

MILESTONE 4: INTERPRETION OF DATA

Link to Youtube: <https://youtu.be/5BxlmfrkF4k>

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

Decision Tree and Regression Models Application on Stock Price Data

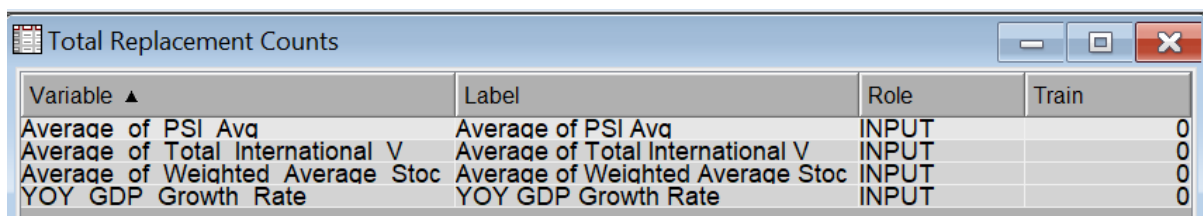
The main goal of this project is to determine the impact of haze on the Singapore's tourism stock price data. At the same time, other variables such as number of tourist arrival at Singapore and year on year Singapore GDP Growth Rate were analysed to see the relationship between these variables and the tourism stock price. These data were extracted from year 2009 until year 2019. Two models were deployed in this project, which are Decision Tree and Linear Regression Model.

Top 5 tourism stock and their daily historical prices were extracted and weightage was calculated based on their market capitalism (Table 1). These top 5 stock prices were aggregated and their weighted average stock price were calculated.

Table 1 Top 5 Tourism Stocks in Singapore (Source: <https://sginvestors.io/sgx-mygateway/2019/01/tourism-related-stocks-ride-growth-in-visitor-arrivals>)

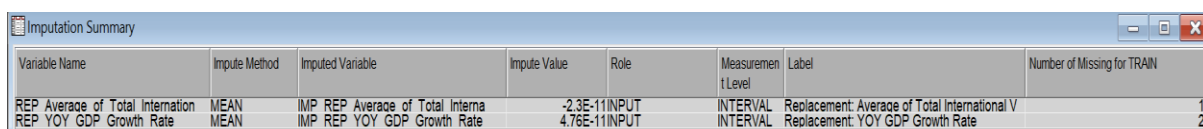
Stock Name	SGX Code	Market Cap (S\$ mln)	Weigtage Calculated
Genting Singapore	G13	13,009	40%
Singapore Airlines	C6L	11,541	36%
Mandarin Oriental	M04	3,259	10%
Ascott Residence Trust	A68U	2,576	8%
Hotel Properties	H15	1,937	6%

Before applying the models, the data were examined to check the existence of any outliers or missing values. Replacement tool from SAS Enterprise Miner was used to replace extreme percentile data as missing value. Impute tool was used to replace the missing values with the means of the non-missing values of that category. The variable with new input value will be named as IMP_original_variable_name with missing value replaced by the new generated input and non-missing values copied from the original input. Results showed that there's no any extreme value in the dataset but there were missing values (Figure 1 and Figure 2). After transforming the data, data partition was carried out to split the total of 44 observations into 50% training data and 50% validation data (Figure 3).



Variable ▲	Label	Role	Train
Average of PSI Avg	Average of PSI Avg	INPUT	0
Average of Total International V	Average of Total International V	INPUT	0
Average of Weighted Average Stoc	Average of Weighted Average Stoc	INPUT	0
YOY GDP Growth Rate	YOY GDP Growth Rate	INPUT	0

Figure 1 Replacement tool showing no extreme values in the dataset



Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP Average of Total Internation	MEAN	IMP REP Average of Total Interna	-2.3E-11INPUT	INTERVAL	INTERVAL	Replacement: Average of Total International V	1
REP YOY GDP Growth Rate	MEAN	IMP REP YOY GDP Growth Rate	4.76E-11INPUT	INTERVAL	INTERVAL	Replacement: YOY GDP Growth Rate	2

Figure 2 Impute tool showing missing values in the dataset

Partition Summary

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

Figure 3 Data partition

Next, interactive Decision Tree Model was built to predict the upward and downward trend of stock price. Average of PSI readings and number of tourist arrival were part of the split rules (Figure 4). However, the subtree assessment plot in Figure 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.

