SINGAPORE MANAGEMENT UNIVERSITY

School of Information Systems

ANLY482 - Analytics Practicum

Project Proposal

Koh Ying Ying Trecia
Luqman Haqim Bin Ab Rahman

Table of Contents

Table of Contents

1. Version

Version	Change Description	Author	Date
1.0	Initial Draft on ALOS	Trecia, Luqman	24/01/2015
2.0	Draft for LTA	Trecia, Luqman	24/02/2015

2. Background

Singapore is a small country, yet it has a complex but comprehensive public transportation network. Consisting of train (known as Mass Rapid Transit, hereinafter known as MRT), bus, light and rapid trains (Light Rail Transport, hereinafter known as LRT), and taxis, the public transport in Singapore employs the hub-and-spoke strategy; busses serve as the means of transportation within a town, and MRT trains are used for long distance travel.

The demand for MRT ridership has significantly increased since 1997 as it served as a cheaper or faster alternative to car or taxi for long distance travel. However, since 2011 to the time of this paper, confidence in the MRT system have dropped as it has been plaque with service breakdowns. Some of these breakdowns can be as short as 45 minutes and some as long as a full day. Most Singaporeans feel that the train breakdown is attributed to the sudden increase of foreign workers in the country and that the MRT infrastructure cannot cope with the sudden increase of ridership, thus leading to the breakdowns.

Calls from the public to improve the MRT infrastructure have been a priority for the MRT operators. It is important that the operators understand the traffic patterns of the MRT ridership to be able to constructively understand and cater or improve the reliability and re-instill confidence in the MRT.

Should the MRT operators cater to the morning peak by increasing the frequency of trains in the morning, or should they increase the train frequency in the evenings when commuters end the day? Should policies be applied across all stations or should each station have different policies?

With the Government's plans to have 6.9 million citizens in Singapore by 2020, we hope to use analytics to be able to understand the travel patterns of the MRT so as to improve the MRT services.

This paper attempts to explore the travel patterns of the MRT ridership in Singapore for the first week of November of 2011. This paper will continue the work done by Roy LEE's Master Thesis and we seek to explore the areas that LEE do not cover in his Master Thesis.

3. Objective

- Business objective: To identify the MRT ridership patterns of the various station to improve the MRT services.
- Technical objective: To use data analytics techniques such like exploratory data analysis (EDA), and statistical methods to study and gain insights from the data to identify patterns that aid business objective. We will then use time series data mining methods to explore the different patterns.

4. Scope

- Perform data cleaning on the data set received to consolidate the important fields that are required for analysis.
- Perform EDA to identify patterns that will help in the study of MRT ridership.
- Use time series data mining to explore the patterns of the MRT ridership.

5. Data Required

For the project, Land Transport Authority (LTA) provides the data sets through LARC research labs. The dataset is a weeks' worth of smart card (EZ-Link) transaction used in Singapore's public transport. The data consist of both bus and also MRT transaction. For this project we will require only MRT transactions.

6. Deliverables

- A detailed report to explain the study and recommendations to improve MRT services
- A detailed description and interpretation of the analysis procedures that has been used in time series data mining.

7. Dependency

Dependency	Description
Data	Data has been retrieved from a database provided by LTA and made available for LARC research initiative. It is however a big dataset.
Technical Skills	No dependencies

8. Stakeholders

a. Project Supervisor

Prof Kam Tin Seong, Associate Professor of Information Systems; Senior Advisor, SIS Programmes in Analytics

b. Project Members

i. Koh Ying Ying Trecia
ii. Luqman Haqim Bin Ab Rahman

c. Project Sponsor

Prof Kam Tin Seong, Faculty Staff of Learning Analytics Research Centre (LARC)

9. Schedule

	Weeks/ Date	Task	Milestone
Midterm	Week 6 09/02/2015	Source and analyse projects available	
	Week 7 16/02/2015	Finalized on project topic Readings related to project Proposal development Data exploration and cleaning EDA Process Draft mid-term report	
	Week 8 23/02/2015	Finalize EDA Process Update mid-term report + power point slides + wiki Decide on tool to use Decide on time series data mining methods	Mid-term Presentation Progress Report + Wiki Due Date: 26 February 2015
	Week 9 02/03/2015	Perform time series data mining methods	
	Week 10 09/03/2015	Perform time series data mining methods with forecasting methods.	

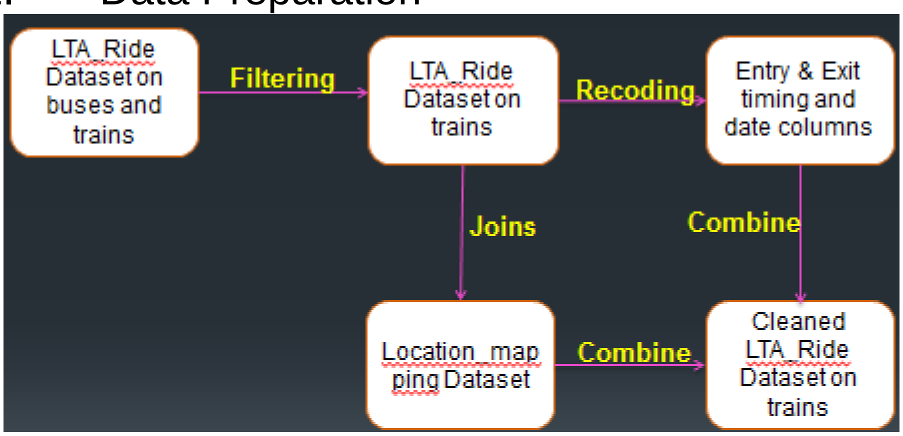
Finals	Week 11 16/03/2015	Analysis & Reporting of the results from time series data mining Draft final research paper, power point slides	
	Week 12 23/03/2015	Study all research and analysis findings Interpreting and comparing models Record findings and documentation Poster creation Update research paper + wiki	
	Week 13 30/03/2015	Finalized analysis result Finalized Research Paper Finalized wiki	
	Week 14 06/04/2015	Submission of Final Report, Poster	Final Presentation Final Report, Poster, Wiki Due Date: TBC

10. Tools Used

For data preparation, descriptive statistics, we use SAS JMP Pro and SAS Enterprise Guide. We used both tools as we are familiar with SAS Enterprise Guide as the Analytics Foundation course uses SAS Enterprise Guide; therefore we are well versed in the tool. We use SAS JMP Pro as recommended by our project supervisor as a faster alternative. However, as we use both tools interchangeable as fit the task.

