MILSTONE 3: DATA PROCESSING

Link to Youtube: https://youtu.be/YAVwkkRMKlQ

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

INTRODUCTION

Southeast Asia countries suffer from haze occasionally that caused by illegal agricultural fires due to industrial-scale slash-and-burn practices in Indonesia. Few studies had proved that this smoke event caused few billion loss in economy and income opportunities. Upon on the event of haze that just happened last month, an analysis about the impact of haze on tourism industry and public sentiment to haze are carried out. Singapore is chosen as our analysis target location due to the availability and completeness of data in this country.

The objectives of this analysis are:

1. To determine the relationship between Pollutant Standard Index (PSI) reading and the tourism stock market prices

2. To identify the effect of haze on the number of tourist arrival

3. To understand the public sentiment to haze through Twitter

DATA DESCRIPTION

Multiple datasets were crawled or extracted from different data sources in order to achieve the analysis goals mentioned. These are the datasets:

Dataset	Description	Data Source
PSI Data	The PSI data was collected on hourly basis and by major regions in Singapore from year 2009 January to year 2019 October.	https://data.gov.sg/dataset/psi
Stock Price Data	The top 5 tourism stock share in Singapore were identified and their stock prices were collected on daily basis from year 2009 January to year 2019 October.	https://sg.finance.yahoo.com/
International Tourist Arrival Data	The number of international visitor arrival was recorded from year 1978 January to year 2019 August on monthly basis.	https://www.tablebuilder.singstat.gov.sg/publicfacing/createDataTable.action?refId=1991
Tweets Data	Tweets regarding key words like *haze* and *jerebu* were collected from 20th September 2019 to 9th October 2019.	https://twitter.com/

DATA PRE-PROCESSING

The data were compiled into one csv file and cleaning such as removing missing data was carried out. After cleaning, PSI data, stock price and tourist arrival number 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.

After data cleaning and pre-processing in Excel, the file was split into three files which are:

I. PSI data and stock price data file
II. PSI data and tourist arrival data file
III. Tweet Data

These three files were imported into SAS Enterprise Miner for further analysis.

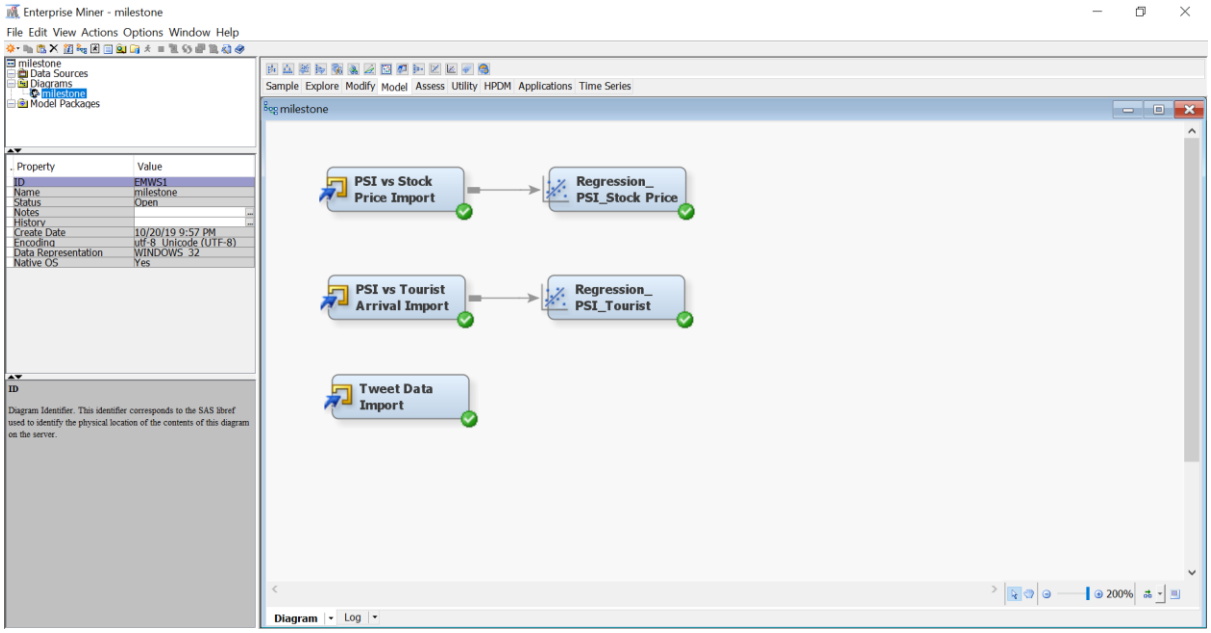
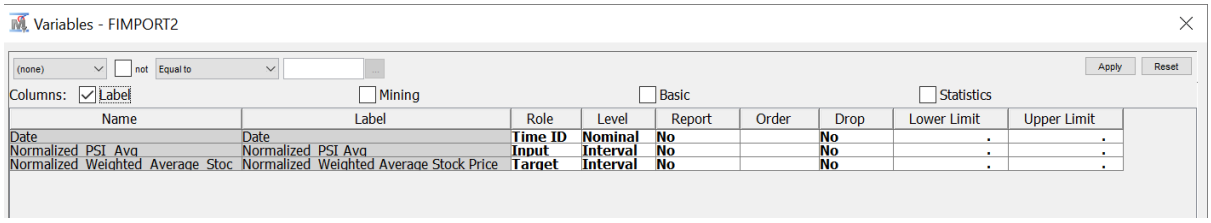


