Deciding whether Opening a Café Alone or next to Competitors

1. Introduction

1.1 Background:

When deciding where to open a café it is important to take into consideration if there are any competitors around that location. At first thought it may be better to open without any competitors nearby, because that will mean less competition for the business owner. But, maybe opening a café next to competitors will create a hotspot in the mind of the consumer and both cafes will benefit from it.

1.2 Problem

Data will contribute to decide on if we should open the café next to others or alone. We will find out if cafes next to neighbors receive more likes than others.

1.3 Interest

This problem will interest potential café openers. It will help them decide the ideal place to open one.

2. Data Collection

2.1 Data Sources

The data will be selected from Foursquare. We will find as many cafes in Athens, Greece as possible. Then find their lat and lng location and call the api to find how many cafes exist around that location in a radius of 35 meters. This will be the number of neighboorhs of that café. Then we will find the number of likes each café received.

2.2 Data Cleaning

The main problem after picking the data was that some cafes had too high values for example 5000 likes while the average was around 17 likes. This would dramatically increase the average in certain cases. To compensate for that we took the cubic root of likes in order to reduce the impact those high scores had on averages.

Also some neighborhoods were clearly more advantageous than others. So we further adjust the likes of each café by dividing with the median of each neighborhood.

We also removed big chain cafes out of that because they would clearly have an advantage due to the reputation.

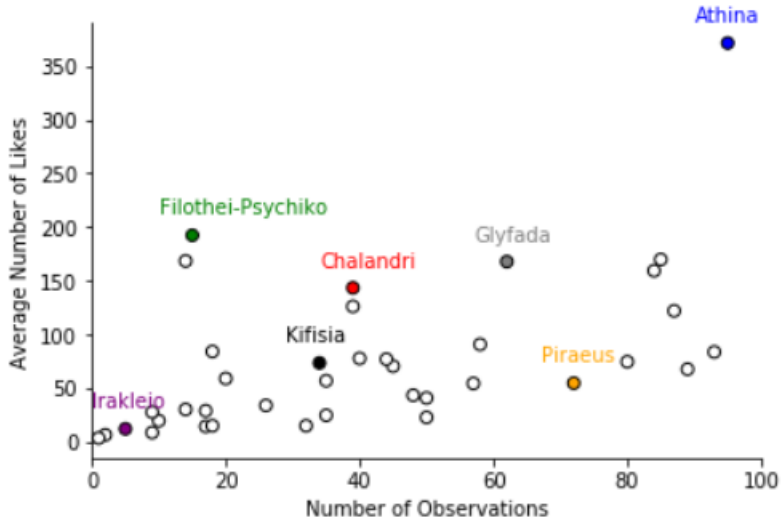
2.3 Feature Selecting

The features that were important were three. How many neighborhoods the café has, in which neighborhood was located and how many likes it received.

3. Explanatory Data Analysis

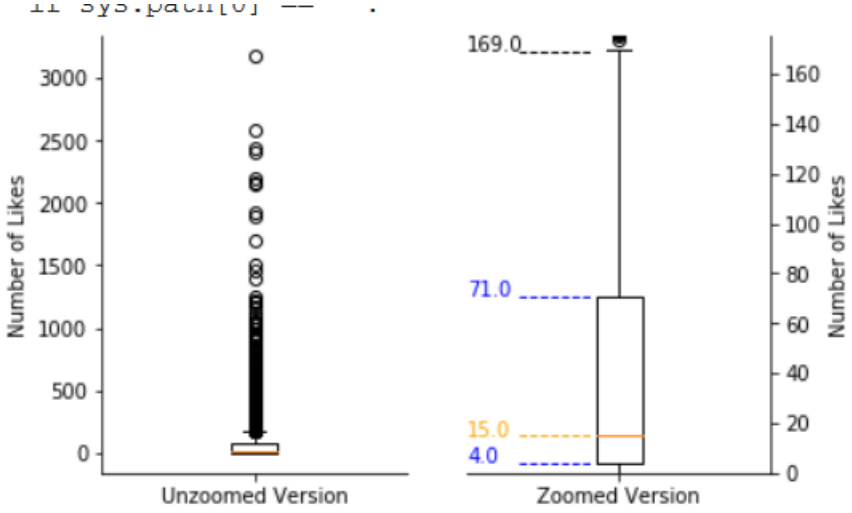
3.1 Examining Neighborhoods

The first thing was to see the neighborhoods. So we plotted the neighborhoods versus the number of observations and we pointed out some of the examples. We clearly see that some neighborhoods receive high number of likes in average. So we had to readjust for the neighborhood.



3.2 Examining Data Set

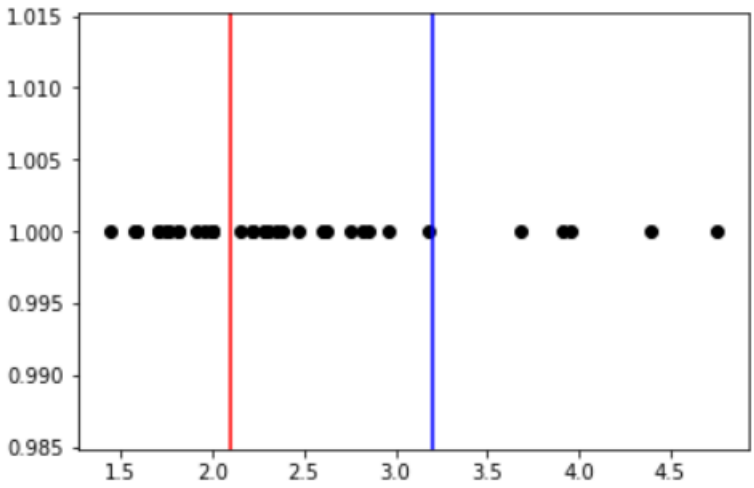
Then we plotted all the data points in one boxplot. The result is clearly ugly and some further adjustments have to be made.