For the data-mining portion, we will use SAS Enterprise Miner as the tool for time series data mining.

11. Data Preparation



The dataset provided by LARC is currently from a MySQL database. We extracted the data by taking a database dump. As we are only interested in the MRT transactions, we added a conditional statement to only include the train dataset.

These are the tables we used:

1. Location_mapping. This table contains the human readable name of a station and the date it was commissioned

2. Lta_ride: this contains the time series transaction table that contains the transaction for the first week of November 2011.

Using SAS Enterprise Guide, we performed these data preparation steps:

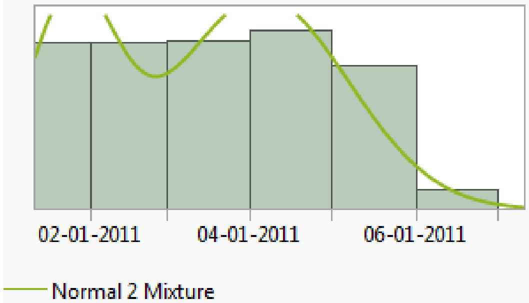
1. Extracting the hour of entry_time and exit_time. This is to analyse the hour of which ridership is the most.
2. Extracted the minutes from the entry_time and exit_time
   a. We then segregated the minutes into quarterly intervals where:
      I. 0-15 = 1
      II. 16-30 = 2
      III. 31-45 = 3
      IV. 46-59 = 4
   b. We choose a 15 minute interval as it would be less time consuming and meaningful to analyze travel patterns in quarters instead of per minute/second
3. We recoded the entry and exit time of midnight (currently represented as 00 hours) to 24. We then added a calculated field “new_duration” where we take the recoded exit time minus the recoded entry time to get an accurate duration of travel for each transaction. While the original dataset has a travel_time field, we found this to be an unreliable field as per this example from the dataset:
   a. Original
      i. entry_time: 23:45:00
      ii. exit_time: 00:10:00
      iii. This results in duration of 1400 minutes. This is incorrect.
   b. Corrected
      i. entry_time: 23:45:00
      ii. exit_time: 24:10:00
      iii. This results in duration of 25 minutes. This is correct.
4. We then extracted the day of the week from the entry_date as to understand the travel patterns. These are the extracted information
   a. 1st November 2011 is Tuesday (Weekday)
   b. 2nd November 2011 is Wednesday (Weekday)
   c. 3rd November 2011 is Thursday (Weekday)
   d. 4th November 2011 is Friday (Weekday)
   e. 5th November 2011 is Saturday (Weekend)
   f. 6th November 2011 is Sunday (Weekend)
5. Finally, we joined the lta_ride table with the location_mapping table to be able to analyze the dataset with the human readable name of the stations.
6. This has resulted in approximately 11 million rows of time series transaction data.

12. Exploratory Data Analysis for Knowledge Discovery

In this phase of the project we attempt knowledge discovery from the provided dataset using a mixture of basic statistics and visualization methods.

a. Commuters Patterns by Date

entry_date



Quantiles

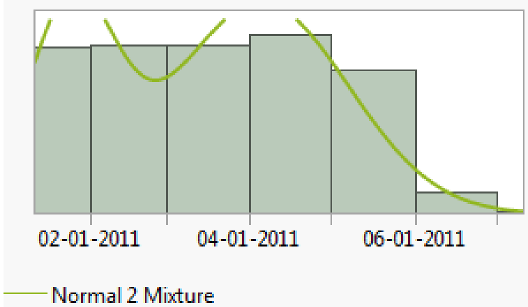
100.0%	maximum	3.39e+9
99.5%		3.39e+9
97.5%		3.39e+9
90.0%		3.39e+9
75.0%	quartile	3.39e+9
50.0%	median	3.38e+9
25.0%	quartile	3.38e+9
10.0%		3.38e+9
2.5%		3.38e+9
0.5%		3.38e+9
0.0%	minimum	3.38e+9

Summary Statistics

Mean	3.3828e+9
Std Dev	3729964.8
Std Err Mean	1123.3722
Upper 95% Mean	3.3828e+9
Lower 95% Mean	3.3828e+9
N	11024582
Skewness	0.0897487

The entry date patterns suggest that there are more ridership in the weekday compared to the weekend.

exit_date



Quantiles

100.0%	maximum	3.39e+9
99.5%		3.39e+9
97.5%		3.39e+9
90.0%		3.39e+9
75.0%	quartile	3.39e+9
50.0%	median	3.38e+9
25.0%	quartile	3.38e+9
10.0%		3.38e+9
2.5%		3.38e+9

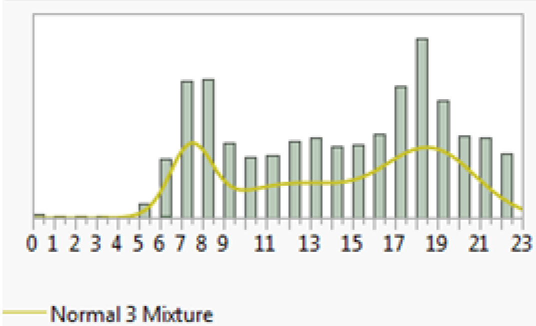
Summary Statistics

Mean	3.3828e+9
Std Dev	3739508
Std Err Mean	1126.2464
Upper 95% Mean	3.3828e+9
Lower 95% Mean	3.3828e+9
N	11024582
Skewness	0.0952385

The exit date patterns suggest that there are more ridership in the weekday compared to the weekend.

b. Commuters Patterns by Hour

hourEnter



Quantiles

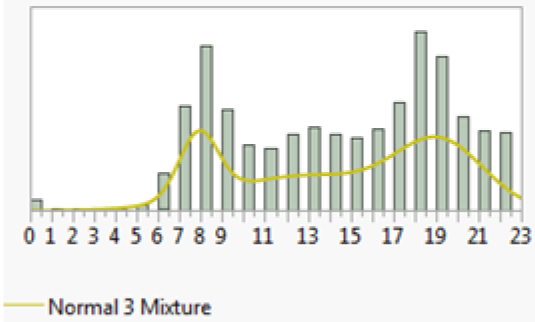
100.0%	maximum	23
99.5%		23
97.5%		22
90.0%		21
75.0%	quartile	18
50.0%	median	15
25.0%	quartile	9
10.0%		7
2.5%		6
0.5%		5
0.0%	minimum	0

Summary Statistics

Mean	14.198121
Std Dev	5.0479003
Std Err Mean	0.0015203
Upper 95% Mean	14.201101
Lower 95% Mean	14.195142
N	11024582
Skewness	-0.158553

