**Gas station prevalence prediction**

**Introduction/Business Problem**

Gas stations are required throughout the city to supply fuel for transportation. They can often be found in dense retail sectors that have a variety of other venues present within a certain radius. Furthermore, you often see different companies opening gas stations in close approximation to each other.

The problem is, how many gas stations are typically built to meet the needs of the community? This is a crucial problem for not only gas companies but also city planning. Allowing too many gas stations results in less area for local businesses to open their stores. The stations also have negative impacts on the environment as the sites they are built on cannot be used after the gas station is removed. Therefore, it would be beneficial to predict how many gas stations are typically built in any given area around an existing gas station so that it can inform future design and build recommendations. (<https://www.theglobeandmail.com/report-on-business/industry-news/property-report/life-after-corner-gas-the-challenges-of-developing-old-service-stations/article32219739/>)

Ottawa, ON Canada is going to be used as the test city as it has a variety of different retail spaces and city densities that can be a good representation of how gas stations are distributed.

# Data

The data that’s going to be used is primarily the Foursquare API data. First the location and label of the venues will be used to identify gas stations using the venues search endpoint and limiting the category. Secondly the other venues in the area will be identified grouped by their and counted. Secondly, population data at an FSA (forward station area) level will be used. This data will be collected from Statistics Canada ( Example can be found here: <https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&S=22&O=A>).

THe population is calculated by finding all of the FSAs within a 2Km radius and summing their populations. This was done using the website (<https://www.freemaptools.com/find-canada-postcodes-inside-radius.htm>). Requests were made to that website to obtain the FSAs and the Stats Canada data was used to group and sum the populations.

For each gas station in Ottawa, the number of gas stations located within a 2km radius will be identified and used as the dependent variable. The population of the gas station will be the total population of every FSA within a 2km radius. Similarly, only the other venues within a 2km radius will be used as predictors. The venues will be differentiated by category. Since there are many categories, the categories will be grouped by the highest level present in the FourSquare category hierarchy.

After the data is gathered and prepared a variety of regression based models will be used to predict the number of surrounding gas stations with the collected data.

**Data Collection**

**Gas Stations**

Data collection was broken down into a number of phases. First, Wikipedia was used as a source to get the FSAs present in Ottawa. This was used later to filter out gas stations not in the city. Next the Foursquare API was used to collect the gas stations in Ottawa. Im used the specific category ID for gas stations obtained from the foursquare website. Since the foursquare API search only returns the top 50 gas stations in a given area, I used google maps to choose overlapping points surrounding Ottawa. I then did an API call looking for the top 50 gas stations within a 25km radius.

After the initial gas stations were discovered and filtered with the FSAs acquired from Wikipedia, the number of surrounding gas stations in a 2Km radius was calculated using the FourSquare API as well. Additionally, gas stations not originally present in the initial query (due to the limitation of 50 venues) were found and nearby gas stations calculated.

**Population**

Population was obtained by getting the population of every FSA within a 2Km radius. The K0A FSA is being omitted due to how large it is and its disproportional population compared to others in the area. The population data was obtained from Statistics Canada and is present as a .csv file that is being read in. To get the FSAs within 2Km of each gas station the freemapstools.com tool for getting the FSAs in a radius was used.

**Nearby Venues**

Similar to counting the nearby gas stations, other venues within a 2Km radius surrounding each gas station were found and counted. Since the primary category for each of the venues results in many sparse groups, the parent category was obtained. After this was completed the number of venues in each parent category was counted.

