What is the relative risk of contracting COVID-19 at venues within different neighborhoods in San Francisco, CA?

IBM Professional Data Science Certification: Capstone Project

By Benjamin Cruz

Problem Introduction

- **Problem:** when utilizing location-based apps such as Google Maps, Lyft, or Uber, the app user does not know the relative risk of contracting COVID-19 in the restaurant's neighborhood they wish to eat at.
- **Potential Solution**: Cluster the neighborhoods by venue category and assign a relative risk of contracting COVID-19 score to each cluster. Now, the user can see which neighborhoods have the venues they to visit and the neighborhoods relative risk of contracting COVID-19 score.
- **Business Interest**: Applications that would like to inform their customers of potential health risks due to COVID-19 when travelling to different venues.

Project Goal

- Create prototype code that defines the methodology for creating neighborhood clusters by venue categories commonly found and defining these clusters' relative risk of contracting COVID-19 scores.
- Use San Francisco, CA venue and COVID-19 data for prototype.

Data Sources

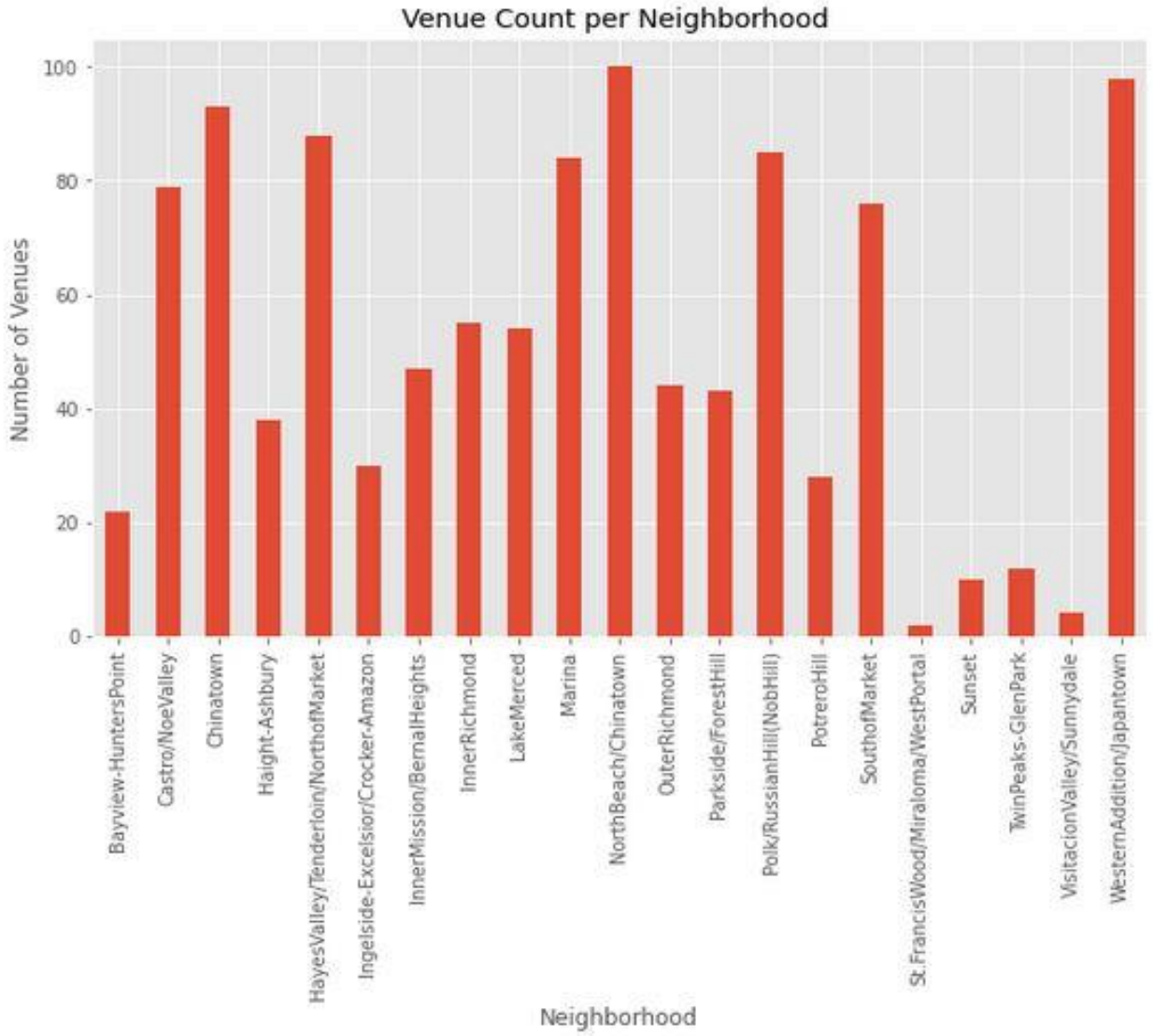
- **Foursquare Venue Data**:
  - Accessed using Foursquare API, getting venues for San Francisco.
- **OpenDataSoft Data for California**:
  - Openly available data accessed here: https://public.opendatasoft.com/explore/dataset/us-zip-code-latitude-and-longitude/table/?refine.state=CA.
- **City of San Francisco COVID-19 Cases by Geography and Date:**
  - Openly available data Accessed here: https://data.sfgov.org/COVID-19/COVID-19-Cases-by-Geography-and-Date/d2ef-idww.
- **San Francisco Department of Public Health, Burden of Disease and Injury Study:**
  - Openly available data Accessed here: http://www.healthysf.org/bdi/outcomes/zipmap.htm .