Figure 1 Three files that imported to SAS Enterprise Miner

The column metadata of each file are shown as below:

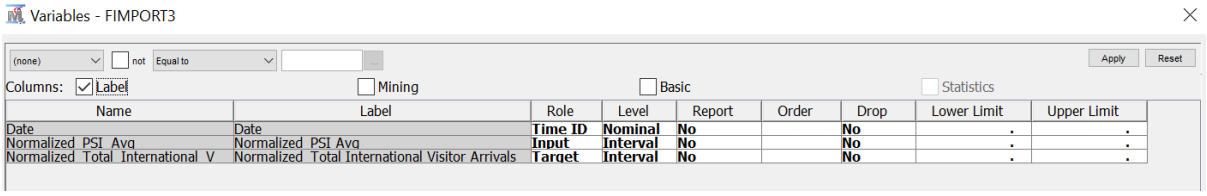


Name	Label	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Date	Date	Time ID	Nominal	No		No	.	.
Normalized_PSI_Avg	Normalized_PSI_Avg	Input	Interval	No		No	.	.
Normalized_Weighted_Average_Stoc	Normalized_Weighted_Average_Stock_Price	Target	Interval	No		No	.	.

Figure 2 PSI and Average Stock Price Data

- Date: Period of the data that were aggregated to monthly basis from year 2009 to year 2019.
- Normalized PSI Average: The daily PSI data from year 2009 to year 2019 were aggregated to average monthly value and was normalized.

- Normalized Weighted Average Stock Price: The weighted average value of the top 5 share data was calculated based on their market capitalism in the share market and the data was normalized. The stock prices collected were from year 2009 to year 2019.


Variables - FIMPORT3

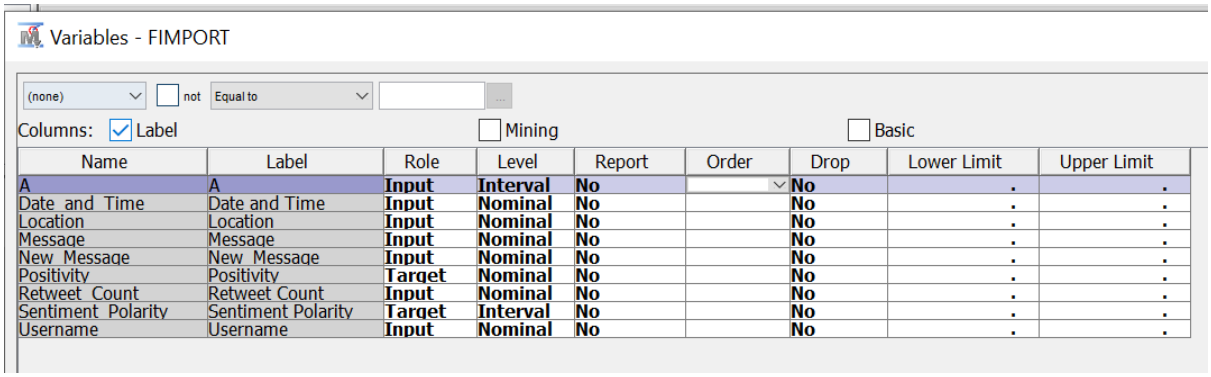
(none) not Equal to Apply Reset

Columns: Label Mining Basic Statistics

Name	Label	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
Date	Date	Time ID	Nominal	No		No	.	.
Normalized_PSI_Avg	Normalized_PSI_Avg	Input	Interval	No		No	.	.
Normalized_Total_International_V	Normalized_Total International Visitor Arrivals	Target	Interval	No		No	.	.

