Question 1 (a)

The mission of Pittsburgh Bike Share is to provide easy-to-use, affordable bike share transportation accessible to all. This research was carried out to understand rider patterns and also finding ways to bring ridership figures up and help the organization to reach their goals. The aim is to serve additional Pittsburgh communities and to expand the established bike-sharing network in order to make Healthy Ride convenient for everyone.

The business problem faced by Pittsburgh Bike Share is keeping the ridership rate up. There was a huge drop in usage by subscriber in 2020 as compared to 2019. They face a challenging issue to ensure that every station would have enough bike and rack for user to use and to dock the bike when bike was returned back to the stations. Pittsburgh Bike have difficulty to ensure users to return the bike to the nearest stations.

The organization aims to increase station density, increase ridership especially increase the use for registered users or the subscribers and also system expansion. The organization would like to consistently outreach and educate riders on safety measures.

These are important to support the initiatives of the project which is to reduce congestion and improve air quality via the reduction of car emission. In addition, to improve public health and fitness and reducing the overall street maintenance cost.

Question 1 (b)

Association Rule Mining would be able to assist in identify the relationship between each variable using the data collected. The objective is to determine the usage patterns of the users of bike sharing and how this information can be used to assist Healthy Ride in making decision especially on the marketing promotions to the correct target audience. This method allows rules with multiple consequents (for CARMA) depending on the needs.

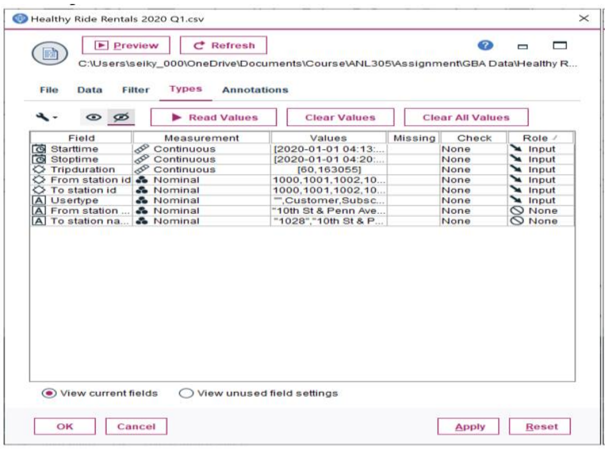
However, Clustering method was used as the Healthy Ride would need to look at the users' usage pattern and what would the important variables and to group it into different clusters. This method would be useful for the Organization to look into different marketing promotion to target different types of users. For this analysis, variables such as the stations name where the rides begin and the rides end, the user types, trip duration and so on were used to group into different clusters.

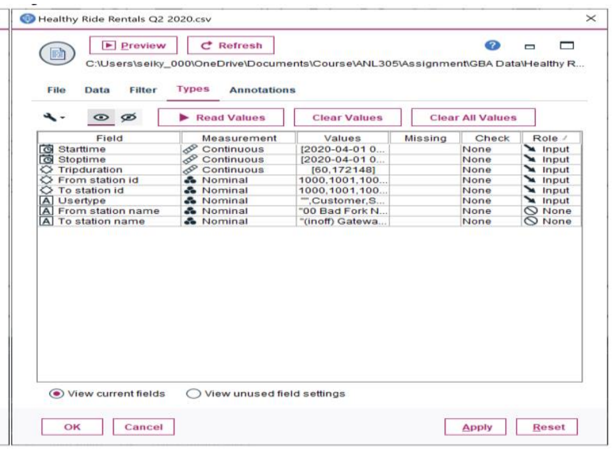
The results generated from Association Rule Mining and or Clustering will be able to help the organization to make key decisions such as how to increase the ridership, when to arrange marketing promotions, what are the target audience they would like to attract and also whether the current rack at the stations are enough to dock all the bikes after the trip ended.

Question 1 (c)

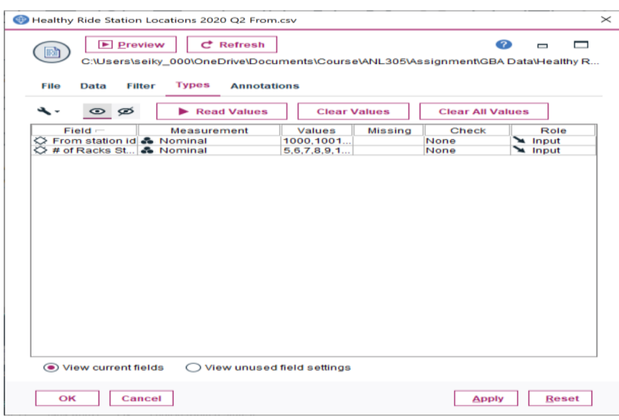
i) In order to study the business problem, we obtained the following 2020 data from the website (https://healthyridepgh.com/data/)

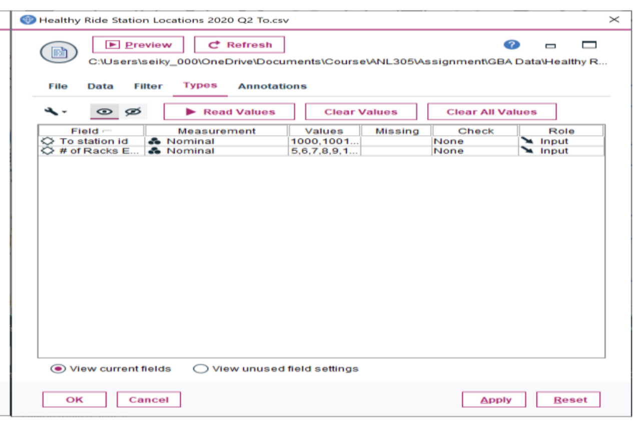
* Healthy Ride Rentals 2020 Q1.csv & Healthy Ride Rentals Q2 2020.csv



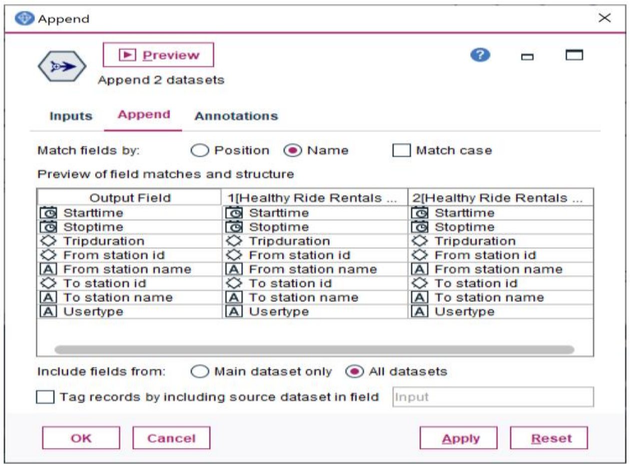


* Healthy Ride Station Locations 2020 Q2.csv – Change Field Header by Excel from "Station ID" to "From Station ID" / "To Station ID" & From "# of Racks" to "# of Racks Start" / "# of Racks End"

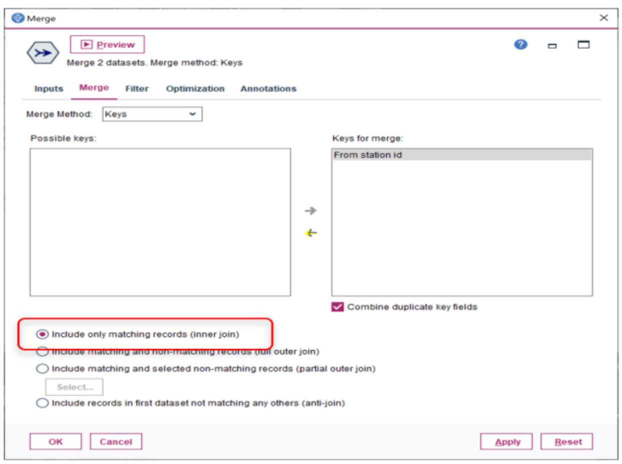


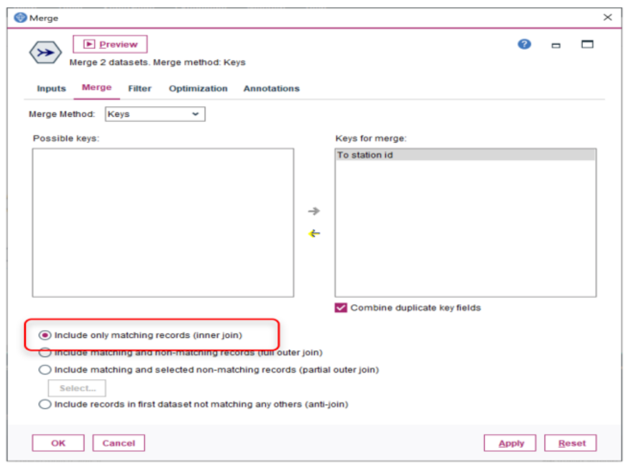


ii) Before we can start the modelling, we selected the potential fields and combine the datasets into one. To do that, we will first use "Append" node to combine 2020 Q1 and Q2 rental records.

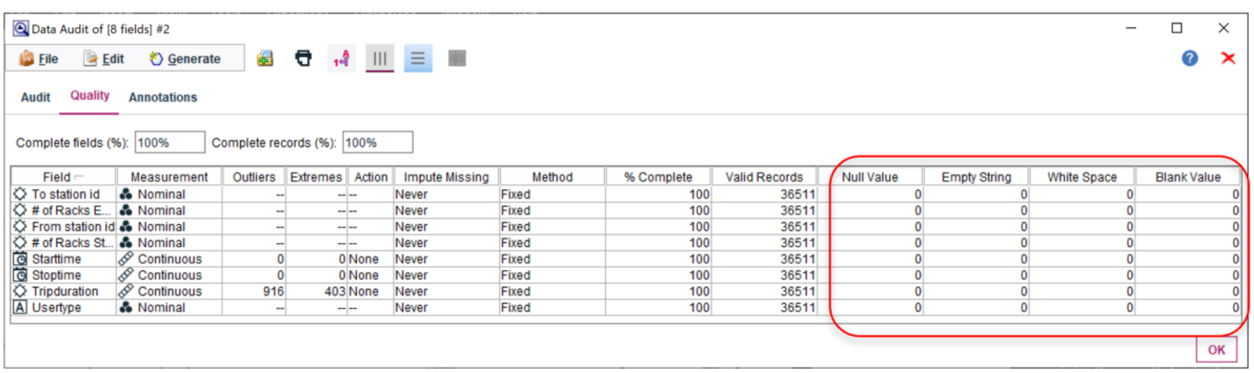


In additional to the rental records, we would also like to include "No of Racks" into each "From Station ID" & "To Station ID" transactions. As the Healthy Ride Station Locations 2020 data does not have same Field for Merging to take place, we use Excel to change the field header from "Station #" to "From Station ID" or "To Station ID". Once the changes made in excel, we will be able to use "Merge" node to combine data into the datasets. Based on the Merging condition below, Null Value are removed from the datasets.





iii) After all the required fields were merged, we performed "Data Audit" Node to confirm that the data has been cleaned.



Field	Measurement	Outliers	Extremes	Action	Impute Missing	Method	% Complete	Valid Records	Null Value	Empty String	White Space	Blank Value
To station id	Nominal	--	--	--	Never	Fixed	100	36511	0	0	0	0
# of Racks E...	Nominal	--	--	--	Never	Fixed	100	36511	0	0	0	0
From station id	Nominal	--	--	--	Never	Fixed	100	36511	0	0	0	0
# of Racks St...	Nominal	--	--	--	Never	Fixed	100	36511	0	0	0	0
Starttime	Continuous	0	0	None	Never	Fixed	100	36511	0	0	0	0
Stoptime	Continuous	0	0	None	Never	Fixed	100	36511	0	0	0	0
Tripduration	Continuous	916	403	None	Never	Fixed	100	36511	0	0	0	0
Usertype	Nominal	--	--	--	Never	Fixed	100	36511	0	0	0	0

iv) In order to study the usage patterns in more details, we further breakdown the "Starttime" and "Stoptime" Timestamp into the following with the Formula:-

* Year : datetime_year(Starttime), datetime_year(Starttime)
* Month : datetime_month(Starttime), datetime_month(Stoptime)
* Date : datetime_day(Starttime), datetime_day(Stoptime)
* Day : datetime_weekday(Starttime), datetime_weekday(Stoptime)
* Hours : datetime_hour(Starttime), datetime_hour(Stoptime)

v) After Data Preparation had completed, we selected the following datasets for analysis.

* Excluded Field For Analysis:
  => Healthy Ride Rentals 2020 Q1 & Q2 Dataset :
  "Starttime", "Stoptime", "Bikeid", "From Station Name", "To Station Name"
  => Healthy Ride Station Locations 2020 Q1 & Q2 Dataset :
  "Station Name", "Latitude", "Longitude"

* Details of Datasets:

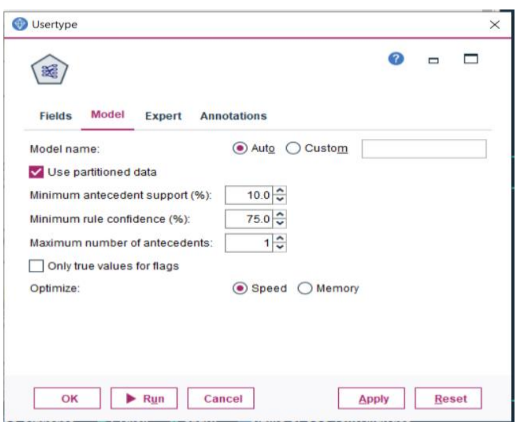
Field	**Type**	**Description**
From Station ID	Nominal	ID of each Station name at the start of rental
# of Racks Start	Nominal	No of Bicycle racks at the start of rental.
To Station ID	Nominal	ID of each Station name at the end of rental
# of Racks End	Nominal	No of Bicycle Racks at the end of rental
Trip Duration	Continuous	Usage timing of the customer
Usertype	Flag	Membership status during rental. (Customer/Subscriber)
Start Year	Nominal	Year when the rental started. (e.g. 2020, 2019, etc)
Start Month	Nominal	Month when the rental started. (1 to 12 represent January to December)
Start Date	Nominal	Date when the rental started. (1^{st} to 31^{st} of the month)
Start Day	Nominal	Day when the rental started. (1 to 7 represent Monday to Sunday)
Start Hours	Nominal	Hours when the rental started. (0 to 23 represent 12AM to 11PM)
Stop Year	Nominal	Year when the rental ended. (e.g. 2020, 2019, etc)
Stop Month	Nominal	Month when the rental ended. (1 to 12 represent January to December)
Stop Date	Nominal	Date when the rental ended. (1^{st} to 31^{st} of the month)
Stop Day	Nominal	Day when the rental ended. (1 to 7 represent Monday to Sunday)
Stop Hours	Nominal	Hours when the rental ended. (0 to 23 represent 12AM to 11PM)

Question 1 (d)

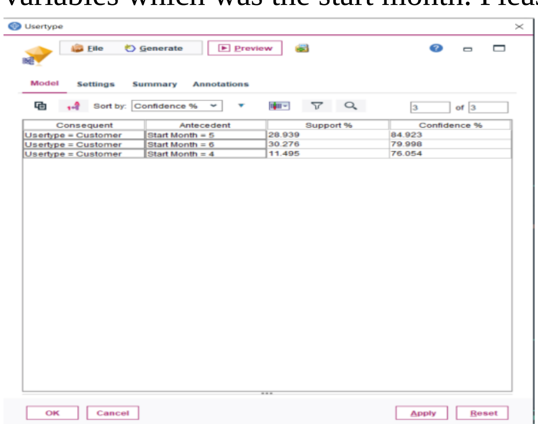
Data mining methods that were used in analysing the data are the association rule mining model (Apriori) and clustering solutions such as (K-means, two-steps and Self-organising Map (SOM)). Mainly due to the advantages of each method for example in Association rule mining, the target would be finding the relationship between the variables and predict what the confidence level users, the customer or subscribers. This could potentially help the business to focus and target customers to become subscribers. On the other hand, clustering solutions such as (K-means, two-steps and SOM), the aim was to look at the clustering records to either support the association rule mining findings or gather further insight and identify other actionable information.

The sections below illustrate the design decision with parameter settings for ARM:
For Association Rule Mining (ARM), "apriori node" was used and the aim was to predict and find if there were any relationship between the variables and the user type which is the customer or subscriber.

In the type nodes setting, the Usertype was set as the Target and the measurement was changed to "Flag". The rules in "Apriori" nodes were configured and the confidence level was set to 75%. A higher confidence level threshold was set to ensure the accuracy of the prediction and also to target the correct audience.



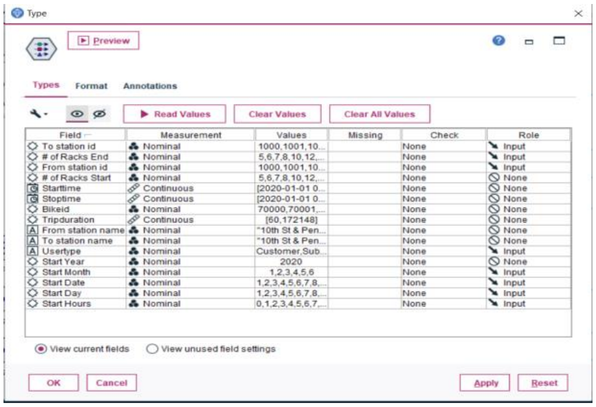
The result from running Apriori nodes showed only one type of relationship between the target and the variables which was the start month. Please refer below screenshot on the results from "Apriori".



From the results above, we can conclude that the usage of bikeshare was high in the month April, May and June 2020 were contributed by customers. Hence, Healthy Ride could potentially market their products to these group of customers to be subscribers in the coming years as these group of users would be using bikeshare in the same period next year.

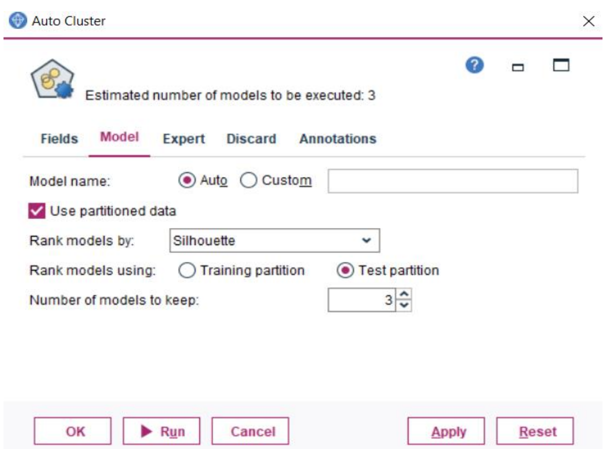
The section below illustrates the parameter settings for Clustering.

This analysis focuses on finding out the possible interesting variables that can be grouped together into a cluster. Type node was used to set the variables.

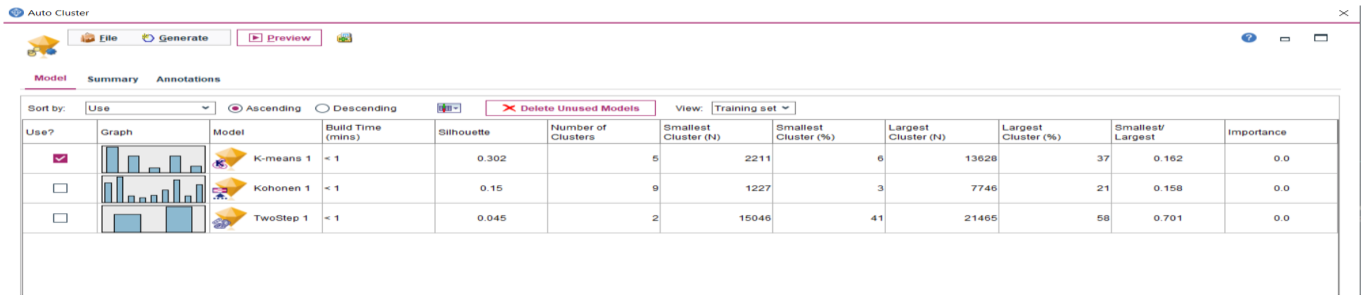


Field	Measurement	Values	Missing	Check	Role
To station id	Nominal	1000,1001,10...		None	Input
# of Racks End	Nominal	5,6,7,8,10,12,...		None	Input
From station id	Nominal	1000,1001,10...		None	Input
# of Racks Start	Nominal	5,6,7,8,10,12,...		None	None
Starttime	Continuous	[2020-01-01 0...		None	None
Stoptime	Continuous	[2020-01-01 0...		None	None
Bikeid	Nominal	70000,70001,...		None	None
Tripduration	Continuous	[60,172148]		None	None
From station name	Nominal	"10th St & Pen...		None	None
To station name	Nominal	"10th St & Pen...		None	None
Usertype	Nominal	Customer,Sub...		None	Input
Start Year	Nominal	2020		None	None
Start Month	Nominal	1,2,3,4,5,6		None	Input
Start Date	Nominal	1,2,3,4,5,6,7,8,...		None	Input
Start Day	Nominal	1,2,3,4,5,6,7,8,...		None	Input
Start Hours	Nominal	0,1,2,3,4,5,6,7,...		None	Input

Auto clustering nodes were used to look at different clustering methods according to the rank of silhouette coefficient as shown in the screenshot below.



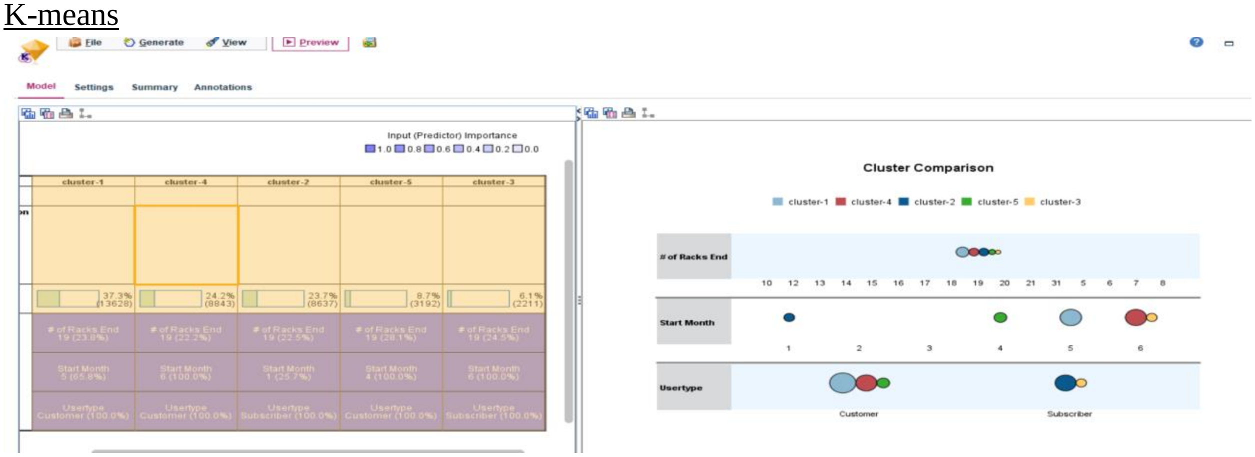
The results below were used to support and gain an overview of what could be the possible clustering method to use as well as comparison between each model.



Use?	Graph	Model	Build Time (mins)	Silhouette	Number of Clusters	Smallest Cluster (N)	Smallest Cluster (%)	Largest Cluster (N)	Largest Cluster (%)	Smallest/ Largest	Importance
☑		K-means 1	< 1	0.302	5	2211	6	13628	37	0.162	0.0
☐		Kohonen 1	< 1	0.15	9	1227	3	7746	21	0.158	0.0
☐		TwoStep 1	< 1	0.045	2	15046	41	21465	58	0.701	0.0

From the results, K-means would seem to be a better clustering model as confirmed by silhouette coefficient. Further analysis and comparison to gain insights of the data of each model suggested by the Auto Cluster node.

<u>K-means</u>



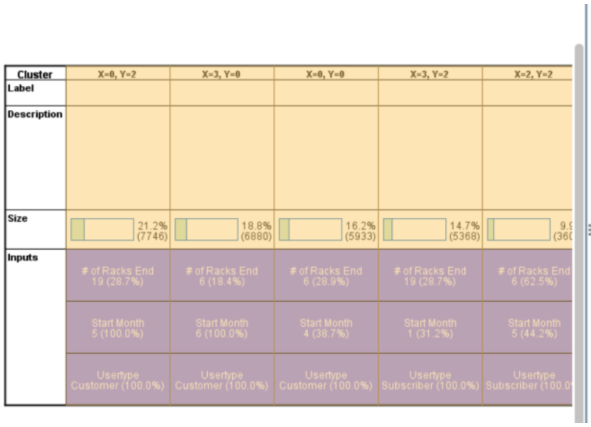
From the Auto Cluster node, it was suggested that K=5 should be used and the results above showed that usage of bike were high in the month of April, May and June. From the results we could see that the user type was high for Customer category. This result helps provide insights that Healthy Ride should look into promoting customers to join as a subscriber in the month of April to June. In additions, it is also important to analyze the end stations available racks especially at popular stations. The clustering results showed that number of racks at each ending stations would be an important variable to look at. On average, the total racks available were between 18 to 20. Hence, if Healthy Ride intend to increase people to sign up or increase ridership, which resulted in increased of bikes this would be an avenue to take into consideration to have more racks for docking.

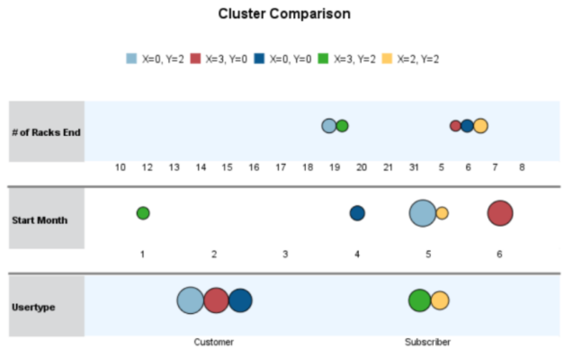
The other two clustering methods as suggested by the auto-clustering nodes, SOM has similar insights as the K-means while TwoSteps has only two clusters to compare. TwoSteps does not seem to be providing useful insights as compared to the other clustering methods.

The screenshots below show the results and the comparison.

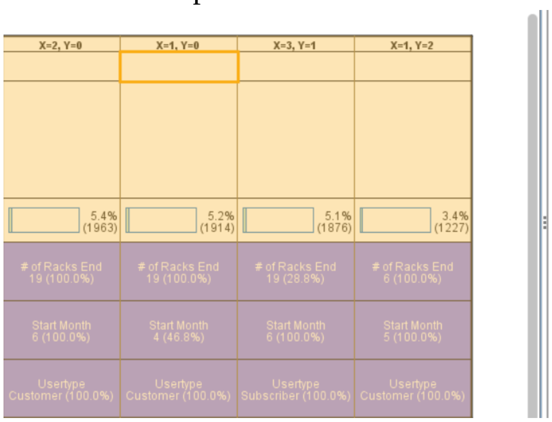
<u>SOM</u>

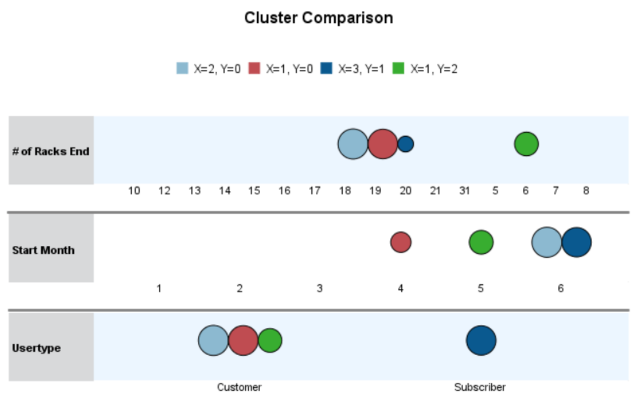
The first 5 cluster comparison



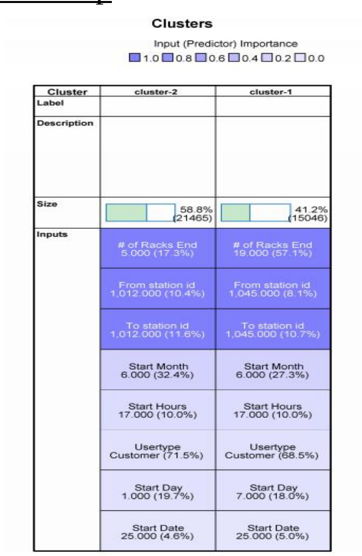


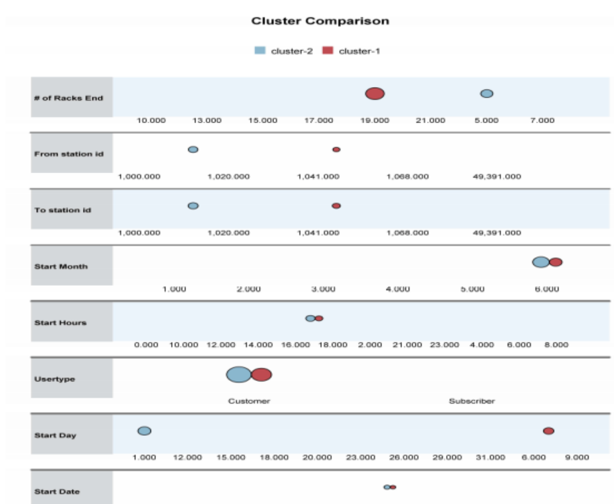
The last 4 comparison





<u>TwoStep</u>



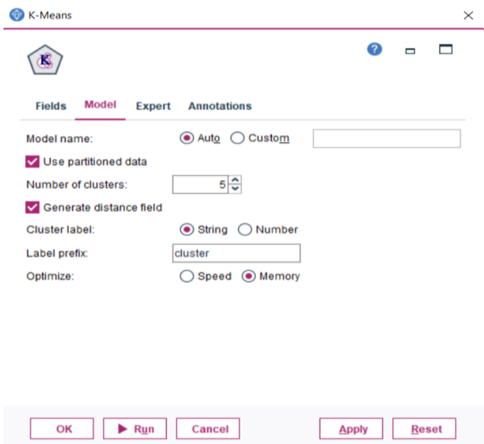


Additionally, a deep dive of analysis was performed on each method, refer to the annex A for more details of the results.

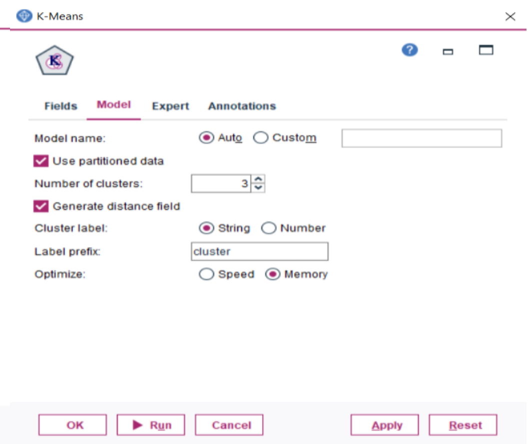
As the auto-cluster suggested, K-Means would be a better solution, hence the section below would further discuss and compare K-means clustering.

Firstly, comparisons were made between K=3 and the auto-cluster suggestion of K=5 as summarized below:

K-Means, Cluster (3)

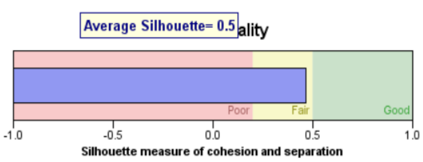


K-Means, Cluster (5)



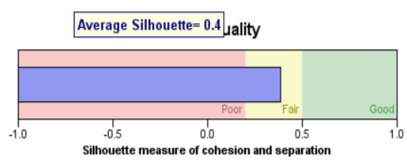
Model Summary

Algorithm	K-Means
Inputs	4
Clusters	3



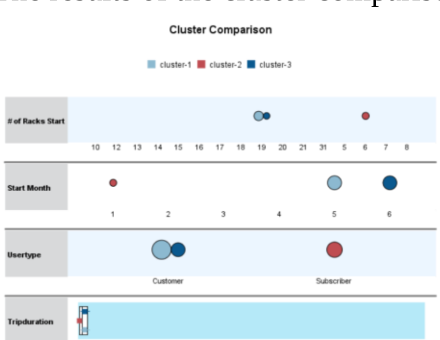
Model Summary

Algorithm	K-Means
Inputs	4
Clusters	5



The Average Silhouette coefficient suggested K=3 would be a better modeling solution to choose from. However, comparison between each cluster provided the same results where the findings suggested that the high usage was between May and June period and the high usage was contributed by customer category.

The results of the cluster comparison for K-means (number of cluster 3) is shown below:

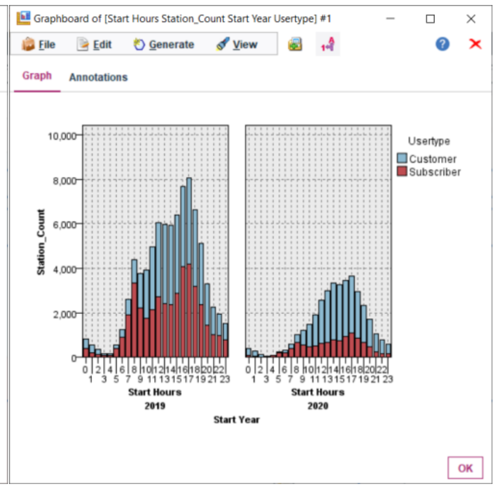


In conclusion, the ARM method would help the business decision to predict and target the confidence level of people using bike share would be customer category in the month of May.

On the other hand, Clustering solution would help to identify what is the cluster of people using bike share where it is shown that during the April to June period more customers are using the bike share.

Question 1 (e)

ARM has confirmed that the customer would uses the bike sharing more frequently in the month of May to Jun. The same usage pattern can be confirmed while comparing 2019 and 2020 data. 2019 usage by Customers was extended till Oct 2019. In addition, there was a huge drop of usage from November 2019 to April 2020 for both subscriber and customer category. From the graph, we noticed that subscribers could be working adults who have a fix timing to report to work and the usage were at the highest at 8am and 5pm. For customer, the high usage would be at 2pm and 5pm.



The recent drops of the subscribers as shown in the graph above as compared to 2019, the company could look into different marketing campaign in the month of May and Jun to attract user to convert from customer category to subscriber category. Heathy Ride could introduce different tiers package which targeted at customers' needs and to encourage them to convert from customer to subscriber category. Different tiers package that Healthy Ride could proposed are as follows:

1) Working adults: usage from Monday to Friday and usage hour would be from 8am to 6pm.
2) Students: usage from Monday to Friday and usage hour would be from 5am to 2pm.
3) Weekenders: usage for Saturday and Sunday and usage hour would be from 12pm to 7pm.
4) Senior citizens: half pricing for senior citizens with no limitation of usage hour and usage days.
5) To introduce daily charge fee. This would be applicable to those users who wish to rent for whole day.

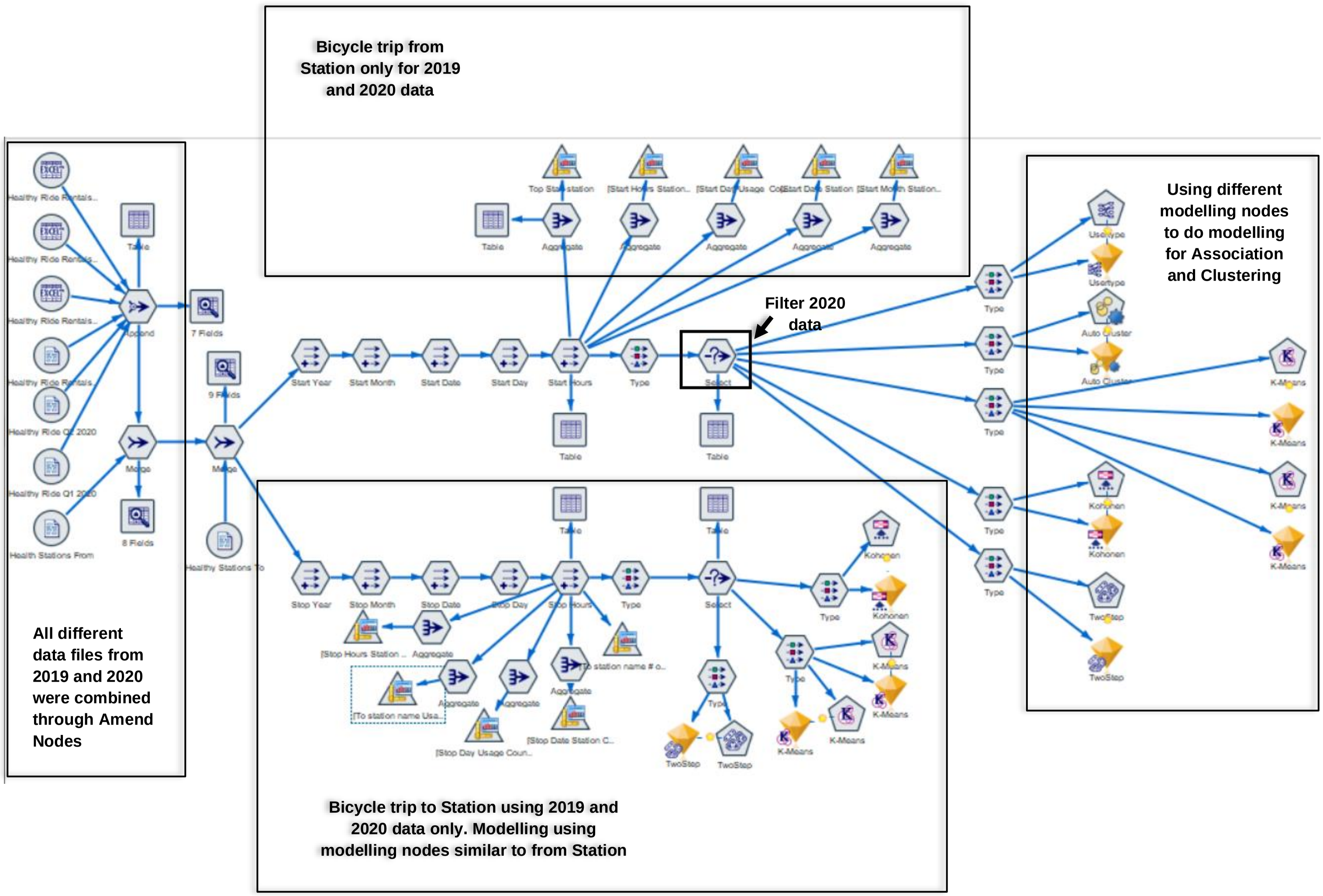
Once these packages finalized and taken up by users, it is important that Healthy Ride would continue to gather the data of the usage. After 6 months of deployment of these packages, these data should be analyzed again to see if there was any movement in usage pattern in terms of user type, usage hours, usage month and also usage days. If there was a slight increase in the ridership, Healthy Ride should continue with the proposed packages and data gathering should continue and further analysis using the same method should be done every 6 months.

As mentioned in 1(a), Healthy Ride has an issue to manage the bicycle that was returned elsewhere and not at the stations. This can be proved from the data that were collected where the return station ID was the bike ID. There could be two probability where users which mainly contributed by customer category did not return the bicycle back to the station. First probability could be there were not enough rack at the popular stations. Hence, the bicycles were not dock at the rack which the bicycle could recognized when it was returned. Second probability would be that the ignorance of the user itself which refused to return to the bicycle stations. In order to solve this business problem, Healthy Ride could implement the following:

1) Have more racks at the popular stations since the Clustering analysis showed that Rack would be an important variable for users.
2) To implement discount fee for every successful return of the bike to the stations for customer category.
3) For unsuccessful return of the bike to the stations, the fee charges should continue until the bike was return with a maximum range up to the daily charge fee.
4) For subscriber category, Healthy ride could implement point usage for every successful return, each subscriber would entitle 10 points which can be accumulated and exchange for discount voucher which can be used to offset their monthly subscription fee.

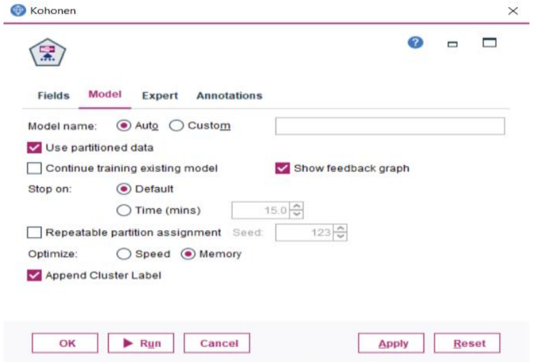
The above should be deployed at popular stations first. Upon successful of each deployment, collection of data should continue to ensure that this deployment should encourage users to dock the rented bike back to any Healthy Ride stations. Healthy ride should perform another round of analysis to look into changes in the users' behaviors. If the data collected were cleaned, then more accurate analysis can be performed.

Question 1 (f)

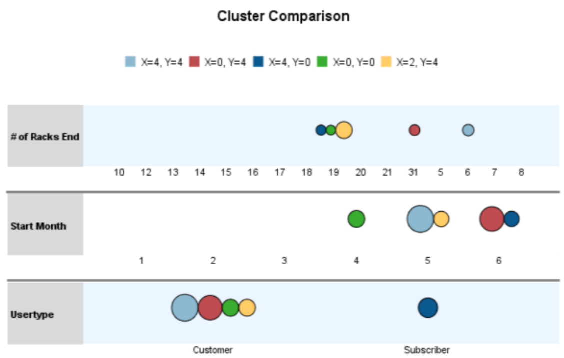

Bicycle trip from
Station only for 2019
and 2020 data
Using different
modelling nodes
to do modelling
for Association
and Clustering
Filter 2020
data
All different
data files from
2019 and 2020
were combined
through Amend
Nodes
Bicycle trip to Station using 2019 and
2020 data only. Modelling using
modelling nodes similar to from Station

Annex A

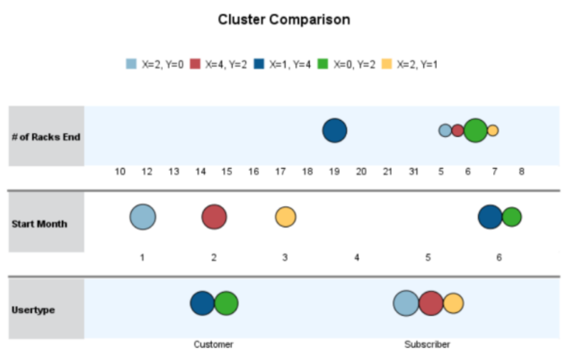
SOM - deep dive



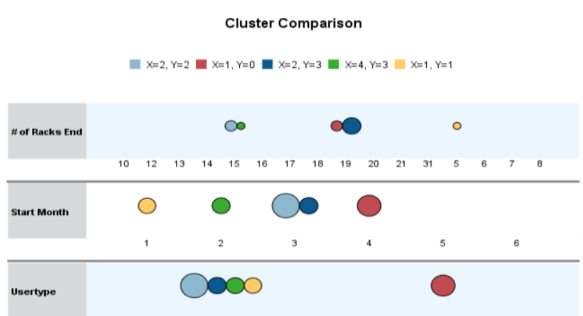
Cluster	X=4, Y=4	X=0, Y=4	X=4, Y=0	X=0, Y=0	X=2, Y=4
Label					
Description					
Size	18.5% (6750)	15.4% (5616)	9.8% (3562)	7.3% (2658)	7.2% (2627)
Inputs	# of Racks End 6 (18.2%)	# of Racks End 31 (19.2%)	# of Racks End 19 (26.6%)	# of Racks End 19 (33.7%)	# of Racks End 19 (100.0%)
	Start Month 5 (100.0%)	Start Month 6 (100.0%)	Start Month 6 (62.1%)	Start Month 4 (100.0%)	Start Month 5 (84.6%)
	Usertype Customer (100.0%)	Usertype Customer (100.0%)	Usertype Subscriber (100.0%)	Usertype Customer (100.0%)	Usertype Customer (100.0%)

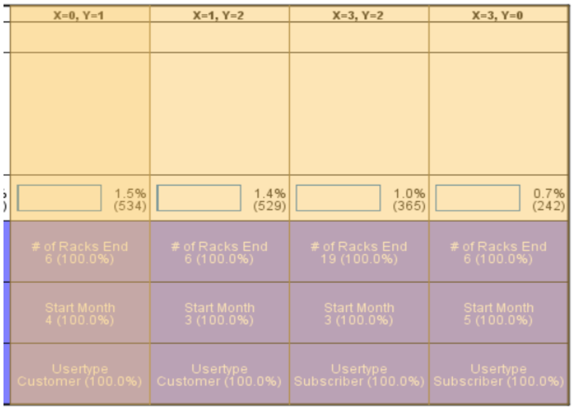


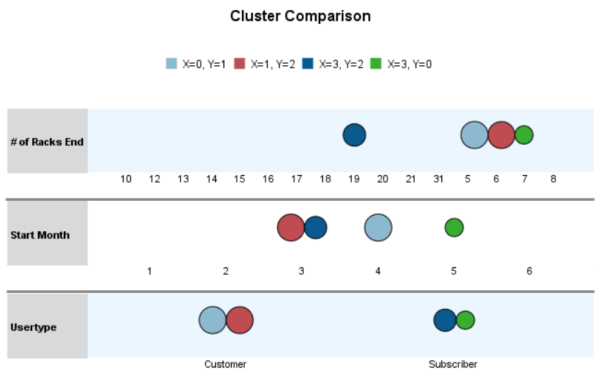
X=2, Y=0	X=4, Y=2	X=1, Y=4	X=0, Y=2	X=2, Y=1
6.1% (2222)	5.6% (2057)	5.4% (1963)	5.3% (1917)	3.8% (1395)
# of Racks End 6 (24.7%)	# of Racks End 6 (24.4%)	# of Racks End 19 (100.0%)	# of Racks End 6 (100.0%)	# of Racks End 6 (29.6%)
Start Month 1 (100.0%)	Start Month 2 (100.0%)	Start Month 6 (100.0%)	Start Month 6 (65.9%)	Start Month 3 (100.0%)
Usertype Subscriber (100.0%)	Usertype Subscriber (100.0%)	Usertype Customer (100.0%)	Usertype Customer (100.0%)	Usertype Subscriber (100.0%)



X=2, Y=2	X=1, Y=0	X=2, Y=3	X=4, Y=3	X=1, Y=1
3.7% (1338)	2.8% (1005)	1.7% (614)	1.6% (579)	1.5% (538)
# of Racks End 15 (18.7%)	# of Racks End 19 (23.4%)	# of Racks End 19 (100.0%)	# of Racks End 15 (19.2%)	# of Racks End 5 (19.1%)
Start Month 3 (100.0%)	Start Month 4 (100.0%)	Start Month 3 (100.0%)	Start Month 2 (100.0%)	Start Month 1 (100.0%)
Usertype Customer (100.0%)	Usertype Subscriber (100.0%)	Usertype Customer (100.0%)	Usertype Customer (100.0%)	Usertype Customer (100.0%)







Two-step- dive deep