Capstone Report

I. Introduction/Business Problem

In this project, we will compare Foursquare data for Toronto and demographic data on Toronto neighborhoods to see if Foursquare data can be a good predictor of any demographic data (and vice versa). The purpose of this exploration is to help citizens and governments use proxy data to understand neighborhoods when desired data is not available. For instance, can the proportion of check-ins of a particular age group for trending restaurants in a particular neighborhood predict the age range of residents in that neighborhood.

II. Data

This project uses two main sources of data. The first is from Foursquare, a local search-and-discovery app which provides personalized recommendations of places to go near a specific location. To get this information we signed up for a Foursquare developer account to be able to use their API. We will then make calls to the API using a list of Toronto neighborhoods.

The second data source contains demographic information on Toronto's neighborhoods. This data comes from the city of Toronto's Open Data Portal (https://portal0.cf.opendata.inter.sandbox-toronto.ca/).

III. Methodology

Data Collection

About the Demographic Data

Demographic data on Toronto neighborhoods was collected from the City of Toronto's Open Data Portal. The csv file was read directly from the download link and stored into a dataframe. To make the data usable for our purposes, we removed unnecessary rows and transposed the matrix so that the neighborhoods represented observations (rows), and the demographic information represented features (columns). Afterwards we extracted the list of neighborhoods for use in searching Foursquare. See Example of Demographic Data for collection code

About the Foursquare Data

Foursquare data is accessible through the company's API. After creating a developer account, Explore calls to the API were made for various neighborhoods to get a list of recommended venues near the searched location. Information about the Explore calls can be found at https://developer.foursquare.com/docs/api/venues/explore.

After collecting recommended venues for each neighborhood, we made Details calls for each venue in the search results to get a set of details about a venue including location, tips, and categories. Because the Details call is a premium call, we were only able to make 500 calls a day. As such, we decided to limit the number of neighborhoods included in our analysis and search 50 venues (the max limit per Explore call) for each neighborhood. The list of neighborhoods was determined randomly.

Note: Searching the Foursquare app using neighborhood names is inconsistent, but often returns results. Using neighborhood names with the API, however, rarely returns results (even when the app did). As such Google Maps was used to identify the center coordinates of neighborhoods and manually entered as a dataframe.

Searched Neighborhoods: Agincourt North, Alderwood, Annex, Bathurst Manor, Bayview Village, Cliffcrest, Dorset Park, Flemingdon Park, Forest Hill North, Guildwood, Henry Farm, Highland Creek, Hillcrest Village, Humber Summit, Ionview, Kennedy Park, Little Portugal, Long Branch, Malvern, Markland Wood, Morningside, Mount Dennis, New Toronto, Oakridge, Regent Park, Roncesvalles, Rouge, Scarborough Village, The Beaches, Thorncliffe Park, West Hill, Weston, and Woburn

The API calls returned JSON files, which were converted to data frames and stored as csv files over several days.

Data Analysis

Exploratory

We first looked at a correlation matrix to determine which demographic features had the highest correlations with the Foursquare features.

Correlation Matrix

	likes	price	rating
Population density	0.576117	0.425788	0.6593
Population Change	0.286917	0.230166	0.402191
Percent children	-0.47105	-0.23649	-0.29716
Percent youth	-0.28073	-0.40864	-0.46793
Percent working Age	0.585589	0.396203	0.723585
Percent pre-retirement	-0.38745	-0.38722	-0.45412
Percent seniors	-0.09204	0.006017	-0.28308
Percent older Seniors	-0.02579	0.060563	-0.14327
Percent married	-0.24157	0.017862	-0.2614
Percent living alone	0.620684	0.49744	0.706935
Average household size	-0.58235	-0.49385	-0.65863
Average income	0.468617	0.54716	0.463065
Percent homeowners	-0.28857	-0.3199	-0.44765
No degree	-0.38821	-0.34849	-0.31516
Secondary diploma	-0.64034	-0.66253	-0.67425
Postsecondary degree	-0.6357	-0.67858	-0.65023
Bachelors degree	0.617754	0.611373	0.610611
Postgraduate degree	0.587558	0.597439	0.551277

Participation rate	0.425325	0.418792	0.632085
Employment rate	0.464837	0.464394	0.647313
Unemployment rate	-0.4803	-0.49766	-0.50397

The correlation matrix shows that the following features generally have a medium to strong correlation with likes, price, and rating:

- population density
- change in population
- age (children to pre-retirement)
- percentage of persons living alone
- percentage of home owners
- education
- workforce participation rate
- workforce employment rate
- workforce unemployment rate

