Alex Yeung

10/04/2020

Abstract

Investigation in

Battle of the Neighborhood

Capstone Project

Contents

[1.0 Introduction 2](#_Toc37516582)

[1.1 Background 2](#_Toc37516583)

[1.2 The Business Problem 2](#_Toc37516584)

[1.2.1 Criteria 2](#_Toc37516585)

[1.3 Interest 2](#_Toc37516586)

[2. Data 3](#_Toc37516587)

[2.1 Data Source 3](#_Toc37516588)

[2.2 Data Cleansing and Manipulation 3](#_Toc37516589)

[2.2.1 Columns Drop 3](#_Toc37516590)

[2.2.2 Spelling of Ward and District 3](#_Toc37516591)

[2.3.3 Rows Drop 4](#_Toc37516592)

[Methodology 5](#_Toc37516593)

[Exploratory Data Analysis – Ward Population Data 5](#_Toc37516594)

# 1.0 Introduction

## 1.1 Background

An international fashion chain, XYZ Limited has decided to open their first store in Asia and is considering Tokyo as a potential candidate.

Being the capital of Japan with a local population of 13.9 million people and an annual 31.2 million foreign tourist in 2019 the city of Tokyo has a very large potential customer base, on top of this, the Olympics is also scheduled to be held in Tokyo with a further expectation of an additional 10 million tourists and $283 billion boost to the Japanese Economy. XYZ Limited want to establish the store as soon as possible to capture this opportunity.

## 1.2 The Business Problem

To gather information from internet sources and investigate which of the 23 wards of Tokyo would be ideal to open a fashion retail store in the ward’s main district based on local population and tourist population as main selection criteria, high population density indicates strong potential customer base and foot traffic hence maximise profit.

In choosing the ward and it’s associated main district, existing shops and business demographic of the district should also be considered as the stakeholder do not wish to open the store in a district with a lack of established venues.

A set of criteria has been provided by the stakeholder as guidelines in conducting this study

### 1.2.1 Criteria

The criteria provided by the stakeholder:

1. The store location should consider the potential tourist as well as local population in the area
2. The store location should be located in a established leisure suburb where there is a number of existing clothing stores, cafe and restaurants.
3. The weighting in decision making when deciding the district should be:

0.4 (Density score) + 0.7 (Airbnb Count score) + 0.2 (Population score)

## 1.3 Interest

The sponsor of this project is XYZ Limited and the stakeholder is the General Manager Business Strategies. The stakeholder has a high interest in maximizing the return of this investment.

# 2. Data

## 2.1 Data Source

The sponsor has provided no data to support this study. In identifying the information required to solve this problem, I will need data on the population within the wards of Tokyo which is available from Wikipedia with data at date of 2016. Useful data fields obtained:

|  |  |
| --- | --- |
| Field | Description |
| Name | Name of ward |
| Population | Population of the ward as of October 2016 |
| Density | Population density of the ward |
| Area | Size of the ward in km^2 |
| Major District | A list of the district within the ward |

Tokyo ward and district population demographic information can be acquired at: <https://en.wikipedia.org/wiki/Special_wards_of_Tokyo>.

I will also need information on the tourist population in Tokyo, as there are no datasets found detailing this, I have used Tokyo Airbnb data found on Kaggle with data at date of 2019 as a substitute to represent this information.

|  |  |
| --- | --- |
| Field | Description |
| neighbourhood | Which district the listing is located |
| latitude | Listing’s latitude |
| longitude | Listing’s longitude |

Airbnb Tokyo data can be acquired at:  
<https://www.kaggle.com/fuyutaro/tokyo-airbnb-open-data>.

With the combination of data provided by the two sources, I am able to conduct the analysis to investigate the criteria set out by the stakeholder.

## 2.2 Data Cleansing and Manipulation

As the information is scrapped from the internet from multiple sources, it required cleaning and merging into one data frame.

### 2.2.1 Columns Drop

Both data sets obtained contains columns that I did not require for this analysis, I have dropped the columns that will not be used in analysis.

### 2.2.2 Spelling of Ward and District

The spelling of the ward name in the raw source files contain variance which need to be fixed to eb used as key when joining the two data frames together.

## 2.3.3 Rows Drop

Datasets contain rows that needs to be dropped as they will affect the analysis of data.

In particular to the Tokyo Ward dataset, overall or grand total row that has been scrapped into the data frame. In particular to the Tokyo Airbnb dataset, rows with no value in key analysis columns are dropped.

# 3.0 Methodology

In analysing the data from multiple sources, I have decided to first look at the relationship within the datasets independent to the other before joining the datasets together and analysing the full dataset.

