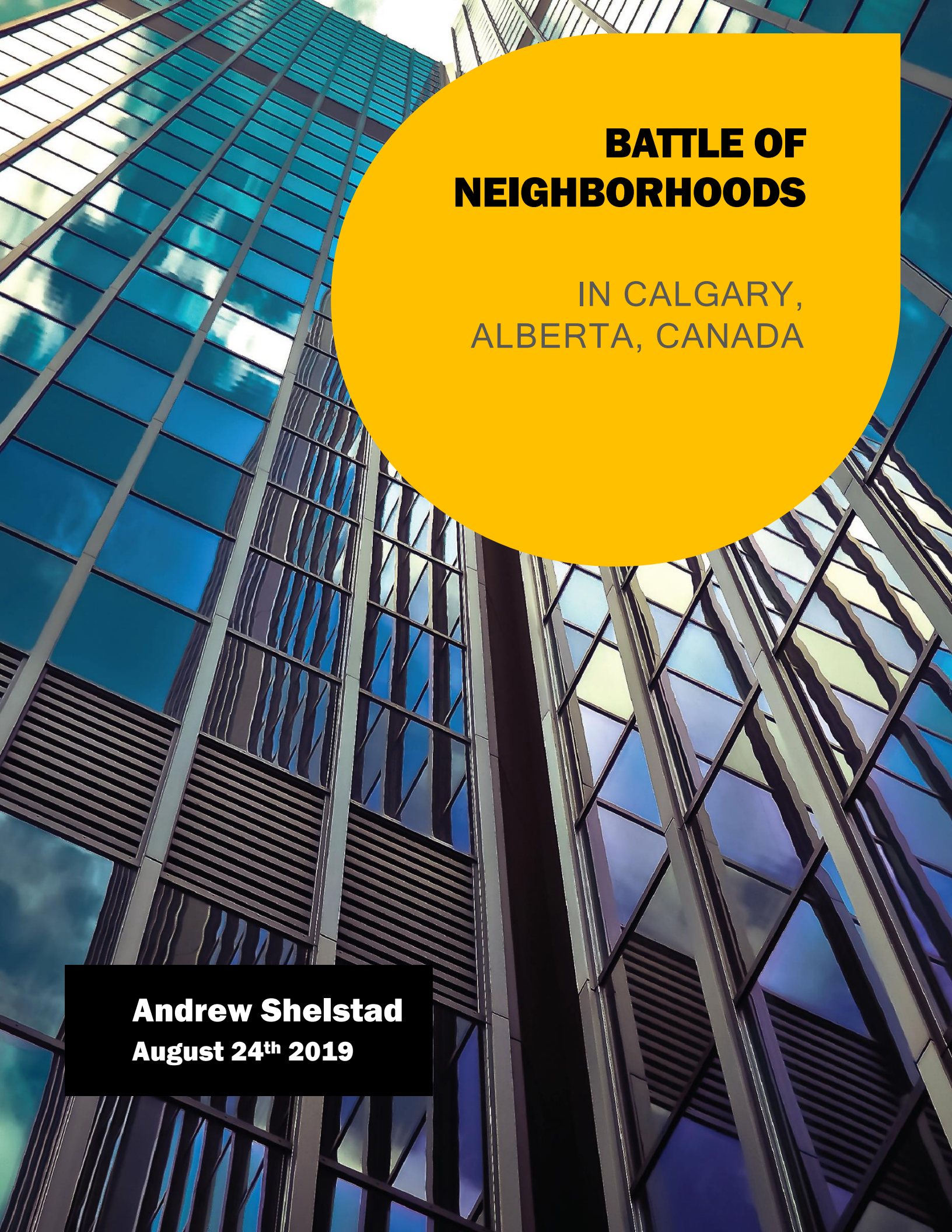

BATTLE OF NEIGHBORHOODS

IN CALGARY, ALBERTA, CANADA

Andrew Shelstad

August 24th 2019

Contents

INTRODUCTION 3
DATA 3
Neighborhood Data 3
Demographics Data 3
Location Data 4
METHODOLOGY 6
Descriptive Statistics and Correlation Between Variables 6
Clustering 9
Evaluating Clusters Using a Classification Algorithm 11
RESULTS 12
DISCUSSION 14
CONCLUSION 14
REFERENCES 15

Table of Figures

Figure 1: Sample of Neighborhood Data Frame ____ *3*
Figure 2: Sample of Demographics Data ____ *4*
Figure 3: Sample of Restaurant Data ____ *4*
Figure 4: The 178 Calgary Neighborhoods Used in the Analysis ____ *5*
Figure 5: Box Plot of Restaurant Category vs. Median Income ____ *6*
Figure 6: Scatter Plots of Neighborhood Features vs Number of Restaurants ____ *7*
Figure 7: Linear Regression Plots of Population Density & Median Income vs. Number of Restaurants ____ *8*
Figure 8: Residual Plots of Population Density & Median Income vs. Number of Restaurants ____ *8*
Figure 9: 3D Plot of Median Income & Population Density vs Number of Restaurants ____ *9*
Figure 10: Box Plot of Median Income & Population Density Across Clusters ____ *10*
Figure 11: Confusion Matrix for SVM Model ____ *11*
Figure 12: Confusion Matrix for KNN Model ____ *11*
Figure 13: Map of Calgary Neighborhoods with Cluster Labels Colored ____ *12*

INTRODUCTION

The purpose of this report is to explore the restaurants in the city of Calgary, Alberta to determine if there is a statistical link between the community demographic data of Calgary neighborhoods and the quantity and type of restaurants that are found within. Calgary is the largest city in the province of Alberta and the 4th largest city in Canada [1]. It has a population of about 1.2 million [1] and a land mass of 825.56km^2 [2]. Calgary is well known for hosting the "Calgary Stampede", an outdoor rodeo and fair that attracts thousands of people from around the world each year. The city is also an attractive city to immigrate to due to a large amount of available jobs in the energy industry. For these reasons (among many others), there is a very diverse range of restaurants that can be found within the city. This report will explore what factors influence the type and quantity of restaurants in Calgary neighborhoods including: neighborhood median income, neighborhood area (km^2), neighborhood population density and number of dwellings within each neighborhood. This information may be useful for anyone looking to open a restaurant in Calgary as it can be used to understand what restaurants are popular in each area and which ones are over saturated. This information can also be useful for anyone who is thinking of moving to Calgary and would like to know what restaurants are popular in each neighborhood of the city.

DATA

The data used for this report falls under the 3 categories of neighborhood data, demographics data, and location data.

Neighborhood Data

The neighborhood data was scraped from a Wikipedia page containing information on all the neighborhoods in Calgary [3]. A table was extracted from this webpage containing information such as the neighborhood name, population, population density, city quadrant and other information. This information is useful to determine the coordinates of each neighborhood to retrieve location data. Some of the columns in this table are also relevant to use in the analysis. Below is a screenshot of a sample of the data frame.

