have more users during each months or year. Besides, peak detection algorithm is used to detect the peak from the graphs based on the time series to make data analytics work more dynamically.

3.2 System Design / Overview

Flowchart

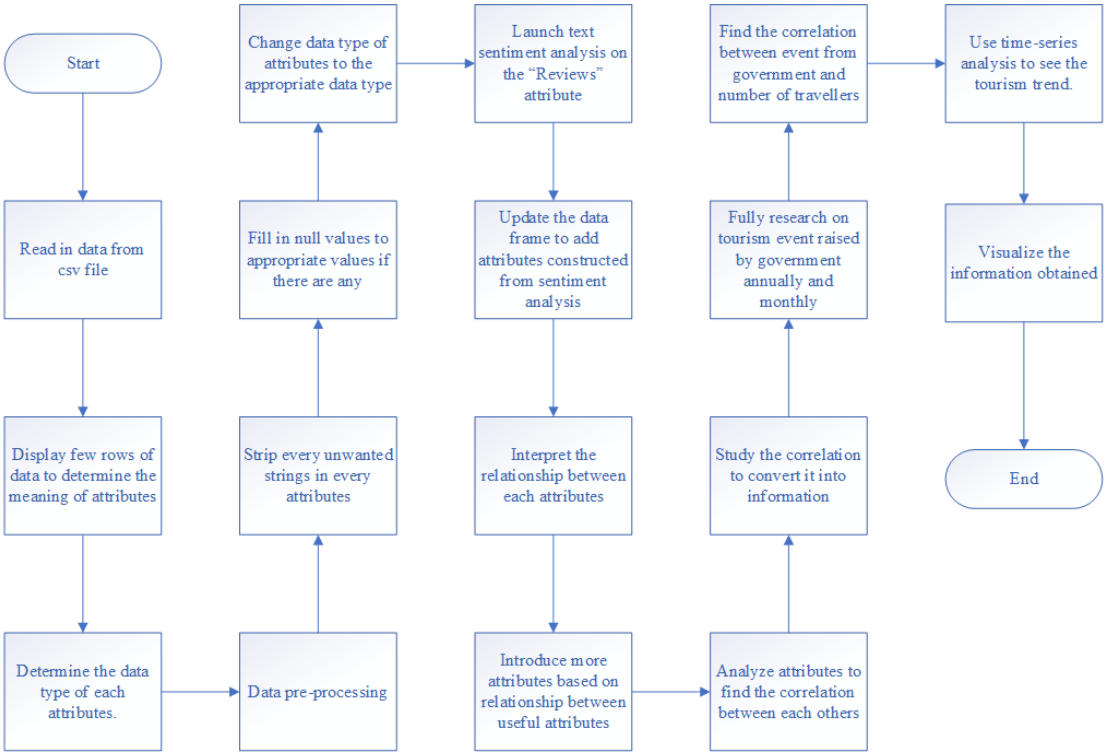


Figure 3.3: Flowchart of project's process

Based on the flowchart shown, this project is started by reading the csv file from the public data crawled to Python using pandas library so that the data frame can be formed. Head and tail of the data is displayed to understand the content of data, data type of every attributes is also displayed afterwards. After having a brief understanding of the data, data pre-processing phase starts to strip every unwanted noises in data. The data type of the attributes are also changed to the appropriate data type from the data frame. For example, date attributes is initially declared as object, it is then converted to datetime data type. As the attributes and the data set is larger, the more time would be spending on data pre-processing. According to (Munkova, D., Munk, M. and Vozar, M., 2013), data pre-processing is the phase which spends the most time in the process of discovering useful information. This is because different data have different set of attributes which requires different type of solution to deal with the unwanted or faulty values that lies in the dataset. There are only some general solutions that can be found to deal with the data for pre-processing phase. But when it comes to a special or different dataset, it requires different method and lots of procedure to be done to validate and verify the data is fully pre-processed so that the information is true while

having data analytics process. After that, text sentiment analysis is launched on the processed data to determine intensity of the reviews made by customers in the public data. Besides, external data are also collected from different sources which might be related to tourism trend such as weather, humidity, web search query and also immigration data from government.

After having more attributes, data correlation analysis can be made to study the correlation between every attributes. Then, the useful information studied from correlation analysis is visualized using graphics such as histograms and bar chart to provide an easy to understand graphs. However, by only visualizing the graph 1 by 1 is not the best way to fully see through all the patterns based on different combinations. Humans do have limitation in a certain amount of graph to visualize it and understand it. According to (Basulto, D, 2013), humans used to be a very good at recognizing patterns that involves brain neurons firing for differentiating objects since humans were young. However, there are machines that can recognize patterns better than humans now. Although much effort made by human in beating machine in data analysis, consistency and real time optimizations, there was no match between human and machine(Ferguson, A, 2018). For an example, the machine that built by IBM to play chess can recognize patterns as many as how much the chess grandmaster does. This proves human might overlook themselves in pattern recognition. So, a better approach to recognize patterns on a huge dataset is using machine learning approach which can outperform human.

So, peak findings algorithm is used to find the peak of the graphs. The peak findings algorithm that is used in this project is from PeakUtils library by (Lucashn, 2016). PeakUtils peak finding algorithm provides 1 dimension peak detection. By using PeakUtils, the function allows to tune the hyperparameter to detect each peak with minimum height and also distance filtering. In this project, PeakUtils is worked together with Pyplot to plot the graphs more interactively.

3.3 Implementation Issues and Challenges

3.3.1 Data attributes and data size challenges

The main issue of this project is the limited number of data attributes. Due to crawling public data is the only way to obtain data, there are less attributes to collect. This is because public data do not have any customer's personal details such as gender, age, height and weight. With less attributes, correlation analysis that can be made will be lesser. To overcome this issue, obtaining private data can increase the number of attributes for data. There are 2 type of ways to get private data which is request from government for government private data, or creating a mobile application which allow users to register for obtaining their personal details dynamically. Unfortunately, government data is not that easy to request for it. This is because of private and confidential issues and some data requires us to pay for it. For mobile application, it requires much time to develop a good one so there will be users to use it. It is also very dependent on quantity of users to collect large data sets. This issue leads to the challenge for this project which is increase data attributes as much as possible.

Besides, the other issue for this project is the number of data is not large enough for correlation analysis. According to (Zamboni, J. 2018), larger size of data will have higher accuracy mean values. In other words, less data will leads to lower accuracy of data. This is because the mean values have a lower accuracy, especially when there is existing outliers and misleading the statistics in a data set. As we set our scope to analyse tourism data in Malaysia, but the data that we currently have is only in Kuala Lumpur. This leads to smaller picture that can be seen as we can only focus to analyse the data in Kuala Lumpur.

3.3.2 Data pre-processing and cleaning challenges

Other than size and attributes of dataset, the challenge for this project is to create an accurate correlation analysis by using the number of data that are currently prepared. The phase of data pre-processing and data cleaning is important to overcome this challenge to avoid unwanted noises appear in the data. This project is challenging as the data pre-processing and cleaning phase consumed most of the time to ensure a clean data is processed before analysing. According to (Munkova, D., Munk, M. and Vozar, M., 2013), data pre-processing is the most time consuming phase in the process of analysing useful information. Without data pre-processing and data cleaning, even though data analysis can be made, but it would not be reliable as there are unwanted

data and outliers that need to be clean and some of the data attributes need to be pre-processed to suitable format before visualized.

3.3.3 Data analysis challenges

Besides, there are no exact direction to find out a suitable way to analyse a data within a short period of time. This is because there are many external data that can be put into considerations on data analysis for tourism and the area is very wide. For example, air humidity, air pollution index, safety index and quality of life index can be taken into considerations because this data are closely related to the tourism trends.

Other than that, there are also seasonal events, annual events and also events raised by government in Malaysia such as Visit Malaysia 2014 that are also related to tourism trends to be collected to determine the impact of the tourism events to the number of arrivals and positive experience to Malaysia. The information of other countries visiting Malaysia is also important to determine the underlying trend based on particular country. For example, the semester break for a particular country and the 4 seasons' temperature in their country that will leads to travellers' arrivals to Malaysia.

With this wide field, it is very challenging to determine the tourism trend to find out the correlations between each other. This is because it can be very specific on a trend of particular tourism location, or very general of trend on an area such as determining arrivals based on country.

Chapter 4: Data pre-processing and cleaning

4.1 Attributes meaning from data collected

Attributes	Meaning
Type	Type of tourism. e.g.: Restaurant, Attraction, Hotel
Low Price	Restaurant lowest price range indicator.
High Price	Restaurant highest price range indicator.
Ranking	Hotel's stars
Name	Name of the restaurant/hotel/attraction.
Date	Date reviewed by the user
Rating	Rating given by user
Review Length	Number of words written in the review.
Country	Tourism location country
State	Tourism location state
City	Tourism location city
Postcode	Tourism location postcode
Date Of Stay	Date of experienced by the user. Month and Year
TravellerType	Type of travellers. E.g.: Couples, Families, Business, Solo
Compound	Sentiment analysis score from reviews
Response Date	Date of respond by the party
User State	User's origin state
User Nation	User's origin country
Contribution	The number of reviews that made by particular users in the website.
Helpful Votes	The number of votes that provided by other users to support the reviews made.
Helpful Ratio	(Helpful Votes/Contribution)
Latitude	Tourism location latitude
Longitude	Tourism location longitude

Table 4.1 Meaning of data attributes

4.2 Datetime data type attributes format

Firstly, after obtained data using web crawler to get public data from tourism recommendation websites, the next step is to do data pre-processing. Every attributes needs to be pre-processed to make sure the correctness of the data can be kept. Otherwise, the information learned from data analysis that shown would not be accurate. Initially, the data crawled for attribute dates are in string data type, it is also has unwanted string in front of the attribute which the whole attribute appeared as "Reviewed 12 April 2017". Furthermore, for the reviews that are recently added within 1 month, it will appeared as "Reviewed 3 days ago" for within 1 week, and "Reviewed 2 weeks ago" within 1 month. So, this kind of attributes which does not have the same pattern requires more steps to handle it. To solve the problem, "Reviewed" string is stripped away first, so the normal case which has exact date of reviewed date can be change to date format easily. After that, the "weeks" and "days" need to be classified because "weeks" needs to multiply by 7 whereas "days" is not needed. Then, current date time is used to deduct the "days" and "weeks" ago to determine the date reviewed, and the date reviewed is then replaced to the column. There do have a minor problem for the "weeks" reviewed because the exact date reviewed could not be traced but only the exact week reviewed could be traced. So, there might have slightly difference of the exact date reviewed by customers if the customers reviewed it few weeks ago. Besides, the format of the date which deduct from current date to days/weeks ago is not with the same format with the usual one. Although it does not displayed in the database, but it prompts error when reaching the modified date. This is because the date time format is converted to display only seconds as minimum. While the different one stores milliseconds too. So, it is stripped to maintain the consistency of data. The same format is also used to solve the "Response Date" and "Date of Stay" attributes. "Date of Stay" attributes only consists of month and year provided, while the datetime format can only accept day, month, year format. So, 1 is added to the "Date of Stay" attributes to indicate it as the day of the attributes. There are null values in "Date of Stay" and "Response Date" because some of the users do not input the date that they went for travel. For "Response Date" is indicated as the particular site organization's reply. Some of the organizations do not reply to the review from the customers, so there are null values too. For a highest precision data analysis, the date of review is not likely equals to the date of stay of a travellers. This is because a traveller might write a review based on his/her experience after a week or maybe after a month. This inconsistency leads to

differences of months between date of review and date of stay. So, attribute "Date of Stay" is used mainly in data analytics. For the null values that lies in "Date of Stay", the values from "Date" attribute is then appended to "Date of Stay". This is the highest accuracy that we can reach based on the data we have currently. While this is open data, the highest accuracy that we can reach based on the date is only by month but not week or exact date. This is because this open data do not provide the exact date of the travellers' experience but only provide the month.

