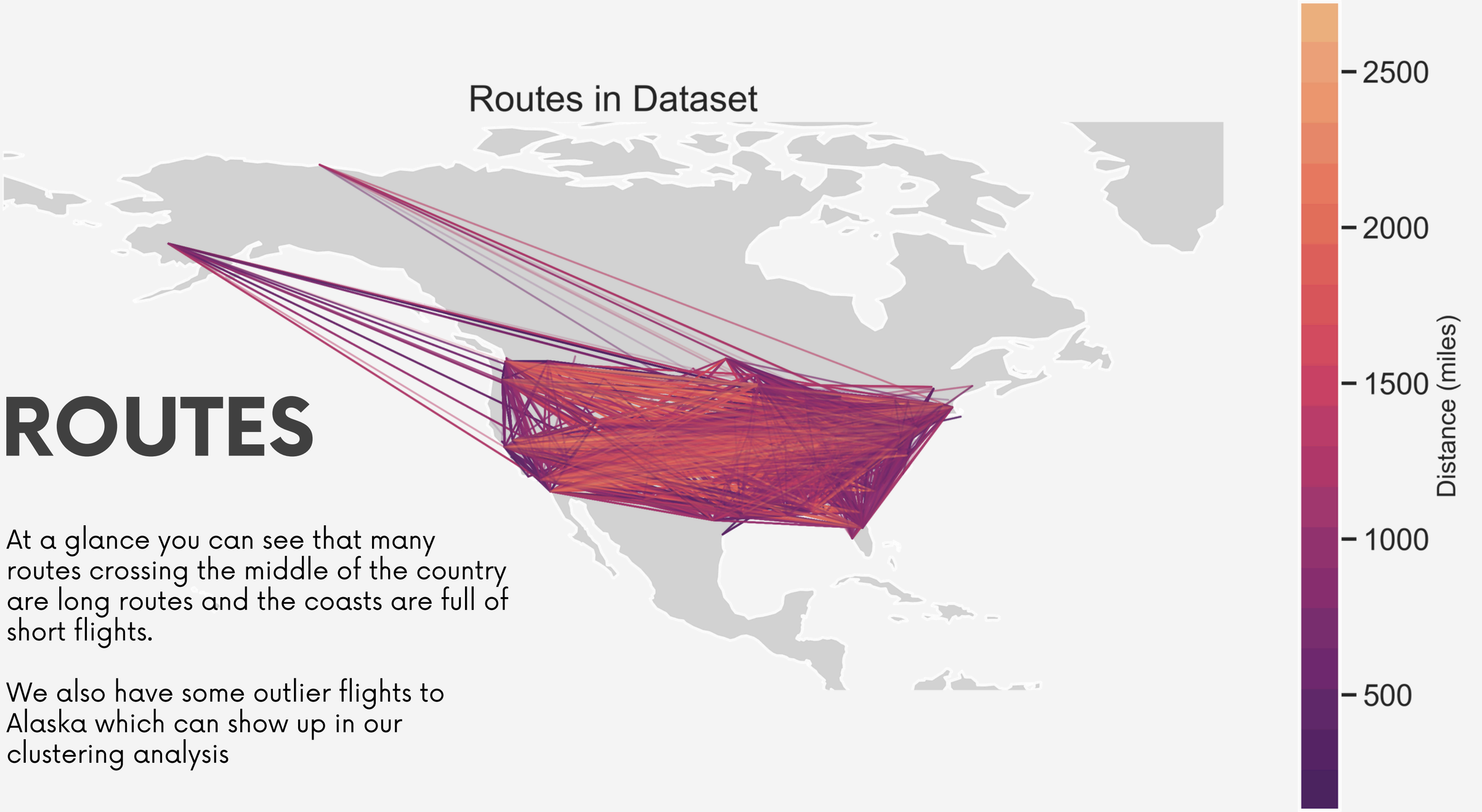
SKY ROUTE

DATE
09/23/2024

CUSTERING ANALYSIS

Identifying distinct groups of city pairs with similar passenger patterns

PRESENTED BY
Byung Jae Bae



ROUTES

At a glance you can see that many routes crossing the middle of the country are long routes and the coasts are full of short flights.

We also have some outlier flights to Alaska which can show up in our clustering analysis

