Link to Github: https://github.com/WQD170093/Data_Mining-MilestoneProject

WQD 7005 DATA MINING
MILESTONE PROJECT:
POLLUTANT STANDARD INDEX AND GENTING SHARE PRICE IN SINGAPORE

LIM LI THEM
WQD 170093

FACULTY OF COMPUTER SCIENCE
UNIVERSITY OF MALAYA
KUALA LUMPUR

2019

TABLE OF CONTENT

CHAPTER 1: INTRODUCTION

Southeast Asia countries suffer from haze occasionally that caused by illegal agricultural fires due to industrial-scale slash-and-burn practices in Indonesia. Experience had proved that every time the smoke event will cause losses in economy and income opportunities due to business interruption. For example, reduced in productivity for certain industries which requires the employee to be outdoors, events or schools being cancelled or postponed and loss of revenue from decrease in number of tourists. According to study carried out by Swiss Re Institute, the haze event in year 1997 and year 2015 had resulted in few million USD losses with transportation industry was impacted the most, followed by tourism, health and lastly was education (figure 1.1).

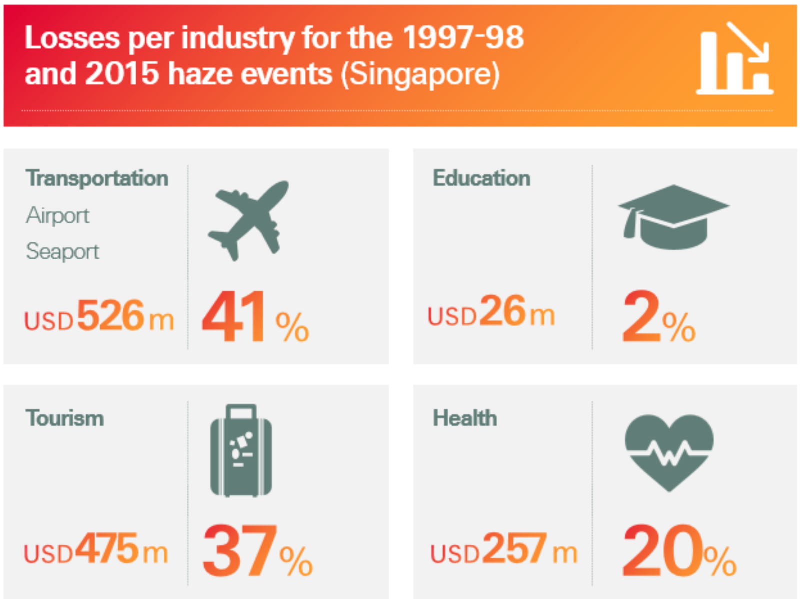

Figure 1.1 Losses caused to some industries by haze outbreak in year 1997 and year 2015

In order to deep dive into the impact of haze on tourism industry, an analysis about the impact of haze on stocks' share prices that related to tourism industry was carried out upon on the event of haze that just happened last month. On the other hand, public sentiment towards this phenomenon was analyzed with aim to understand the impact of haze from the aspect of society other than economics. Pollutant Standard Index, PSI, which is one of the indexes that measure the air quality level, is used in this research to measure the level of air pollutant during this haze breakout in Singapore. It's in the scale of 0 to 500, that enables the public to aware the air pollution level in a particular place. The figures were grouped into different level with its effects to public health as shown in table 1.1.

Table 1.1 PSI level and its health effects

PSI	Descriptor	General Health Effects
0–50	Good	None
51–100	Moderate	Few or none for the general population
101–200	Unhealthy	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects. To stay indoors.
201-300	Very unhealthy	Health warnings of emergency conditions. The entire population is more likely to be affected.
301+	Hazardous	Health alert: everyone may experience more serious health effects

The stocks that were chosen to be studied in this research would be the top 5 largest stocks in the Singapore tourism industry (table 1.2). However, the main focus would be the share price of Genting Singapore, which had the highest market capitalization value of 13,0009 Singapore Dollar as of January year 2019. The relationship between this share price and the PSI value was observed and other factors that affect the Genting share price such as the share prices of other tourism stocks, Gross Domestic Product's (GDP) growth rate and number of international tourist arrival were investigated. Singapore was chosen as our analysis target location due to the availability and completeness of data in this country.

Table 1.2 Top 5 tourism stocks in Singapore sorted by market capitalization value

Name	SGX Code	Market Cap (S$ mln)	Total Return YTD (%)	Total Return 2018 (%)	Total Return 3Y (%)	P/E (x)	P/B (x)	ROE (%)
Genting Singapore	G13	13,009	10.8	-23.2	71.0	18.1	1.72	9.6
Singapore Airlines	C6L	11,541	3.5	-8.3	-3.6	25.0	0.81	3.6
Mandarin Oriental	M04	3,259	-7.1	4.4	42.5	37.9	1.94	5.0
Ascott Residence Trust	A68U	2,576	8.3	-6.1	31.9	21.3	0.85	4.8
Hotel Properties	H15	1,937	3.3	-5.7	10.4	9.8	0.90	10.1

In summary, the objectives of this research are as below:

1. To determine the relationship between Pollutant Standard Index (PSI) reading and the G13 stock price
2. To identify the other factors that affect G13 stock price
3. To understand the public sentiment toward haze through Twitter

CHAPTER 2: RESEARCH METHODOLOGY

Python was used to do web crawling to extract the real time and historical PSI data from Singapore's public data portal. It's used to do extract unstructured data, tweets, from twitter as well. Besides web crawling, some correlation analysis and modelling were carried out in Python to achieve the objectives of this study. On the other hand, SAS Enterprise Miner 9.4 was deployed to perform data pre-processing, data visualization and modelling. Examples of machine learning algorithms used in this study were K-mean Clustering, Decision Tree, Logit Regression and Ordinary Least Square model (Liner Regression). Model comparison was performed to find the best model that is able to predict the Genting share price based on the inputs keyed-in.

CHAPTER 3: DATA COLLECTIONS AND PRE-PROCESSING

Five types of data were collected in this project, which were PSI readings, stock prices, number of international tourist arrival, year-on-year Gross Domestic Product (GDP) growth rate and tweets. Real time and historical daily PSI data were crawled from Singapore's public data portal from year 2009 until year 2019. Same goes to year-on-year GDP growth rate, were collected from the same source and same period as PSI data but were collected on quarterly basis. The daily closing share prices of the 5 stocks mentioned were collected from Yahoo Finance from year 2009 until year 2019 whereas number of international tourist arrival by countries were acquired from Department of Statistics Singapore from year 1978 until year 2019 on monthly basis. Lastly, the real time tweets which consisted of words like 'haze' and 'jerebu' were crawled from Twitter from 20th September 2019 until 9th October 2019, with total about 88,353 tweets were captured.

The metadata of the data acquired were shown in figure 3.1 and figure 3.2 below.

Variables - FIMPORT4

(none) not Equal to

Columns: Label Mining Basic

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Average of A68U	Input	Interval	No		No	.	.
Average of C6L	Input	Interval	No		No	.	.
Average of G13	Input	Interval	No		No	.	.
Average of H15	Input	Interval	No		No	.	.
Average of M04	Input	Interval	No		No	.	.
PSI Avg	Input	Interval	No		No	.	.
Total International Visitor Arri	Input	Interval	No		No	.	.
Trend	Target	Interval	No		No	.	.
YOY GDP Growth Rate	Input	Interval	No		No	.	.
VAR1	Time ID	Nominal	No		No	.	.

