Predicting Service Demand in a City

Introduction

A pattern in the growth of services emerges with the growth of a city or neighborhood, affected along with its cultural and societal values. Kuhnert et al. (2006) had displayed examples of these patterns such as the logarithmic increase of the services such as petrol stations, pharmacies, clinics etc. Geoffrey West in his book called "Scale", has pointed out that in organism-like structures such as animals, cities and companies, there seems to be an optimal curve in relation to its size wherein values above this curve are organisms pushing its limits, and below the curve are organisms underutilizing its resources. With this line of thinking, it can be said that there must be a baseline in which it can be determined if a city or neighborhood still lacks the necessary services.

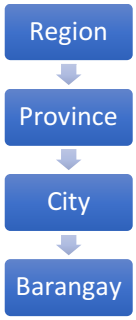
Putting up a business and wondering which type of service is lacking in a certain neighborhood/city is surely a nerve-wracking experience especially for first-time entrepreneurs. Questions such as "Should I put up a laundry service?" or "Is there demand for an Asian restaurant in this neighborhood?" are some motivational examples. If information is available for these entrepreneurs, they will have an idea which of the types of services that are in short demand or in over-supply. In turn, they can have an easier time to weigh the pros and cons and scope into businesses that can potentially be profitable.

Knowing which businesses are appropriate for a certain location/neighborhood are not only useful for entrepreneurs, but also for city governments. Having insight on which neighborhoods could use more or less of a certain type of service can be helpful in careful city planning. Also, questions about what it takes for a neighborhood/city to improve are also of interest. "Does a city need more roads?" or "Does a city need more connectivity to support the number of services/establishments?" are just some of the questions that might be crucial to answer.

Data Desrcription

Overview

In the Philippines, the usual government unit structure is shown below:



Since there is more interest in a more granular view, the ideal level to investigate is at the *Barangay* level which is the most basic government unit in the Philippines. But due to memory constraints and location data constraints, this analysis will be limited only at the City level.

Location

For this analysis, location data in Foursquare will be collected using the Foursquare location API. This will include primarily the category of a certain venue/service established in a city. As a start, the analysis will scope into the Hotel business. More types can be incorporated later on.

Other important fields

Population information about these barangays are important to obtain as this was identified as an important variable in establishment count growth.

Other-related fields that may be useful to describe how connected a city is are the total land mass of the City/Municipality, the number of roads, and the total road length. Latitude and longitude data was also obtained for querying.

Response

To know if a certain city has an oversupply/undersupply of a certain service, the response to be used for this analysis are the number of establishments in a city for a venue category. But since only a portion of the total number of establishments can be queried and can be known in Foursquare, the proportion of establishments of a certain category will be used instead.

Methods

City Information Dataset Preparation

Population and land area data on the cities and municipalities was obtained from a Wikipedia page site[1] . Relevant latitude and longitude data for the Philippine cities and municipalities was obtained from a public github repository[2].

Venues dataset Preparation

Establishments were determined for each city using the Foursquare API. The radius to search on was determined by assuming a circular range from the available latitude and longitude data and using the total area of the City. Due to querying limits, the number of establishments was limited to 500. The Venue category, name and longitude and latitude was obtained from the query.

Statistical Modeling

There is a need to make sure that venue categories to be modelled are well-populated in Foursquare to avoid non-response biases. For example, it maybe possible that bakeries are not

[1] https://en.wikipedia.org/wiki/List_of_cities_and_municipalities_in_the_Philippines
[2] https://github.com/zymartinez/philippines-lat-lng-list.

well-reported due to the fact that not a lot of people review it. Thus, for this proof-of-concept, the venue category **Hotel** was chosen. Furthermore, population and population density (population/land area) were used as features to model the proportion of establishments in a city.

Model evaluation was done using the root-mean squared error. Three variations of the features were tested – (model 1) multiple linear regression (no transformations), (model 2) log population, and (model 3) log population and log population density.

