Exploratory DATA Analysis in R

CSE 587- Data Intensive Computing Project 1

Srinivasan Rengarajan – 50097996

Aravindhan Thanigachalam -50098345

ABSTRACT & INTRODUCTION:

Data Analysis and Mining are some of the most sophisticated techniques available in Computational Processing today. It is an active field of research and focuses on the data modeling and statistical inference. The dependence of knowledge on Complex Engineering Systems relies heavily on the data. Learning from Data hence, assumes significance in Engineering and Computer Sciences, as well as in several other fields like Quantam Mechanics, Astrophysics, Geology, Finance etc.

Data sciences has seen a big revolution in the past decade and is now available in various formats - Images, Videos and meta data in general. Various tech- savvy firms like Google, Facebook, Amazon, Twitter focus much on Data Intensive Computing and there is an increasing paradigm shift towards Algorithms that learn from Data and construct reliable models for further understanding and analysis

In this project, we leverage the computational capabilities of the R programming language in order to understand how to learn from models and infer from data. This report comprises of three parts- Analysis of NY Times Data as given in the Data Sciences book, a cast study based on Real Direct (a Real Estate Firm in New York) and thirdly mining of a dataset from Yelp which contains user reviews, Academic

Exploratory Data Search is a preliminary step in constructing of models for Data and gives us an insight into the breadth and depth of the dataset- also telling us what are the possible inferences that can be made from the data. It also involves cleaning and basic pre-processing of the data.

PROJECT OBJECTIVES :

1. Explore Statistical Modeling in the R Programming language.

2. To understand basic functionality of the R language and use it for processing and analysis of the following datasets :

(i) NY Times Data from Chapter 2 of the TextBook.

(ii) Real Direct Case Study from Chapter 2 of the Textbook.

(iii)Analysis of data from yelp.com – a user review based site. This data contains user and academic reviews, details of businesses , and general data about the city of Phoenix , AZ

3. To make inferences from the Data given- which gives an estimate of the data and the amount of knowledge that can be gained from it.

4. To plot distributions and plots that characterize and summarize the visualization and analysis of the data.

3. PROJECT APPROACH :

As mentioned above, this project is divided into 3 categories – NY Times Data, real Direct Case Study and Analysis of user reviews and business data from yelp.

3.1 – NY Times Data Analysis

The NY Times Data is the one given in Chapter 2 of the Text.

Data Description :
Each one represents one (simulated) day's worth of ads shown and clicks recorded on the New York Times home page in May 2012. Each row represents a single user. There are five columns: age, gender (0=female, 1=male), number impressions, number clicks, and logged- in Chapter 2: EDA on NY Times Datasets

1: Create a new variable, age group, that categorizes users as "<18", "18-24", "25-34", "35-44", "45-54", "55-64", and "65+".

The code for the following is give below :

```
nyt1$age_group[nyt1$Age < 18] <- "<18";
nyt1$age_group[nyt1$Age >= 18 & nyt1$Age <= 24] <- "18-24";
nyt1$age_group[nyt1$Age >= 24 & nyt1$Age <= 34] <- "24-34";
nyt1$age_group[nyt1$Age >= 34 & nyt1$Age <= 44] <- "34-44";
nyt1$age_group[nyt1$Age >= 44 & nyt1$Age <= 54] <- "44-54";
```

```
nyt1$age_group[nyt1$Age >= 54 & nyt1$Age <= 64] <- "54-64";
nyt1$age_group[nyt1$Age >= 65] <- "65+";
```

INFERENCE:

The above code divides the Age column of the NY into sub categories based on the intervals 0-18. 18-24, 24-34, 34-44, 44-54, 54-64 and Above 65.

This gives an impression on how many people viewed Ads in the NY Times are of age 18-24. say. The ads can be categorized according to the stuff they might be interested in- such as clothes, sale of watches, Education. These are examples of Variables which can be inferred from the age_group of people who view the ads.