To satisfy the criteria of maximising customer bases for local population, I first look at the trends within the Tokyo Ward dataset, this showed which ward had the highest population and density which provided insights into which ward would provide a strong local population.

The second step I looked at the Tokyo Airbnb data to understand which ward has the highest number of Airbnb count, this will indicate which ward will best satisfy the criteria for tourist population.

For the third step, I joined the two datasets together and analysed the all the key variables. Here I used k-means cluster to group the wards based on similarity in population. I also normalised the data and applied the weighting formula to identify the top 3 wards which fits the criteria best to focus on in the Foursquare locational data analysis.

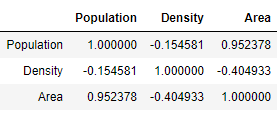
The next step, I have refined the focus from wards to the main district for the selected wards. I use the Foursquare API to obtain the top 100 venues for our selected three districts to understand the culture of the selected ward and its business demographics, all three select districts showed a strong establishment of businesses.

The last step I used Foursquare API again to look at Fashion businesses specifically within each district to find out which district has the hight number of existing fashion business hence indicating that is a popular shopping suburb.

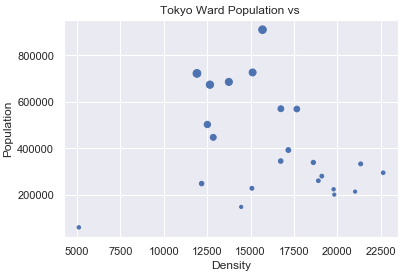
The chart below summarises the process flow:

## 3.1 Exploratory Data Analysis – Local Population Data

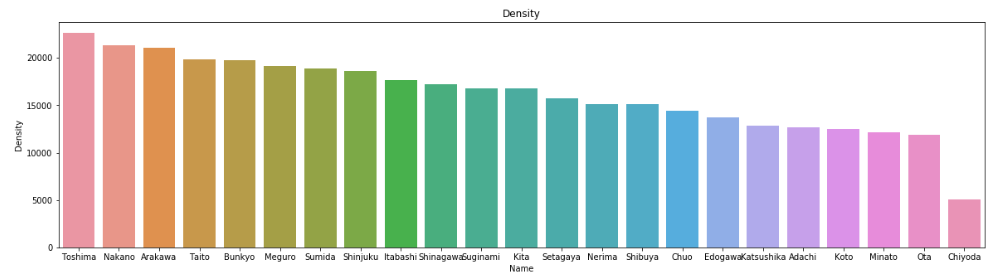
In this section, I looked at the Tokyo data specifically. The correlation table and scatter plot of the key variables do not show a strong relationship, this result was not as expected as the density variable showed low correlation to the other two variables. The variable density is of most interest as it represents the highest concentration of potential customers.



Visualizing the relationship in a scatter plot showed not apparent trend although I can clearly see there is a big difference.



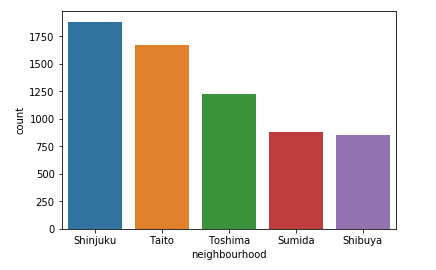
Plotting the density in a bar plot in descending order, I can see that the wards of Toshima, Nakano, Arakawa, Taito and Bunkyo have the highest density.



## 3.2 Exploratory Data Analysis – Tourist Population Data

From the Kaggle I obtained the 2019 data on Tokyo Airbnb, this will be an approximation to potential tourist the ward can provide.

After cleansing the data, I grouped the by the Airbnb listing’s associated ward to summarise the count of listings. The chart below shows the result.



