MATLAB Report

Data Analysis of Dianping

Yezhou Ma
School of Software
14302010052

1. Distribution of different features

1.1. Among Category

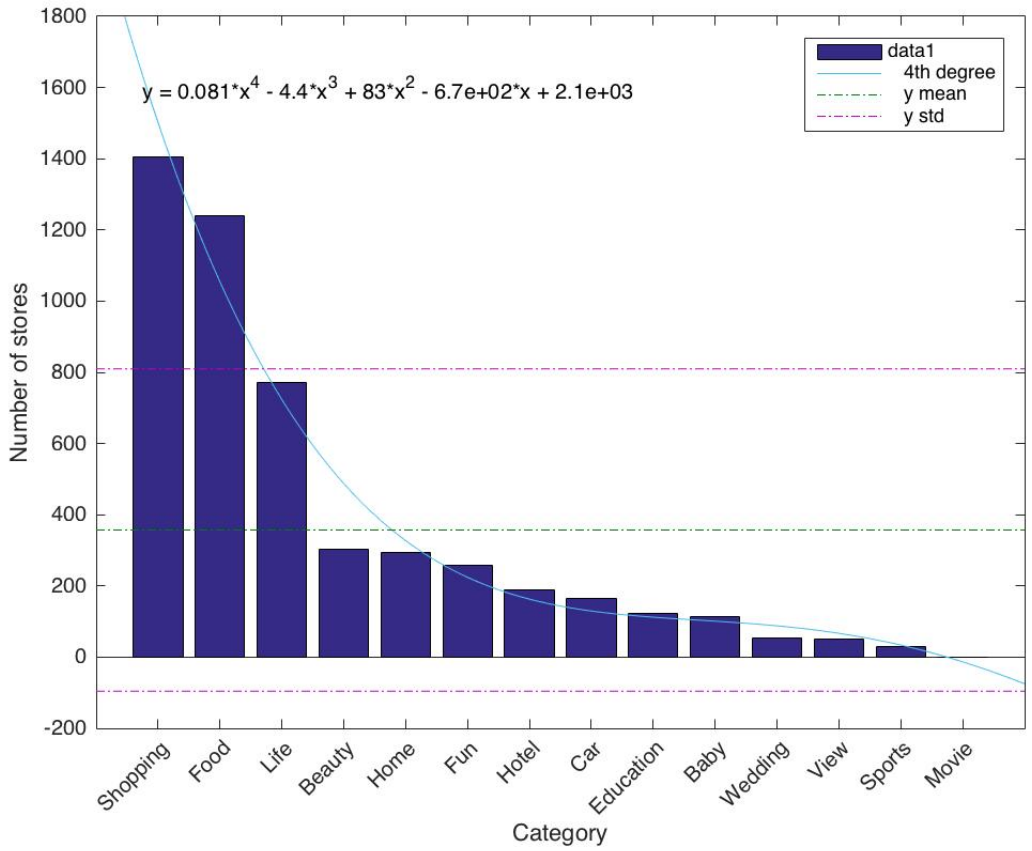


Figure 1. Number of stores per category

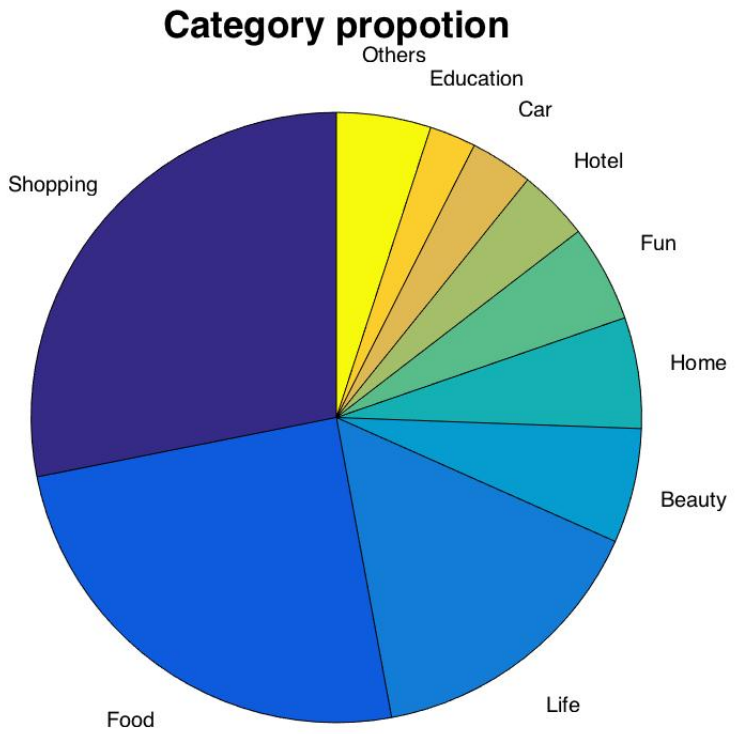


Figure 2. Propotion of Categories

Shopping stores occupy the most of the market, next is Food category. Only these two categories and life stores take up more than the average number, the other categories contain only a small number of stores. The range between y std and y mean is the standard deviation of data. In the figure, numbers of Shopping and Food stores are over one standard deviation greater than average value. Thus the standard deviation is great.

The categories are ranked by descendant order, thus we can fit the number of stores with rank by a 4^{th} polynomial function.

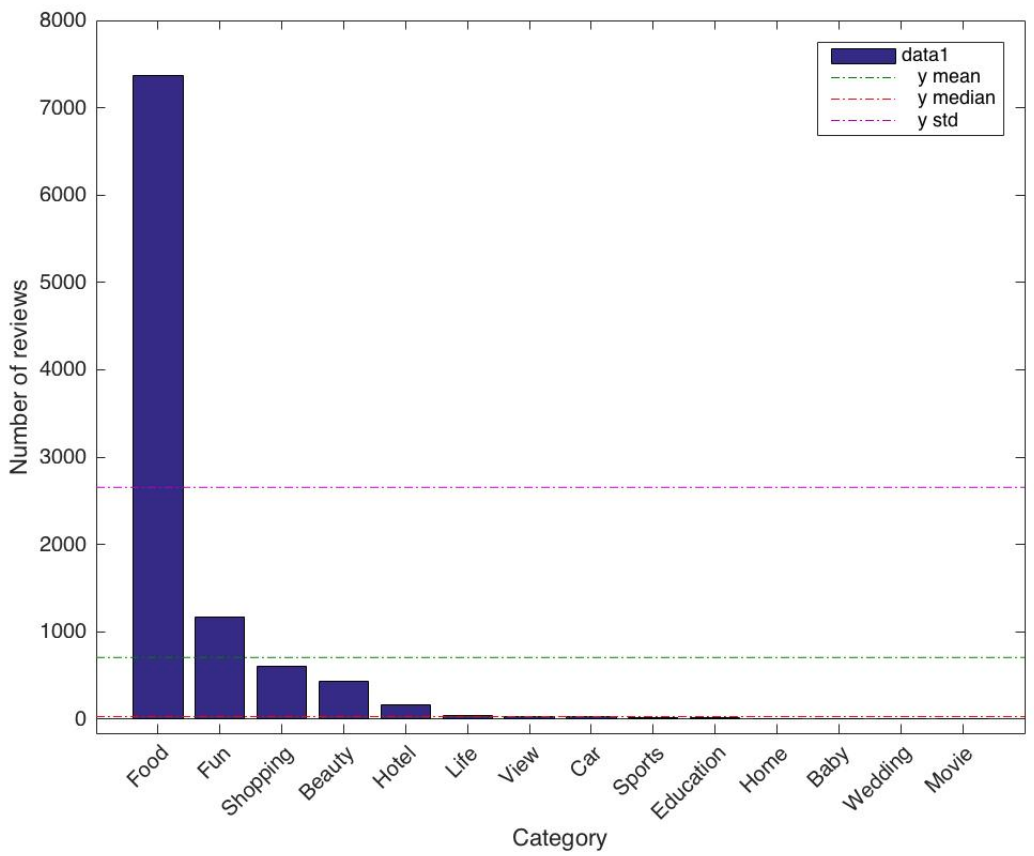


Figure 3. Number of reviews per category

The number of reviews of Food stores goes far beyond the other categories and raise the average

number, thus contributes to the great standard deviation. And there is no review for Home, Baby, Wedding, Movie stores. It makes sense since there is only over a hundred Baby stores, decades of Wedding store and one Movie store. While Home stores rank 5th according to the number of stores per category, it is surprising there is no review for Home stores. Therefore, we feel the urge to analyze the number of reviews per store.

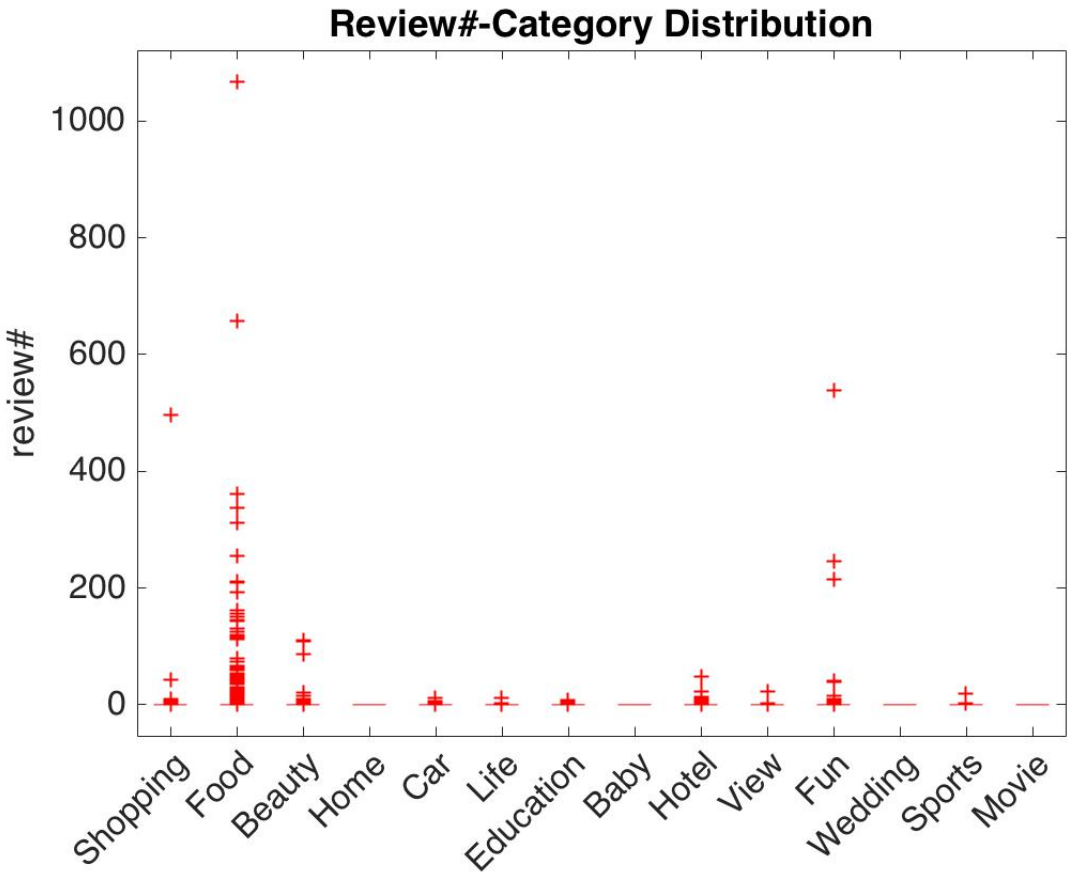


Figure 4. Distribution of review amount

Figure 4 is the box graph of distribution of review amount per category. Abnormally, there is no box in the figure only markers which are supposed to represent abnormal entities. Thus a conclusion can be drawn that the distribution is not Gaussian Distribution, to be more specific, it is skewed and the variance should be great.