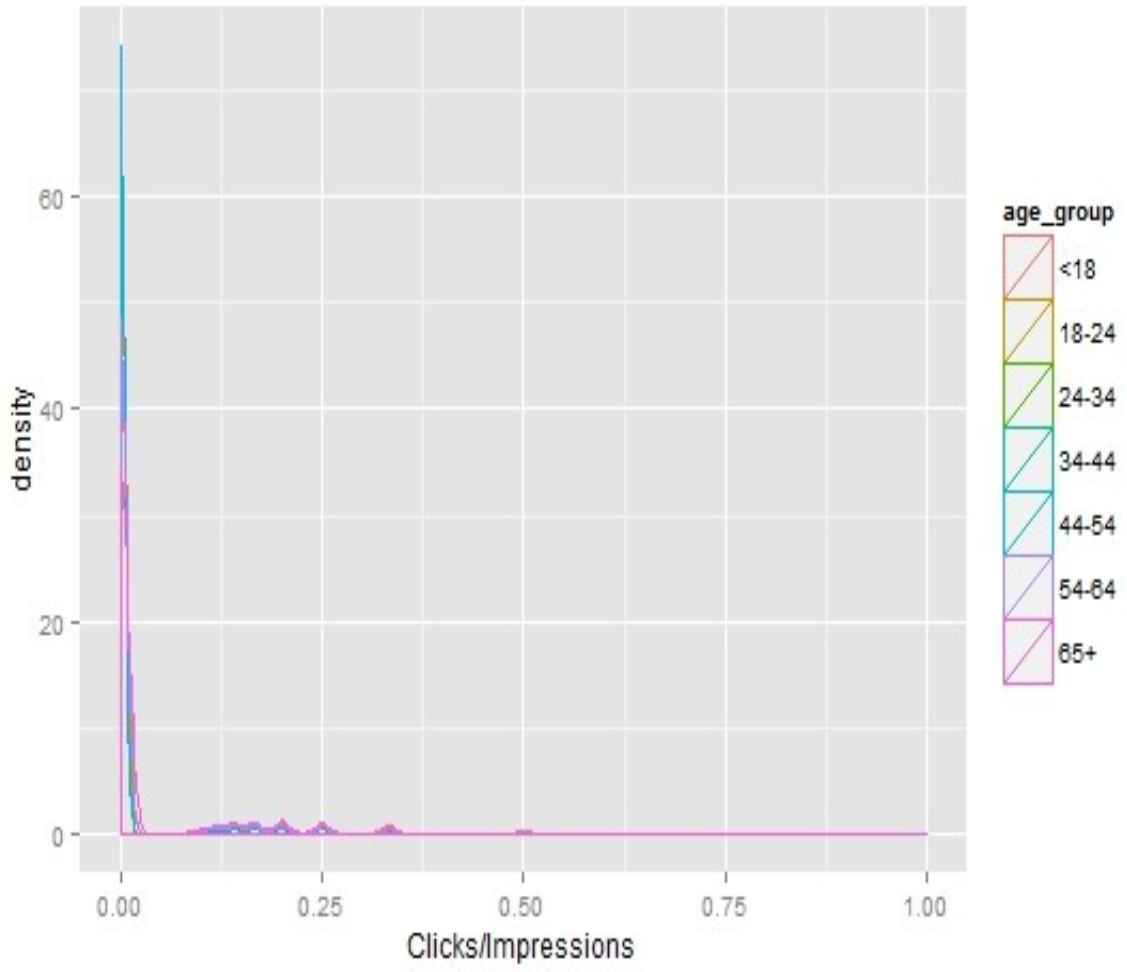
2. **For a single day: • Plot the distributions of number impressions and click- through-rate (CTR=# clicks/# impressions) for these six age categories.**

This questions asks for the rate at which users click on an ad when it is presented to them. Impressions refer to the ads presented to the user. This is again correlated to age categories.

Knowing the Click through rate (CTR) of people in various age categories will reveal the amount of times the particular ad is clicked by people of different age categories. This like, the previous case tells us which add is popular among which age_group.

In addition to the above, it also tells us whether an ad is popular across multiple age groups.

The plot for that is shown below :

