Exploring relationship between venues and population in Toronto

Zhidi Luo

Feb 19, 2019

1.1 Introduction: 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 Introduction: 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.

2.1 Data: 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: Data Pre-processing

- 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.

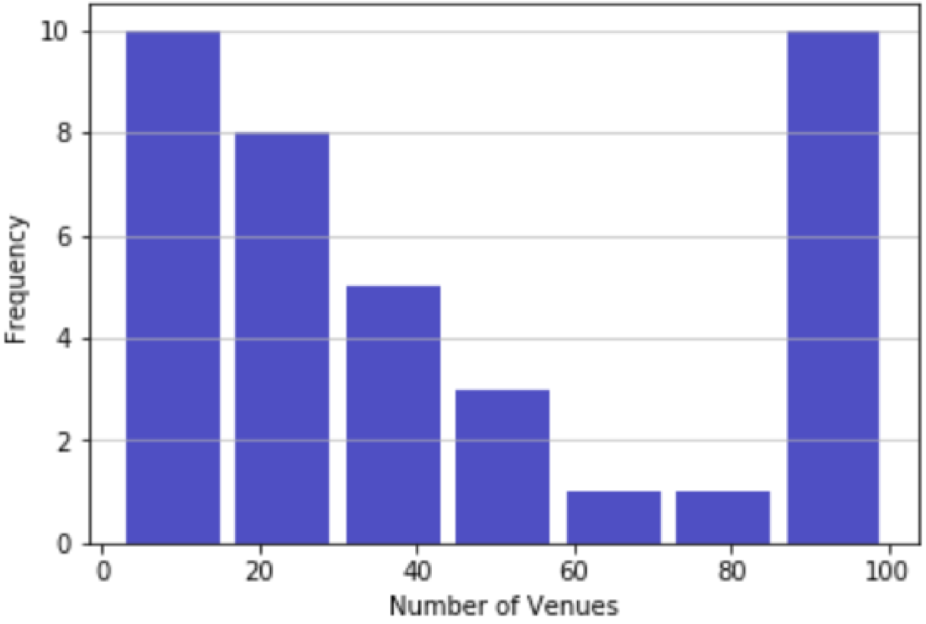
3.1 Methodology and Results: Exploratory Data Analysis

	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

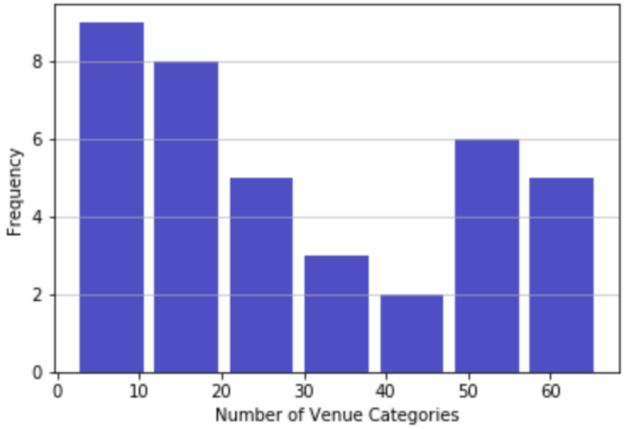
3.1 Methodology and Results: Exploratory Data Analysis

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

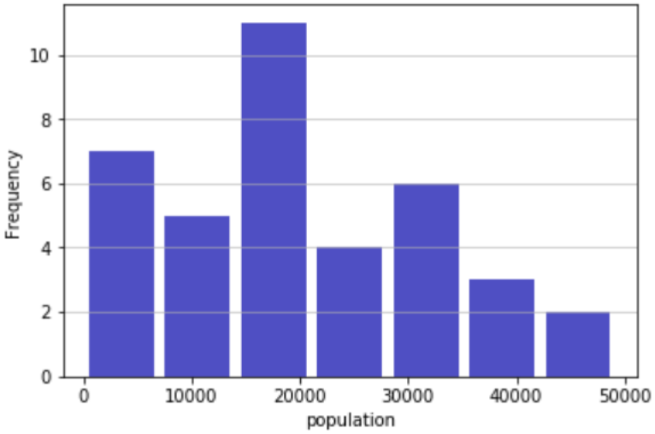
3.2 Methodology and Results: Data Visualization



