Exploring relationship
between venues and population in Toronto

Zhidi Luo

Feb 19, 2019

1. Introduction

1.1 Background

Population is a hot topic that is extremely useful to various studies. Population is the total number of people living in an area, such as a town, a city or a country. Having more people living in the same area could result in crowdedness, increase of noise, and etc. But at the same time, more people could also mean the thrive of the area, with more shops, more gyms, more restaurants, and etc. I believe there is a strong relationship between the amount / diversity of venues in a location and the population of the location. This will help us understand what makes a highly populated location, or what kinds of facilities or venues usually exist in a highly populated location. It would be advantageous if we can leverage this knowledge to adjust the highly populated area or an extremely low populated area. This will help with city planning on housing constructions, venue selections, and etc.

1.2 Problem

Toronto data will be used to explore the relationship between venues and population. Venues include types of venues and the total number of venues in each postal codes, and the venues will be compared to the population of the same postal code.

2. Data

2.1 Data Sources

Venues data can be found from foursquare, as we explored in previous weeks. We will use venue names and venue categories within each postal codes. Population data can be found in Statistics Canada website Population and Dwelling count highlight tables in the link below:
https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/hlt-fst/pd-pl/Table.cfm?Lang=Eng&T=1201&SR=1&S=22&O=A&RPP=9999&PR=0

2.2 Data Preprocessing

Data requested from foursquare will include postal code, borough, neighborhood, and we can also request data for venue name and venue category. After getting rid of unwanted strings and cleaning up the venue category, we will combine everything and have the data frame with each row a venue with name, category and postal codes. Then we will group the venues within the same postal codes, creating the unique number of venue categories and the total number of venues within each postal code. Finally, we will combine this data with population data downloaded from Statistics Canada.

3. Methodology and Results

3.1 Exploratory Data Analysis

First we want to get a sense of what the data looks like. After cleaning up the data, we have two independent variables and one dependent variables that we want to look at. The two independent variables are:

- Total number of venues in a postal code
- Number of unique venue categories in a postal code.

The dependent variable is:

- Population of the postal code

We first look at the description of the data as below:

	Venue #	Venue Category #	Population
mean	44.7	29.1	20064.1
std	36.7	20.5	13766.8
min	2.0	2.0	0.0
25%	15.3	12.5	10695.8
50%	34.0	24.0	18832.0
75%	85.5	51.3	31305.3
max	100.0	66.0	49195.0

The average number of venues and the standard deviation of venues are both higher than venue category number, which makes perfect sense because a location could have multiple venues of the same category. The average number of population is about 20,000 within a postal code, and the median is about 19,000. The mean of venue 44.7 and the mean of venue categories 29.1 are much higher than the median of venue 34 and the median of venue categories 24, suggesting that the two independent variables might be skewed. We will confirm their skewness in the next section by plotting them. Also note that the minimum number of population is zero, which does not make sense to me. Either there's something about the population that I didn't know, or this is a data error. But since the 25% quantile of population is 10,695, which makes perfect sense, we are going to proceed assuming the data is ok.

Next we will look at the correlations of the three variables:

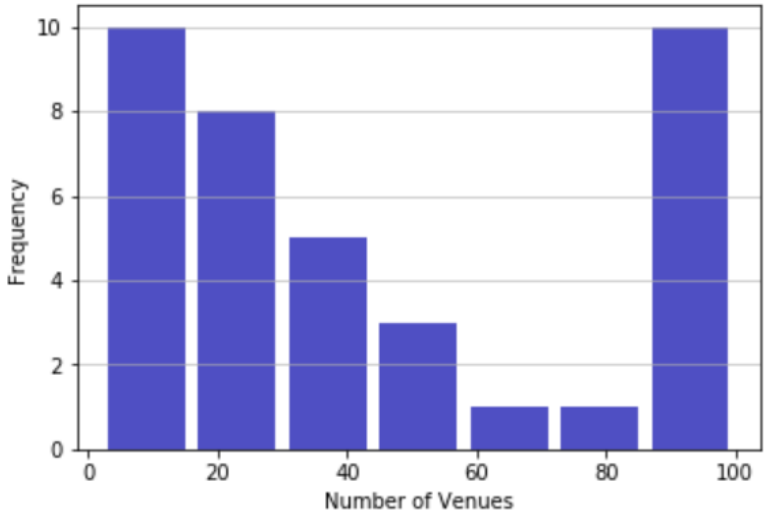
	Venue #	**Venue Category #**	**Population**
Venue #	1	0.984	-0.469
Venue Category #	0.984	1	-0.409
Population	-0.469	-0.409	1

We can see that venue number and the venue category number are highly correlated. The population and these two independent variables are both negatively correlated, both correlations are around -0.4 to -0.5.