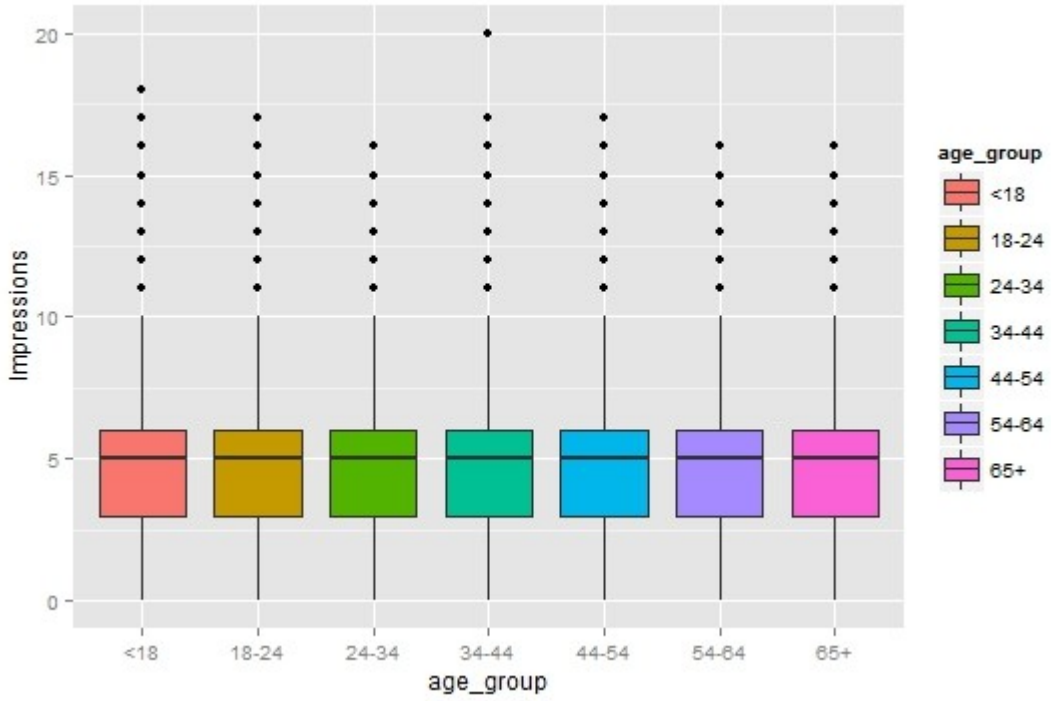


INFERENCE :

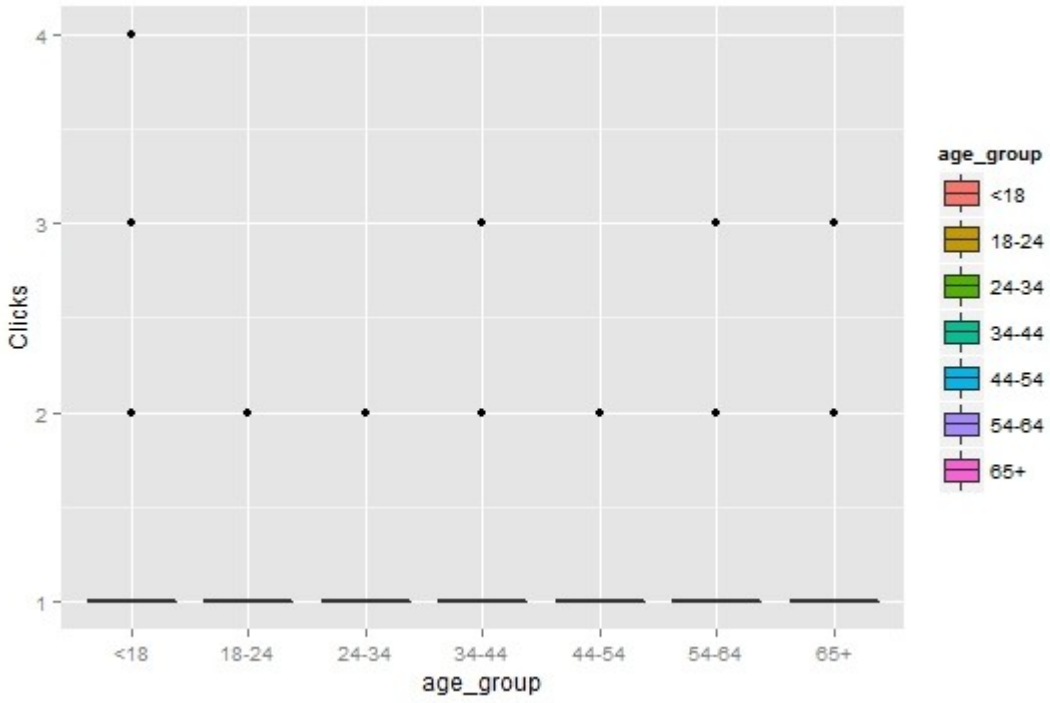
The CTR is more for ages 34-44 and decreases for ages greater than 65.

The figure shown below indicates the plot of Impressions for each age category. This distribution is called the Box plot or the Box and Whisker Plot.