We see that the median number of likes was 15 and the bottom 25 percentile was just 4! Also only 5% of the cafes received more than 169 likes. But those likes are so much more than the median of 15 that change the average in a big way. For that reason we decided to take the cubic root of each like.

3.3 Adjusting for Neighborhood

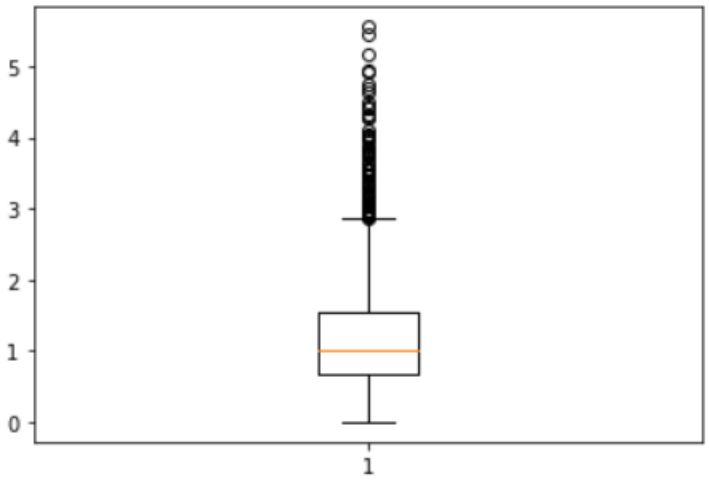
After taking the cubic root of the cafes likes we draw the following graph. Each dot in the graph is the median value of the new rating of some neighborhood. We see that there are three categories of those values.



We found the median of each category and then in each café we divided that median in the category its neighborhood existed.

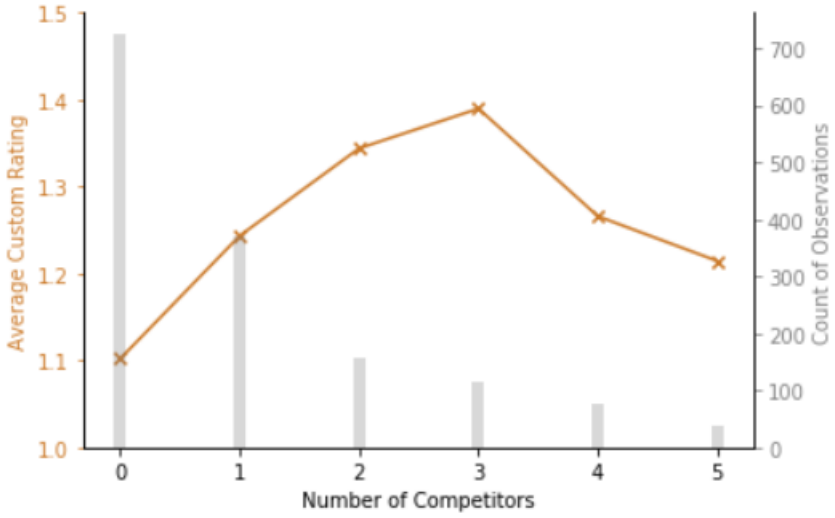
3.4 Final Check

After applying the changes discussed we drew the final boxplot that shows all the cafes. It looks good.



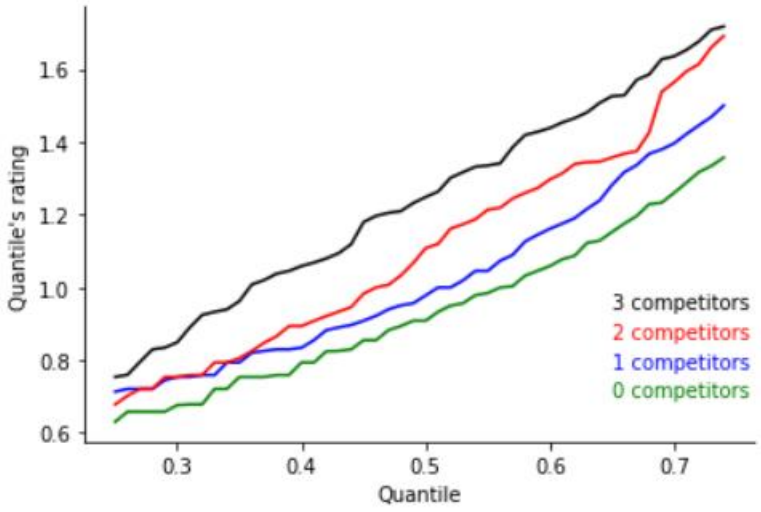
3.5 Results

	Average Rating	count
0	1.102016	726
1	1.243370	370
2	1.343927	157
3	1.389701	116
4	1.265879	78
5	1.214428	39
6	1.514053	24
7	1.369749	11
8	1.794133	8
9	1.727233	8
10	1.913504	4
11	1.309877	6
12	2.000000	1



Those graphs above show that the average number of the new adjusted rating spikes when there are 3 competitors around you. Then fall if there are 4 or 5 and then climbs back up, although there are not many observations for above 5.

But someone may not like taking averages so we drew the following graph.



On the x-axis we have the quantile percentile and on the y-axis we have the rating for that percentile. For example 0 competitors has 0.7 rating for its 40% percentile. We see that for the number of competitors that there is enough data, in all stages the 0 competitors line is dominated by the other 3. This indicates that opening next to no competitor is a bad idea.