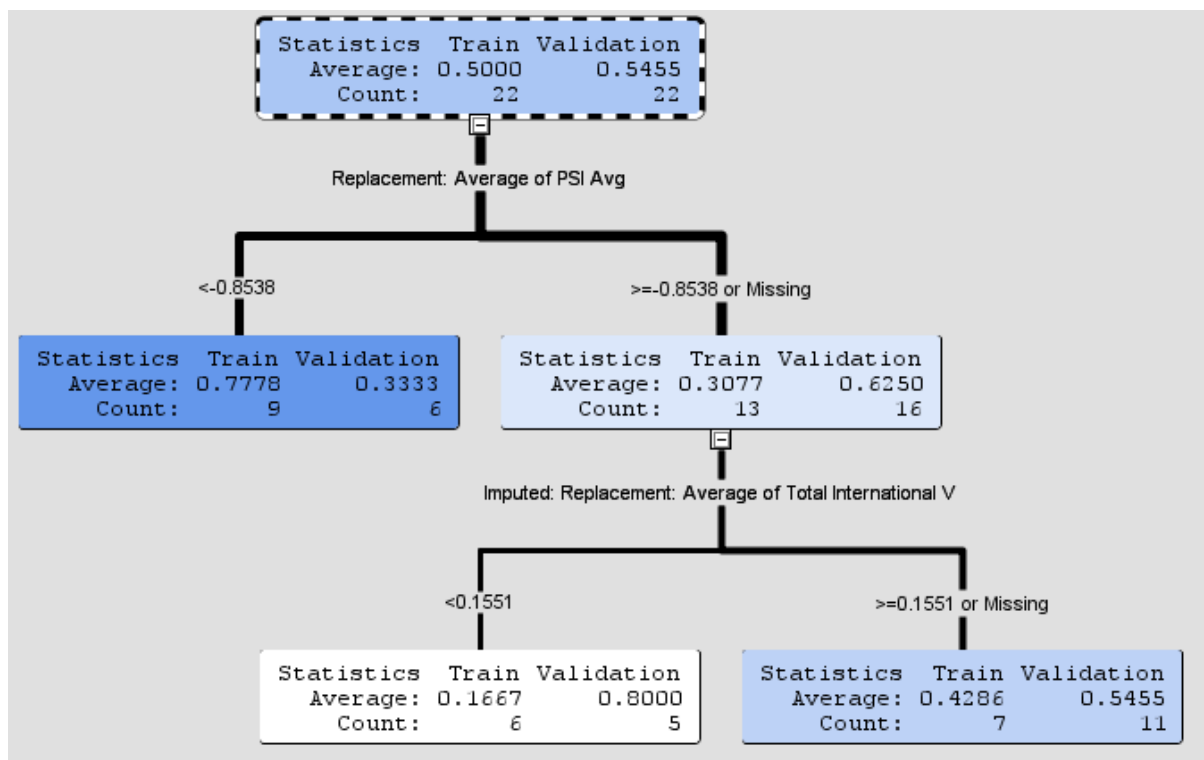


Figure 4 Interactive Decision Tree Model

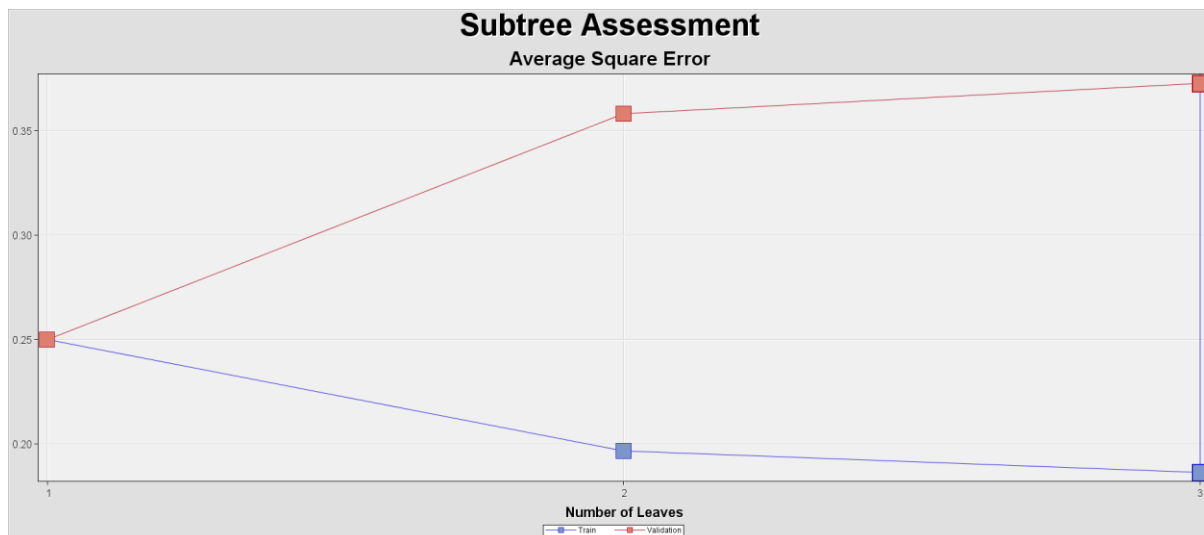


Figure 5 Subtree Assessment Plot

Beside Decision tree model, the correlation between the variables were studied using correlation matrix and regression model. Correlation among the variables and correlation matrix was run in Python and result was shown in Figure 6 and Figure 7. The stock price was negatively correlated to PSI readings and number of visitor arrival but positively proportional to GDP growth rate.

	PSI Avg	Weighted Average Stock Price	Visitor Arrivals	YOY GDP Growth Rate	Trend
PSI Avg	1.000000	-0.481728	0.616889	-0.328231	-0.086786