This indicates the quartiles, mean Impressions for various age groups. The age groups are color coded and split into intervals similar to the data shown above.



The above shows a box plot for the Impressions as compared to the Age group. The Impressions refer to the number of ads that are present in the NY Times ad. It shows the correlation between the number of Impressions at various age groups. It shows that on an average the click through rate is around 5.



The above plot shows the number of Clicks in the Age Group. It shows that there is a large number of clicks in the age group of less than 18. Two clicks are almost constant all over the age groups.

THE NYTIMES Analysis of Data Code :

```
#Final R Script - Project 1 EDA in R #
#Aravindhan Thanigachalam 50098345 , Srinivasan Rengarajan - 50097996#
#input the data of each day in order to calculate the metrics and store them in a
#data frame #
nyt1$age_group[nyt1$Age < 18] <- "<18";
nyt1$age_group[nyt1$Age >= 18 & nyt1$Age <= 24] <- "18-24";
nyt1$age_group[nyt1$Age >= 24 & nyt1$Age <= 34] <- "24-34";
nyt1$age_group[nyt1$Age >= 34 & nyt1$Age <= 44] <- "34-44";
nyt1$age_group[nyt1$Age >= 44 & nyt1$Age <= 54] <- "44-54";
nyt1$age_group[nyt1$Age >= 54 & nyt1$Age <= 64] <- "54-64";
nyt1$age_group[nyt1$Age >= 65] <- "65+";
#Question 2#
summary(nyt1);
install.packages("doBy");
library("doBy");
siterange <- function(x){c(length(x), min(x), mean(x), max(x))};
summaryBy(Age~age_group, data =nyt1, FUN=siterange);
```

 # so only signed in users have ages and genders
summaryBy(Gender+Signed_In+Impressions+Clicks~age_group,data =nyt1);
install.packages("ggplot2")
library(ggplot2)
ggplot(nyt1, aes(x=Impressions, fill=age_group))+geom_histogram(binwidth=1);
ggplot(nyt1, aes(x=age_group, y=Impressions, fill=age_group)) +geom_boxplot();

CTR Calculation :

 # create click thru rate
 # we don't care about clicks if there are no impressions
 # if there are clicks with no imps my assumptions about
 # this data are wrong

CATGORIES BASED ON IMPRESSIONS and PLOTS :
nyt1$hasimps <-cut(nyt1$Impressions,c(-Inf,0,Inf));
summaryBy(Clicks~hasimps, data =nyt1, FUN=siterange)
ggplot(subset(nyt1, Impressions>0), aes(x=Clicks/Impressions,colour=age_group)) + geom_density();
ggplot(subset(nyt1, Clicks>0), aes(x=Clicks/Impressions,colour=age_group)) + geom_density();
ggplot(subset(nyt1, Clicks>0), aes(x=age_group, y=Clicks,fill=age_group)) + geom_boxplot();

# Without subset- Categorize Impressions for the Entire Dataset.
ggplot(nyt1, aes(x=age_group, y=Impressions, fill=age_group)) + geom_boxplot();
ggplot(subset(nyt1, Clicks>0 & Impressions > 0), aes(x=age_group,y=Clicks, fill = age_group)) +
geom_boxplot();

#Plot the distributions of number impressions and click through-rate (CTR=# clicks/# impressions) for these six age categories.
ggplot(subset(nyt1, Impressions>0), aes(x=Clicks/Impressions,colour=age_group)) + geom_density();
#Plot Clicks Impressions versus Clicks for different Age Groups #
ggplot(nyt1, aes(x=Clicks, y=Impressions, fill=age_group)) + geom_boxplot();

#Categorizes based on the number of Clicks - Question 3#
nyt1$ClicksGroup[nyt1$Clicks == 0] <- "No clicks";
nyt1$ClicksGroup[nyt1$Clicks == 1] <- "One click";
nyt1$ClicksGroup[nyt1$Clicks == 2] <- "Two clicks";
nyt1$ClicksGroup[nyt1$Clicks == 3] <- "Three clicks";
nyt1$ClicksGroup[nyt1$Clicks == 4] <- "Four clicks";
nyt1$Category[nyt1$Gender == 1 & nyt1$Age < 18] <- "<18 Males";
nyt1$Category[nyt1$Gender == 0 & nyt1$Age < 18] <- "<18 Females";

#To estimate the density function #
ggplot(subset(nyt1,Clicks > 1), aes(x=Clicks/Impressions, colour=age_group)) + geom_density();
#Plotting impression counts (histogram)
ggplot(nyt1, aes(x=Impressions, fill=age_group)) + geom_histogram(binwidth=1);
# create categories #
nyt1$scode[nyt1$Impressions==0] <- "NoImps"

```
nyt$scode[nyt1$Impressions >0] <- "Imps"
nyt1$scode[nyt1$Clicks >0] <- "Clicks"
```

REAL DIRECT CASE STUDY :

Real direct is a case study for a data given by a Real Estate Firm in New York.