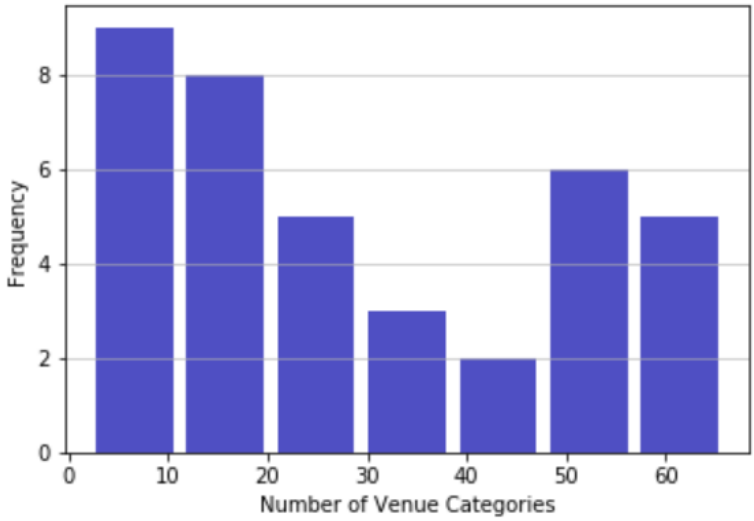
3.2 Data Visualization

From Section 3.1, we have a general sense of what the variables look like. In this section, we will plot some of these graphs to have a better understanding of the data.

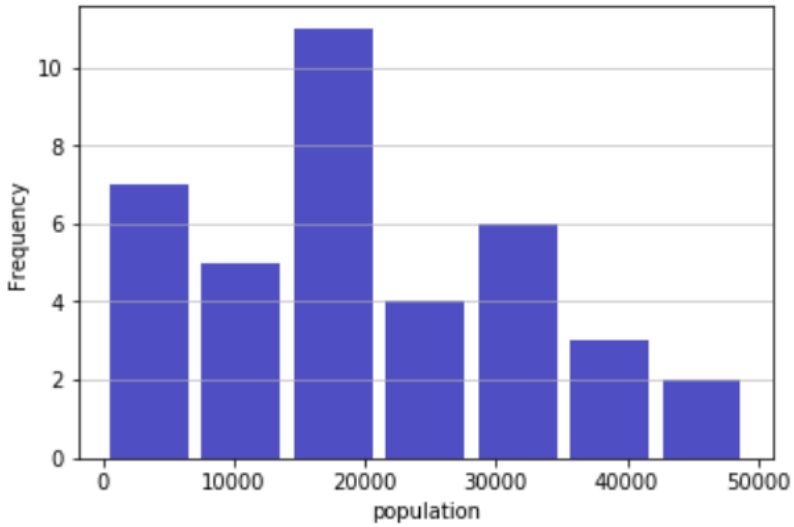
We'll start with histograms for each variable.



The above histogram plots the frequency of number of venues, we can see that the frequency of the number of venues is decreasing in general, but rise sharply at the right tail.