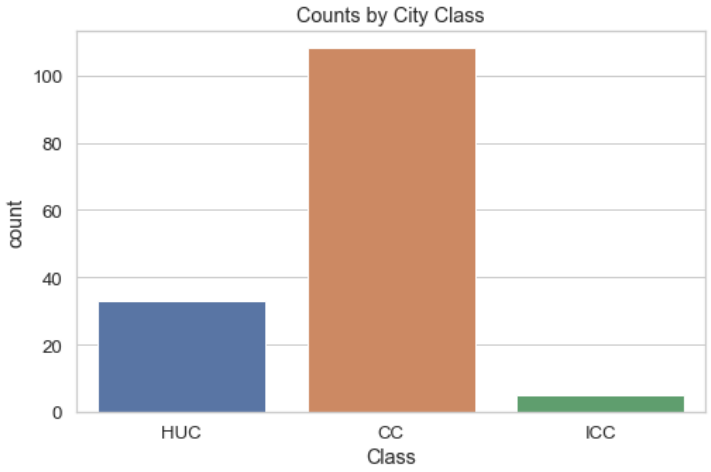
Results

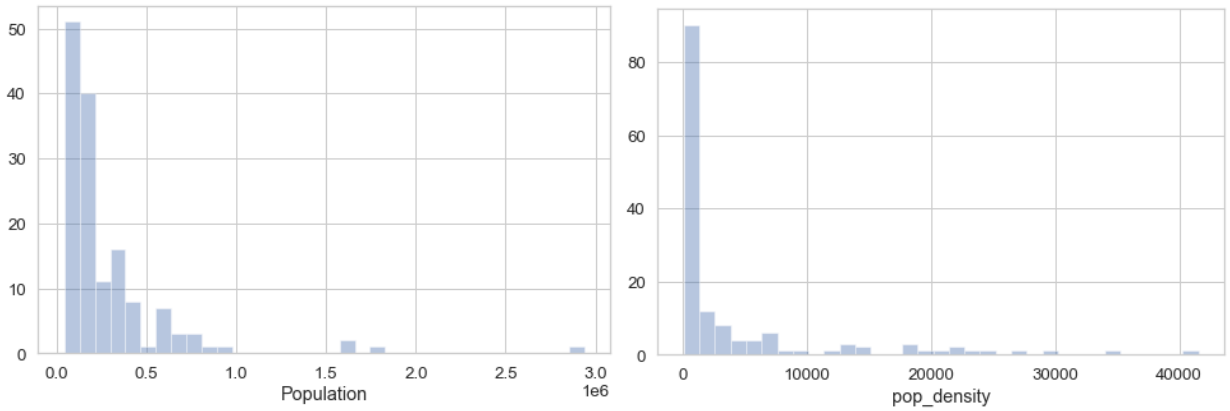
Data Preparation

Population table scraped from Wikipedia with filters on city data contained 146 rows, pertaining to cities. Joining with latitude and longitude data reduced the number of rows to 110 due to missing information. Determining establishment data further reduced the number to 104 cities due to the API not working for some cities.