The commuter pattern by hour for entry shows that there is a peak of commuters entering the train stations at 7 and 8 am, plateau after that and a slight increase during mid-day and then the peak is at 6pm. The first peak suggests the most common time commuters board the train to go to work. The second peak could suggest that the workers board the train to have lunch at a nearby MRT station. The last peak suggests that 6pm is the time most commuters board the train to go home.

hourExit



Quantiles

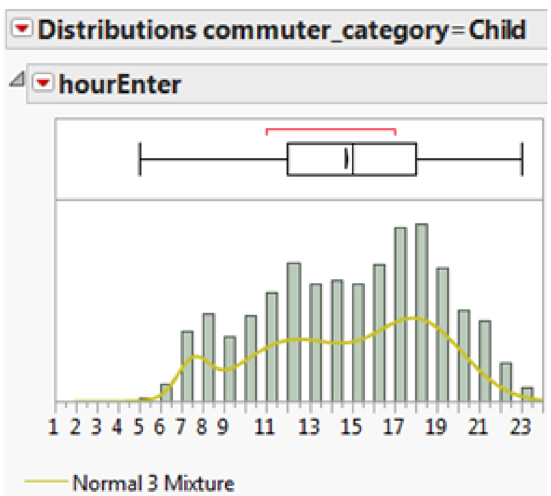
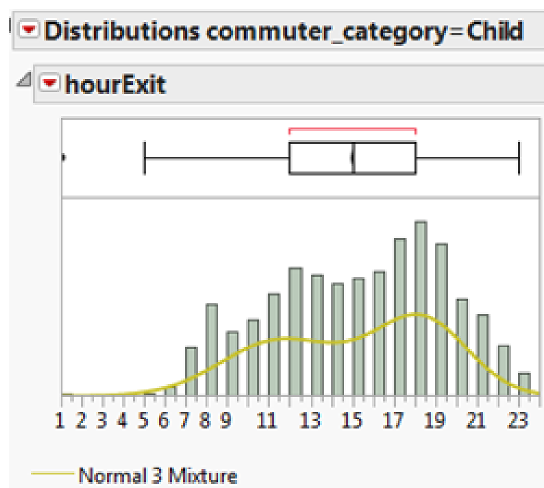
100.0%	maximum	23
99.5%		23
97.5%		23
90.0%		21
75.0%	quartile	19
50.0%	median	15
25.0%	quartile	9
10.0%		8
2.5%		6
0.5%		0
0.0%	minimum	0

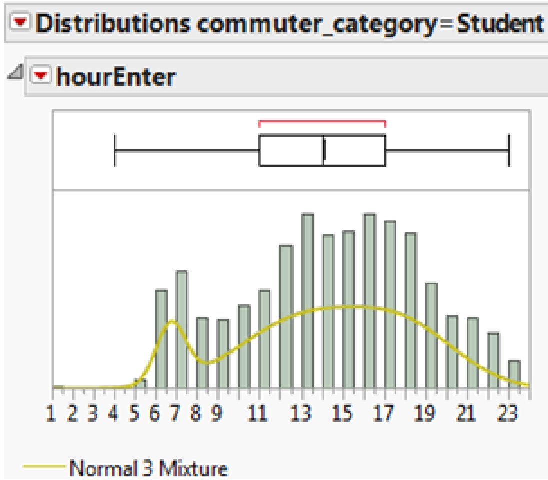
Summary Statistics

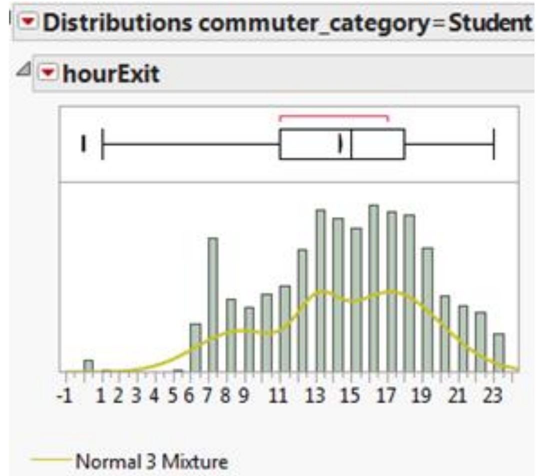
Mean	14.492401
Std Dev	5.1044312
Std Err Mean	0.0015373
Upper 95% Mean	14.495414
Lower 95% Mean	14.489388
N	11024582
Skewness	-0.221068

The exit commuter patterns are similar, if not the same as the entry pattern; therefore the analysis is the same. This also suggests that commuters' average travel time do not exceed an hour.

c. Commuters Hour Patterns grouped by Commuter Type

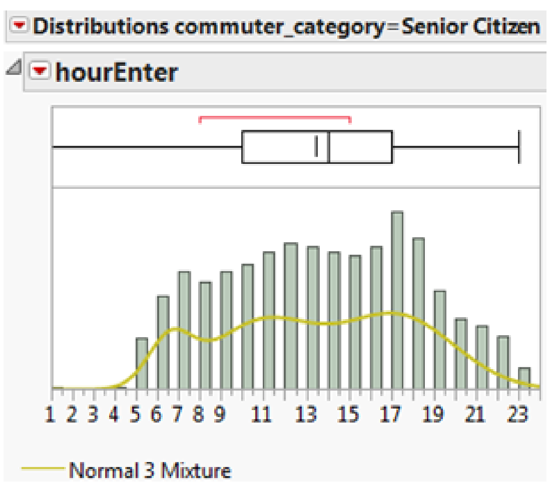
Entry Hour Grouped By Commuter Type	***Exit Hour Grouped by Commuter Type***
	
The travel patterns for child have a peak at 12pm and 6pm. This suggests that most children travel in the afternoons during the school holidays. It is also important to note that the travel pattern for child is similar, if not the same as Senior Citizens. This suggest that senior citizens who travel with their grandchildren. As mentioned earlier that November is the school holidays, this suggest that grandparents look after their grandchildren during the school holidays.	