Figure 3 PSI and Visitor Arrival Data

- Date: Period of the data that were aggregated to monthly basis from year 2009 to year 2019.
- Normalized PSI Average: The daily PSI data from year 2009 to year 2019 were aggregated to average monthly value and was normalized.
- Normalized International Visitor Arrival: The monthly visitor arrival data from year 2009 to year 2019 were normalized.


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 4 Tweet Data

- Date and Time: Period of tweets that were collected.
- Location: The locations of the tweets that were posted.
- Message: The tweet message regarding haze collected.
- New Message: The cleaned tweet message after removing stop words and underwent stemming and lemmatization.
- Positivity: The sentiments detect from the tweet messages. It was categorized into Positive, Negative and Neutral using the TextBlob package in Python.
- Retweet Count: The frequency of retweet the tweet message.
- Sentiment Polarity: The sentiment polarity score of the tweet message was calculated using the TextBlob package in Python.
- Username: The username used to post the tweet message.

DATA VISUALISATION

Graphs are plotted against the three sets of data to describe the distribution of the dataset and to visualize the relationship between the variables. Word Clouds are used to highlights the public thoughts about this smoke event.

I. PSI and stock price data

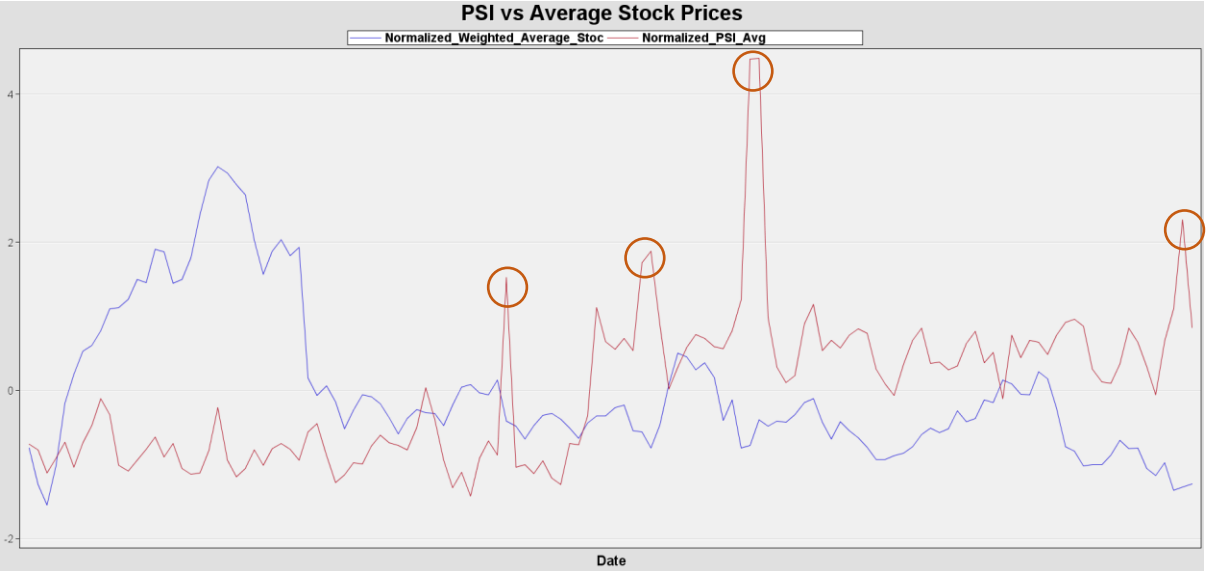


Figure 5 PSI and Average Stock Prices over the years 2009 to year 2019

Figure 5 demonstrates both PSI and stock prices fluctuate over the years. However, the PSI data in overall shows an increasing trend while the stock prices show a decreasing trend over the years. When the PSI are at its spike (as circled), the stock prices dropped. The stock prices are at its highest level when the PSI data are at lower value.