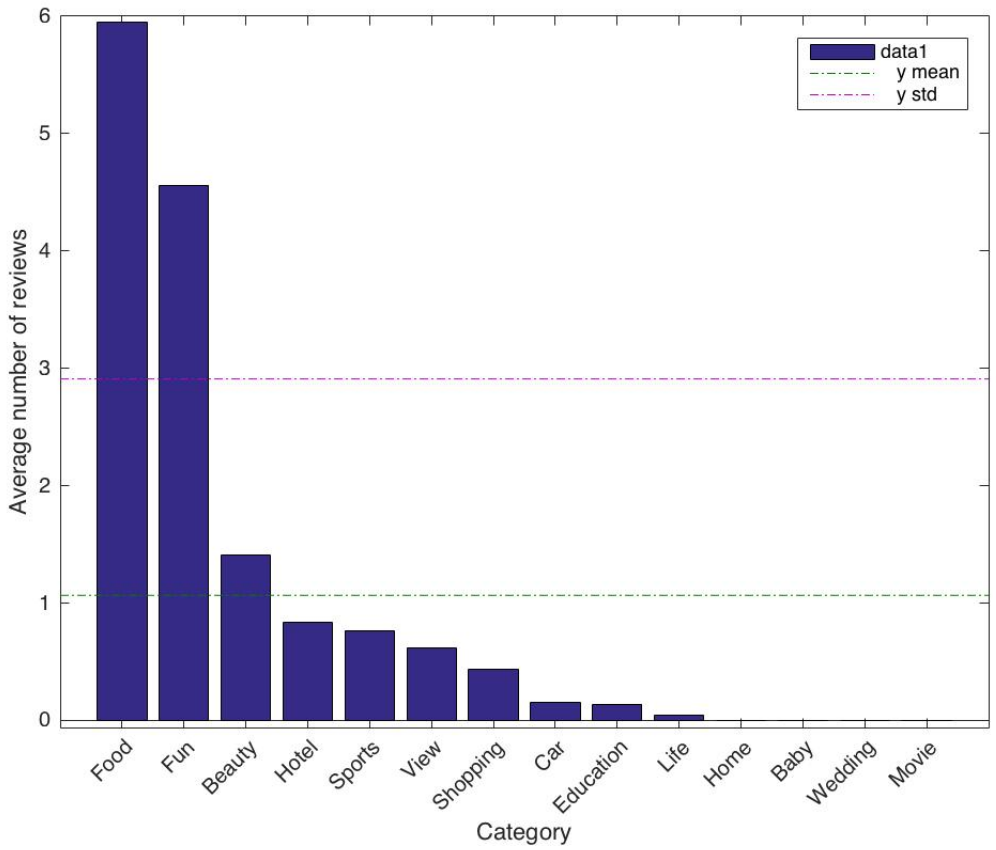


Figure 5. Average number of reviews per store per category

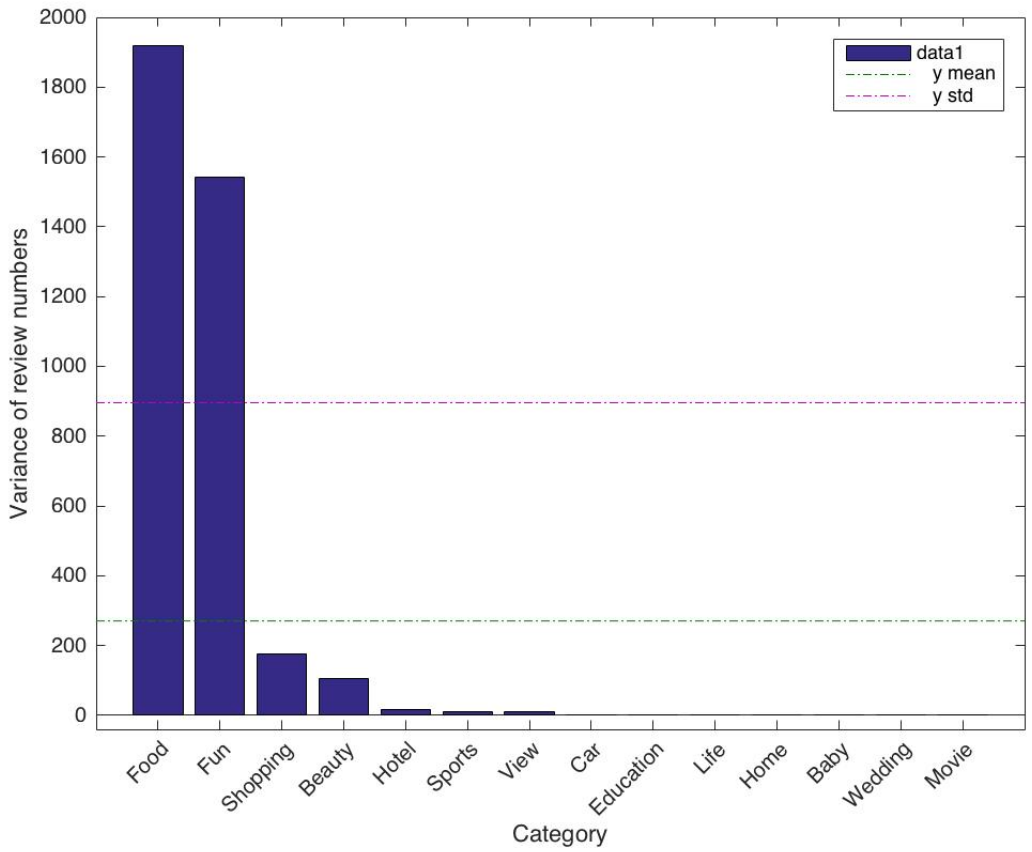


Figure 6. Variance of average number of reviews per store per category

Combine all the diagrams, we can clearly see characteristics of each category.

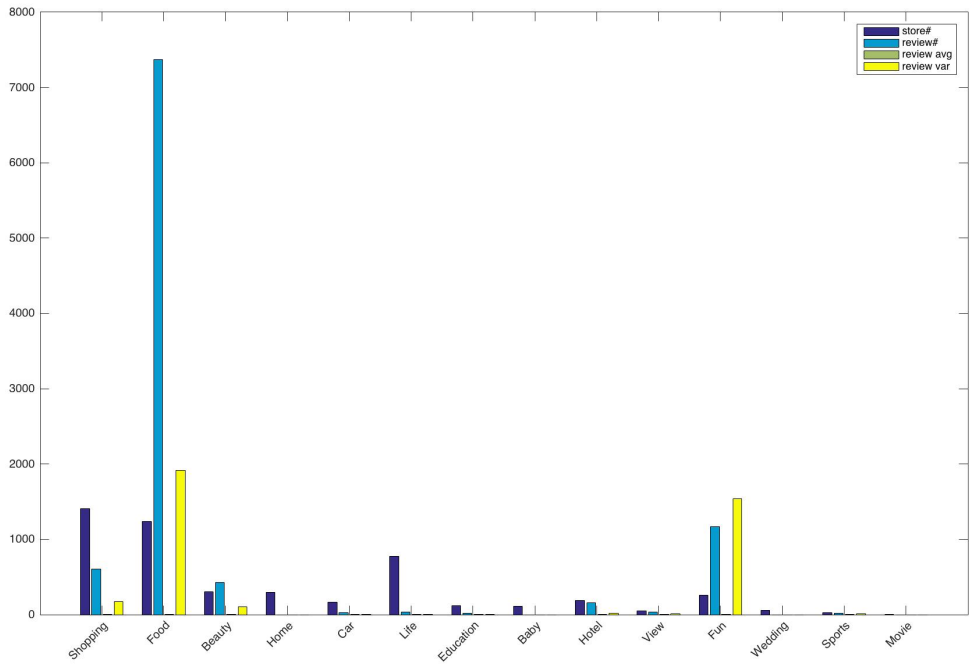


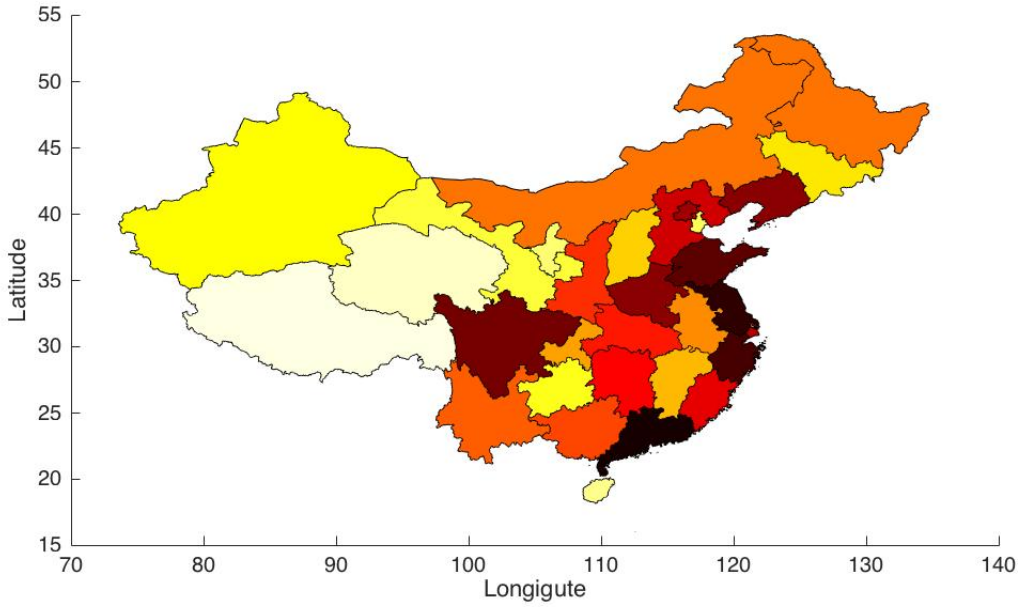
Figure 7. 4 index of each category

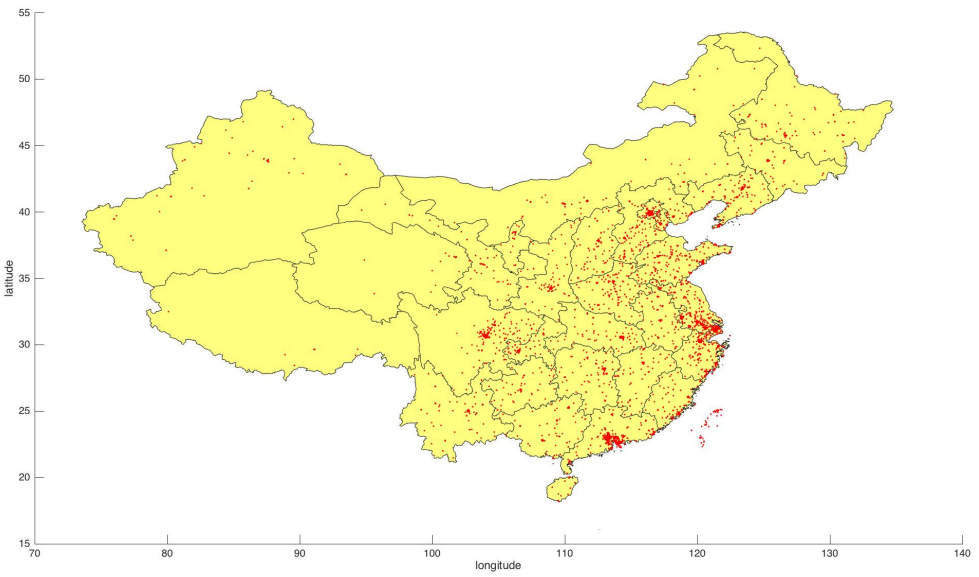
Customers tend to review more for Food stores, for they tend to decide where to have meals or buy food depending on reviews of stores. While its variance of review amounts is also great, popular Food stores can have hundreds of reviews, and new stores or ordinary ones have zero or few reviews. So are Fun stores. On the contrary, reviews count for less for Shopping stores.

Figure 8. Ratio of verified stores in each category

Stores of over 30% categories actually are not verified at all, the ratio in sports stores are the highest, the next is Food.

1.2. Among region





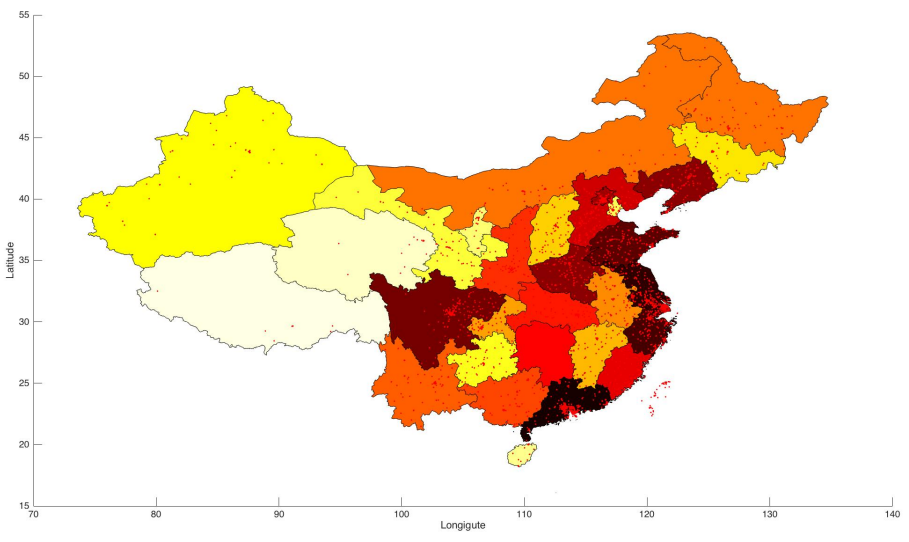


Figure 9. Distribution of stores per province

A dot represents a store, and is shown at the exact point of its location given its latitude and longitude. Color of each province, municipality or special administrative region suggests rank of the region. The deeper the color is, the higher the rank is. The basic China map is plot based on data provided by *StatPlanet*[2], thus misses some important parts of China. Fortunately, the error can be born.

Distribution of stores according to their location conforms to the economic development of each region. The dots cluster are municipalities or capitals of provinces. And deep-colored regions tend to be more international and modern. We can also see the ladder-like distribution from west to east. *The three steps of China's terrain*[2] is famous in Chinese topography.