4.3 Origin of country and state of reviewers

"User Location" from the data is indicated as the place that the user came from. It is an important data so that we can know the preference of foreign travellers or local travellers for travelling in our country. However, the input by the users' review do contains many variations, which leads to inconsistency of data. Some of the users might input the format as city, state, country, while some others users might input only country, or only state. There are also people who used short form of the country to the input. In this case, the data in this attribute are very messy and very hard to be grouped because the same states and countries travellers might having different input format. So, to further reduce the inconsistency input of data, this attribute's data are comma delimited separated to place in different column. After that, the 1st column of the 2D array are used to store in a new dataframe introduced as "User State" because we know that the user will input by the sequence of city to state to country. The last column of the array are used to store in as "User Nation". However, there are also null values in the 2D array, because there some users only input without comma, while some users input it with few commas. We know that the size of 2D arrays need to be kept consistent for their width and height. So, if the users only have input without comma, the input will only falls in the 1st column while the subsequent columns will be null values. It can be described as the illustration below.

User A input = Bukit Mertajam, Pulau Pinang, Malaysia

Bukit Mertajam	Pulau Pinang	Malaysia

User B input = Kuala Lumpur

Kuala Lumpur		

As we can see, if the maximum of comma is 2 in the "User Location" column, it will separate into 3 columns to store the input, whereas for the input that do not have comma, only the 1st column have the input. So, what we can do is check for null values that lies in subsequent columns, if it is null value, the previous column of the values will be copied to the next column until the end of the column which described as below.

User A input = Bukit Mertajam, Pulau Pinang, Malaysia

Bukit Mertajam	Pulau Pinang	Malaysia

User B input = Kuala Lumpur

Kuala Lumpur	Kuala Lumpur	Kuala Lumpur

The best case is the user input states, country with the correct format, while the unfortunate case is the user only input the country they are from. This is because we can track user's country based on the states they inputted, but we cannot track user's state based on the country they inputted. After storing "User State" and "User Nation", 2 columns are used to compare if they both share the same strings. If both strings are the same, and it appeared to be a state/city, the country of the state/city will be changed into to replace the current string in the column "User Nation". There are few libraries that can be used in python to find out their country based on the city/state input such as Geograpy3. However, this library do not support well in determining the country from state/city. For Geograpy3, it is good at determining Europe countries based on city/state given. However, it do not know well the city/state in Malaysia. It can't recognize Penang, Petaling Jaya and etc. Besides, there are city/state that have same name but locates at different country. For example, Brisbane is commonly known at Australia, but there are also a place in America called Brisbane. Geograpy3 can list out all of the countries that have a city named Brisbane. But what we want is usually the highest possibility which is from Australia. We do not know the list of the output will contain how many countries and it would be undesired to just randomly pick 1 of the country in the list. So, it is not suitable for us to use this library for determining country. To solve this problem, the "User Nation" and "User State" which are having the same strings are printed out in a descending order based on the count.

```
raw_data['same'].value_counts()
```

None	39939
Kuala Lumpur	2525
Malaysia	2284
Singapore	2240
Kl	809
London	772
Australia	546
Sydney	542
Melbourne	484
United Kingdom	340
Perth	333
Jakarta	258
India	256
Uk	232
Hong Kong	223
Brisbane	214
Mumbai	200
Indonesia	195
Philippines	195

Figure 4.1 Frequency of similar strings that lies in "User State" and "User Nation"

After that, based on the list, every "User Nation" are changed manually by determining the state input is from which country. For example: "User State" and "User Nation" both are having the input of Kuala Lumpur. We know that Kuala Lumpur is in Malaysia. So, we change the "User Nation" string into Malaysia. By using this approach, most of the "User Nation" can be modified to correct input. However, to further solve the problem dynamically, a library can be created to match the lists of "User State" to have a correct "User Nation" for the purpose of long term usage.

4.4 Restaurant's price range and hotel's star

Besides, there are also data from Restaurant which have dollar sign '$' to indicate the price of the restaurants. In a restaurant, there do have a list of foods with different price range. So, the price indicator from the data collected do have a range too. For example, "$ - $$$" or "$$ - $$$$". More "$" indicates higher pricing based on the restaurants. To store the price range of restaurants, 2 more columns are introduced which is "Low Price" and "High Price" to store the lowest price range and the highest price range. For the restaurants that do not have the price range such as "$" and "$$$" in a particular restaurants, both columns will store the same values. These dollar signs "$" are then converted into numeric form such as from "$$" to 2. By converting into numeric values, these data can be used in data analytics. Other than restaurant price indicator, "Ranking" attribute stores the number of stars of a hotel. It is then normalized to the range between 1 and 5. While there are null values in either hotel and restaurants that do not have a stars or prices indicator, supervised learning method can be used to predict the stars and the prices of the hotel or restaurants that poses null values.

4.5 Precise location data from tourism location

To have an accurate data analysis, a precise of the particular tourism location is essential. An area of particular tourism location is defined using the same postcode. Initially, there are lots of location that do not have a postcode, either it is null value, or it is filled in with words. This leads to inaccurate grouping when we want to group the location based on postcode to determine the trend. So, a python library which named geopy is used to determine the postcode of a particular tourism location. Geopy is a library that can determine the full address of a location after inserted by users. Besides, the latitude and longitude of the particular tourism location can also be found by using geopy. It is easier for geopy to detect the full address of a location if the popularity is higher. So, if the location cannot be detected based on the text input, the text input must be more generalize. For example, based on the data that we collected, we have the location's name, state, country, and postcode. If the location could not be detected, considering not using location's name as input while keep the remaining 3 as attributes for it to determine the location. This can maximize the location's precision on the data. While for the rows that do not have a proper postcode, after having the location determined by geopy, extract the location's postcode from the address from geopy by coding a function to find 5 consecutive integers and print it out as we know postcode consists of 5 consecutive integers which lies in the address string. However, there are some of the address that couldn't be detected by geopy with postcode. So, we have no choice to leave the unfound postcode to null value. With this function, it maximize the accuracy of postcode as majority of the location's faulty or null valued postcode can be found and modified. After maximizing the accuracy of postcode, the accuracy of the latitude and longitude of particular tourism location can be maximized too as the text input to geopy are the same for it to detect the address. With latitude and longitude of the location, the highest precision of the location can be plotted. It can be used to find the exact distance from 1 location to another. It can also be used for visualizing the trends based on the maps by displaying a heat map on it.

4.6 Handling reviews attribute and sentiment analysis

For reviews attribute, due to the length of reviews made by customers might be longer than default size set by the travelling recommendation websites, there do have a toggle button which is “Show More” to click for displaying full text made by customers. The designed web crawler managed to click the toggle button automatically to crawl full text of the reviews. However, “Show Less” appeared as last words for the texts and it is crawled and stored into the database. Due to the reviews are going to go through text sentiment analysis. The words “Show Less” might affect the accuracy of the reviews. This is because for NLTK, “Less” is indicate as negative. So, the “Show Less” in all reviews are stripped away to avoid inaccuracy of text sentiment analysis. Due to limited attributes as these are open data, to increase attributes, text sentiment analysis is used to analyse the reviews made by the customers. There are 2 kinds of approach in text sentiment analysis which is machine learning approaches method or lexicon approaches, both also using NLTK library in Python. There are plenty of time spent to compare whether which is a better way in this project. The main difference is lexicon approach has a ‘dictionary of sentiment’ which already built by others, it will map words to sentiments to determine the intensity of words, while machine learning approaches requires to train a model which is very dependent on volume of data in order to have a high accuracy prediction. In this project, the volume of data that crawled is not considered as large. So, lexicon approaches is used in this project which is VADER sentiment analysis. This is because the defined library is proven to be very accurate as (Table 1.1 pg.18) shown. Even emoticons, acronyms and capital words will also take into consideration to determine intensity of sentences. In the reviews from customers, there do have probability for them to make comments using acronyms and emoticons. So, VADER is used to work well in these reviews from customers. After that, the intensity of the reviews are updated to the database for adding the attributes. These attributes include Negative, Positive, Neutral and Compound. While Compound is the average based on the calculation from Negative, Positive and Neutral, we can just use Compound to analyse the data. Due to the data of Compound are centred, normalization are used to scale the output to range from 1 to 5 which are having the same range as Rating to further justify the correlation between Rating and Compound. After normalize the Compound, we calculate the inaccuracy scores between Rating and Compound which is calculating the absolute differences between each other (lower is better).

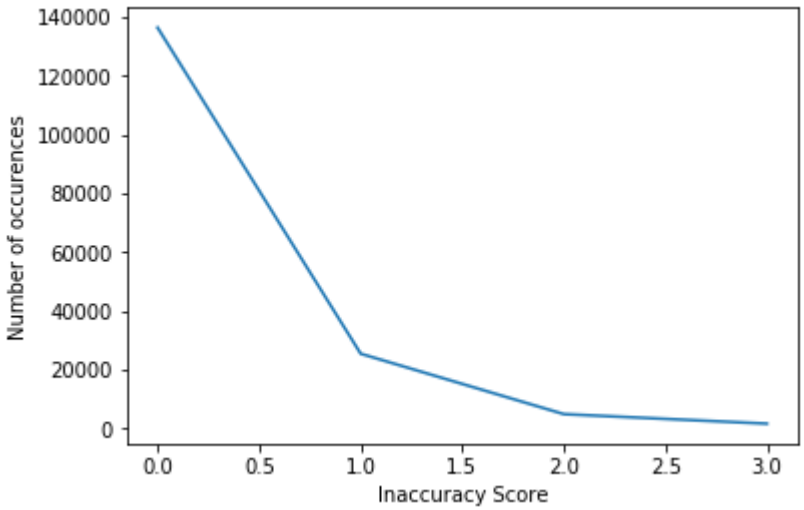


Figure 4.1 Number of occurrence of absolute difference between Compound and Rating

As we can see, majority of the absolute difference between Compound and Rating falls within 1 and 0, which shows that Compound and Rating is directly proportional with each other. Other than sentiment analysis on reviews, the word count of every reviews are stored in "Review Length" so that we can find out the relation between Review Length and Rating.