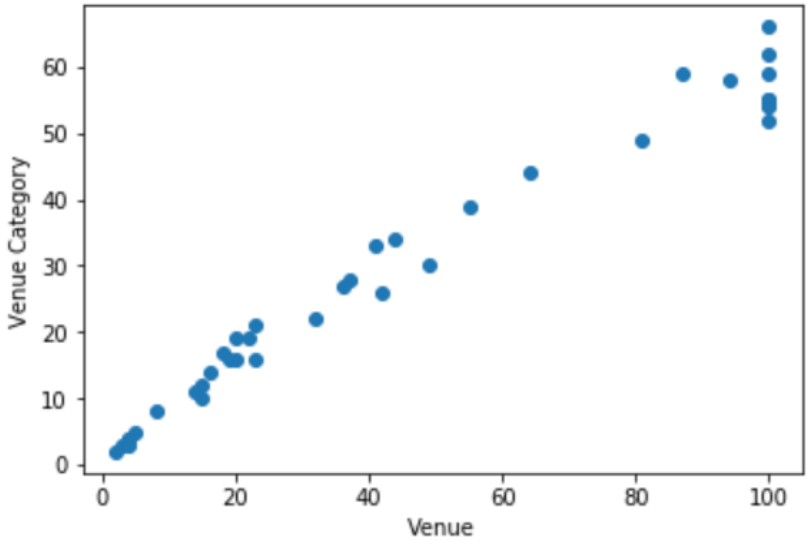


The above histogram plots frequency of venue categories. While this histogram is not as extreme as the venue histogram, we can also see the decreasing trend when the venue category number is smaller, and a sharp rise when the number is higher. This would explain the skewness and the high correlation that we calculated in the exploratory data analysis part of this analysis.



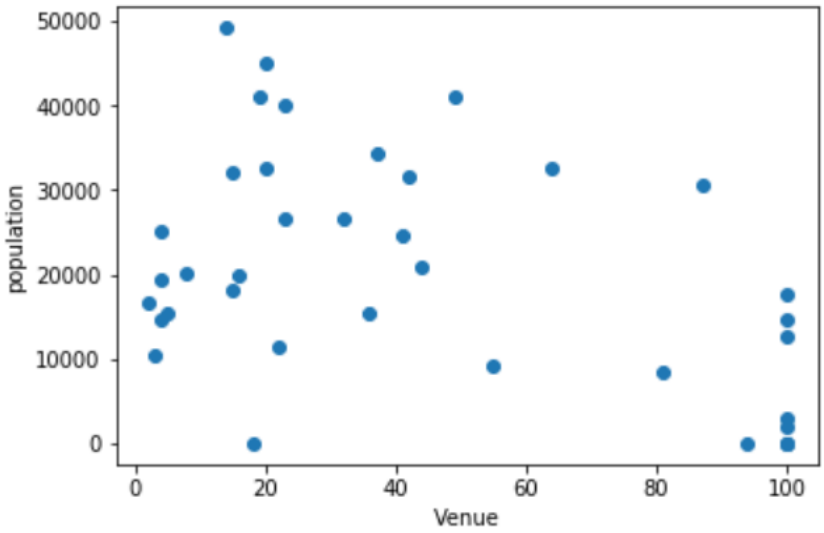
The above histogram plots the population frequency. This plot is much more like a normally distribution histogram compared to the venue and venue category histogram.

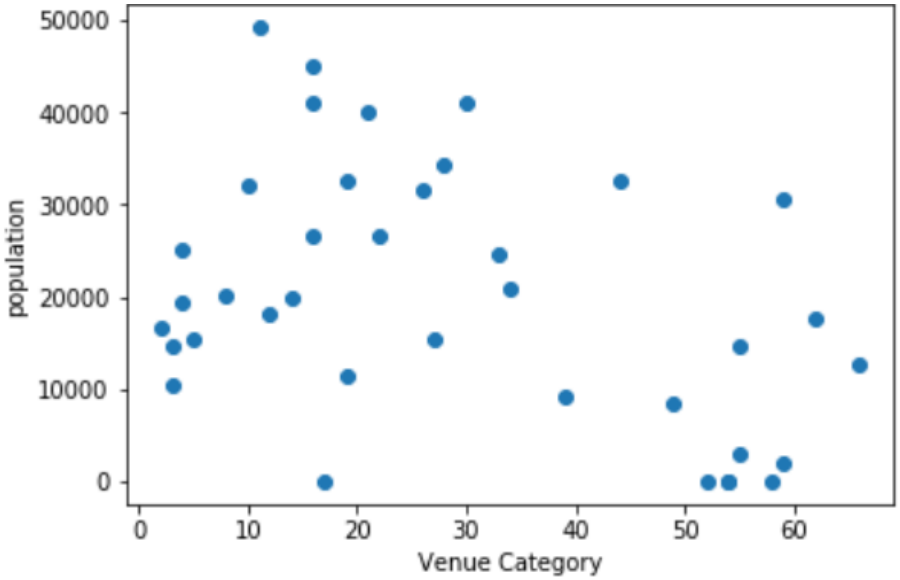
Next, we will take a look at how each of these variables are related to each other. If we plot venue against venue category as below, we can see a clear positive relationship between the two variables. As the number of venues increase, so do the number of venue category. This is consistent to the almost 1 correlation between these two variables.



