Finding the best area to open a Chinese Restaurant in Singapore

Data Science Capstone- Coursera

Presentor: Anmol Saraf

Background and Problem Statement

- Singapore is one of the most cosmopolitan cities in Asia
- However, Chinese population forms the majority in the city-state
- A large variety of cuisines can be found here
- The aim is to find the best location in the city to open a Chinese restaurant
- The idea is to maximize the profit

Data Acquisition and Cleaning

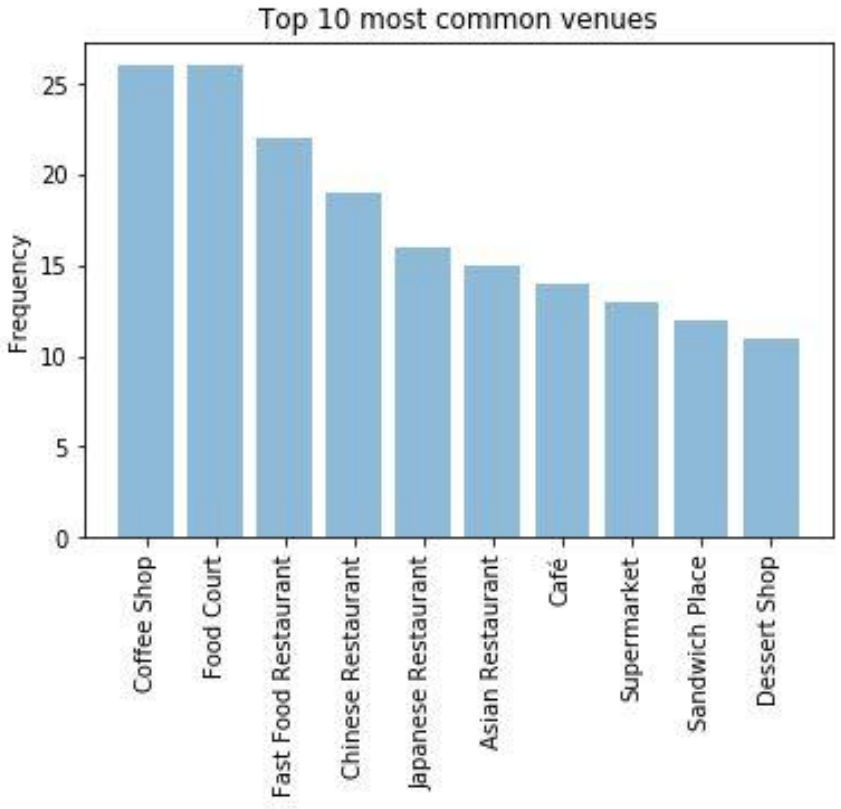
- Planning Areas of Singapore were obtained from the Wikipedia page
- There are 54 planning areas in the city
- Web-scraping was done to read the webpage data
- Using GeoPy client the geographical coordinates were obtained
- 10 most populous areas were shortlisted
- FourSquare API was used to obtain the venues details for the locations

Data Acquisition and Cleaning

	Area	Area Latitude	Area Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Bedok	1.323976	103.930216	Bedok Chwee Kueh 勿洛水粿	1.324903	103.930250	Chinese Restaurant
1	Bedok	1.323976	103.930216	Ya Kun Kaya Toast 亞坤	1.324095	103.929198	Coffee Shop
2	Bedok	1.323976	103.930216	Duke Bakery	1.324691	103.932514	Bakery
3	Bedok	1.323976	103.930216	FairPrice Fínest	1.324140	103.929260	Supermarket
4	Bedok	1.323976	103.930216	Song Zhou Luo Bo Gao 松洲箩卜糕	1.324836	103.930520	Breakfast Spot
5	Bedok	1.323976	103.930216	5 Senses	1.324993	103.932517	French Restaurant
6	Bedok	1.323976	103.930216	Dian Xiao Er 店小二	1.325118	103.930119	Chinese Restaurant
7	Bedok	1.323976	103.930216	Greendot	1.324506	103.930166	Vegetarian / Vegan Restaurant
8	Bedok	1.323976	103.930216	Din Tai Fung 鼎泰豐	1.324475	103.930164	Dumpling Restaurant
9	Bedok	1.323976	103.930216	Hai Di Lao 海底捞	1.324152	103.929106	Hotpot Restaurant