Community	Quadrant	Sector[10]	Ward[11]	Type[10]	2012 PopulationRank	Population(2012)[9]	Population(2011)[9]	% change	Dwellings(2012)[9]	Area(km2)[10]	Populationdensity
Abbeydale	NE/SE	Northeast	NaN	Residential	NaN	5,917	5,700	3.8	2,023	1.7	3,480.6
Acadia	SE	South	NaN	Residential	NaN	10,705	10,615	0.8	5,053	3.9	2,744.9
Albert Park/Radisson Heights	SE	East	NaN	Residential	NaN	6,234	6,217	0.3	2,709	2.5	2,493.6
Altadore	SW	Centre	NaN	Residential	NaN	9,116	8,907	2.3	4,486	2.9	3,143.4
Alyth/Bonnybrook	SE	Centre	NaN	Industrial	NaN	NaN	NaN	−5.9	NaN	3.8	4.2

Figure 1: Sample of Neighborhood Data Frame

Demographics Data

The census data was scraped from a table on the "Great News" website containing information on community demographics in Calgary [4]. The table contains information such as community name, median income by community, and other information.

The data is relevant to use in the analysis to see if median household income influences restaurants in the neighborhood. Below is a screenshot of a sample of the data frame.

Community	Newsletter	Med_Income	Med_Age	Population	Dwellings	Quadrant	Med_Home_Price
Abbeydale	-	$55,345	34	6,071	2,031	SE	$305,000
Acadia	Acadia	$46,089	42	10,969	5,067	SE	$447,000
Albert Park/Radisson Heights	-	$38,019	37	6,529	2,936	SE	$349,900
Altadore	The Source	$53,786	37	9,518	4,537	SW	$925,000
Applewood Park	-	$65,724	33	6,864	2,228	SE	$380,000

Figure 2: Sample of Demographics Data

Location Data

There are two kinds of location data that will be leveraged in the analysis. The first is geographical coordinates of each neighborhood. The coordinates of each neighborhood were retrieved using the geocoder library in python by looking up the neighborhood names in the neighborhoods data frame. The second kind of location data is data on the restaurants in each neighborhood. This was retrieved using the Foursquare API by looking up each coordinate set for neighborhoods in the data frame and returning nearby venues with the search query "food". This data will be used in conjunction with the neighborhood and demographics data frames in the analysis. Below is a screenshot of a sample of the restaurant data frame.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Abbeydale	51.05976	-113.92546	Subway	51.059215	-113.934836	Sandwich Place
1	Abbeydale	51.05976	-113.92546	roadside pub	51.059277	-113.934529	Wings Joint
2	Abbeydale	51.05976	-113.92546	Redbox	51.059108	-113.934845	Pizza Place
3	Acadia	50.97227	-114.05843	Bolsa Vietnamese	50.968895	-114.070252	Vietnamese Restaurant
4	Acadia	50.97227	-114.05843	A&W Canada	50.971985	-114.070898	Fast Food Restaurant

Figure 3: Sample of Restaurant Data

After preparing and combining the data into one data frame, the outcome was 178 neighborhoods to use in the analysis. Figure 4 shows a map of these neighborhoods below:

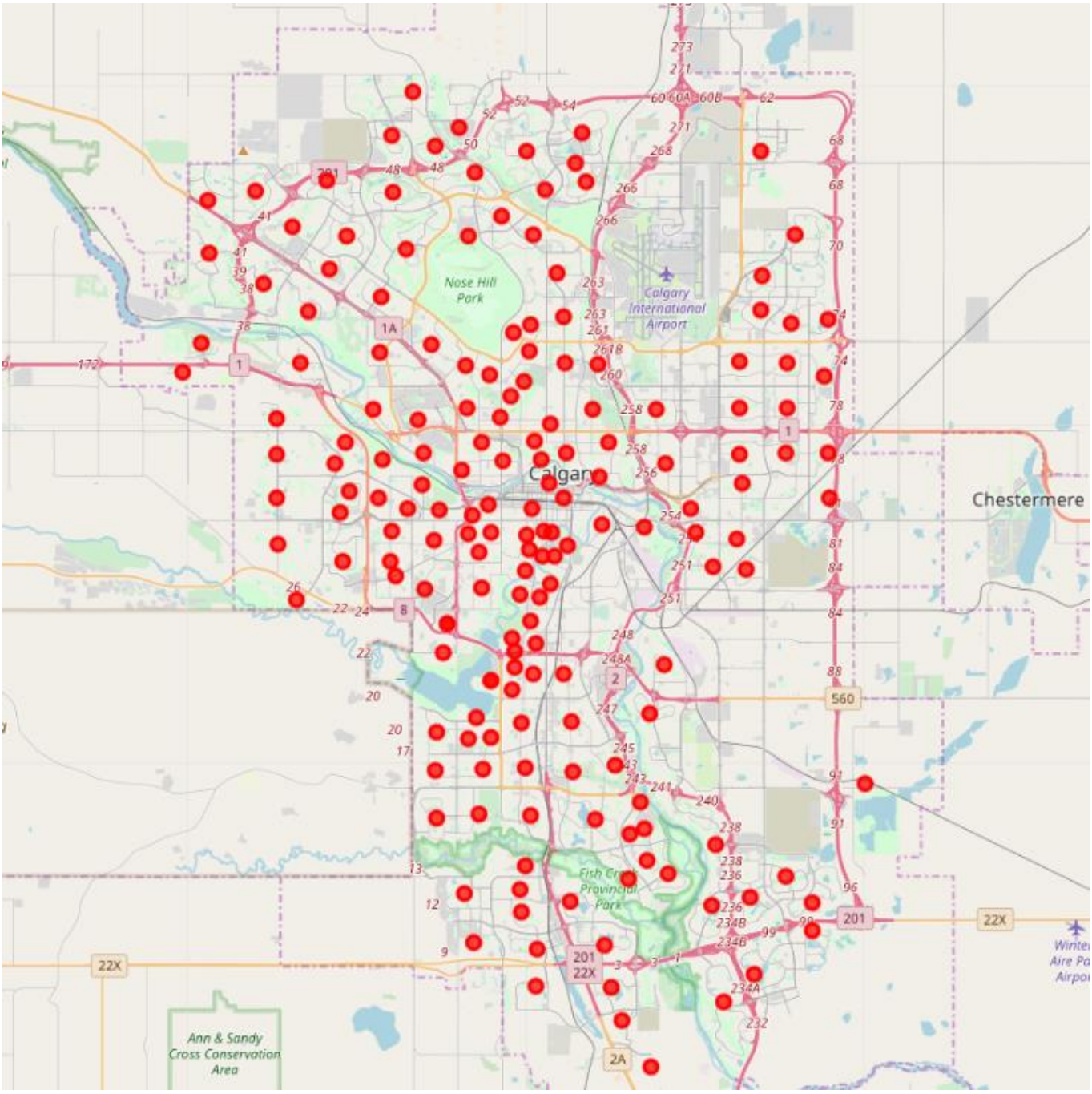

Figure 4: The 178 Calgary Neighborhoods Used in the Analysis

METHODOLOGY

Descriptive Statistics and Correlation Between Variables

The first step in the analysis was to describe more information on the dataset and find out which variables are correlated and statistically significant.