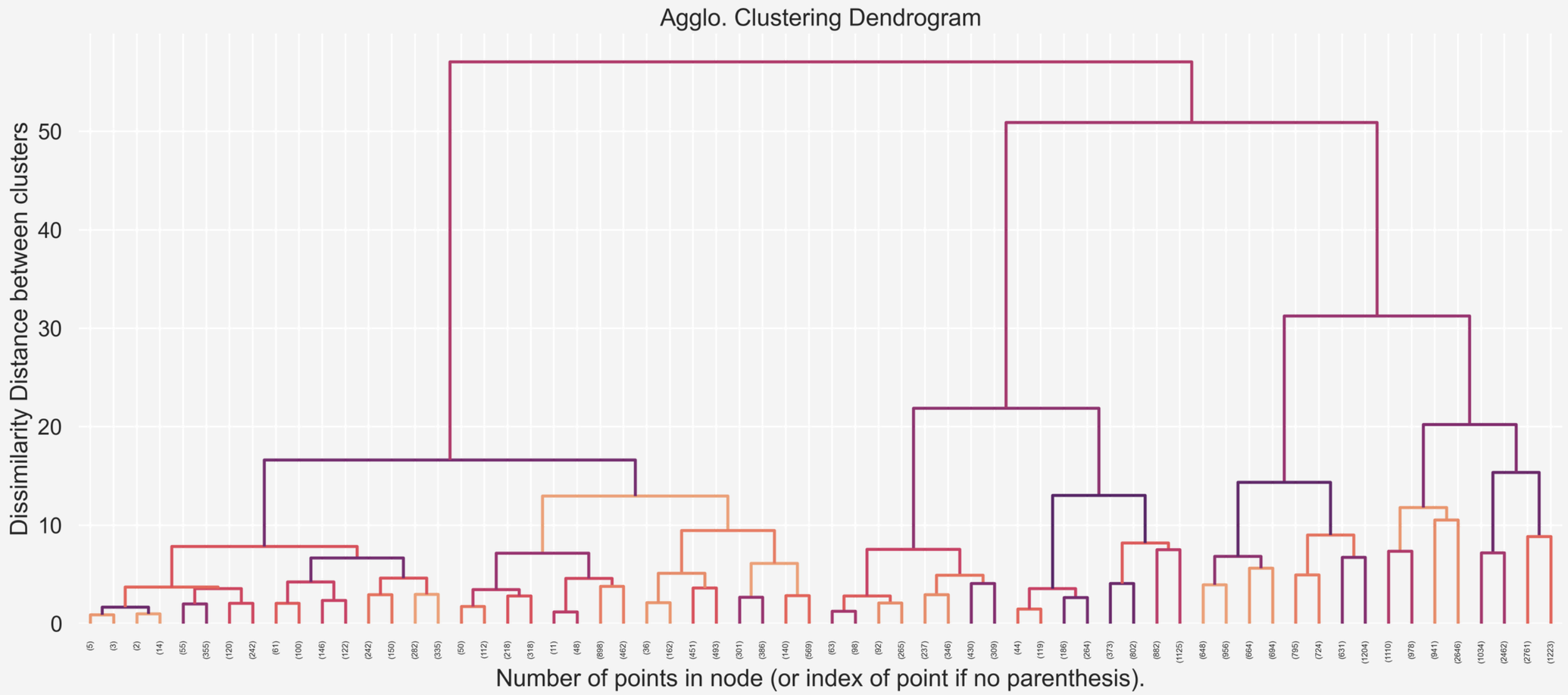
LIMITATIONS OF ANALYSIS

This clustering analysis does not reflect direct customer behavior and uses metrics like average fares and market share to reflect market demand

CLUSTERING ON CONTINUOUS FEATURES

Agglormarative Hierarchacal Clustering was used to see how well clusters differenciate from different clusters.

We can see that there are about 2-3 clusters that have very high dissimilarity from each other.



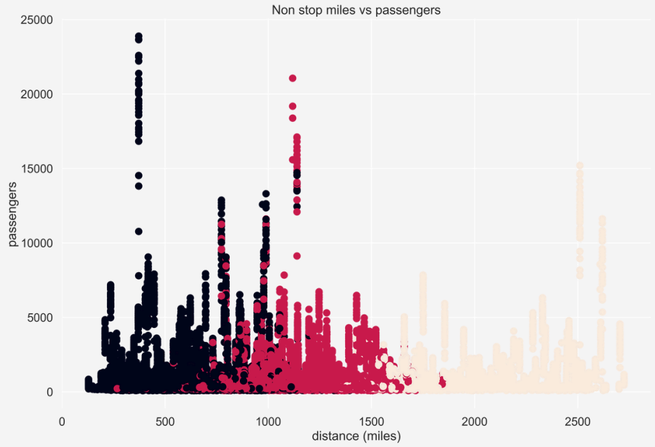
COMPARISON

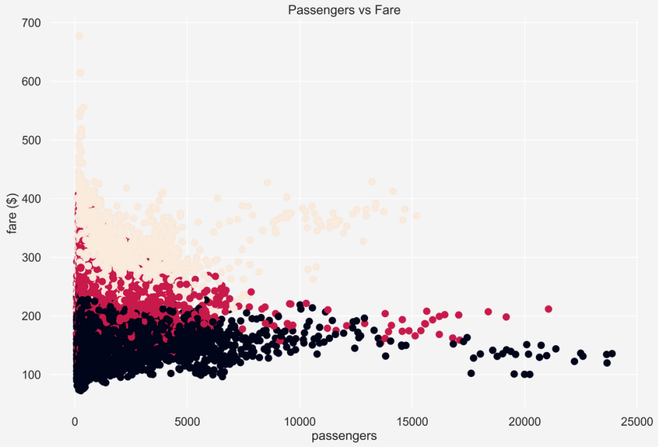
The top two charts show K-Means with 3 clusters.

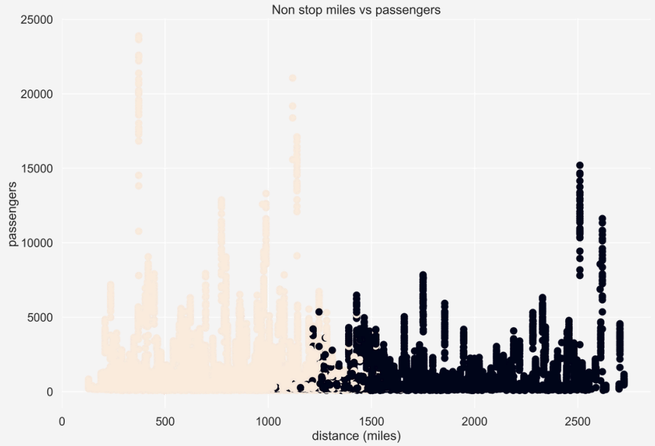
The bottom two charts show K-Means with 2 clusters.

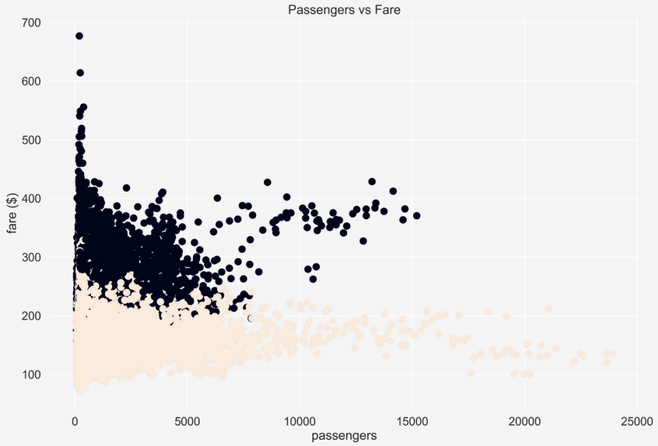
Silhouette score (higher is better):
0.32236013543110364 (Top)
vs
0.4894144534509019 5 (Bottom)