Sample rows from the data frame after collecting venues data

Data Acquisition and Cleaning



- *10 most visited venues in all the locations combined*

Data Manipulation

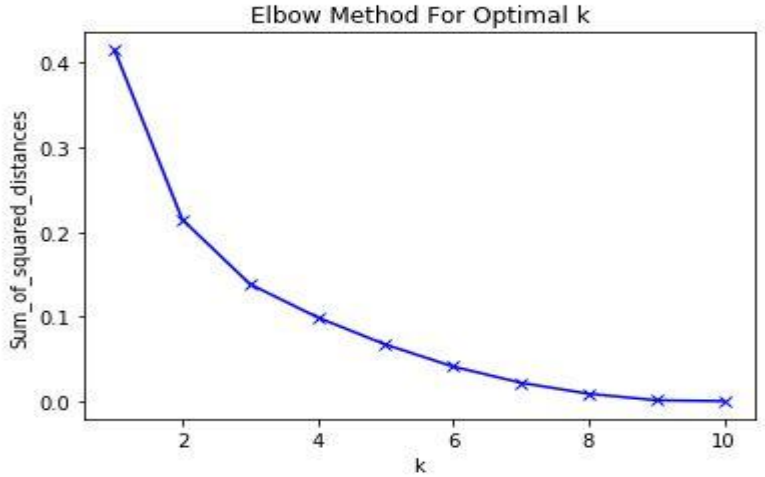
- One-hot encoding was used on the data set
- Categorical values were replaced with numbers

	Area	American Restaurant	Arcade	Asian Restaurant	Athletics & Sports	BBQ Joint	Bakery	Bank	Bookstore	Boutique
0	Ang Mo Kio	0.000000	0.000000	0.021277	0.000000	0.000000	0.021277	0.021277	0.000000	0.000000
1	Bedok	0.017544	0.000000	0.035088	0.000000	0.000000	0.017544	0.000000	0.017544	0.000000
2	Choa Chu Kang	0.000000	0.000000	0.083333	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	Hougang	0.000000	0.000000	0.043478	0.043478	0.043478	0.000000	0.000000	0.000000	0.000000
4	Jurong West	0.015873	0.000000	0.095238	0.000000	0.015873	0.000000	0.000000	0.015873	0.015873
5	Punggol	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000

