MTR Station Clustering

Based on Foursquare Data

View Jupyter Notebook

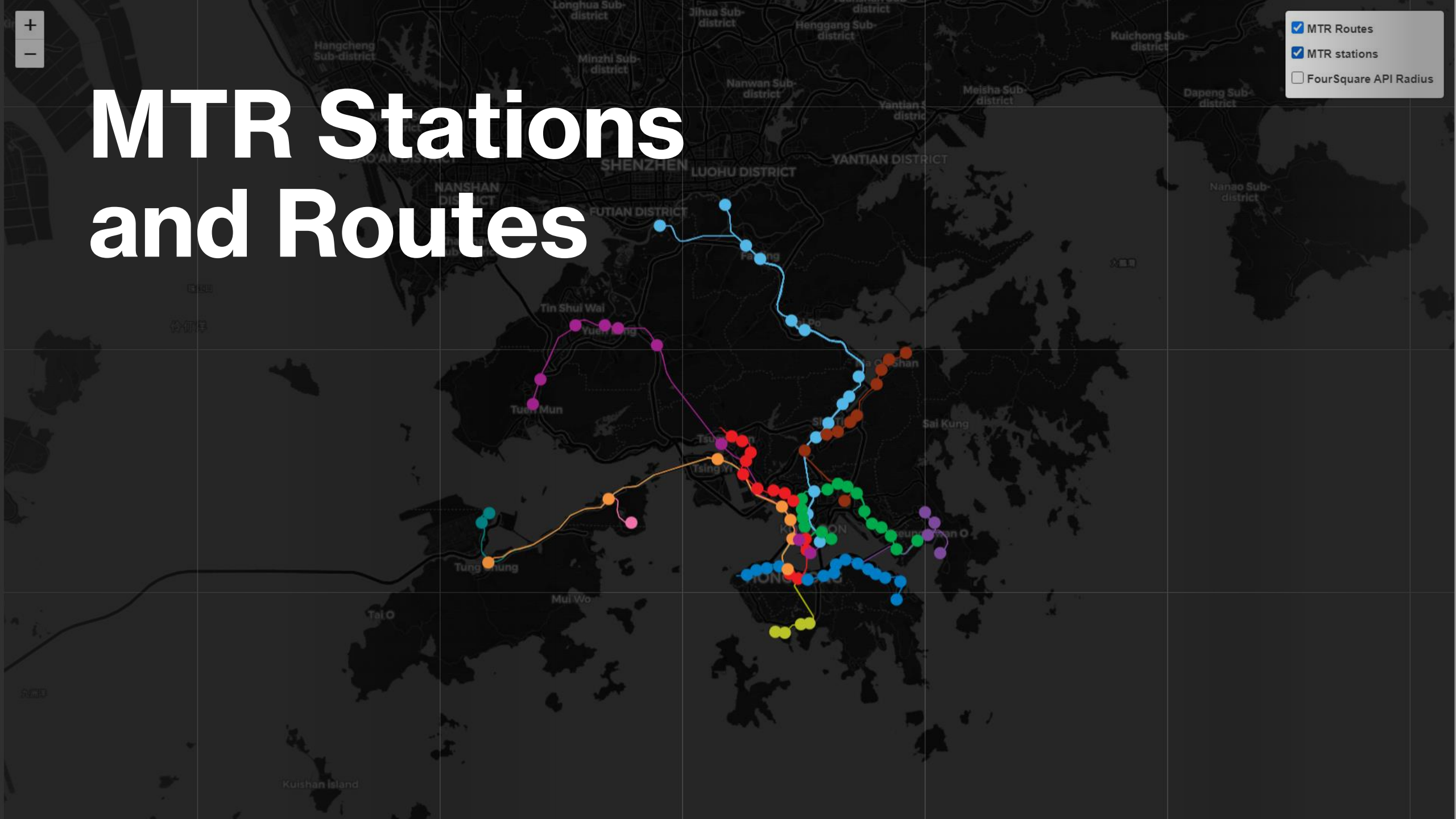
Introduction

- This is an exploration into how to cluster the Hong Kong MTR network using Foursquare data.
- This is interesting to those living or visiting Hong Kong, the MTR corporation and other data scientists.
- The result of the clustering will show which stations are most similar/dissimilar to each other.

