A

Data Science

Project Report

On

The Battle of Neighborhoods

by IBM-Coursera

Topic: Top wards in Tokyo to open a Restaurant

Table of Contents:

INTRODUCTION

In this project we will try to find an optimal location for a restaurant. Specifically, this report will be targeted to stakeholders interested in opening a restaurant in Top 5 busiest wards of Tokyo.

Tokyo is the most populous metropolitan area in the world. Currently ranked 3rd in the global economic power index, Tokyo is definitely one of the best places to start up a new business. During the daytime, especially in the morning and lunch hours, office areas provide huge opportunities for restaurants. Reasonably priced, shops are usually always full during the lunch hours and, given this scenario, we will go through the benefits and pitfalls of opening a breakfast cum lunch restaurant in highly dense office places. Usually the profit margin for a decent restaurant lie within 15−20% range but, it can even go high enough to 35%. The core of Tokyo is made of 23 wards (municipalities) but, we will concentrate on 5 busiest business wards of Tokyo to target daily office workers.

We will use our data science powers to generate the best promising criteria we could so that best possible final location can be chosen by stakeholders to start or expand their business.

For Business personnel who wants to invest or open a restaurant, this analysis will be a comprehensive guide to start or expand restaurants targeting the large pool of office workers in Tokyo during lunch hours.

For Freelancer who loves to have their own restaurant as a side business, this analysis will give an idea, how beneficial it is to open a restaurant and what are the pros and cons of this business.

Also new graduates can find reasonable lunch/breakfast place close to office.

DATA

Based on definition of our problem, factors that will influence our decision are:

- average land price
- venue
- distance

We will be using appropriate tools to differentiate and choose among the most appropriate location for the investors or stakeholders to open or invest in a restaurant making a good profit out of it.

We will be using the foursquare API to generate location details and wikipedia files to load data and deal with latitude and longitude coordinates. Further with the help of libraries and their in-built functions we will generate the list of top 5 busiest wards which will make it easier for stakeholders to invest or open a restaurant at a good location.

Data Preparation:

I have made use of Special Wards of Tokyo page from Wiki to scrap the table to create a data-frame. For this, I have used the required libraries and then have created a data-frame containing name of 23 wards of Tokyo, Area, population and 1st Major District.

After manipulation and using the required libraries like pandas the final output or the data frame will look somewhat like the one shown in the figure given below:

Data Manipulation:

After this I have used the libraries to obtain the coordinates of some major districts in wards of Tokyo and now by using the land market value area in Tokyo web page, we will be determining the land price in various districts.

After this by the use of foursquare API the location details are determined which are further mapped using the existing data to obtain the final output or the desired table required from which we can select the top 5 busiest wards or location to open or expand a restaurant generating high revenues or profits.

After manipulating the data according to the rules and regulations and as per our requirements the final output will somewhat look like the one shown below in the figure.

METHODOLOGY

We know that to implement any idea there are a lot of ways to do. One can go with many desirable options or ways to implement a certain thing or say a project.

Our project can also be implemented using various ways but for simplicity we have here used the python concepts and libraries defined for machine learning algorithms such as clustering, classification, etc. Here we have used python as it is an easy to code language and mostly used for data science projects to work with because of its versatile nature.

Here we have used the famous unsupervised machine learning algorithm that is Clustering to implement the scenario and the project in an efficient and easy way.

Clustering:

Clustering is a Machine Learning technique that involves the grouping of data points. Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group. In theory, data points that are in the same

group should have similar properties and/or features, while data points in different groups should have highly dissimilar properties and/or features. Clustering is a method of unsupervised learning and is a common technique for statistical data analysis used in many fields.

In Data Science, we can use clustering analysis to gain some valuable insights from our data by seeing what groups the data points fall into when we apply a clustering algorithm.