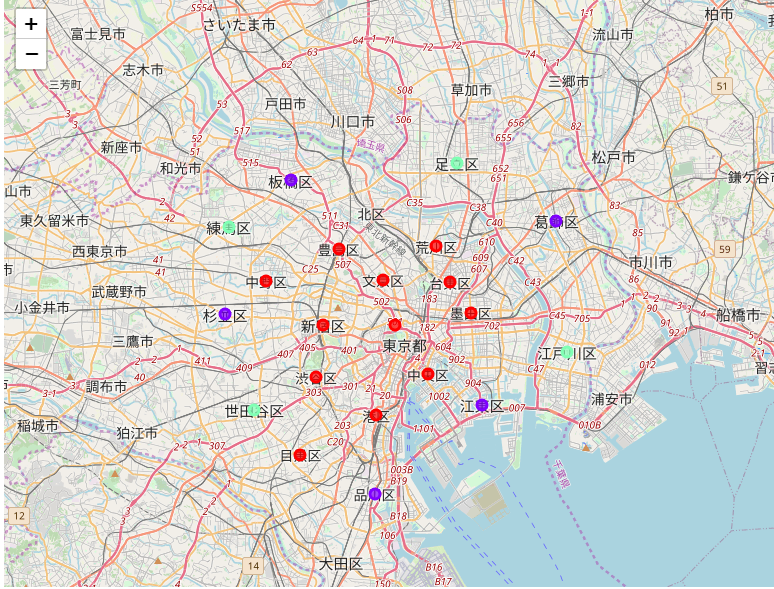
The ward of Shinjuku is ranked top in terms of tourist potential however the wards of Toshima and Taito are in the top 5 in both the tourist and local population analysis.

## 3.3 Exploratory Data Analysis – All Data

Merging the two dataset together on name of the ward, I performed a cluster analysis on the full dataset then I normalised the data and applied the weighting from the stakeholder to identify the top 3 wards.

### 3.3.1 Cluster Analysis

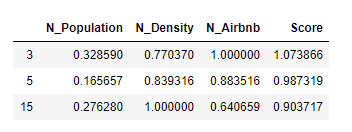
The results from performing cluster analysis using k-means with a k value of 3 shows the result below, the wards to be selected for locational data analysis should be of the same cluster.



### 3.3.2 Choosing Three Wards

Normalising the data provides a common scale on the dataset to apply the weighting.





The top 3 score from the weight are index 3,5 and 15 which associate to Shinjuku, Taito and Toshima.

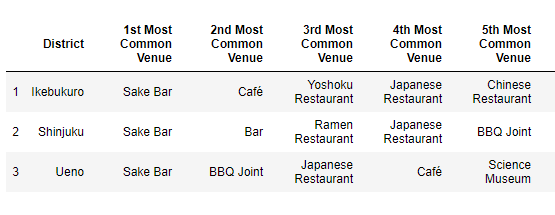


## 3.4 Foursquare Location Data – All Venue

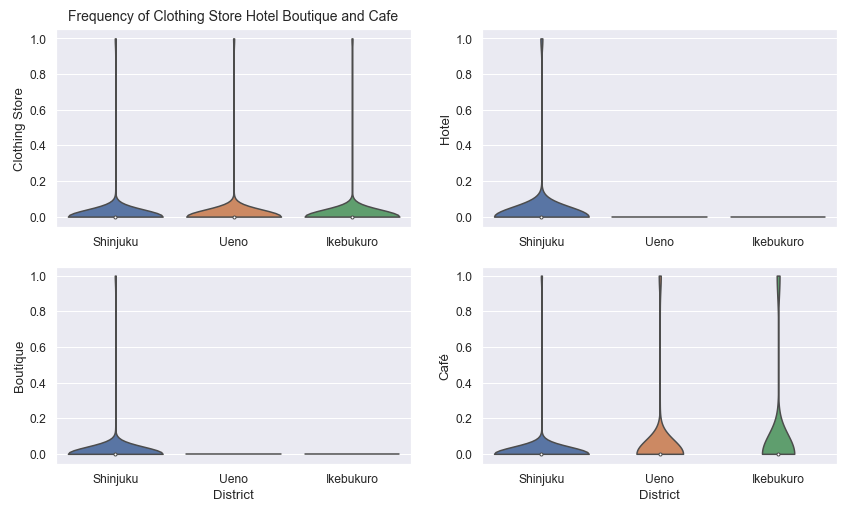
Using Foursquare API, I obtained a list of venue of each ward’s main district, Here I have switched the analysis from looking at wards to the main district within each ward.

|  |  |
| --- | --- |
| Ward | District |
| Shinjuku | Shinjuku |
| Taito | Ueno |
| Toshima | Ikebukuro |

Applying One Hot coding and summarizing the frequency of each venue category, all district show a good range of existing entertainment venues.