Davies-Bouldin score (lower is better):
1.1110435866127217 (Top)
vs
0.7982241607373859 (Bottom)



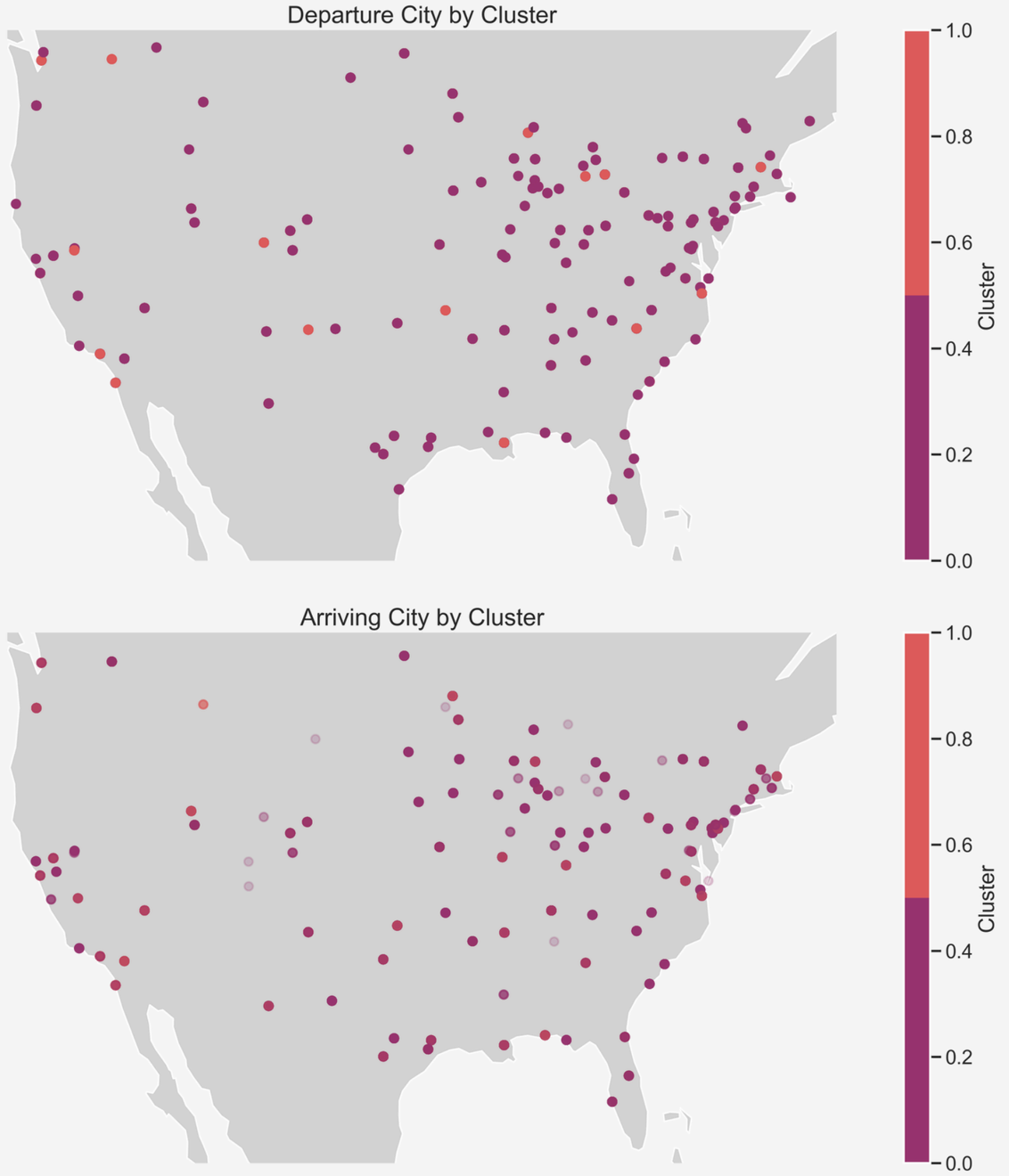






CHECKING HOW WELL IT SPLITS GEOGRAPHIC REGIONS

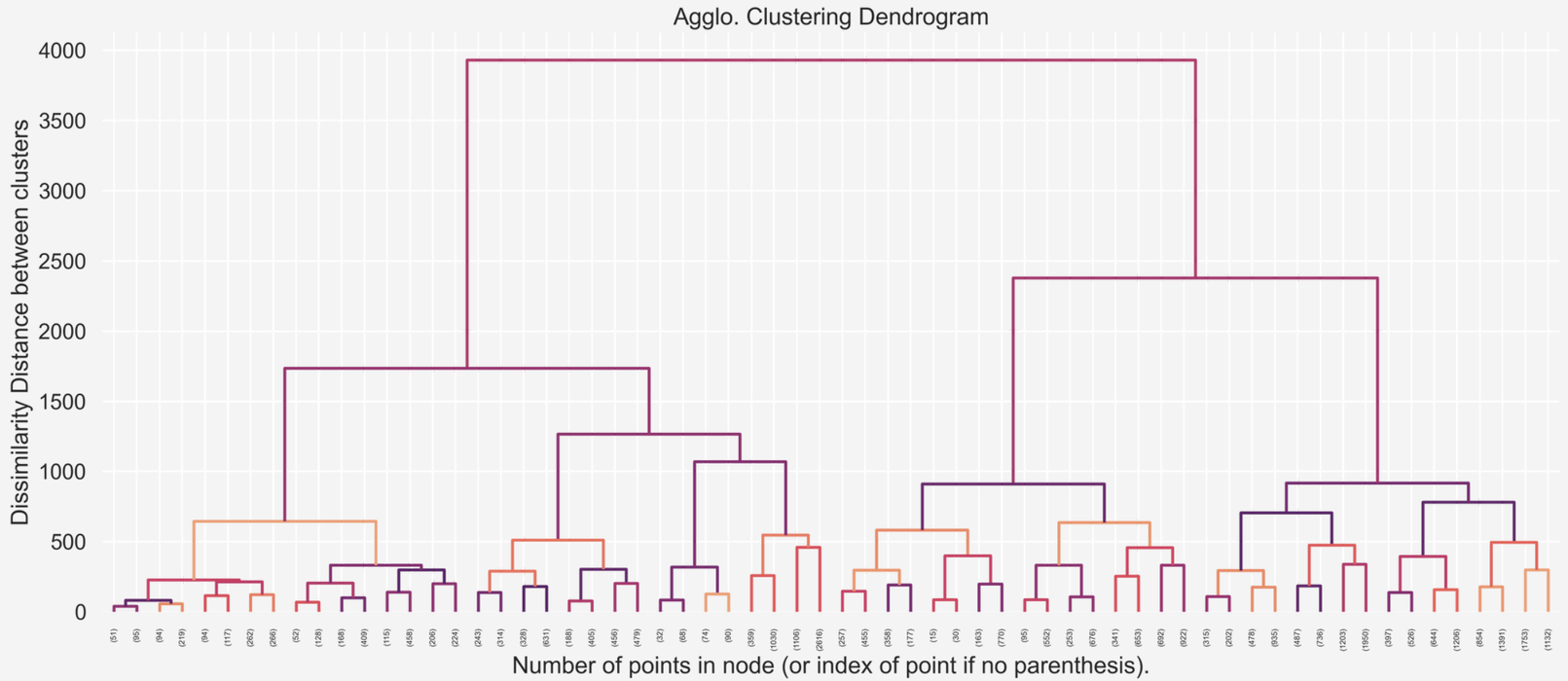
It seems that clustering on non geographic data doesn't reflect much clustering in terms of geographical regions.