It was determined that the top 5 restaurant categories by number of restaurants in the city are:

1. Pizza Place (246)
2. Sandwich Place (165)
3. Fast Food Restaurant (154)
4. Restaurant (140)
5. Vietnamese Restaurant (127)

A boxplot was produced for the distribution of median income by venue category to see how much the median income ranged in the neighborhoods where these restaurants were located. The boxplot was a useful tool in identifying outliers that may skew the data.

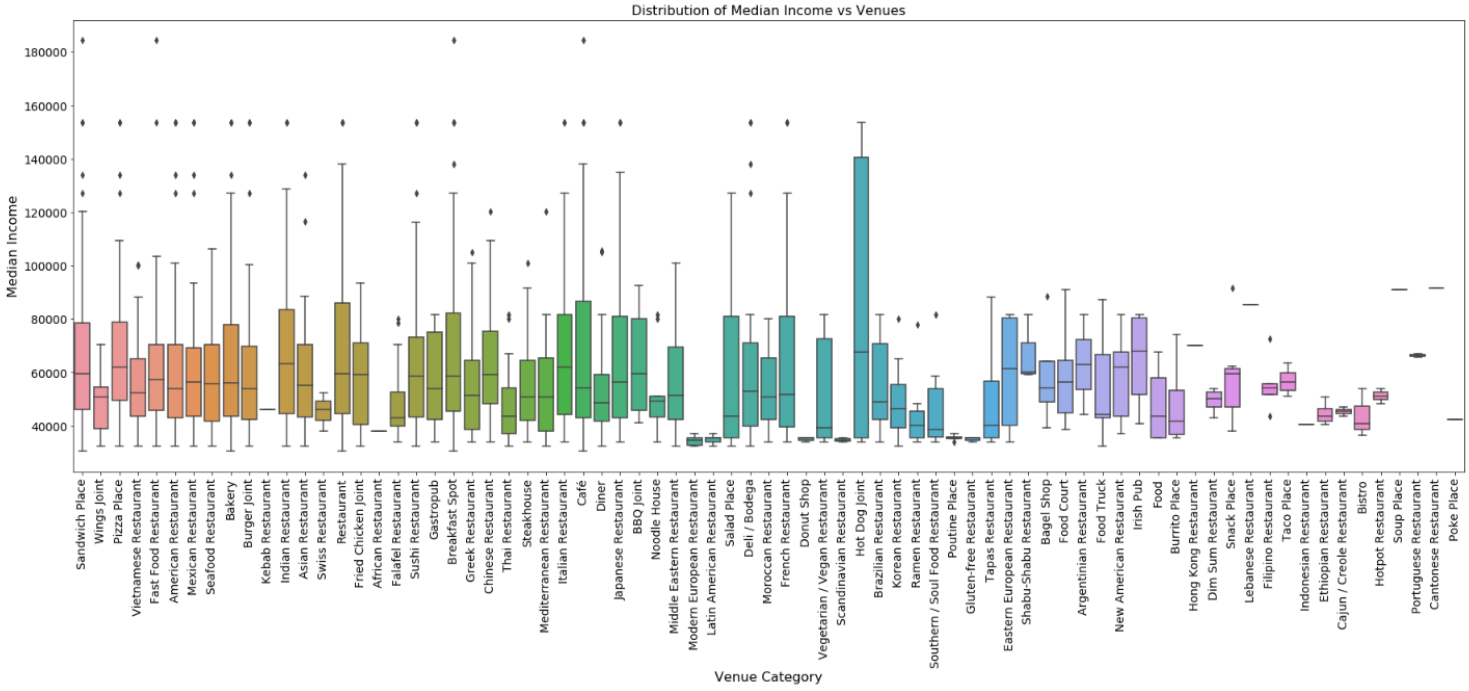

Figure 5: Box Plot of Restaurant Category vs. Median Income

For example, if the average median income by neighborhood is taken per restaurant category the category with the highest income would be Cantonese Restaurant. However, this is misleading because when looking at the box plot and a bar chart of the count of restaurants, it is realized that there is only one Cantonese Restaurant in the data set and it just happens to be in a wealthy neighborhood. Therefore, this is not a good way to determine what impact median household income has on what restaurant categories are in the neighborhood.

Rather than further looking at the neighborhood features in comparison with restaurant categories, it was necessary to first explore the total number of restaurants regardless of category in relation to neighborhood features to determine if there is a correlation between variables. The 4 features that were examined include Population Density, Median Household Income, Number of Dwellings, and Neighborhood Area. Scatter plots were produced with the features as the independent variable and the number of restaurants as the dependent variable. Figure 6 on the next page shows these 4 scatter plots.