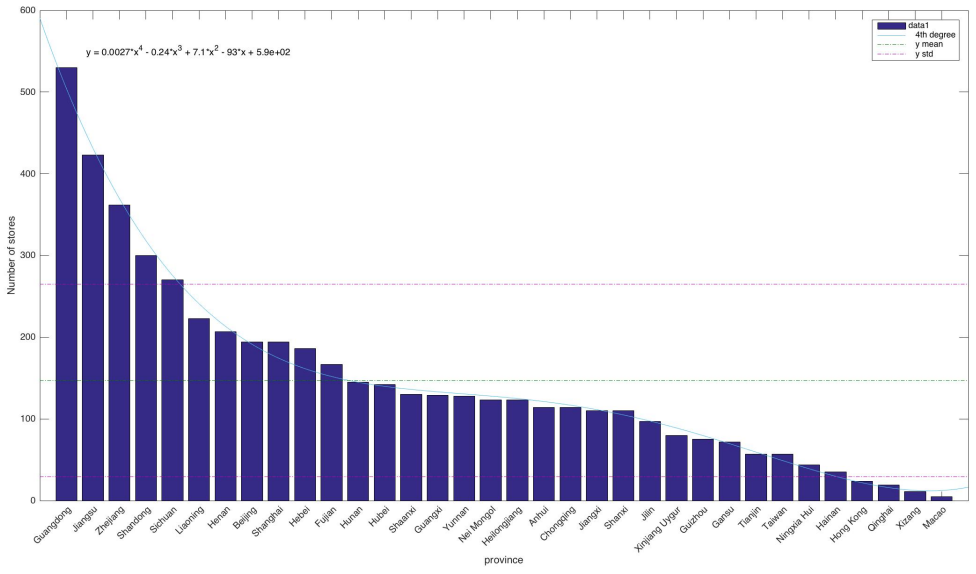
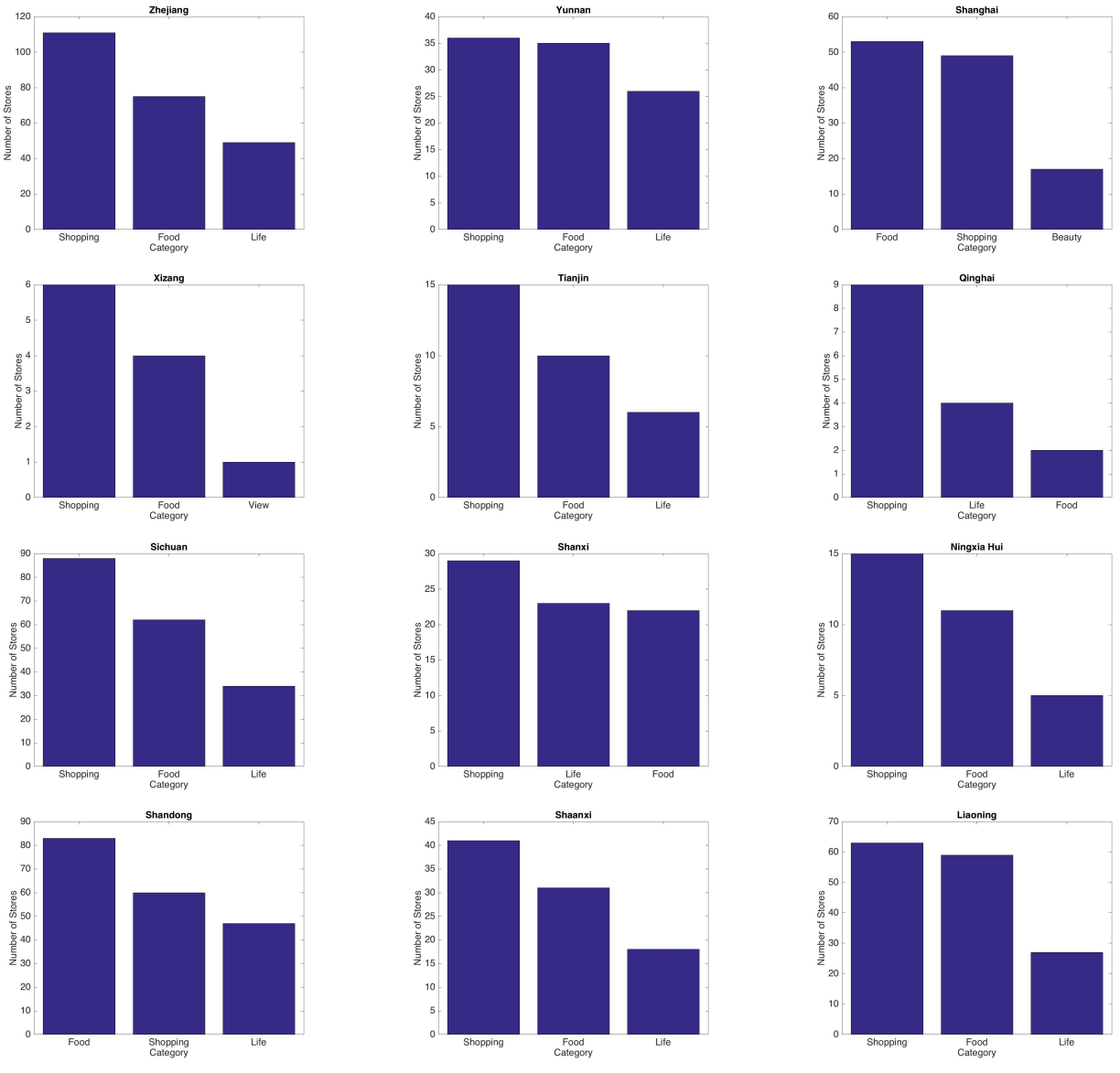


Figure 10. Rank of province

The bar diagram show the rank of provinces more directly. And I do a basic fitting on relation between number of stores and rank. We can see Guangdong, Jiangsu and Zhejiang province rank the top 3, while Beijing and Shanghai rank a bit lower. And later we will look deep into each province to see what contributes most to its rank.



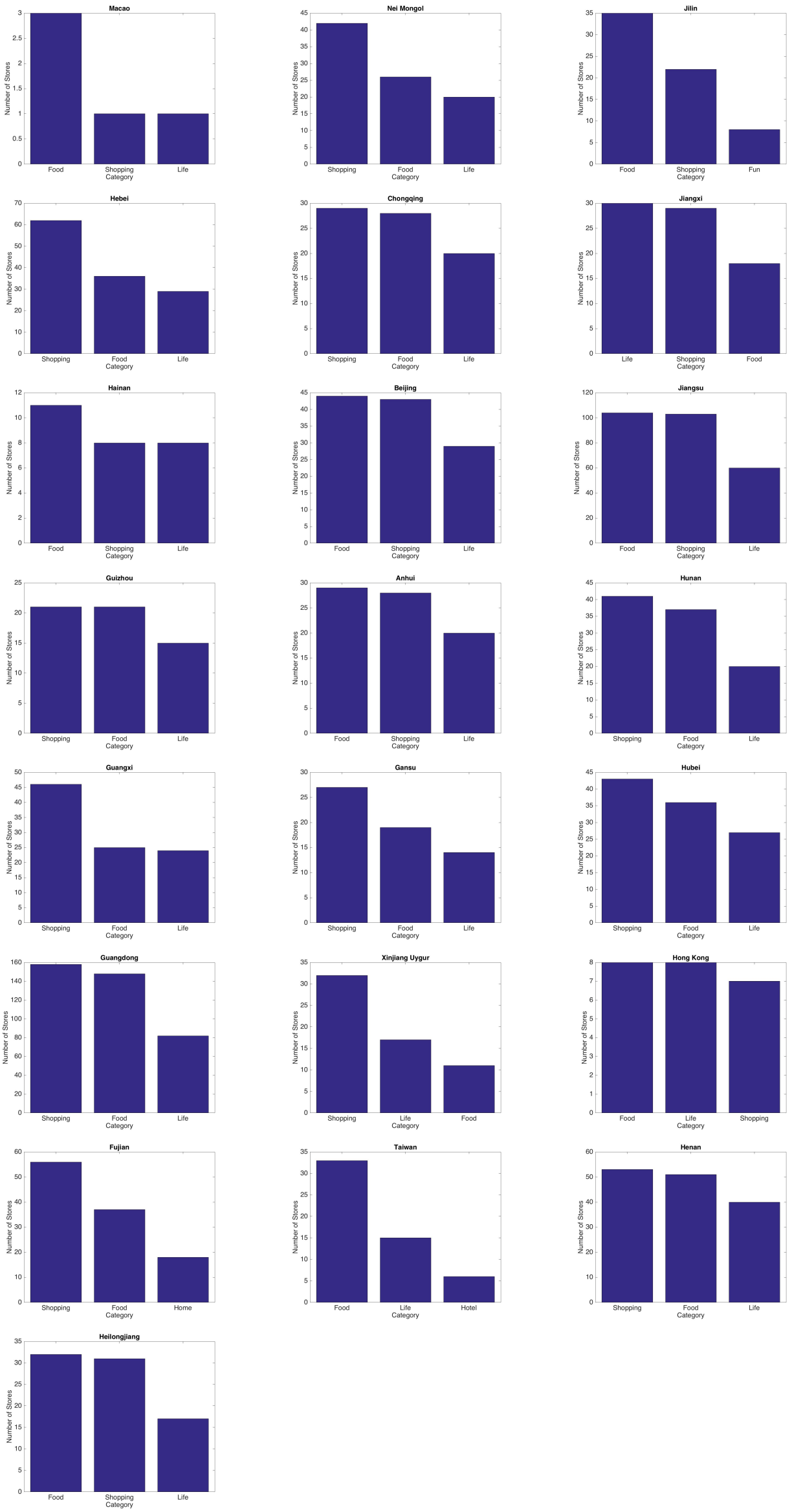


Figure 11. Top 3 categories per province

For most region, top 3 categories are Food, Shopping and Life, no matter how these three categories rank. But in Fujian, the third category is Home and it is Beauty in Shanghai, Fun in

Jilin, Hotel in Taiwan, View in Xizang. Regional peculiarity can be seen.

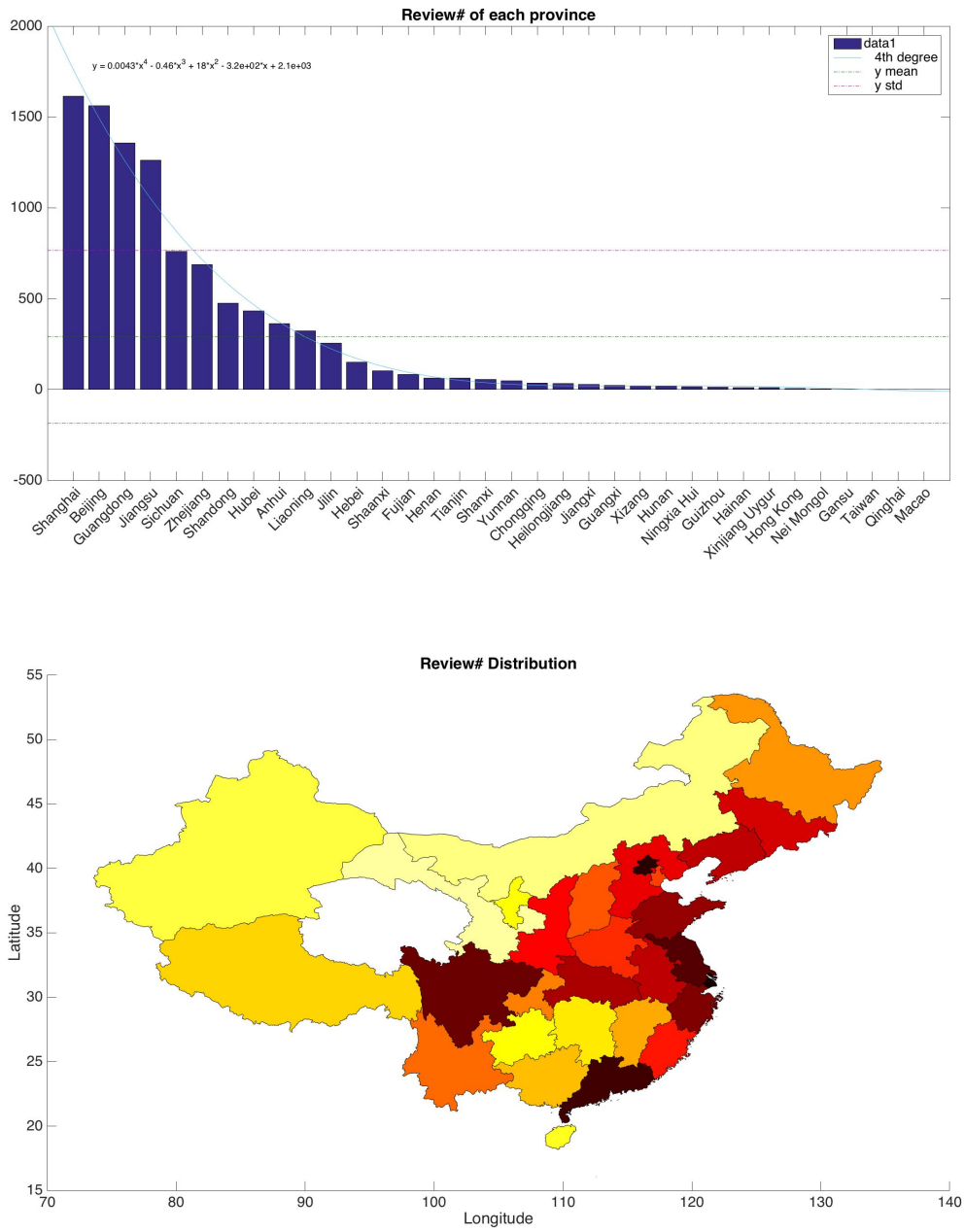
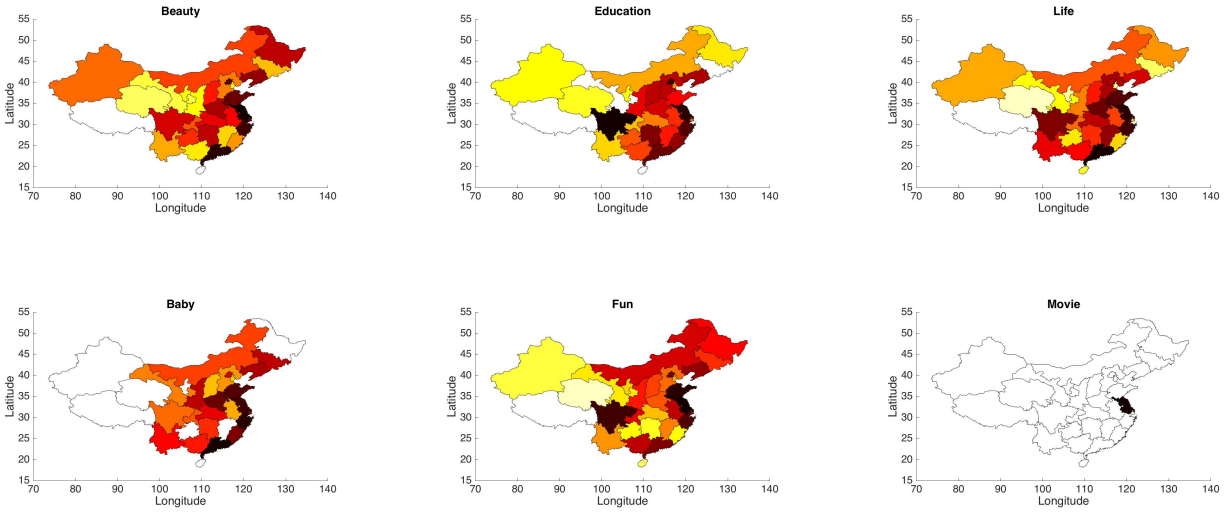