4.7 Outcome of attributes after data pre-processing and cleaning

Before:

```
Address          object
Comment          object
Contribution     object
Country          object
Date             object
Date Of Stay     object
Helpful Votes    object
Local            object
Name             object
Postcode         object
Price            object
Ranking          object
Rating            int64
Reply            object
Response Date    object
State            object
TravellerType    object
Type             object
Unnamed: 0        int64
User Location    object
Username         object
dtype: object
```

After:

```
Unnamed: 0              int64
Type                   object
Low Price             float64
High Price            float64
Ranking               float64
Name                   object
Date           datetime64[ns]
Rating                float64
Review Length           int64
Country                object
State                  object
City                   object
Postcode              float64
Date Of Stay   datetime64[ns]
TravellerType          object
Compound              float64
Response Date  datetime64[ns]
User State             object
User Nation            object
Contribution          float64
Helpful Votes         float64
Helpful Ratio         float64
Latitude              float64
Longitude             float64
dtype: object
```

Figure 4.3 Differences of data type before and after data pre-processing and cleaning

Initially, most of the data are classified as object as there are unwanted strings inside the data. After pre-processing, every attributes that should be numeric values are converted from object data type to numeric data type such as integer or float. This phase is very important to allow further data analytics.

Chapter 5: Data analytics

5.1 Correlation analysis on time series

After data pre-processing and cleaning, data analytics phase is used to determine the tourism trend. Initially, every external data that might have relation to tourism trend are collected from different trusted sources such as government, Numbeo, Google Trends and timeanddate. From these sources, we collected data which is annual arrivals of travellers to Malaysia, monthly average temperature, humidity, pressure at Kuala Lumpur and quality of life index, purchasing power, safety index, health care index, cost of living index, property price to income ratio, traffic commute time index and pollution index. There are few graphs plotted to determine the correlation between each attributes after all of the data are normalized using min max normalization before plotting it. Then, the most related correlation graph is plotted which are related to currency of MYR, Quality of life index and number of reviews count.

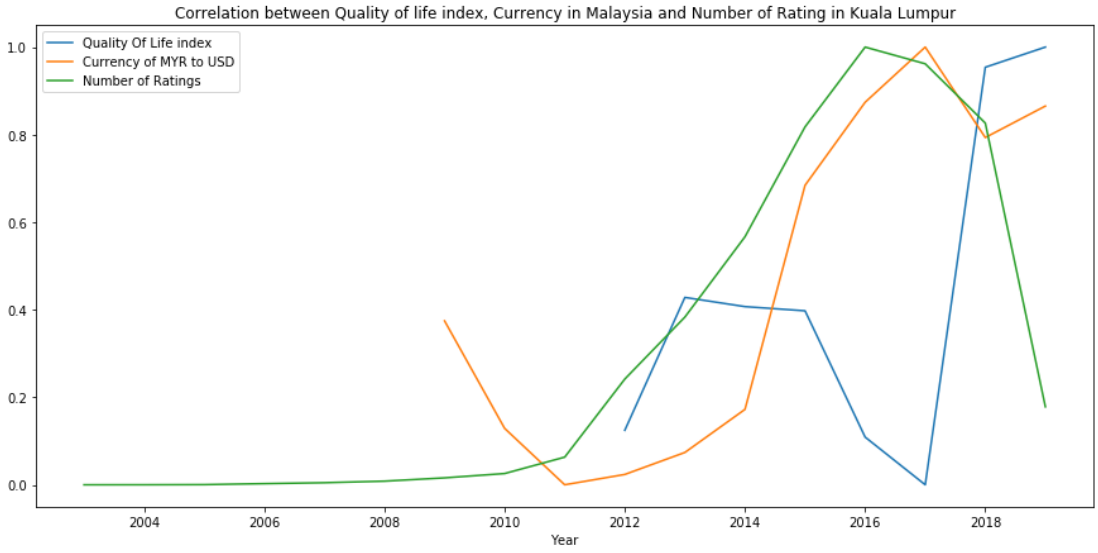


Figure 5.1 Correlation between Quality of life index, Currency in Malaysia and Number of Rating in Kuala Lumpur

As we can see from the graph, the quality of life index is inversely proportional with the reviews count based on year count. This can be imply that the higher the quality of life index in Kuala Lumpur, the lesser the travellers to visit Kuala Lumpur. Besides, we can see that the currency of Malaysian ringgit to US dollar is directly proportional with reviews count. This is because when the currency of MYR drops, the exchange rate of foreign currency to MYR increases, and for foreign travellers, it is cheaper for them to travel to Malaysia during this period. Generally, the lower the currency of MYR leads to lower quality of life index in Kuala Lumpur and it leads to the higher reviews count.

5.2 Data analytics based on country

To determine the reviews are come from which country's travellers, the number of reviews made based on different country's travellers are plotted. Due to there are too many different countries, there are only top 10 countries based on the reviews count are visualized.

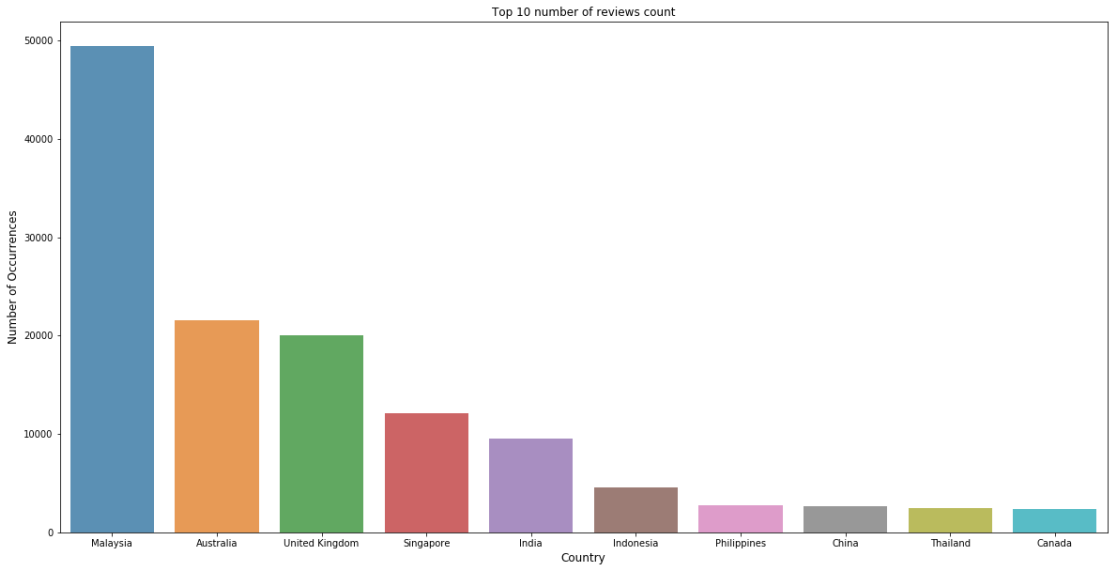


Figure 5.2 Top 10 number of reviews count based on different country's travellers

It is clearly can be seen that majority of the reviews are made by Malaysian. It is then followed by Australia, United Kingdom and Singapore. However, it does not correlated with total arrivals to Malaysia based on the government's data stated.

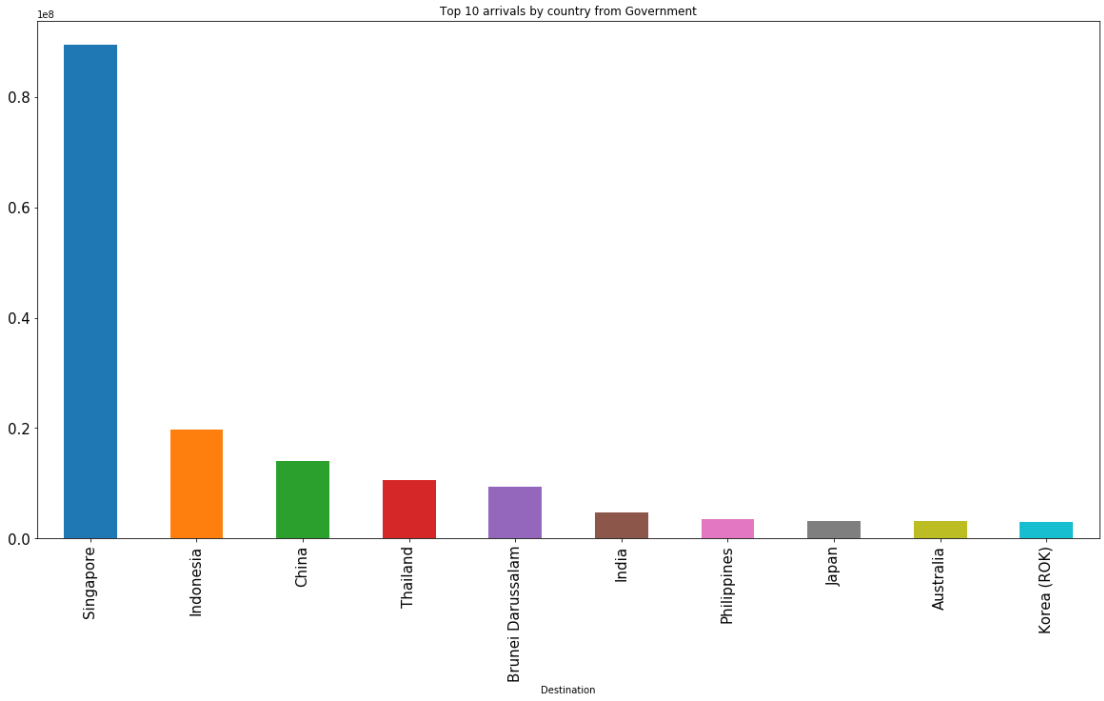


Figure 5.2 Top 10 average arrivals on different country from government's arrival data

Based on the reviews collected, Australia is ranked 2nd at the total number of reviews, however, it appears to be the 9th at government's arrival data. While for United Kingdom which ranked 3rd from the data we collected, it does not even ranked top 10 arrivals to Malaysia. This can be infer that the European cultural countries are more likely to travel at Kuala Lumpur, or they are more likely to share their opinion and experience in public as references for the others. Whereas for the Asia country such as Indonesia, China and Thailand, travellers are not likely to travel to Kuala Lumpur, or they do not have much interest to make a review. For Singapore, which ranked the 1st at government's arrival data, the number is greatly larger than any other country's arrivals. This might be caused by Singaporeans are most likely to travel to Johor which is very close to their country and the exchange rate of SGD to MYR is very high that causes them to frequently visit to Malaysia.

After that, we want to study about the reviews count from particular country based on month as shown below.