Data Cleaning

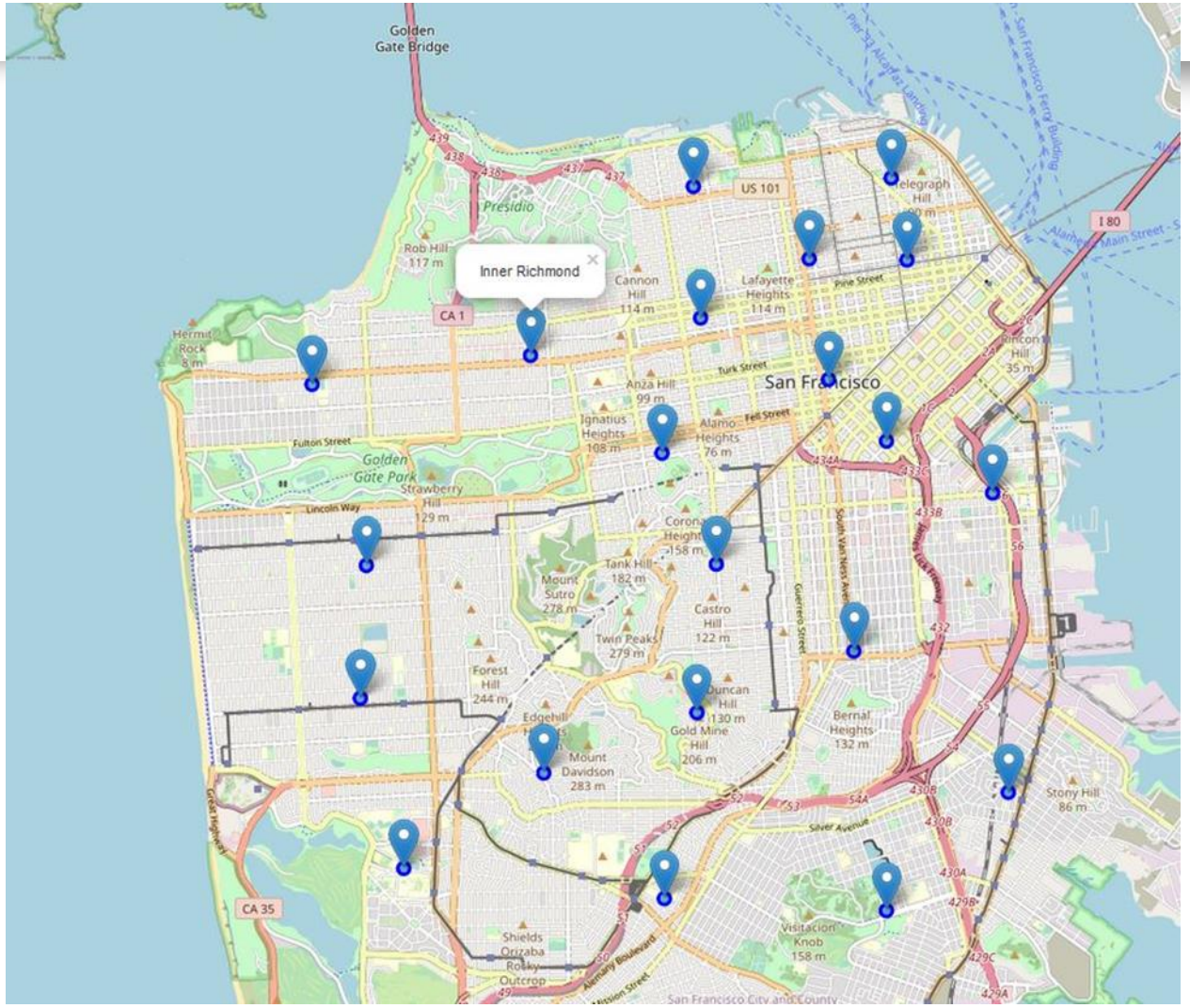
- **Foursquare Venue Data**:
  - We requested the neighborhood name, latitude, longitude, venue name, latitude, longitude, and category from Foursquare's data warehouse. We then grouped the venues by neighborhood.
- **OpenDataSoft Data for California**:
  - Imported excel file of zip-code, longitude, and latitude information into a dataframe and dropped the "State," "Timezone," "Daylight savings time flag," and "geopoint" columns from the dataframe.
- **City of San Francisco COVID-19 Cases by Geography and Date:**
  - Imported excel file of zip-code and Average New COVID-19 Cases per Day into a dataframe.
- **San Francisco Department of Public Health, Burden of Disease and Injury Study:**
  - Got the San Francisco Department of Public Health data of neighborhood name and zip codes using the Beautiful Soup python package to extract the HTML table within the web page to a dataframe (full code link here).