Figure 12. Distribution of review amount to region

A bit different from Figure 10, in the distribution of review amount, Beijing, Shanghai, Guangzhou these three most popular cities for job hunters ranked just top 3. While surprisingly, Qinghai province has no review.

1.2.1. Combine category and region



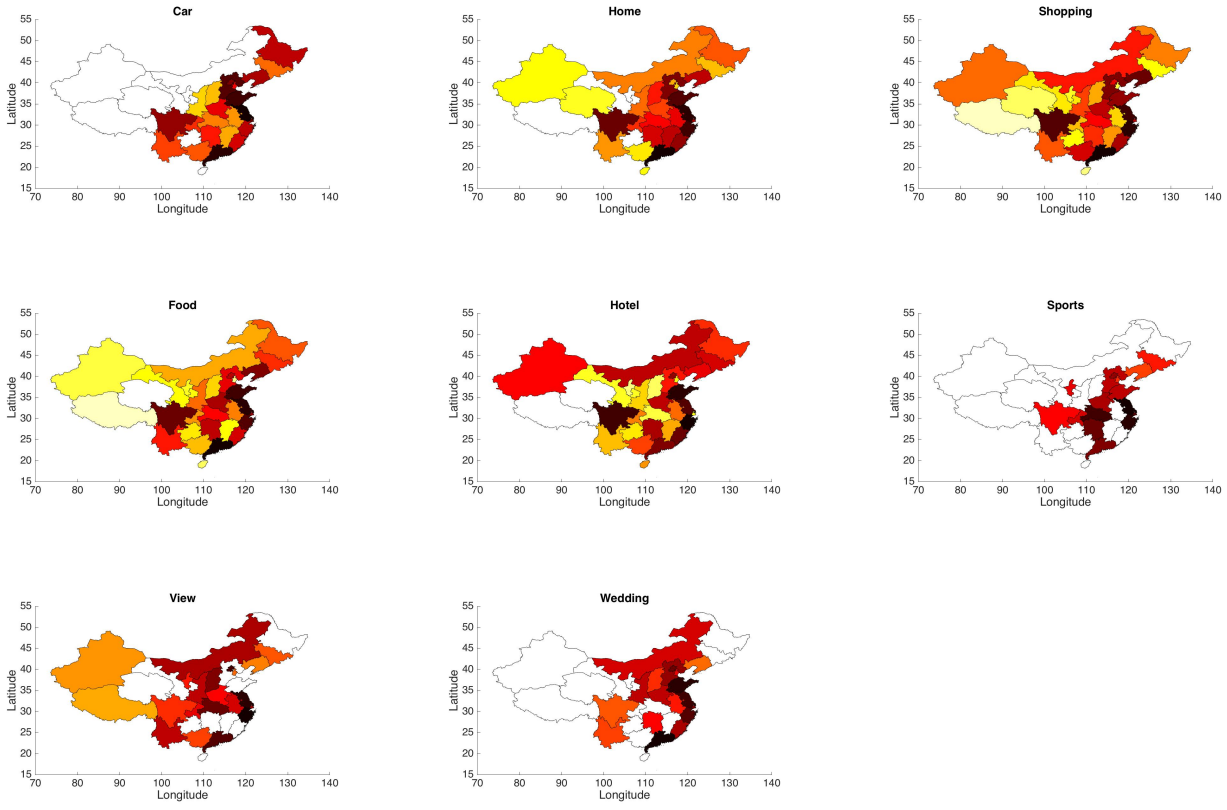


Figure 13. Distribution of each category among province

If a region is white, it means there is no store.

1.3. Temporal

1.3.1. register date

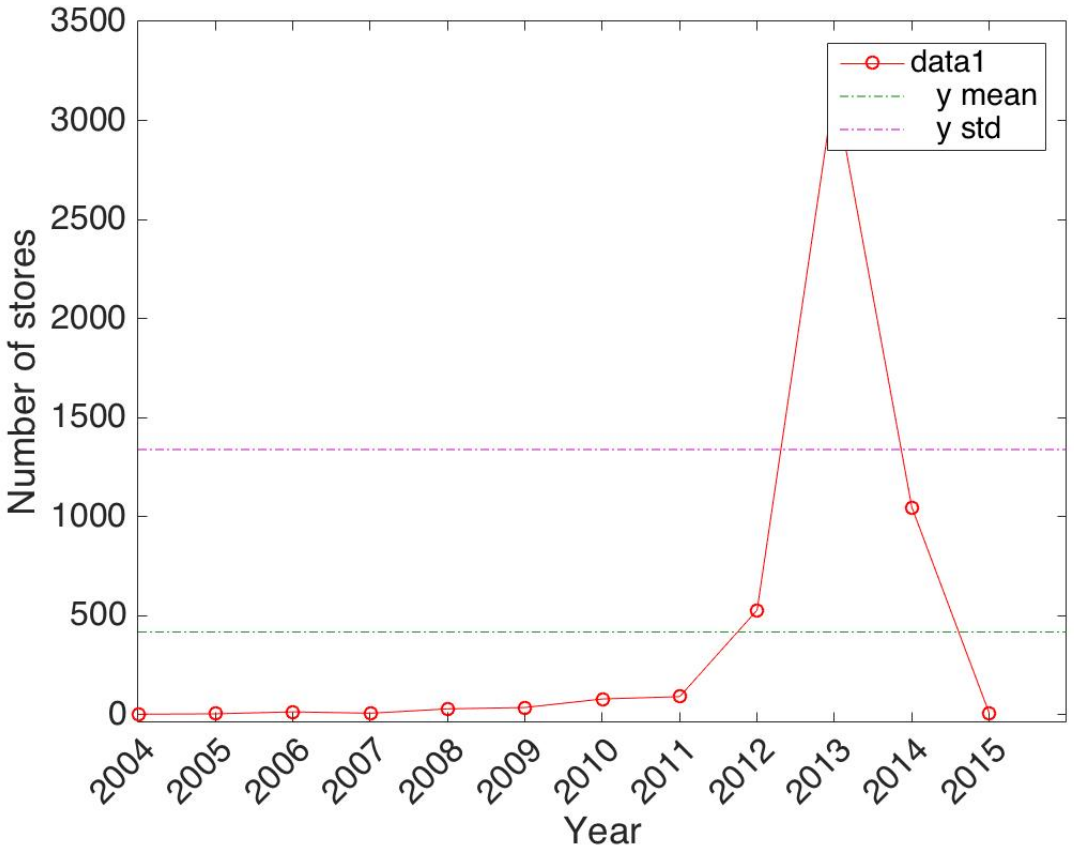


Figure 14. Number of stores registered per year

Since Dianping enterprise was established in April 2003, and experienced hard time at the beginning, there are few stores registered between year 2004 to 2011. Actually, before 2011 Dianping only focused on UGC (User Generated Content), while later online groupon was introduced. Given that online paying sprung up at that time, there was a reasonable increase in customers. It began to earn its popularity in year 2012, a remarkable increase of new stores

registered could be witnessed. However, after most of the stores completed their registration, there were only a few new stores registered in 2015.

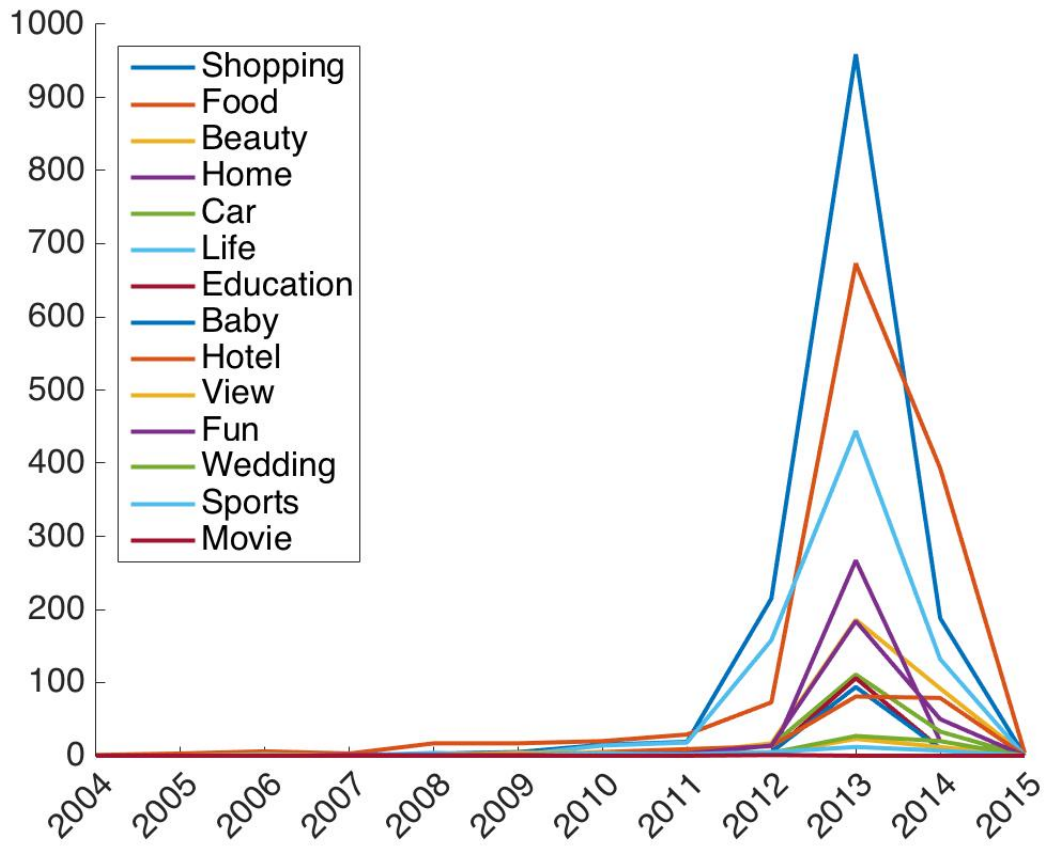
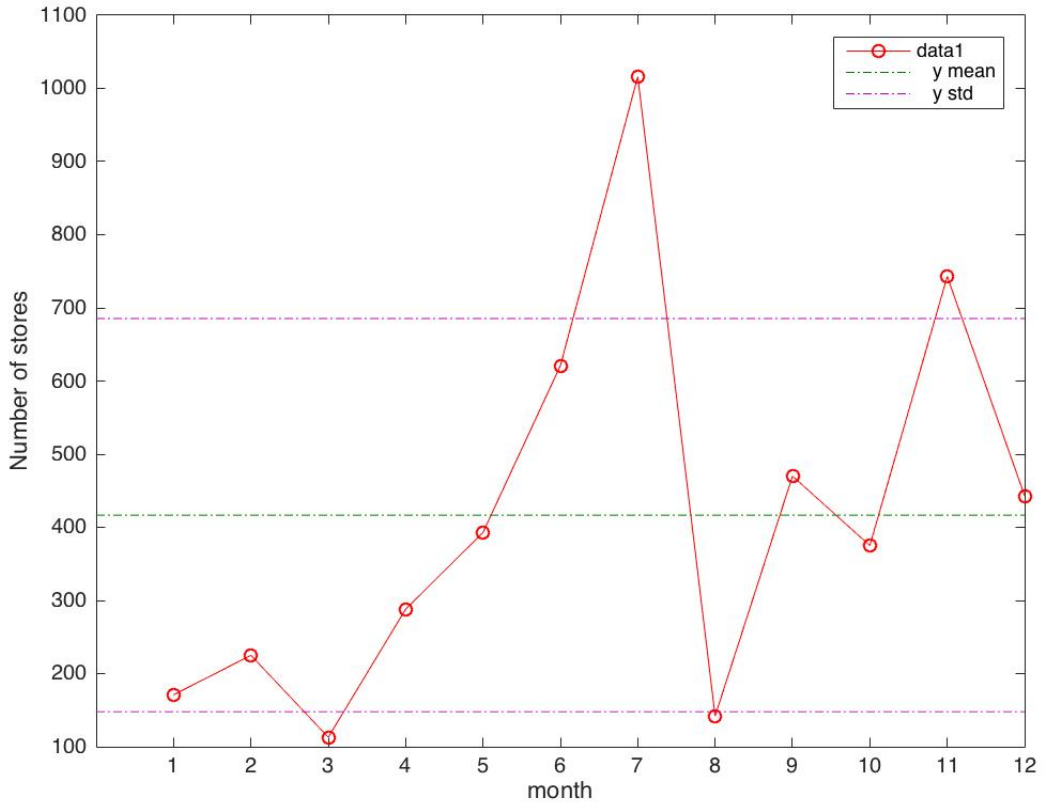