5.2.1 Reviews count in Australia based on month

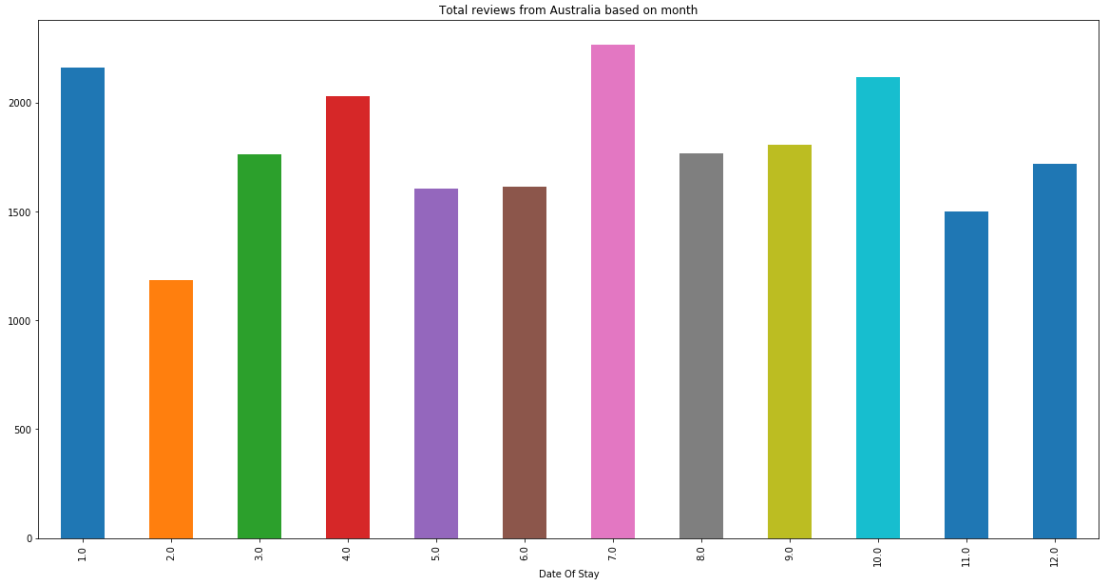


Figure 5.4 Total reviews from Australia based on months

Based on the graph shown, it can be seen that during January, April, July and October, the reviews count are higher compared to the other months. Although it can be visualized with human eyes, it is not desirable to compare the graphs sequentially without any peak indication. So, a peak detection graph is plotted using library from plotly for finding the peak points in the graph. This can further lessen the burden for human to make comparison.

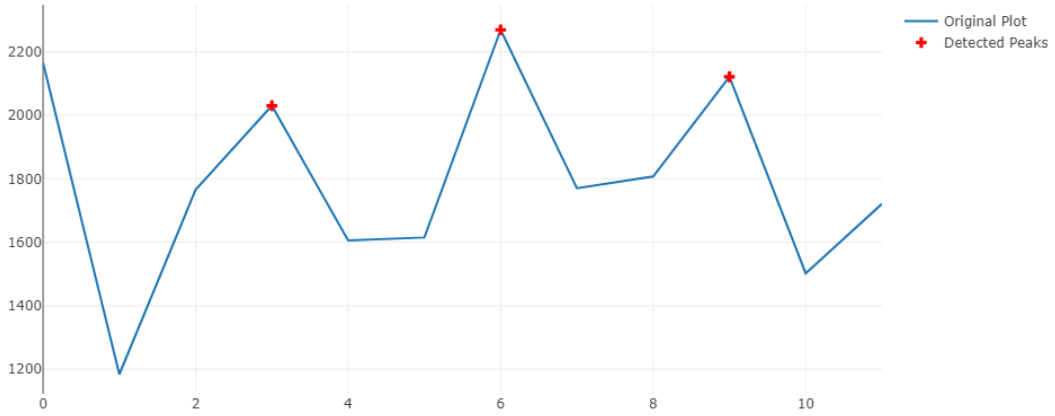


Figure 5.5 Total reviews from Australia based on months with peak detection

Same graphs based on different years are then plotted to justify the consistency of peak months.

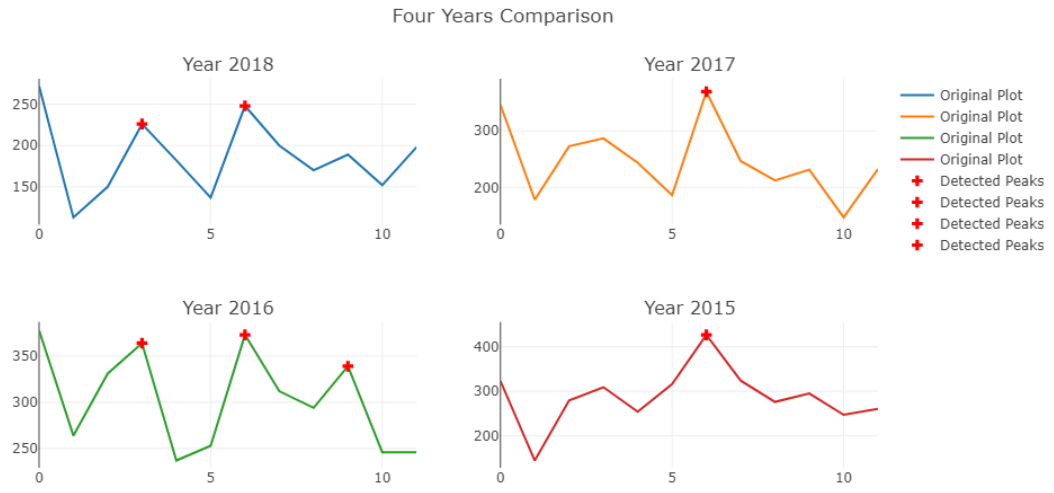


Figure 5.6 Reviews count from Australia based on months with different Year's data peak detection

Based on the graph shown above by yearly basis, it clearly shows that during January, April, July and October, the number of reviews count are higher than the others. This implies there are certain reasons lying behind the pattern of the reviews count from travellers in Australia. To further study the reason about the peak of these months, the 4 seasons of Australia is taken into considerations. Based on Australia's tourist website, December to February is summer season, March to May is autumn, June to July is winter and September to November is spring season. Based on these 4 seasons, we can't really see the trends of peak arrivals from Australia. This is because all of the peak arrivals falls between the middle of each season. So, the school holidays of Australia are taken into account to determine the correlation of the peak arrivals from Australia to visit Kuala Lumpur. According to Australia's tourism website, there are 4 terms of school holidays in Australia which falls on April, July, October and January. Thus, it can be clearly seen that during Australia's school holidays, they will be more likely to visit Kuala Lumpur. By having this prove, it further justifies that the number of reviews count are correlated with the number of arrivals to Kuala Lumpur.

5.2.2 Reviews count in United Kingdom based on month

The same graphs are also plotted to determine the peak months of United Kingdom.

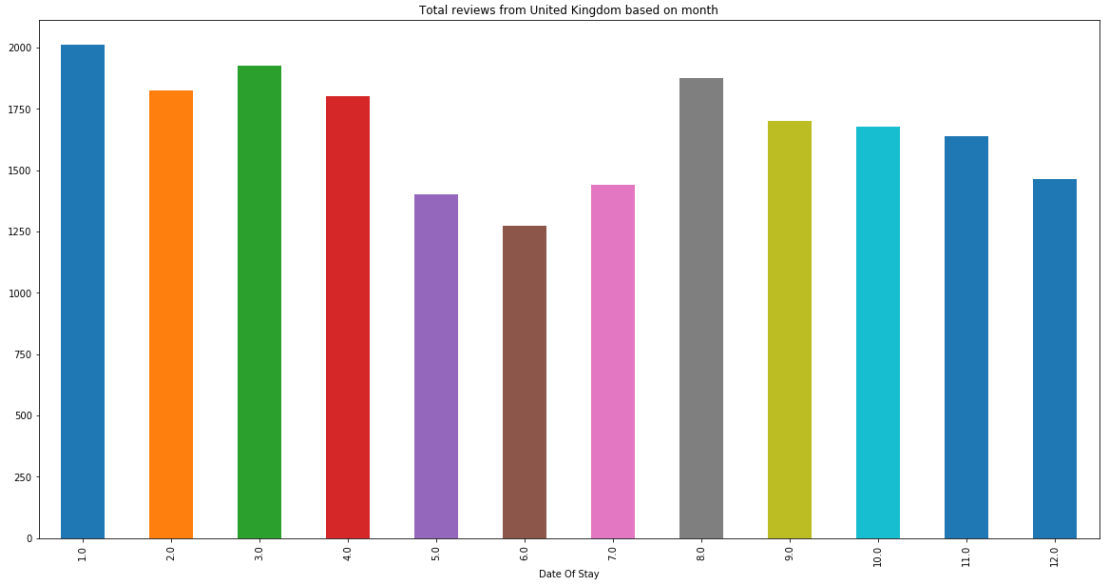


Figure 5.7 Total reviews from United Kingdom based on months

Based on the graph shown above, we can see a wave like pattern graph for the total reviews count in United Kingdom. It begins high from January and slowly decreases until June, and from June grows gradually until August, and decreases slightly until December. To further visualized, peak detection graphs are plotted for justifying the subsequent months do have the same graph patterns.

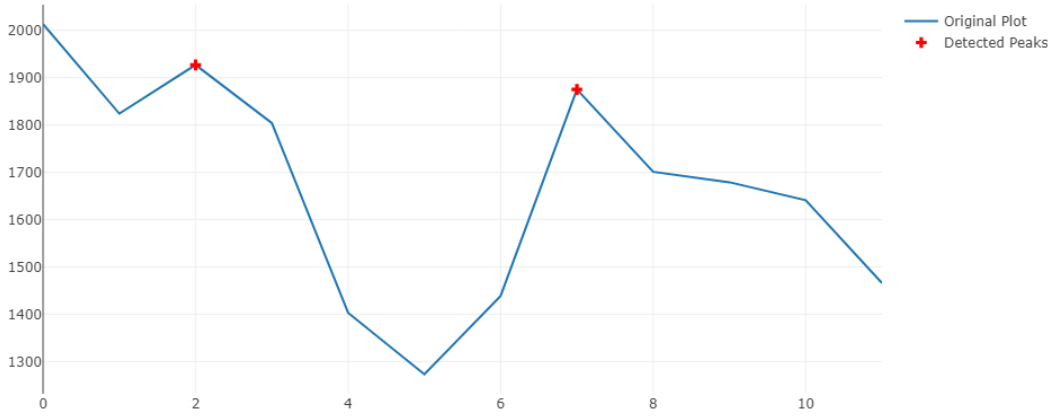


Figure 5.8 Total reviews from United Kingdom based on months with peak detection

In this case, we took average temperature of United Kingdom from (HolidayWeather) to compare with the reviews count based on months.