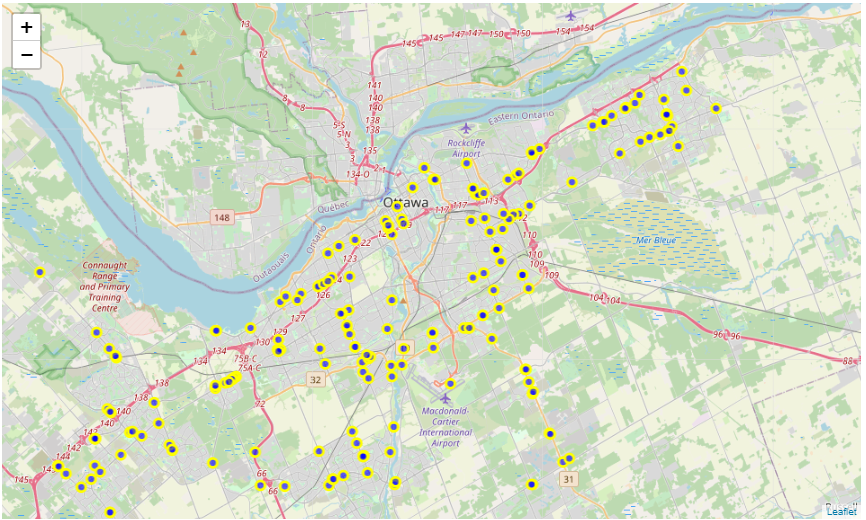


Figure 1: Gas stations found in Ottawa using the FourSquare API. Map displayed using Folium.

**Methodology**

**Data Exploration**

The dependent variable is a count variable within a fixed distanced. Since this is the case it was hypothesized to follow a Poisson distribution. To test this hypothesis a histogram was done for the surrounding gas station count. Poisson distribution’s mean is equal to its variance. While this isn’t the case for this particular distribution they are close enough that the over all population could still follow that distribution.

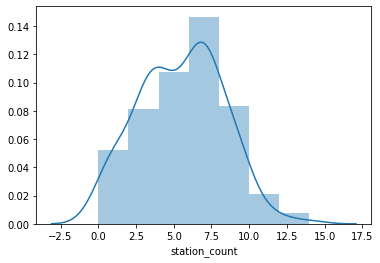


Figure 2: Histogram of number of gas stations within a 2Km radius of each gas station within Ottawa.

The histogram suggests a Poisson distribution. To check further a Quantile-Quantile plot can be done to compare the quantiles of an actual Poisson distribution to this one.

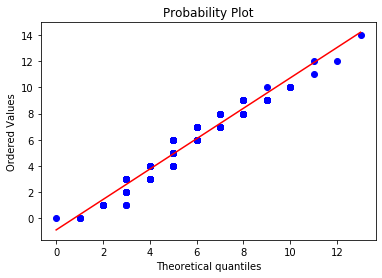


Figure 3: QQ-Plot of the Number of gas stations within a 2Km radius of a gas station compared to a Poisson distribution with a mu value of 5.471.

The QQ Plot also suggests a Poisson distribution. A final check using a Kolmogorov-Smirnov test was done to check for goodness of fit for the distribution. The test resulted in a p value of 4.19e-9 meaning it is statistically similar to the Poisson distribution.

**Feature Analysis**

A seaborn pair plot was done on all of the features. Using this pair plot College and University, Travel and transport, profession and other places and Arts and Entertainment were removed from the analysis due to their sparsity and lack of any significant correlations. The pair plot of the resultant features is shown below.