Sample rows from the data set

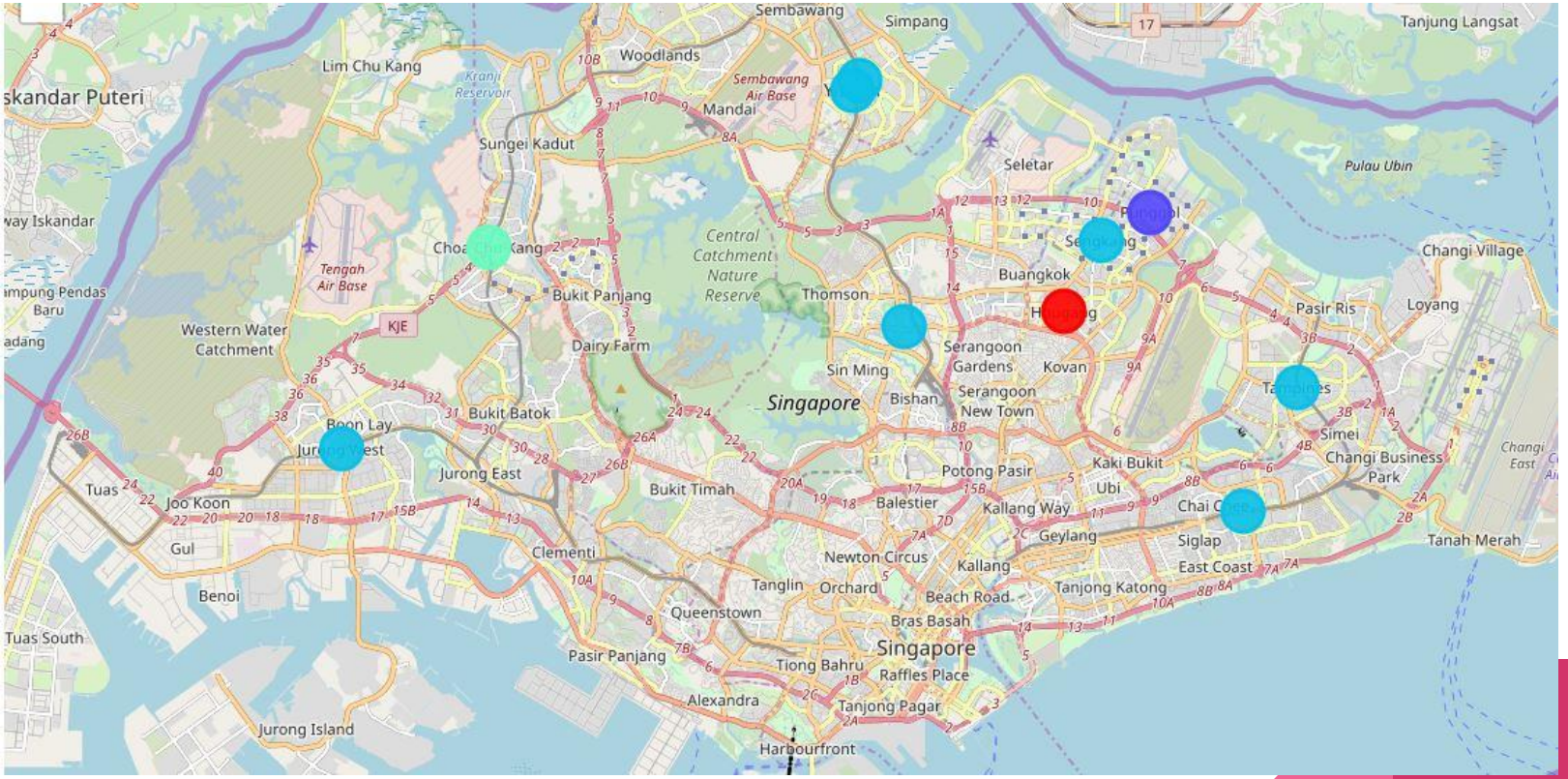
Exploratory Data-Analysis

- Top 5 most visited venues in each location were determined
- K-Means Clustering was used to group together similar locations
- Optimum number of clusters was determined using Elbow Plot



- 4 clusters were created

Results and Discussion



Clusters displayed on Singapore Map

Results and Discussion

- Coffee Shop is the most visited venue
- Chinese Restaurant is the 4th most visited venue.
- *Bedok* and *Ang Mo Kio*
  - are both Cluster 2 areas
  - don't have Chinese Restaurant as one of the top 5 most visited venues
- *Bedok* being the most populous area of all, has eateries as the most common venues, but Chinese Restaurant is not one among them

Results and Discussion

- **Bedok** would be highly profitable because
  - it attracts the most number of customers
  - eating joints are the most visited venues here
- **Jurong West** could also be considered because
  - it is the second most populous area
  - belongs to the same cluster as Bedok
  - fewer Chinese Restaurant are here currently--> less competition

Conclusion

- Data Analysis was used to arrive at the conclusion
- K-Means clustering was the chosen analysis mode
- Model could be improved with more data like the rentals
- Insightful starting point for beginners and explorers