Venue Count Per San Francisco Neighborhood

- Almost all neighborhoods have more than 20 venues.
- Eight out of 20 neighborhoods have more than 60 venues.



San Francisco Neighborhoods Plotted on San Francisco

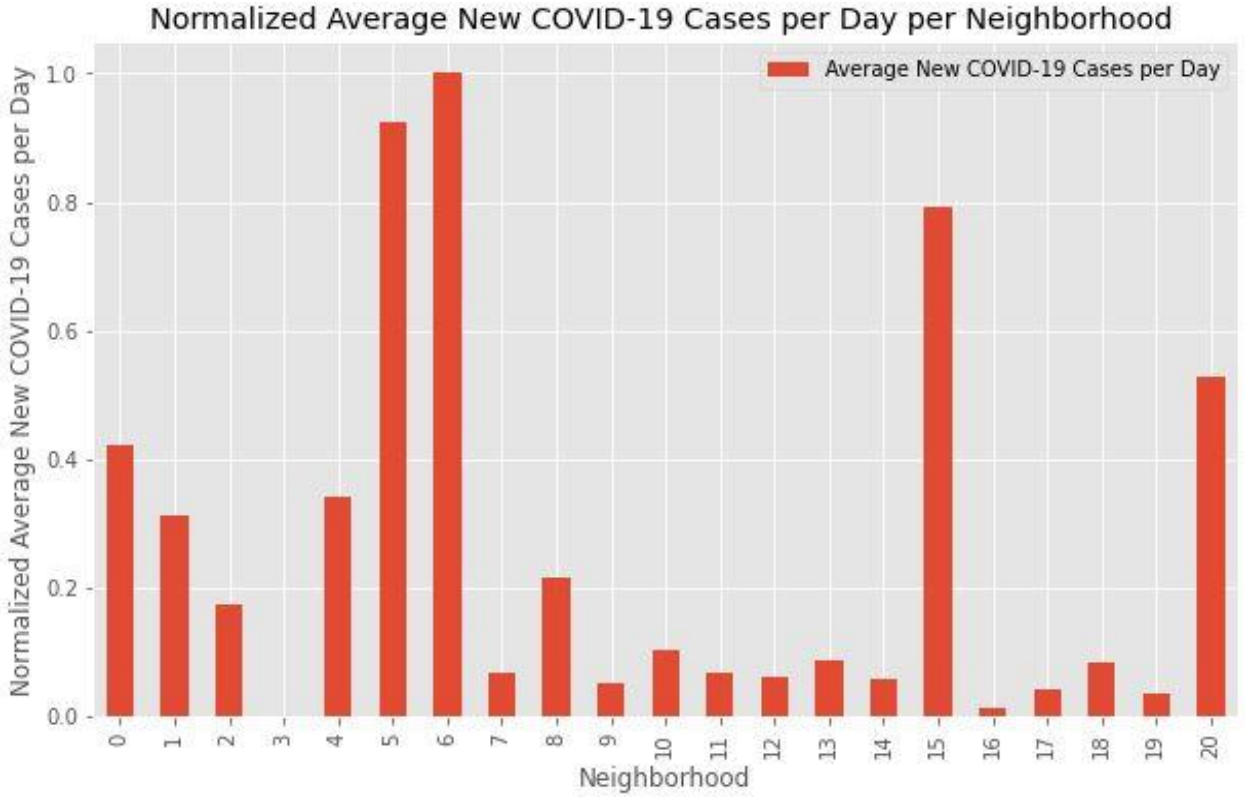


Methodology

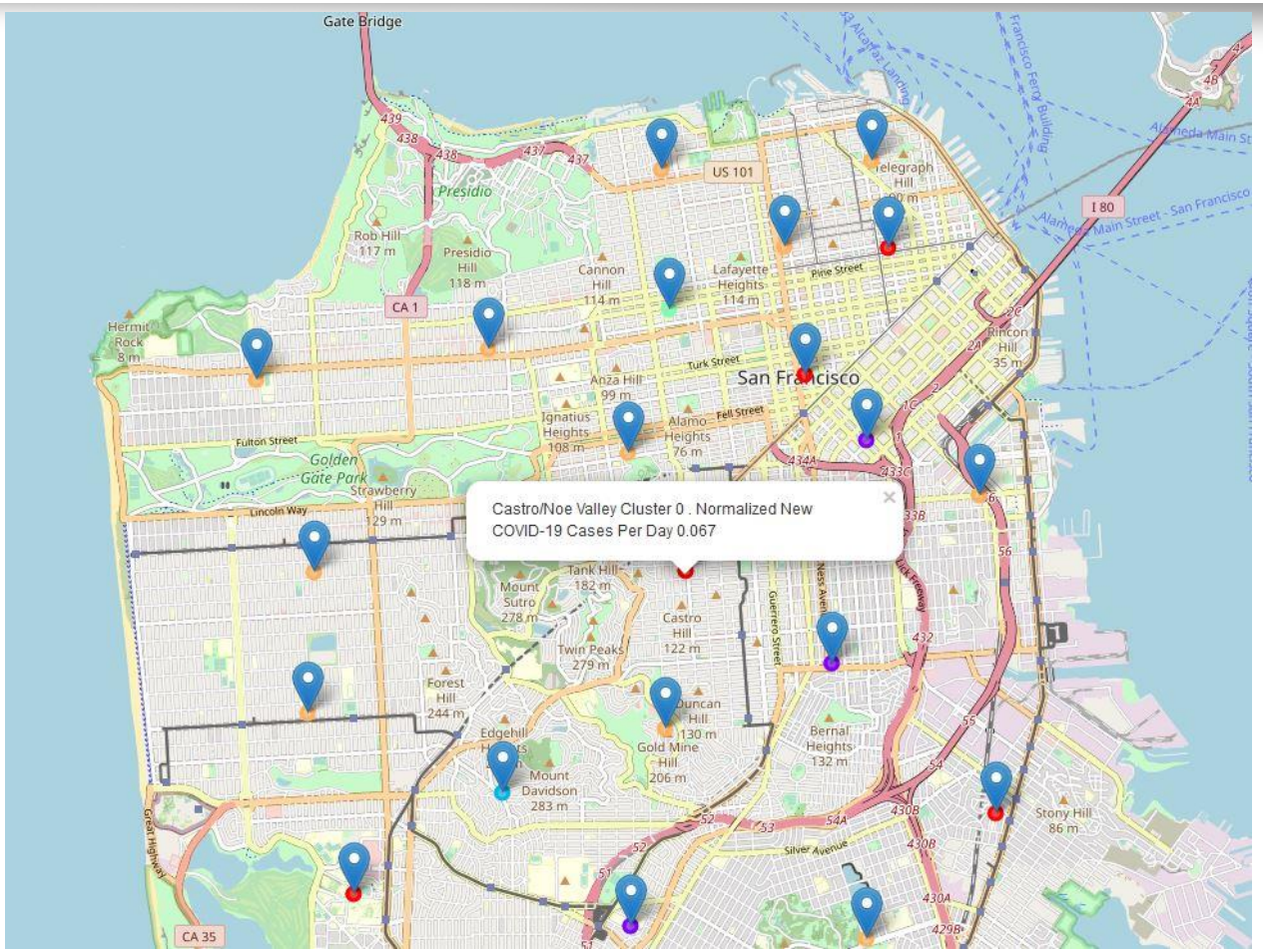
- Obtained all venues categories/frequencies and grouped them by neighborhood.
- I then appended the average new COVID-19 cases per day data to this venue dataframe.
- Our next methodology task was to perform k-means clustering to cluster the neighborhoods into five clusters using the venue and COVID-19 dataframe.
- We appended the cluster number for each neighborhood to the dataframe and used this dataframe to plot the clusters on a map of San Francisco.
- We then analyzed each of the five clusters and classified them based on their three most common venue types.
- We also attached the relative risk of contracting COVID-19 score to each neighborhood cluster by taking the mean of all the risk scores for neighborhoods within a cluster. We passed the score to a function that determined if the risk was “Low” (risk<0.333), “Medium” (0.333> risk <0.667), or “High” (risk >0.667).