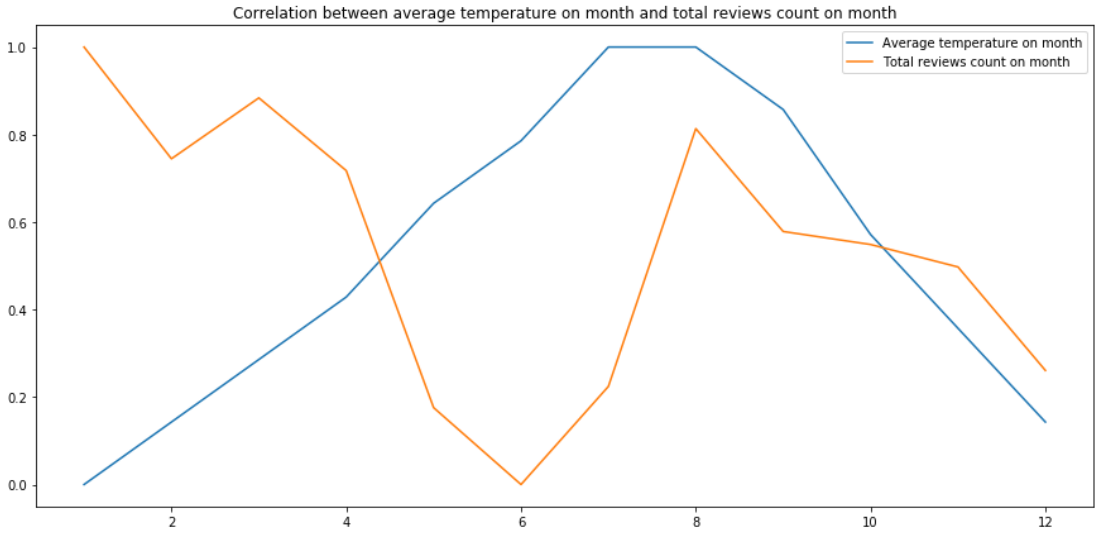


Figure 5.9 Correlation between average temperature on month and total reviews count on month in United Kingdom

As we can see, the average temperature is inversely proportional with reviews count from January to July. Whereas is proportional with from July to December. Their correlation score is -0.41686 which shows that the reviews count from United Kingdom based on months might due to the average temperature in United Kingdom. After that, graphs based on different years are then plotted to justify the consistency of peak months.

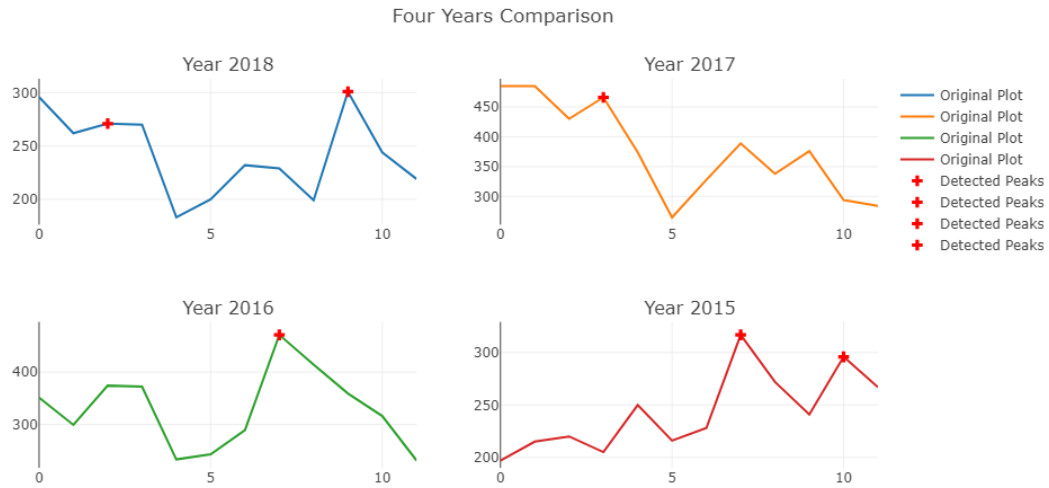


Figure 5.10 Reviews count from United Kingdom based on months with different Year's data peak detection

According to the graph of four years comparison peak finding graphs, there are no consistent peak based on months found in the graphs. Thus, no peak conclusion can be

drawn. However, based on the four graphs, it can be seen that during there are consistent trough appears which falls between May and July (4 to 6 in the graph). According to (Seasonsyear), the four seasons of United Kingdom from March to May is Spring season, June to August is Summer, September to November is Autumn and December to February is Winter. This shows that the trough appears is in between Spring and Summer season. However, it does not correlates with the trough in reviews count. For the average temperature in United Kingdom based on year's month, the data can be found but it requires payment. So, further analysis could not be made and the underlying reasons to have that consistent trough is unable to determine.

5.3 Data analytics based on particular tourism location

By further magnifying the scope to more specific location, every improving or weakening tourism location can be found by comparing the reviews count from the previous years. By comparing the differences of reviews count based on a particular tourism location between year 2017 and 2018, the best and worst performance tourism location is visualized. By referring to the graph, we can see that the best performance tourism location in year 2018 is a restaurant named "Dinner In The Sky Malaysia".

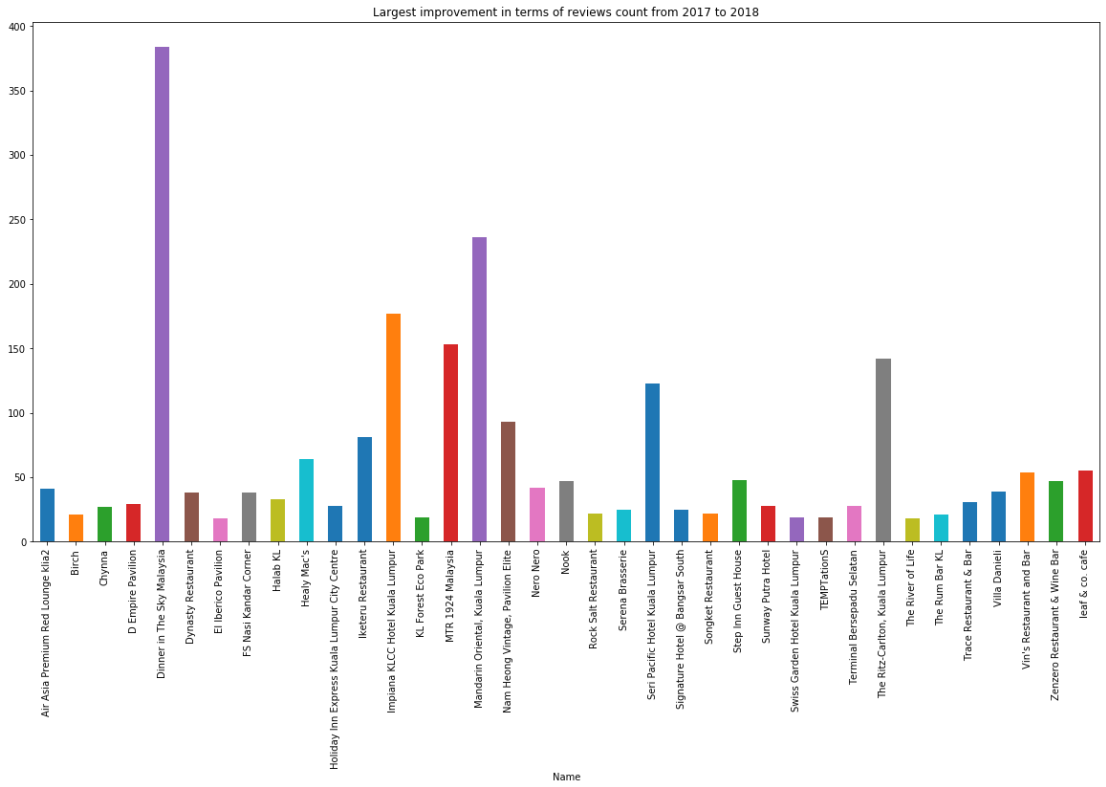


Figure 5.11 Largest improvement in terms of reviews count from year 2017 to 2018

In year 2018, we can see the huge improvement of “Dinner In The Sky Malaysia” compared to the others. To further investigate the peak result of “Dinner In The Sky Malaysia”, we plot the graph based on monthly reviews count to determine the peak months that caused “Dinner In The Sky Malaysia” to have such huge improvement.

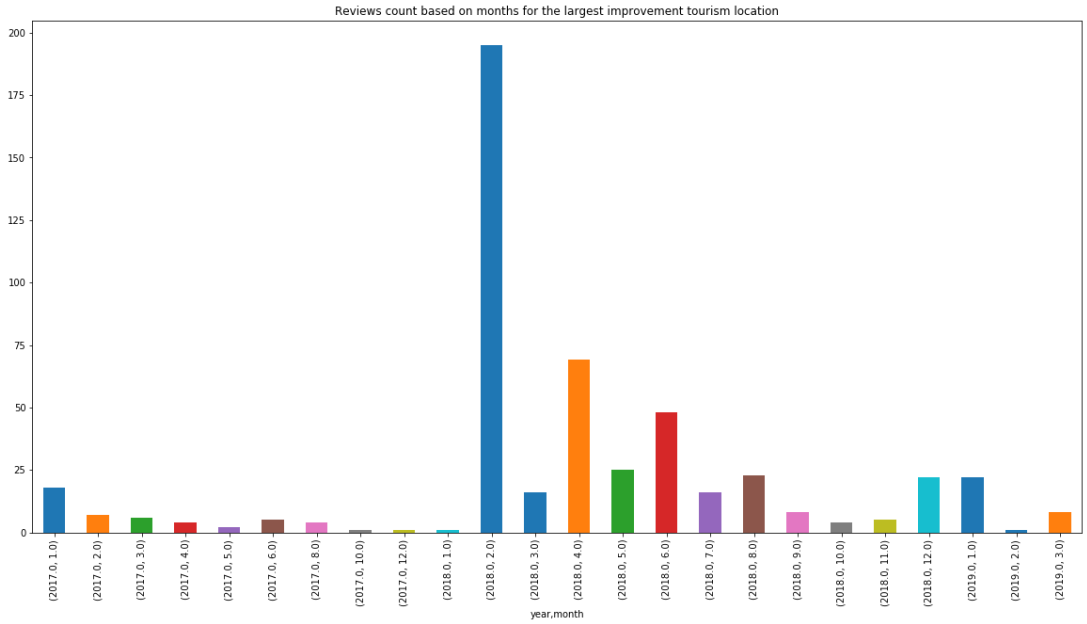


Figure 5.12 Reviews count based on months for the largest improvement tourism location

By visualizing the graph on monthly basis, we found out there are huge performance improvement on the review counts at February 2018. In first thought, it might because of some seasonal events that promote such improvements as if there are no any errors in this public data. However, in the subsequent February 2019, the review counts are under performed compared to February 2018. It should not have such regression if the seasonal events are annual events. In this case, we want to know which will be the traveller type that mainly reviews this particular tourism location in February 2018. We also need to find out the reasons of the regressions of February 2019. Firstly, the traveller type of reviewers is visualized at February 2018 to determine the majority traveller type at February 2018.