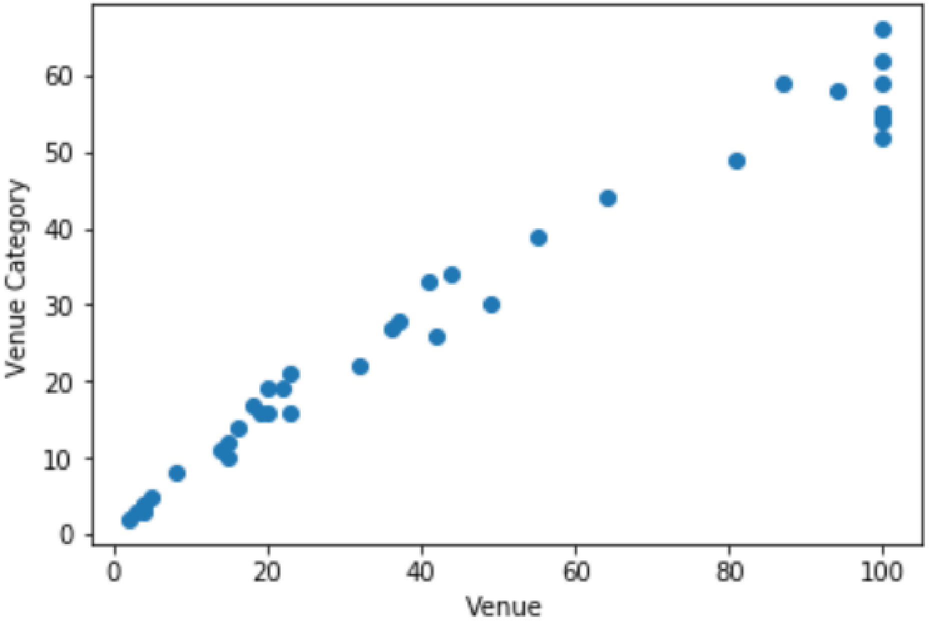
3.2 Methodology and Results: Data Visualization



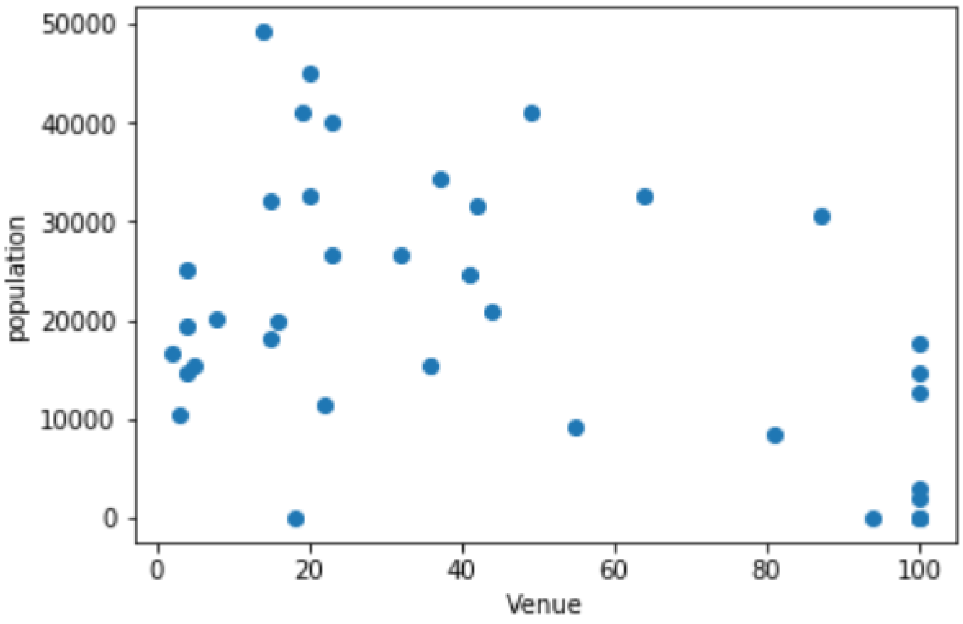
3.2 Methodology and Results: Data Visualization