Normalized Average New COVID-19 Cases per Day San Francisco Neighborhood

- It appears neighborhoods 5 (Ingleside-Exelsior/Crocker-Amazon), 6 (Inner Mission/Bernal Heights), and 15 (South of Market) represent the neighborhoods with the highest relative risk of contracting COVID-19.
- Haight-Ashbury represents by far the safest neighborhood. However things can change over time, so we would need to continually update the data.



San Francisco Clustered Neighborhoods Plotted on San Francisco



Results

- Taking a look at the results, it is clear that the relative risk of contracting COVID-19 in most venue neighborhoods is low. .
- However, one could argue to avoid venues in cluster 1, the "Pizza Place, Mexican Restaurant, and Nightclub/Bar" due to its relatively high COVID-19 Risk Score.

Table 1: Summary of Results

Cluster Number/Color	Cluster Classification	COVID-19 Risk Score
Cluster 0/Red	Coffee Shop, Café, and Hotel	Low
Cluster 1/Purple	Pizza Place, Mexican Restaurant, and Nightclub/Bar	High
Cluster 2/Light Blue	Bus Line, Trail, and Yoga Studio	Low
Cluster 3/Light Green	Bakery, Spa, and Café	Low
Cluster 4/Orange	Chinese Restaurant, Grocery Store, and Grocery Store	Low

Conclusion

- This project addresses this problem by providing the user relative COVID-19 Risk Scores for different neighborhood clusters that contain different venues.
- For example, a user may want to get food from a pizza place and try to go to a restaurant in venue cluster 1.
- However, after seeing that the relative COVID-19 Risk Score is High, they may choose to order delivery from that restaurant instead of going in person.
- An application such as this helps inform consumer decisions about going to a restaurant or shopping in the age of COVID-19.