I tried two MapReduce jobs to profiling. Firstly, I used MapRe- duce to count the total number of records in the data set, which is 268205 lines. Secondly, I tried to count the number of columns in each record. Ideally, there would be 27 columns for each record, in which each column is filled with valid data. However, after MapReduce job, I found that the range of number of columns is ranging from 6 to 67. There are 79029 records have length greater than 28; 99586 records have length between 25 and 28; 89590 records have length smaller than 25. The reason behind this is firstly because some records have missing values. Because the file is comma separated files, during MapReduce job, the data in one column are counted multiple times. Also, OpenData mentions that there will be duplicate data in the data set. Therefore, two MapReduce jobs are executed to improve data accuracy. Then, I used another MapReduce job to filter out the abnormal records. Since the column I were going to use is DCA License number, Business type, License status, business industry, and borough, I checked if records have valid data for these fields. If a record have valid data for the columns we need, we put them into a scheme in Impala for future analytic.

Analytic: In this dimension, I tried to focus on the density of businesses and variety of businesses and calculate a score for each borough to indicate its business index. The first goal is to count the total number of businesses in each borough. I used count(), group by to get result. When calculating the number of businesses in each borough, I9 found that the name of borough varies. There are two name versions for each borough, for example, “Manhat- tan” and “MANHATTAN”, “Brooklyn” and “BROOKLYN”, “Bronx” and “BRONX”, “Queens” and “QUEENS”, “Staten Island” and “STATEN ISLAND”. To solve this problem, I used ‘and’ to add them together. The results are: Brooklyn:47879; Queens:43397; Manhattan:38923; Bronx:23520; Staten Island:8713. However, different boroughs have different land area. If I only use the total number of businesses, it would be unpersuasive. Therefore, I divided the total number of businesses in each borough by its land area. The land area of Manhattan, Brooklyn, Queens, Bronx, and Staten Island are 59.13, 183.42, 281.09, 109.04, 151.18 km square respectively. After we divided the total number of businesses by its land area, the results are: Manhattan: 658.26; Brooklyn:261.03 ; Bronx:215.70 ; Queens:154.39 ; Staten Island:57.63 businesses per km square.