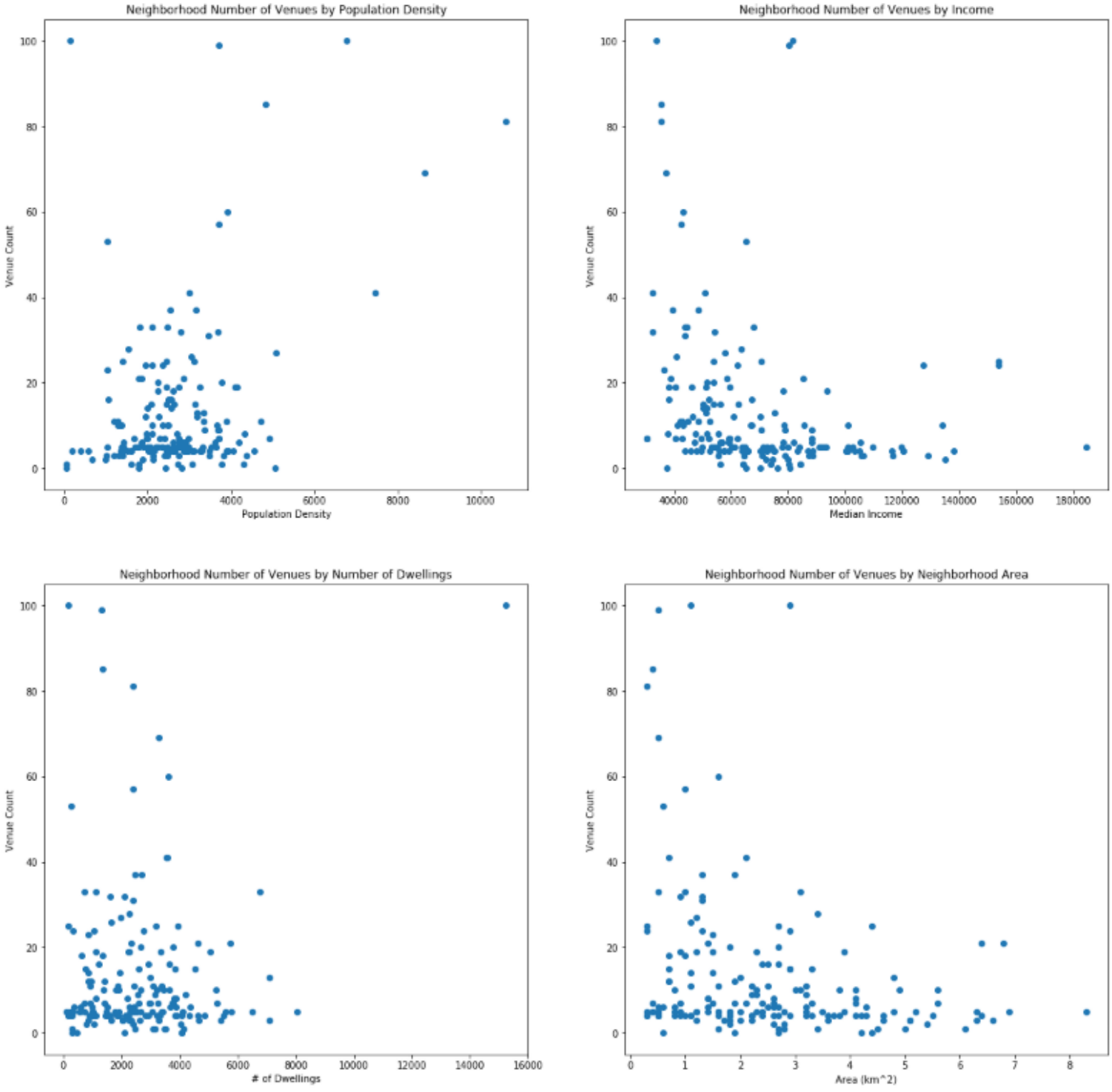

Figure 6: Scatter Plots of Neighborhood Features vs Number of Restaurants

Looking at the scatter plots it seems there is a positive correlation between population density and number of restaurants and a negative correlation between income, number of dwellings and area when each are compared to number of restaurants. The two variables that will be focused on for the remainder of the analysis are Population Density and Median Income. This was chosen because area, population density and number of dwellings are already closely related to each other and only one of the three is necessary for the remainder of the analysis. Figure 7 on the next page shows a linear regression plot of these two variables in relation to number of restaurants.

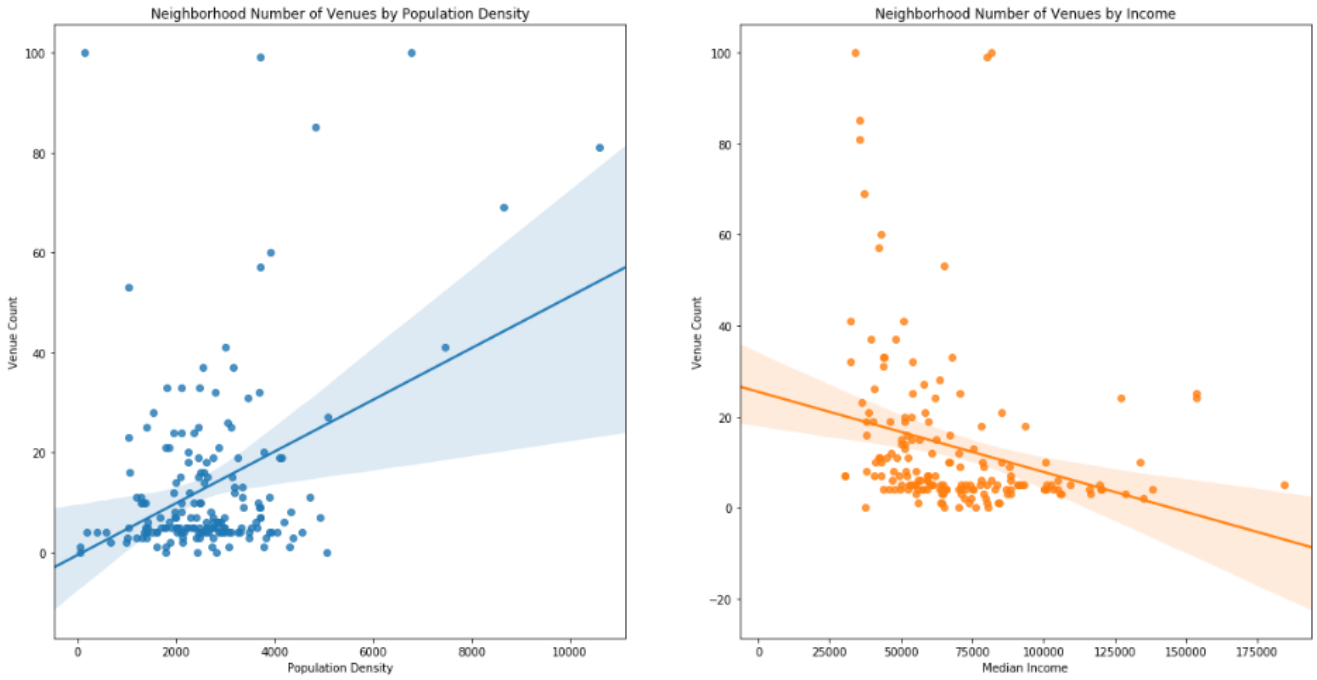

Figure 7: Linear Regression Plots of Population Density & Median Income vs. Number of Restaurants

The results of these regression plots confirm what was hypothesized: there is a correlation between these features and number of restaurants. However, when computing the Pearson Coefficient and R value the results are that there is not a strong linear correlation. The coefficient for population density is approximately 0.4 with a very small R value (<0.001) indicating a weak positive correlation with high certainty. The coefficient for median income is approximately -0.3 also with a very small R value (<0.001) indicating a weak negative correlation with high certainty. To further illustrate the point that a linear model is not suitable for this data set residual plots were created from these features. The plots indicate that there is curvature in the fitted line and a linear model is not suitable.