Figure 15. Number of stores registered of different categories per year

Temporal distribution of store registration of each category is identical to the overall distribution.



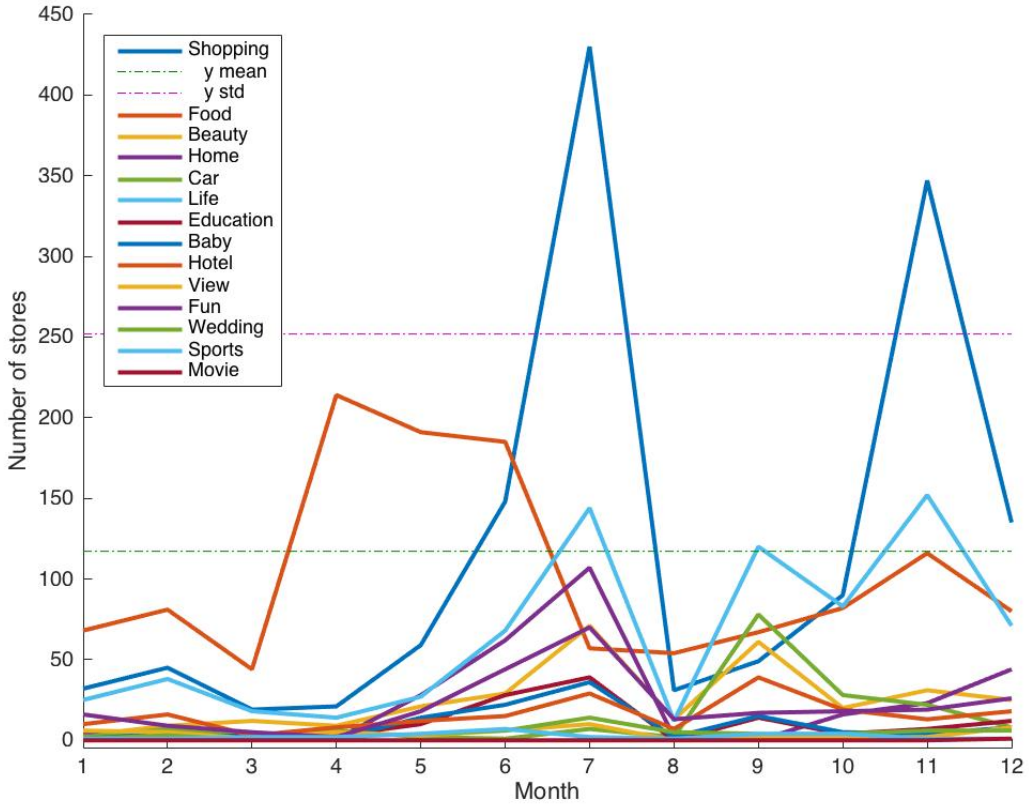
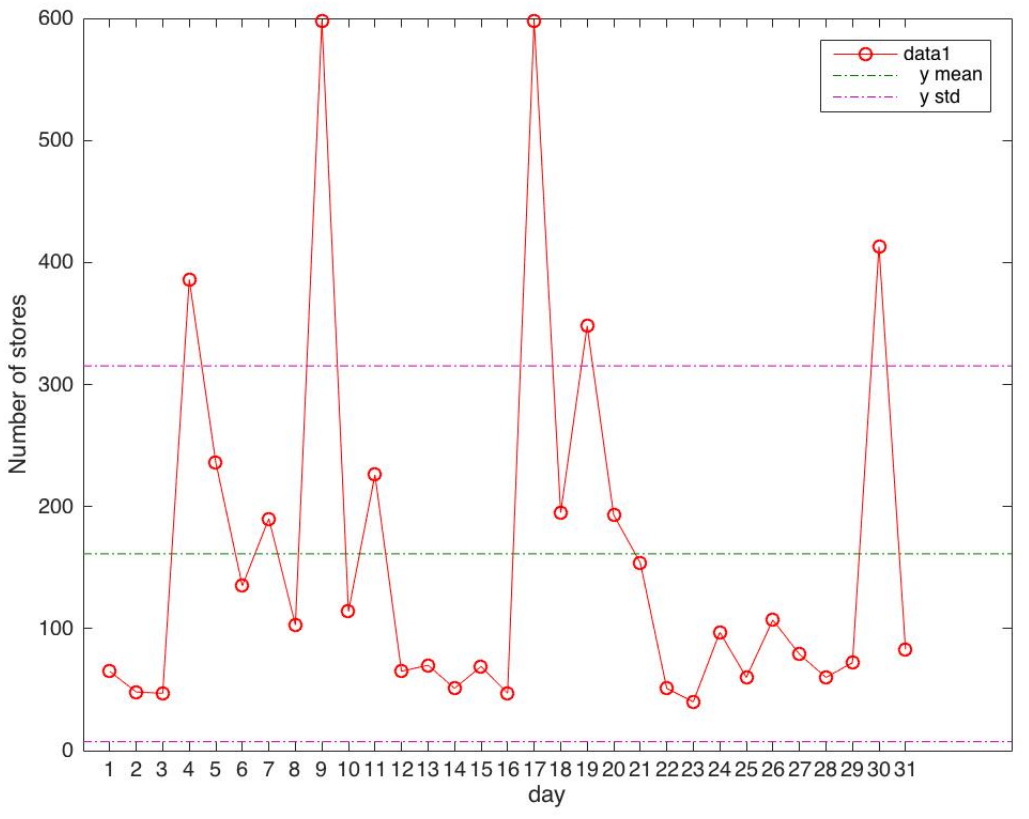


Figure 16. Distribution of store registration among 12 months

A lot of stores registered in summer vacation and November. June, July, September, November and December gained more stores than the average number. Looking into each category, the temporal distributions are similar to some extent except Food stores. Different from the rest, Food store achieved its peak in April and only had a few store registered in July. Furthermore, it should be pointed out that there was not necessarily a increase in November for each category. We can see a large number of stores registered in November mainly because of Shopping stores which weighed heaviest among all.



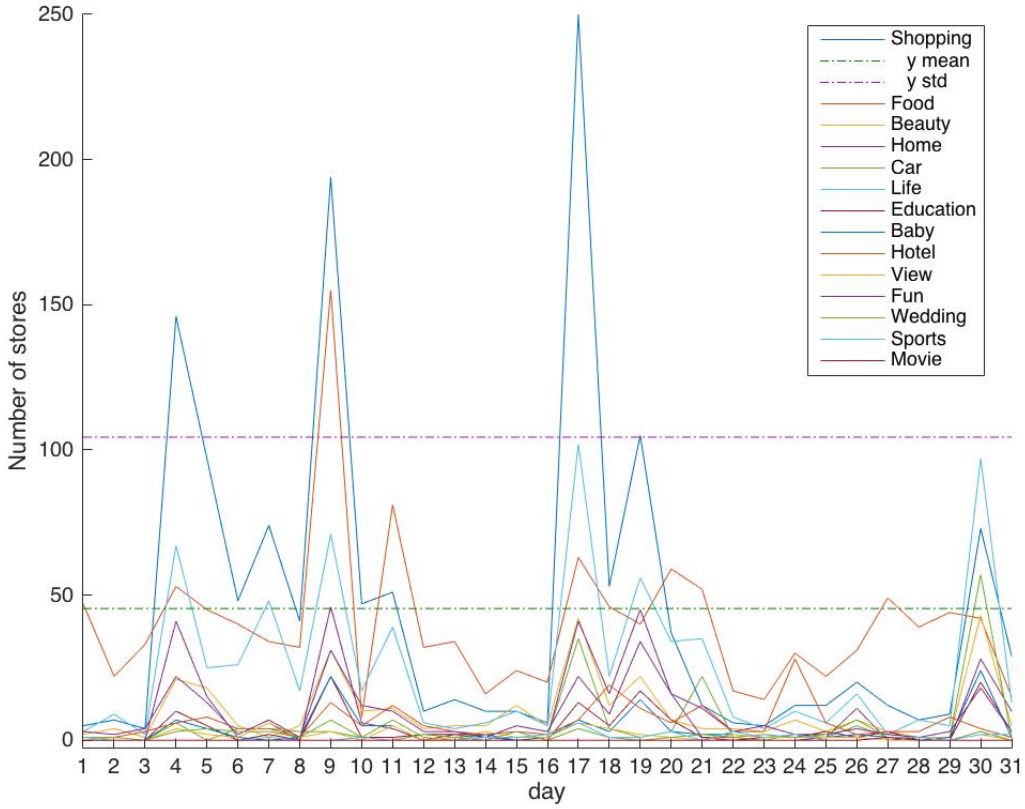


Figure 17. Distribution of store registration among 31 days of a month

Surprisingly, the distribution suggests a period and most categories conform to the overall distribution. The middle of a month had most stores registered and a lot of store registered in the end of a month and the 9th day.