Weighted Average Stock Price	-0.481728	1.000000	-0.555806	0.731913	0.150233
Visitor Arrivals	0.616889	-0.555806	1.000000	-0.238719	-0.194298
YOY GDP Growth Rate	-0.328231	0.731913	-0.238719	1.000000	-0.079573
Trend	-0.086786	0.150233	-0.194298	-0.079573	1.000000

Figure 6 Correlation table among the variables

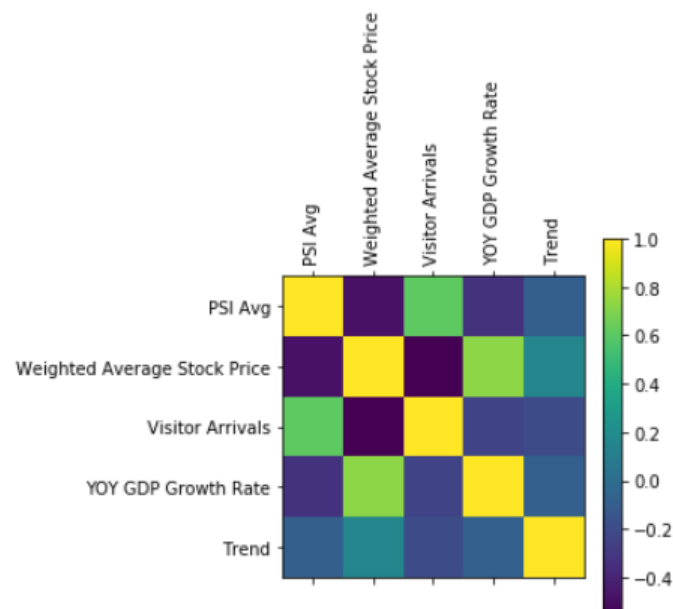


Figure 7 Correlation matrix plot

Linear regression model was applied on this dataset to determine the variables that effect stock price data and for prediction purpose. The regression result in Figure 8 revealed that p-value was less than 0.05, indicating that the group of independent variables (PSI readings, GDP growth rate and number