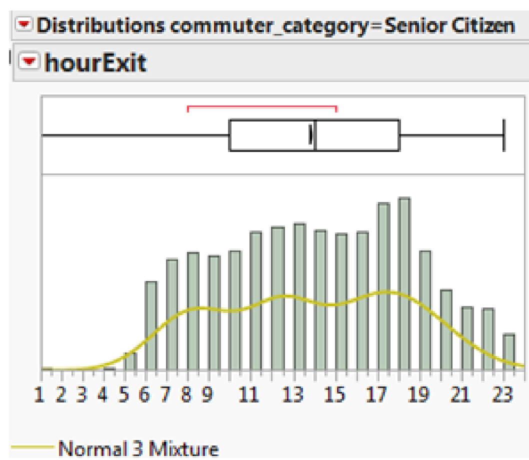




Students who take the train in the have a peak of 7 am and 1 pm. As November is the start of the school holidays, the two peaks of 7am suggest students who are taking their O Levels examinations that start at 8am. For the peak of 1pm, suggest students who are take the train for extracurricular activities in school.

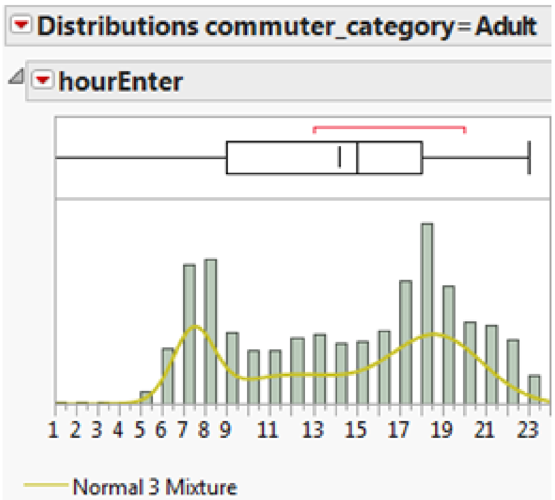
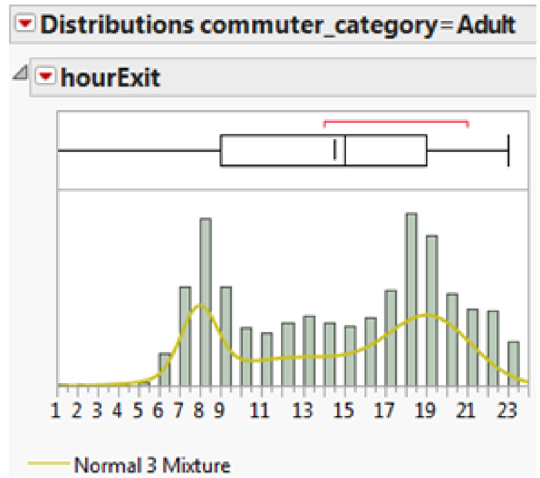
The exit patterns for Students suggest that most student reach school at 7am. This also suggests that the distance between home to school is short, as the duration travelled is within the 7^{th} hour of the day. The second peak, at 5pm, suggests that that it also the time students go home from school.





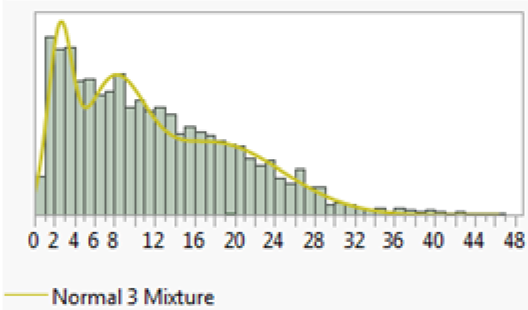
The Senior Citizen hour of entry suggest that most Senior Citizens take the train at 7am and 5pm. This suggests that they take the train to go to work or to babysit their grandchildren. A second peak at 5pm suggest that the Senior Citizens take the train at 5pm to go home to beat the rush hour crowd so as to get a seat on the train.

The exit pattern for senior citizen is similar to the entry pattern. However, it is observed that Senior Citizens that peak for boarding, 5pm, is not the peak for exit. The peak for exit is 6pm. This suggests that Senior Citizens spend a bit more time traveling back home. This suggests that senior citizens are more willing to travel further for work.

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>The adult hour entry pattern for train commuter has a peak of 8am and 6pm. This two peaks suggest the hour they board the train to go to work and go back home.</td>
<td>The pattern for adult exit is similar for entry with the exception for the exit peak of 8am. This suggests that commuters who board the train at 7am and 8am exit the train at 8am for work.</td>
</tr>
</table>

d. Commuter Patterns by Distance

distance_travelled



Quantiles

100.0%	maximum	46.6
99.5%		36.6
97.5%		28.8
90.0%		23.2
75.0%	quartile	17.1
50.0%	median	10
25.0%	quartile	4.9
10.0%		2.2
2.5%		1
0.5%		0.5
0.0%	minimum	0

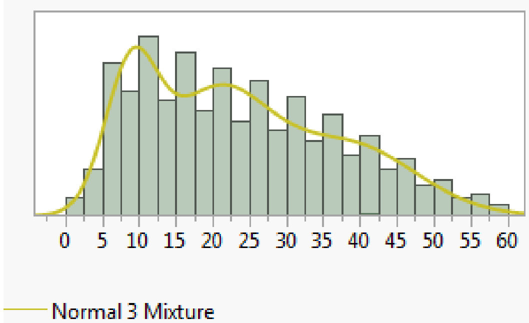
Summary Statistics

Mean	11.611661
Std Dev	8.0827817
Std Err Mean	0.0024343
Upper 95% Mean	11.616432
Lower 95% Mean	11.60689
N	11024582
Skewness	0.7458813

The commuter pattern by distance shows a right skewed distribution with a mean of 11.6KM, standard deviation of 8KM, with a median of 10KM and a maximum of 46.6KM. This suggests that most of the commuters use the MRT for short distance. This suggest that most of the commuters that MRT live near to their destination. As the distance increase, the distribution decreases. This suggests that those living further away from the destination MRT station prefer to take other means of transportation such as bus. Busses that travel long distances are called 'Cross country bus services' where they travel between towns. Such services include 960, 170, 190 and 67.

