CSE590-Fundamentals of Data Science

Mini Project 3

Rahul Rishi Sharma: 110347475

Yelp: Exploratory Data Analysis

Introduction

Today with the help of Yelp, people are able to make quick decisions for which they previously had to think over a lot. Users are now able to make decisions based on average ratings given to various business. There is lot of data Yelp has been providing people for doing analysis and coming out with intersenting prediction and recommendation.

The goal of this project is divided into two parts.Firslty to visualize the data in various forms to predict some patterns.Second is to analyse the data and find come up with fun, innovate and creative ways to generate great new insights with the help of Claims and recommendations. Collectively doing exploratory data analysis.

Dataset

The datasets used for this project are: businesses, reviews, users.

The business dataset, relates to the business establised in United States and registered with Yelp over the past few years and each business has a uniques business_id.The review dataset contains reviews a user gives for a business and each review has a unique review_id. The user dataset has the user information and the each user is identified by a unique user_id

Appraoch Taken

The approach is to analyse all the three datasets and try to present some interesting facts through visualisations. The dataset downloaded was in json formated which also required parsing the data before the analysis could be done.Below are the phases the project was undertaken

- Parse and Clean the data
- Exploratory Data Analysis

Cleaning and Pasing the data

The data downloded was in a json format, it was first parsed into three different csv files each for business,reviews and users. The data was huge and to analyse the data using pandas I needed to reduce the size, this was done by removing the extraneous columns in the files not need for the analysis.. The three datasets were comined into a single dataset with the following relationship model.

(Review(business_id) JOIN Business(business_id))(user_id)JOIN User(user_id)

Exploratory Data Analysis

State Wise Analysis

The datasets had a lot of scope for exploratory analysis with respect to all the three dimension of datasets.The first task was to identify on the map the locations from where the data was collected.

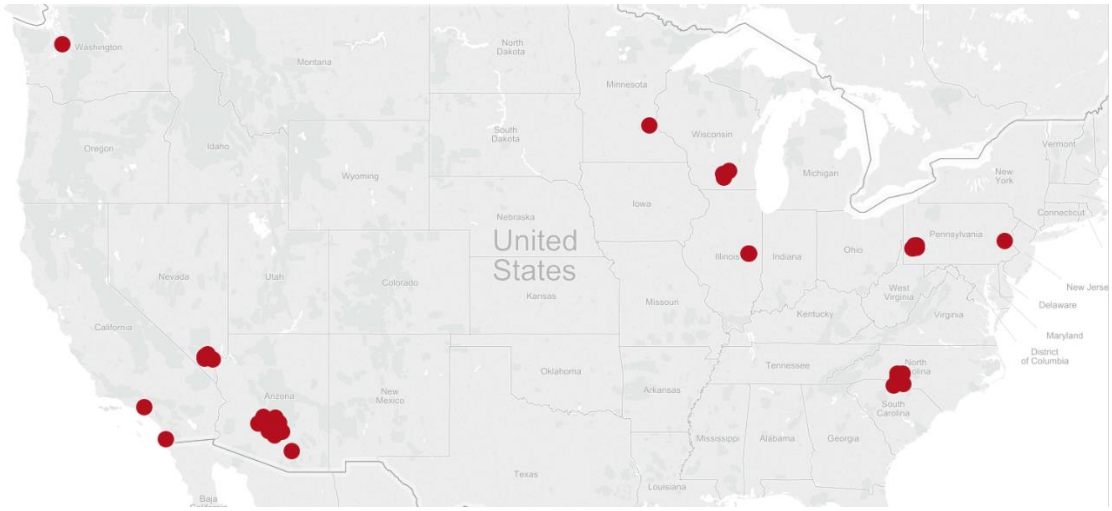


Figure 1

As it can be seen the majority of the data was collected from the states Nevada,Arizona, North Carolina and South Carolina.This data also gives us an idea about the business distibution on the map.