CLUSTERING ON COORDINATES ONLY

Instead of using other features we try clustering on coordinates

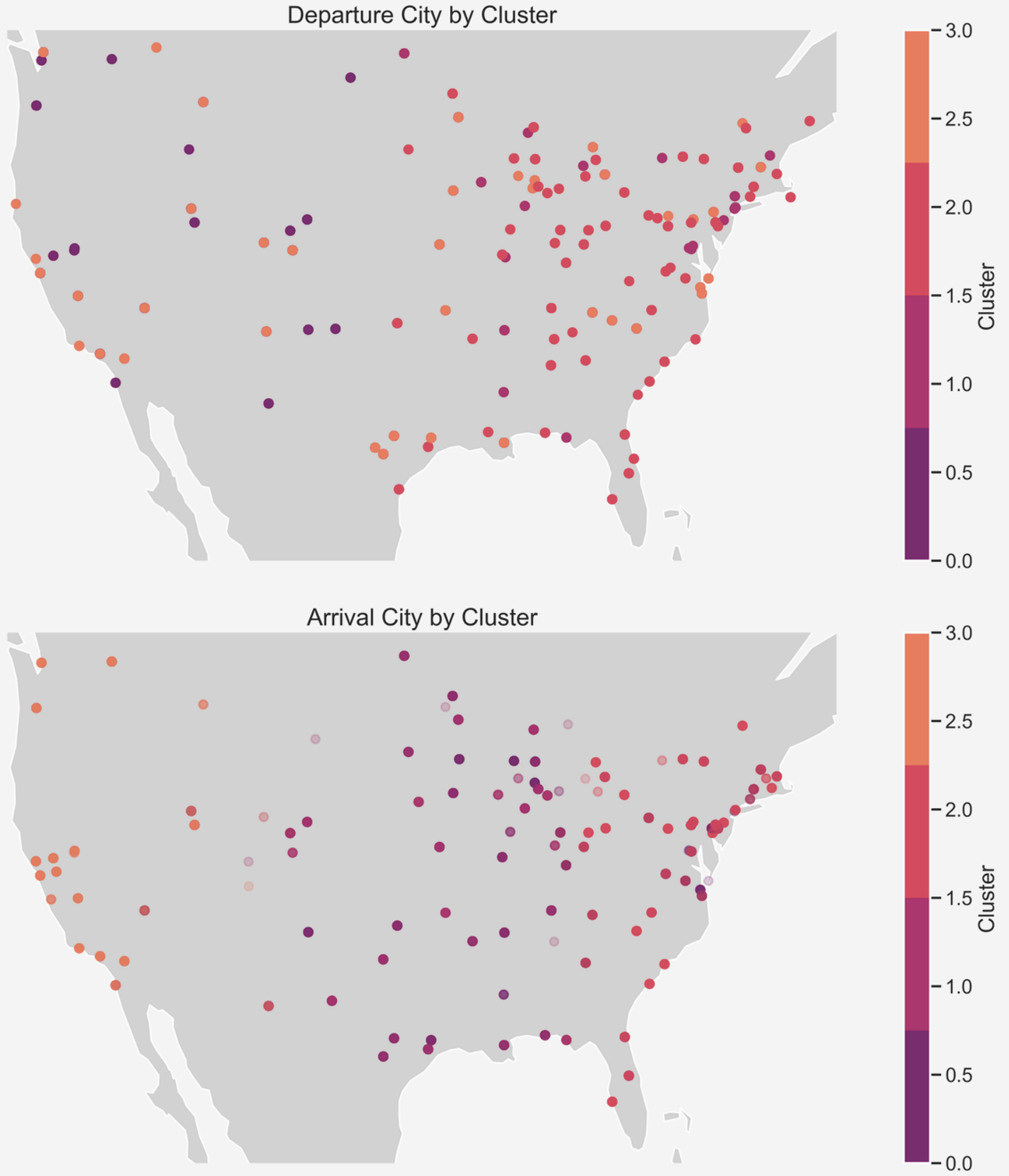
There is 4 nearly equal clusters that seem to appear.



GEOGRAPHICAL CLUSTERS

Even using coordinates, the clusters are not super defined.

There is a clear east and west coast seperation, but the center of the country has some overlap.



CONCLUSION

There are 2 clear clusters/markets and the more clusters create more overlap, performing less optimally.

Metrics also perform well when market share features are removed from the clustering.