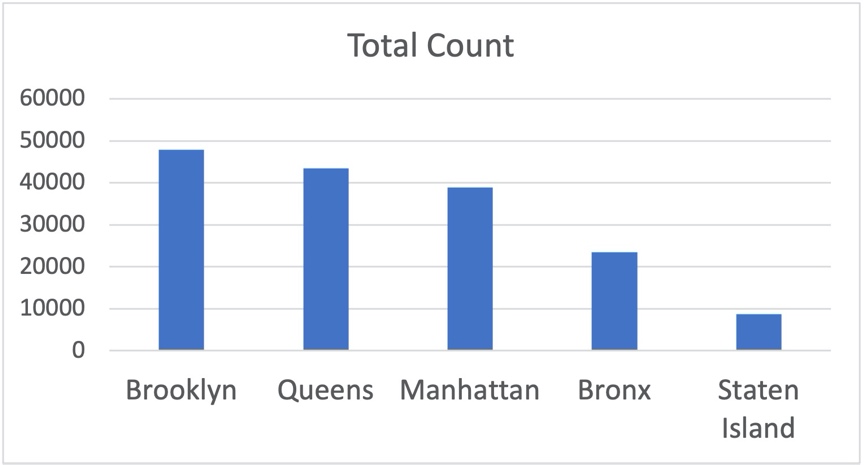


Figure 1: Total Business Count

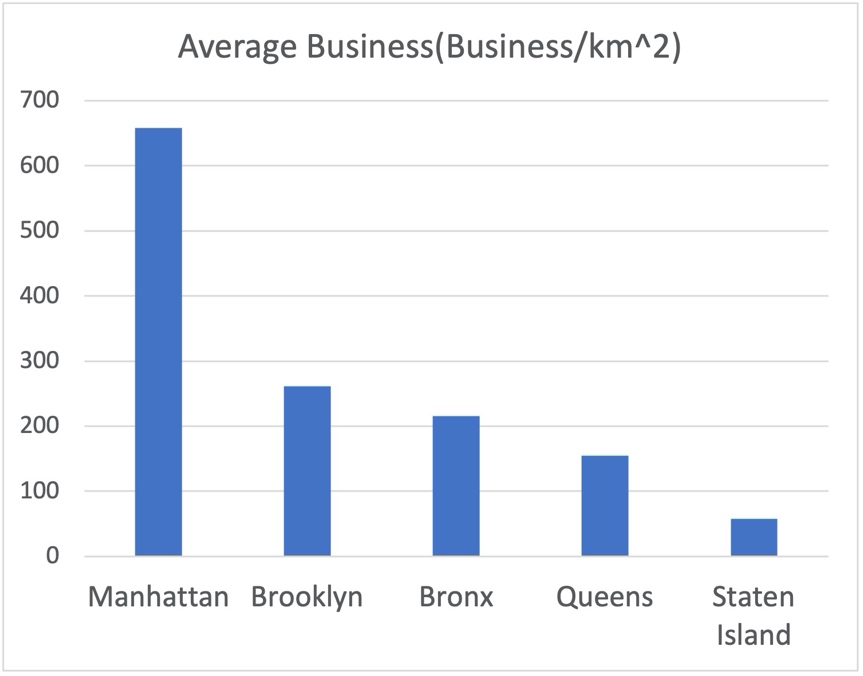


Figure 2: Average Business Count

I can see that Brooklyn ranked first previously, but after divided by land area, Manhattan has a huge advantage over other boroughs. The final ranking is Manhattan > Brooklyn > Bronx > Queens > Staten Island. So far, I figured out the density of businesses in each borough. However, knowing only the density of businesses is not enough for me to conclude which borough has the best business environment. I still need to find out what categories are there among those businesses, i.e., I need to know what kind of business are there. I use distinct keyword to have a vague idea about what kind of businesses in New York City.

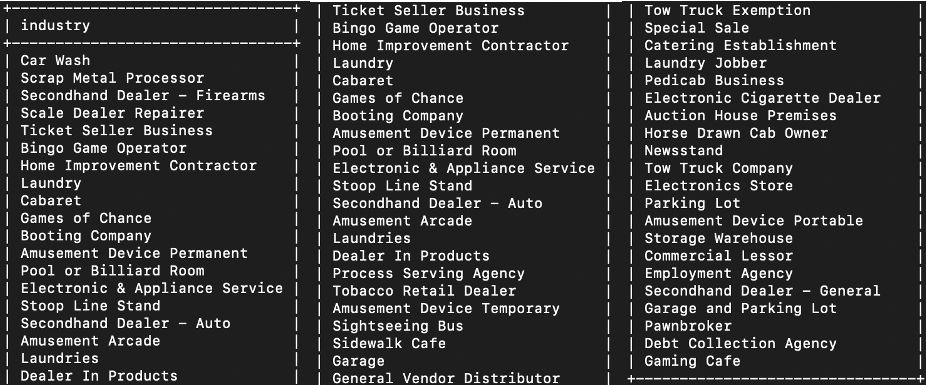


Figure 3: Business Category

The category varies. Therefore, it is hard and cumbersome to count the number of each kind. But I can see that they can be split into three main categories: food, life services, and entertain- ment. So, I analyze the top 10 most frequent kinds in each borough and category them into these three main categories to see the distribution. Through the distribution, I can see that Brooklyn and Manhattan have a balance among food, life services, and entertainment, while Queens, Staten Island, and Bronx put focus on life services the most. For Staten Island, it doesn’t even have a food business holding a DCA License. By zero number on food business doesn’t necessarily mean that there is no food restaurant in Staten Island. It simply means there are no food business holding a DCA License. Finally, I use data above to calculate a final business score for each borough. Three factors need to be considered which are: food, life services, and entertainment. The weight for food is 0.5, the weight for life services is 0.3, and the weight for entertainment is 0.2. The formula I used to calculate the final business score is: Final Business Score = food business per km2 \* 0.5 + life services business per km2 \* 0.3 + entertainment business per km2 \* 0.2 The results are the following:

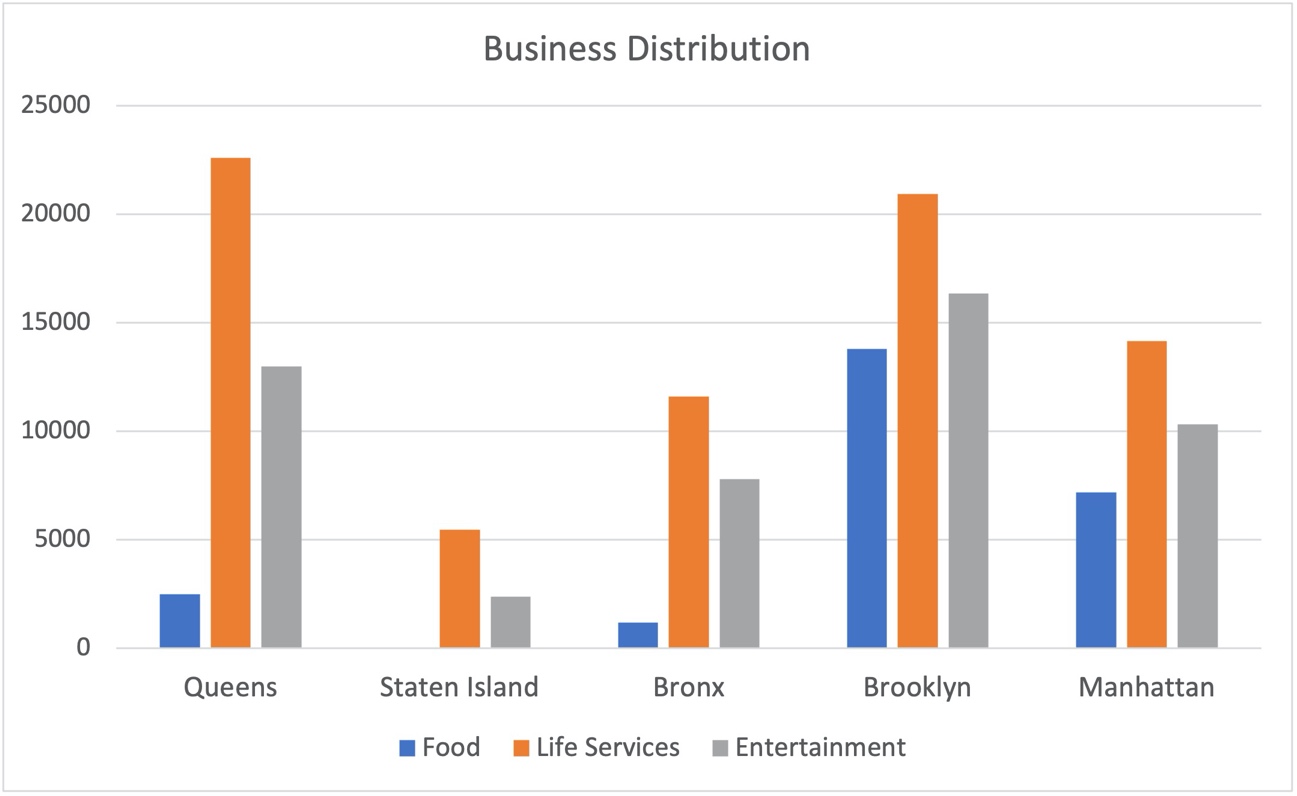


Figure 4: Business Category Distribution

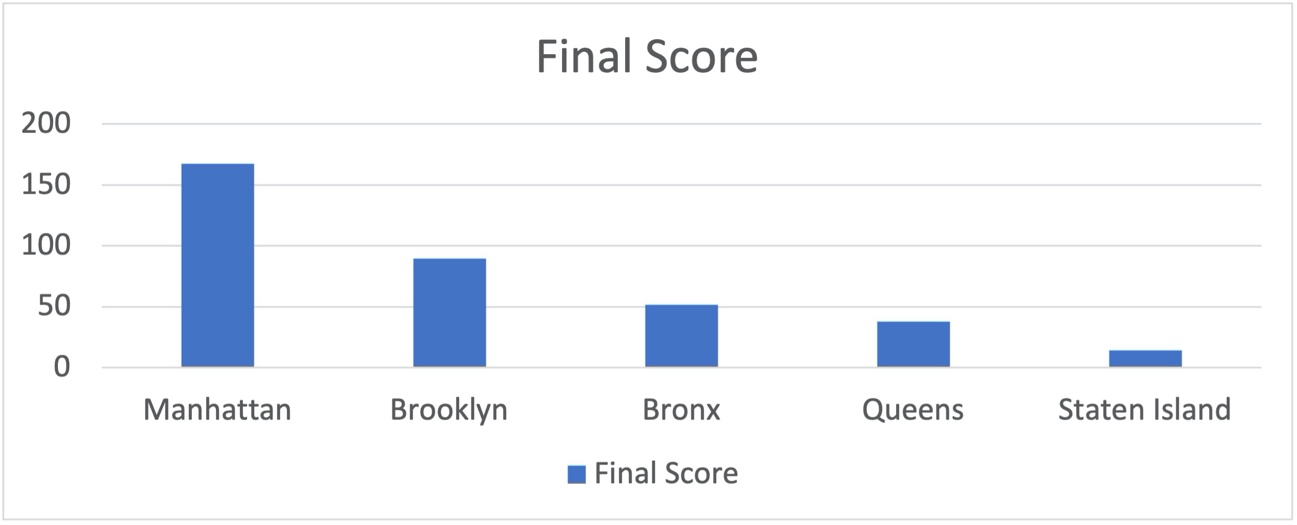


Figure 5: Final Score for each Borough

Manhattan (167.26) > Brooklyn (89.64)> Bronx (51.52)> Queens (37.66)> Staten Island (13.95) Therefore, I concluded that Manhattan is the borough with highest business score, following by Brooklyn, Bronx, Queens, and Staten Island. The results of this analytic are used for final conclusion. • Manhattan(score 5): Highest density of business, great balance on business distribution 17

• Brooklyn(score 4): Highest number of business, great balance on business distribution • Bronx(score 3): Adequate number of business and business distribution • Queens(score 2): Relatively high number of business • Staten Island(score 1): not balance on business distribution distribution