Combine with location

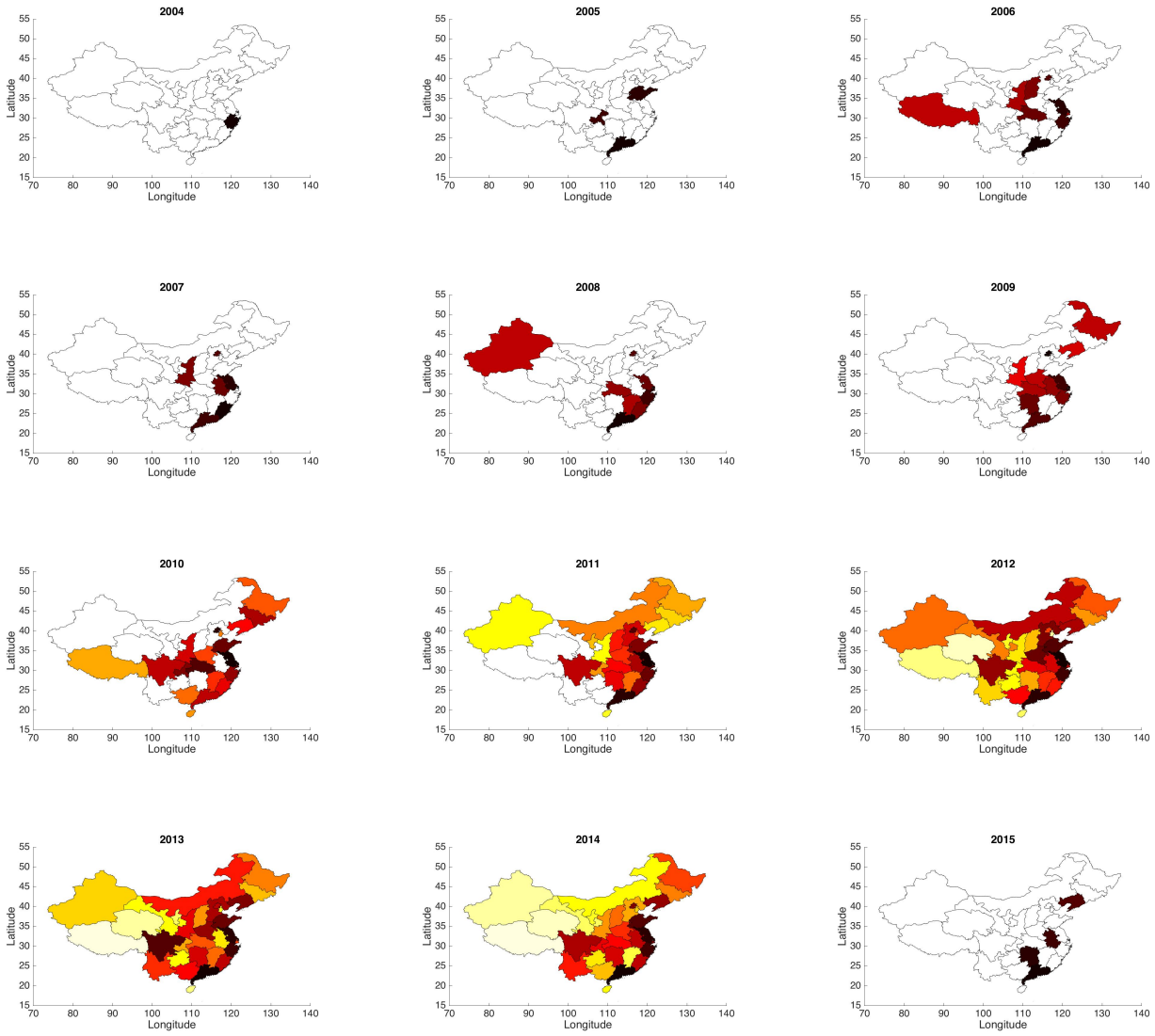
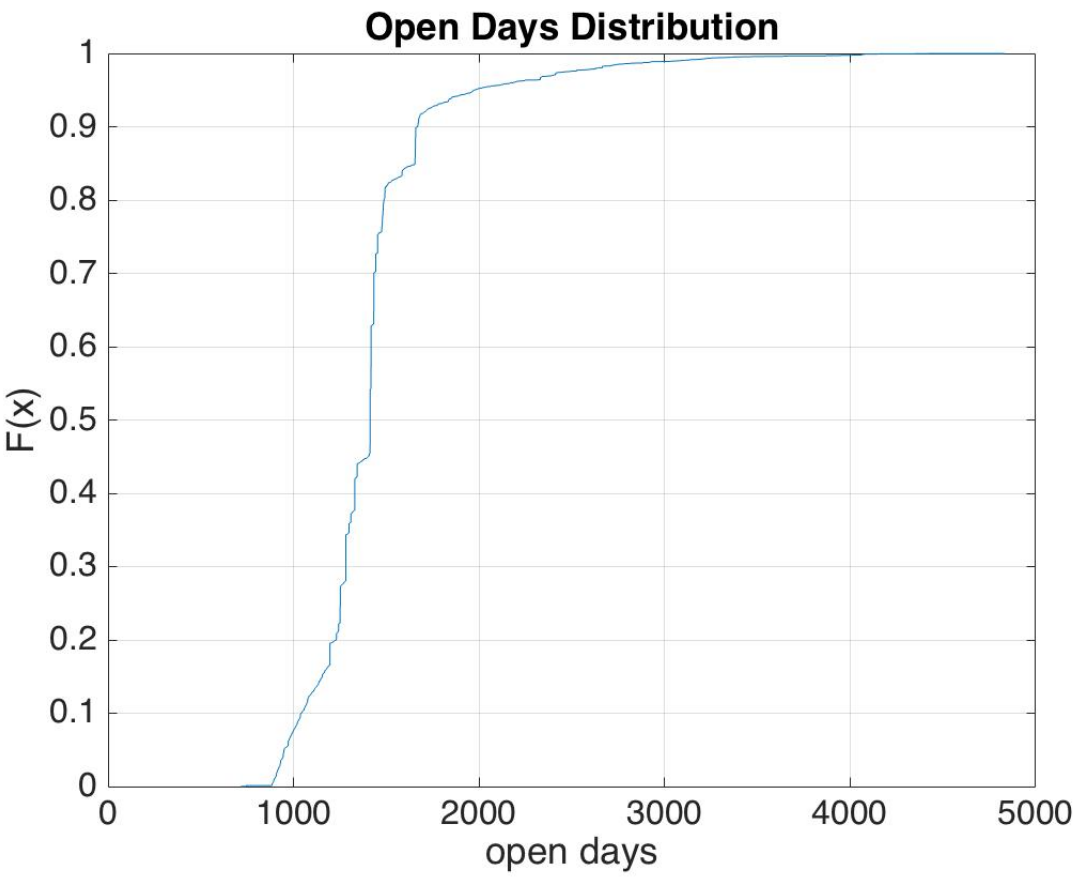


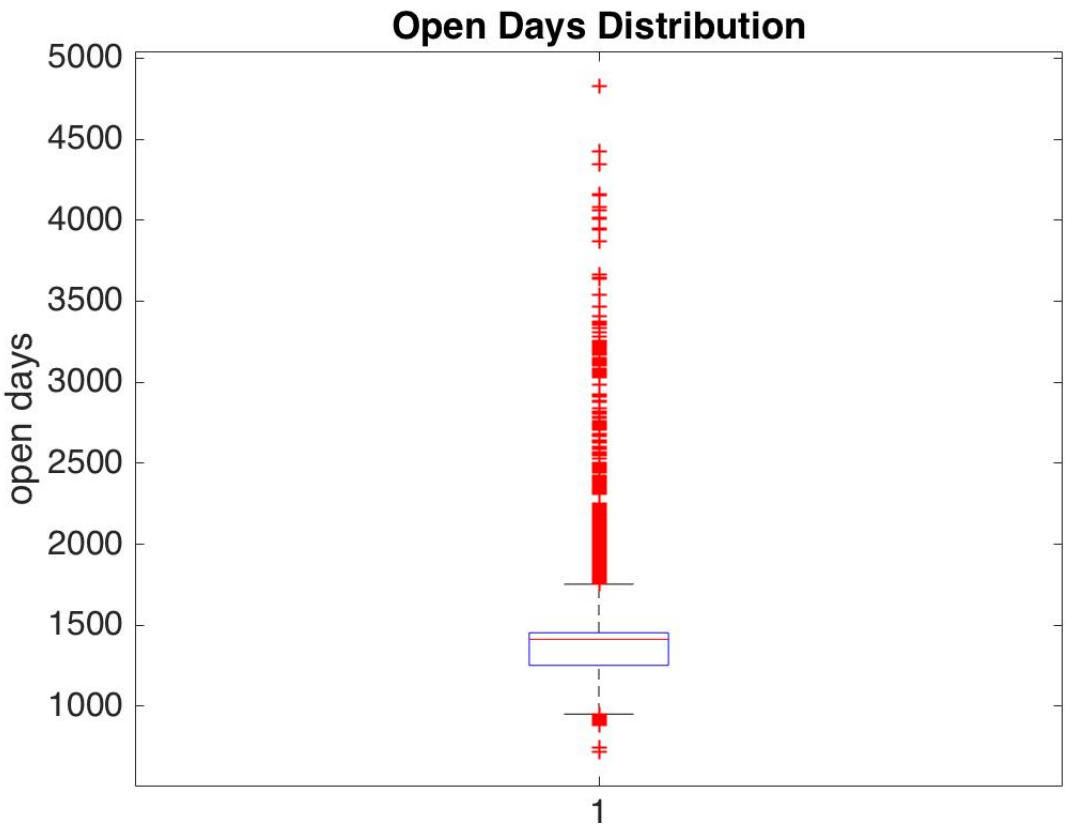
Figure 18. Temporal distribution of store registration in each region

We can see a dynamic development of Dianping in China mainland.

1.3.2. joined days

Use matlab function datenum, we get the joined days of a store, i.e. days between register date and today (May 30th 2017).





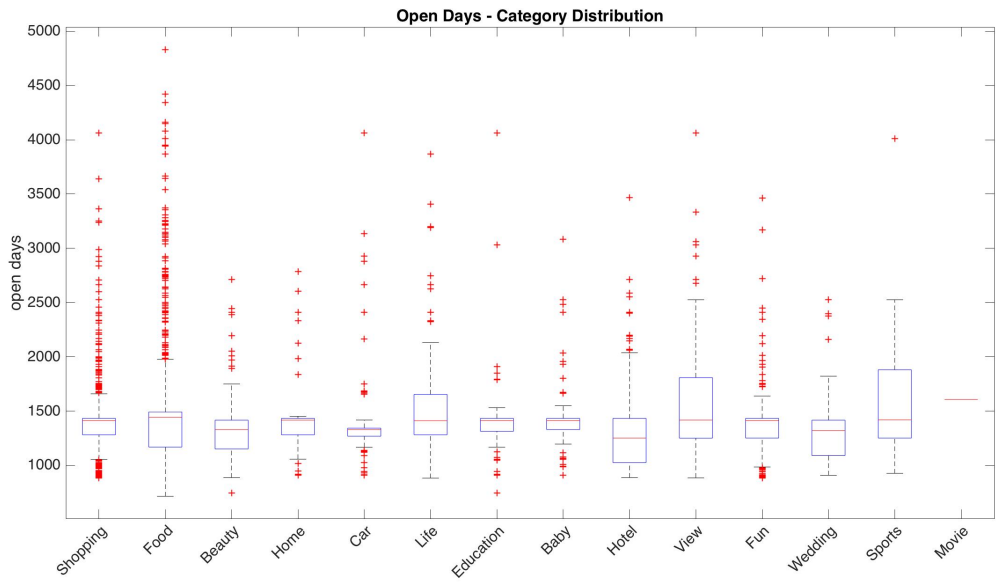


Figure 19. Distribution of joined days of stores

Since our data only includes stores registered before year 2015, the fewest days are 716 days. Figure 17.1 shows the CDF (Cumulative Distribution Function) of joined days, and Figure 17.2 and 17.3 are box graphs. The distribution is still a skewed one.

2. Deep into Review feature

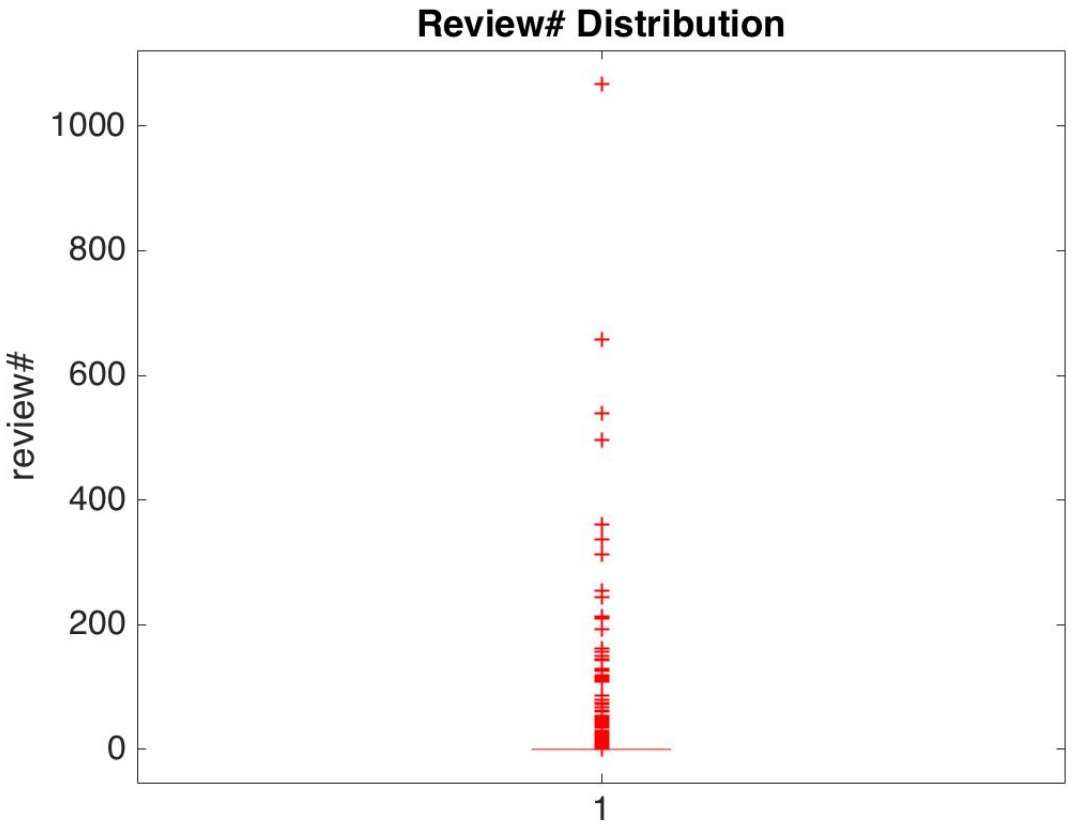


Figure 20. Box graph of review distribution

The review distribution is seriously skewed, thus we should implement a log function. The following CDF use special axis.