Geographical locations have little impact on good clustering.

We can interpret the data that there is high low-cost and high-cost carrier seperation. There are no features that define a middle price tier market. It seems that high-priced low cost carriers and low-priced high cost airlines are both competing in this space. The routes reflects a bimodal distribution.

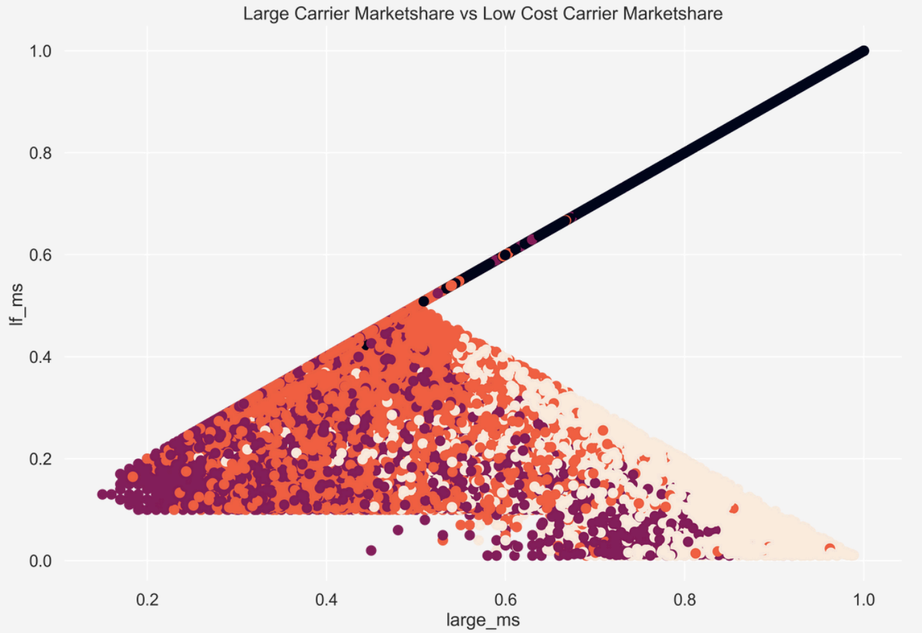
Appendix

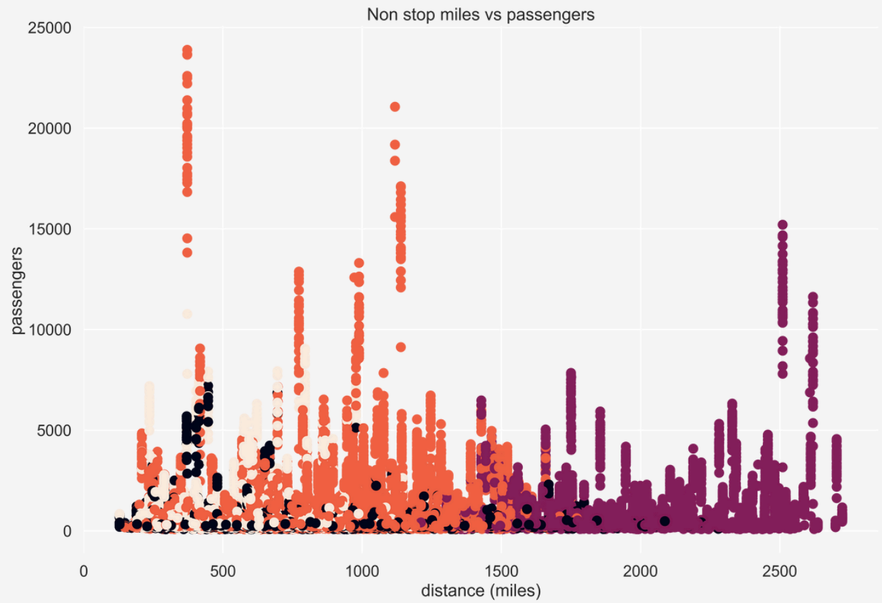
1st Agg. Hierarchical Clustering

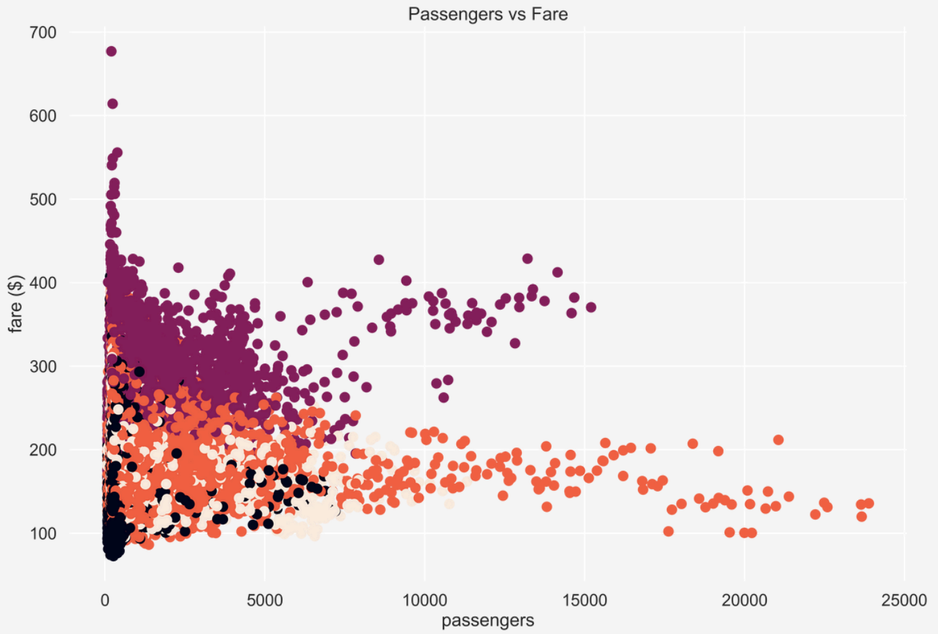
The average silhouette score is:
0.250756540985180 5