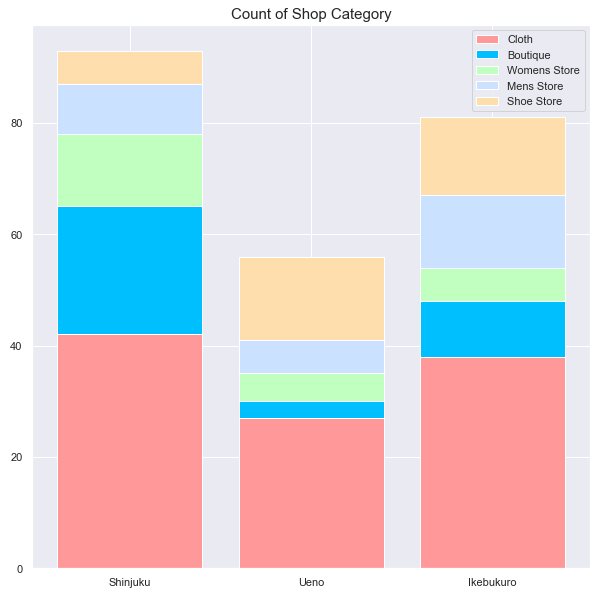


Looking at frequency of key business relating to the fashion and population, I discover that both Ueno and Ikebukuro lacks hotel which is not ideal in consideration of the criteria for tourist population.



## 3.5 Fashion Venues Within Selected District

A separate call to the Foursquare API with only venues relating to fashion show both Ueno and Ikebukuro has less fashion related shops as the limit of the call was 100 and only Shinjuku can reach the max limit of the call.

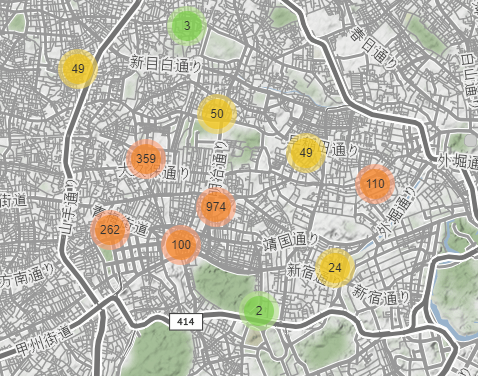


# 4.0 Discussion

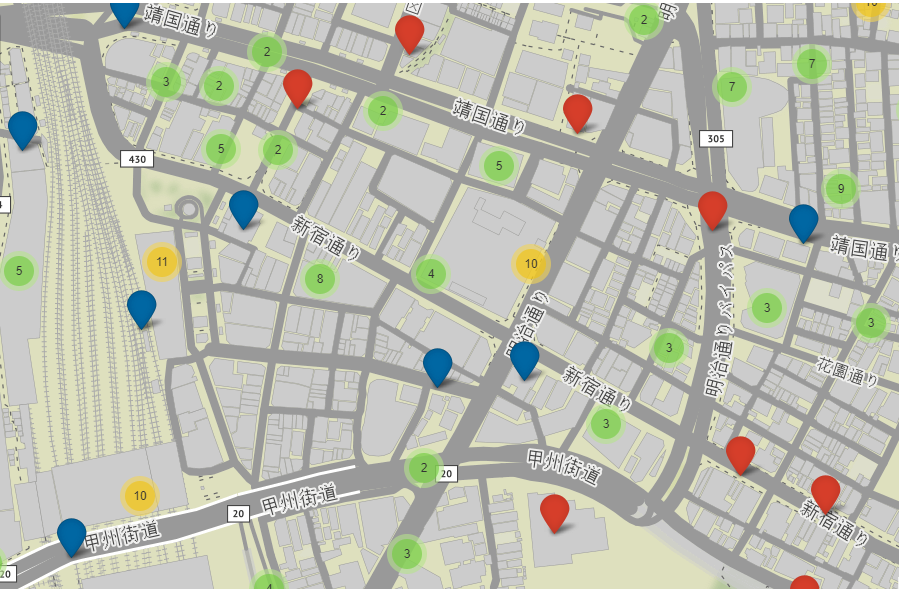
## 4.1 Results

The score of Shinjuku against the weighting formula provided is the highest score at 1.07 in section 3.3 of this report, the analysis into existing business in section 3.4 shows this district has good established entertainment venues as well as an advantage over Ueno and Ikebukuro in hotel. Section 3.5 also showed there is a higher number of fashion related stores in comparison to the other two district, there fore Shinjuku would be the best district for opening the new fashion store.

The map below shows the location within Shinjuku district where highest concentration of location of interests are:



The map below shows the blocks where a good mix of existing businesses and Airbnb exist with the blue marker indicating a business and the red marker indicating an Airbnb.



## 4.2 Limitations

The limitation of 100 returned venue listing when calling the Foursquare API could affect the accuracy of this analysis.

# 5.0 Conclusion

In the event of XYZ Limited opening the district I recommend is Shinjuku as it provides the best mix of potential customer from local and tourist population and it consist of a good range of entertainment venues with a good establishment as a shopping district for fashion.