Data

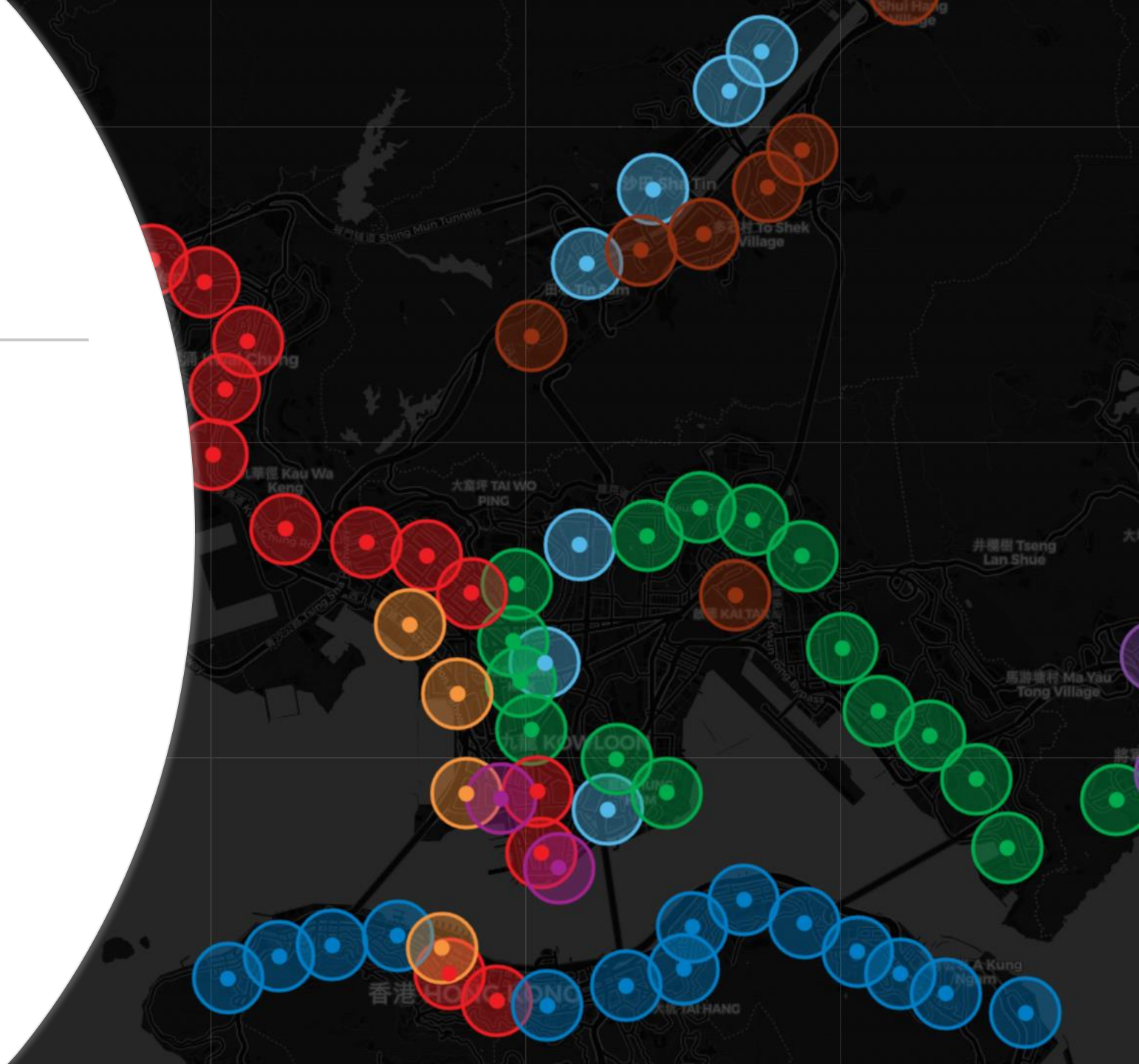
- Three sources of data was used:
  - Overpass Turbo for station coordinates and routes in Geojson format
  - Wikipedia for cross referencing and backup
  - Foursquare API to collect venue information for a given radius around each station.