The Davies-Bouldin score is:
1.2707774363381936

'nsmiles', 'passengers', 'fare', 'large_ms', 'fare_lg', 'lf_ms', 'fare_low'





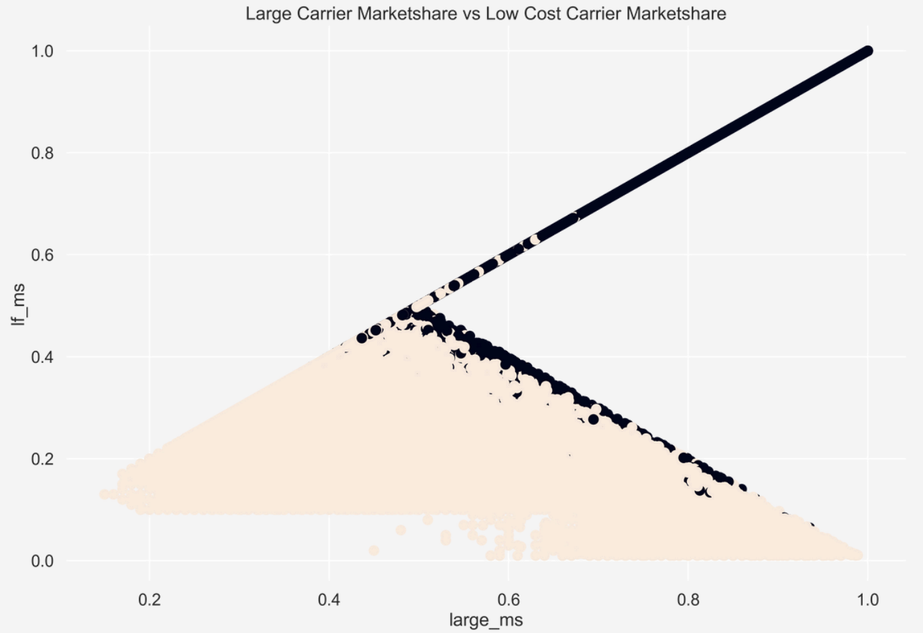


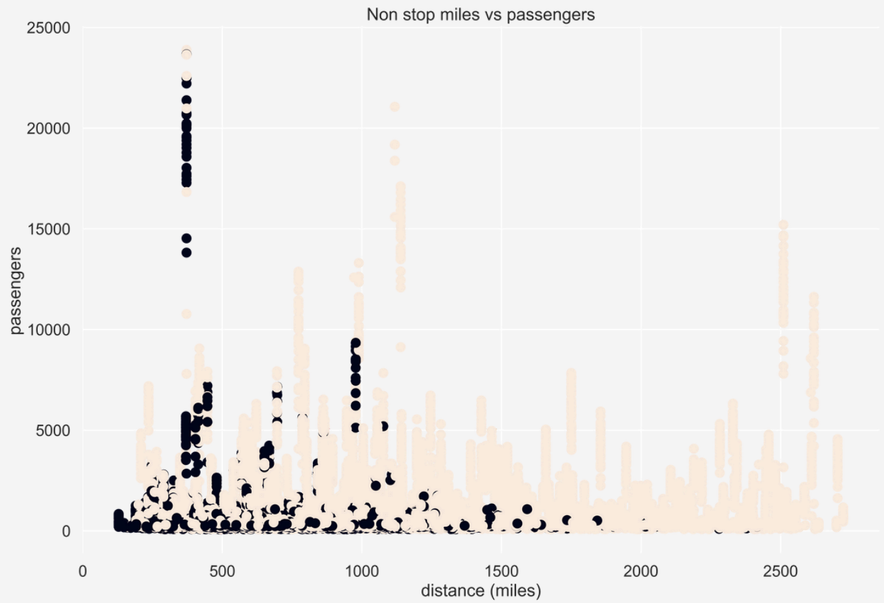
1st KMeans n_clusters=2

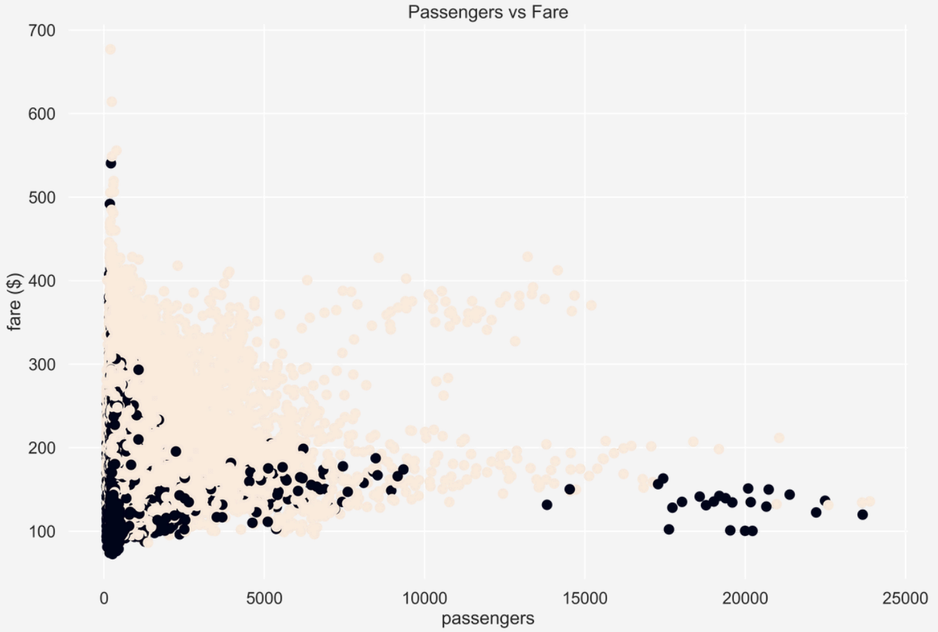
The average silhouette score is:
0.290106230289341 1

The Davies-Bouldin score is:
1.2307796042861823

'nsmiles', 'passengers', 'fare', 'large_ms', 'fare_lg', 'lf_ms', 'fare_low'