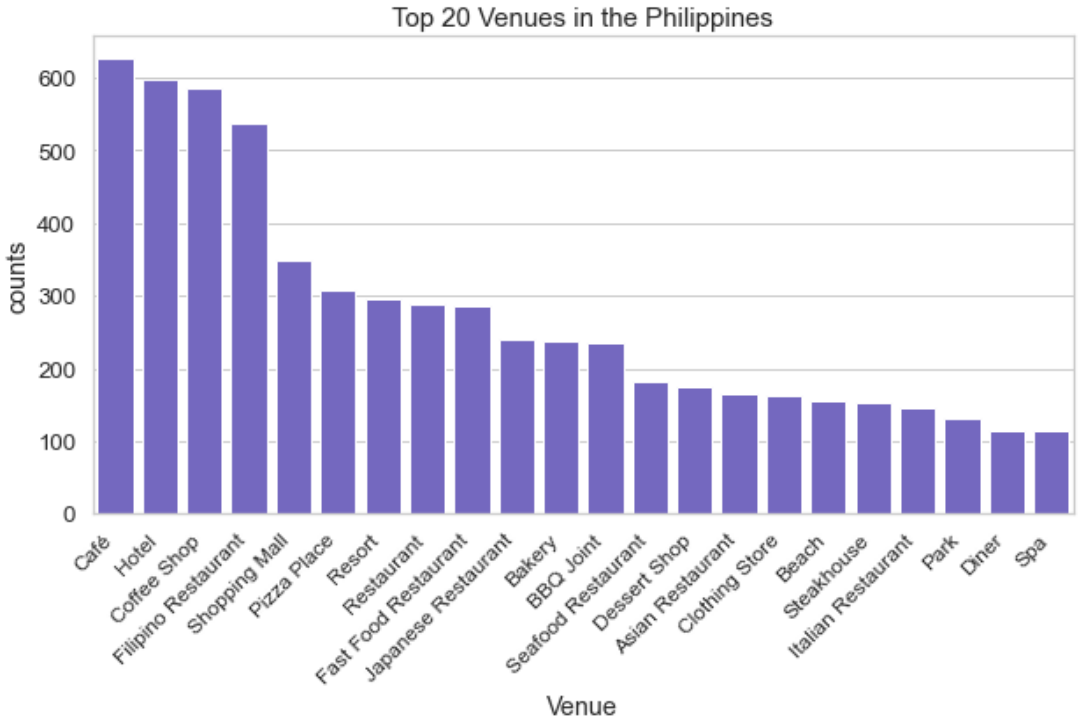
Data Exploration

Show below are distributions of City Class, Population and population density.





A total of 215 venue categories were scraped from a total of 104 cities. Shown below are the top 20 venues in terms of counts.



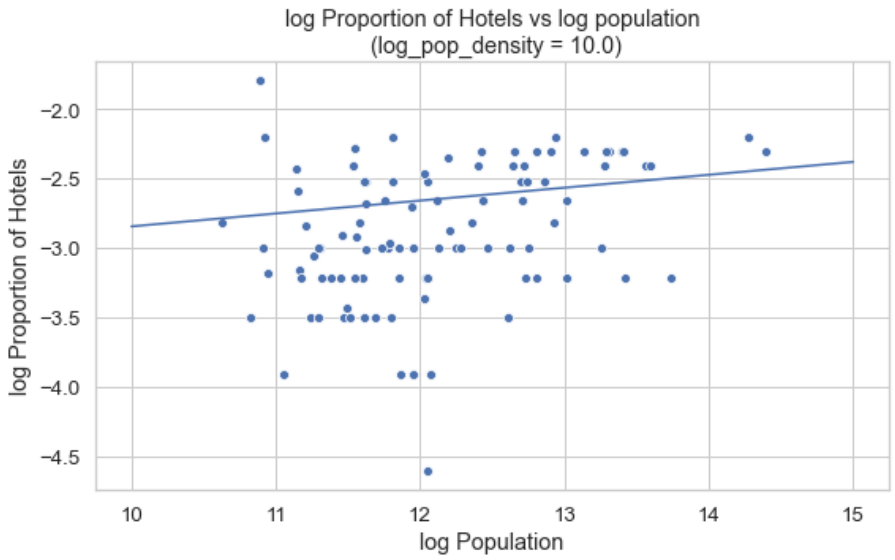
Modeling of proportions of hotels in a city

The response that was used for modeling was the log proportion. It can be observed below that the third model utilizing log-population and log-population density had the lowest RMSE. This will be used for further analyses. The full model has the form:

$$\log(proportion) = 0.093 \times \log(Population) + 0.083 \times \log(Population\ Density)$$