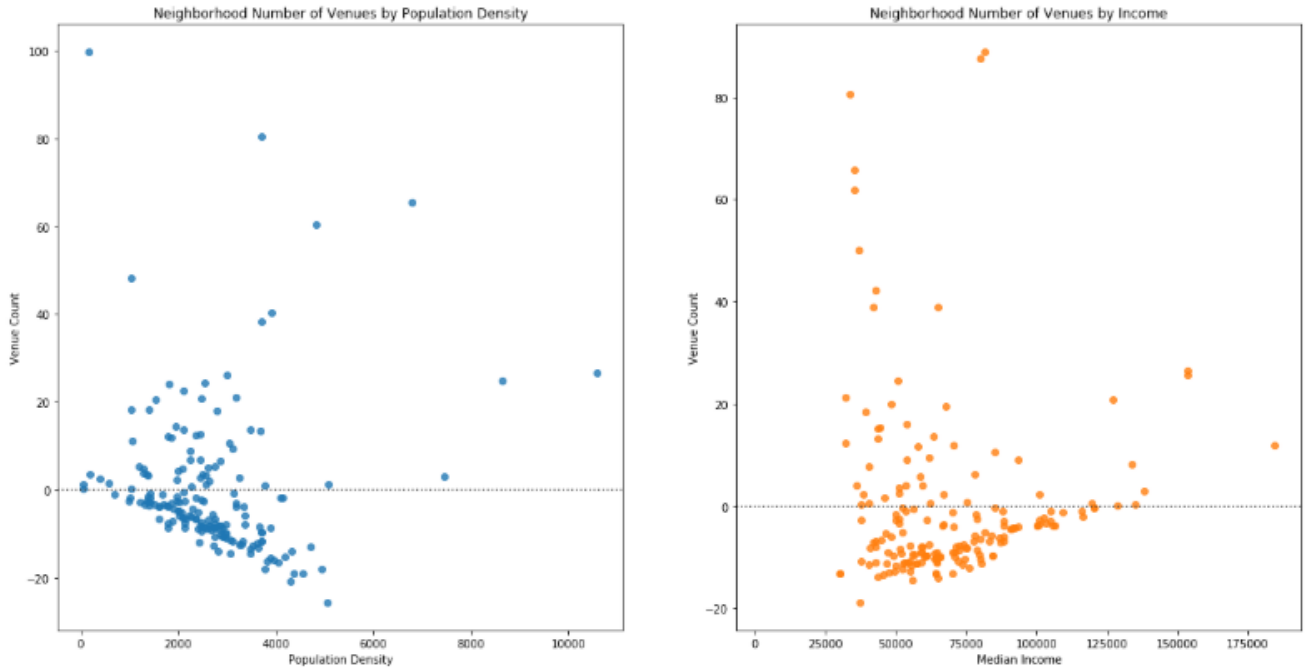

Figure 8: Residual Plots of Population Density & Median Income vs. Number of Restaurants

As a final step in data visualization the features for median income and population density were plotted on a 3D axis against the number of restaurants to see the impact the two features have together on the number of restaurants in a neighborhood.

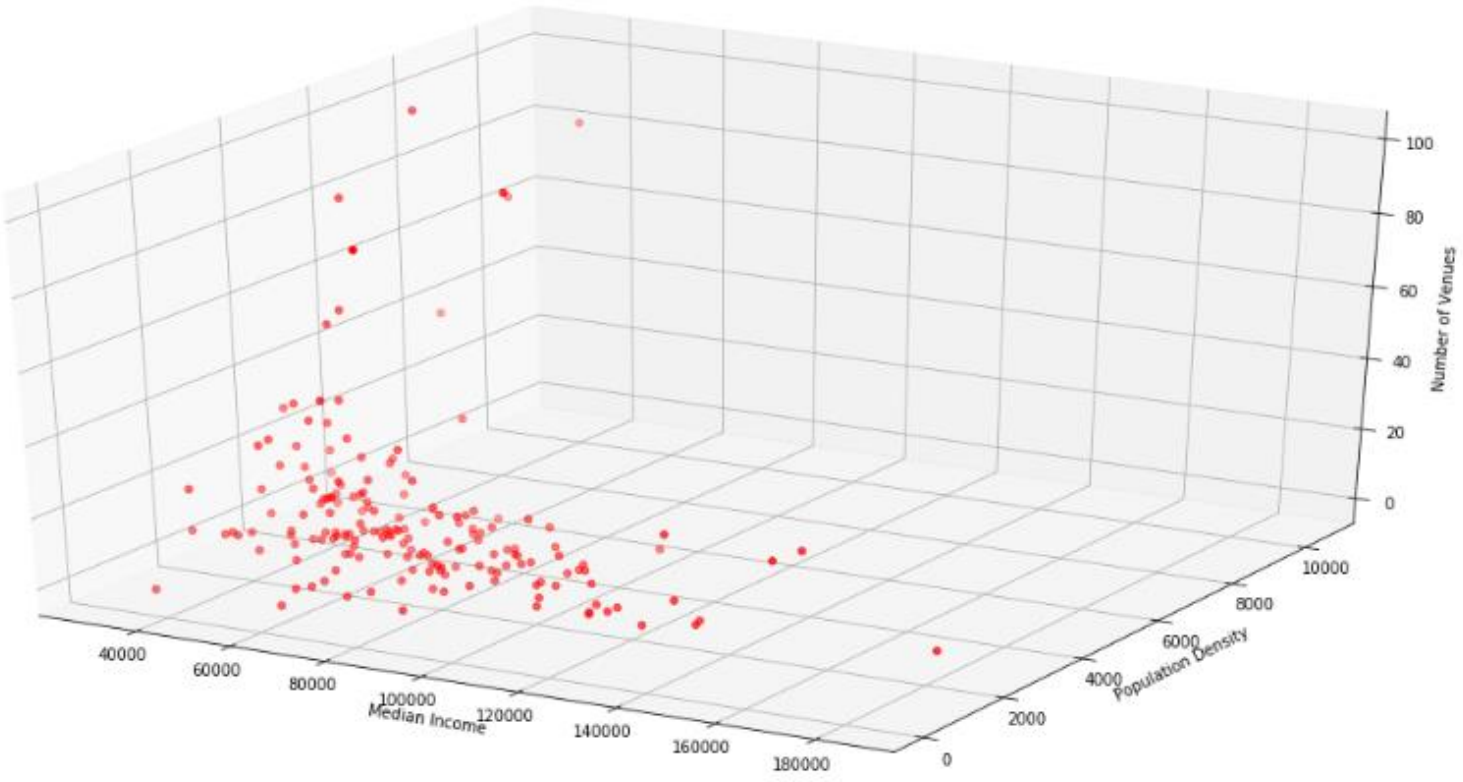

Figure 9: 3D Plot of Median Income & Population Density vs Number of Restaurants

Looking at the 3D plot it can be determined that population density and median income are correlated with the number of restaurants in each neighborhood. The plot shows that there are the most restaurants in lower income areas with higher population density. Although the correlation exists it is not linear and would require a very complex regression model and therefore regression is not an appropriate method to solving the problem. This step was still necessary in the analysis however to determine what variables are statistically significant and can be used in later steps of the analysis.