Figure 3.1 Metadata of PSI, stock prices, number of international arrival and GDP growth rate dataset

Variables - FIMPORT

(none) not Equal to

Columns: Label Mining Basic

Name	Label	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
A	A	Input	Interval	No		No	.	.
Date and Time	Date and Time	Input	Nominal	No		No	.	.
Location	Location	Input	Nominal	No		No	.	.
Message	Message	Input	Nominal	No		No	.	.
New Message	New Message	Input	Nominal	No		No	.	.
Positivity	Positivity	Target	Nominal	No		No	.	.
Retweet Count	Retweet Count	Input	Nominal	No		No	.	.
Sentiment Polarity	Sentiment Polarity	Target	Interval	No		No	.	.
Username	Username	Input	Nominal	No		No	.	.

Figure 3.2 Metadata of tweets dataset

There were missing values in the variables PSI, number of tourist arrival and GDP growth rate. The rows with the missing daily PSI values were removed whereas the missing values in the variables number of tourist arrival and GDP growth rate (figure 3.3) were filled with the mean of the respective classes using the impute function built in the SAS Enterprise Miner as shown in figure 3.3. After data cleaning, PSI data, stock prices and the number of tourist arrival were normalized using z-score formula as below:

$$z_i = \frac{x_i - \mu}{s}$$

where x_i is the data point, μ is the sample mean and s is the sample standard deviation.

Variable ‘trend’ was created to capture the movement of Genting stock price. Value ‘1’ was assigned if the price of Genting stock moved upward (price in the previous day is lower) and value ‘0’ if the price of Genting stock moved downward (price in the previous day is higher).

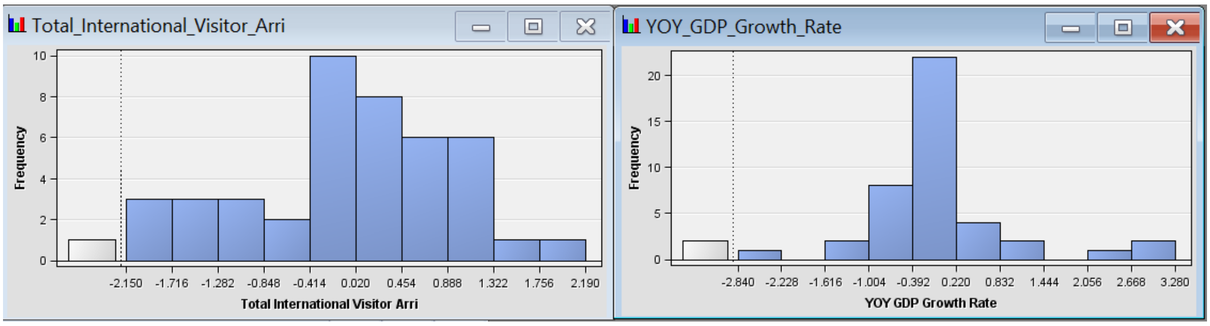

Figure 3.3 Histogram of tourist arrival and YOY GDP growth rate

For tweets’ sentiment analysis, data pre-processing such as convert to lower capital letters, remove stop words, numbers and symbols and also remove words’ length that were less than three digits were executed. This was to ensure the sentiment polarity score can be calculated correctly to avoid wrong sentiment labelling. In figure 3.4, the column ‘Message’ was the original tweets collected and the column ‘New Message’ was the tweets after cleaning, which were tokenized and converted to list. Sentiment polarity scores and its sentiment categories (positive, negative and neutral) were appended to the datafile.

Message	Username	Date and Time	Location	Retweet Count	Sentiment Polarity	New_Message	Positivity
@VapeTheBud @7ACRESmj @liftandco Yeah I just tried to find "Jack Haze" using the @liftandco search function and it didn't find anything.	Matt Fenton	10/8/2019 22:56	Ontario, Canada	0	0	['yeah', 'tried', 'find', 'jack', 'haze', 'using', 'search', 'function', 'find', 'anything']	Neutral
@MarkPinnix @Gina2God @MAGA_Randi @looneym7701 @velbryant1 @nvrimmie @MagaKatnip @TDigornio @sexyasspatriotâ€¦ https://t.co/ek2AvP4rdt	Nobodybutme IncorrigibleDeplorable #KAG2020!	10/8/2019 22:55	USA	0	0	['randi']	Neutral
@MarkPinnix @Gina2God @MAGA_Randi @looneym7701 @Nobodybutme17 @velbryant1 @nvrimmie @MagaKatnip @TDigornioâ€¦ https://t.co/BVD6CwdTdQ	Karen Ladybug1	10/8/2019 22:55	Tennessee, USA	0	0	['randi']	Neutral
RT @Genius: drive around with your friends, smoke a gram of that haze	konney	10/8/2019 22:54	233	346	0	['drive', 'around', 'friends', 'smoke', 'gram', 'haze']	Neutral
RT @evefabrics: â€œPurple Hazeâ€Apron Top | 1 of 1 | Exclusive https://t.co/biufS5pw9t	ðŸ	10/8/2019 22:54	None	2	0.5	['purple', 'haze', 'apron', 'top', 'exclusive']	Positive
RT @jacobinmag: It's a big club, and you're not in it. https://t.co/sqcEmHC5dy	Tania	10/8/2019 22:54	Los Angeles, CA	2216	0	['big', 'club']	Neutral
RT @Genius: drive around with your friends, smoke a gram of that haze	D	10/8/2019 22:53	Catalonia, Spain	346	0	['drive', 'around', 'friends', 'smoke', 'gram', 'haze']	Neutral
@IanColdwater Evil at scale..... really https://t.co/n7BxTQM9J0	Sean Patrick Condon	10/8/2019 22:53	on the highway to hell	0	-0.4	['evil', 'scale', 'really']	Negative
@aveuhree haze by Tessa violet, Pure by Hey violet, and idk	ð‰ð¶â„¸ (but 10x spookier)	10/8/2019 22:53	jung hoseok love bot	0	0.214285714	['haze', 'tessa', 'violet', 'pure', 'hey', 'violet', 'idk']	Positive
@OtiMabuse @kelvin_fletcher Sensual, simplistic & to make us all *feel in the moment & get carried away in a haze oâ€¦ https://t.co/dAlFYvuguU	Julie Anne Cotterill	10/8/2019 22:53	Sutton Coldfield	0	-0.5	['fletcher', 'sensual', 'simplistic', 'amp', 'make', 'feel', 'moment', 'amp', 'get', 'carried', 'away', 'haze']	Negative

Figure 3.4 Tweets after data cleaning

CHAPTER 4: DATA DESCRIPTION AND VISUALIZATION

As mentioned previously, Genting stock price was the main focus in this project. The relationship between PSI value and Genting Stock Price was shown in the graph 4.1. The PSI values, in overall, showed an increasing trend whereas the Genting stock price was decreasing and fluctuating from year 2009 until year 2019. The spike in PSI data was due to a serious haze outbreak that happened in Singapore in year 2013 to year 2015 and the recent haze event. The opening of Resort World and Casino, on the other hand, had caused the Genting stock price to spike in year 2010. The PSI was inversely proportional to Genting stock price at first glance. When the PSI was at its spike (as circled), the stock prices dropped. The stock prices are at its highest level when the PSI data was at the lower value. In addition, the graph 4.2 indicated the relationship between number of tourist arrival, GDP growth rate and Genting stock price. Despite the number of tourists increased over the years, the Genting stock price was decreasing and moving up and down. The GDP growth rate seemed like moving proportional with Genting stock price. There's a peak in GDP growth rate in year 2010 because of global economic recession in year 2009 and the year before but Singapore managed to bounce back in year 2010 due to its fast growing of manufacturing sector.