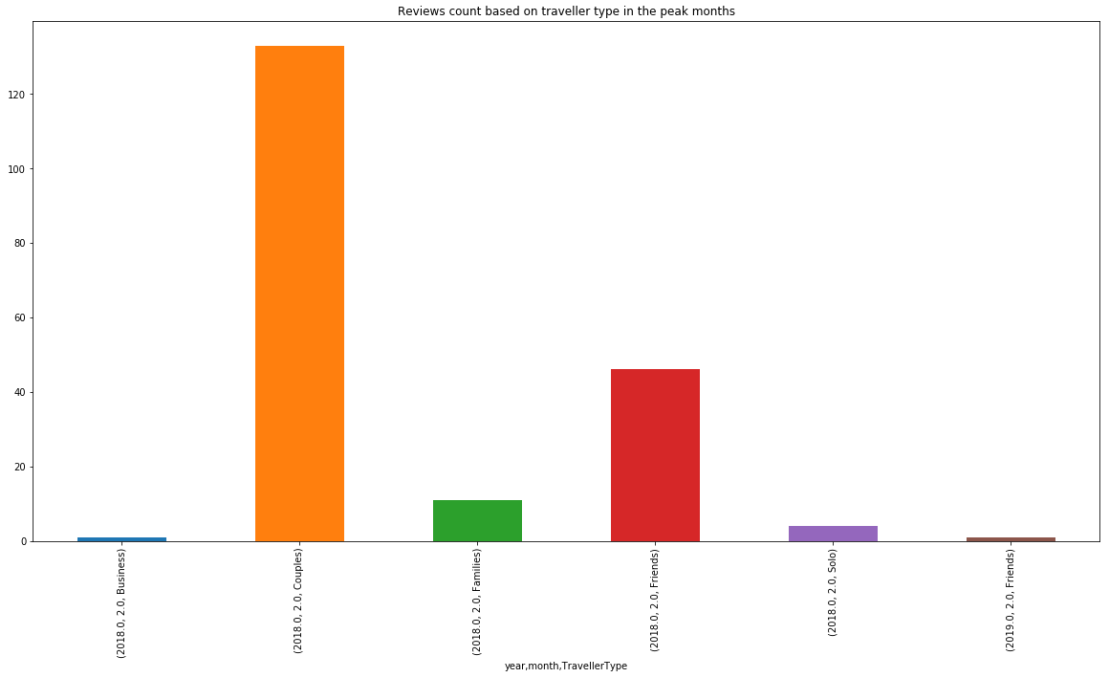


Figure 5.13 Reviews count based on traveller type in the peak months

The graph shows that majority of the reviewers at February 2018 are couples. As there do not have any special events for couples in February 2018 other than Valentine's Day, it can be expected that most of the couples are celebrating their Valentine's Day at "Dinner In The Sky Malaysia" in February 2018. After knowing the celebrations of Valentine's Day, there are questions to study why February 2019 do not have the same peak performance from February 2018. To further investigate the problem, reviews count of the restaurants with same postcode with "Dinner In The Sky Malaysia" in February 2019 are visualized.

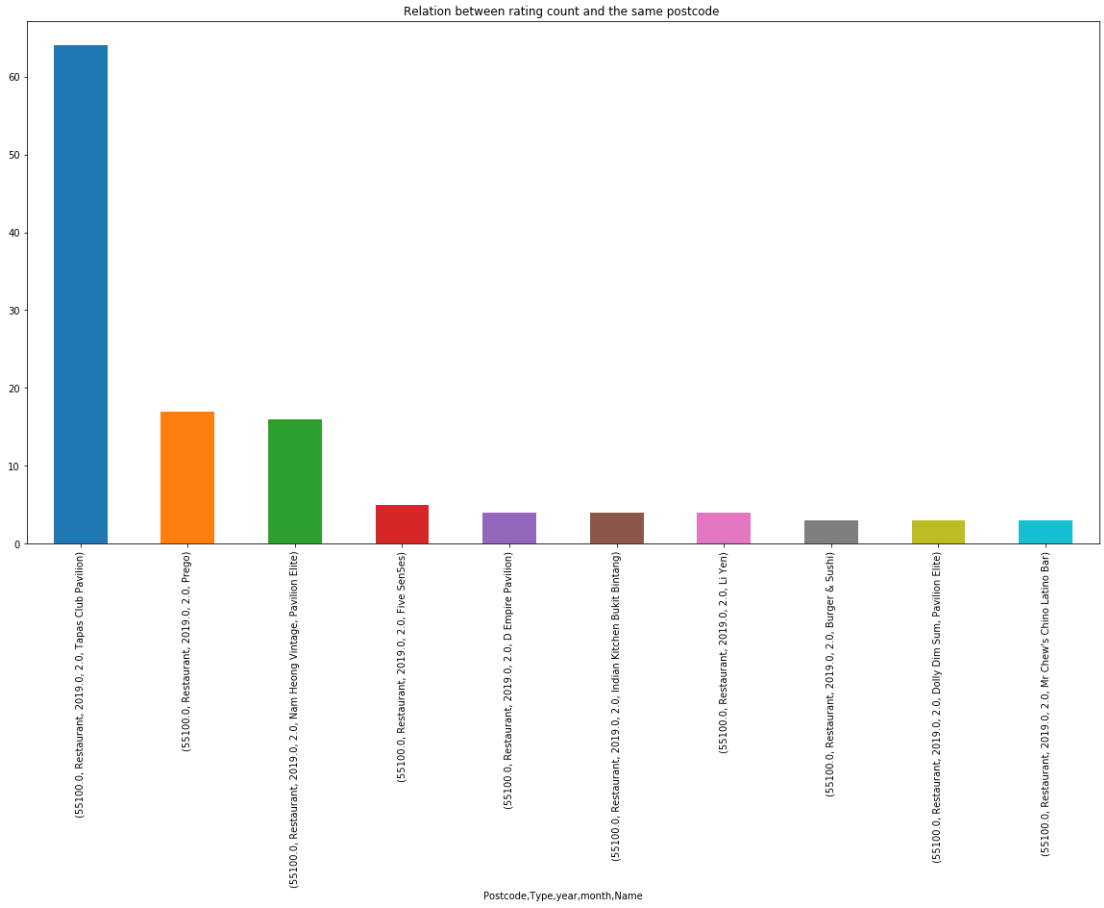


Figure 5.14 Relation between reviews count for the same postcode area

The graph above shows that at February 2019, a restaurant named "Tapas Club Pavilion" have much higher reviews count compared to other restaurants with the same postcode. It might be one of the reason of the regression of reviews count from "Dinner In The Sky Malaysia" in February 2019. By comparing the differences of reviews count between each year on a particular tourism location, the reviews count growth or fall anomalies can be detected

5.4 Peak detection for data analytics

The method of visualizing data and filtering 1 by 1 by comparing differences is able to show the trends of particular tourism location. However, there are too many tourism locations in the data. It is impossible to visualize and compare every differences of reviews count and average rating for a particular tourism location sequentially. To further reduce the human's workload, every tourism locations are separated based on years and months to group the reviews count and average ratings. Then, library from scipy named find_peaks is used to determine the peaks of every tourism location separately. When the function "find_peaks" is triggered, a list of array which consists of the peak points will be outputted. The length of the arrays are then stored into a new dataframe based on the particular tourism location name to store the frequency of peak points appeared in a particular tourism location. Then, the dataframe of storing peak point count is sorted in descending order to find out the most peak points tourism location. With this function, most of the filtering job can be saved and time can be used to focus on those tourism location with high number of peak points as shown below.

	Name	Peak count	Reviews count
2088	NZ Curry House	17	165
1916	Matahari Lodge	17	201
2880	Royale Chulan Bukit Bintang	15	927
1148	Hakka Restaurant	14	476
362	Bijan	14	710
3178	Sri Nirwana Maju Restaurant	12	220
3767	Villa Samadhi	12	825
1761	Little Penang Cafe	12	323
1439	Jogoya	12	190
3258	Sungei Wang Plaza	12	437
994	Fraser Place Kuala Lumpur	12	1485

Table 5.1 Highest peak count based on grouped reviews count from years and months which falls above quartile 0.95 from the total reviews count from particular tourism location

Besides, the reviews count is also taken into considerations because the graphs of tourism location with low reviews count are not reliable as these graphs are normalized before plotting. Therefore, the fluctuation between the low reviews count tourism locations' graphs will be higher as their maximum and minimum differences is not a large value when the month and year reviews count are grouped together. So, these

tourism location are chosen based on the reviews count which are higher than 0.95 quantile and peak count in descending order. The reviews count based on years and months are normalized for peak findings to allow a fair peak algorithms with the same hyperparameter scales to be used. For the find_peaks algorithm, the hyperparameter is set to threshold 0.15 and height 0.5.

```
peaks, _ = find_peaks(np.array(norm),height=0.5, threshold=0.15)
```

Threshold means that the point will be determined as a peak if the vertical distance of its neighboring points are at least higher than 0.15. While height is determined as the required height of peaks must be higher than 0.5 only will be classified as a peak. Then, the top 10 peak counts location tourism can be determined based on the grouped years and months reviews count as shown.

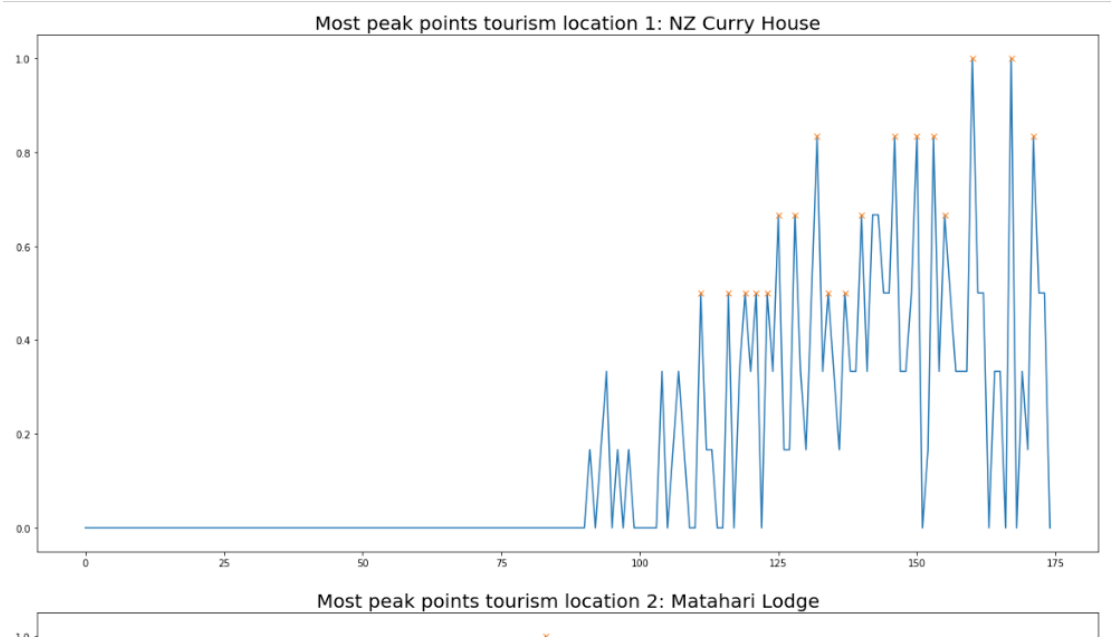


Figure 5.15 Highest peak points tourism location plotted

As we can see, the peak plotted are all above 0.5 of height and the differences between neighboring points are at least higher than 0.15. These hyperparameters can be fined tune again to have a more suitable input. With this methodology, much visualization works can be saved as this method is very dynamic which do not consists of any static input needed. We can directly choose few tourism location from this algorithm to study the trends of a particular tourism location or even tourism trends based on 'TravellerType' and 'User Nation'.