Predicting Foursquare data

We used linear regression models to identify which features of demographic data can predict the three main types of Foursquare data: likes, price, and rating.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  likes   R-squared:                       0.619
Model:                            OLS   Adj. R-squared:                  0.580
Method:                 Least Squares   F-statistic:                     15.71
Date:                Wed, 20 Mar 2019   Prob (F-statistic):           2.95e-06
Time:                        18:30:57   Log-Likelihood:                -97.964
No. Observations:                  33   AIC:                             203.9
Df Residuals:                      29   BIC:                             209.9
Df Model:                           3
Covariance Type:            nonrobust
=========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                   37.7569      8.311      4.543      0.000      20.759      54.755
Population density       0.0008      0.000      2.579      0.015       0.000       0.001
Postsecondary degree   -66.2331     28.721     -2.306      0.028    -124.974      -7.492
Unemployment rate       -1.4507      0.426     -3.408      0.002      -2.321      -0.580
==============================================================================
Omnibus:                        2.425   Durbin-Watson:                   1.976
Prob(Omnibus):                  0.297   Jarque-Bera (JB):                1.298
Skew:                           0.148   Prob(JB):                        0.522
Kurtosis:                       3.926   Cond. No.                     2.20e+05
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.346
Model:                            OLS   Adj. R-squared:                  0.302
Method:                 Least Squares   F-statistic:                     7.926
Date:                Wed, 20 Mar 2019   Prob (F-statistic):            0.00172
Time:                        18:34:26   Log-Likelihood:                 20.054
No. Observations:                  33   AIC:                            -34.11
Df Residuals:                      30   BIC:                            -29.62
Df Model:                           2
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      2.0665      0.165     12.488      0.000       1.729       2.404
Average household size    -0.1425      0.067     -2.120      0.042      -0.280      -0.005
Unemployment rate         -0.0271      0.013     -2.161      0.039      -0.053      -0.001
==============================================================================
Omnibus:                        0.215   Durbin-Watson:                   2.196
Prob(Omnibus):                  0.898   Jarque-Bera (JB):                0.407
Skew:                           0.120   Prob(JB):                        0.816
Kurtosis:                       2.511   Cond. No.                         70.6
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 rating   R-squared:                       0.790
Model:                            OLS   Adj. R-squared:                  0.769
Method:                 Least Squares   F-statistic:                     36.44
Date:                Wed, 20 Mar 2019   Prob (F-statistic):           5.74e-10
Time:                        18:44:20   Log-Likelihood:                 14.531
No. Observations:                  33   AIC:                            -21.06
Df Residuals:                      29   BIC:                            -15.08
Df Model:                           3
Covariance Type:            nonrobust
==========================================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      8.3900      0.338     24.839      0.000       7.699       9.081
Unemployment rate         -0.0855      0.016     -5.424      0.000      -0.118      -0.053
Percent pre-retirement    -8.4761      1.620     -5.231      0.000     -11.790      -5.162
Percent living alone       0.0234      0.005      4.279      0.000       0.012       0.035
==============================================================================
Omnibus:                        0.068   Durbin-Watson:                   1.458
Prob(Omnibus):                  0.967   Jarque-Bera (JB):                0.283
Skew:                          -0.003   Prob(JB):                        0.868
Kurtosis:                       2.546   Cond. No.                         986.
==============================================================================
```

IV. Results

Predicting Venue Likes

We found that population density, postsecondary degree, and unemployment rates were the best predictors for venue likes.

Predicting Venue Price

Price was difficult to predict with the available demographic data. The best we found was using average household size and unemployment rate, both of which are negatively correlated and account for only a third of the variability in price.

Predicting Venue Rating

We see that percent of working age adults accounts for half of the variability of rating and that the higher percentage, the higher the rating will be. This suggests new businesses should open in areas with large working-age populations. The highest R-squared value we found, however, was for a model looking at unemployment rate, percent living alone, percentage of pre-retirement individuals, and percentage of people living alone. The lower the unemployment rate and the percent of pre-retirees, along with a slight increase in percent of people living alone, accounts for greater ratings.

Predicting Demographic Information

In testing various regression models using Foursquare data to predict demographic information, we noticed that individual features can be fairly good predictors (accounting for 30 to 50% of targets), but combining predictors achieve no significant increase in R-squared. This is likely because the three Foursquare predictors are themselves highly correlated and don't contribute much to the model when added together.

Nevertheless, here are the outcomes of various regression models:

A higher average number of likes for venues in a particular neighborhood indicates,

- a greater population density (R-squared = 31)
- a higher percent of working age people (R-squared = 32)
- a higher percent of people living alone (R-squared = 37)
- a lower average household size (R-squared = 32)
- a lower percent of people who have not completed at least a bachelor's degree (R-squared = 39)
- a higher percent of people who have completed at least a bachelor's degree (R-squared = 36)

A higher average price for venues in a particular neighborhood indicates

- a lower percent of people who have not completed at least a bachelor's degree (R-squared = 42)
- a higher percent of people who have completed at least a bachelor's degree (R-squared = 35)

A higher average rating for venues in a particular neighborhood indicates

- a greater population density (R-squared = 42)
- a higher percent of working age people (R-squared = 51)
- a higher percent of people living alone (R-squared = 48)
- a lower average household size (R-squared = 42)
- a lower percent of people who have not completed at least a bachelor's degree (R-squared = 44)
- a higher percent of people who have completed at least a bachelor's degree (R-squared = 35)
- a higher workforce participation rate (R-squared = 38)
- a higher workforce employment rate (R-squared = 40)

V. Discussion

None of the regression models we created had high enough R-squared values for us to make any meaningful recommendations. However, they do show some general trends. For instance, neighborhoods that have venues with higher than average ratings and likes, are likely populated with a higher proportion of working age people and lower household sizes. Does this indicate that working age people in smaller (or no) families are more likely to rate or like a venue? Or does this mean that they are more likely to rate it higher? Further analysis is needed.

Other interesting, although logistical trend is that neighborhoods with more higher-priced venues (higher average price) are more likely to have more highly educated people living in them. Interestingly, though, average household income was not a good predictor for price, which may suggest that people don't patron expensive venues because they have more money, but because they are more educated.

VI. Conclusion

We began this project with the hope that we could find some linkages between Toronto's demographic data and Foursquare data for venues in Toronto. Although some of the results were promising, neither data sets gave complete conclusions. We recommend that if both governments and business would like to use these results for their own predictions, that they do so carefully. The results show general trends, but cannot be used to predict exact amounts. There is still too much variability that is not explained by the predictors used here, both for predicting demographic information and Foursquare information. However, we do believe these results can be used for a quick check for new businesses to see if their other research matches with the results found in this project.