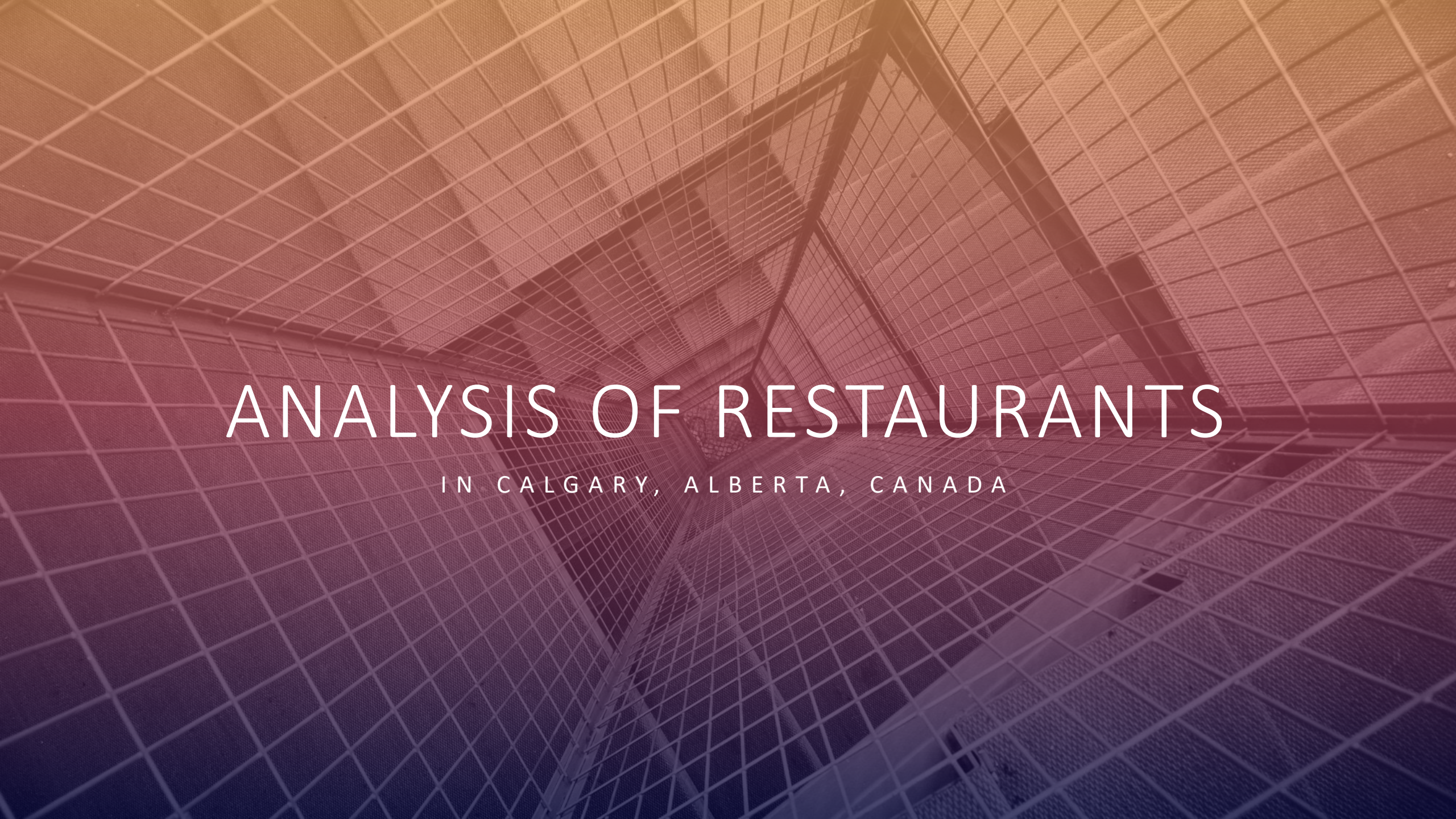


ANALYSIS OF RESTAURANTS

IN CALGARY, ALBERTA, CANADA

INTRODUCTION

CITY OF CALGARY

BACKGROUND INFORMATION

- Located in Alberta, Canada.
- Near the rocky mountains.
- Population of 1.2 million.
- 4th largest city in Canada.
- 198 neighborhoods.
- Culturally diverse.

THE PROBLEM

Based on factors such as median household income and population density, can we explain the quantity and category of restaurants in each neighborhood?

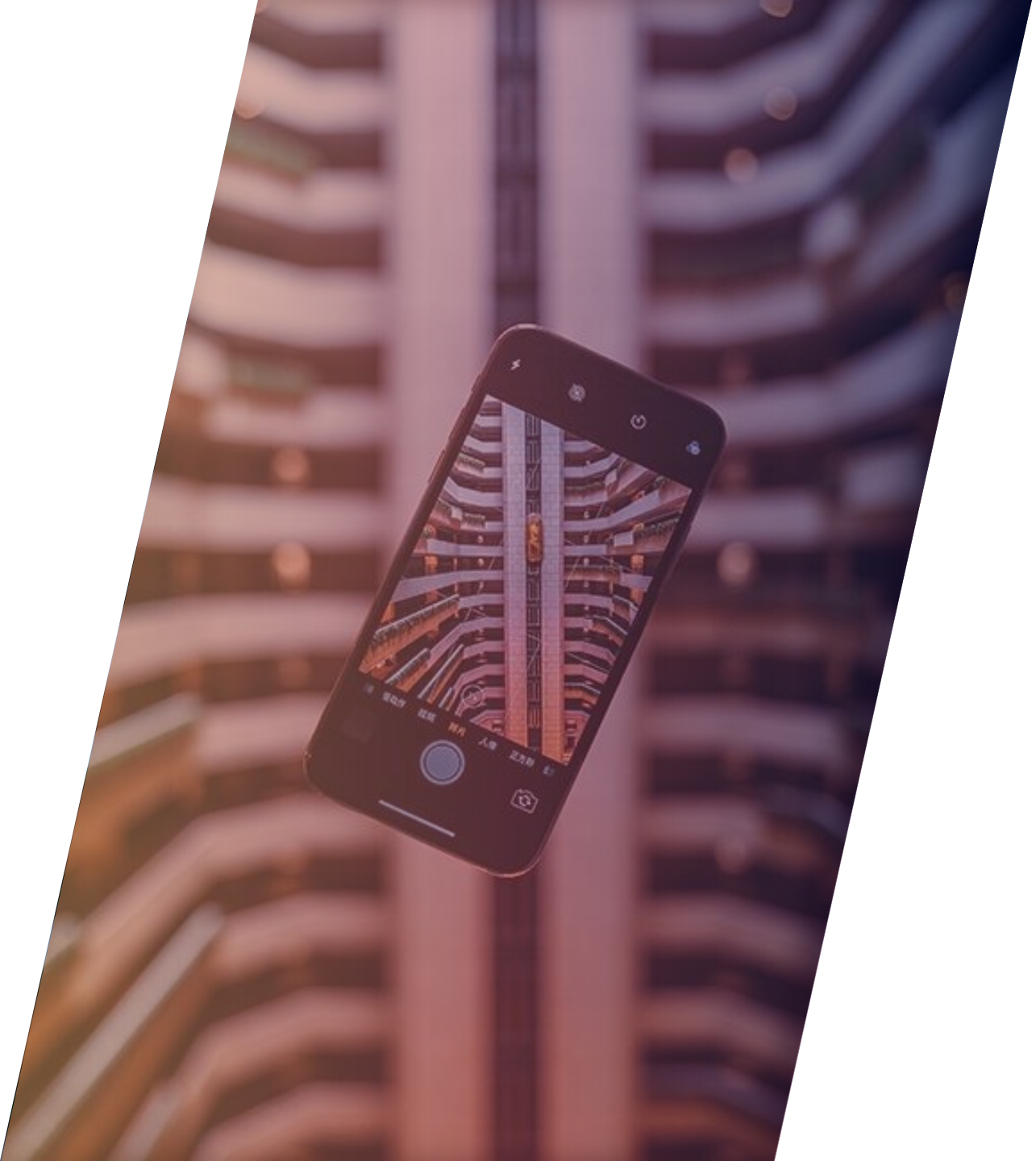
RELEVANCE

- This information may be useful for those who are interested in opening a restaurant in Calgary.
- It may also be interesting for those who are thinking of moving to the city.

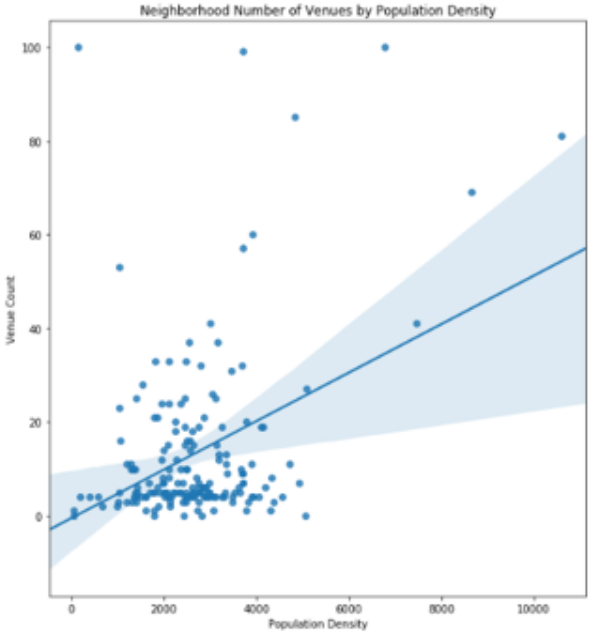
DATA

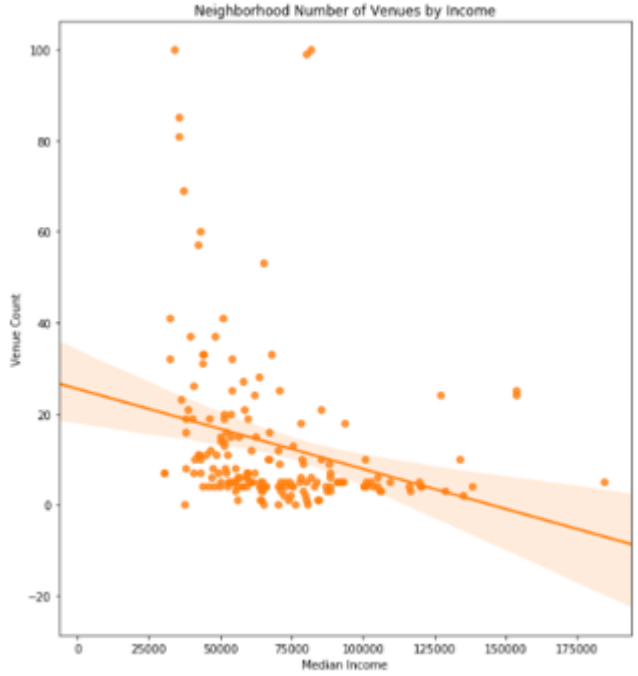
DATA SOURCES

- Neighborhood Data
  - Collected from scraping HTML source from Wikipedia page (https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary)
  - Includes information such as Population, Area (km2), Population Density, etc.
- Neighborhood Demographics Data
  - Collected from scraping HTML source from webpage (https://great-news.ca/demographics/)
  - Includes information such as Median Household Income, Median Age, Median Home Sale Price, etc.
- Coordinate Data
  - Collected from Geocoder library in Python.
  - Used for plotting neighborhoods on folium map and for retrieving venue information from Foursquare.
- Location Data
  - Collected from Foursquare to retrieve venue information on nearby restaurants in each neighborhood.



METHODOLOGY & ANALYSIS





CORRELATION

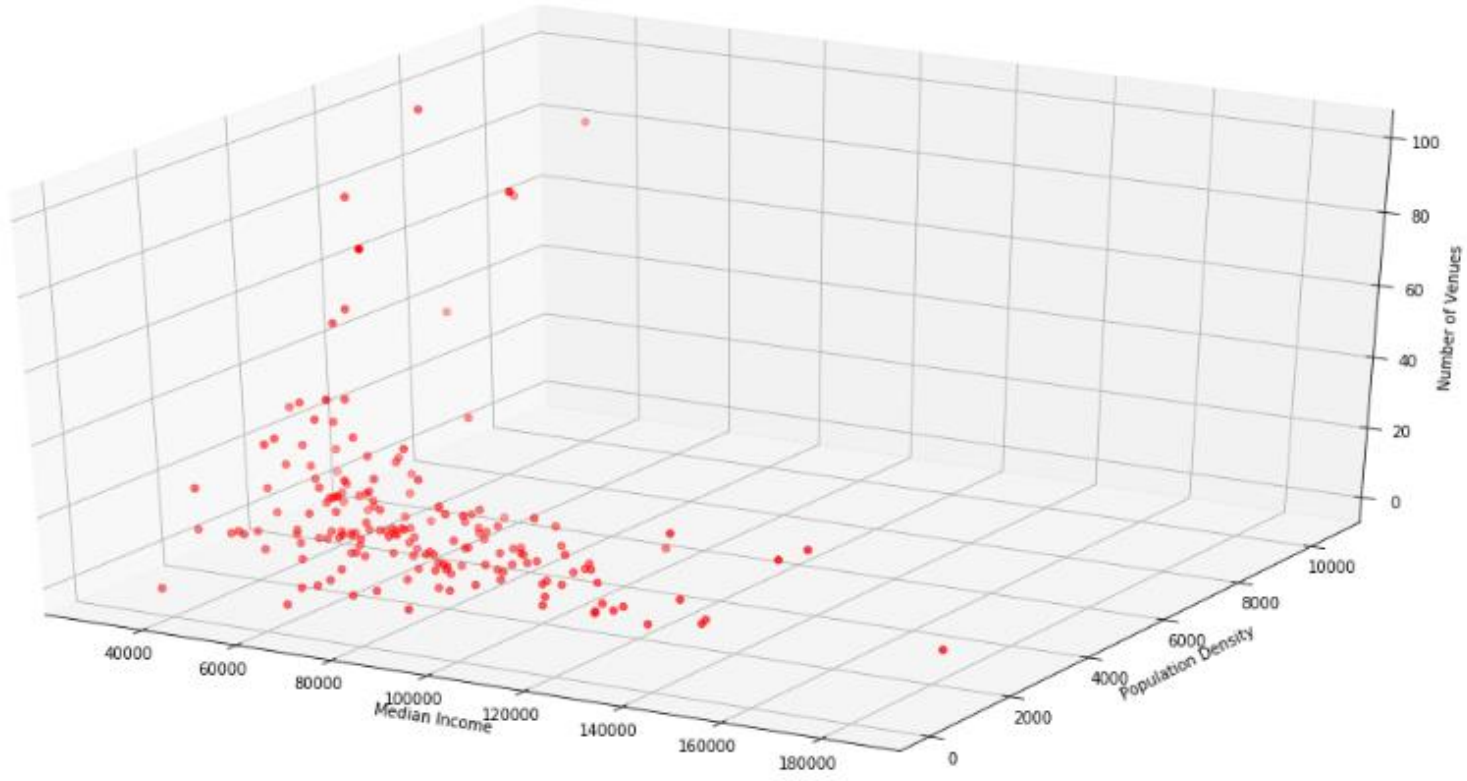
REGRESSION PLOTS

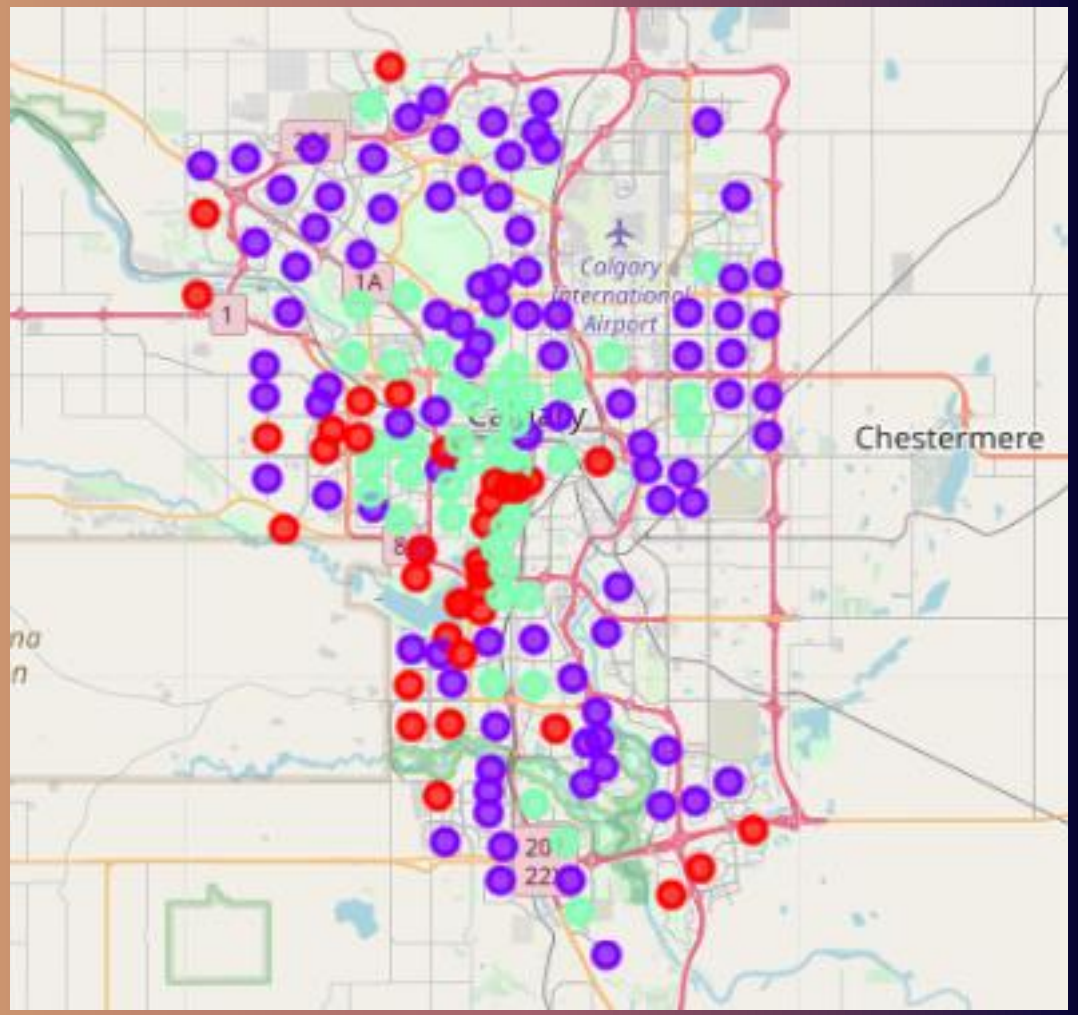
- Plotted population density (blue) and median income (orange) against the number of restaurants in each neighborhood.
- Population density has positive correlation.
- Median income has negative correlation.
- Weak correlation for both variables.
- Pearson coefficient of approximately 0.4 and -0.3 respectively.
- Linear model may not be suitable.

CORRELATION

3D PLOT

- 3D plot of median income and population density against number of restaurants.
- Neighborhoods with highest number of restaurants had lowest income and highest population density.
- Suggests a correlation between the three variables.





CLUSTERING

FIRST ATTEMPT

- Clustered neighborhoods using a K Means model (n clusters = 3).
- Features include:
  - Median Income.
  - Population Density.
  - 5 Most Common Venue Categories.
- Scaled data using Min Max scaler.
- Red = Cluster 0, Purple = Cluster 1, Green = Cluster 2.

CLUSTERING

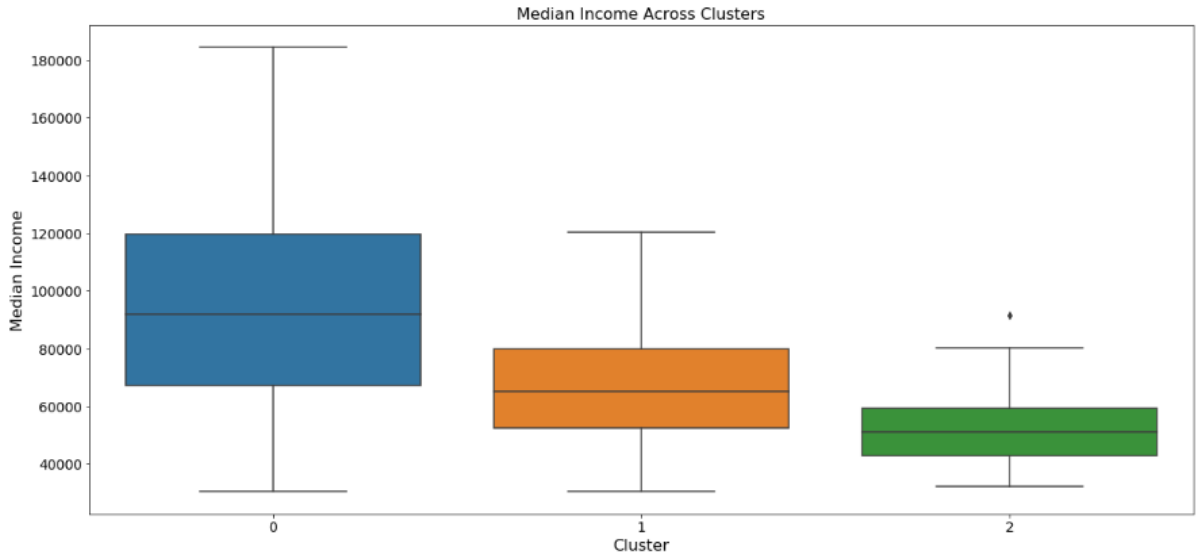
RESULTS: FIRST ATTEMPT

- Results were somewhat evenly distributed across city. Did not significantly differentiate restaurant categories between clusters.
- Cluster 0 had high median income and low population density, cluster 1 had medium income and medium population density, cluster 0 had low income and medium population density.

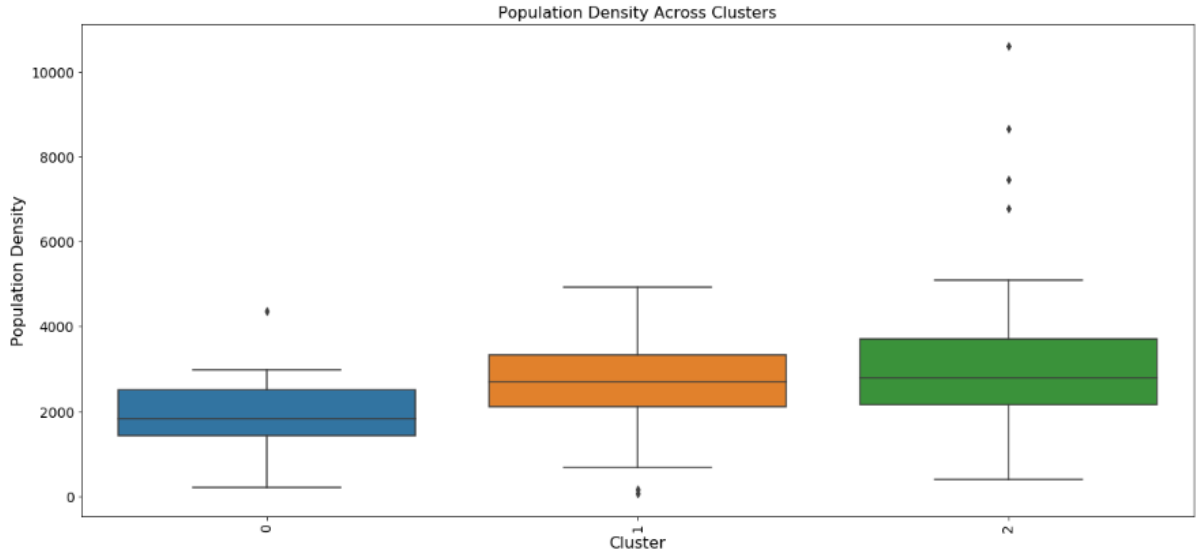
Cluster	# of Neighborhoods	Avg Income	Avg Pop Density	Most Common
Cluster 0	36	$96,000	1951	Café
Cluster 1	92	$68,000	2686	Pizza Place
Cluster 2	45	$51,000	3229	Pizza Place

CLUSTERING

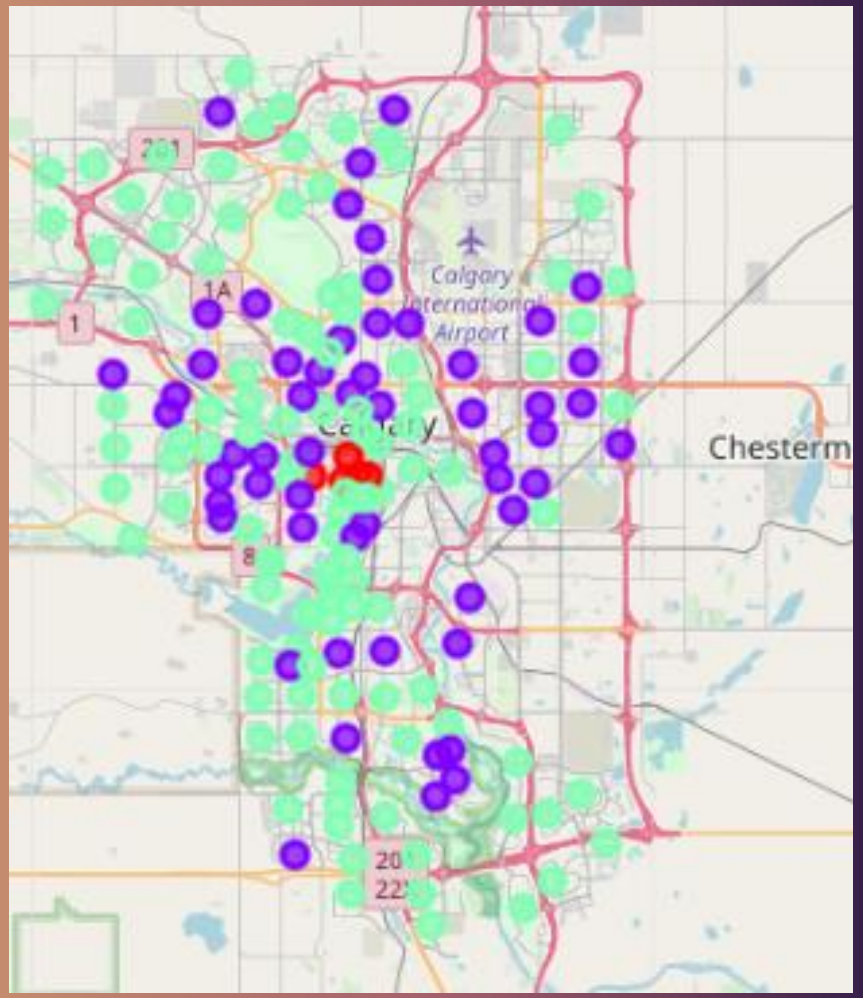
BOX PLOTS: FIRST ATTEMPT



Median income across clusters 0, 1, and 2.



Population density across clusters 0, 1, and 2.



CLUSTERING

SECOND ATTEMPT

- Removed venue categories: pizza place and restaurant.
  - Pizza place is most common venue in city. Causing results to be biased.
  - Restaurant is too general as a category.
- Data scaled using standard scaler.
- Ran model again with new data.
- Red = Cluster 0, Purple = Cluster 1, Green = Cluster 2.

CLUSTERING

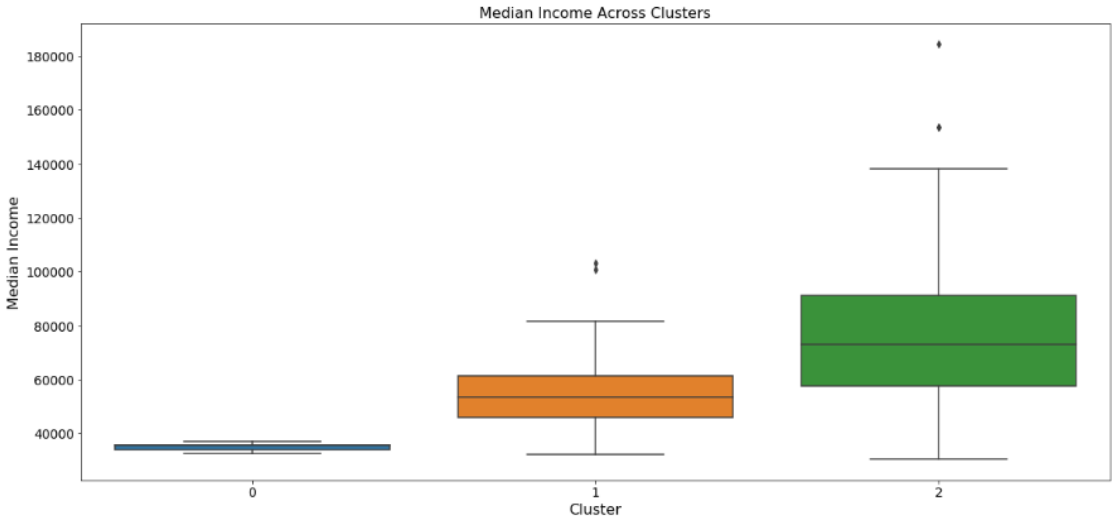
RESULTS: SECOND ATTEMPT

- Better results yielded on second try.
- Cluster 0 had low median income and high population density, cluster 1 had medium income and medium population density, cluster 2 had high income and low population density. Significant difference between clusters.

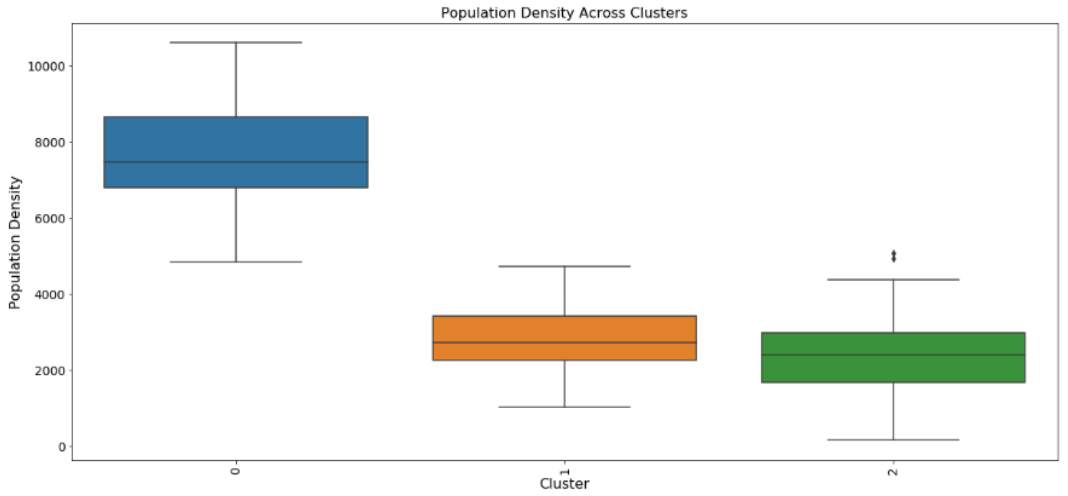
Cluster	# of Neighborhoods	Avg Income	Avg Pop Density	Most Common
Cluster 0	5	$35,000	7667	Vietnamese Restaurant
Cluster 1	56	$56,000	2821	Fast Food Restaurant
Cluster 2	110	$77,850	2393	Sandwich Place

CLUSTERING

BOX PLOTS: SECOND ATTEMPT



Median income across clusters 0, 1, and 2.

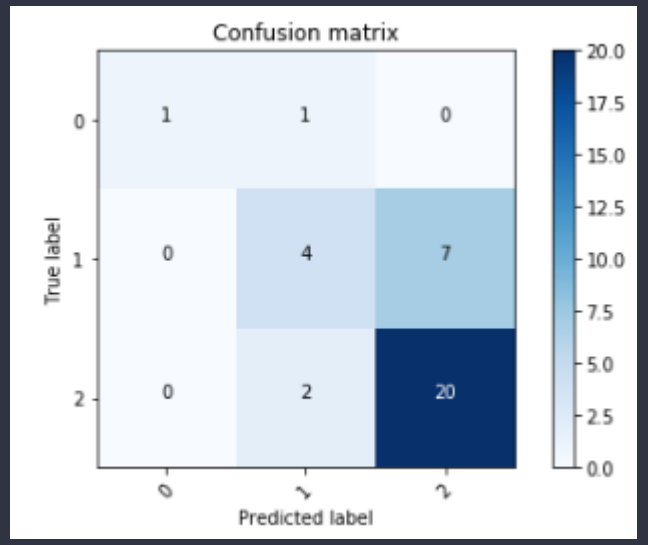


Population density across clusters 0, 1, and 2.

CLASSIFICATION

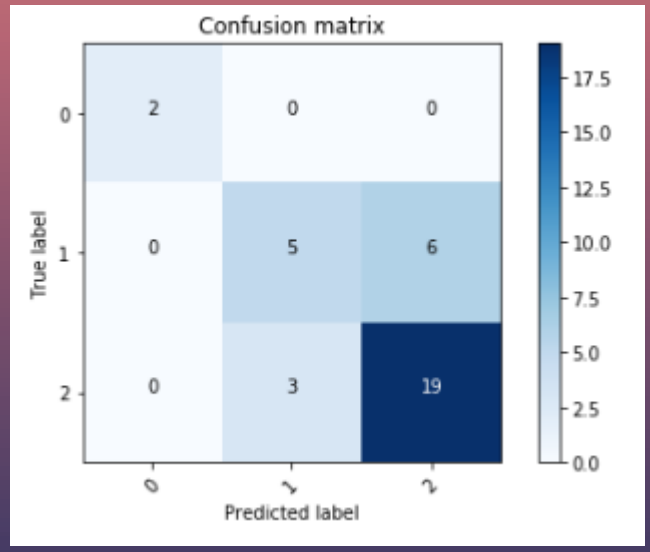
SUPPORT VECTOR MACHINES

- Used a classification model to test the clusters made with the K means algorithm. Tried SVM model first with linear kernel.
- Overall cross validation accuracy score of 68%.



K NEAREST NEIGHBORS

- K = 3 yielded best results.
- Overall cross validation accuracy score of 69%.
- Very similar results to SVM model.



RESULTS

RESULTS

- Cluster 0 has the lowest income and highest population density. This is because they are all neighborhoods close to downtown and consist mainly of apartments and condos. The mean age of the neighborhoods is quite young (31) and these neighborhoods largely consist of rental properties so it makes sense that it would have low income.
- Cluster 1 has medium income and medium population density. The neighborhoods are spread evenly across the city. The average age is 38. The cluster is mainly made up of the older inner-city neighborhoods but also contains some neighborhoods on the outskirts of the city as well.
- Cluster 2 has high income and low population density. The neighborhoods are also spread across the city. The average age is 39. The cluster is made up largely of suburban and wealthy neighborhoods.

CONCLUSION

Based on the results, the analysis was successful in clustering the neighborhoods based on median income, population density, and number of restaurants per neighborhood. It was not as successful at determining what factors in each neighborhood influenced what categories of restaurants were more common. This may be due to the abundance of large chain restaurants that are commonly found in each neighborhood regardless of income or population density. The results may have been better if fast food / chain restaurants were not included. There is also the possibility that these factors have no influence on what category of restaurants are more common in each neighborhoods.