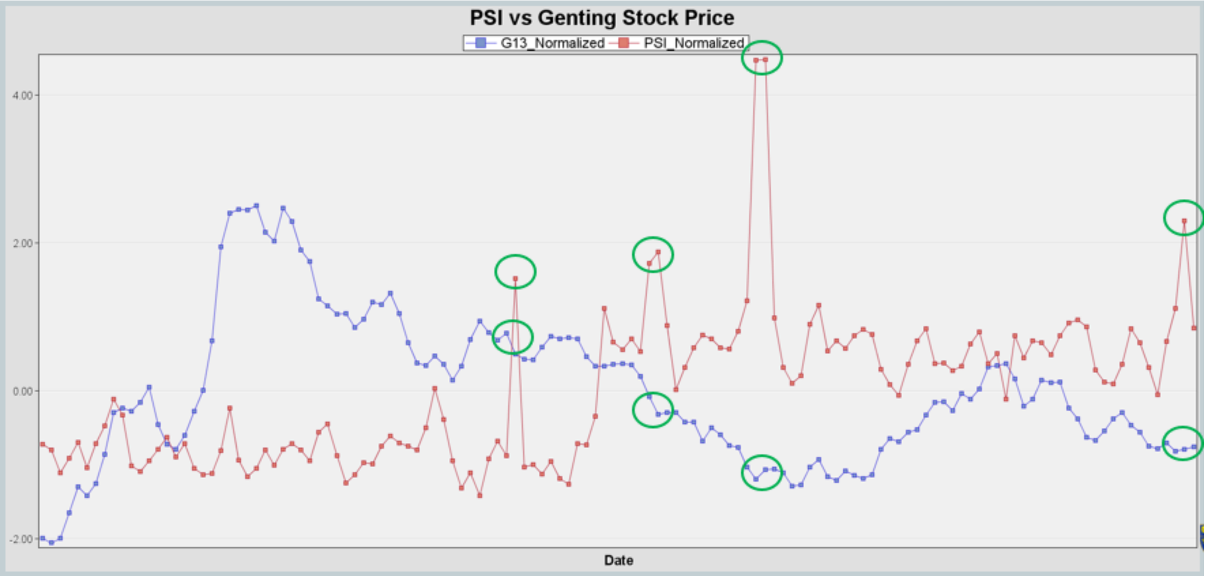

Figure 4.1 Relationship between PSI value and Genting Stock Price

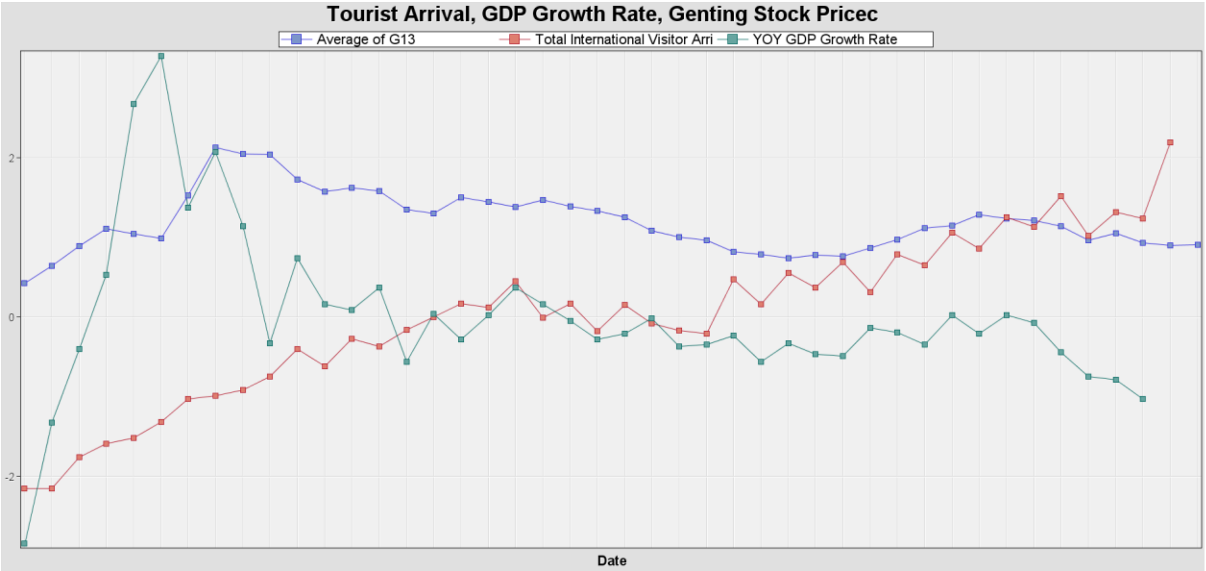

Figure 4.2 Relationship between number of tourist arrival, GDP growth rate and Genting stock price

K-mean clustering was performed to look for the hidden pattern in the dataset. Three clusters were chosen after few trials as pattern was spotted when 3 clusters were formed (figure 4.3). In cluster 1, when the level of PSI, number of tourists and GDP growth were lower, all the share prices were at their minimum. In cluster 2, however, PSI values and the number of tourists were lower than average but GDP growth rate was higher than usual, which result in some of the share prices such as C6L and G13 were higher than average level. Cluster 3 profile was the opposite of cluster 2. PSI values and the number of tourists were skewed to the right but GDP growth rate was at its average level. This resulted in stock prices like H15, M04 and A68U higher than average. Stock prices of C6L and G13 were lower in cluster 3.