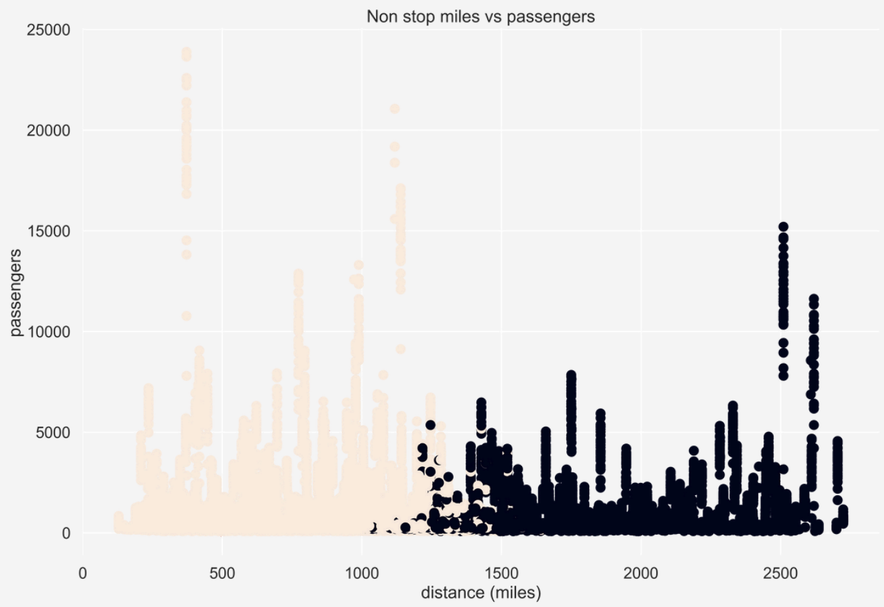


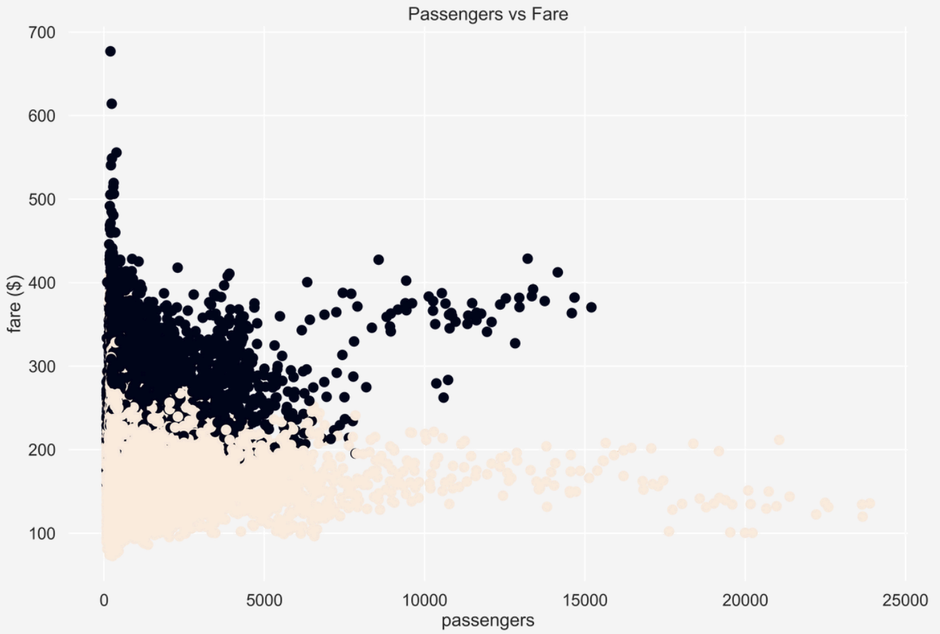
2nd KMeans n_clusters=2 (No market share)

The average silhouette score is: 0.4897933628767886

The Davies-Bouldin score is: 0.797307522739757

'nsmiles', 'passengers', 'fare', 'fare_lg', 'fare_low'