e. Commuters Patterns by Time

travel_time_min



Quantiles

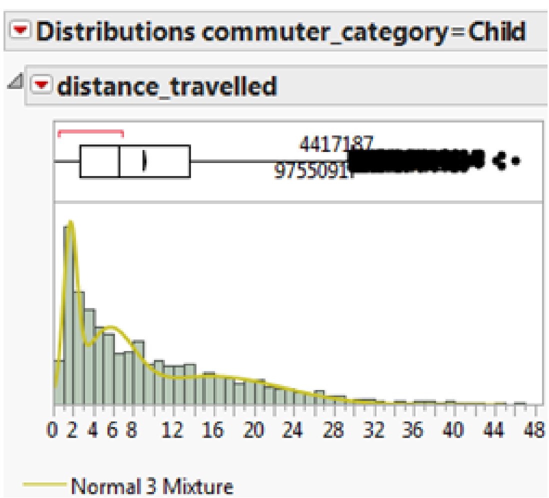
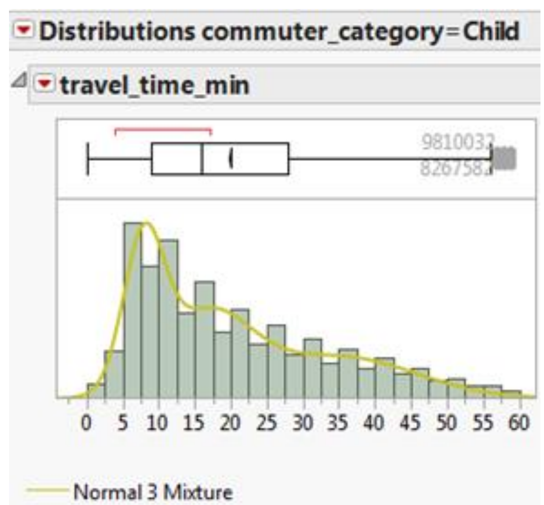
100.0%	maximum	59
99.5%		57
97.5%		52
90.0%		43
75.0%	quartile	33
50.0%	median	22
25.0%	quartile	12
10.0%		7
2.5%		4
0.5%		1
0.0%	minimum	0

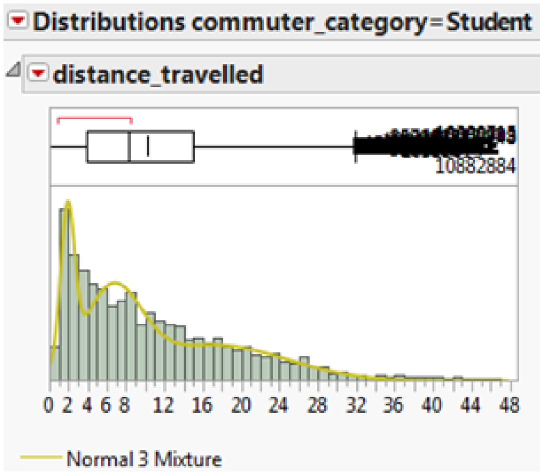
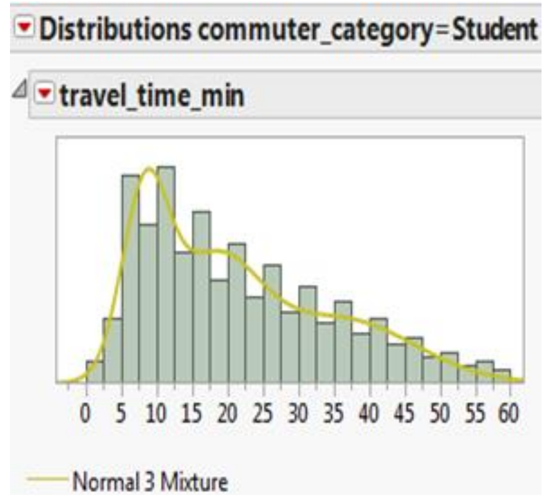
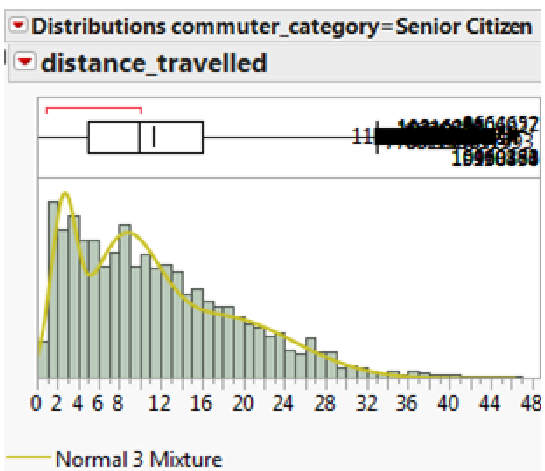
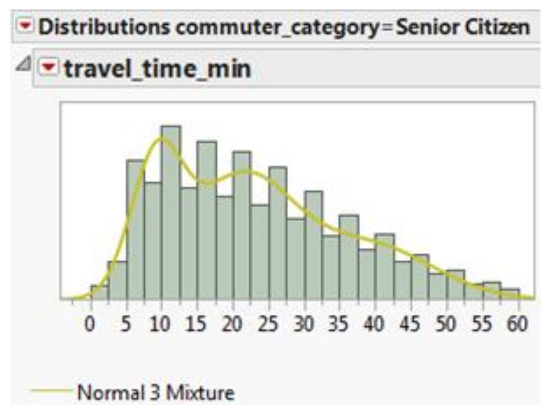
Summary Statistics

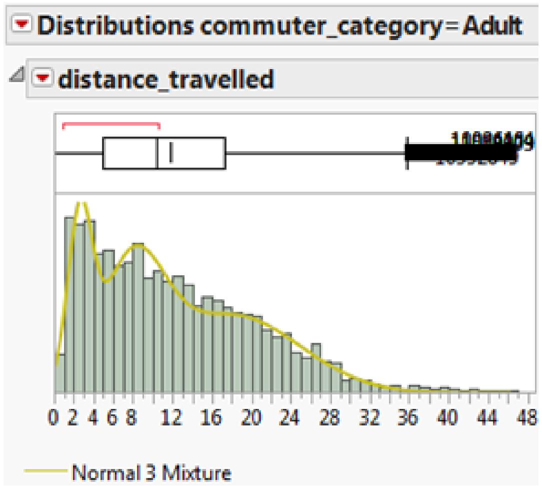
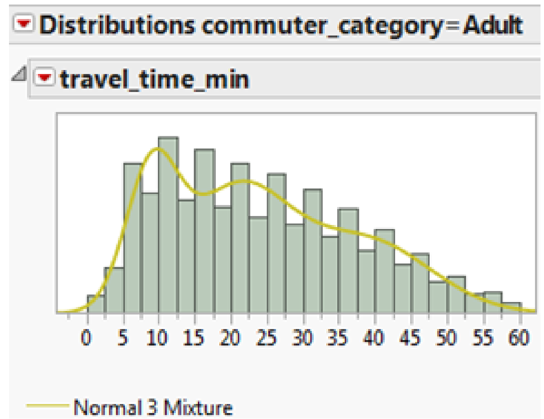
Mean	23.460301
Std Dev	13.427014
Std Err Mean	0.0040439
Upper 95% Mean	23.468227
Lower 95% Mean	23.452375
N	11024582
Skewness	0.4879247