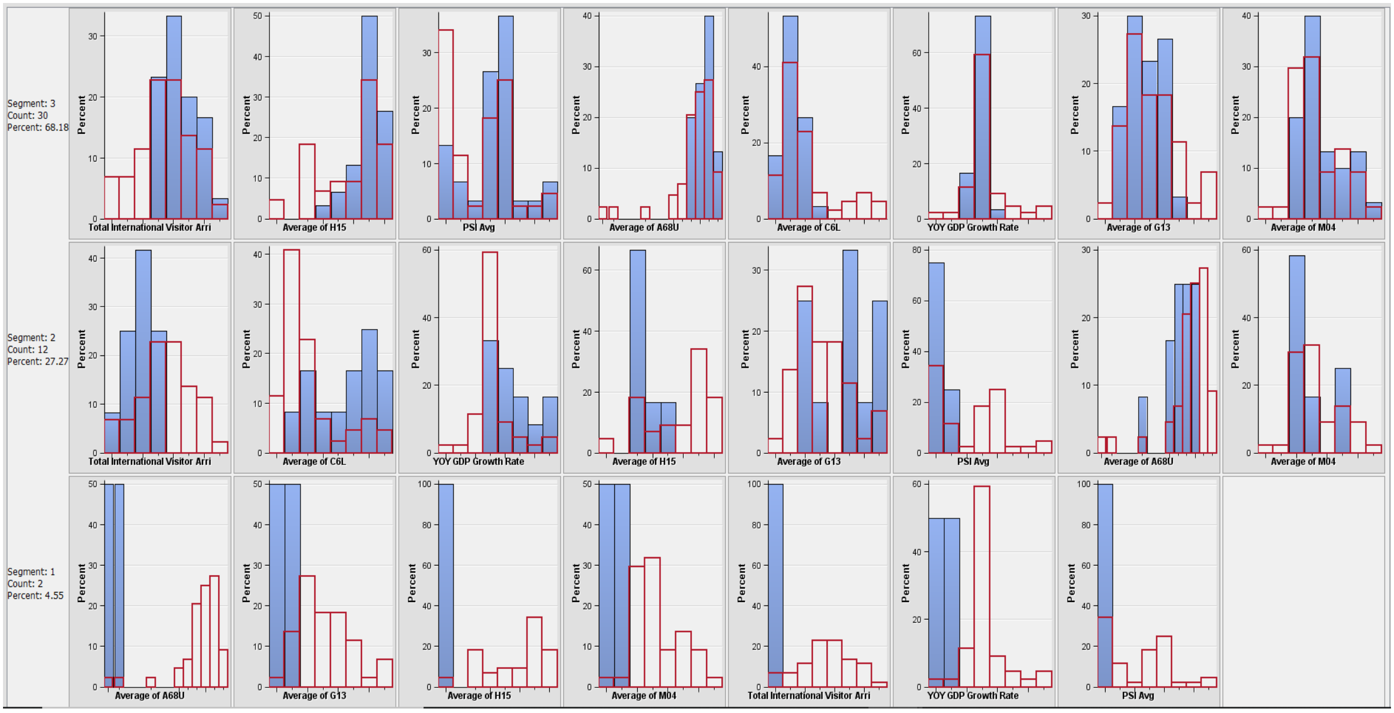

Figure 4.3 K-mean clustering on dataset

Besides predicting the Genting stock price trend and its value, the project also aimed to understand the public sentiment towards haze phenomenon. The sentiments in these tweets were analysed and were categorized as Positive, Negative and Neutral sentiments. The sentiment polarity scores were also calculated in the range of -1.0 to 1.0. Tweets with positive sentiments would have score more than zero while tweets with negative sentiments would have score less than zero. Result showed that about 48.77% of the tweet messages were neutral in sentiment while the rest of the messages showed emotion in which positive messages made up of 41.51% and negative messages made up of 9.72% (figure 4.4). The histogram in figure 4.5 indicated that most of the tweets fell into bin between score 0 and 0.1, which were neutral in sentiment. Word clouds were created to show the top terms in every sentiment category. The positive word cloud contained terms such as ‘appreciate’, ’better’ and ‘hikmah’ while the negative tweets were having terms like ‘worse breath’ and ‘getting worse’ (figure 4.6). On the other hand, the word cloud of neutral sentiment consisted of terms like ‘friends’, ‘drive around’ and ‘joke’ that carried no emotion. The period of tweets collected was at the end of September and beginning of October year 2019, in which the haze phenomenon was getting better and some regions were clear of haze. This can explain why the tweet messages were displaying more neutral and positive sentiments compared to the negative emotion.

