**Finding out Good Places to Open the New Restaurant**

**Dong-yeong Kim**

**Jan 18, 2021**

**1. Introduction**

**1.1 Background**

In my country, South Korea, self-employment is more common way than other countries. In fact, 25% of Korean are self-employed, and this situation is largely involuntarily because of cut-back and employment difficulties. You can find that the rate is very high, compared to other countries such as Japan(10%), US(6%), and OECD average(15%).

**1.2 Problem**

A lot of the self-employed in South Korea prefer to open restaurant or cafe, but in so many cases, they start up without the know-how to make their business prosper. So, in this project, I try to find out a good location, especially near some station, to open the new restaurant in the metropolitan area of South Korea.

**1.3 Interest**

As I mentioned above, it will be helpful for those who are self-employed or who will soon become. And some social-value ventures related economy and labor problem may have interest in it.

**2. Data**

**2.1 Sources**

I used data below.

· Number of famous restaurants near the each station, from Foursquare

: I use the 'explore' endpoint to count only 'famous' restaurants

· User number(of Nov 20) and location data of the each station, from Seoul Metropolitan Government

**2.2 Cleaning and Selection**

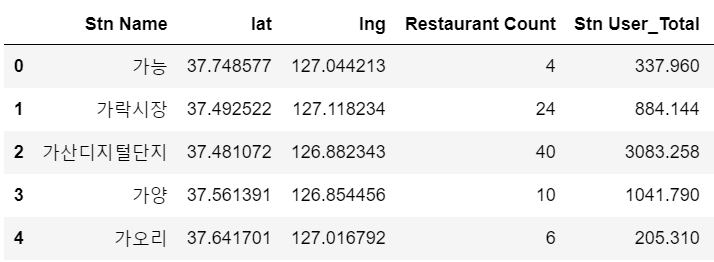
The Data I found was well-cleaned to use instantly, so I just conduct a few steps below.

· User number data included information that I did not need(ID number of stations and dates when the data was collected), so I dropped them.

· User number data was classified according to the date of use and whether to get on or off, so I summed them only by each station to get the monthly user number.

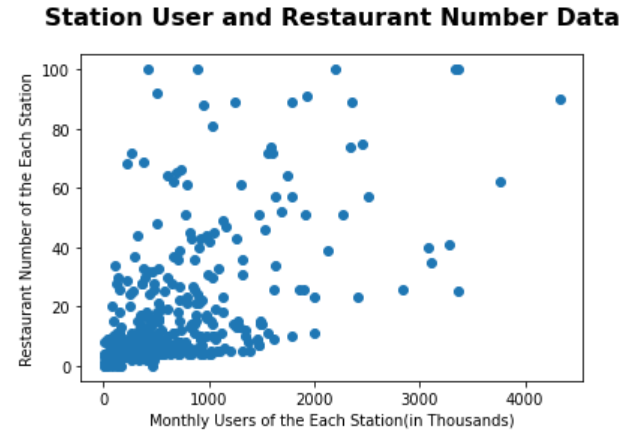
· Location of stations and the restaurant number data seemed not to need to be cleaning, so I merged them with the dataframe of user number data.

below is the head of the dataframe.



**2.3 Exploratory Analysis**

Let's see the scatter plot, where the number of subway users is the x value and the number of restaurants is the y value.



It seemed I can use linear regression, not classification nor clustering because data dots on the plot are clustered between (0,0) ~ (1000, 20) area.

**3. Methodology**

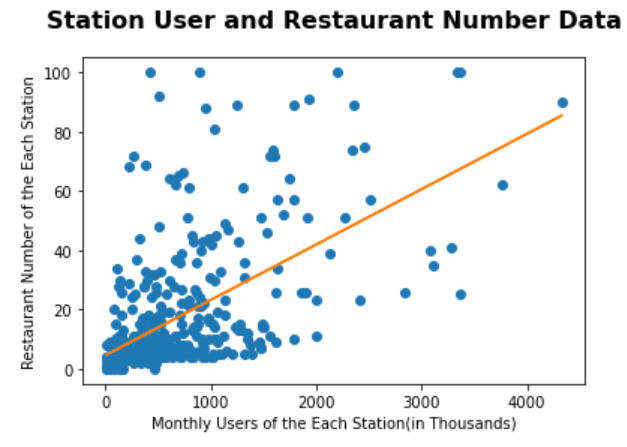
**3.1 Premise**

In this study, I assumed that sales would be proportional to the floating population per restaurants. I know that there are many other variables that affect on the sales, I think the floating population is critical and and its data seems to be an intuitive explanation.

**3.2 Linear Regression Modeling**

I did it by using the linear regression package in scikit-learn. And I did not divide the data into train sets and test sets, because I found I cannot use this model to predict and just can use it to describe the current status.

The result is shown as the linear line in the plot below, and the equation following.



*y = 0.0187X + 4.5950*

**4. Results and Discussion**

I defined 'restaurant capability' as subtracting real restaurant number from prediction, and it means the number of additional stores that can be opened, considering the average.

When stations are listed in the order of the highest number, the list of the top ranks is as follows. This ranking can be considered as a list of stations that are likely to be relatively less competitive to open the new restaurant.

**5. Conclusion**

In this project, I found places to be good location to open the new restaurant, by using some data. As a result of the analysis, it seems to be easier to open up the restaurant near some stations(like 신림, 신도림, 연신내, ...) because those have fewer famous restaurants compared to other, considering floating population and thus relatively less competition is expected.

Nevertheless, this project has some blind spots as below, because of the limit of the conductor's data analyzing knowledge and skill.

· It did not consider other variables, such as rent, purchasing power of main users of each station.

· Koreans usually use other applications(Naver Maps, Kakao Maps, etc.), not Foursquare, so the data about popular restaurants may be distorted.

(and I think the distortion made the result counterintuitive, but I want to use Foursquare to consider the estimator of this project, who may not be familiar with korean platforms.)

If I have the opportunity to develop this project, I will use more kinds of data including factors affecting self-employed sales, and other application to crawling social data.