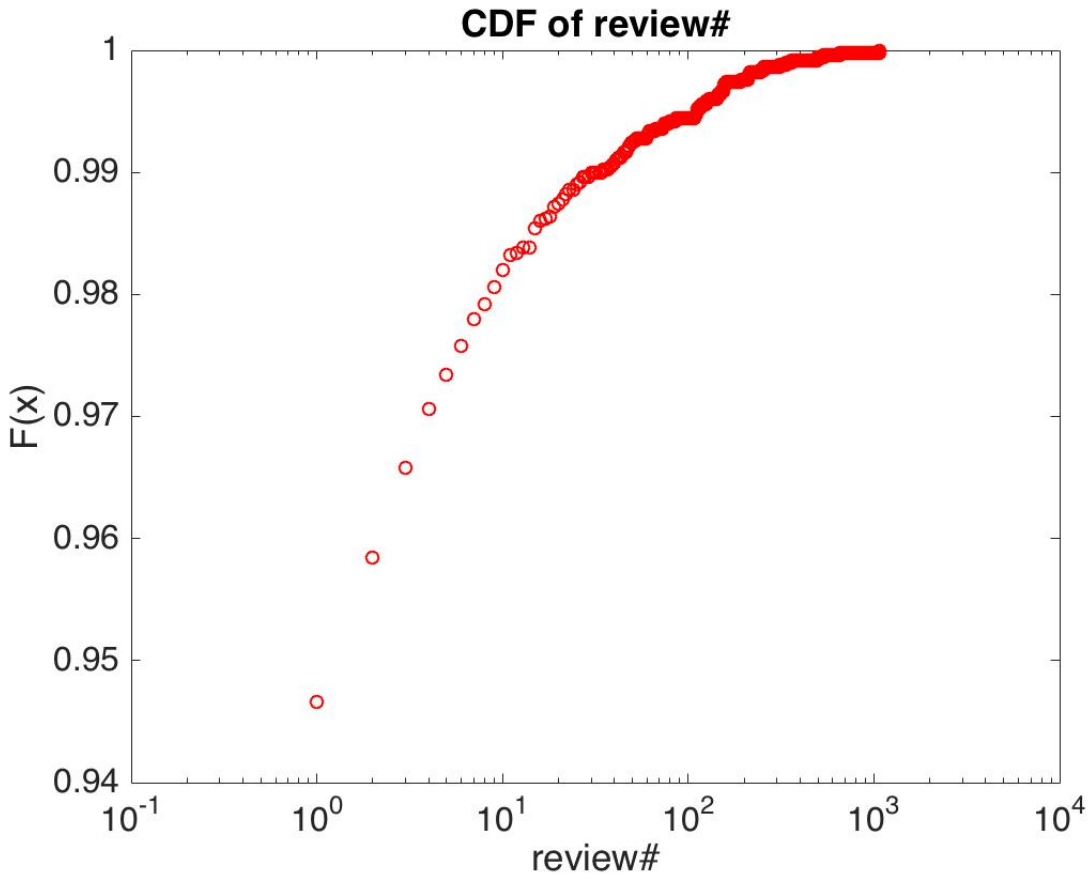
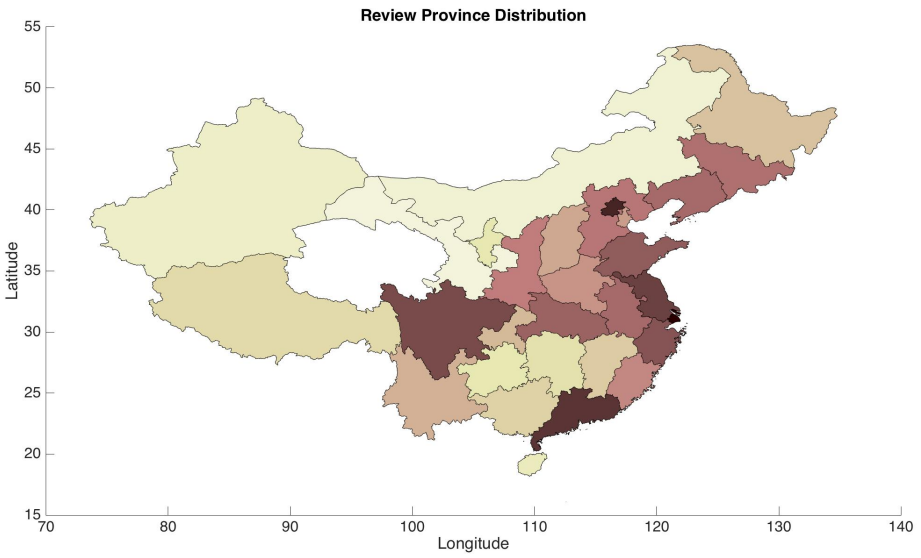


Figure 21. Log-x axis of CDF of review amount

Most stores only have 0 or several reviews, only an extremely small part have hundreds of reviews. And the maximum review amount is over 1000.

2.1. Relationship between Review and Verification



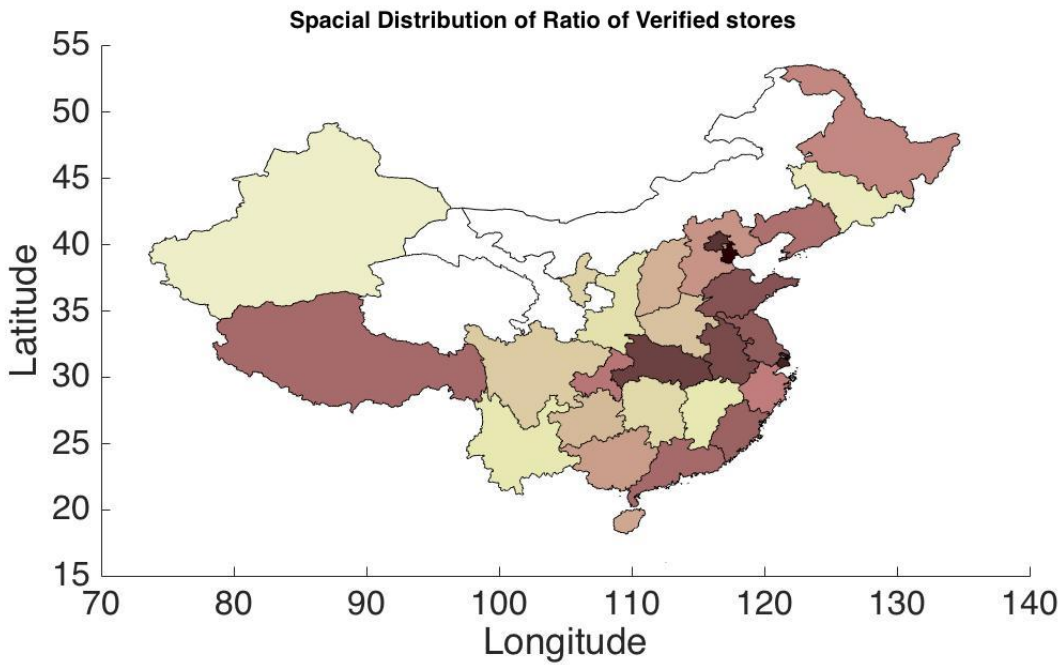
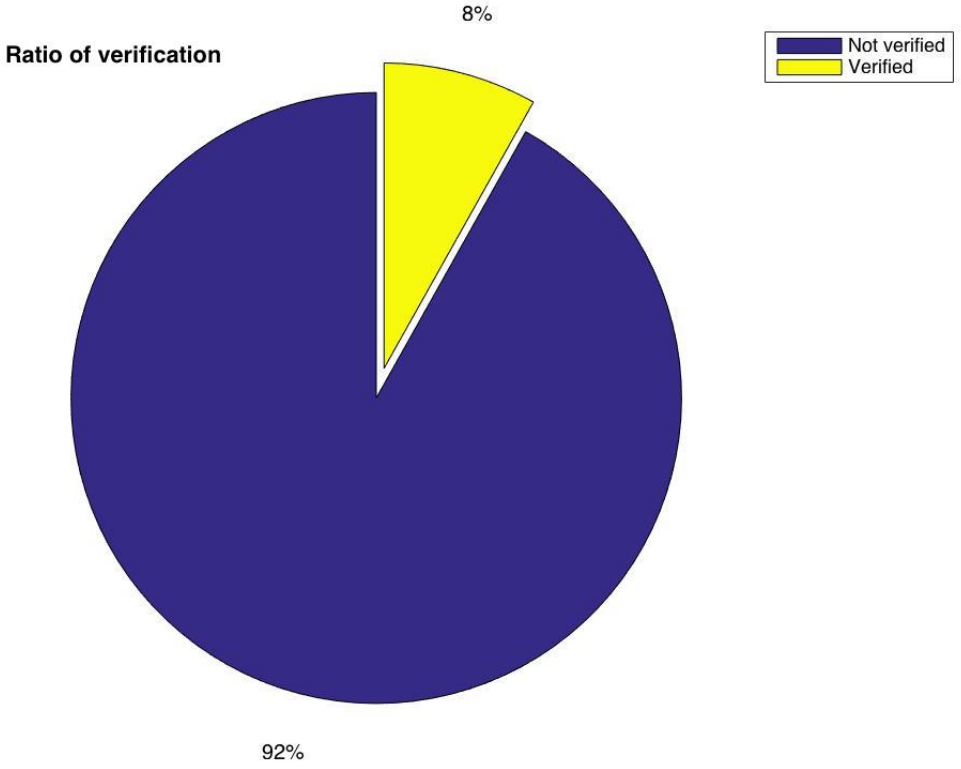


Figure 22. Spacial Distribution

I find the similarity between spacial distribution of review amount and that of ratio of verified stores. The deep-colored provinces are similar, and for those rank low in Figure 20.1 rank low in Figure 20.2 as well. Thus I assume that review amount and verification is positively related.



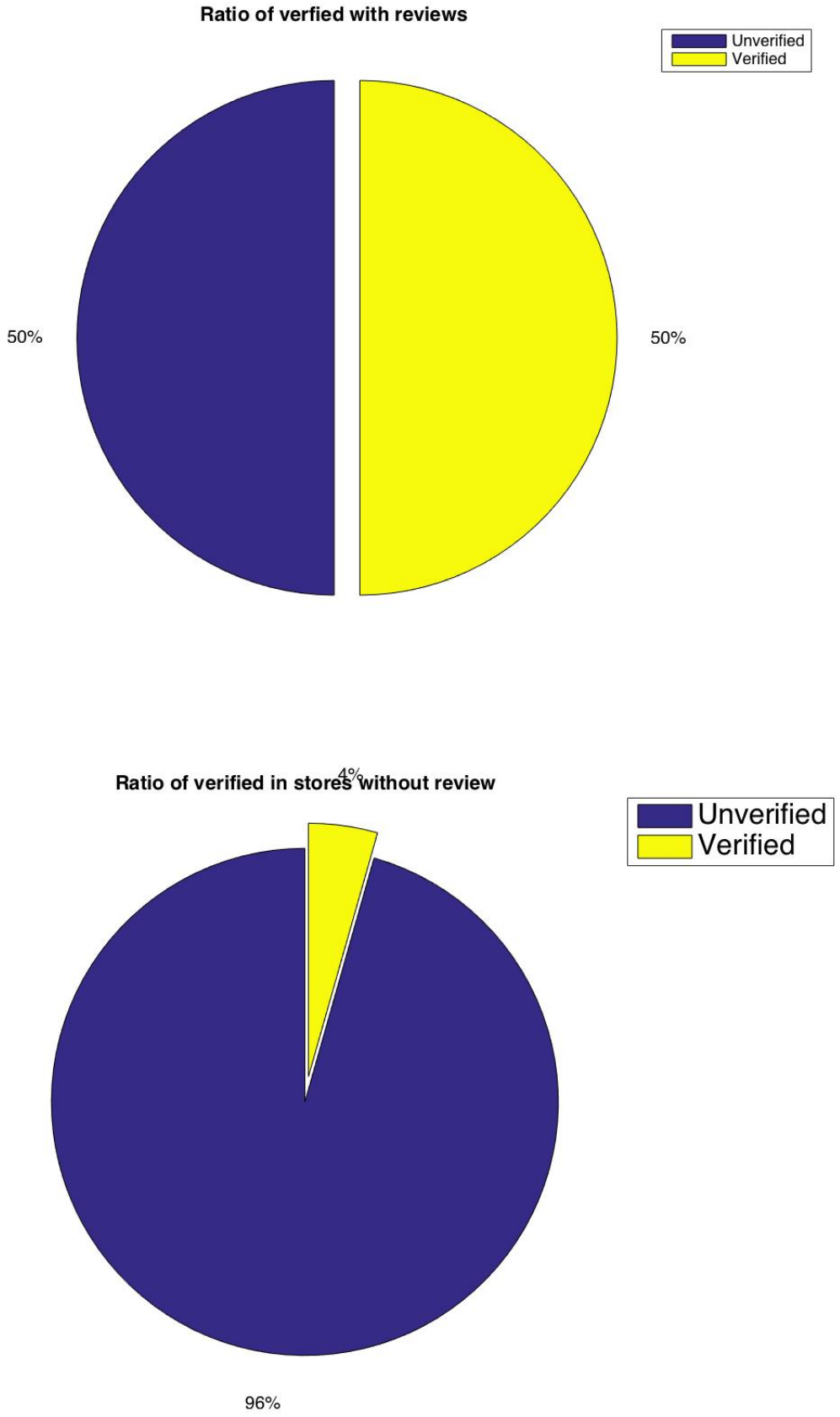


Figure 23. Ratio of verification

Ratio of verified stores in stores with at least one review is much higher than that in stores without any review.