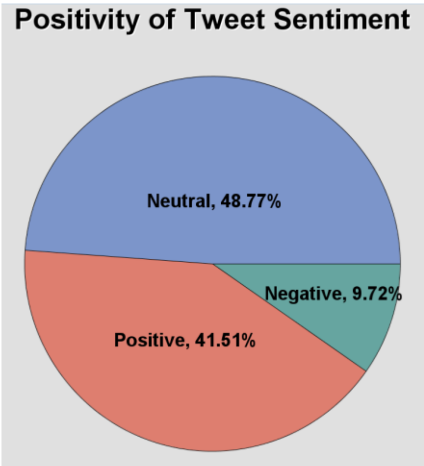

Figure 4.4 Positivity of tweet Sentiment

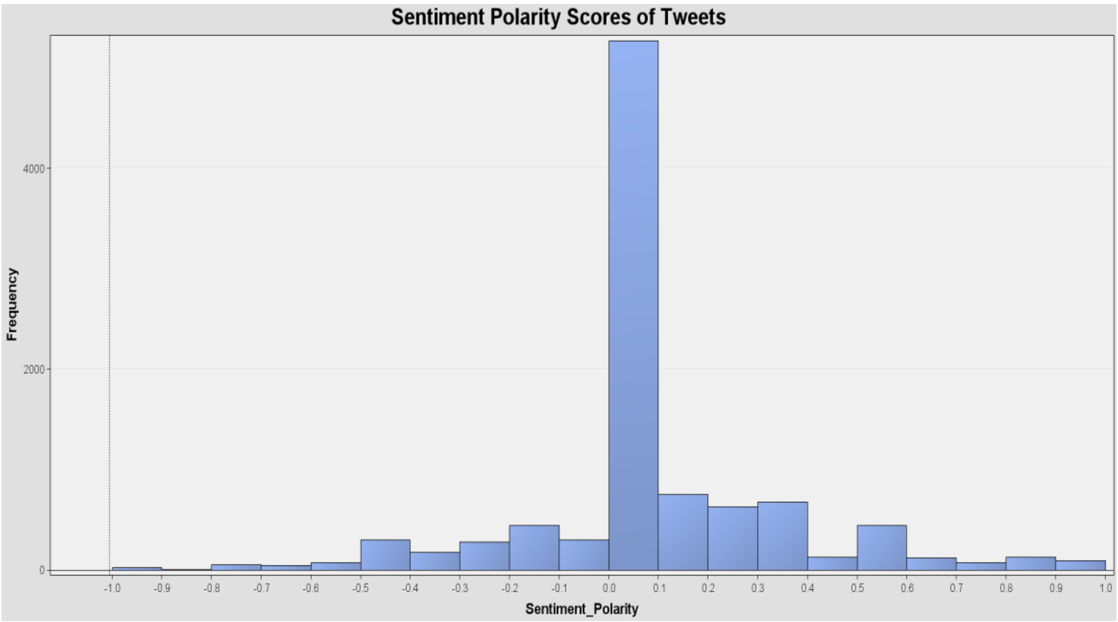

Figure 4.5 Sentiment polarity scores of tweets

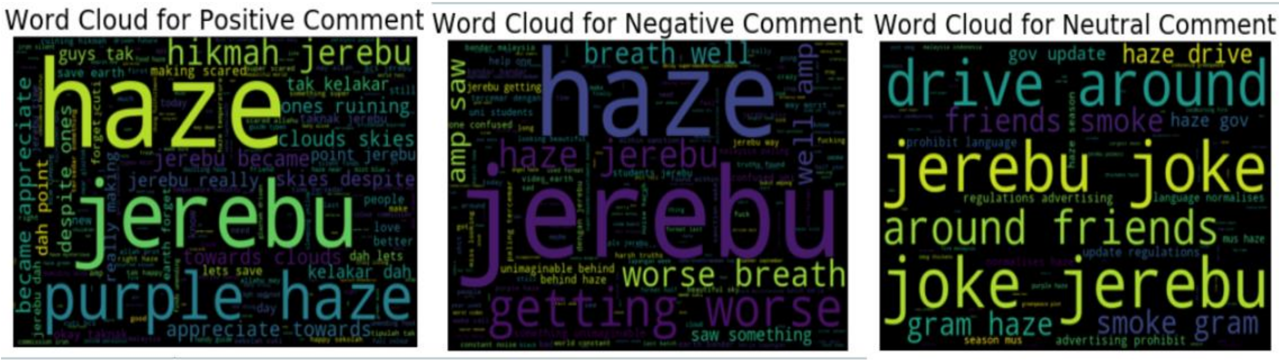

Figure 4.6 Word clouds for every sentiment category

CHAPTER 5: MODELLING AND RESULTS

The correlation between the variables were checked using correlation matrix in Python before applying any models (figure 5.1 and figure 5.2). The Genting share price, G13, was negatively correlated to PSI readings, number of visitor arrival and share price of H15 but positively proportional to GDP growth rate and other stocks' prices.

	PSI Avg	Visitor Arrivals	GDP Growth	A68U	C6L	H15	M04	G13	Trend
PSI Avg	1.000000	0.617179	-0.327364	0.394310	-0.486339	0.765864	0.314759	-0.503952	-0.086500
Visitor Arrivals	0.617179	1.000000	-0.238562	0.644469	-0.684337	0.765720	0.598053	-0.112230	-0.194125
GDP Growth	-0.327364	-0.238562	1.000000	0.337235	0.677897	-0.112583	0.091828	0.464736	-0.079359
A68U	0.394310	0.644469	0.337235	1.000000	-0.184453	0.737588	0.539104	0.263823	-0.217090
C6L	-0.486339	-0.684337	0.677897	-0.184453	1.000000	-0.451879	-0.114509	0.396142	0.193808
H15	0.765864	0.765720	-0.112583	0.737588	-0.451879	1.000000	0.558161	-0.210956	-0.040619
M04	0.314759	0.598053	0.091828	0.539104	-0.114509	0.558161	1.000000	0.346116	-0.014276
G13	-0.503952	-0.112230	0.464736	0.263823	0.396142	-0.210956	0.346116	1.000000	-0.088996
Trend	-0.086500	-0.194125	-0.079359	-0.217090	0.193808	-0.040619	-0.014276	-0.088996	1.000000

Figure 5.1 Correlation table among the variables

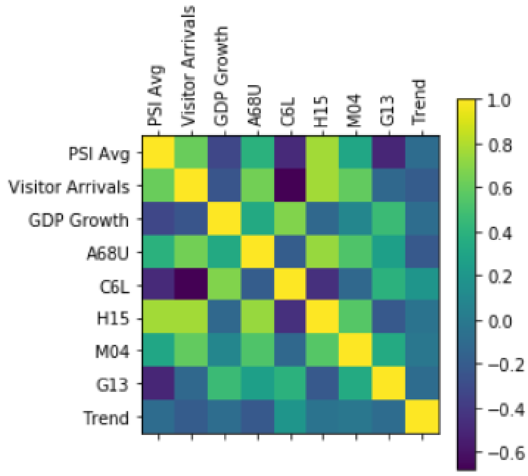

Figure 5.2 Correlation matrix plot

To confirm on the relationship between these variables, the first model applied was interactive decision tree model. A total of 44 observations were split into 50% training data and 50% validation data (figure 5.3). Interactive Decision Tree Model was then built to predict the upward and downward trend of Genting share price. Average of PSI readings and H15 stock price, which had the highest information gain, become the split rules (figure 5.4). Based on the decision tree built, PSI value that less than -0.855 will lead to downward moving trend of the Genting share price. PSI value that greater than -0.855 and having H15 stock price higher than 3.669 will lead to upward moving trend of Genting share price and vice versa. However, the subtree assessment plot in figure 5.5 showed that the performance of validation sample only improved up to a tree of one leaf and then diminished as the complexity of the model increased. The fit statistics table 5.1 showed an average squared error of 0.178 for training sample but 0.309 for validation data. This might due to overfitting of the model in the training data. The model achieved accuracy rate of 69% based on the average squared error of validation dataset in overall.

Partition Summary

Type	Data Set	Number of Observations
DATA	EMWS1.Impt_TRAIN	44
TRAIN	EMWS1.Part_TRAIN	22
VALIDATE	EMWS1.Part_VALIDATE	22