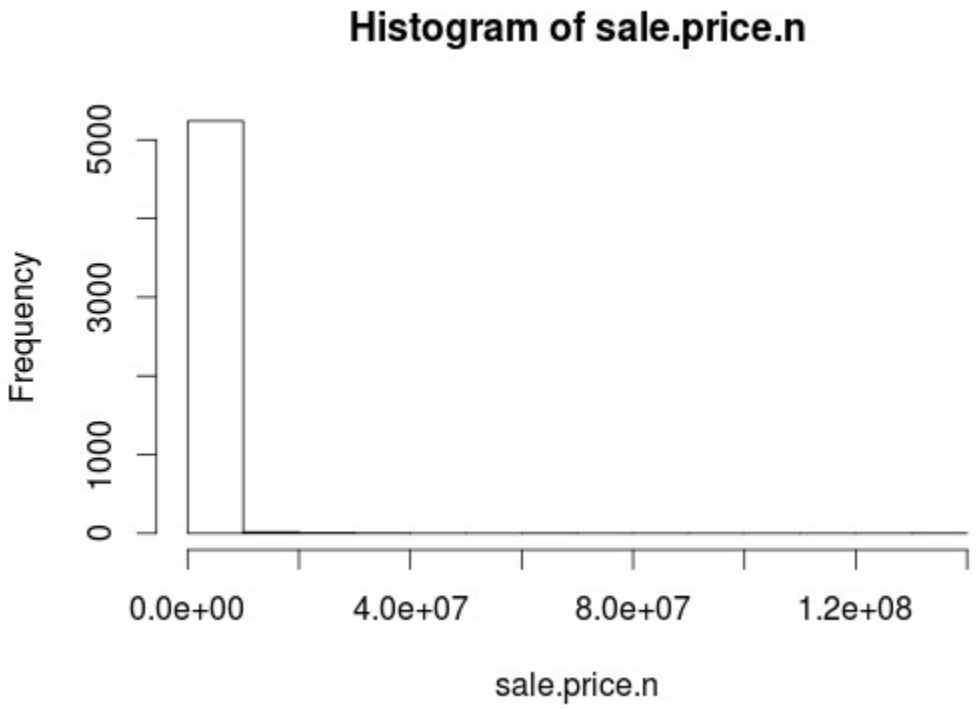
The analysis of the data has been asked for various real estate terms like House Price, Land Price,

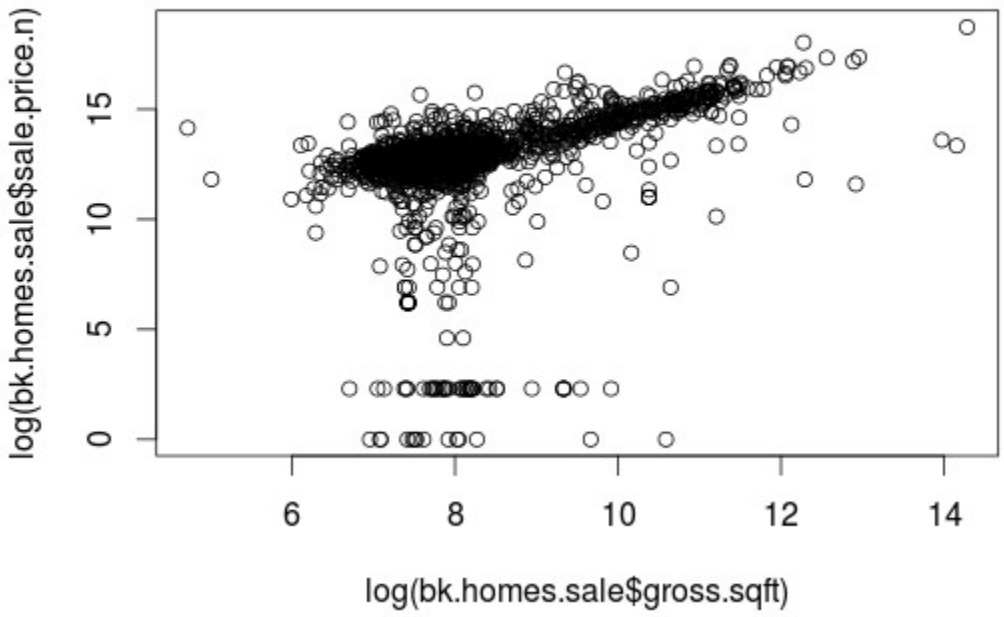
Year of house build etc.