The commuter travel time is also a right skewed with a mean of 23 minutes, standard deviation of 13 minutes, a median of 22 minutes and a maximum of 59 minutes. This suggests that commuters taking train spend an average of 23 minutes in the train, suggesting that they use the train to travel near distances, as suggested by the Commuter Pattern by distance.

f. Commuters Distance Patterns grouped by Commuter Type

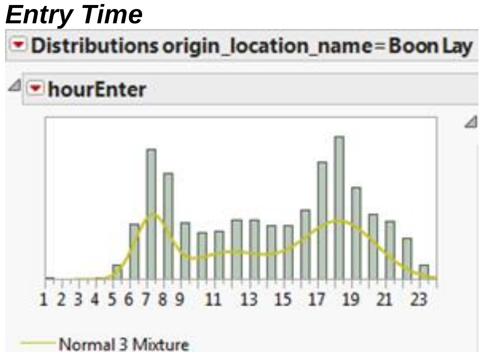
Distance Travelled Grouped By Commuter Type	***Travel Time Grouped By Commuter Type***
	
This distance travelled by Child suggest that most of the Child Travel to nearby stations. For example, Children board the train from Yew Tee MRT and disembark at Bukit Batok MRT station to visit the arcade or watch a movie at West Mall.	The time travelled by Child is right skewed. This suggests that most Child do not travel for long duration.

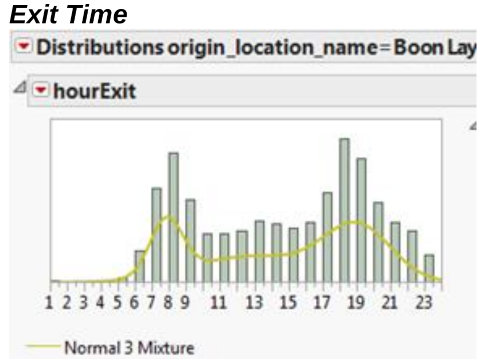
<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td colspan="2">The travel time and distance travelled for Student is similar with Child. This suggest that both groups have similar interest and reasons to travel. The travel pattern is similar as it is during the school holidays.</td>
</tr>
<tr>
<td colspan="2"></td>
</tr>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>The travel pattern for Senior Citizens is right skewed. This suggests that Senior Citizens do not travel far. It could be that senior citizens travel nearby to babysit their grandchildren while their children work.</td>
<td>The trend for the travel time by Senior Citizens indicates that they do not travel far. The peak of the travel time is 10 minutes. This matches the theory that the distance travel is short; therefore the travel time is shorter.</td>
</tr>
<tr>
<td colspan="2"></td>
</tr>
</table>

<table>
<tr>
<td>
</td>
<td>
</td>
</tr>
<tr>
<td>The travel distanced by Adult is right skewed. This suggests that most adults do not live far away from their work place or home.</td>
<td>The time travelled by adults is right skewed. This similarly confirms the theory that most adults do not live far away from workplace or home.</td>
</tr>
</table>

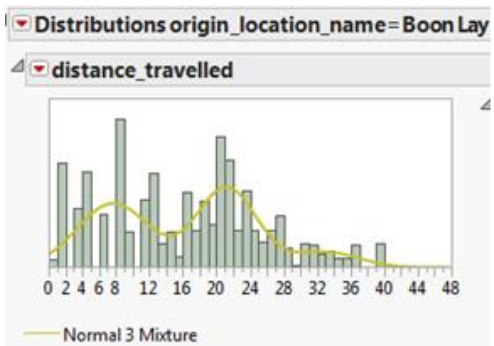
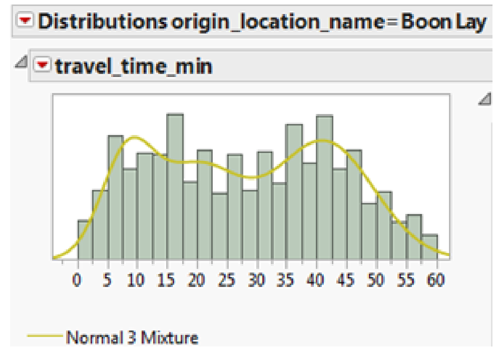
g. Commuters Travel Patterns grouped by Different zones

i. Boon Lay

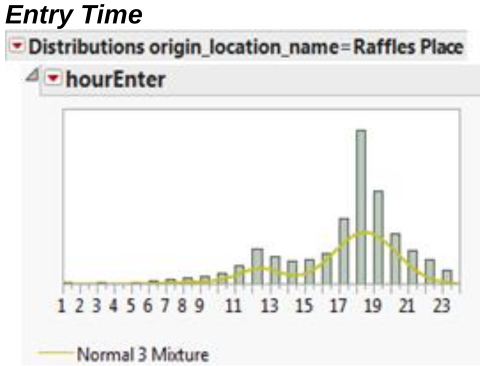
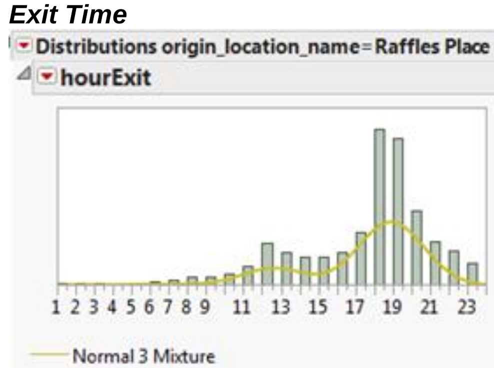
<table>
<tr>
<td>Entry Time


There are two peaks, one at 7AM and one at 6PM. At 6PM this suggests that heart landers are traveling to work. For the 6PM peak, this suggests that people who work around Boon Lay are going home from work and are mostly factory workers who are Malaysians. This is most likely the case as Boon Lay is close to factories and industrial estates.</td>
<td>Exit Time