Figure 5.3 Data partition

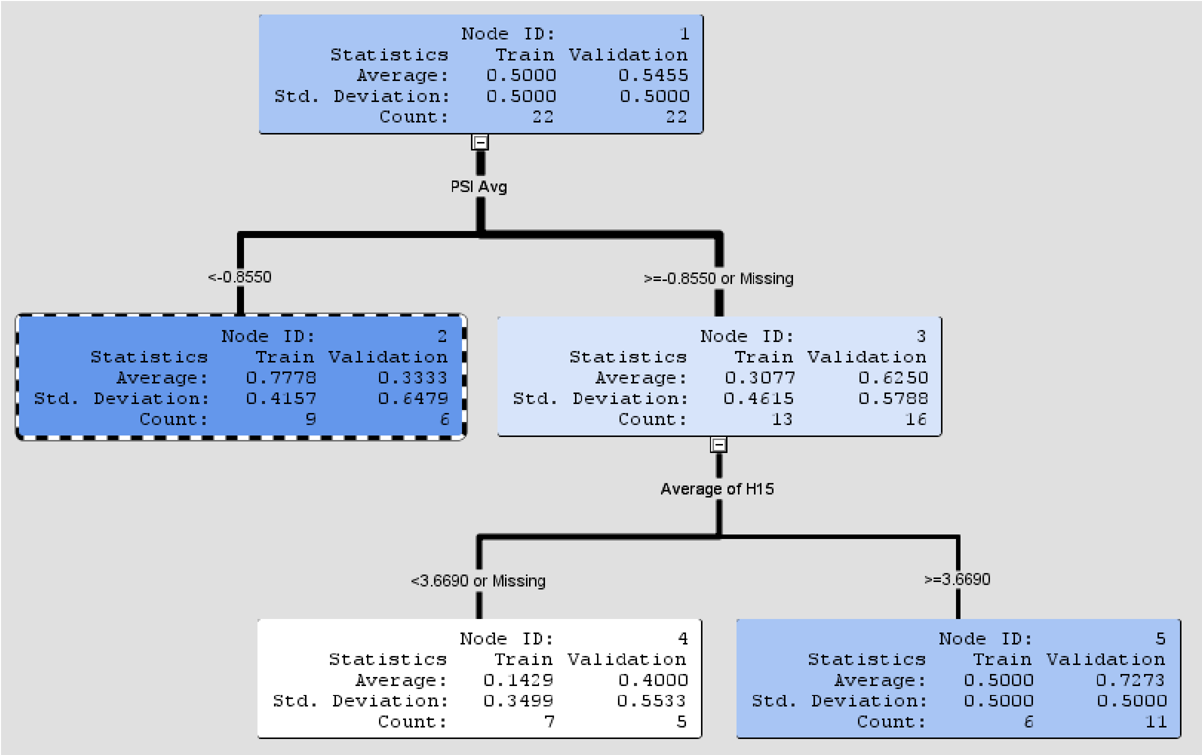

Figure 5.4 Interactive Decision Tree Model

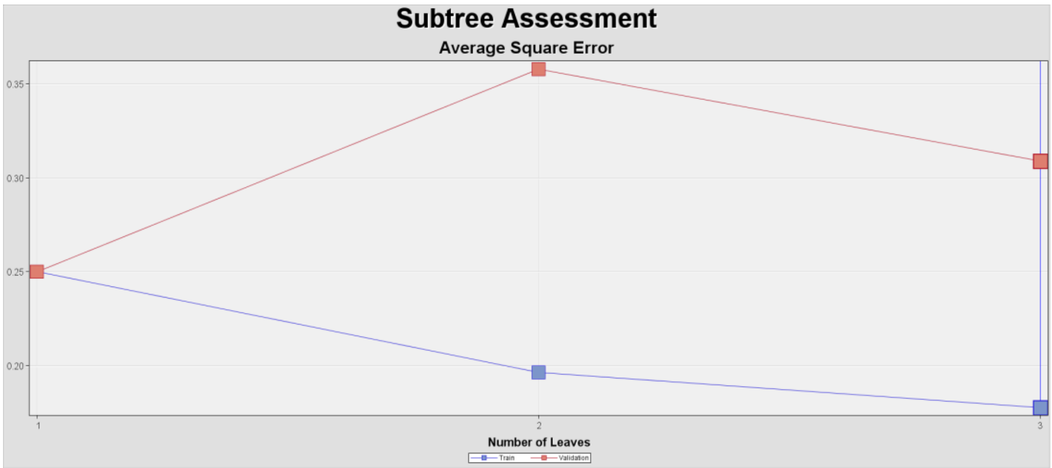

Figure 5.5 Subtree Assessment Plot

Table 5.1 Fit statistic table of Interactive Decision Tree Model

Target ▲	Fit Statistics	Statistics Label	Train	Validation
Trend	NOBS	Sum of Frequencies	22	22
Trend	MAX	Maximum Absolute Error	0.857143	0.857143
Trend	SSE	Sum of Squared Errors	3.912698	6.799131
Trend	ASE	Average Squared Error	0.17785	0.309051
Trend	RASE	Root Average Squared Error	0.421723	0.555924
Trend	DIV	Divisor for ASE	22	22
Trend	DFT	Total Degrees of Freedom	22	.

Other than decision tree, logit regression model was deployed also to predict the moving trend of Genting share price. The result in figure 5.6 revealed that p-value was higher than 0.05, indicating that the group of independent variables (PSI readings, other stocks' prices, number of tourist arrivals and GDP growth rate) were unable to predict the dependent variable (Genting stock price) reliably at 5% significant level. R-square value of 0.5459 also indicated that only 54.59% of the variance in Genting share price can be forecasted using this model. Although the average squared error of 0.1135 is lower than the average squared error of decision tree for training data, but its validation data had higher average squared error value of 0.5894 (table 5.2). Precision, recall and f1-score of logit regression model were calculated using Python and it found that only 10 out of 22 samples were classified correctly using the model (figure 5.7). In short, this model only managed to acquired an accuracy rate of 41% based on the average squared error of validation sample.

Model comparison was carried out on these two models and it turned out that decision tree was better than logit regression model in forecasting the Genting share price's trend. Both models had similar average squared error values for training data but decision tree had smaller average squared error in validation sample compared to logit regression model (figure 5.7). In summary, the classification models gave us the highest accuracy rate of 69% in forecasting the moving trend of Genting share price.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	3.002676	0.375335	1.95	0.1361
Error	13	2.497324	0.192102		
Corrected Total	21	5.500000			

Model Fit Statistics

R-Square	0.5459	Adj R-Sq	0.2665
AIC	-29.8681	BIC	-16.3651
SBC	-20.0487	C(p)	9.0000

Figure 5.6 Logit regression model result

Table 5.2 Fit statistics table of logit regression model

Target	Fit Statistics	Statistics Label	Train	Validation
Trend	AIC	Akaike's Infor...	-29.8681	.
Trend	ASE	Average Squa...	0.113515	0.589425
Trend	AVERR	Average Error ...	0.113515	0.589425
Trend	DFE	Degrees of Fr...	13	.
Trend	DFM	Model Degree...	9	.
Trend	DFT	Total Degrees...	22	.
Trend	DIV	Divisor for ASE	22	22
Trend	ERR	Error Function	2.497324	12.96735
Trend	FPE	Final Predictio...	0.270689	.
Trend	MAX	Maximum Abs...	0.682174	1.83706
Trend	MSE	Mean Square ...	0.192102	0.589425
Trend	NOBS	Sum of Frequ...	22	22
Trend	NW	Number of Est...	9	.
Trend	RASE	Root Average ...	0.336919	0.76774
Trend	RFPE	Root Final Pre...	0.520278	.
Trend	RMSE	Root Mean Sq...	0.438294	0.76774
Trend	SBC	Schwarz's Bay...	-20.0487	.
Trend	SSE	Sum of Squar...	2.497324	12.96735
Trend	SUMW	Sum of Case ...	22	22

Fit Statistics
Model Selection based on Valid: Average Squared Error (_VASE_)

Selected Model	Model Node	Model Description	Valid: Average Squared Error	Train: Average Squared Error
Y	Tree	Decision Tree	0.30905	0.17785
	Reg3	Regression_Overall	0.58943	0.11351

Figure 5.7 Model comparison result between decision tree and logit regression model

Ordinary Least Square (OLS) method was deployed in this project too to predict the Genting share price value. It turned out that this linear regression model had better accuracy rate than the two classification models with accuracy rate of 95%. The model also had small p-value that's less than 0.05 and high R-squared value of 0.971, indicating that the PSI readings, stocks' prices, GDP growth rate and number of tourist arrivals were able to predict the Genting share price reliably at 5% significant level. Figure 5.10 showed that the predicted values forecasted by the OLS model were plotted against the actual value, the values were quite close to each other. The individual p-value in figure 5.8 demonstrated the ability of each individual independent variable to predict the dependent variable. Variables such as PSI, stock prices of H15, A68U and M04 were statistically significant as they had p-value less than 0.05. Number of tourist arrival, GDP growth rate and stock price of CL6, on the other hand, were not statistically significant as its p-value was greater than 0.05.