Chapter 6: Conclusions

In conclusion, there are too much useful data wasted for the tourism sector in Malaysia. Malaysia's government did not use the data they get from tourism sector effectively to apply data analytics on the data they have. This caused the tourism sector of Malaysia to make no progress on improving customers' experience because they have no idea on customers' preference while travelling in Malaysia. Because of Malaysia is a well-known tourist attraction country and there are still many improvements can be made to improve tourism sector in Malaysia. The solution for improving Malaysia's tourism sector is to apply data analytics on it, so that customers' experience can be improved by understanding every customers' preference when they are travelling in Malaysia. So, this project is started by the main motive to study the tourism trends according to tourism data collected from the tourism recommendation website so that travellers' behaviour can be understood based on different nation, traveller type and also particular tourism location's growth or loss.

By understanding travellers' behaviour and preferences, suitable marketing strategy or advertising can be applied to them by following the information obtained. For example, we studied that Australians are more preferred to visit Kuala Lumpur during January, April, July and October. So, government can have some seasonal promotions for Australians during this 4 months period to encourage them to visit Malaysia. More costs can be saved for digital marketing when the target audiences visiting months are determined. Thus, by having this strategy, the investment made for marketing can reach the break-even point as soon as possible.

On the other hand, for tourism location in Malaysia, recommendation can be given to the particular business holder by having time series analysis on their tourism location. For example, we know the main customer based on "Dinner In the Sky Malaysia" are majority consists of couple, and the reason that might affect the regression of business for that particular restaurant is from the same area which named "Tapas Club Pavilion". By providing such meaningful information for business perspective, a business holder can understand their strength and weakness in their business. They can make any changes to improve their business model so that customer's experience can be improved too.

With large data set analysed, the accuracy of analysing the travellers' behaviour is higher. The higher the accuracy, the higher the precision of precise marketing can be made on the correct customers. Thus, this can also greatly improve customers' experience when travelling in Malaysia, so they will more likely to visit Malaysia again with such structured tourism sector in Malaysia that can understand what they want upfront before the they need it.

Generally, this project involves data pre-processing, cleaning and analytics and these phases are proved to be successful as there are information obtained from data analytics phase with a strong justification proved. This shows that the data pre-processing and cleaning phases involved in this project improved the precision and accuracy of the data and leads to a reliable and accurate data analytics phase. Since this project involves many different aspects and requires different perception to study the tourism trends, this project can be further extended by collecting more data from different tourism recommendation websites to build up a larger database at first. Then, after data pre-processing and cleaning, different algorithms or templates can be created or used to determine different attributes for trends and anomaly detection.

References

Booking, n.d., The largest selection of hotels, homes, and vacation rentals. Available from: < https://www.booking.com/> [10 April 2018].

Expedia, n.d., Cheap Hotels, Resorts, and Flights Booking | Travel with Expedia Malaysia. Available from: <https://www.expedia.com.my/> [10 April 2018].

PowerBi, n.d., Business intelligence like never before. Available from: <https://powerbi.microsoft.com/en-us/> [10 April 2018].

Tate, J., 2017, Top 4 Advantages of Power BI. Available from: <http://digital.withum.com/blog/top-4-advantages-of-power-bi> [9 April 2018].

Zoho, n,d,. The operating system for your business. Available from: <https://www.zoho.com/> [9 April 2018].

Statravel, 2018, 6 reasons why Malaysia is the ultimate travel destination. Available from: <http://www.statravel.co.uk/travel-blog/2013/04/6-reasons-why malaysia-is-the-ultimate-travel-destination/> [10 April 2018].

Tourism Statistics, 2016, Malaysia Tourism Statistics in Brief. Available from: <https://www.tourism.gov.my/statistics> [10 April 2018].

Data Analytics, n.d., What is Data Analytics? - Definition from Techopedia. Available from: <https://www.techopedia.com/definition/26418/data analytics> [10 April 2018].

Sharpley, R.2000, The influence of the accommodation sector on tourism development: lessons from Cyprus. International Journal of Hospitality Management, 19(3), pp.275–293.

Travel Packages, n.d., Travel Packages | Tourism Malaysia. Available from: <http://www.malaysia.travel/en/my/travel-packages> [10 April 2018].

Trivago, n.d., trivago. Available from: < https://www.trivago.com.my/> [10 April 2018].

Poon, W.C. and Low, K.L.T. 2005, Are travellers satisfied with Malaysian hotels? International Journal of Contemporary Hospitality Management, 17(3), pp.217–227.

The Gurney, n.d., The Gurney Resort Hotel & Residences. Available from: <https://www.gurney hotel.com.my/> [10 April 2018].

Calderon, P., 2018, VADER Sentiment Analysis Explained. Available from: <http://datameetsmedia.com/vader-sentiment-analysis-explained/> [1 August 2018].

Eru, O. and Yakin, V., 2017, Emoji, Dijital Pazarlamada Yeni Bir Oyuncu. International Refereed Journal Of Marketing And Market Researches, (10), pp.19–38.

Zamboni, J., 2018, The Advantages of a Large Sample Size. Available from: <https://sciencing.com/advantages-large-sample-size-7210190.html> [3 August 2018].

Hutto, C.J. & Gilbert, Eric., 2015, VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

Anon, 2017, 'You will like it!' using open data to predict tourists' response to a tourist attraction[Online]. Available at: https://www.sciencedirect.com/science/article/pii/S0261517716302680 [Accessed: 1 March 2019].

Anon, Media Centre[Online]. Available at: https://tripadvisor.mediaroom.com/caen-about-us [Accessed: 1 March 2019].

Anon, TripAdvisor: number of reviews 2014-2017 | Statistic[Online]. Available at: https://www.statista.com/statistics/684862/tripadvisor-number-of-reviews/ [Accessed: 1 March 2019].

Basulto, D., 2013, Humans Are the World's Best Pattern-Recognition Machines, But for How Long?[Online]. Available at: https://bigthink.com/endless-innovation/humans-are-the-worlds-best-pattern-recognition-machines-but-for-how-long [Accessed: 2 March 2019].

Brownlee, J., 2017, Why One-Hot Encode Data in Machine Learning?[Online]. Available at: https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/ [Accessed: 2 March 2019].

Ferguson, A. and Ferguson, A., 2018, Human vs Machine – iProspect – Medium[Online]. Available at: https://medium.com/iprospect/human-vs-machine-6b11a9d4ed8 [Accessed: 2 March 2019].

Fischer, Thomas & Krauss, Christopher & Treichel, Alex, 2018. "Machine learning for time series forecasting - a simulation study," FAU Discussion Papers in Economics 02/2018, Friedrich-Alexander University Erlangen-Nuremberg, Institute for Economics.

Munková, D., Munk, M. and Vozár, M., 2013. Data Pre-processing Evaluation for Text Mining: Transaction/Sequence Model. Procedia Computer Science, 18, pp.1198–1207.

Scott, Daniel & Lemieux, Christopher. (2010). Weather and Climate Information for Tourism. Procedia Environmental Sciences. 1. 146-183. 10.1016/j.proenv.2010.09.011.

dfrc, 2018, Big Data and Weather: The Impact on Touristic Destinations[Online]. Available at: http://www.dfrc.com.sg/big-data-weather-tourist-destinations/ [Accessed: 24 March 2019].

Anon, Australian school holidays - Tourism Australia[Online]. Available at: https://www.australia.com/en/facts-and-planning/australian-school-holidays.html [Accessed: 2 April 2019].

Anon, Seasons in England[Online]. Available at: https://seasonsyear.com/England [Accessed: 5 April 2019].

Lucashn, 2016, lucashn/peakutils[Online]. Available at: https://github.com/lucashn/peakutils [Accessed: 6 April 2019].

Anon, London, United Kingdom: Annual Weather Averages[Online]. Available at: https://www.holiday-weather.com/london/averages/ [Accessed: 6 April 2019].

Press, G., 2016, Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says[Online]. Available at:

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#21099f846f63 [Accessed: 7 April 2019].

UTAR
UNIVERSITI TUNKU ABDUL RAHMAN

Faculty of Information and Communication Technology

WEB DATA CLEANING AND ANALYTICS FOR MALAYSIA TOURISM

Project Objectives

- Data pre-processing and data cleaning thoroughly for an accurate and analyzable data which can achieve a higher precision.
- Data analytics to find out tourism trends and correlation between attributes

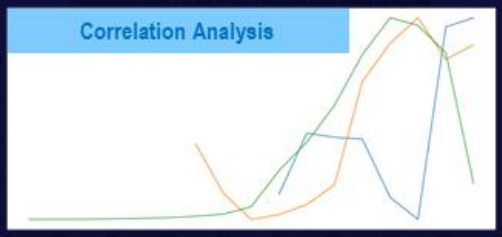
Results in Data Pre-processing and Cleaning

Address	object
Comment	object
Contribution	object
Country	object
Date	object
Date Of Stay	object
Helpful Votes	object
Local	object
Name	object
Postcode	object
Price	object
Ranking	object
Rating	int64
Reply	object
Response Date	object
State	object
TravellerType	object
Type	object
Unnamed: 0	int64
User Location	object
Username	object
dtype: object	

Unnamed: 0	int64
Type	object
Low Price	float64
High Price	float64
Ranking	float64
Name	object
Date	datetime64[ns]
Rating	float64
Review Length	int64
Country	object
State	object
City	object
Postcode	float64
Date Of Stay	datetime64[ns]
TravellerType	object
Compound	float64
Response Date	datetime64[ns]
User State	object
User Nation	object
Contribution	float64
Helpful Votes	float64
Helpful Ratio	float64
Latitude	float64
Longitude	float64
dtype: object	