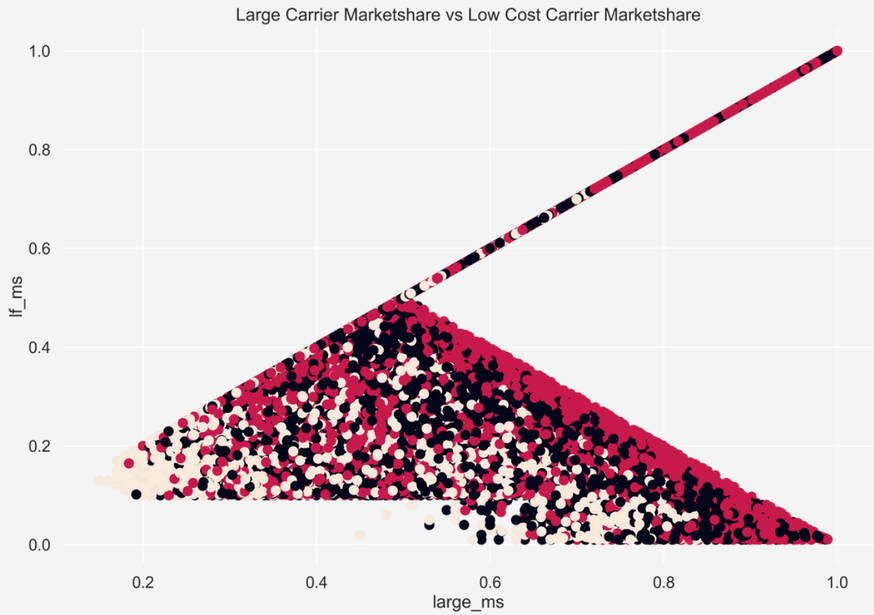


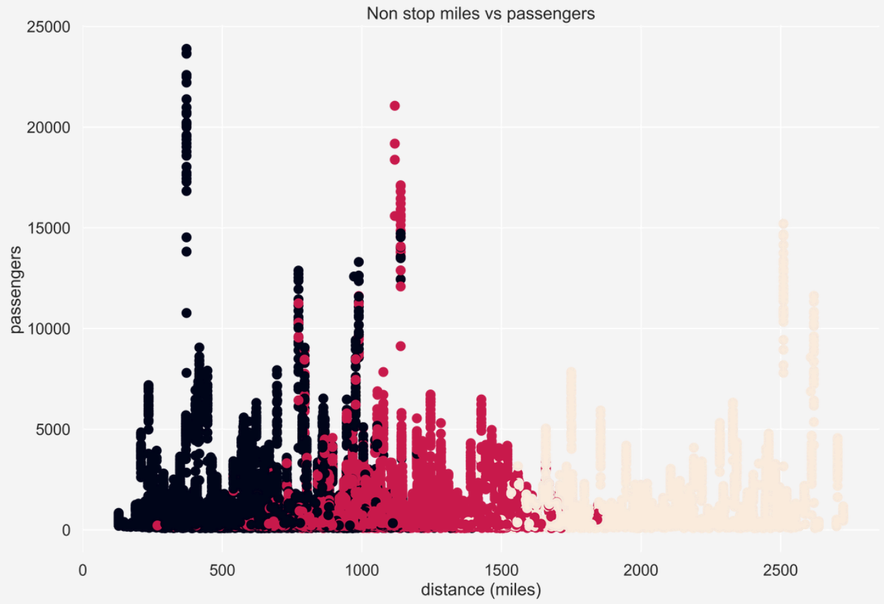
3rd KMeans n_clusters=3 (No market share)

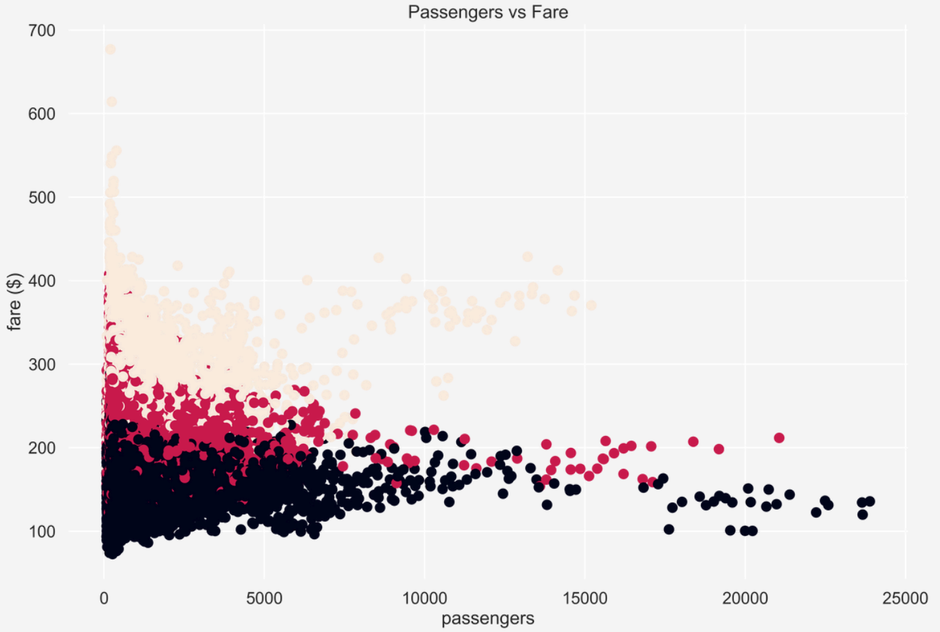
The average silhouette score is:
0.3223820720789846 7

The Davies-Bouldin score is:
1.1109552052058145

'nsmiles', 'passengers', 'fare', 'fare_lg', 'fare_low'



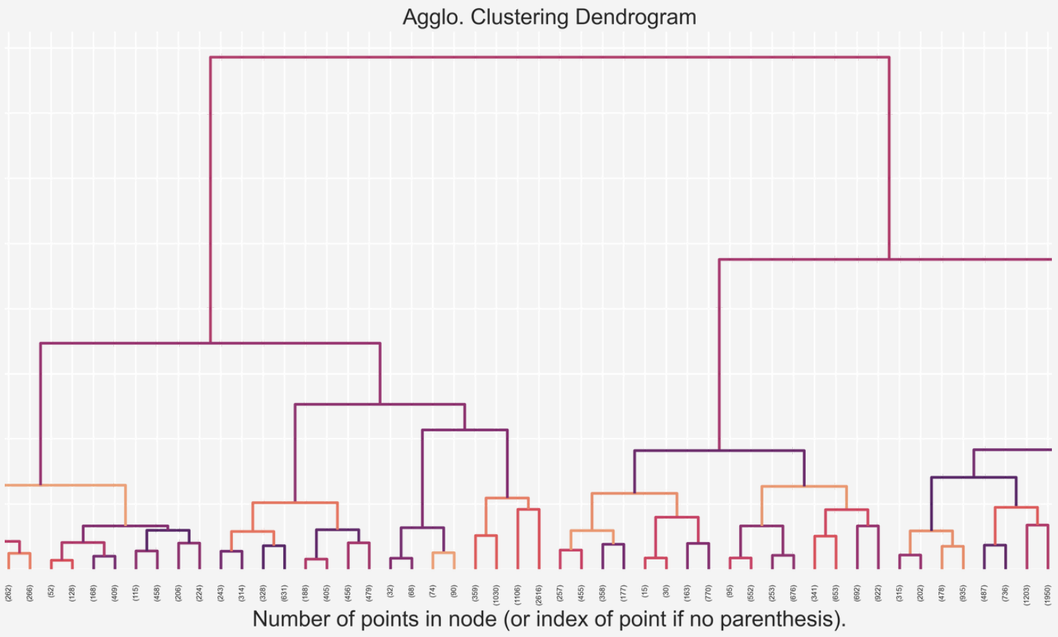




2nd Agg. Hierarchical Clustering (Coordinates Only)

The average silhouette score is:
0.3629596570519873 5

The Davies-Bouldin score is:
0.9915806024341887



"city1_lat",
"city1_long",
"city2_lat",
"city2_long"