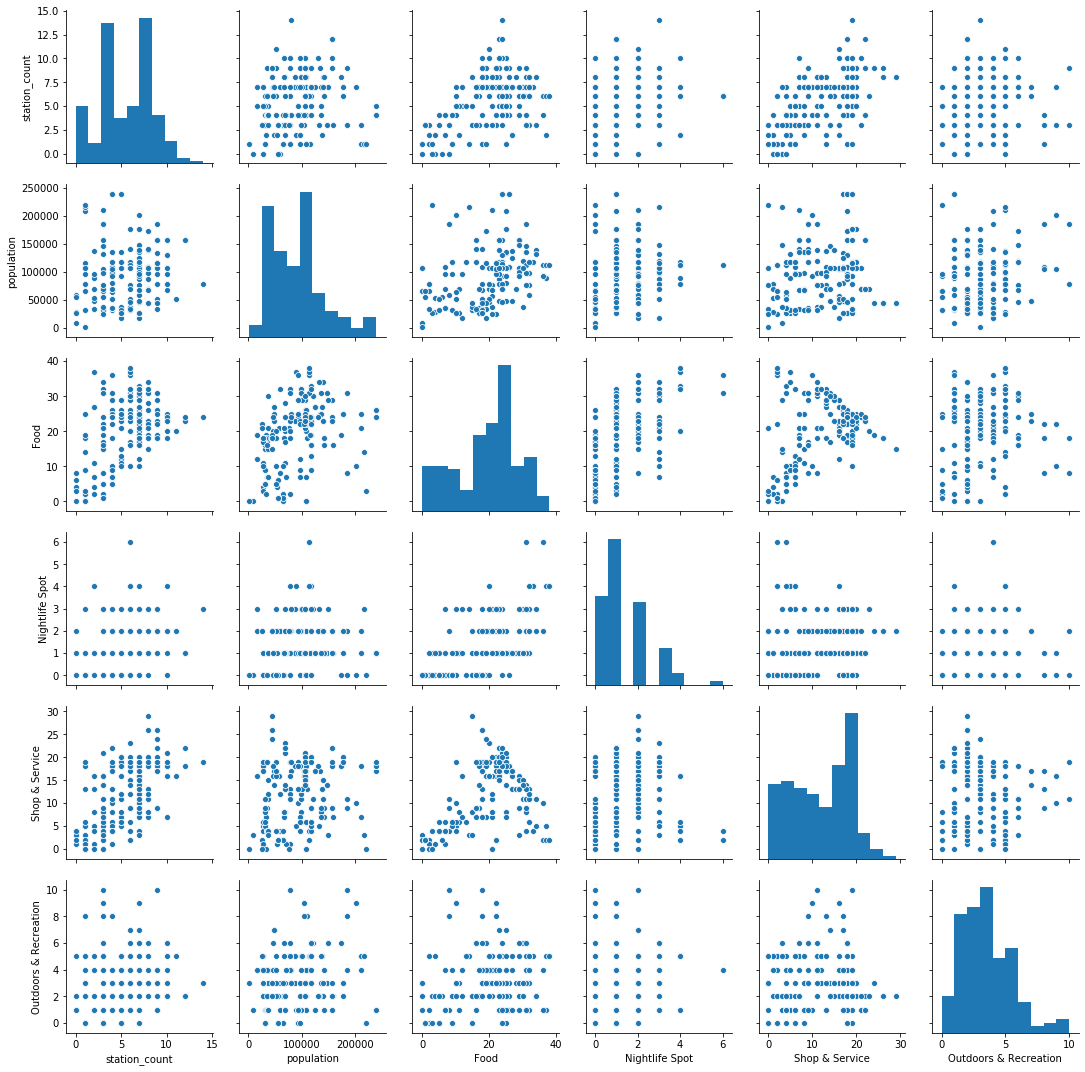


Figure 4: Pair plot of features collected to predict the number of gas stations within a 2Km radius.

**Model Development**

The data was split into a 20:80 train and test split. Due to this being a regression analysis 4 different models were compared. OLS Linear regression, ridge regression, random forest regression and Poisson regression. The models were compared by using a 5 fold cross validation and a average root-mean squared error was calculated using each fold. SkLearn was used for all of the models with the exception of the Poisson regression which was used from statsmodels and wrapped in a SkLearn class. The Random forest regression max depth parameter was optimized by iterating through choices from 2-10 and comparing the cross validated scores. Ultimately a max depth of 5 performed the best.

Lastly each of the models were trained on the full train set and a final test was performed on the hold out test set. Models were compared using a root-mean squared error analysis.

**Results**

The random forest regression performed the best on the test data with a RMSE value of 1.87. This aligned with the scores obtained in the train data cross fold validation comparison as well indicating the performance was likely not due to overfitting. Surprisingly the Poisson regression performed the worst despite the dependent variable following that distribution.

Table 1: Comparison of predictive models performance on the test data set.

| **models** | **Train CV Results** | **Test Results** |
| --- | --- | --- |
| Linear OLS | 2.1725 | 2.083854 |
| Linear Ridge | 2.1729 | 2.075226 |
| Random Forest | 2.145 | 1.874256 |
| Poisson | 2.1744 | 2.154290 |

**Discussion**

Using the data collected on the number of venues and the population the model using the regression tree regression could predict the number of gas stations within a 2Km radius on average within approximately 2 gas stations. The analysis confirmed that the number of gas stations surrounding a gas station within 2Km follows a Poisson distribution. With this model city planning can be done to deny an additional gas station to be built where the model outputs a significantly lower number (< 2 already present) in the 2Km area. Therefore these areas can be used for other venues such as local businesses. It would also prevent that site from becoming unusable due to any leakages or accidents, further damaging the local environment and potential for business.