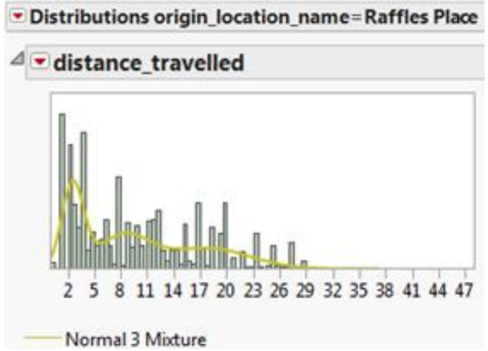
There are two peaks, one at 8AM one at 6PM. This suggests that commuters who board the train from Jurong East spend an hour traveling to work. For the second peak, this indicates that the factory workers live near Boon Lay MRT station. This is most likely he case as most factory workers are Malaysians and it is more economical to live near their workplace.</td>
</tr>
<tr>
<td>Distance Travelled – Origin station from Boon Lay</td>
<td>Duration Travelled – Origin station from Boon Lay</td>
</tr>
</table>

	
There are two peaks, one at 8KM one at 20KM. This could mean two possible scenarios: the 8KM peak could indicate factory workers who travel to Jurong East MRT station to take the bus back to Johor Baharu (SBS Service 160). For the 20KM peak, this could suggest that the Malaysian Factory workers travel to Kranji MRT station to take the bus back to Johor Baharu (SBS Service 170).	There are two peaks, 15 minutes and 40 minutes. The travel duration is also indicative of the previous hypothesis that the factory workers are heading to either Kranji or Jurong East MRT Station.

ii. Raffles Place

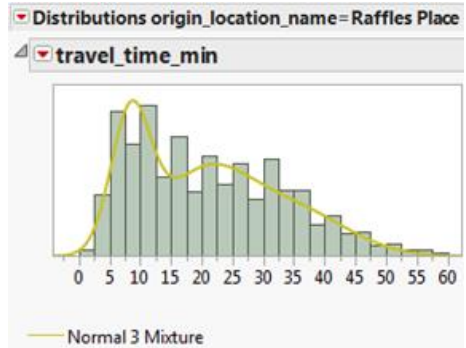
Entry Time 	***Exit Time*** 
The distribution for commuters entering Raffles Place MRT follows a left skewed distribution. This suggests office workers are making their way home.	Similar to the entrance pattern, the exit pattern for Raffles Place follows a left skewed distribution. This could suggest that commuters who work around Raffles Place take between an hour or two to travel away from Raffles Place. This suggests that commuters from Raffles Place are heading back home to the heartlands after work. This could also suggest that more people are ending work towards the later quarter of 6PM. For example, a commuter who boards the train at 6.55PM might reach home at 7.04PM – which might explain the reason the 7PM peak.

Distance Travelled – Origin Station from Raffles Place



The distribution is right skewed. This suggests that commuters who travel from Raffles place live near Raffles Place. However, there is also a second peak at 8KM and 20KM. This two indicate that there are people who live further from Raffles Place, such as Ang Mo Kio and Queensway.

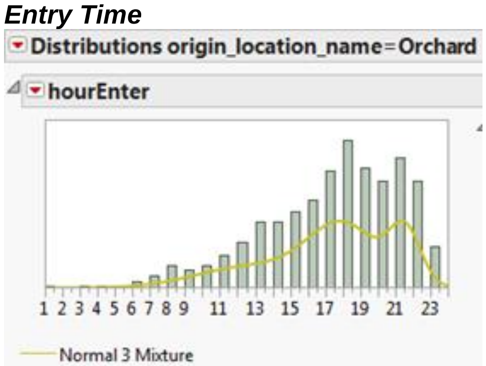
Duration Travelled – Origin Station from Raffles Place



Confirming the hypothesis from the distance travelled, most of those who start their journey from Raffles Place live nearby. This is evident as most of those who start their journey from Raffles place take about 10-15 minutes to their destination.

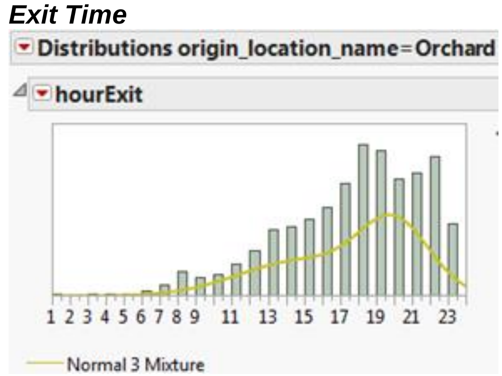
iii. Orchard

Entry Time

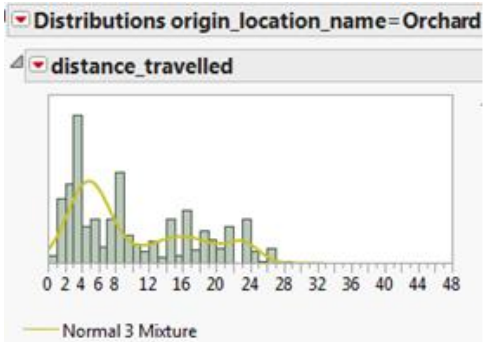
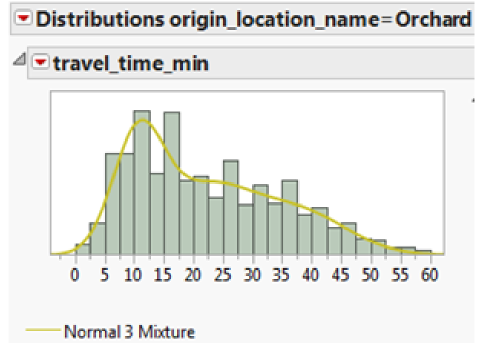


The graph is left skewed with a peak at 6PM. This is similar to the Raffles Place graph that has a peak at 6PM. This suggest that commuters who work around Orchard MRT station heading back home. A second peak is also seen at 9PM. This suggests that retail staffs working around Orchard MRT are heading back home. This also suggests that there is more commuter traffic towards the later part of the day.