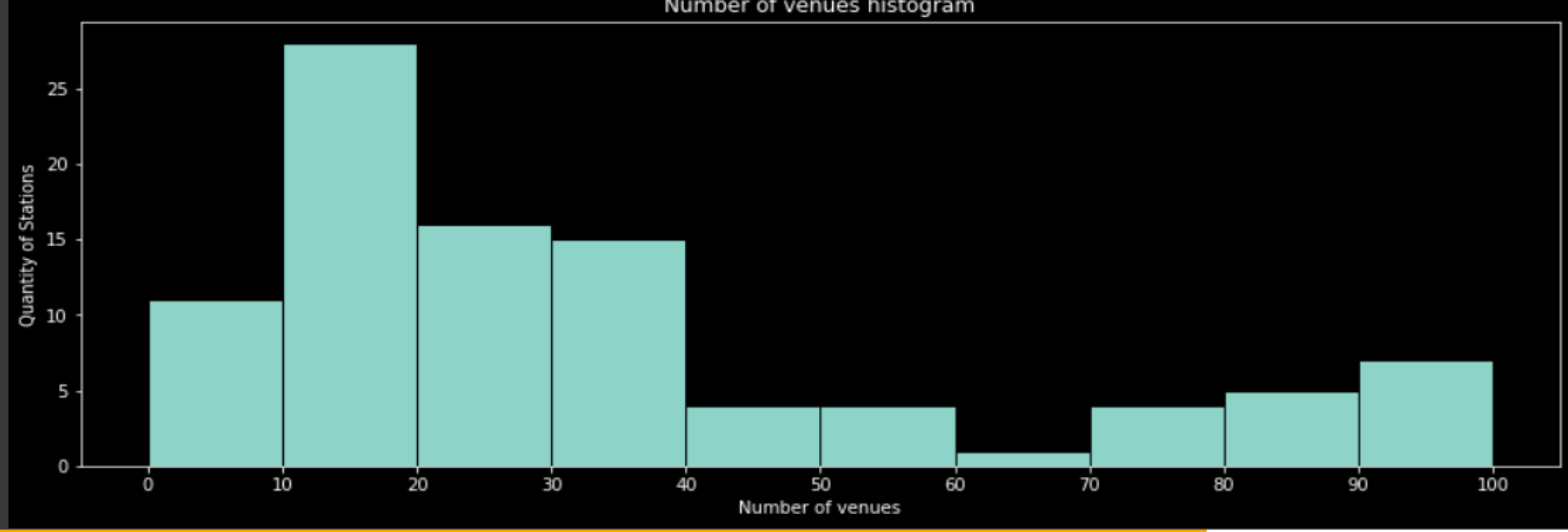
MTR Stations and Routes



Foursquare API Settings

- Explore venues surrounding coordinate
- 500m search radius
- 100 venues return limit
- Response in JSON format
- Zoomed example as shown





Foursquare Results Histogram

- Problem:
  - Some stations received less than 10 venues
- Solution:
  - Replace missing data with corresponding top 10 data of the whole dataset

Top 10 Venues

1. Chinese Restaurant
2. Coffee Shop
3. Café
4. Fast Food Restaurant
5. Japanese Restaurant
6. Hotel
7. Cha Chaan Teng
8. Cantonese Restaurant
9. Noodle House
10. Shopping Mall