The below two scatter plots are population plot against venue and venue category. The relationships are less obvious, there seem to be a weak negative correlation between population and the two independent variables. This makes a lot of sense

because we should expect population is not simply linearly related to venue number. It is a complicated metrics that can be influenced by a number of things.





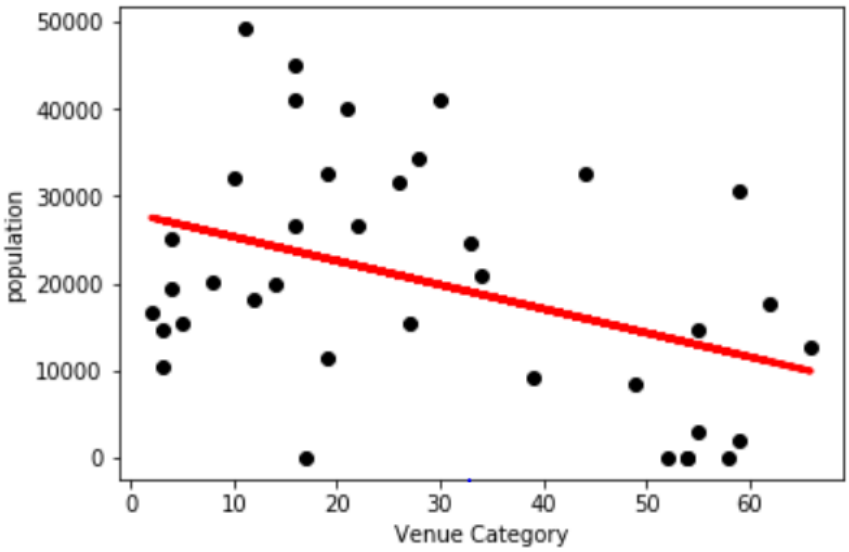
3.3 Predictive Modeling

In this section, we will build a predictive model for population against venue and venue category. As we have seen above, the data is quite simple and straightforward, and since we don't have a lot of data points, we'll use a simple linear regression to explore the relationship of population and venue.

In addition, since we don't want to really predict population based on venue and venue category, we just want to explore their relationships, this model is not really for predictive purposes. Thus we don't need to split the data to training and testing set. We'll simply run the model on the whole data and look at what their relationship is like.

Finally, since venue number and venue category numbers are highly correlated. We can just pick one of them to put in the model. If we use both of these independent variables, we will have a multicolliearity issue, which means we cannot trust the beta of either of these variables. We will use venue category since it has a slightly higher correlation to our dependent variable.

After preparing the data in the previous sections and the discussion above, we plug the data into a linear regression and have a beta of -274. This indicates that every increase in venue category will result in, on average, a decrease of 274 population in the postal code area. The mean squared error of the regression is 153657504.79, and the variance score is 0.17.



The above plot is the regression plot. As we can see, there is a negative relationship between the population and venue category. But the regression error is pretty huge as expected because population is a complicated metric that cannot be predicted using venue alone.

4. Discussion

This study gave a simple view on how population is related to venues in different postal code areas in Toronto. But we are limited by having too few data available and thus the model is simple. For further exploration, we can include more data, such as data in whole Canada, or even worldwide. In addition, with more data, we can do more complicated analysis such as neural network or random forest to further dissect this relationship. Finally, population is can be connected to a number of aspects of a location besides venues. This could another direction worth exploring.

5. Conclusion

In this study, I explored the relationship between population in different postal code areas in Toronto and the venue data in the same postal code area. There is a negative relationship between population and venue numbers. I built a linear regression model to accomplish and illustrate this goal.