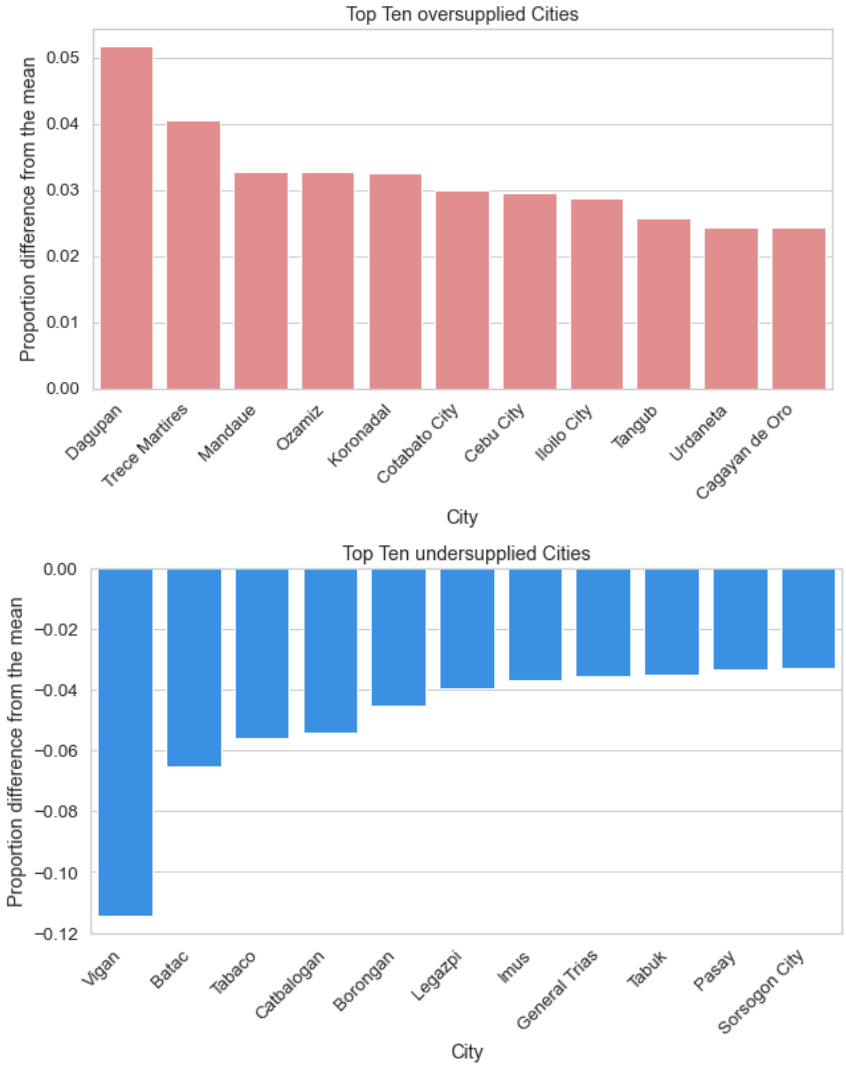
Model	Root mean squared error
Multiple Linear Regression (no transformation)	0.457
Log-population	0.455
Log-population; log-density	0.447

To check the fit, a scatter plot of the actual values with the fitted line of the model at the average log population density was created.



Proportion Differences relative to the mean

Using the fitted linear model, differences in proportions from the mean of the regression line (the predicted value) was determined. Shown below are the top ten oversupplied and undersupplied cities. Oversupply means that the actual proportion of hotels are greater than the predicted proportion. On the other hand, undersupply means that the actual proportion of hotels in a city are below the predicted proportion.



Discussion

A proof-of-concept modeling of establishment proportions in a city was done in this analysis. From various literature, population was deemed to be an important factor which relates to the growth of establishments. Population density on the other hand, can serve as a variable to represent possible connectivity in a city, although this may be better represented by features concerning road/transpiration networks.

The modelled regression line as estimated in model 3 shows a baseline curve to which “ideally” a city with such population and density will expect to have in terms of the number/proportion of establishments. Points below the curve represent cities that can possibly still accommodate for more establishments of a certain kind. Possibly, it can also mean that these cities are still in need of such establishment to support the population. On the other, points above the curve represent cities that are over the limit of the proportion of establishments that can be handled/needed by

the population. These explanations, as reasonable as they are, cannot be explained reasonably in the current analysis.

The model predicts the top five oversupplied cities in terms of hotels: Dagupan, Trece Martires, Mandaue, Ozamiz and Koronadal. While the top five undersupplied cities are as follows: Vigan, Batac, Tabaco, Catbalogan and Borongan with proportion differences from the baseline ranging from 5%-12%. The former five cities are possible locations where setting up a business might not be a good idea. The latter five cities represent potential cities to start up a hotel business due to a negative discrepancy between the expected proportions to the actual.

These results can also be taken in another way. The oversupplied cities might have qualities that could have helped the city support a disproportionate number of establishments. This may be due to good governance, which may in turn be reflected in the number of networks supporting this industry. Undersupplied cities on the other hand, are cities which need to improve its infrastructure to support such industry.

This analysis could benefit in the addition of the detailed description of road/transport networks in a city. Although the population and the population density could somehow be correlated with transport structure, there is high variation among similarly sized cities especially across countries. Furthermore, other types of establishments can also be explored, especially those that have minimal non-response bias. Lastly, there can also be merit in looking into a more comprehensive and granular dataset, which can further refine model accuracy and utility.

Conclusion

This analysis demonstrated a method to identify cities that are undersupplied or oversupplied in relation to the baseline, as estimated by a regression curve relating proportion of establishments and city characteristics. This was shown by obtaining the proportion difference between predicted and actual proportions, where positive differences mean oversupply while negative differences mean undersupply.