Top 10 Result Table per Station

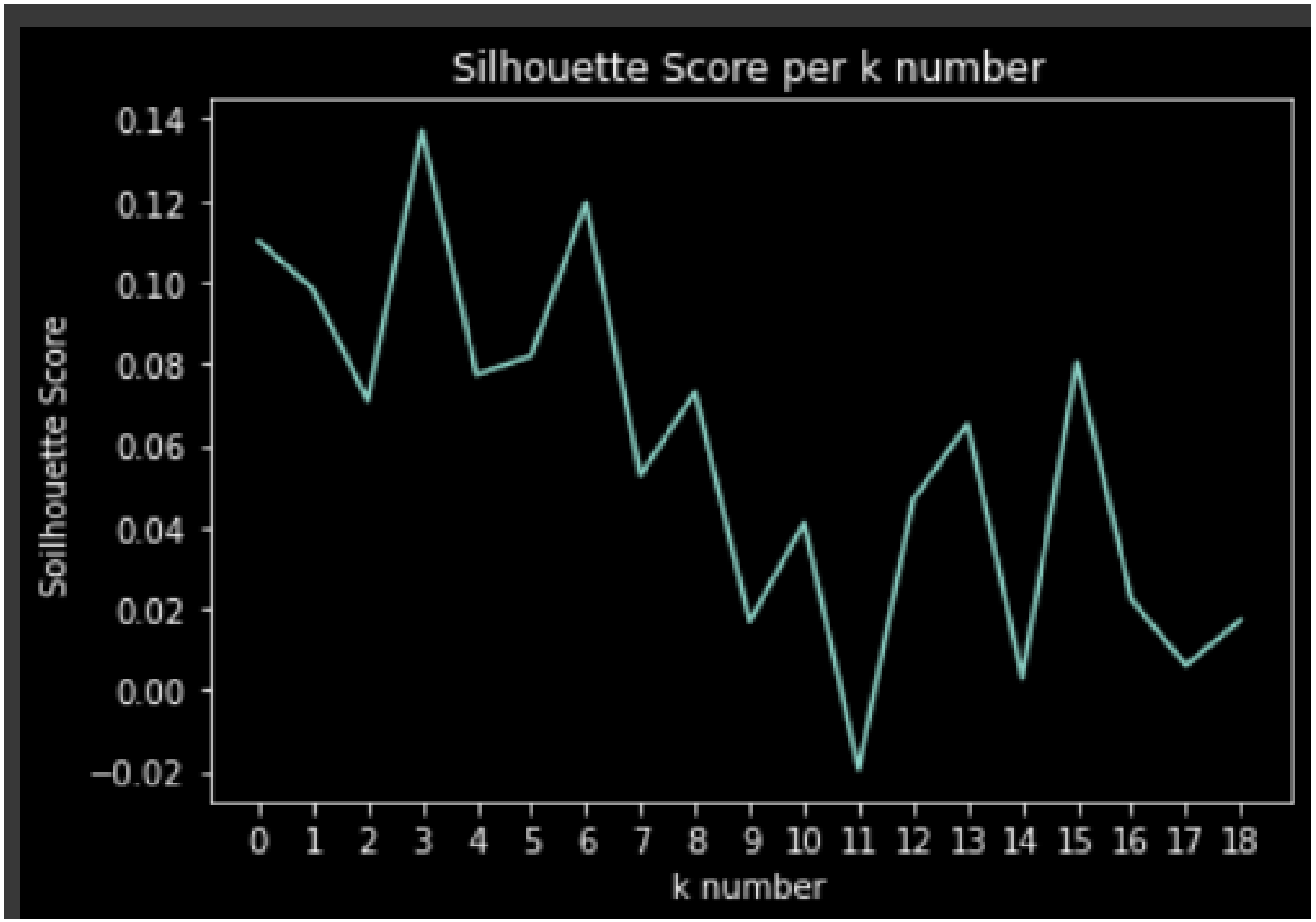
Station	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Admiralty	Café	Italian Restaurant	Hotel	Burger Joint	Clothing Store	Steakhouse	Bakery	Seafood Restaurant	Cantonese Restaurant	Tea Room
~	~	~	~	~	~	~	~	~	~	~

Methodology

- Two K means Clustering models were created:
  - K Best Model: based on best silhouette score over a range of K= 1 to 20
  - K Lines Model: based on number of MTR lines.