The code which has been given for basic functions has been verified for a different dataset.ie Brooklyn Bronx Dataset

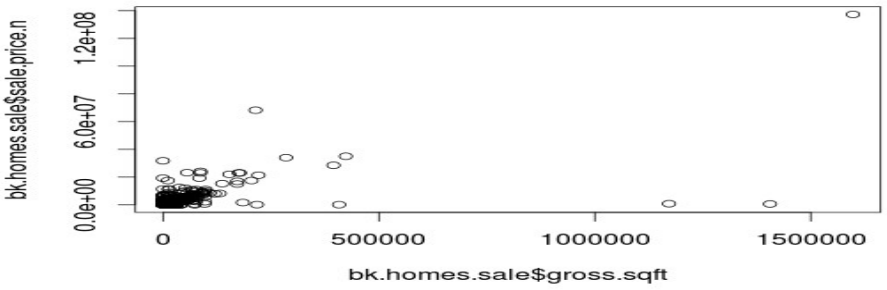
The data is initially cleaned to remove commas and other unwanted data which makes analysis harder. Conversion of Lower Case etc are examples of data cleaning.

The various plots and their analysis are given below.

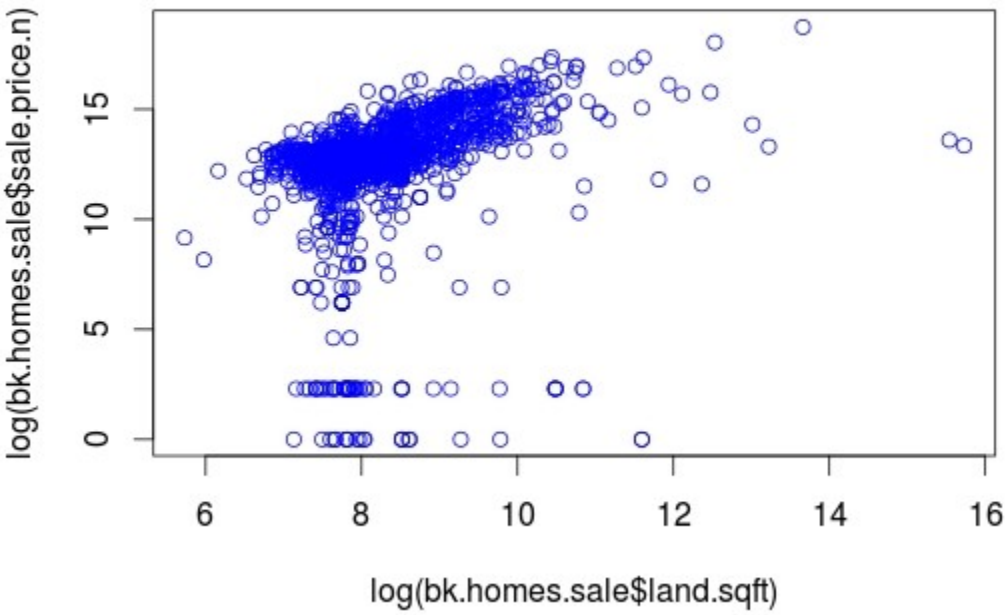




The above pot shows the log plot of Home's gross square feet versus the sales price. This shows that there is more sales at an average square feet. This is probably due to the other factors influencing locality of the house, closeness to schools, colleges and universities, state Tax etc.



4.



The above plot shows the land square feet versus the home sales price. The price of the land and home are different influencing factors for the price of the house.

This shows that more land square feet is not proffered for house purchase. These are probably agricultural lands.

YELP DATASET :

This dataset is part of the third round of Yelp Challenge announced recently. Yelp.com is a company based in San Francisco, CA and it gives makes publicly available ratings and reviews of the user on a particular place for example a restaurant, a spa, a bar etc.