OLS Regression Results

Dep. Variable:	Average of G13	**R-squared:**	0.971
Model:	OLS	**Adj. R-squared:**	0.965
Method:	Least Squares	**F-statistic:**	176.8
Date:	Sun, 08 Dec 2019	**Prob (F-statistic):**	1.88e-26
Time:	17:11:45	**Log-Likelihood:**	5.7261
No. Observations:	44	**AIC:**	2.548
Df Residuals:	37	**BIC:**	15.04
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PSI Avg	-0.1421	0.061	-2.318	0.026	-0.266	-0.018
Total International Visitor Arrivals	-0.0589	0.085	-0.692	0.493	-0.231	0.113
YOY GDP Growth Rate	0.0359	0.041	0.884	0.382	-0.046	0.118
Average of A68U	1.1957	0.369	3.241	0.003	0.448	1.943
Average of C6L	-0.0239	0.031	-0.775	0.443	-0.087	0.039
Average of H15	-0.2498	0.106	-2.350	0.024	-0.465	-0.034
Average of M04	0.5879	0.158	3.731	0.001	0.269	0.907

Figure 5.8 OLS regression result

```
Mean Absolute Error: 0.16336200933909206
Mean Squared Error: 0.04513249785536299
Root Mean Squared Error: 0.2124441052497409
```

Figure 5.9 Mean squared errors produced in OLS model

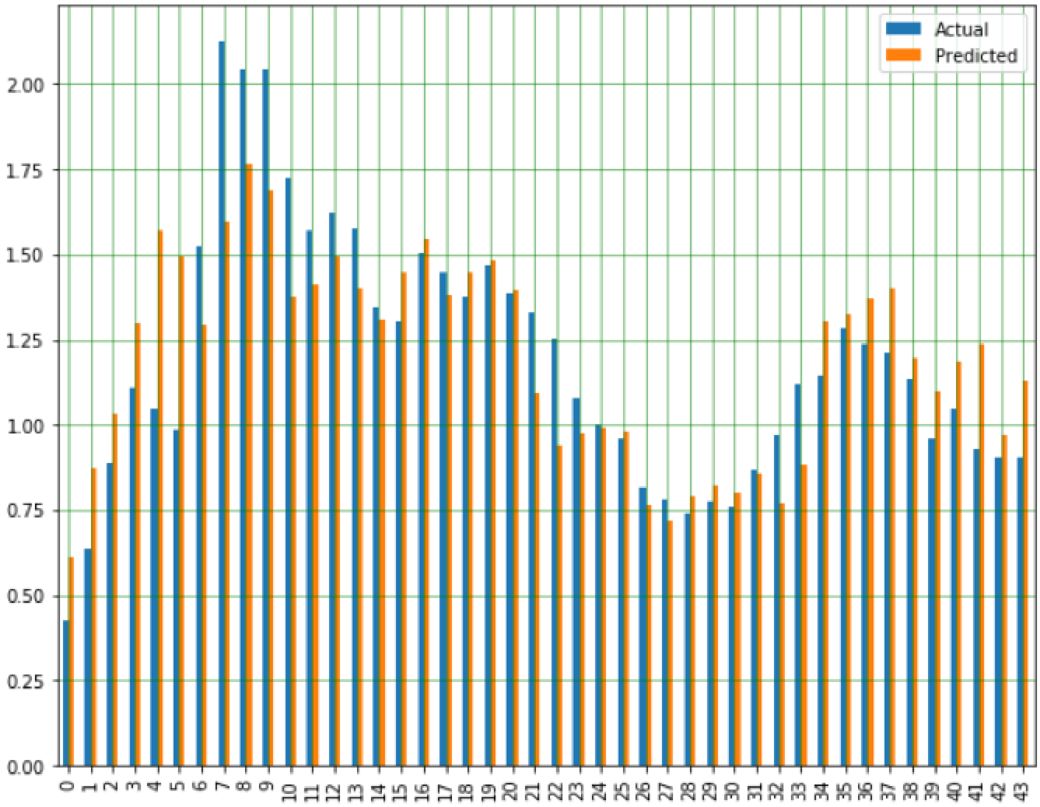

Figure 5.10 Actual value vs predicted value forecasted by OLS model

Extra analysis was carried out to investigate the relationship between PSI and the other stocks' prices. It turned out that most of the stocks' prices were not significant related to PSI except for stock H15, which was having p-value that's less than 0.05 (refer to figure 5.11, figure 5.12, figure 5.13, figure 5.14).

OLS Regression Results

Dep. Variable:	Average of H15	**R-squared:**	0.991
Model:	OLS	**Adj. R-squared:**	0.989
Method:	Least Squares	**F-statistic:**	581.8
Date:	Tue, 10 Dec 2019	**Prob (F-statistic):**	7.77e-36
Time:	17:17:05	**Log-Likelihood:**	-10.390
No. Observations:	44	**AIC:**	34.78
Df Residuals:	37	**BIC:**	47.27
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PSI Avg	0.2555	0.085	3.015	0.005	0.084	0.427
Total International Visitor Arrivals	0.0366	0.123	0.297	0.768	-0.213	0.287
YOY GDP Growth Rate	-0.0710	0.058	-1.222	0.229	-0.189	0.047
Average of G13	-0.5197	0.221	-2.350	0.024	-0.968	-0.072
Average of A68U	2.7140	0.406	6.693	0.000	1.892	3.536
Average of C6L	-0.0105	0.045	-0.234	0.816	-0.101	0.080
Average of M04	0.5573	0.250	2.226	0.032	0.050	1.065

Figure 5.11 OLS model with H15 share price as target variable

OLS Regression Results

Dep. Variable:	Average of A68U	**R-squared:**	0.994
Model:	OLS	**Adj. R-squared:**	0.993
Method:	Least Squares	**F-statistic:**	926.5
Date:	Tue, 10 Dec 2019	**Prob (F-statistic):**	1.52e-39
Time:	17:20:43	**Log-Likelihood:**	46.789
No. Observations:	44	**AIC:**	-79.58
Df Residuals:	37	**BIC:**	-67.09
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PSI Avg	-0.0235	0.026	-0.923	0.362	-0.075	0.028
Total International Visitor Arrivals	0.0687	0.032	2.166	0.037	0.004	0.133
YOY GDP Growth Rate	0.0106	0.016	0.658	0.515	-0.022	0.043
Average of G13	0.1849	0.057	3.241	0.003	0.069	0.301
Average of H15	0.2018	0.030	6.693	0.000	0.141	0.263
Average of C6L	0.0376	0.011	3.562	0.001	0.016	0.059
Average of M04	-0.1137	0.070	-1.619	0.114	-0.256	0.029

Figure 5.12 OLS model with A68U share price as target variable

OLS Regression Results

Dep. Variable:	Average of C6L	**R-squared:**	0.990
Model:	OLS	**Adj. R-squared:**	0.988
Method:	Least Squares	**F-statistic:**	539.3
Date:	Tue, 10 Dec 2019	**Prob (F-statistic):**	3.12e-35
Time:	17:24:11	**Log-Likelihood:**	-67.485
No. Observations:	44	**AIC:**	149.0
Df Residuals:	37	**BIC:**	161.5
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PSI Avg	-0.2010	0.345	-0.583	0.563	-0.899	0.498
Total International Visitor Arrivals	-2.3066	0.246	-9.376	0.000	-2.805	-1.808
YOY GDP Growth Rate	0.2017	0.214	0.941	0.353	-0.233	0.636
Average of G13	-0.6674	0.861	-0.775	0.443	-2.412	1.077
Average of H15	-0.1410	0.601	-0.234	0.816	-1.359	1.077
Average of A68U	6.7846	1.905	3.562	0.001	2.926	10.644
Average of M04	3.1486	0.827	3.806	0.001	1.472	4.825