Clustering

The next step of the analysis was to cluster the neighborhood data using a K Means algorithm and clustered into 3 clusters. The goal of the clustering was to classify the neighborhoods into 3 clusters: low income/high population density, medium income/medium density, and high income/low density. If the results from the correlation step were correct this is how the neighborhoods should be clustered and the differences in venue types and counts can be analyzed between each cluster hopefully yielding different results from each other. The features chosen for clustering were the median income, population density, and the top 5 most common restaurants in the community. The data was scaled using a min-max scaler before being fit into the model. Upon the first test of the clustering algorithm the results did not show a very noticeable difference between the average income and population density between clusters. Furthermore, each cluster had “Pizza Place” as either the highest or second highest instance of restaurants in each community. It was assumed that since Pizza Place was the most common restaurant category in the city and was nearly evenly distributed between neighborhoods, it was skewing the data and causing unwanted results. In response, all venue categories equal to Pizza Place were removed from the data set. It was decided to also remove the venue category “Restaurant” from the data set as this is a very general category and could refer to many different categories of restaurants.

After cleaning the data set, the previous steps were ran again with more promising results. Looking at the below box plots, it is apparent that there is a significant difference in median income and population density between clusters. Cluster 0 has a mean income of $35,000 and a mean population density of 7,700. Cluster 1 has a mean income of $56,000 and a mean population density of 2,800. Cluster 2 has a mean income of $78,000 and a mean population density of 2,400.

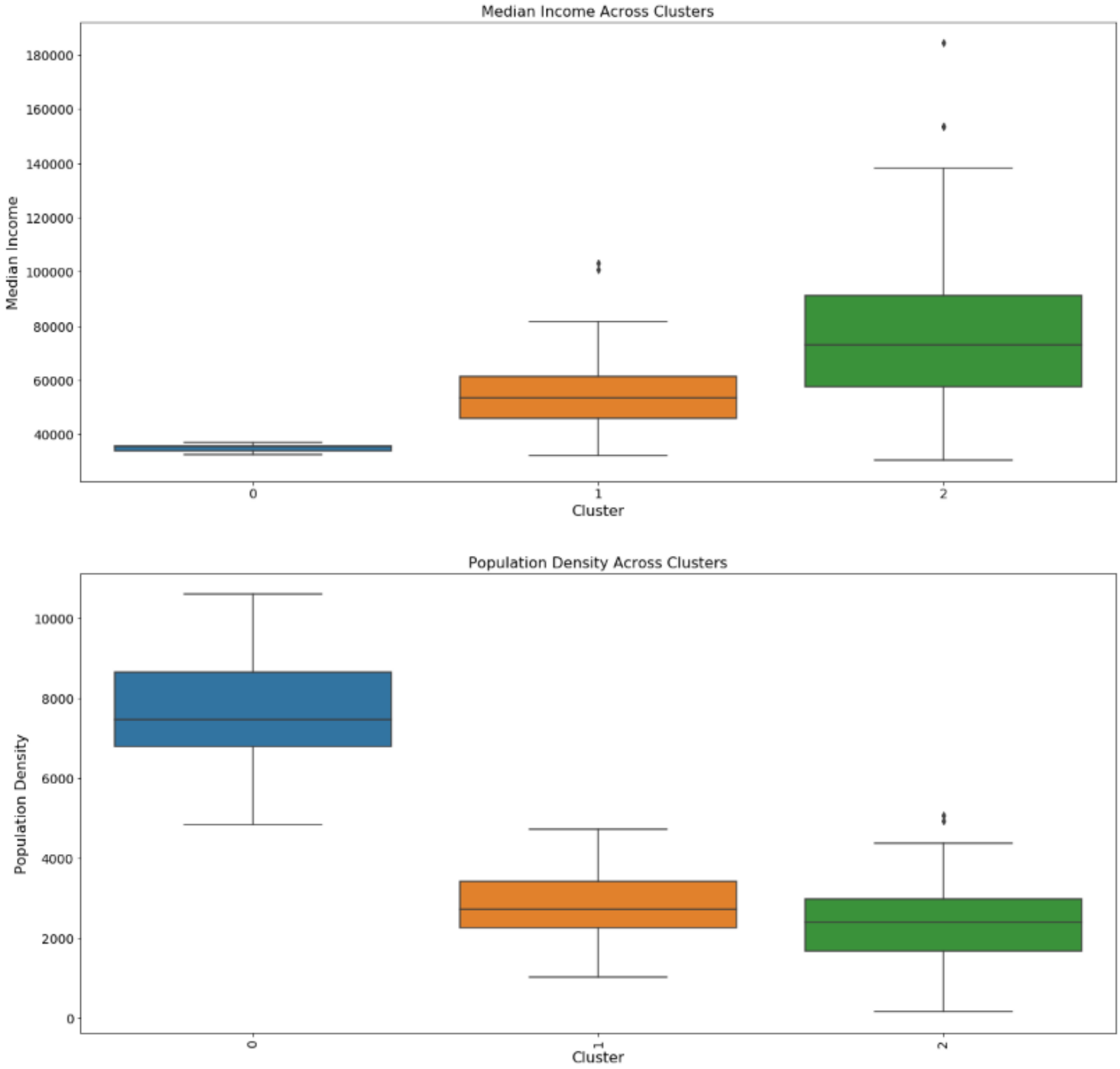

Figure 10: Box Plot of Median Income & Population Density Across Clusters

Evaluating Clusters Using a Classification Algorithm

In order to evaluate how well the unsupervised K Means algorithm segmented the neighborhoods into different clusters, a supervised classification algorithm was trained and tested on the data to see how apparent the difference between clusters are. One important difference between the dataset used in the K Means algorithm and the dataset used in the classification algorithm is that the venue category feature was not used and instead replaced with the count of venues in each neighborhood. The reason for this is that there would be too many features if venue category was selected and the model would potentially overfit or run very slowly. The data was split into training and testing sets (80%, 20%).

The first model that was tested was a support vector machines model. This was selected because there are multiple features and the data set is small. SVM is a good starting point in this application and is much easier to implement than a logistic regression model since the dependent variable is not binary. Each kernel type was tested but Linear was found to yield the best results for this application. Looking at the confusion matrix below it shows that the SVM model predicted almost all the cluster 2 labels correctly but predicted over half of the cluster 1 labels incorrectly and half of the cluster 0 labels incorrectly. The train set accuracy score was 0.68, the test set accuracy was 0.71 and weighted average F1 score was 0.69.

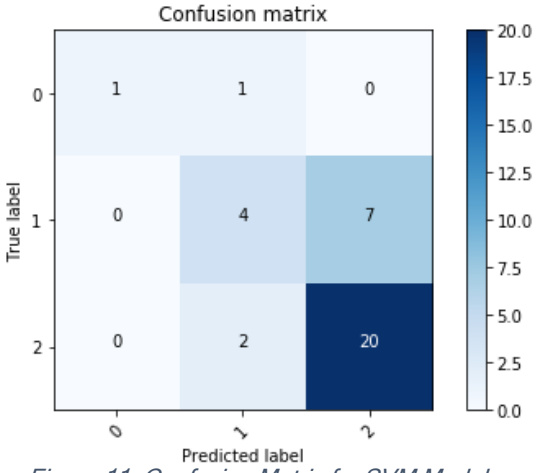

Figure 11: Confusion Matrix for SVM Model

The second model that was tested was developed using the K Nearest Neighbors algorithm for the same reason of ease of use. Different K values were tested but a K value of 3 was found to work the best. Looking at the confusion matrix below it shows that like the SVM model, the KNN model predicted almost all the cluster 2 labels correctly, just under half of the cluster 1 labels correctly and all of the cluster 0 labels correctly. It seems that there is a significant improvement in the performance of the KNN model which has a train set accuracy of 0.88, test set accuracy of 0.74 and weighted average F1 score of 0.73.

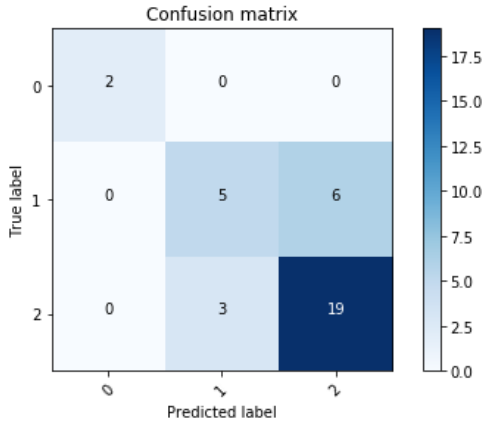

Figure 12: Confusion Matrix for KNN Model

Although the results of the classification models were somewhat accurate, the models could still be prone to overfitting so in order to make sure that they are performing desirably, a cross validation score was taken with 4 folds used. The results were an average score of 0.68 for the SVM model and an average score of 0.69 for the KNN model. The results are that both algorithms are nearly the same in predicting the clusters from the K means model. This instills confidence in the K means model in separating the data into clusters. Although the accuracy is not that high its import to note that the data set is not very big and one of the features used to predict the cluster label was for the count of venues instead of the venue category. With these things considered the classification models performed surprisingly well.

RESULTS

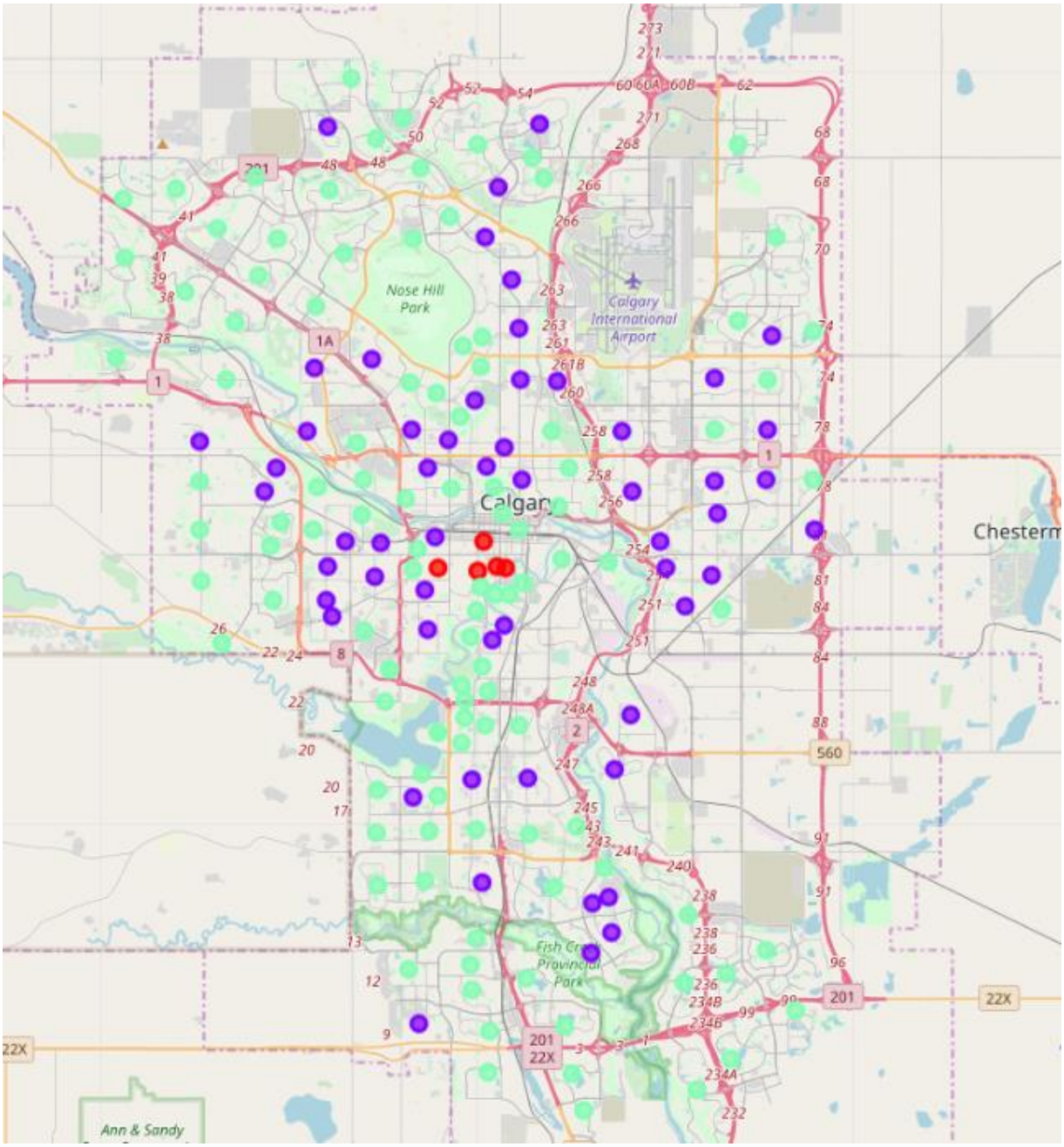

Figure 13: Map of Calgary Neighborhoods with Cluster Labels Colored

For the map on the previous page, the markers in red represent Cluster 0, which is the cluster with the lowest income and highest population density. There are only 5 neighborhoods in this cluster which may seem like the clustering model did not perform well but these neighborhoods are close to downtown Calgary, have lower income and mainly consist of apartment buildings and condos which explains the high population density. These are among the only neighborhoods in Calgary where these characteristics can be found because generally speaking, as the neighborhoods get closer to downtown the median income increases with the population density because the areas become more expensive to live in. These neighborhoods are unique in that they have the lowest median income in the city and highest population density. These communities consist largely of rental properties because they are among the nicer neighborhoods in the city but very expensive to buy a house in. The median age in these neighborhoods is early thirties which is considerably younger than most other neighborhoods in the data set, but the model was not trained using age as a feature. The top five restaurant categories in order in Cluster 0 are Vietnamese Restaurants, Café's, Mexican Restaurants and Japanese restaurants. These results make sense because most of these restaurants are relatively inexpensive. Café's are abundant in urban areas and these neighborhoods are very close to downtown. With this information it can be deduced that Cluster 0 represents neighborhoods with low income, high population density and inexpensive restaurants. These neighborhoods represent only 3% of the neighborhoods in the data set but they have 16% of the restaurants in the data set making these neighborhoods the highest density of restaurants in the city.

The markers in purple represent Cluster 1, which is the cluster with medium income and medium population density. There are 56 neighborhoods in this cluster making it the second largest cluster in the data set. The cluster consists of neighborhoods spread across the city and consists of an average age of 38. One observation that can be made is that none of these neighborhoods are in downtown and most of them aren't as close to it as Clusters 0 and 2. Another interesting observation from the map is that it looks like most of these neighborhoods are adjacent to a major road or highway. While there are some neighborhoods in Cluster 2 that share this characteristic, it appears that this is more common in Cluster 1 and this is especially prevalent in the northwest and southwest of the city. It also seems as if Cluster 2 stretches more outward to the outskirts of the city and surrounds Cluster 1 in some areas, particularly the northwest quadrant. This makes sense because the outer neighborhoods of the city are generally newer suburb communities that have a lot more space than the older row communities. The top 5 restaurant categories in order in Cluster 1 are Fast Food Restaurants, Vietnamese Restaurants, Sandwich Places, Chinese Restaurants, and Sushi Restaurants. Like cluster 0, these restaurants are all relatively inexpensive places to eat. This may not support the theory that the restaurant category has a noticeable change in areas of different income and population density because the difference in income and population density between Clusters 0 and 1 is drastic yet the restaurant categories are not. With this information it can be assumed that Cluster 1 represents neighborhoods with medium income, medium population density, and inexpensive restaurants. These neighborhoods represent 32% of the neighborhoods in the data set and 32% percent of the restaurants in the data set making the ratio of neighborhoods to restaurants relatively the same. Overall this cluster is average in almost every attribute.

The markers in green represent Cluster 2, which is the cluster with the highest income and lowest population density. There are 110 neighborhoods in this cluster making it the largest cluster by far. Like Cluster 1, Cluster 2 is spread across the city relatively evenly and the average age is 39. Generally speaking, the neighborhoods in this cluster are in neighborhoods with a lower population density however there are a few neighborhoods in downtown which has a high population density. All these neighborhoods are in north downtown however and the majority of restaurants in downtown Calgary are located in south downtown, so it makes sense that the model didn't group these neighborhoods with Cluster 0. As previously mentioned, it appears that there are a higher number of neighborhoods on the outskirts of the city than Cluster 1. This makes sense because many of the neighborhoods on the outskirts of the city have houses with acres of land meaning the population density is much lower and the price of housing is higher. The price of owning a house in these outer neighborhoods is comparable to the price of houses in or very close to downtown and since the neighborhoods are fairly evenly distributed between the two it makes sense that this cluster has a smaller difference between mean population density but a larger difference between mean income when compared to Cluster 1. The top 5 restaurant

categories in order for Cluster 2 are Sandwich Places, Café's, Fast Food Restaurants, Chinese Restaurants, and Sushi Restaurants. Again, the difference in popular venue categories are not very dissimilar from the previous cluster at all. In fact, most of these restaurant categories overlap with previous categories although in different order. With this information it can be assumed that Cluster 2 represents neighborhoods with high income, low population density, and inexpensive restaurants. These neighborhoods represent 64% of the neighborhoods in the data set but 52% of the restaurants making it the cluster with the lowest density of restaurants per neighborhood.

DISCUSSION

It can be observed from the results section that the analysis performed on the data for restaurants within Calgary was successful at clustering different neighborhoods based on median income, population density, and number of restaurants per neighborhood. The analysis was not successful in determining what factors in each neighborhood influence what categories of restaurants are more prevalent in each community. A source of error for this could be due to the fact that there are many large chain restaurants such as Subway (Sandwich Place), Starbucks (Café), and McDonalds (Fast Food Restaurant) to name a few that have many restaurants in every major city and are spread out everywhere regardless of income and population density. A recommendation that can be made is to run the analysis without these restaurant categories included (the same way pizza places were dropped from this analysis). However, this may reduce the size of the data by a considerable amount. Another possible source of error is the categorization of venues on Foursquare. As mentioned previously there were many venues with the category of "Restaurant". This is a very ambiguous category that could refer to any category of restaurant so the labelling of data on Foursquare may not always be reliable.

CONCLUSION

This report attempted to answer two questions:

1. Can demographic data for Calgary neighborhoods explain why some neighborhoods have more restaurants than others?
2. Can demographic data for Calgary neighborhoods explain the popularity of certain restaurant categories in certain neighborhoods?

The analysis was successful in answering question 1 by showing that there is a higher number of restaurants in lower income, higher density areas and a lower number of restaurants in higher income, lower density areas. Cluster 0 especially proved this point as it had the lowest number of neighborhoods out of the three clusters but the highest density of restaurants. It was also the neighborhood with the lowest income and highest population density by a considerable amount. The analysis partially answered question 2 but there was not a substantial enough difference between venue categories in clusters for the results to be satisfactory.

REFERENCES

[1] A. Pariona, "The Largest Cities in Canada," 28 May 2019. [Online]. Available: https://www.worldatlas.com/articles/biggest-cities-in-canada.html. [Accessed 24 August 2019].

[2] W. contributors, "Calgary," 23 August 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Calgary&oldid=912097839. [Accessed 24 August 2019].

[3] W. contributors, "List of neighbourhoods in Calgary," 10 May 2019. [Online]. Available: https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary. [Accessed 24 August 2019].

[4] Great News, "Calgary Community Demographics," 2019. [Online]. Available: https://great-news.ca/demographics/. [Accessed 24 August 2019].