There are various clustering algorithms used to implement the technique. One such famous and most widely used technique is k-means and the same in being used in this project as well.

k-means Clustering Algorithm:

K-Means is probably the most well-known clustering algorithm. It's taught in a lot of introductory data science and machine learning classes. It's easy to understand and implement in code.

To begin, we first select a number of classes/groups to use and randomly initialize their respective center points. To figure out the number of classes to use, it's good to take a quick look at the data and try to identify any distinct groupings. The center points are vectors of the same length as each data point vector.

Each data point is classified by computing the distance between that point and each group center, and then classifying the point to be in the group whose center is closest to it.

Based on these classified points, we recompute the group center by taking the mean of all the vectors in the group. We repeat these steps for a set number of iterations or until the group centers don't change much between iterations.

Now after applying the required algorithm and manipulating data we will form the clusters which we can be shown with either graphs or maps as per requirements. Here we have used map to show the locations and clusters. The gist of how the final output may like is given in the figure below. We can use various libraries like folium to plot the map.

Major districts of Tokyo segmented into 3 clusters based on the most common venues are shown in the figure.

ANALYSIS AND VISUALIZATION

Visualization:

Folium is a python library that can create interactive leaflet map using coordinate data. Since I am interested in restaurants as popular spots firstly I created a data-frame where the venue category column in previous data-frame contains the word Restaurant and the with the coding and use of the folium library to create maps accordingly as per the requirements the output somewhat looked the way as shown in the figure below.

Circular marks represent the most frequently visited restaurants in the 5 Major (Nihombashi- Green, Nagatacho- Red, Shibuya- Orange, Shinjuku- Magenta, Shinagawa- Blue) districts of Tokyo, according to Foursquare data.

Analysis

From the several data-frames that we created for exploratory data analysis, using one of them, we have plotted which district has restaurants among the most frequently visited places and, Nagatacho of Chiyoda ward comes on top with 56 restaurants. The figure explains more efficiently the analysis work.

The analysis and visualization are an important part of any project as with the help of these two things one can be sure of what one has created with a confidence of project working well and not failing even in bad situations. It somewhat gives an acceptable solution yet.

RESULTS AND DISCUSSION

We reached at the end of the analysis, where we successfully got the 5 major wards of Tokyo and, as the business problem started with benefits and drawbacks of opening a lunch restaurant in one of the busiest districts, the data exploration in most case gave accurate results. I used data from web resources like Wikipedia, python libraries like Geopy, and Foursquare API, to set up a very realistic data-analysis scenario.

Nagatacho district in Chiyoda ward and Nihombashi in Chuo ward are dominated by restaurants as the the most common venue whereas Shibuya and Shinjuku areas are dominated by bars, pubs, and cafe as most common venues. Nagatacho has maximum number of restaurants as the most common venue whereas has Shibuya area has the least.

Since the clustering was based only on the most common venues of each district, Shinjuku, Shibuya fall under the same cluster and, Nagatacho, Nihonbashi fall under another cluster. Shinagawa is separated from both of these clusters as convenient stores stand out as the most common venue with a very high frequency. It certainly gives us some very important preliminary information on possibilities of opening restaurants around the major districts of Tokyo. One pitfall of this analysis could be consideration of only one major district of each ward of Tokyo, taking into account of all the areas under the 5 major wards would give us an even more realistic picture.

According to this analysis, Shinagawa area will provide least competition for an upcoming lunch restaurant as convenience store is the most common venue in this area and, the frequency of restaurants as common venue are very low compared to the remaining districts.

CONCLUSION

Finally, to conclude this project, we have got a small glimpse of how real-life data-science projects look like. I’ve made use of some frequently used python libraries to scrap web-data, use Foursquare API to explore the major districts of Tokyo and saw the results of segmentation of districts using Folium leaflet map. Potential for this kind of analysis in a real-life business problem is really a great thing. Also, some of the drawbacks and chance for improvements to represent even more realistic pictures are mentioned. Finally, with the help of appropriate tools and knowledge it was fun to work on this project and generate the desired outputs.