of tourist arrivals) able to predict the dependent variable (stock price) reliably at 5% significant level. R-square value of 0.7057 also indicated that 70.57% of the variance in stock price can be predicted from these independent variables. Nevertheless, this was an overall measure of the strength of association, and did not reflect the extent to which any particular independent variable was correlated with the dependent variable. The analysis of maximum likelihood estimates, on the other hand, showed the ability of each individual independent variable to predict the dependent variable. Variables such as number of tourist arrival and GDP Growth were statistically significant as they had p-value less than 0.05. PSI data was not statistically significant as its p-value was greater than 0.05. The score rankings matrix assesses the model's performance for different score values. The means of the predicted values were fluctuated around the means of the target values for each bin, showing that the regression model built was still good fit in overall.

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	3	10.173211	3.391070	14.39	<.0001	
Error	18	4.242132	0.235674			
Corrected Total	21	14.415343				

Model Fit Statistics			
R-Square	0.7057	Adj R-Sq	0.6567
AIC	-28.2115	BIC	-24.5325
SBC	-23.8473	C(p)	4.0000

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Pr > t	
Intercept	1	-0.0881	0.1063	-0.83	0.4181	
IMP_REP_Average_of_Total_Interna	1	-0.3909	0.1541	-2.54	0.0206	
IMP_REP_YOY_GDP_Growth_Rate	1	0.6511	0.1137	5.73	<.0001	
REP_Average_of_PSI_Avg	1	0.0518	0.1110	0.47	0.6463	

Figure 8 Regression Model Result

Tweets Sentiment Analysis

Besides predicting the stock price trend and its value, the project also aimed to understand the public sentiment towards haze phenomena. Tweets that contained key words like *haze* and *jerebu* were collected from 20th September 2019 to 9th October 2019. 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.

Positivity of Tweet Sentiment

A pie chart titled "Positivity of Tweet Sentiment" illustrating the distribution of sentiment in tweets. The chart is divided into three segments: a large blue segment for "Neutral" at 48.77%, a large red segment for "Positive" at 41.51%, and a small teal segment for "Negative" at 9.72%. Each segment is labeled with its category and percentage.