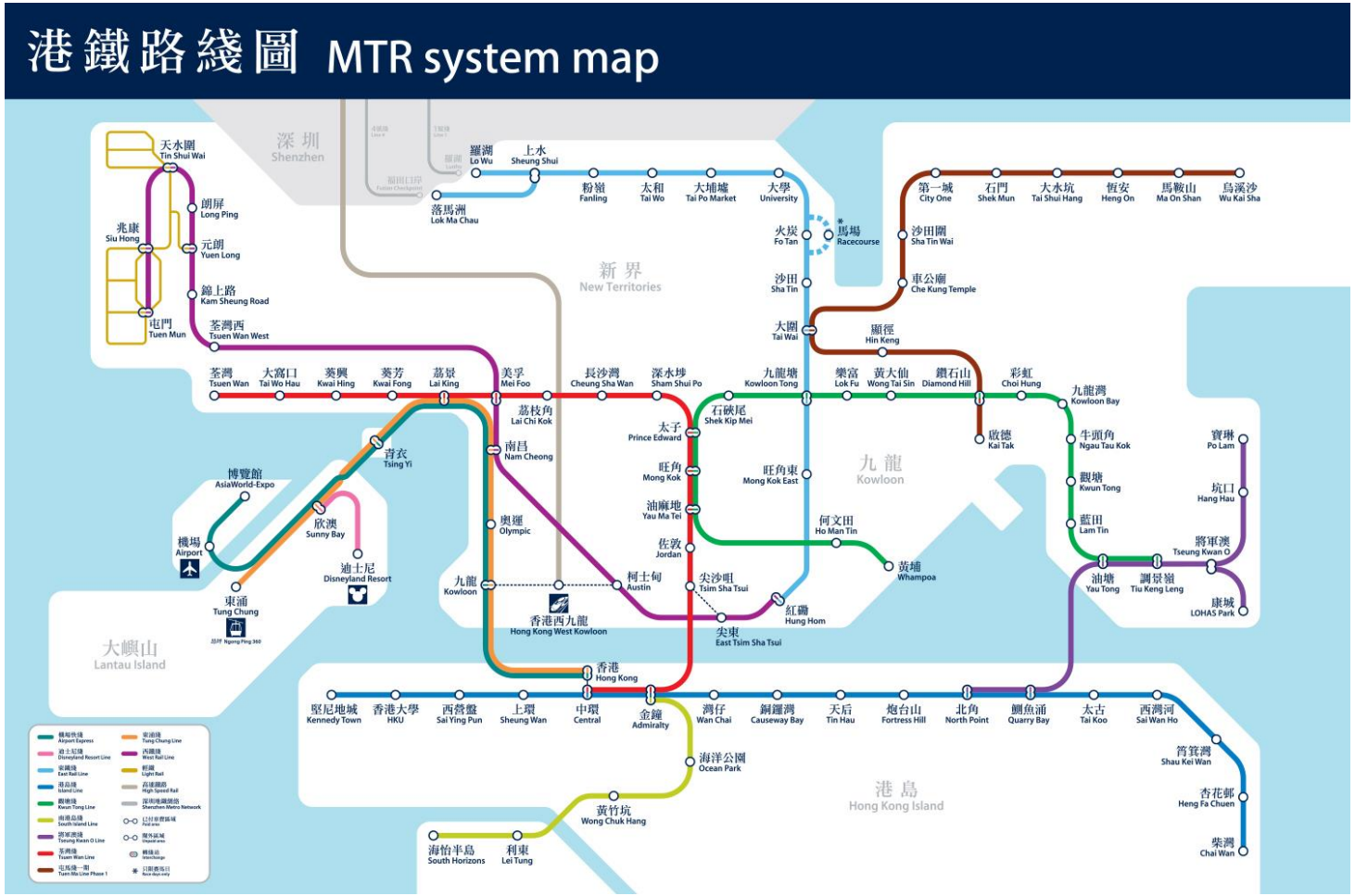
K Best Model

- K was chosen based on best silhouette score.
- In this case it was K=3

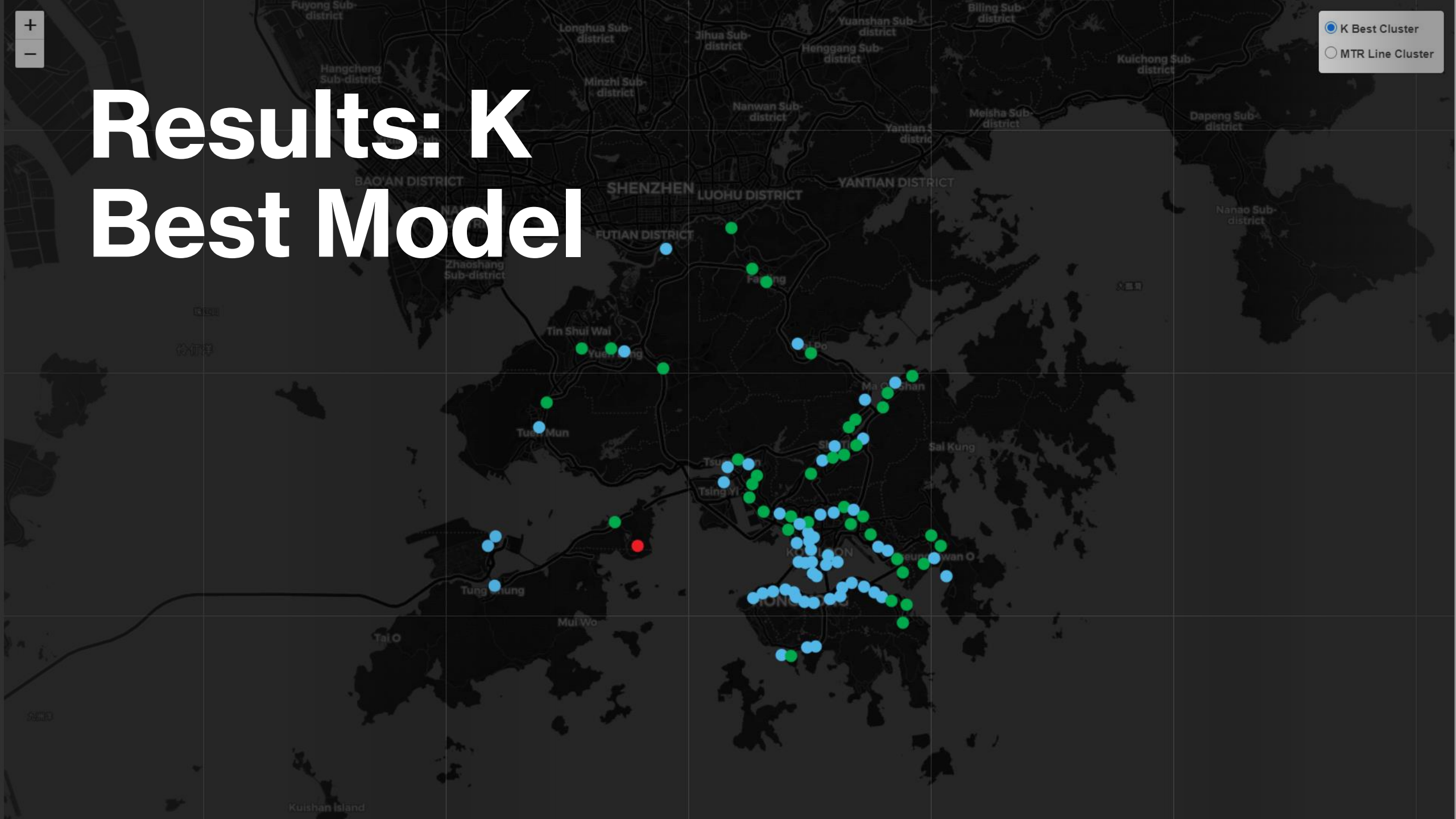


K Lines Model

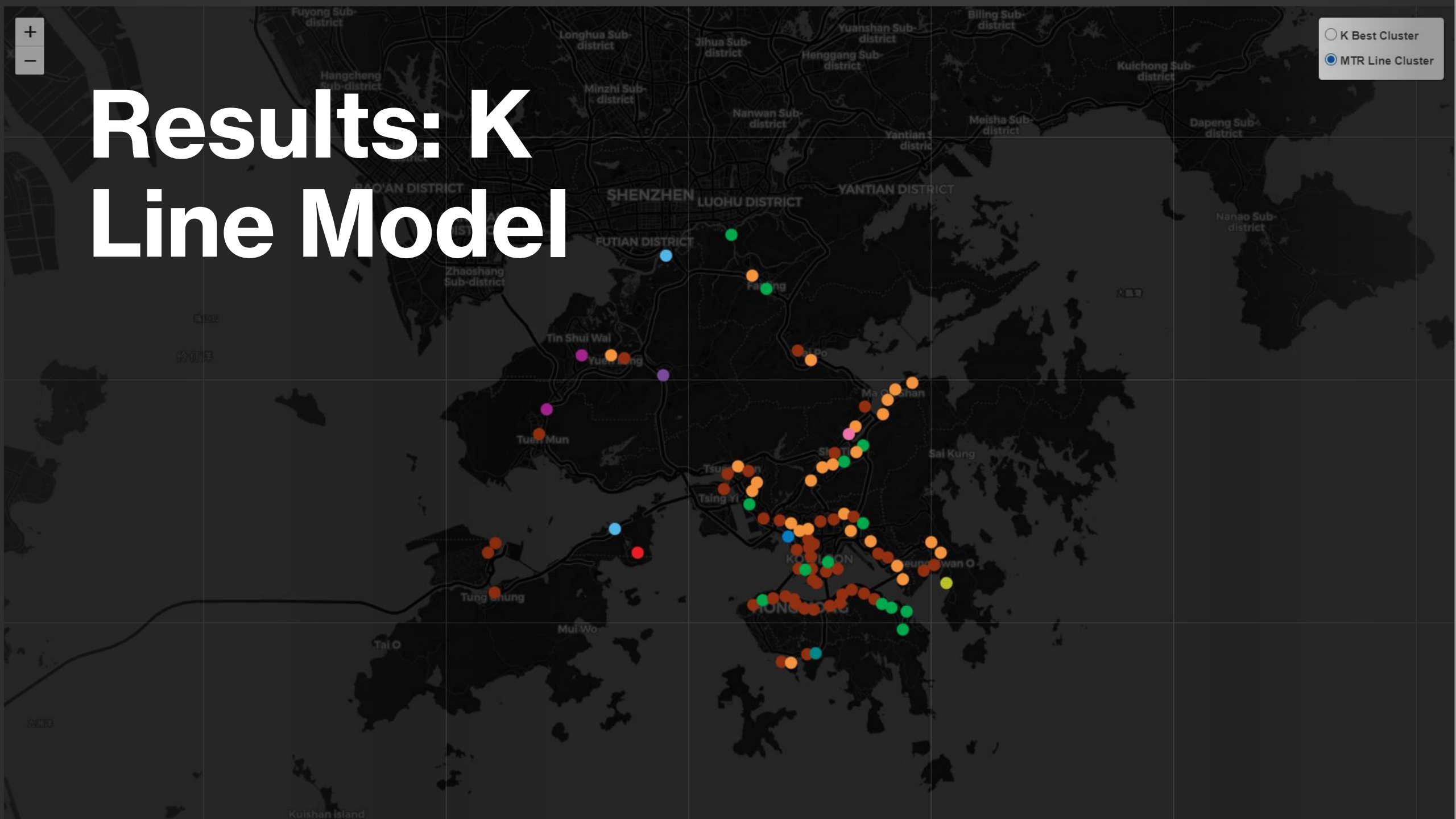
- Chosen based on the number of MTR Lines
- In this case K=11



Results: K Best Model



Results: K Line Model



Discussion: K Best Model

- Cluster 0 - Entertainment and some F&B venues. Represents the types of venues considered in "central" Hong Kong.
- Cluster 1 - Heavy F&B related venues. Represents venues outside of "central" Hong Kong.
- Cluster 2 - The Disneyland resort station. Different to other stations. However, interesting that another amusement park, Ocean Park, is not included.

Discussion: K Line Model

- Cluster 8:
  - Largest Cluster
  - Mostly "central" HK stations.
  - Almost evenly divded between F&B and shopping venues.
- Cluster 4:
  - Second largest cluster.
  - Top majority is "fast food" and other F&B venues.
- Cluster 1:
  - Third largest
  - Mostly comprised of "Chinese/Cantonese" restaurants.
  - Bottom venues mostly transportation and shopping venues.
- Other clusters:
  - Mostly outliers as they are vastly different from the others.
  - Due to the low silhouette score, some maybe "forced" into the wrong cluster due to how the model requires K=11.

Conclusion

- Two Clustering models created:
  - Based on K best, which was K=3
  - Based on # of MTR lines, which was K=11
- Many stations are similar, but most have an abundance of F&B venues.

Future Recommendations

- Additional data sources such as:
  - District population density,
  - surrounding property prices,
  - ethnic/nationality breakdown, etc.
- Alternative API such as the Google Map API
- Manually or automatic tweaking of search radius to minimize overlap and maximize number of results.
- Experimenting with larger top N venues to include in algorithm.

Thank You for Your Time!

Created for the IBM Data Science Professional Certifical Captstone Project