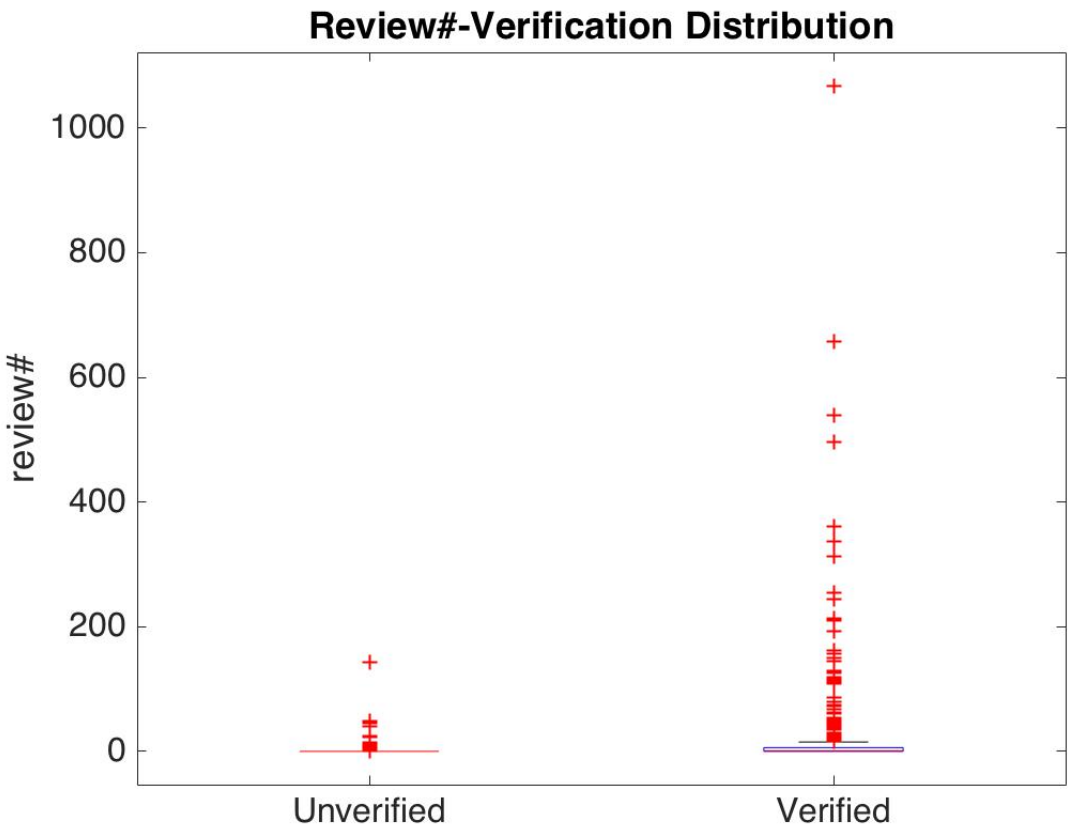


Figure 24. Box graph to show relationship between review amount and verification

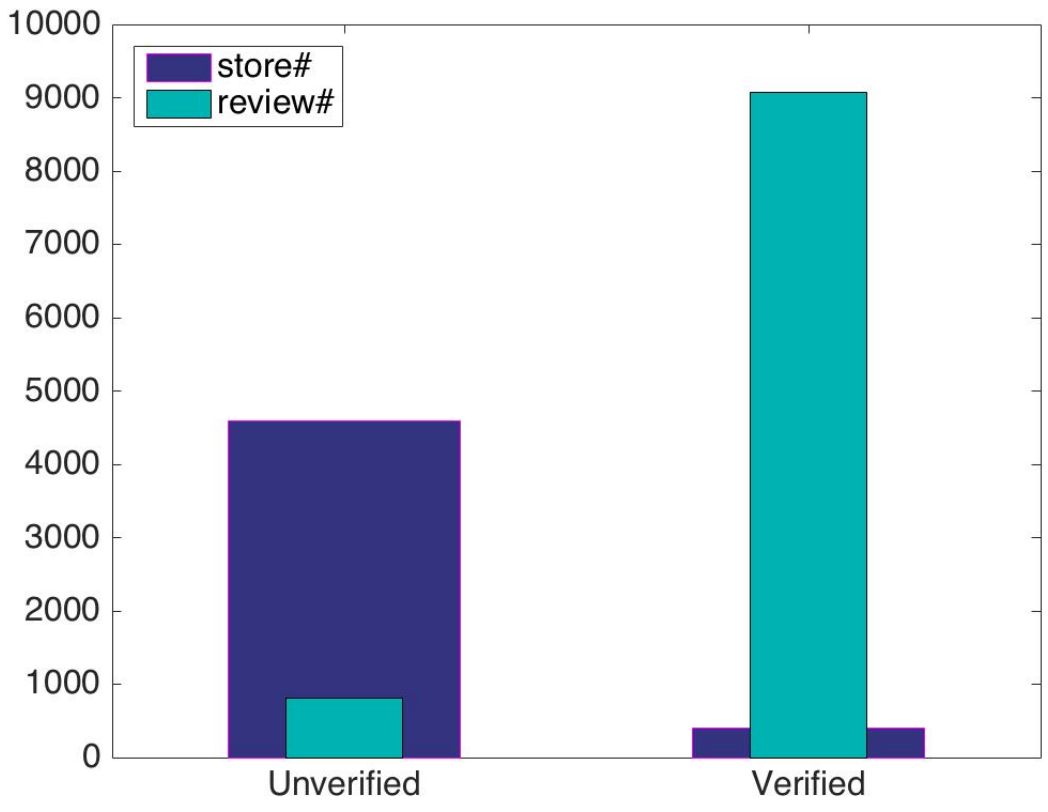


Figure 25. Distribution of store amount and total review amount based on verification

The majority of the stores are not verified, but the majority of the reviews were for verified stores. Thus the average review amount per store for verified ones and unverified ones differ greatly.

Remark. Verification feature should be included to predicate whether a store have reviews.

2.2. Relationshop between review amounts and joined days

Based on common sense, I assume there should also be a positive correlation between review amounts and joined days. The longer a store joined Dianping, the more reviews it receives. Combine with verification feature, a heap map is needed.

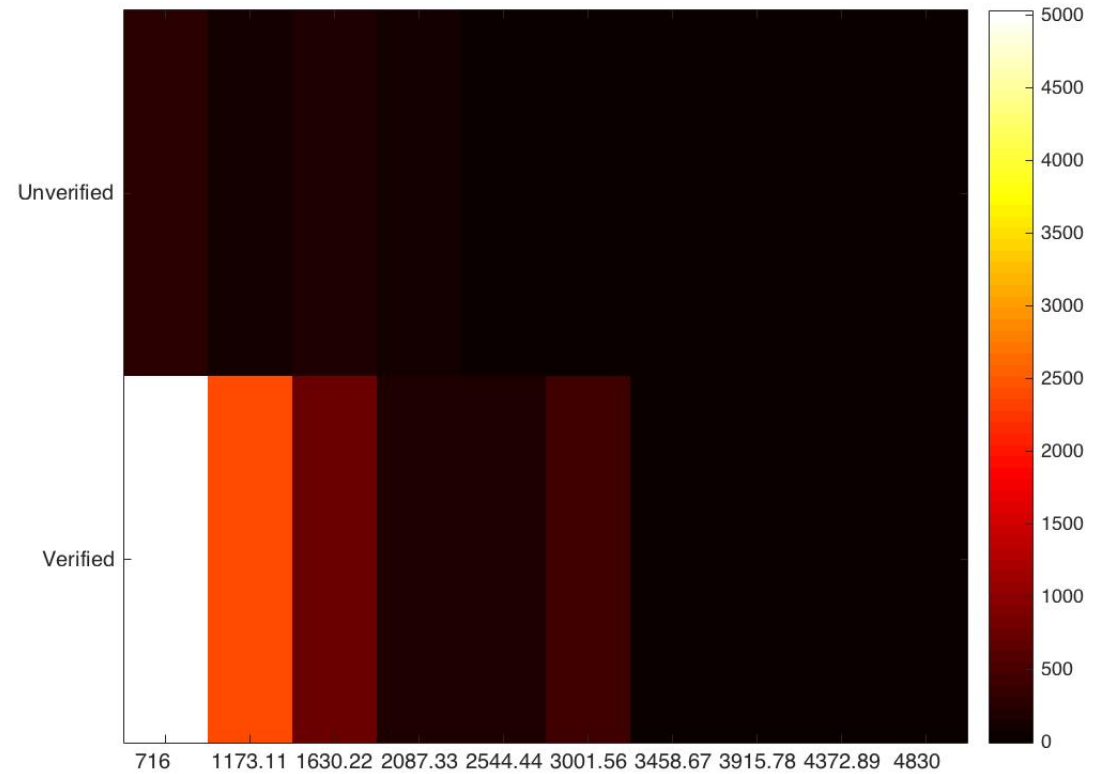


Figure 26. Heap Map of review amounts with verification and joined days

The minimum joined days are 716 days, and the maximum is 4830 days, I divide the range by 10, thus put all data into 10 buckets. For each bar, the left value is included but the right not.
As the color bar shown, the light the bar is, the more reviews it represents. To my surprise, review amount does not increase with joined day. The heat map once again prove the positive correlation between review amounts and verification.

Note. However, different distribution of review amounts with verification is shown, which indicates there is correlation between joined days and verfication.

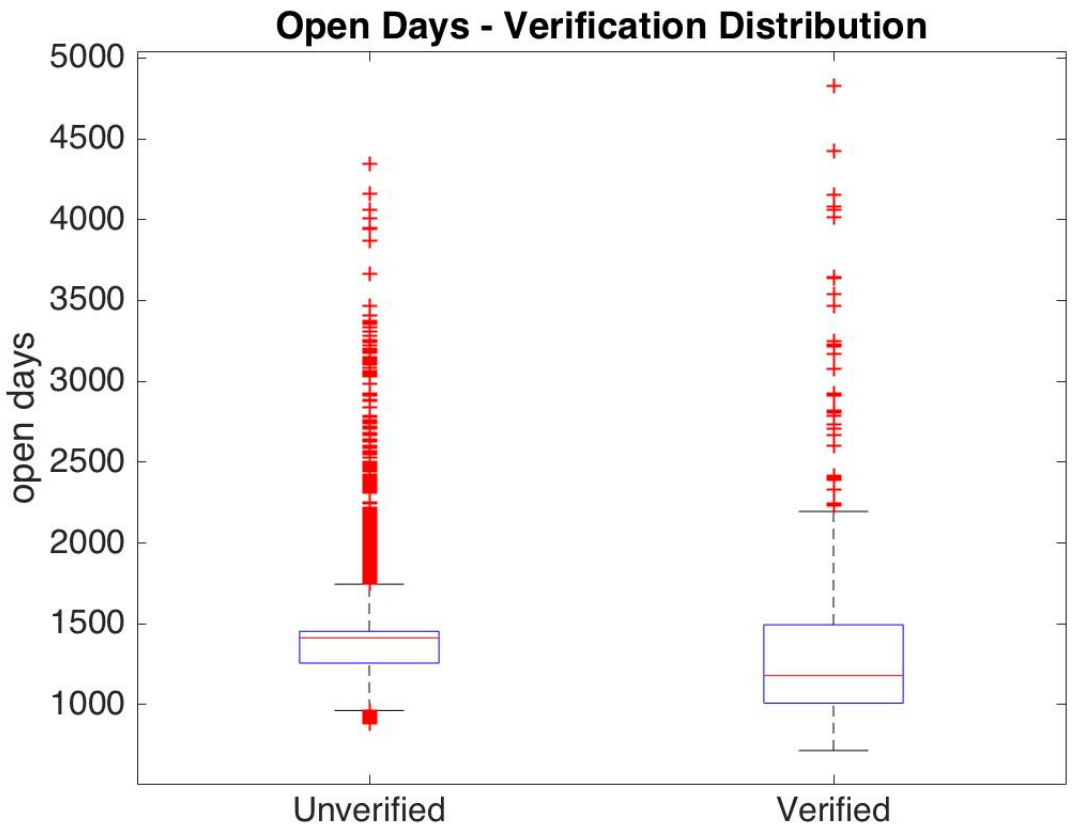


Figure 27. Box graph of distribution of joined days to verification

Distribution of unverified stores clusters, while that of verified stores is sparse, with a larger range and greater standard deviation. And the mean value of joined days of unverified stores is greater.

---

1. *StatPlanet* is an app for plotting world map, it provides data of countries territory. However, since it is not made by Chinese, the China map is not accurate and the name of some region is kind of weird. For example, *Shaanxi* is 陕西省, *Shanxi* is 山西省, *Nei Mongol* is 内蒙古自治区, *Ningxia Hui* is 宁夏回族自治区, *Xinjiang Uygur* is 新疆维吾尔自治区. And Taiwan, Hong Kong and Macao are not included. For convenience, Paracel Islands/ Xisha Islands are not plotted. In another word, all my diagram focus on China mainland. ↩

2. A nomenclature in topography. From the North-West of China to the South-East, the land descends in steps like a terrace. ↩