Sentiment	Percentage
Neutral	48.77%
Positive	41.51%
Negative	9.72%

Figure 9 Positivity of Tweets' Sentiment

Word Cloud for Positive Comment

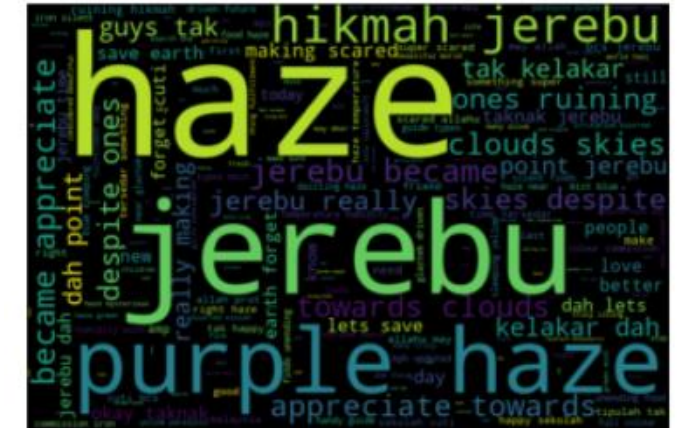


Figure 10 Word Cloud for Positive Comments

Word Cloud for Negative Comment



Figure 11 Word Cloud for Negative Comments

Word Cloud for Neutral Comment

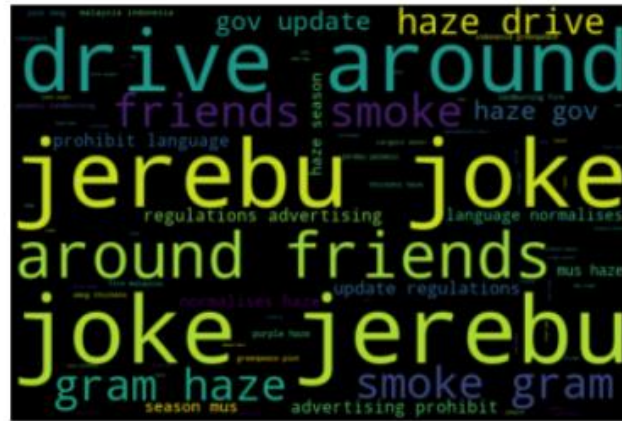


Figure 12 Word Cloud for Neutral Comments

Figure 9 showed the positivity of tweets' sentiments. 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%. The positive tweets contained terms such as 'appreciate', 'better' and 'hikmah' while the negative tweets were having terms like 'worse breath', 'getting worse' and 'confused' as displayed in word clouds (figure 10 and figure 11). On the other hand, the word cloud of tweets with neutral sentiment consisted of terms like 'friends', 'drive around' and 'joke' that carried no emotion (figure 12). The period of tweets collected was at the end of September and beginning of October year 2019, in which the haze phenomena 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.

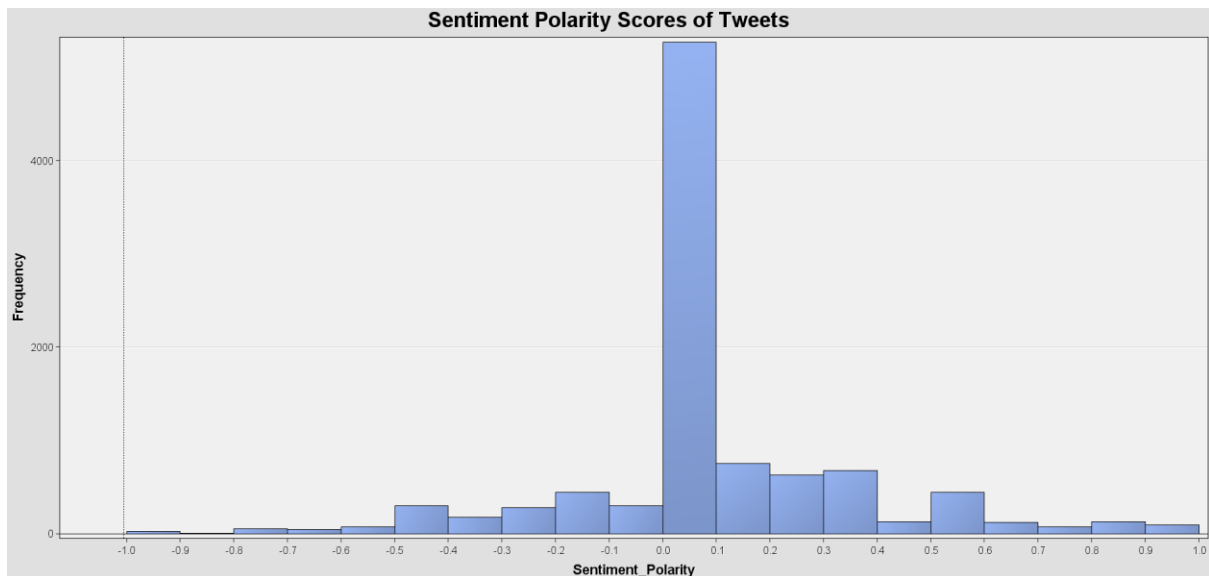


Figure 13 Sentiment Polarity of Tweets collected

The histogram in figure 13 indicated that the distribution was slightly skewed to the right, which matched with the sentiment category of the tweets that most of the tweets collected were in neutral and positive sentiments. Most of the polarity scores fall between 0 and 0.1, in which about 45,987 tweets were neutral sentiment. Around 33,776 tweets were having polarity score more than 0.2 while there were only 8,590 tweets with scores less than zero.