II. PSI and tourist arrival data

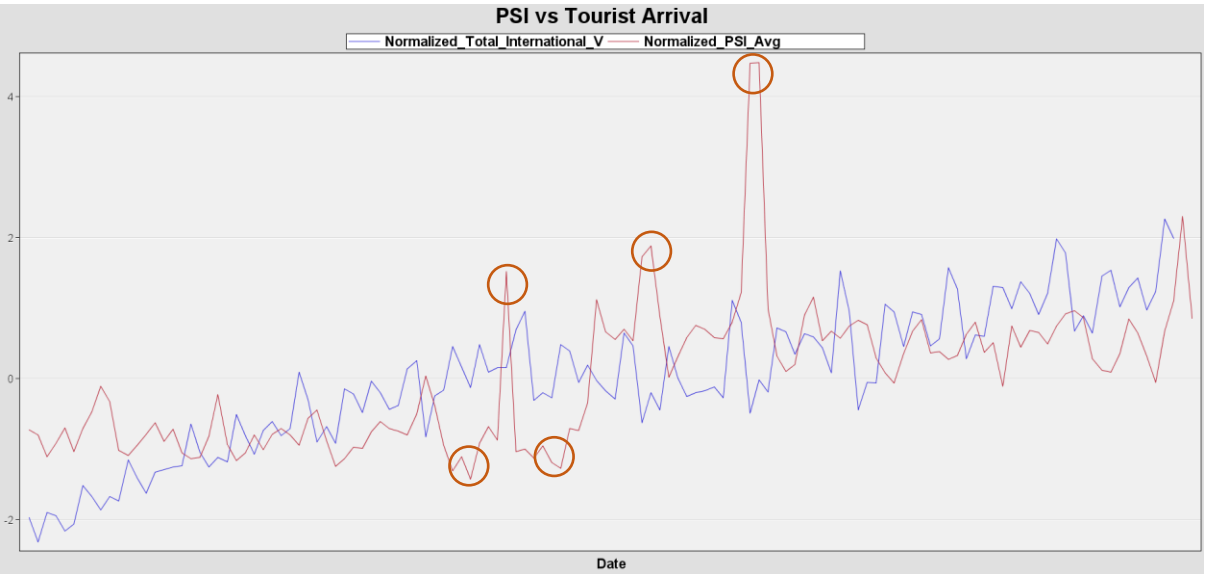


Figure 6 PSI and Tourist Arrival data over the years 2009 to year 2019

Figure 6 displays both PSI and tourist arrival data are in an upward trend over the years. Similar to the relationship between PSI and stock price data, number of tourists arrived is inversely proportional to the PSI data. When the PSI are at its spike (as circled), the number of tourists fell. The tourist's arrival data is increasing when the PSI value dropped.