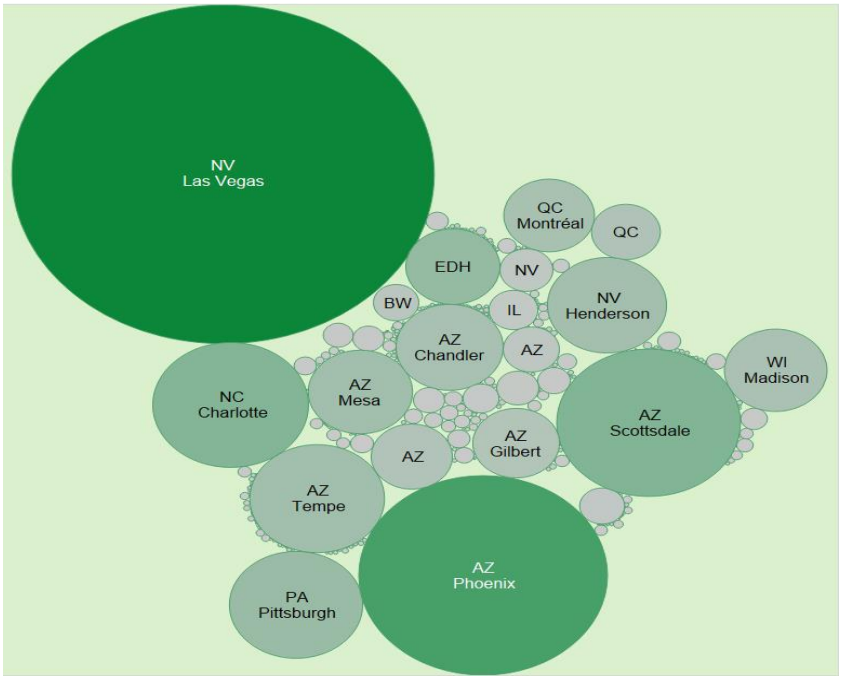


Figure 2

Figure 2 gives us a clear picture about the famous places with respect to review count, i.e people there are truly active in using Yelp. Las Vegas and Phoenix stand out from others substantially in the above graph and it makes sense to delve into the data of these states more to get more intresting insights.

In order to understand which category business was contributing more towards the high review count of both the states , I divided the combined dataset based on catergories. Figure 3 tells us which category dominates in the review data.