With the advent of social media, there is increasing need for popularity of business establishments across the world. This is typically through collection of data via an online survey or a feedback form. The user is allowed to post comments and give ratings (stars) to data- which will enable other people who want to visit the place get an idea of how the service is. This is very similar to the IMBD rating we see of Hollywood movies and TV Shows!

Yelp regularly releases its datasets for research and analysis purposes and this is one such version from the city of Phoenix, AZ

DATASET DESCRIPTION :

The link to the dataset is http://www.yelp.com/dataset_challenge/

This dataset consists of a collection of data about the city Phoenix, AZ United States. This gives a list of the following :

15,585 businesses

- **111,561** business attributes
- **11,434** check-in sets
- **70,817** users
- **151,516** edge social graph
- **113,993** tips
- **335,022** reviews

These datasets are also co-related. It consists of five JSON Object files . Each line in each file is a single individual JSON object. JSON objects are similar to trees. Each object is a collection of a series

of Objects which are composed of various arrtibutes. For example. Time object is composed of three more variables – HH:MM:SEC . Each business may be sub categorized as two or more business types. For example – a restaurant may also be categorized as a bar. Each user has a unique user Id and each business corporation has a unique business id. This enables correlation between the five different JSON files as shown :

1. Business :

```
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

This data is composed of business id. This is unique to a particular business (Although not to a particular category!! - Each business may be sub-categorized into two or more categories). It also gives the full address, City, State, Latitude, Longitude – enabling us to find it via GPS!

Information of no of reviews given to each establishment, number of hours it is open, the user rating are typically useful in judging the place. In addition to this, we also have data called Attributes – which defines whether the place has Wifi, is suitable to children etc.

2 .Review

```
{
   'type': 'review',
   'business_id': (encrypted business id),
   'user_id': (encrypted user id),
   'stars': (star rating, rounded to half-stars),
   'text': (review text),
   'date': (date, formatted like '2012-03-14'),
   'votes': {(vote type): (count)},
}
```

The second sub category is based on the Review. This gives a business_id(correlated with the Business object given above). It gives a user id, text(Comments given about the place, Type of review(Funny, Useful etc...), the date on which the review was given.

The most important data here is the Star rating.- which is the first rating that is everybody's center of focus.

3. USER

```
{
   'type': 'user',
   'user_id': (encrypted user id),
   'name': (first name),
   'review_count': (review count),
   'average_stars': (floating point average, like 4.31),
   'votes': {(vote type): (count)},
   'friends': [(friend user_ids)],
   'elite': [(years_elite)],
   'yelping_since': (date, formatted like '2012-03'),
   'compliments': {
      (compliment_type): (num_compliments_of_this_type),
      ...
```

```
    },
    'fans': (num_fans),
}
```

The USER Object gives the userid, names, count,number of friends, compliments. Compliments again is an attribute which is positive or negative. The important data we get from this is since when the user started giving Yelp reviews. Frequent reviewers may be in the same business line as the establishment or may have a similar interest.

For example a person who is a frequent reviewer is given an **“elite”** may be a professional in the field.

This also tells us how many people have a particular common hobby.

It gives us important information vital about the user's social profile

4. Check -in

```
{
    'type': 'checkin',
    'business_id': (encrypted business id),
    'checkin_info': {
        '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
        '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
        ...
        '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
        ...
        '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)
    }, # if there was no checkin for a hour-day block it will not be in the dict
}
```

This is another important aspect of the dataset. This dataset consists of check-in's performed by people who are at a particular place. It is a common practice now to Check in (tell people you are here, with who etc...) in sites like Facebook and Four Square. This gives the number of checkin's per hour on Monday to Friday and also weekends.

A good inference which can be made from the checkin data is how many people are at the place on Weekdays, weekends etc. This gives the place a knowledge of the amount of business transacted per day and to bolster their popularity. This data file can be correlated with user and business id's JSON files.

5. tip