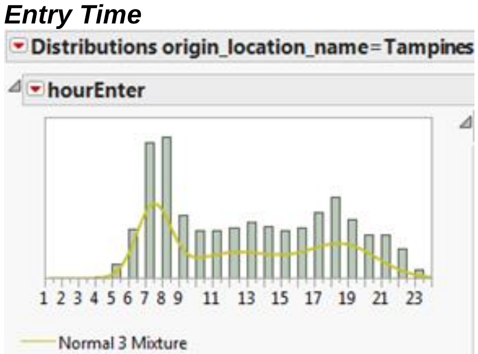
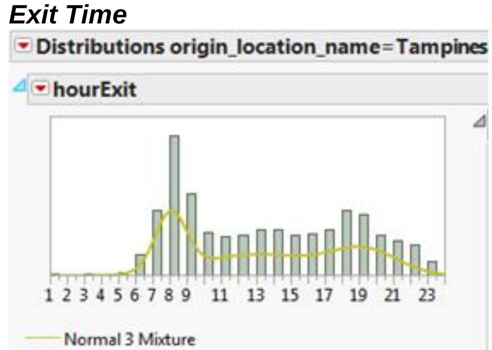
Exit Time

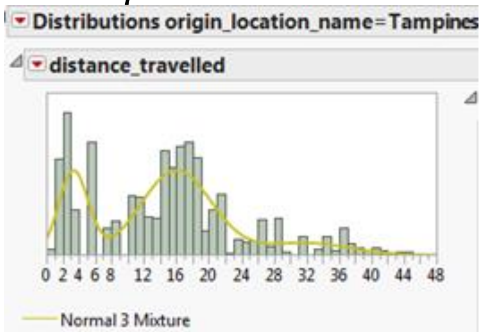
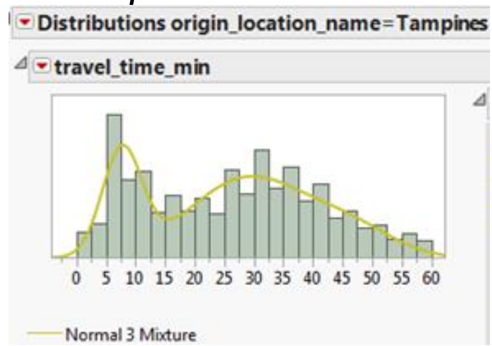


The graph is left skewed. This suggests that there are commuters who travel away from Orchard take about an hour to go back home. The patterns are similar to the entrance, with two peaks at 6PM and 10PM. This suggests workers reaching home at 6PM, and retail staff reaching home at 10PM. This also suggests that retail workers working in Orchard stay far away from Orchard. This could also be interpreted that most of the retail staff working in Orchard leave after the second quarter of 9PM, thus reaching home at 10PM.

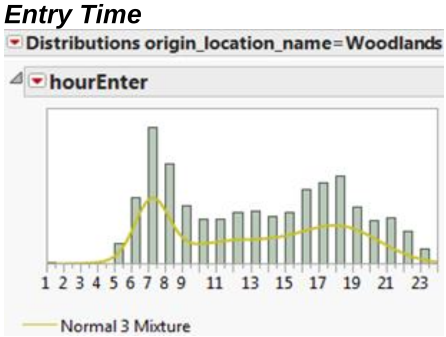
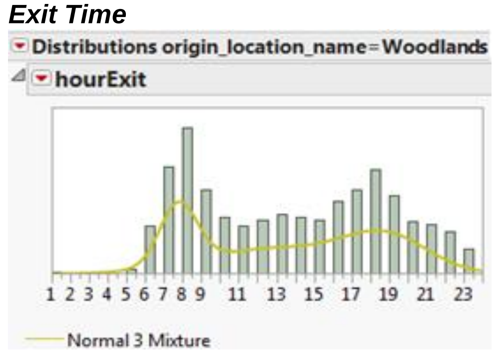
Distance Travelled – Origin station from Orchard	*Duration Travelled – Origin station from Orchard*
 This graph suggest that the distance travelled away from Orchard is about 3 KM.	 There are two peaks for this, one at 10 minutes, another one at 15 minutes. This suggests that most of the commuters take about 15 minutes to their destination from Orchard.

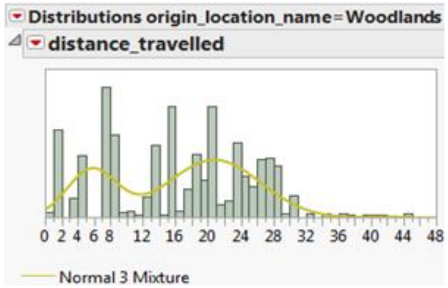
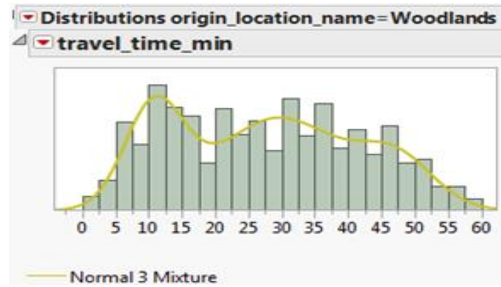
iv. Tampines

Entry Time	*Exit Time*
 There are two main peaks for Tampines, one at 7AM one at 8AM. This suggests that heart landers are heading to work. Another peak is also seen at 6PM. This suggests that workers who work around Tampines MRT station are heading back home.	 There is one main peak for the hour exit, 8AM. This suggests that commuters who travel from Tampines could be travelling to work. As there are two peaks for hour enter, but one peak for exit time, this suggest that commuters from Tampines spend Boon.

Distance Travelled – Origin station from Tampines	*Duration Travelled – Origin station from Tampines*
 There are two peaks for distance travelled from Tampines. One is at 2KM, the other around 12-19KM. This matches the hypothesis based on the exit timing where there are two types of commuters in Tampines, those who work far away from Tampines (and spend more than a hour travelling) and those who work around Tampines.	 There are two peaks, one at 5 minute one at 30 minutes. This confirms the theory in the distance travelled that there are two types of commuters in that start their journey in Tampines, long distance commuters and those who travel short distances.

v. Woodlands

Entry Time	*Exit Time*
 There are two peaks for the hour of entrance for the Woodlands station, one at 7am the other at 6pm. This suggests that most commuters from Woodlands are heart landers who are taking the train to work. The second peak at 7pm suggests that the workers from the industrial estates in Woodlands are heading home.	 There are two peaks, similar to the entrance patterns, however the morning peak is at 8am. This suggests that commuters from Woodlands take about an hour to travel to work.
Distance Travel – Origin Station from Woodlands	***Duration Travelled – Origin Station from Woodlands***

 The distance travelled from Woodlands is right skewed. This suggests that commuters in Woodlands work in nearby places such as Yio Chu Kang industrial estate. As there are other peaks, this could suggest that there are commuters working further such as Jurong or Orchard.	 The travelling time for users originating from Woodlands is between 10 to 35 minutes. This suggests that commuters from Woodlands do not work near Woodlands. This suggests that those who work in Woodlands take the bus to work. This is further supported that Woodlands is a Regional Bus Interchange. Therefore the train users are people who work further such as Orchard or Raffles Place.

This concludes the EDA. We will now proceed to analyse the different time series data mining techniques before picking on one before running data mining on the dataset.

13. References