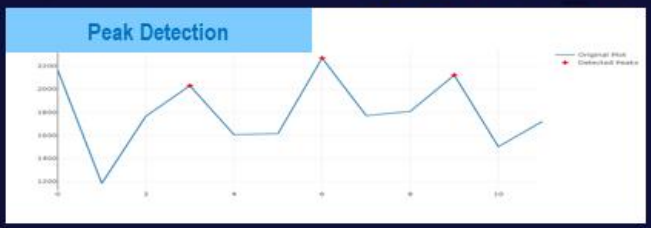
Results in Data Analytics

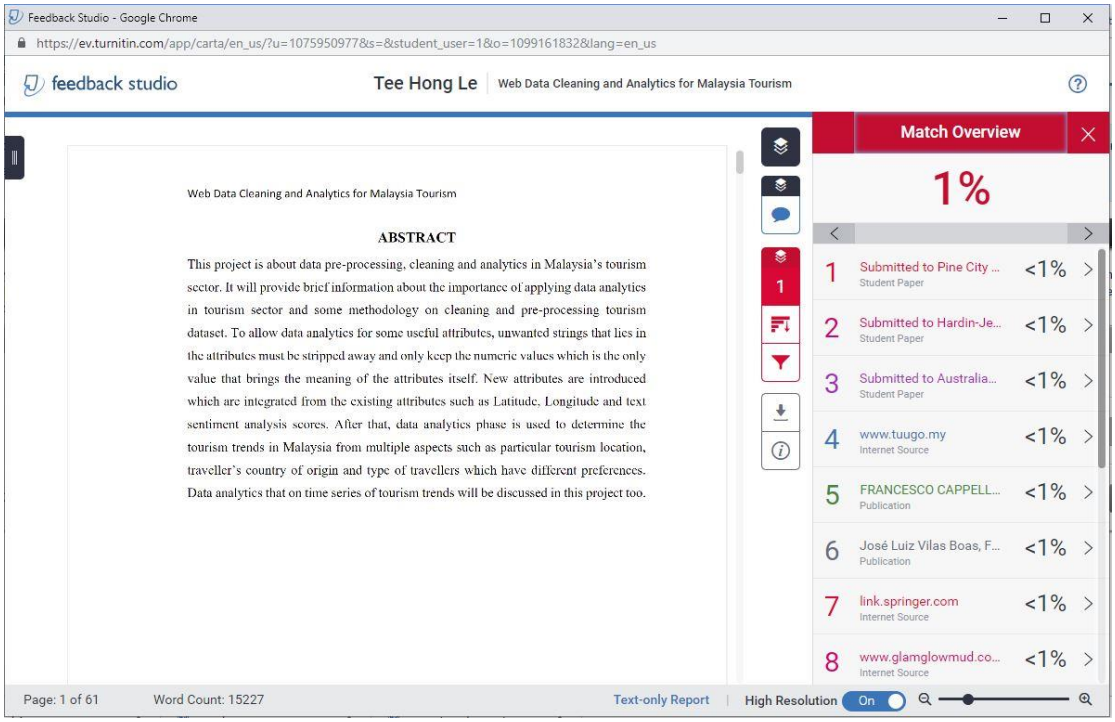
Correlation Analysis

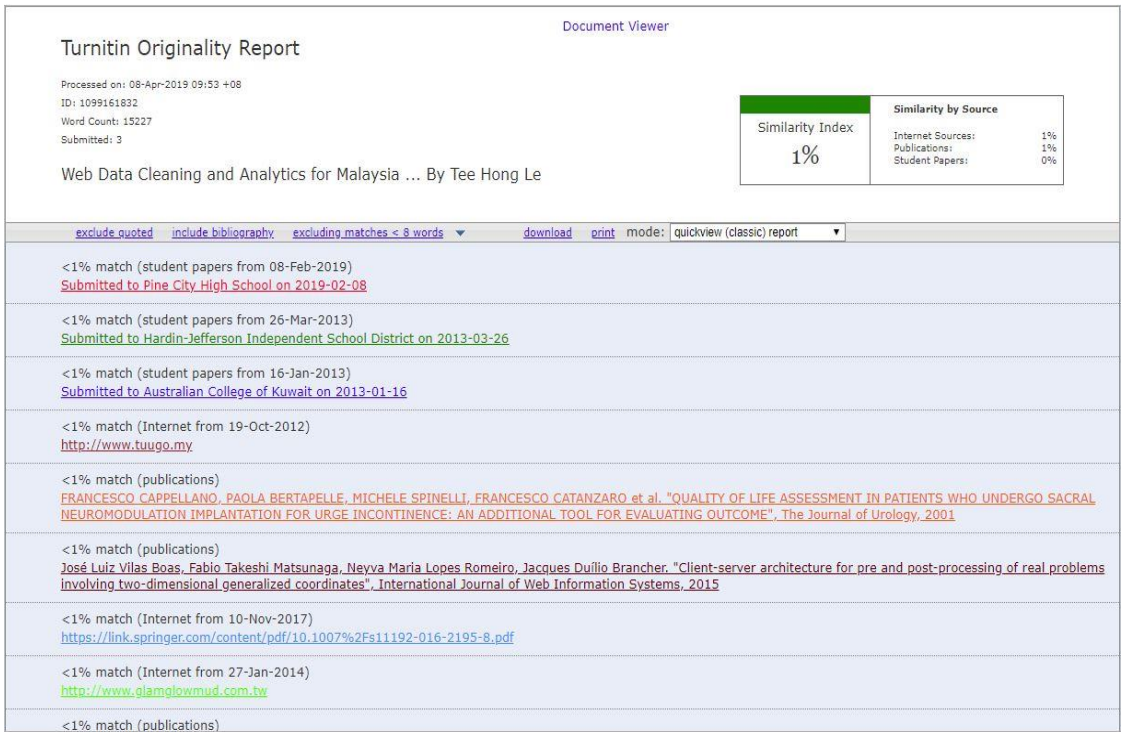


Tourism Trends Finding

Peak Detection




Feedback Studio - Google Chrome
https://ev.turnitin.com/app/carta/en_us/?u=1075950977&s=&student_user=1&o=1099161832&lang=en_us
feedback studio
Tee Hong Le
Web Data Cleaning and Analytics for Malaysia Tourism
Web Data Cleaning and Analytics for Malaysia Tourism
ABSTRACT
This project is about data pre-processing, cleaning and analytics in Malaysia's tourism sector. It will provide brief information about the importance of applying data analytics in tourism sector and some methodology on cleaning and pre-processing tourism dataset. To allow data analytics for some useful attributes, unwanted strings that lies in the attributes must be stripped away and only keep the numeric values which is the only value that brings the meaning of the attributes itself. New attributes are introduced which are integrated from the existing attributes such as Latitude, Longitude and text sentiment analysis scores. After that, data analytics phase is used to determine the tourism trends in Malaysia from multiple aspects such as particular tourism location, traveller's country of origin and type of travellers which have different preferences. Data analytics that on time series of tourism trends will be discussed in this project too.
Match Overview
1%
1 Submitted to Pine City ... <1% Student Paper
2 Submitted to Hardin-Je... <1% Student Paper
3 Submitted to Australia... <1% Student Paper
4 www.tuugo.my <1% Internet Source
5 FRANCESCO CAPPELL... <1% Publication
6 José Luiz Vilas Boas, F... <1% Publication
7 link.springer.com <1% Internet Source
8 www.glamglowmud.co... <1% Internet Source
Page: 1 of 61
Word Count: 15227
Text-only Report
High Resolution On


Document Viewer
Turnitin Originality Report
Processed on: 08-Apr-2019 09:53 +08
ID: 1099161832
Word Count: 15227
Submitted: 3
Web Data Cleaning and Analytics for Malaysia ... By Tee Hong Le
Similarity Index
1%
Similarity by Source
Internet Sources: 1%
Publications: 1%
Student Papers: 0%
exclude quoted
include bibliography
excluding matches < 8 words
download
print
mode: quickview (classic) report
<1% match (student papers from 08-Feb-2019)
Submitted to Pine City High School on 2019-02-08
<1% match (student papers from 26-Mar-2013)
Submitted to Hardin-Jefferson Independent School District on 2013-03-26
<1% match (student papers from 16-Jan-2013)
Submitted to Australian College of Kuwait on 2013-01-16
<1% match (Internet from 19-Oct-2012)
http://www.tuugo.my
<1% match (publications)
FRANCESCO CAPPELLANO, PAOLA BERTAPELLE, MICHELE SPINELLI, FRANCESCO CATANZARO et al. "QUALITY OF LIFE ASSESSMENT IN PATIENTS WHO UNDERGO SACRAL NEUROMODULATION IMPLANTATION FOR URGE INCONTINENCE: AN ADDITIONAL TOOL FOR EVALUATING OUTCOME", The Journal of Urology, 2001
<1% match (publications)
José Luiz Vilas Boas, Fabio Takeshi Matsunaga, Neyva Maria Lopes Romeiro, Jacques Duílio Brancher. "Client-server architecture for pre and post-processing of real problems involving two-dimensional generalized coordinates", International Journal of Web Information Systems, 2015
<1% match (Internet from 10-Nov-2017)
https://link.springer.com/content/pdf/10.1007%2Fs11192-016-2195-8.pdf
<1% match (Internet from 27-Jan-2014)
http://www.glamglowmud.com.tw
<1% match (publications)

Universiti Tunku Abdul Rahman			
Form Title : Supervisor's Comments on Originality Report Generated by Turnitin for Submission of Final Year Project Report (for Undergraduate Programmes)			
Form Number: FM-IAD-005	Rev No.: 0	Effective Date:	Page No.: 1of 1

FACULTY OF INFORMATION AND COMMUNICATION TECHNOLOGY

Full Name(s) of Candidate(s)	TEE HONG LE
ID Number(s)	15ACB03273
Programme / Course	BACHELOR DEGREE OF COMPUTER SCIENCE (HONS)-(CS)
Title of Final Year Project	Web Data Cleaning and Analytics for Malaysia Tourism

Similarity	**Supervisor's Comments (Compulsory if parameters of originality exceeds the limits approved by UTAR)**
Overall similarity index: 1 % **Similarity by source** Internet Sources: 1 % Publications: 1 % Student Papers: 0 %	
Number of individual sources listed of more than 3% similarity: _____	
Parameters of originality required and limits approved by UTAR are as Follows: **(i) Overall similarity index is 20% and below, and** **(ii) Matching of individual sources listed must be less than 3% each, and** **(iii) Matching texts in continuous block must not exceed 8 words** ***Note: Parameters (i) – (ii) shall exclude quotes, bibliography and text matches which are less than 8 words.***	

Note Supervisor/Candidate(s) is/are required to provide softcopy of full set of the originality report to Faculty/Institute

Based on the above results, I hereby declare that I am satisfied with the originality of the Final Year Project Report submitted by my student(s) as named above.

_____ _____

Signature of Supervisor Signature of Co-Supervisor

Name: Name: _____

_____

Date: Date: _____

_____

UNIVERSITI TUNKU ABDUL RAHMAN

FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY (KAMPAR CAMPUS)

CHECKLIST FOR FYP2 THESIS SUBMISSION

Student Id	15ACB03273
Student Name	Tee Hong Le
Supervisor Name	Dr. Liew Soung Yue

TICK (√)	DOCUMENT ITEMS Your report must include all the items below. Put a tick on the left column after you have checked your report with respect to the corresponding item.
	Front Cover
	Signed Report Status Declaration Form
	Title Page
	Signed form of the Declaration of Originality
	Acknowledgement
	Abstract
	Table of Contents
	List of Figures (if applicable)
	List of Tables (if applicable)
	List of Symbols (if applicable)
	List of Abbreviations (if applicable)
	Chapters / Content
	Bibliography (or References)
	All references in bibliography are cited in the thesis, especially in the chapter of literature review
	Appendices (if applicable)
	Poster
	Signed Turnitin Report (Plagiarism Check Result - Form Number: FM-IAD-005)

*Include this form (checklist) in the thesis (Bind together as the last page)

I, the author, have checked and confirmed all the items listed in the table are included in my report. _____ (Signature of Student) Date:	Supervisor verification. Report with incorrect format can get 5 mark (1 grade) reduction. _____ (Signature of Supervisor) Date: