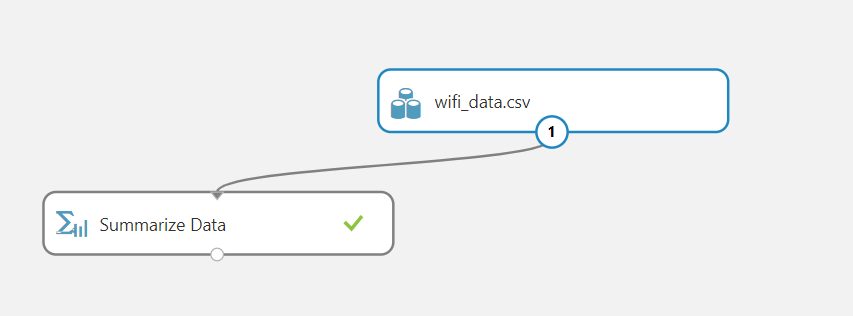
**PROJECT – 4**

1. Preprocessing the data. ( 4 points)

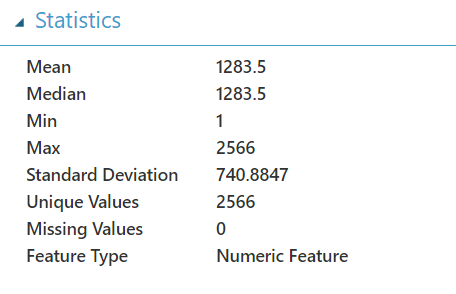


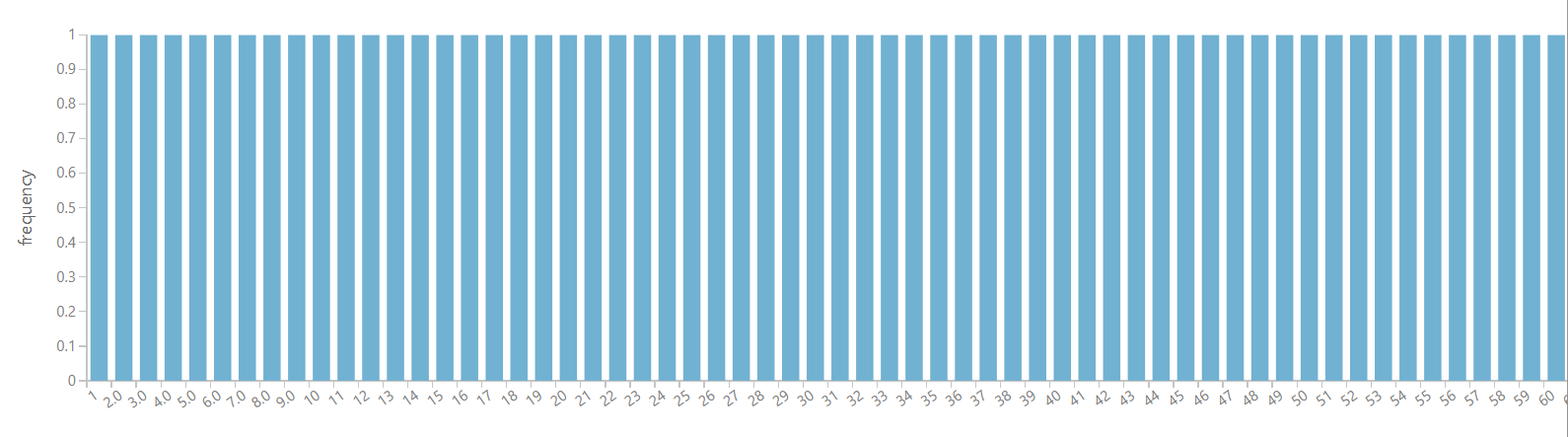
* 1. Report the size of the data set and check frequency distribution of the columns and write your inferences. ( 2 points)



Dataset have 2566 rows and 10 columns.

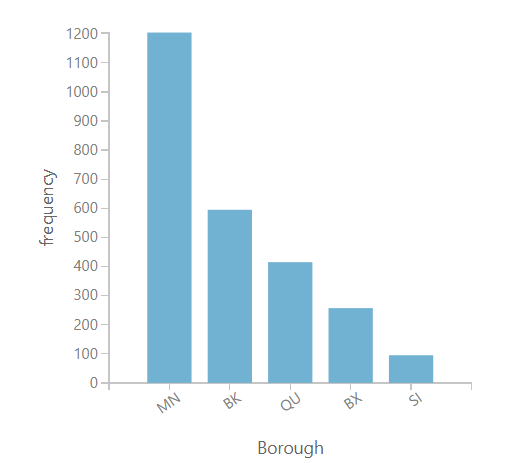
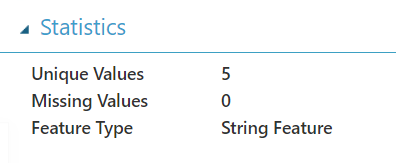
1. OBJECT-ID

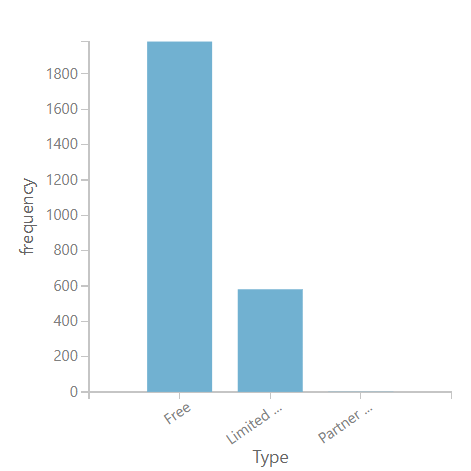
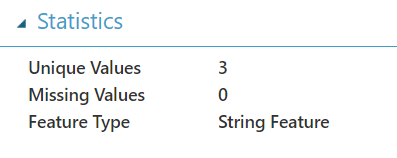


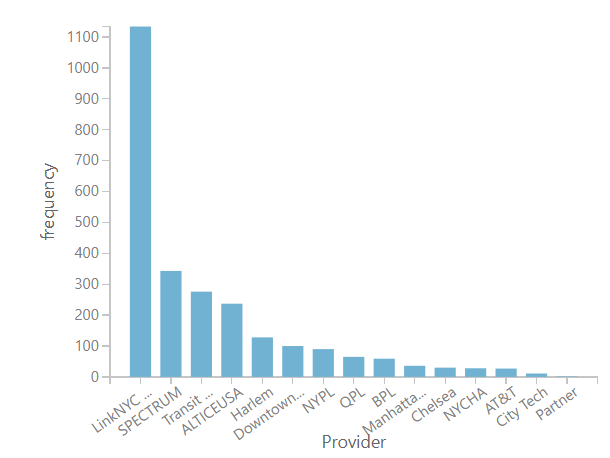
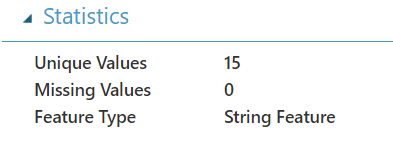


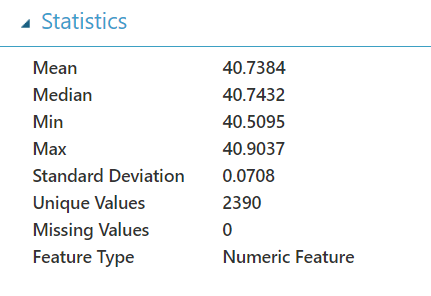
All the values are unique.

2. BOROUGH

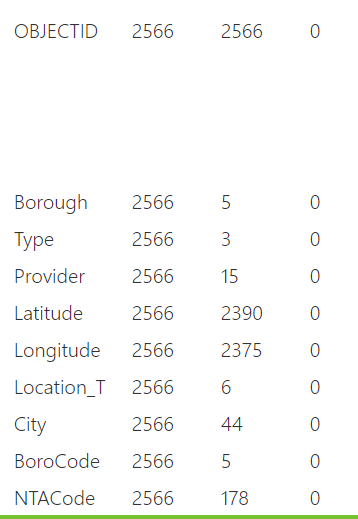








There are no missing values in dataset.



And the above are the unique value count for each column

Data summary

------------------------------------------------------------------------------------------------------------

Feature Count Unique Value Count Missing Value Count Min Max Mean Mean Deviation 1st Quartile Median 3rd Quartile Mode Range Sample Variance Sample Standard Deviation Sample Skewness Sample Kurtosis P0.5 P1 P5 P95 P99 P99.5

OBJECTID 2566 2566 0 1 2566 1283.5 641.5 642.25 1283.5 1924.75 {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,43,44,45,46,47,48,49,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,72,73,74,75,76,77,78,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,101,102... 2565 548910.166667 740.884719 0 -1.2 13.825 26.65 129.25 2437.75 2540.35 2553.175

Borough 2566 5 0

Type 2566 3 0

Provider 2566 15 0

Latitude 2566 2390 0 40.509531 40.903723 40.738396 0.056115 40.693069 40.743196 40.796066 40.687191 0.394192 0.005019 0.070845 -0.404158 -0.065426 40.568045 40.572667 40.591861 40.84605 40.874629 40.880524

Longitude 2566 2375 0 -74.244107 -73.714838 -73.947564 0.048173 -73.985954 -73.958354 -73.922968 -73.769559 0.529269 0.004509 0.06715 0.582311 1.67407 -74.138192 -74.112341 -74.011847 -73.808307 -73.755022 -73.7441

Location\_T 2566 6 0

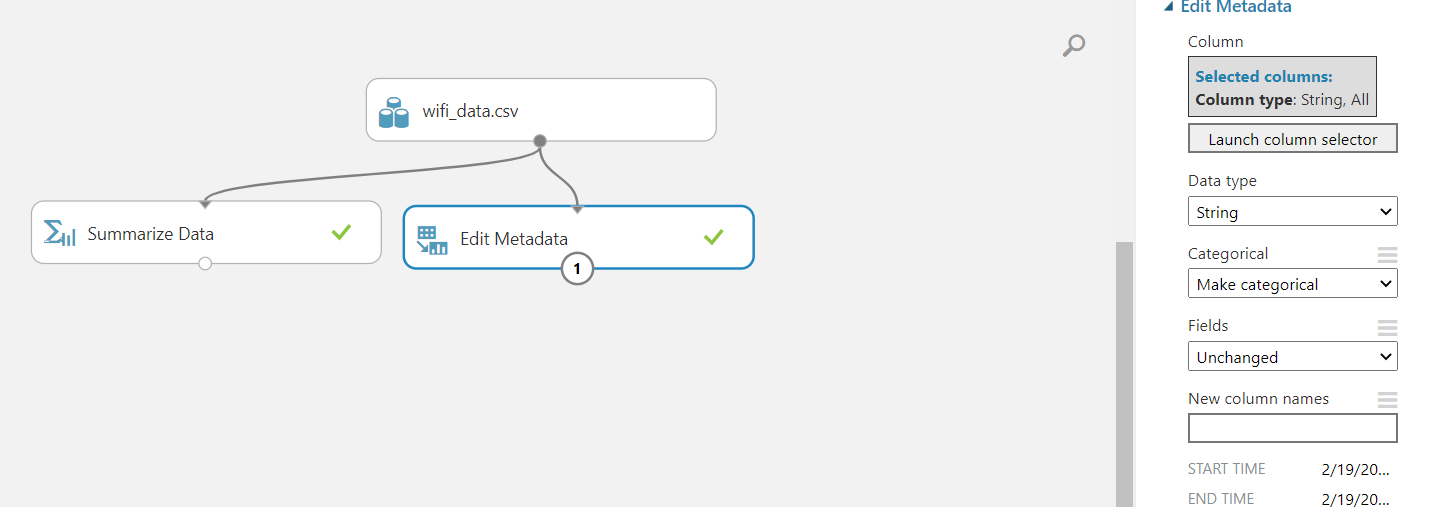
City 2566 44 0

BoroCode 2566 5 0 1 5 2.196804 1.163467 1 2 3 1 4 1.644294 1.2823 0.512822 -1.126972 1 1 1 4 5 5

NTACode 2566 178 0 ------------------------------------------------------------------------------------------------

My inference is that this data set contains the details of different wifi hotspot providers which can be either free, limited free or partnered and the code and their locations are given. This dataset will help us find the regions where wifi hotspot will tend to cluster and regions where there are no free wifi providers so that we can find the target audience for paid wifi providers.

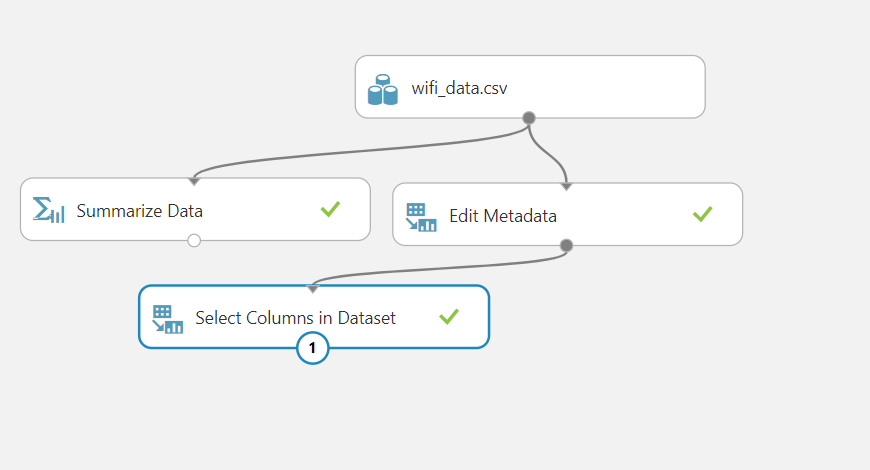
* 1. Convert string features into categories. (1 points)



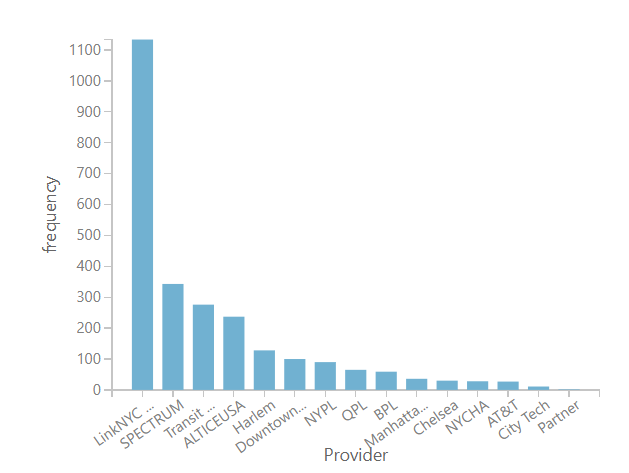
Made string data categorical.

* 1. Select relevant features for model building. Drop the redundant and irrelevant features. ( 1 point)

Object ID, latitude and longitude are not useful so we are dropping that column out. There so many unique values and less common ones.



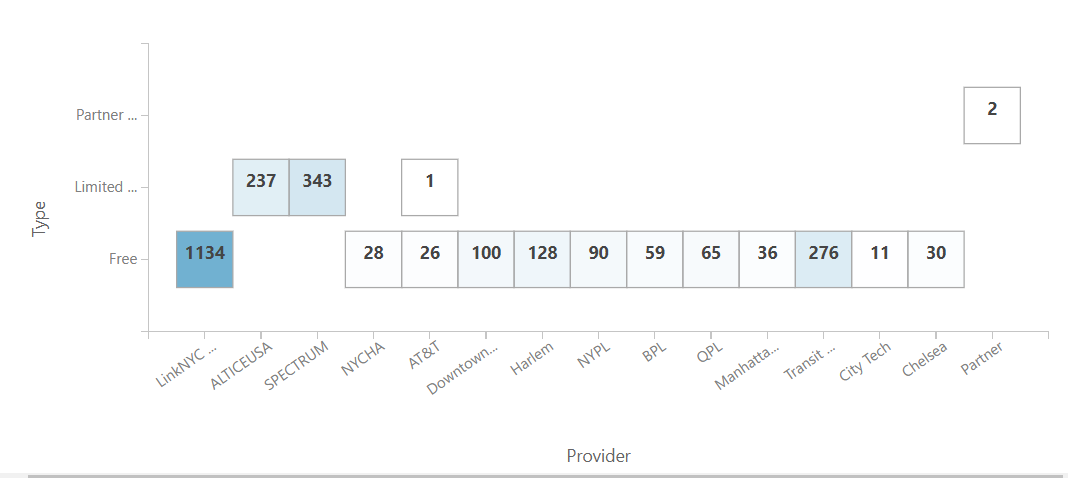
1. Perform univariate and bivariate analysis and answer the following questions. ( 8 points)
   1. Which provider has the highest number of wifi hotspots? ( 1 point)



LinkNYC – Citybridge has highest no. of wifi providers that is 44% i.e. 1134 wifi hotspots.

* 1. Which provider provides the highest number of free wifi hotspots? ( 1 point)

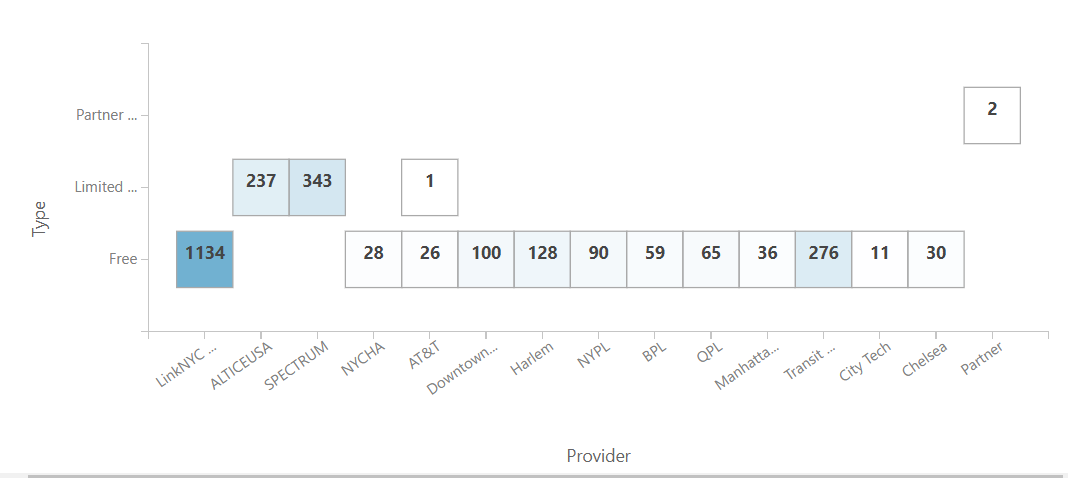
Wifi provider with highest no. of free wifi hotspots is LinkNYC – Citybridge and no. is 1134



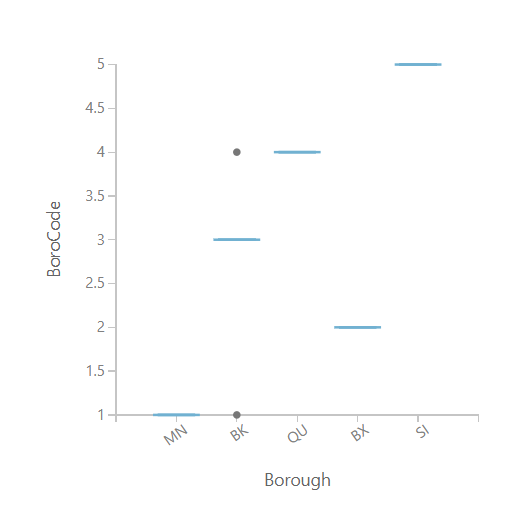
* 1. List down the name of providers that provide “Limited Free” type of wifi. ( 1point)

There are 3 wifi providers who provide Limited Free wifi and they are –

* + 1. ALTICEUSA
    2. SPECTRUM
    3. AT&T



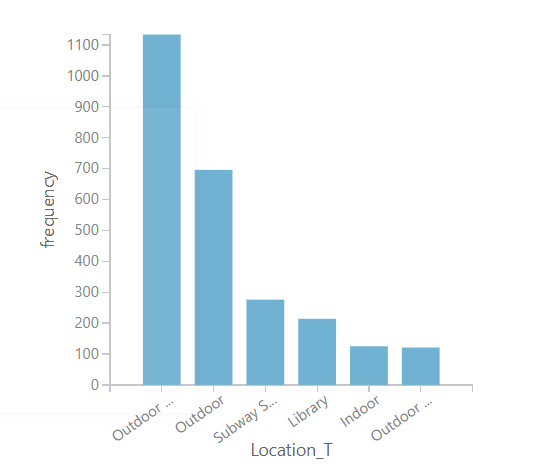
* 1. What is the correlation coefficient between borough and BoroCode. ( 2 points)

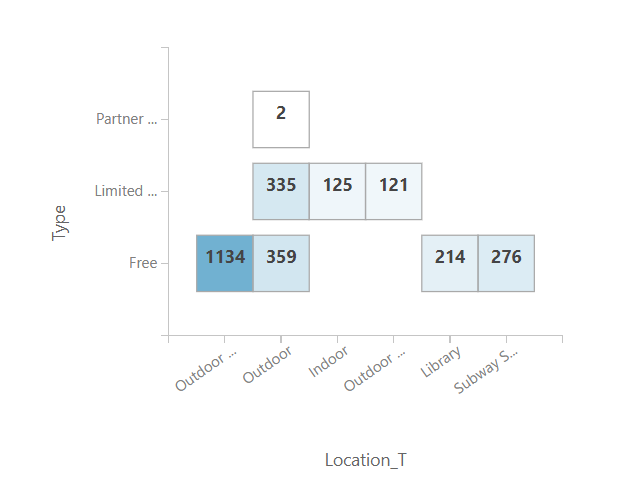




As boro code increases there are less number of borough and also a particular borough as usually same boro code. Only BK has boro codes 1, 3, 4 rest have same boro code for a given borough.

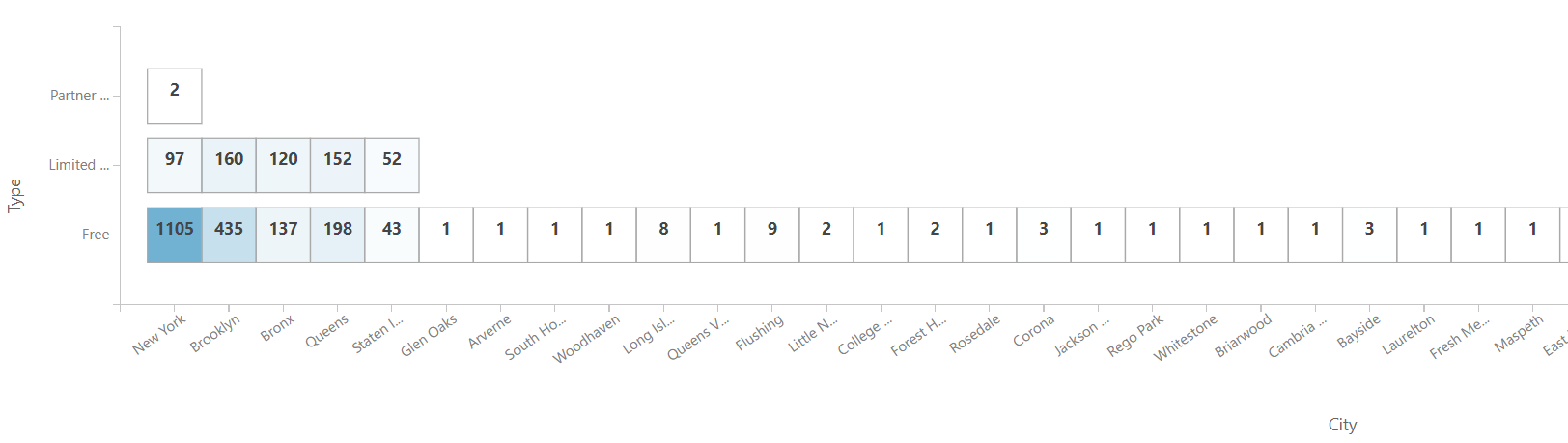
* 1. Are there any indoor free wifi hotspots? ( 2 points)

There are 129 i.e. 4.9% indoor wifis



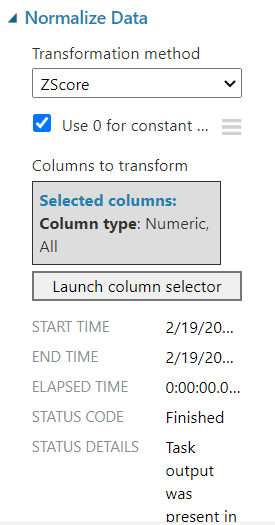
No, there are no free wifi provider which are indoor

* 1. Which city has the “partner\_site” type wifi hotspot? ( 1 point)



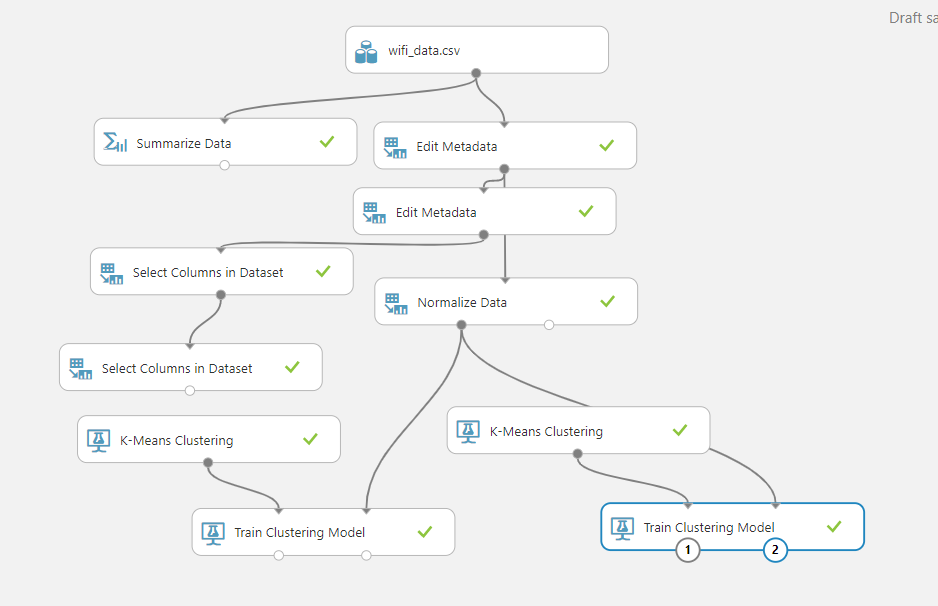
Only New York have “partner\_site” type wifi.

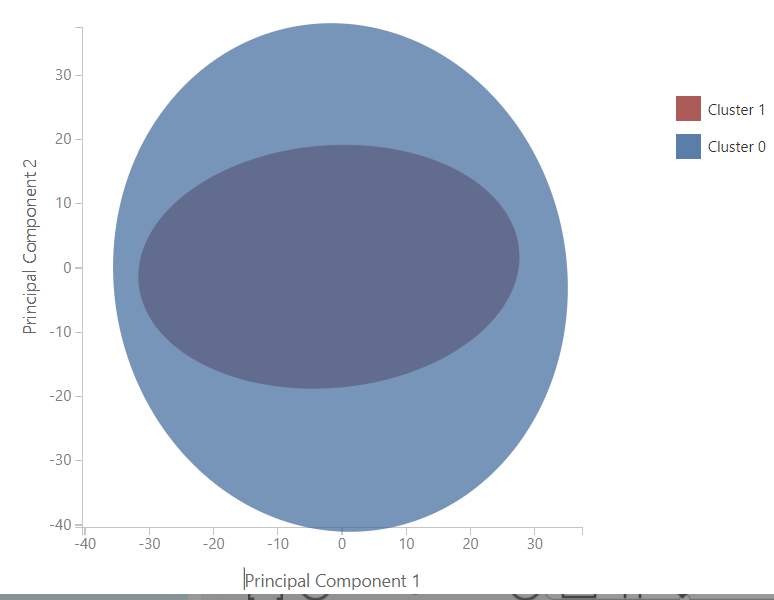
1. Apply following techniques to identify the natural cluster of wifi hotspots locations. And mention the number of optimal clusters and their properties. ( 6 points)
   1. Normalize the data. (1 point)

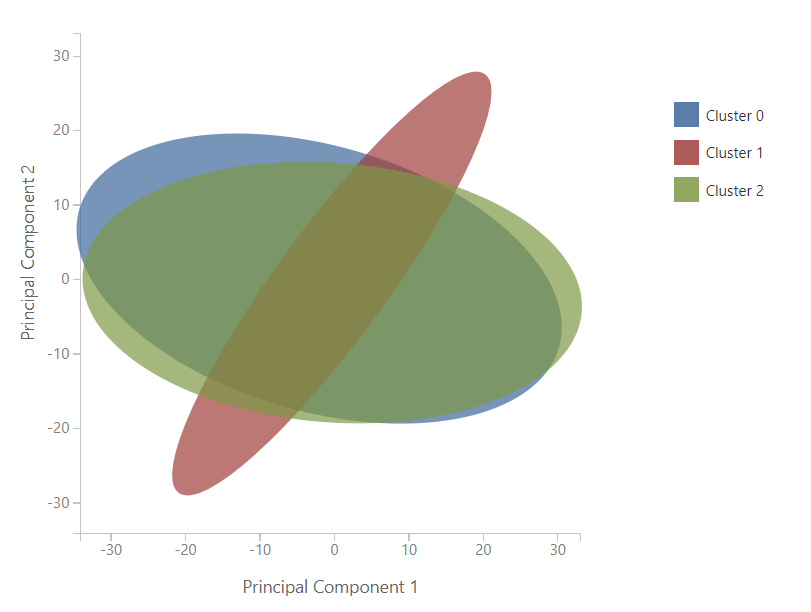


Did Z-score normalization

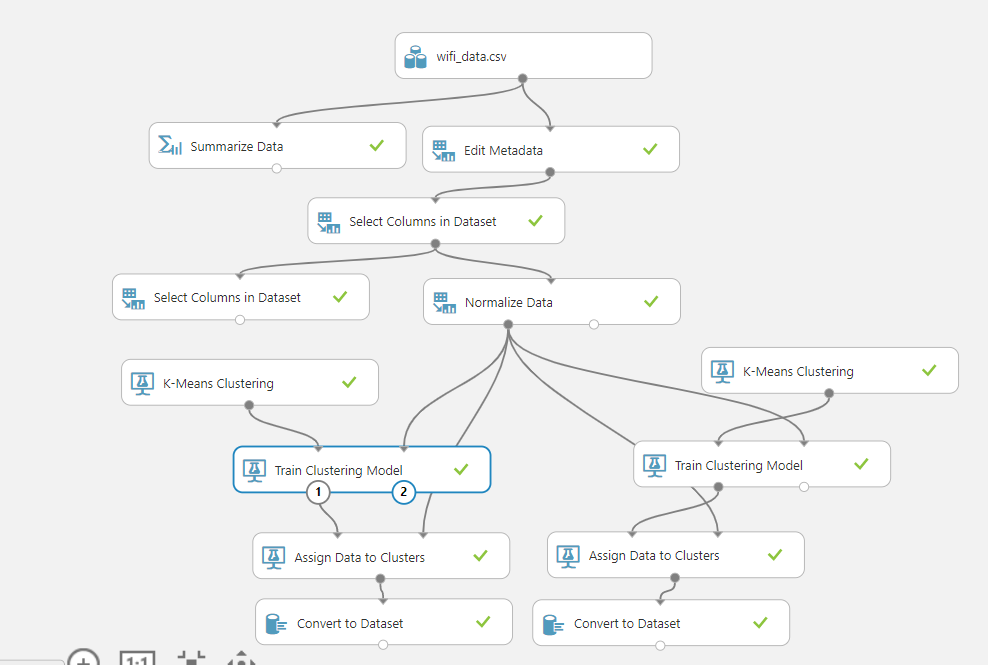
* 1. Build a K-means clustering model for k=2 and 3. ( 2 points)



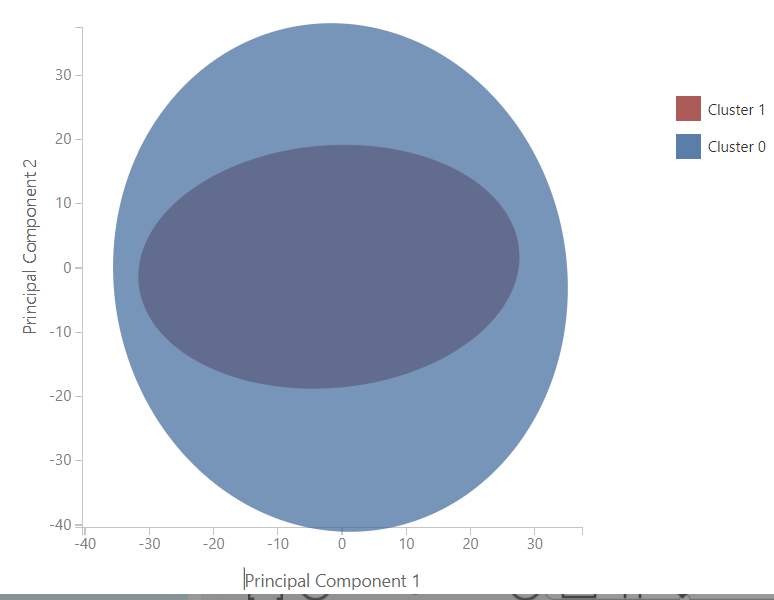


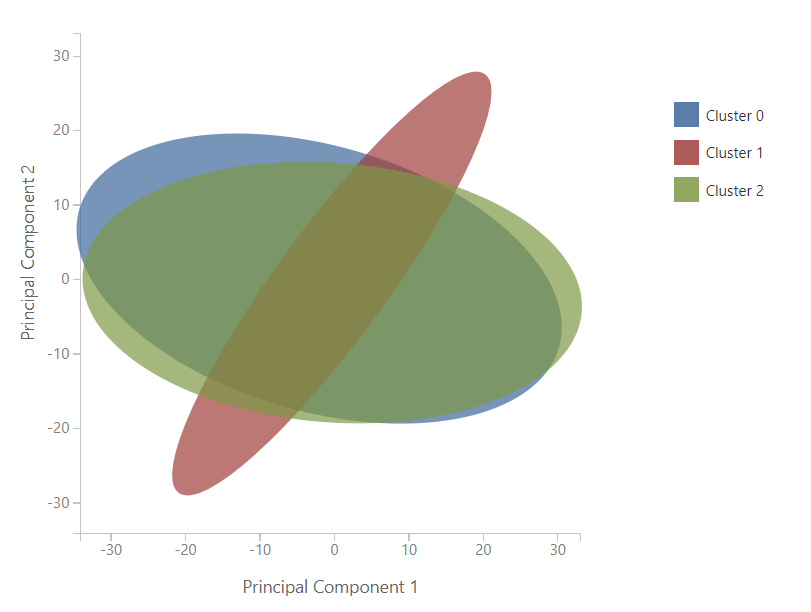


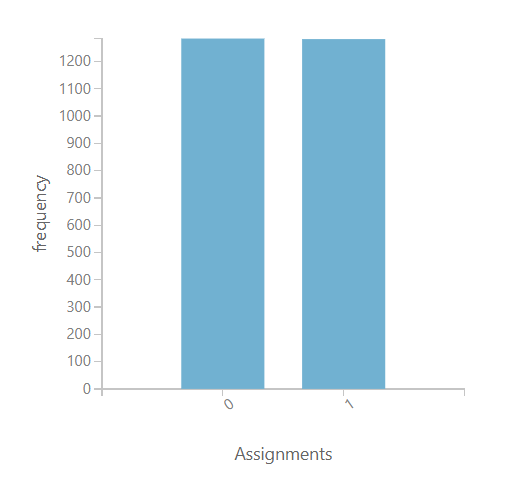
Trained 2 models with 2 and 3 clusters.

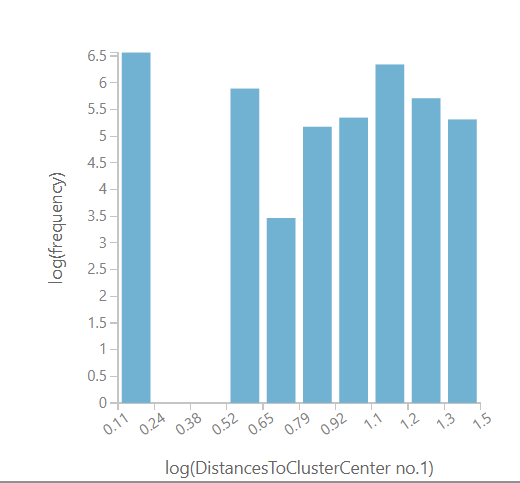
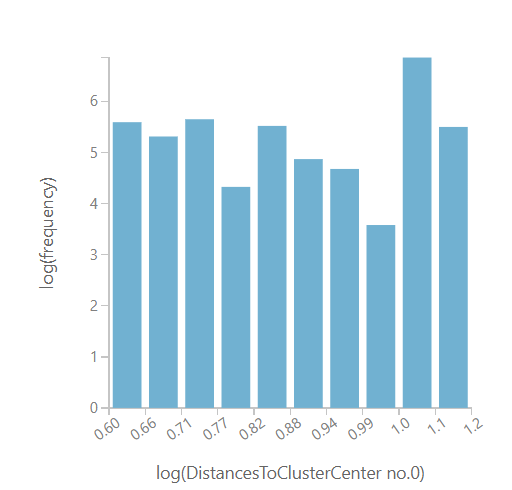


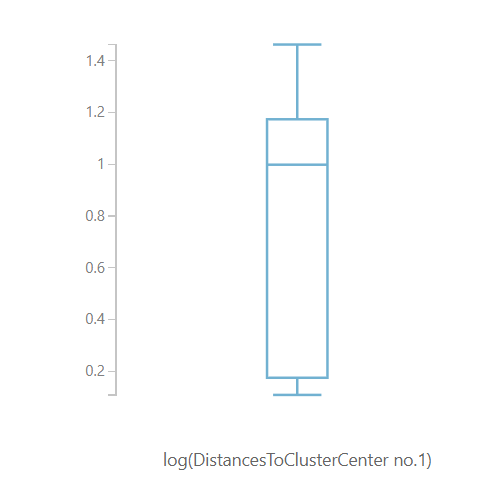
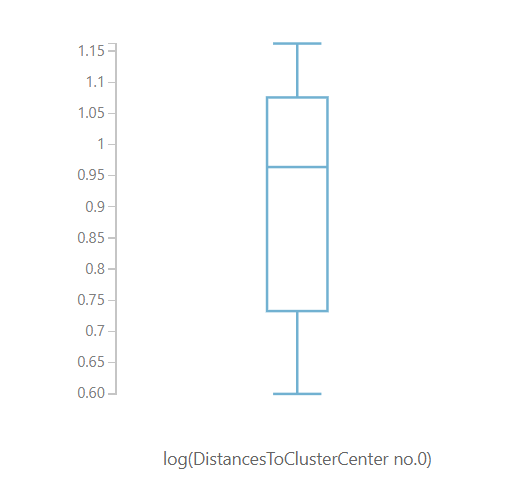
* 1. Visualize the clusters. ( 1point)

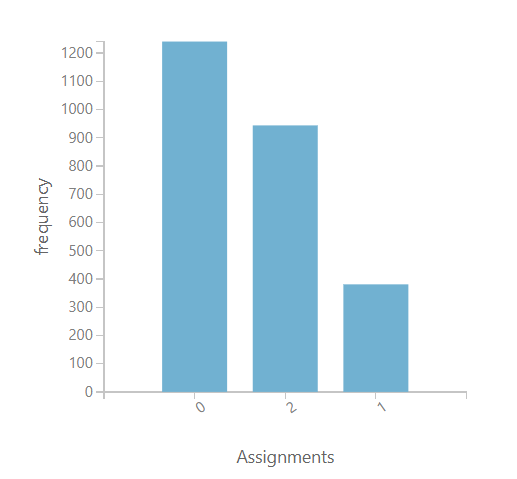


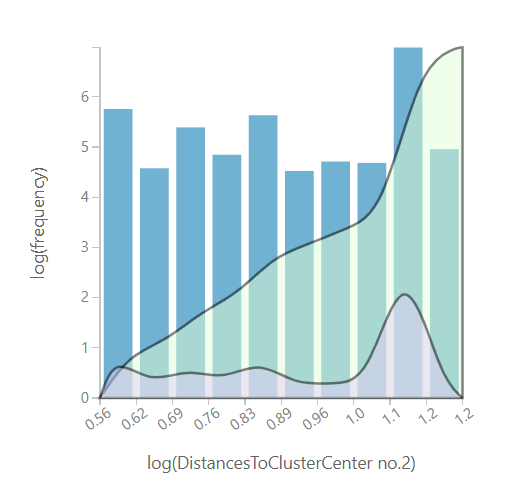
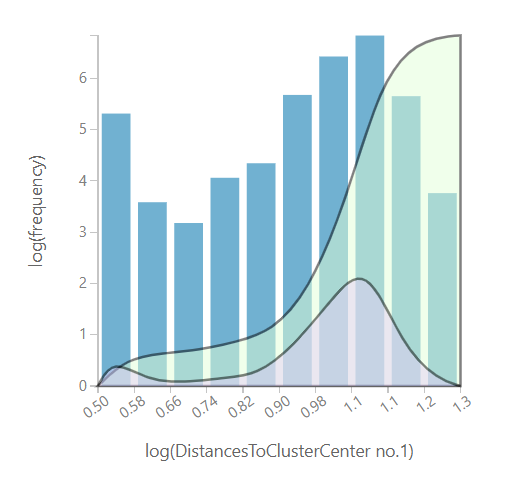
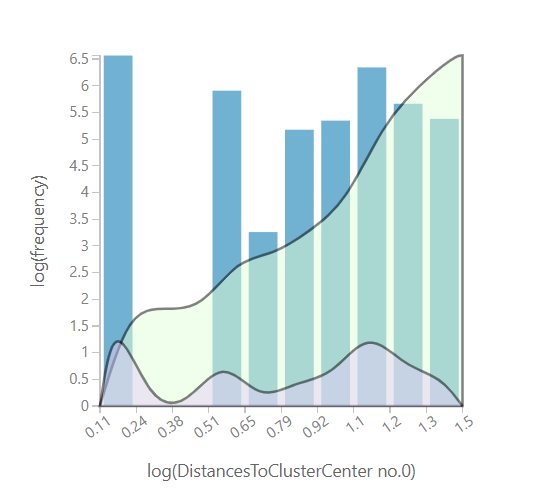


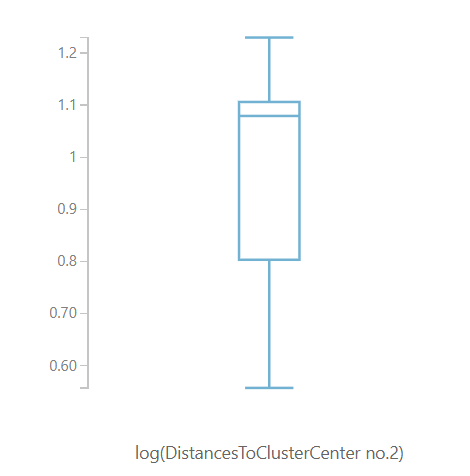
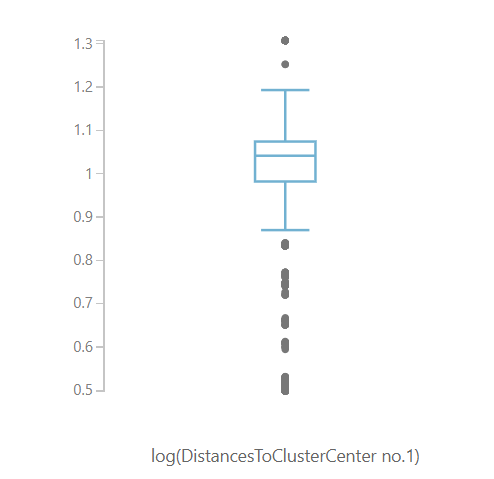
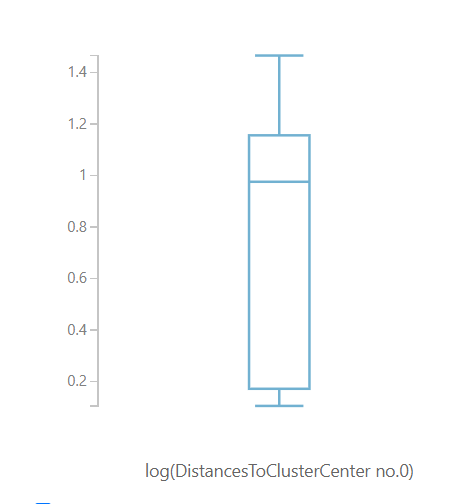




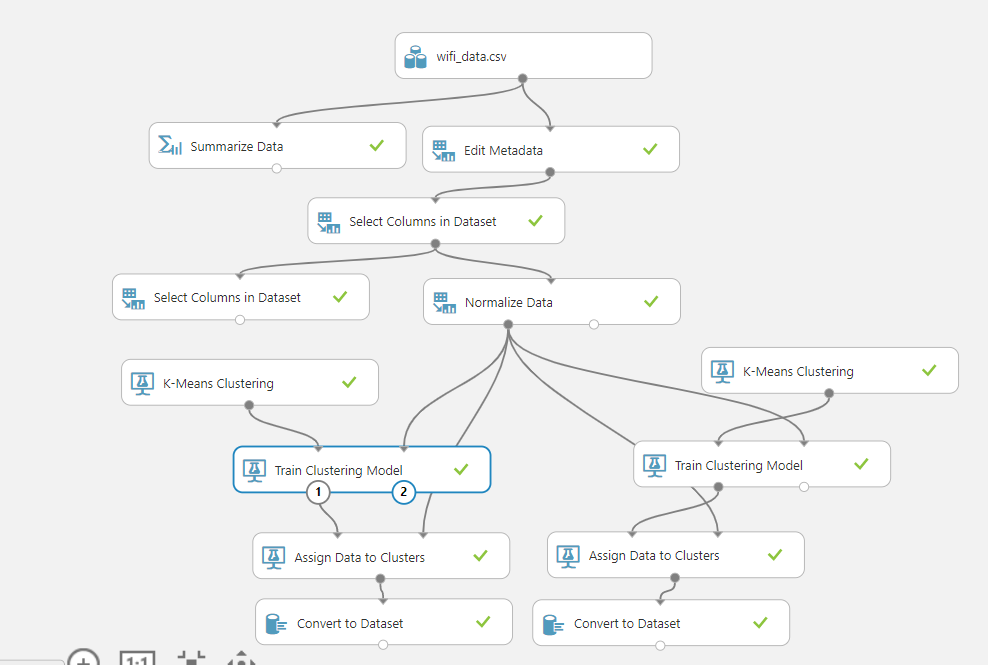




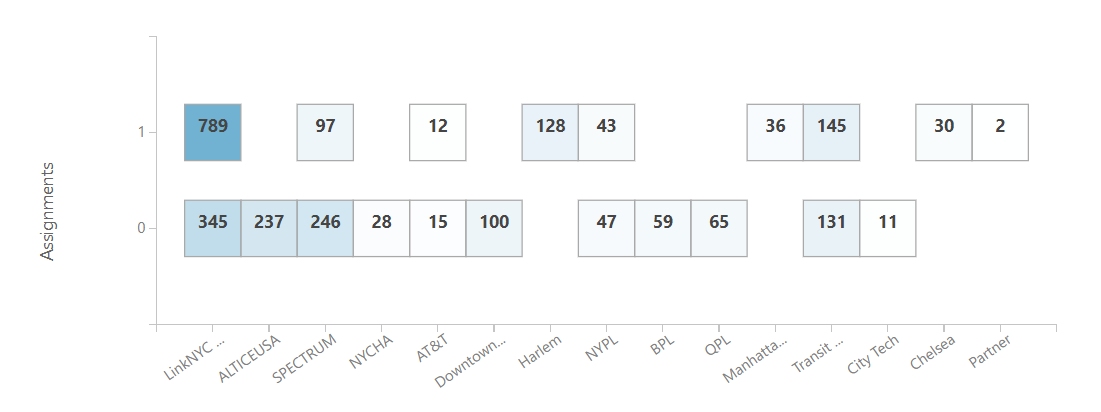
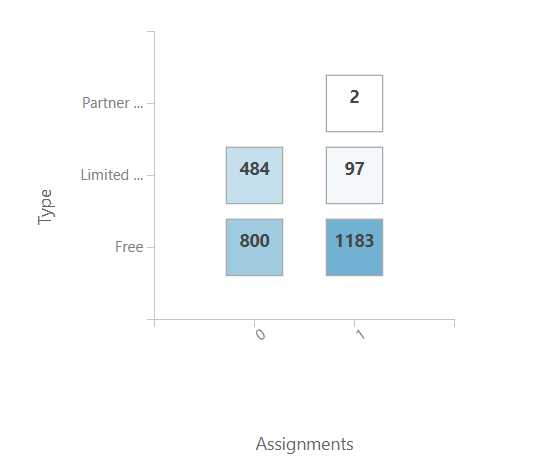
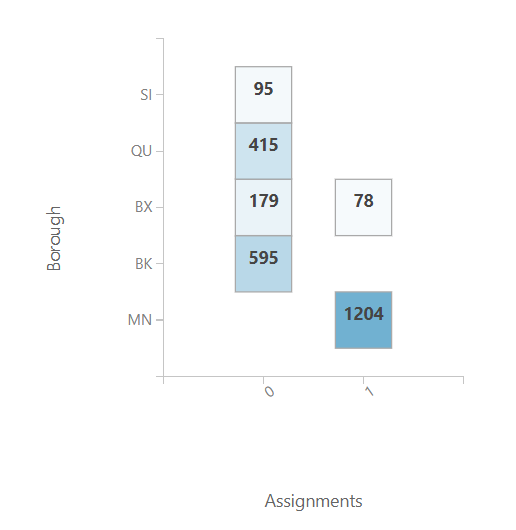


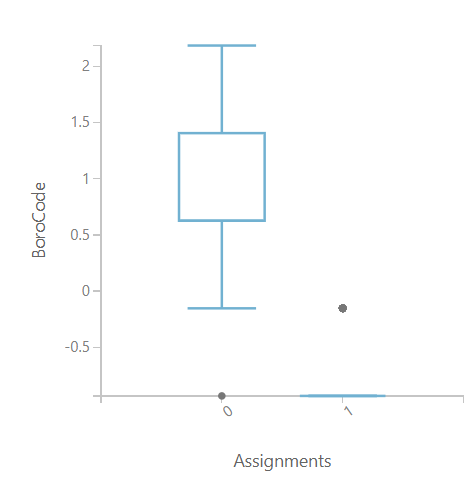


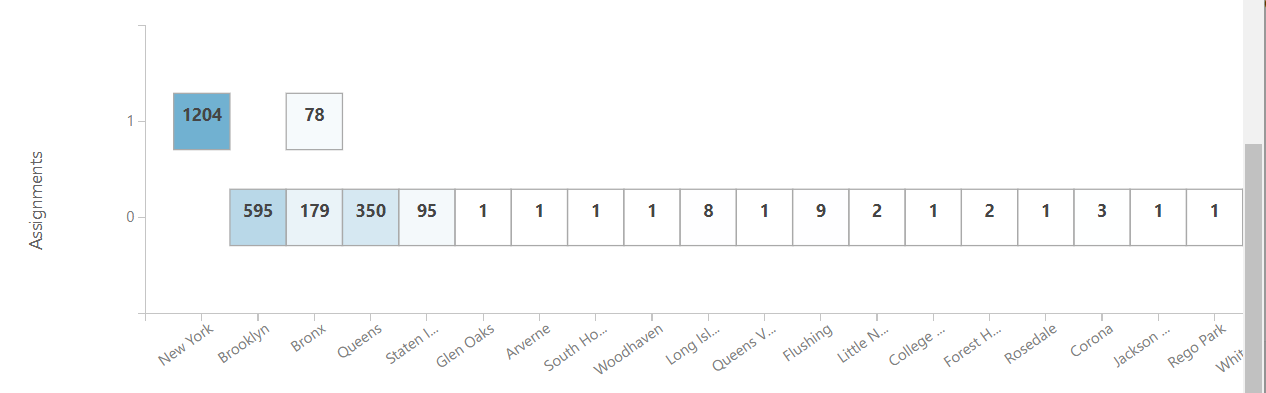
* 1. Assign cluster labels to the dataset and perform bivariate analysis between cluster labels and various features and write your inferences. ( 2 points)

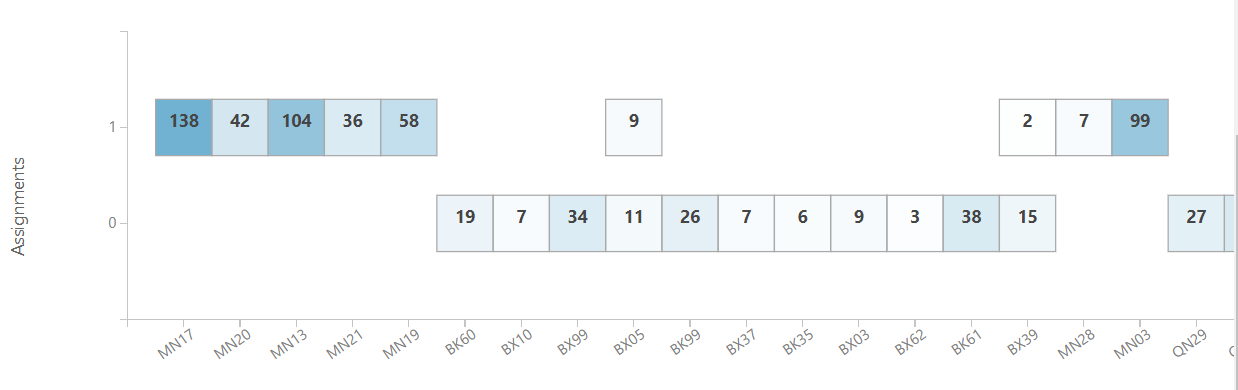


For 2 cluster classification

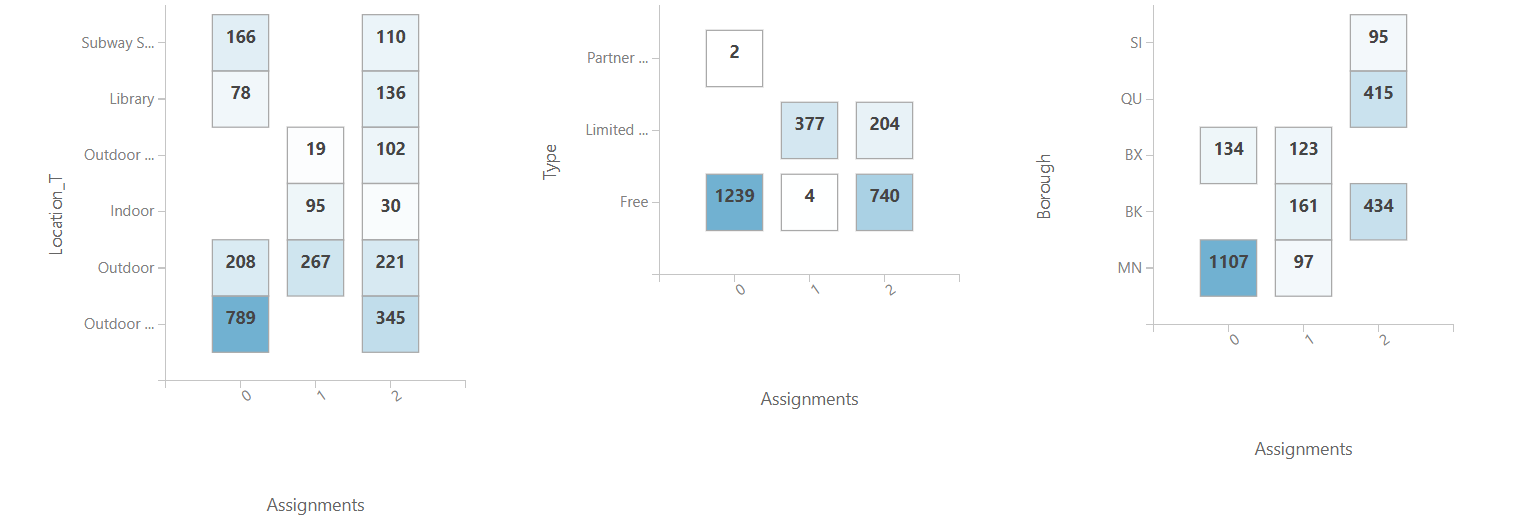


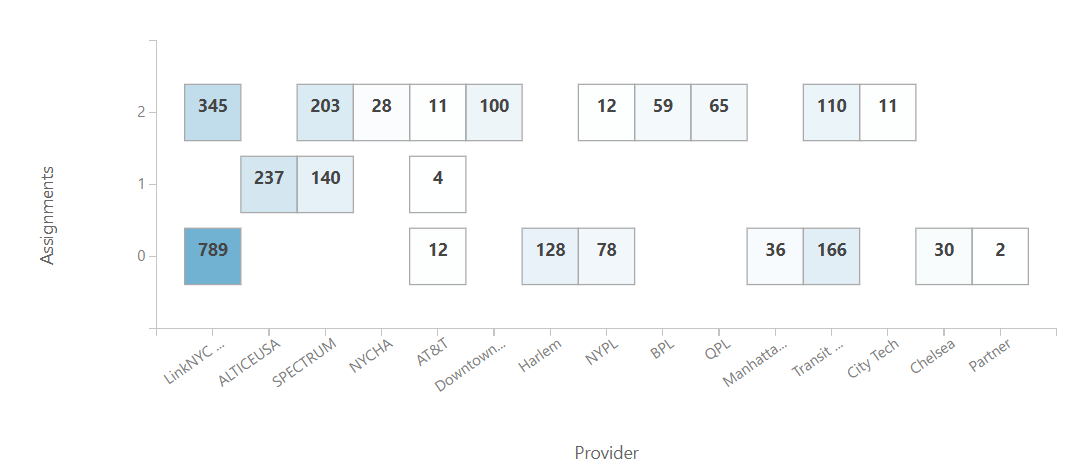
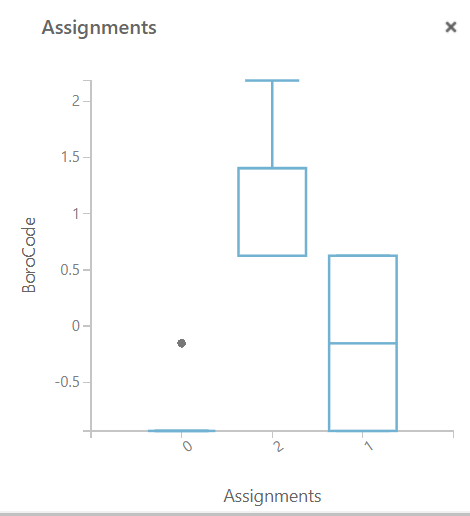


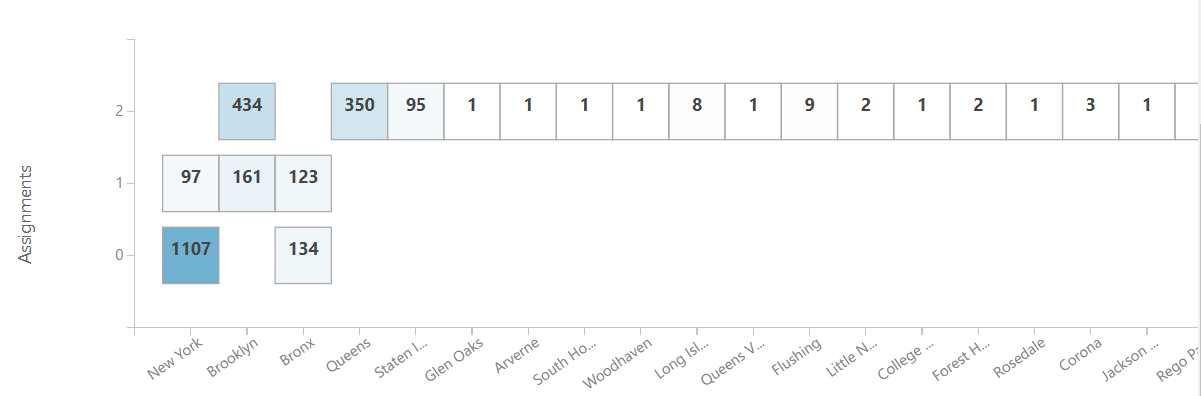


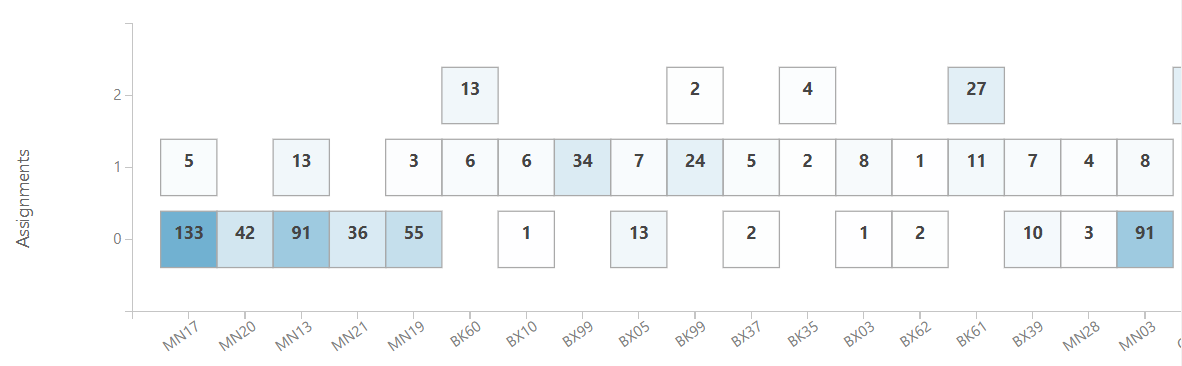


One of the major thing we notice here is that in the city column the assignment of cluster 0 are none and all are from cluster 1 and also there is only 1 more city having cluster 1 so we can say city id not a major classification feature.









I prefer 3 cluster classification as there is less overlap between the the clusters in the first 1 there are so many overlaps that its kind of intersecting so 3 cluster classification is better.