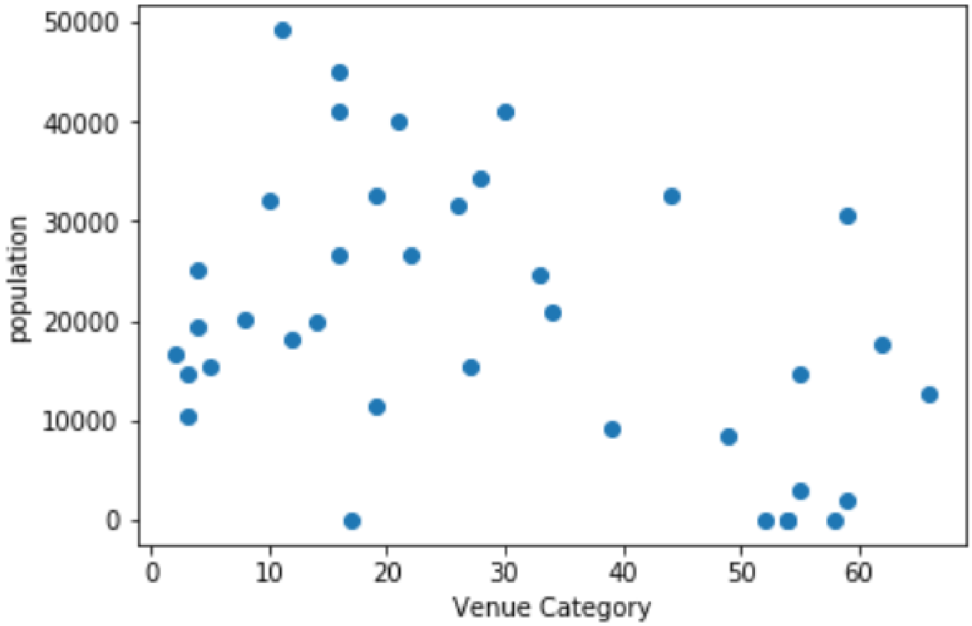
3.2 Methodology and Results: Data Visualization



3.2 Methodology and Results: Data Visualization



3.2 Methodology and Results: Data Visualization



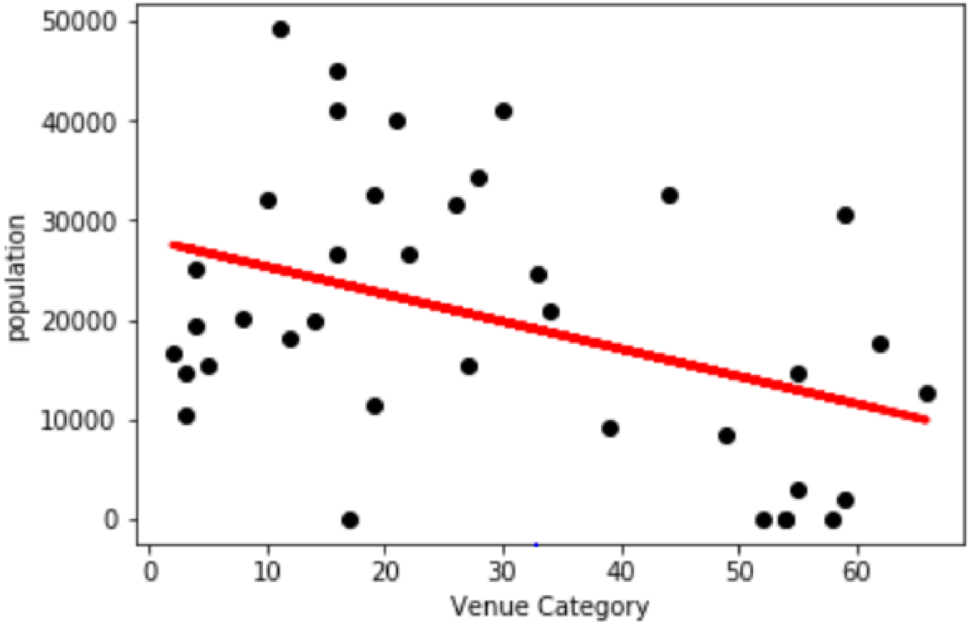
3.3 Methodology and Results: Predictive Modeling

- 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 multicollieарity 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.

3.3 Methodology and Results: Predictive Modeling

- 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.

3.3 Methodology and Results: Predictive Modeling



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.