Figure 5.13 OLS model with C6L share price as target variable

OLS Regression Results

Dep. Variable:	Average of M04	**R-squared:**	0.986
Model:	OLS	**Adj. R-squared:**	0.984
Method:	Least Squares	**F-statistic:**	382.0
Date:	Tue, 10 Dec 2019	**Prob (F-statistic):**	1.69e-32
Time:	17:25:41	**Log-Likelihood:**	10.881
No. Observations:	44	**AIC:**	-7.762
Df Residuals:	37	**BIC:**	4.727
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
PSI Avg	0.0174	0.058	0.299	0.767	-0.101	0.135
Total International Visitor Arrivals	0.2293	0.066	3.466	0.001	0.095	0.363
YOY GDP Growth Rate	-0.0444	0.036	-1.241	0.222	-0.117	0.028
Average of G13	0.4651	0.125	3.731	0.001	0.213	0.718
Average of H15	0.2119	0.095	2.226	0.032	0.019	0.405
Average of A68U	-0.5816	0.359	-1.619	0.114	-1.310	0.146
Average of C6L	0.0894	0.023	3.806	0.001	0.042	0.137

Figure 5.14 OLS model with M04 share price as target variable

Besides Python, the score function in the SAS Enterprise Miner was deployed to predict the Genting share price using the linear regression model. The first 10 rows of data were extracted from the datafile and was imported into the SAS Enterprise Miner. Linear regression was used and score function was applied in order to compare the predicted share price and the actual share price (figure 5.15). The result in figure 5.16. showed that the variance of predicted Genting share price to the actual price can be as close as 9% or at the maximum gap of 50%.

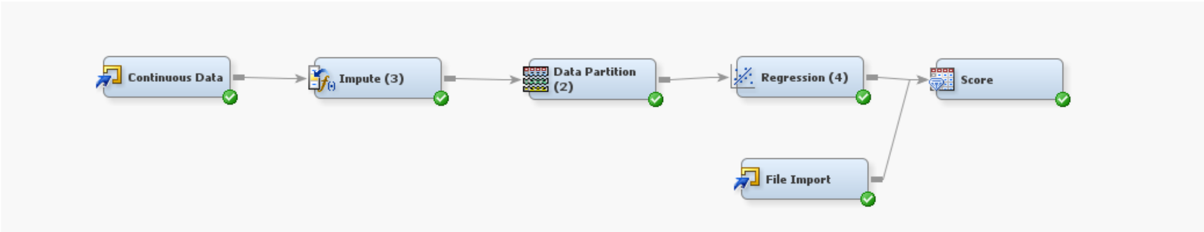

Figure 5.15 Score function with linear regression model on Genting share price

Ob... ▲	Date	PSI_Avg	...	...	Imputed: Total Int...	Imputed: YOY ...	Average_of_A68U	Average_of_C6L	Average_of_H15	Average_of_M04	Average_of_G13	Predicted: Average_of_G13
1	2009-Qtr1	-1.02	...	...	-2.15	-2.84	0.420563	10.5577	0.997131	0.749211	0.425223	0.544633
2	2009-Qtr2	-1.02	...	...	-2.15	-1.33	0.543369	11.865	1.388906	1.075301	0.639106	0.703298
3	2009-Qtr3	-0.5	...	...	-1.76	-0.4	0.792262	13.22431	2.009846	1.260672	0.887738	0.868427
4	2009-Qtr4	-0.94	...	...	-1.59	0.53	0.988924	13.88	2.124426	1.243763	1.111475	1.171342
5	2010-Qtr1	-0.91	...	...	-1.52	2.67	1.138073	14.82852	2.169508	1.347203	1.04541	1.424079
6	2010-Qtr2	-1.03	...	...	-1.32	3.28	1.065912	14.80548	2.352581	1.394387	0.986048	1.368202
7	2010-Qtr3	-1.18	...	...	-1.03	1.37	1.075943	15.51406	2.799844	1.365845	1.52625	1.245735
8	2010-Qtr4	-0.9	...	...	-0.99	2.07	1.122821	15.80719	2.837656	1.834341	2.126094	1.401419
9	2011-Qtr1	-1.1	...	...	-0.92	1.14	1.110006	14.2771	2.577097	1.992852	2.045161	1.524882
10	2011-Qtr2	-0.89	...	...	-0.75	-0.33	1.09356	14.09774	2.135161	1.862024	2.041613	1.541194
11	2011-Qtr3	-0.76	...	...	-0.4	0.74	1.039547	12.22266	2.102578	1.34966	1.722578	1.430008

Figure 5.16 Result of score function

CHAPTER 6: CONCLUSIONS

In a nutshell, haze outbreak causes losses to some of the industries, especially tourism industry. The OLS linear regression model in this research had revealed that Genting Singapore's share price, which has the largest market capitalization value in the tourism sector, was affected by the PSI value and was negatively correlated to it. Other tourism stocks' share price such as Hotel Properties (H15) was affected by the PSI value as well. Thus, the issue of haze should be taken seriously by the countries and a long-term solution should be discussed to prevent further loss in term of economics and to the society. On the contrary, other factors like number of tourist arrival and year-on-year GDP growth rate which were deemed to impact the Genting share price turned out to be correlated insignificantly to the Genting share price. The Genting share price was affected by some of its competitor stock prices (A68U, M04 and H15) as well. The classification models, interactive decision tree and logit regression model, were used to predict the movement trend of Genting share price in this research. However, unlike OLS model which achieved accuracy rate of 95%, the interactive decision tree and logit regression models only achieved accuracy rate of 69% and 41%. This might due to small sample size (only 44 rows of data) or the nature of the dataset is not suit for the classification model. It's recommended that more data sample is collected for a better accuracy rate and more attributes can be investigated to improve the completeness and accuracy of the models.

In this project, the public sentiments about the recent haze outbreak mostly were neutral and positive due to the tweets collection period was at the end of the September 2019 which was almost the end of the phenomenon. In order to have better understanding on the public sentiments, more tweets should be collected when the haze outbreak is just started. Else, Google trend can be used instead of collecting the tweets as Twitter has limitation in the collection period (not many historical tweets can be collected).

CHAPTER 7: REFERENCES

Gross Domestic Product in Chained (2015) Dollars, Year on Year Growth Rate, Quarterly. (n.d.). Retrieved from https://data.gov.sg/dataset/gross-domestic-product-in-chained-2015-dollars-year-on-year-growth-rate-quarterly.

HazeShield: Swiss Re. (2017, July 11). Retrieved from https://corporatesolutions.swissre.com/innovative-risk-solutions/non-physical-damage-business-interruption/hazeshield.html.

Pollutant Standards Index. (2019, September 25). Retrieved from https://en.wikipedia.org/wiki/Pollutant_Standards_Index.

Pollutant Standards Index (PSI). (2016, April 8). Retrieved from https://data.gov.sg/dataset/psi.

Singapore Department Of Statistics: SingStat Table Builder - Variables/Time Period Selection. (n.d.). Retrieved from https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=1991.

Tourism-Related Stocks Ride Growth in Visitor Arrivals: SGX My Gateway. (2019, January 31). Retrieved from https://sginvestors.io/sgx-mygateway/2019/01/tourism-related-stocks-ride-growth-in-visitor-arrivals.

Twitter. It's what's happening. (n.d.). Retrieved from https://twitter.com/.

Yahoo Finance – stock market live, quotes, business & finance news. (n.d.). Retrieved from https://sg.finance.yahoo.com/.