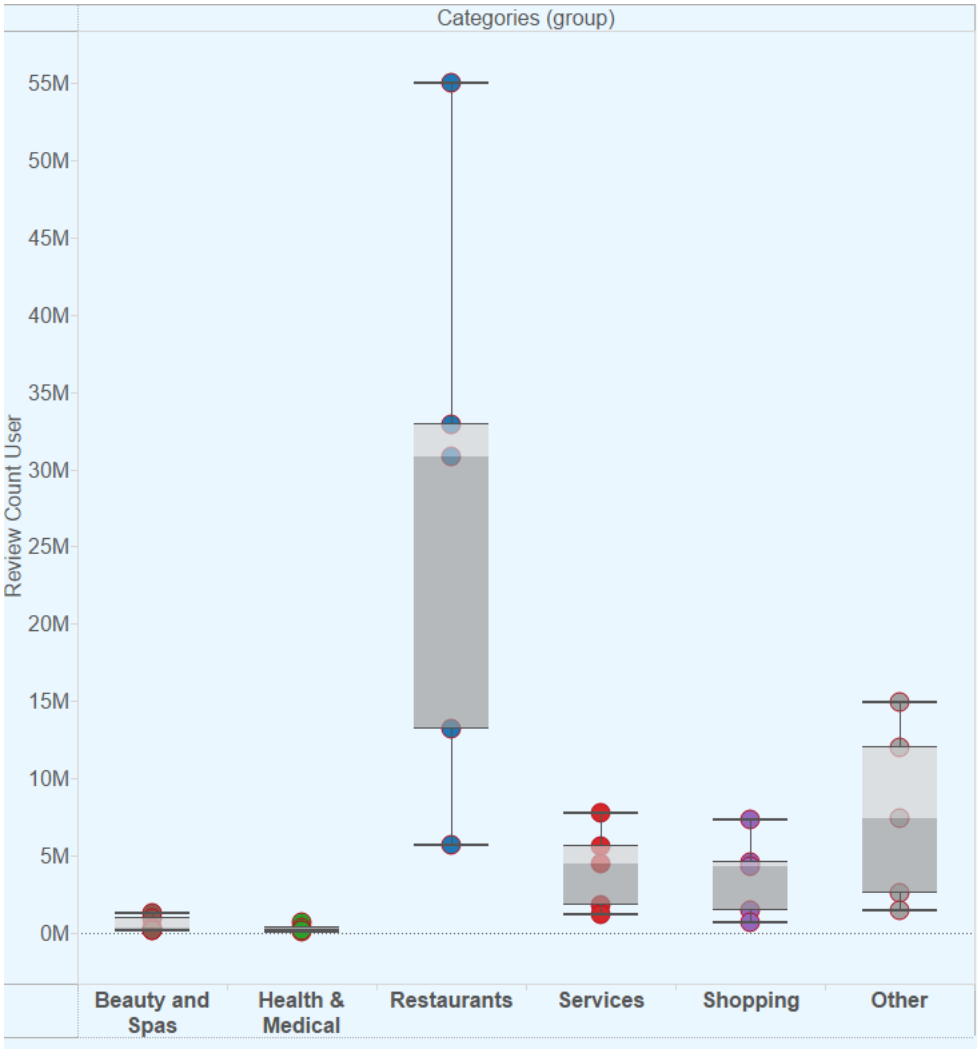


Figure 3

The data was divide based on what values were seen in the categories column of the business data. The other group also looks big, but it contains data from various fields like education and arts , Bakeries , etc.. as the Restaurant data was already greater than other I did not group it further.The dots in the box plot signify the stars(1- 5) 1 being the closest to x – axis.

So now since we knew that restaurant data is the largest, I decided to analyse this data for the states AZ and NV.

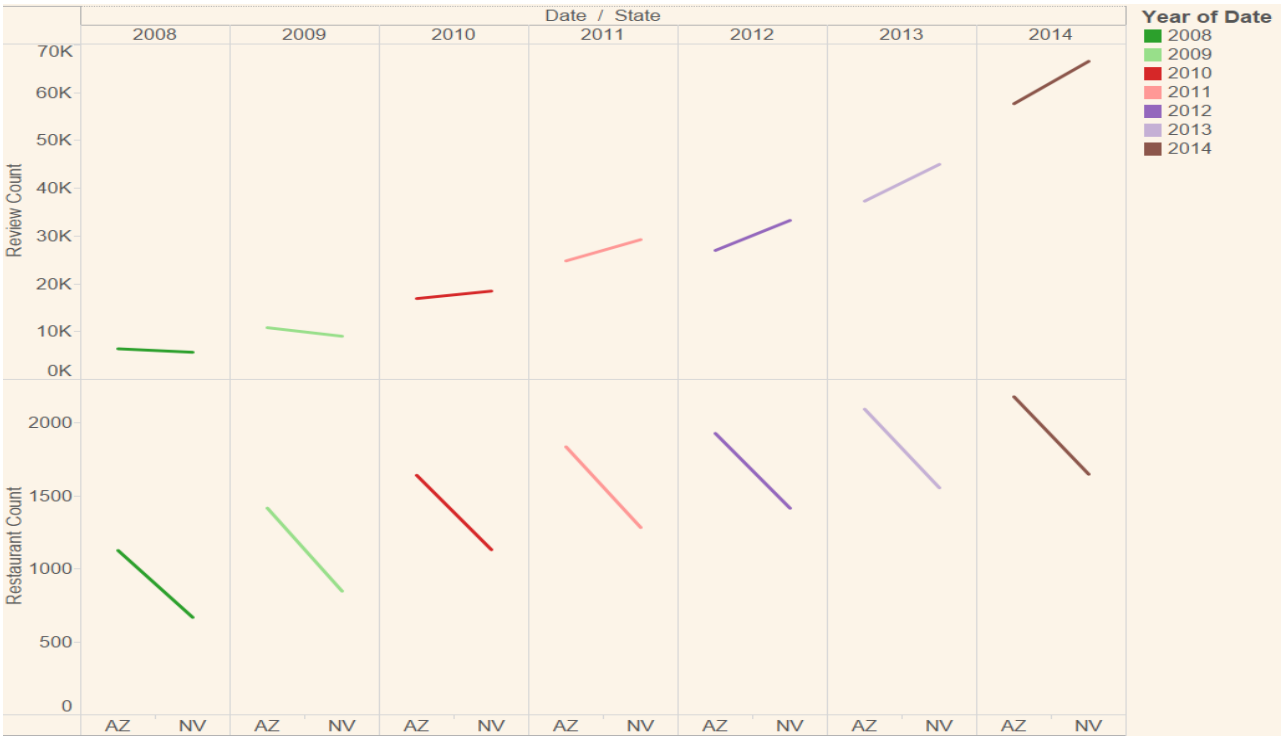


Figure 4

The top graph shows, how the review count has increased over the years[2008 to 2014] and bottom graph shows how the restaurant count has increased for both the states. There is an intresting fact this graph depicts: During all the years AZ seems to have more number of restaurants than NV, but the review counts for them are opposite from the year 2010 onwards. There is a sudden change for the review count from the year 2009 to 2010, and there after NV has been topping in restaurant reviews. From this data we can interpret NV has been developing over the years, and also places like Las Vegas attract a lot of people and so the restaurant reviews are increasing. Figure 5 below justifies this fact, it shows the top 10 hot restaurants in the given dataset, the hotness was classified by the number of reviews obtained.

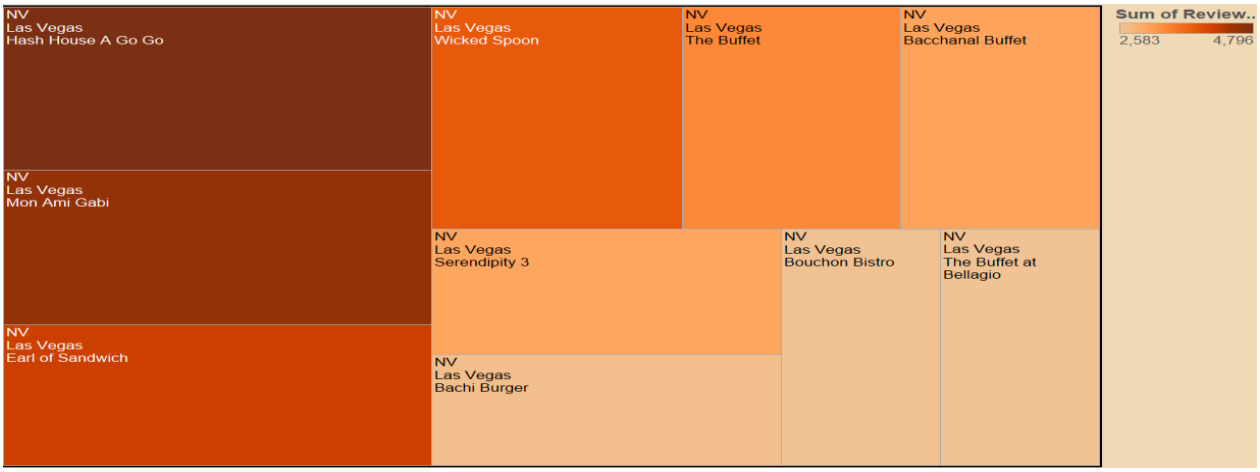


Figure 5

Reviews Analysis

As we see in Figure 4, how over the years the reviews have been increasing in both the states, I tried to analyse the review count over the years.

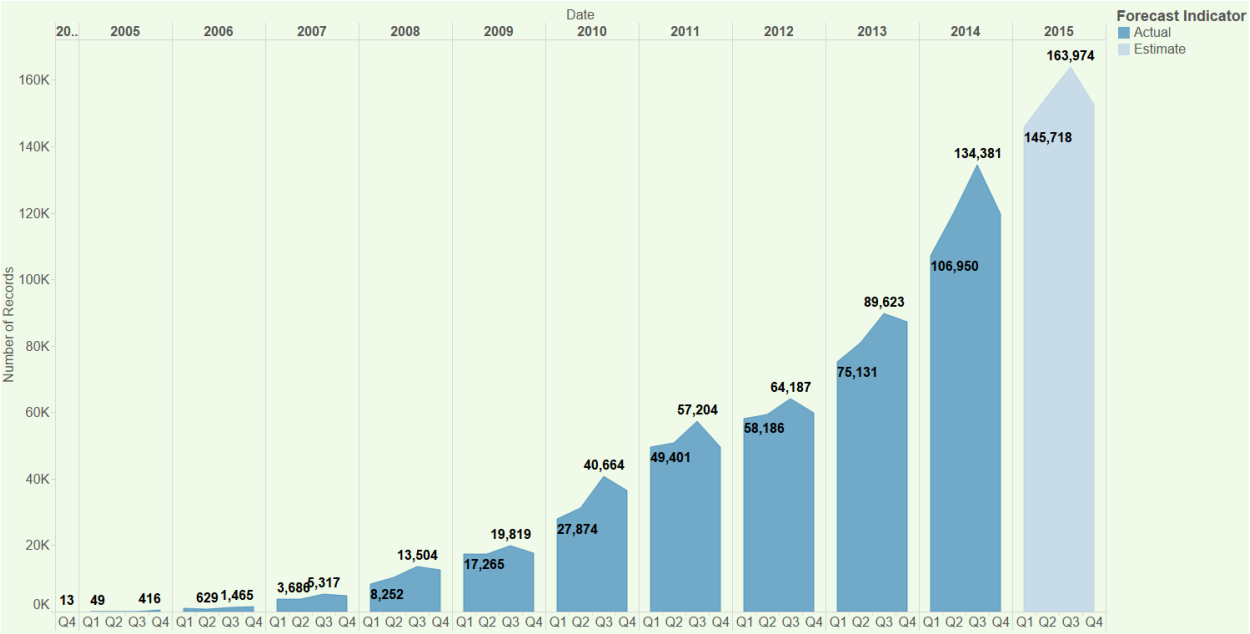


Figure 6

It can be seen from Figure 6 that the total number of reviews in a year has been increasing. Also an intersting trend to note is that people tend to review less during the first quarter and the count increases till third quater of the year, and there is visble drop in the fouth quarter. Taking into advantage the increasing and decreasing ratio per quarter over the years, I was able to predict an approximate value of business reviews that could be seen during the year 2015.

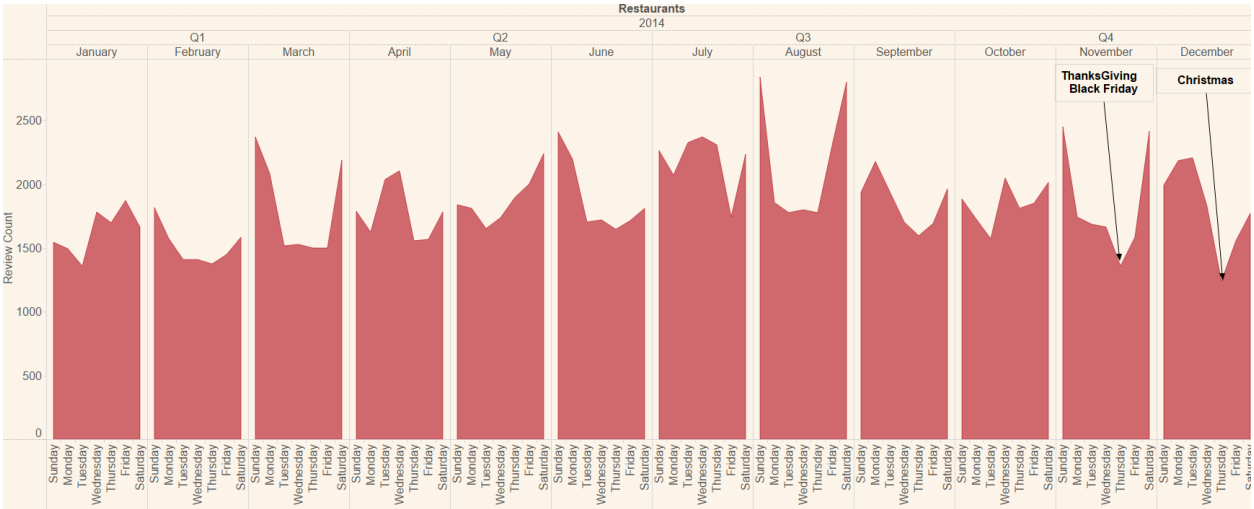


Figure 7

In order to analyze the treand per quarter,I took the year 2014 and plotted the graph in Figure 7. It can be seen how the number of reviews are more during 3rd quarter as compared to the other quarters. Also another trend to node here is the 'U' shapped curve on the top obtained in most of the months. The lower axis shows the weekdays during the months and it can be seen that people tend to write reviews more during the weekends(Saturday and Sunday). I also tried analyzing the certain dips seen during particular months, that has been justified by national holidays, As seen people tend to be busy with families on occasion like Christmas and the review count decreases as expected.

Lets us now assume the below claim for getting more insight into reviews and stars relationship:
H_o : As the restaurants,people,reviews are increasing year by year , so is the competition in the market. This implies it would be very difficult for restaurants to get a complete 5 start rating, owing to the competitive market.
H_a : This is the alternative hypothesis , by saying that the above statement is false and that better reviews for restaurants are obsereved in the competative market.

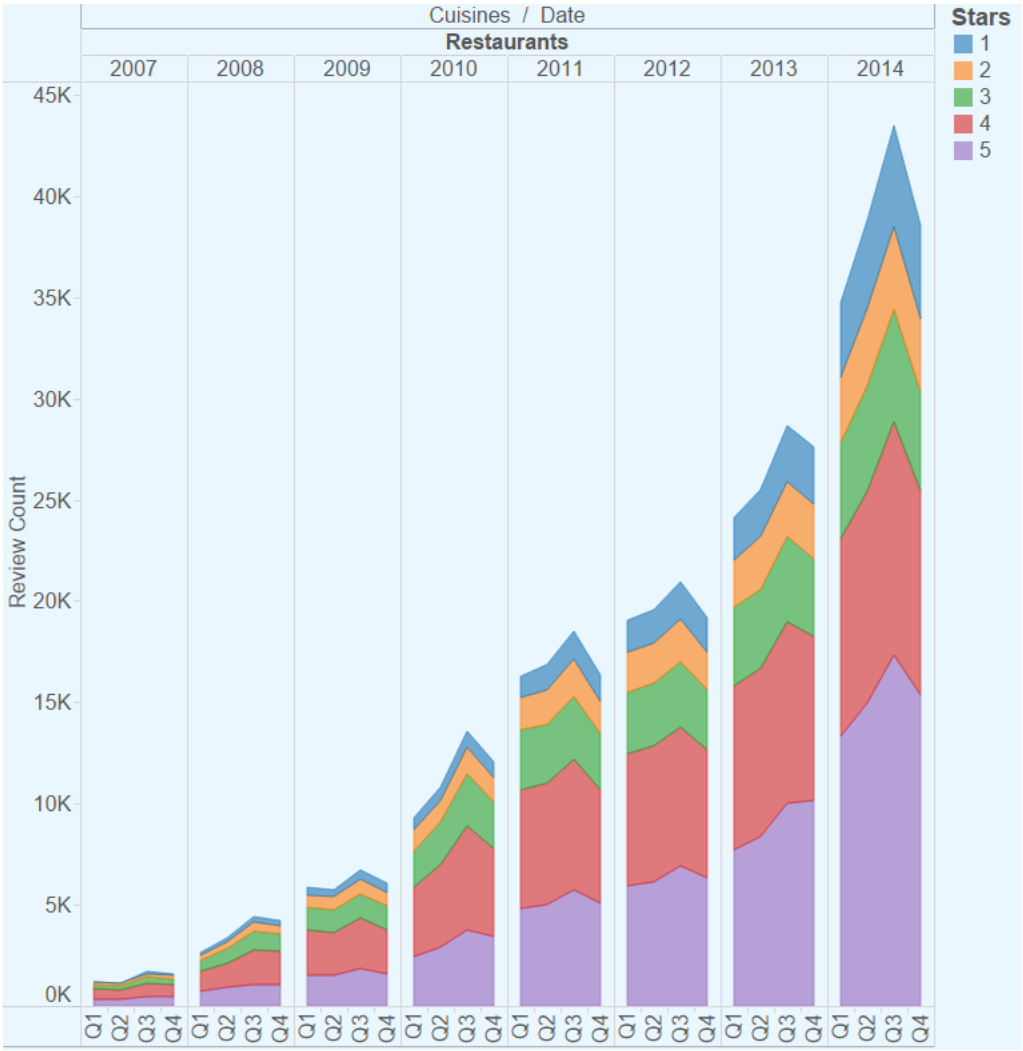


Figure 8

We prove the alternate hypothesis by Figure 8.

Figure 8 , gives more analysis on the reviews across the years clusterd on the basis of stars given. It is quite evident that the number of reviews for 5 star and 4 star has been increasing more compared to 1,2 and 3 star rating reviews. We can infer that the reviews are more because there are more restaurants opening as the years pass, also the restaurants have been improving as the time runs. Another thing to notice is that the ratio of 5 star reviews to the total reviews is way better than the other rating again justifying our claim that restaurants quality has been improving

Miscellaneous Analysis

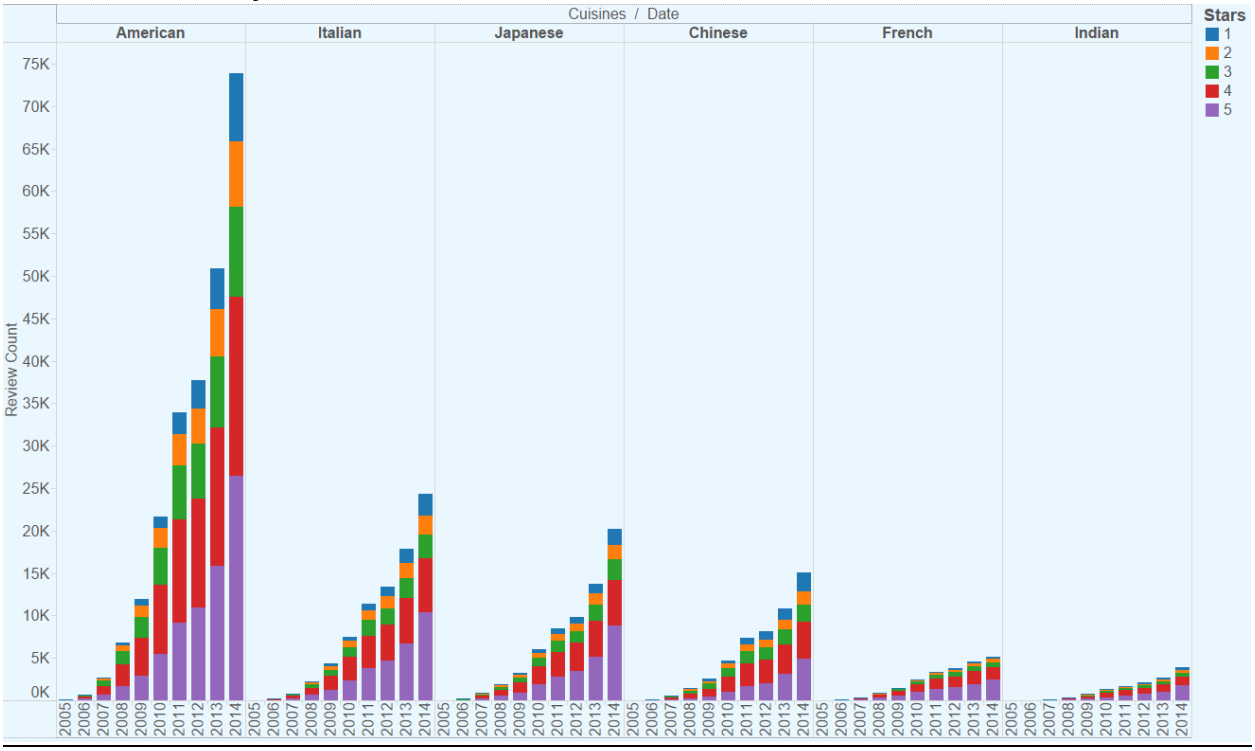
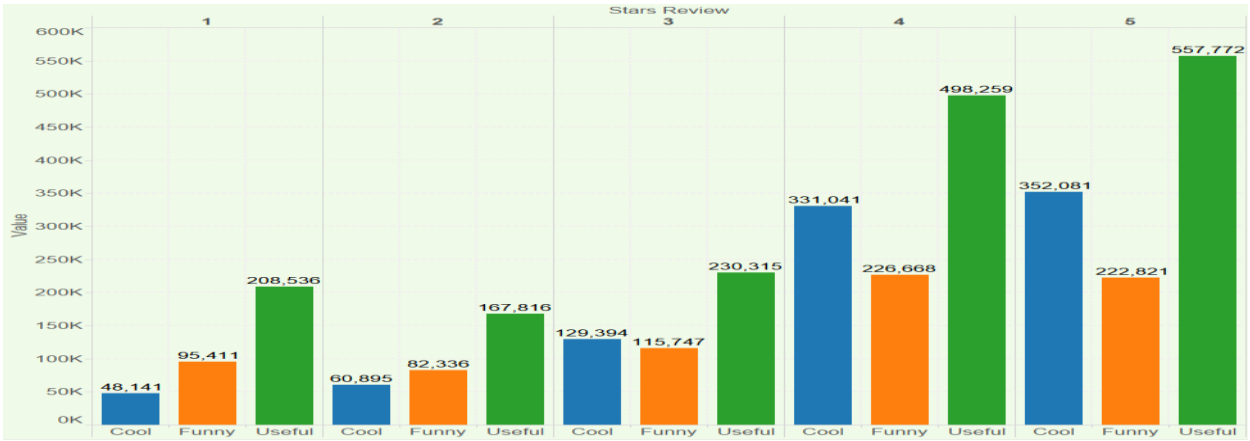


Figure 9

I above graph represents what kind of cusine is liked by the people of USA. Its clreary evident that American food is liked the most , that's is obvious. But the fact to note is some of the other cuisines, there seem to a close competition amongst others. The inclusion of Indian and Chinese cuisines in the top 6 out of the 14 different cuisines is justified by the fact that China and Indian contribute to the 2nd and 3rd largest foreign population.



The above graph given some insights into how what users like to rate other users on their reviews and usefulness of a review seems to be of at most import to the users. May be Yelp should considering sorting reviews on how useful it is for others and mark it as more relevant.

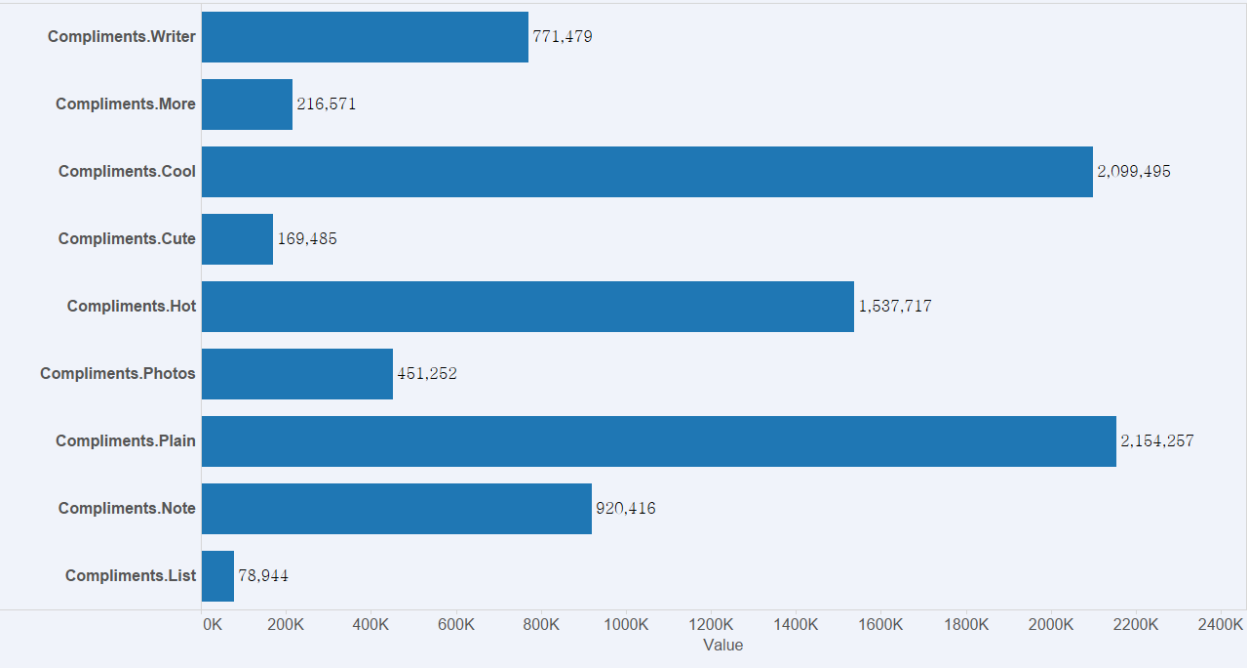


Figure 10

Figure 10 represents total number of complements received for the users over all the years and its relevant to say that List, More are some of the least used compliments attributes featured on the yelp page, this could be because of the not ambiguity of these terms used in daily conversations.

Conclusion

The Yelp dataset has a great scope of data analysis that can be done by just good visualizations. Due to the constraint size on the report, I have tried to present few of the interesting finding. I had also attempted to find linear regression analysis using scikit libraries by predicting the price range of a restaurant based on the location and restaurant attributes. Was able to get a prediction accuracy of approximately 70%. This prediction can be used for upcoming restaurants in deciding their price range.