III. Tweet Sentiments

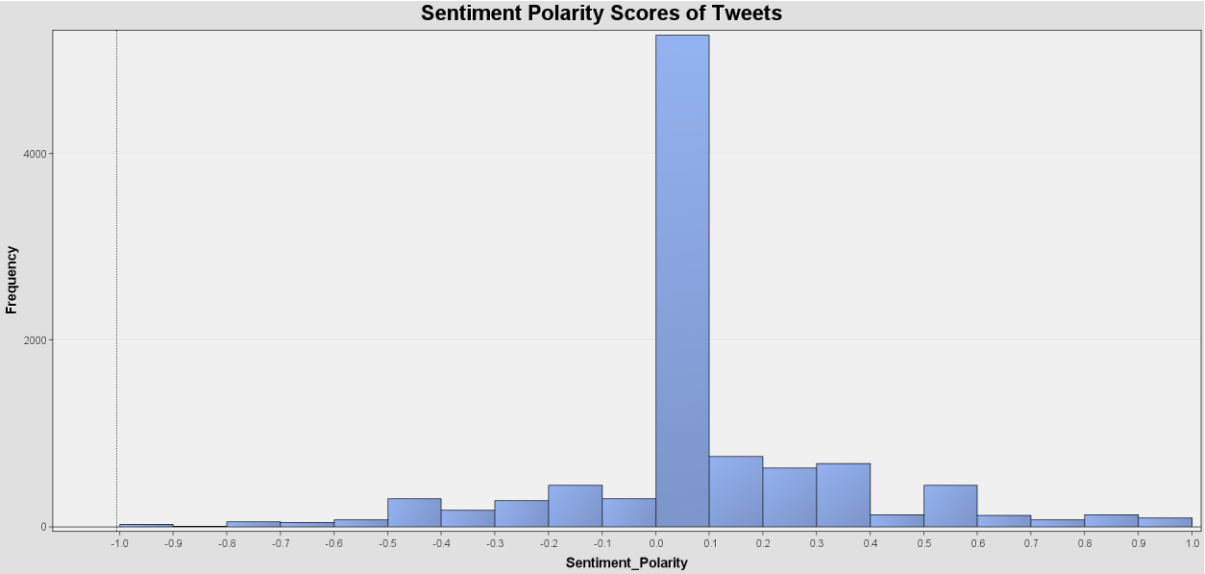


Figure 7 Sentiment Polarity of Tweets collected

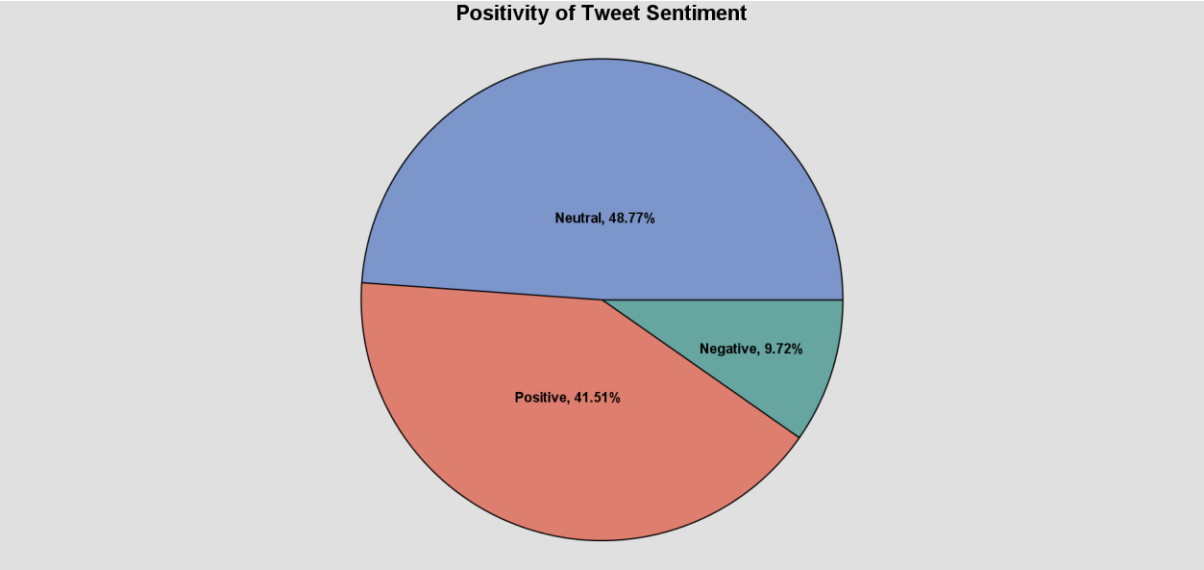


Figure 8 Positivity of Tweet Sentiment

Tweets messages discussed about *haze* or *jerebu* were captured and the sentiments in the text message were analysed. The sentiments were made up of Positive, Negative and Neutral sentiment and the sentiment polarity score were calculated. The histogram shows that the distribution was slightly skewed to the right, which surprisingly indicates that most of the tweets collected were in neutral and positive sentiments. Most of the polarity score fall between 0 and 0.1, which were neutral sentiment. Pie chart clearly reveals that 48.77% of the text message is neutral in sentiment while the rest of messages show emotion in which positive messages made up of 41.51% and negative messages made up of 9.72%. 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. Missing data appeared in the sentiment polarity scores which were displayed in the histogram as missing bins.

Figure 9 Word Cloud of Positive Tweet Messages

Figure 10 Word Cloud of Negative Tweet Messages

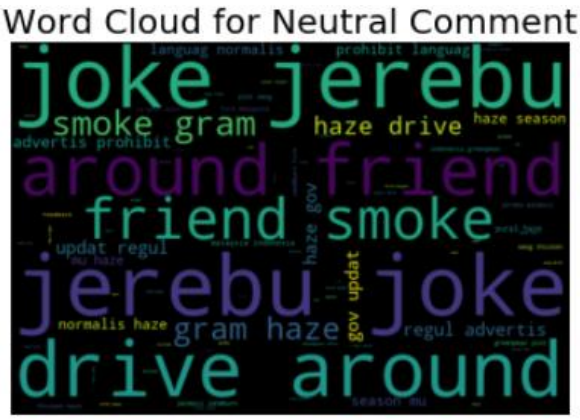


Figure 11 Word Cloud of Neutral Tweet Messages

The word clouds by sentiments generated from the tweet messages reveals that words such as ‘haze’ and ‘jerebu’ were mentioned the most in all these three sentiments messages. In the positive comment’s word cloud, words such as ‘appreciate’ appeared as people were commenting that they’re appreciating the clear air and sky when the haze event happened. In the world could with negative comments, words such as ‘breath well’ and ‘worse breath’ were emerged showing the situation faced by the public during haze. In the world cloud with neutral sentiment, words without any sentiments such as ‘friend’, ’drive’ and ‘joke’ were found.