```
{
   'type': 'tip',
   'text': (tip text),
   'business_id': (encrypted business id),
   'user_id': (encrypted user id),
   'date': (date, formatted like '2012-03-14'),
   'likes': (count),
}
```

The last and final JSON file is the “tip”. This gives the people working in the place a feedback on how to improve their skills in the particular service and also indicates popularity of a particular item on the Menu (in case of a restaurant) etc.. This can be correlated with business and user id's.

PRELIMINARY ANALYSIS : What can you get from this data? - Insights.....

To summarize , this is a behemoth of convoluted data !!!.

The amount of analysis and knowledge which can be extracted from this is tremendous. The variables which are independent of each other , both inside a particular JSON file or across the various files can be estimated. As mentioned above, the main advantage of this data set is the intricate interdependencies gives the plausibility and wide scope for analysis.

This is largely a machine learning and Information retrieval problem, which is also the original intention of the competition. However, in order to extract useful information, we need to know what information there is to be extracted . This is why a preliminary Exploratory data analysis has been done for this in order to carry forward the problem with much more deeper insights of the variables and the correlations involved.

Another important aspect of this problem which increases the computational complexity of the problem is the fact that it is in form of a JSON object. This is normally different from the usual CSV format that we essentially use for data analysis. JSON allows us to have a tree like data structure which makes processing complex, but analysis and visualization very easy.

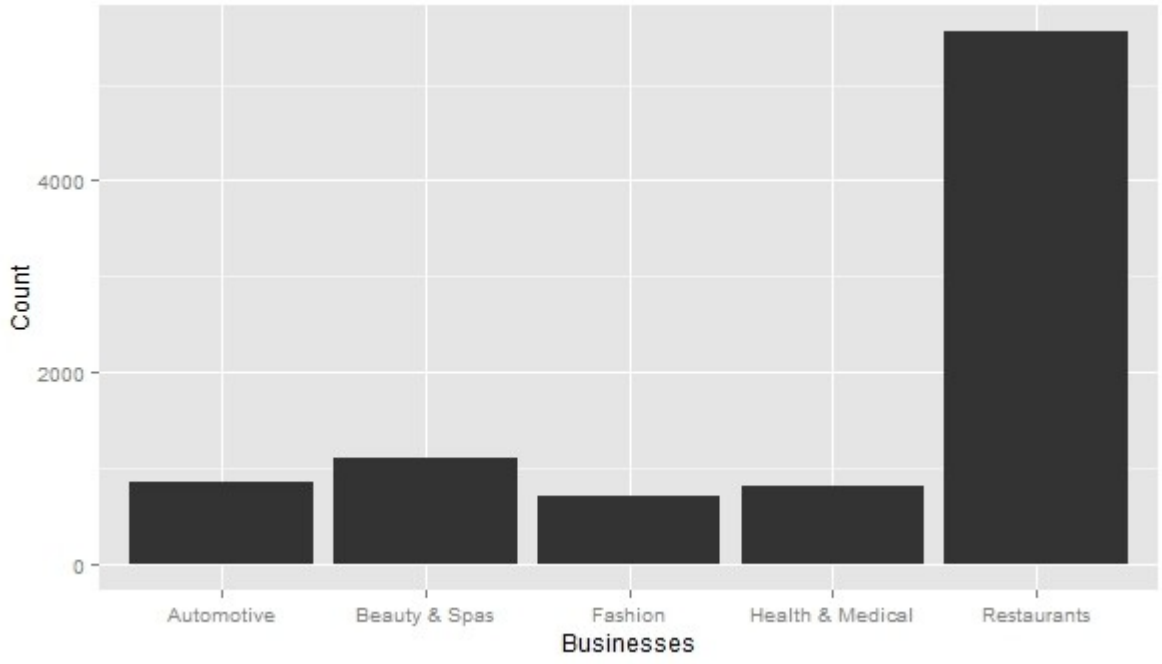
R has a package (rjson) which can be loaded into R Studio which has functions fromJSON which can be used for easy parsing. This delays execution time in the project as there are many for loop constructs.

Hence the purpose of this is to extract information enough to formulate this as a learning problem which gives an inference on favorite places, ratings , stars when queried with a particular information.

ANALYSIS OF DATA :

Since the dataset is humongous, we decided to carry out analysis of a particular subset via manual sampling. Instead of writing a sampling algorithm, a graph of popularity of select business types has been selected manually and certain attributes like Price Range, Business type versus Rating etc are compared. This indicates the most popular business, and we go in depth to analyze it's parameters in order to generate a constructive visualization. This enables the user to know the distribution of the data across the most popular business type. It is plotted in a map of Phoenix, AZ using the Lat Long co-ordinates mentioned in the Business JSON object file.

This graph indicates the number of sample of the number of well known Business Cateogories in Phoenix, AZ. The important business establishments are namely, Automotive Industry, Beauty and Spas , Fashion, Health and Medical and also Restaurants.



Fig(1) : Count of Businesses in Phoenix, AZ

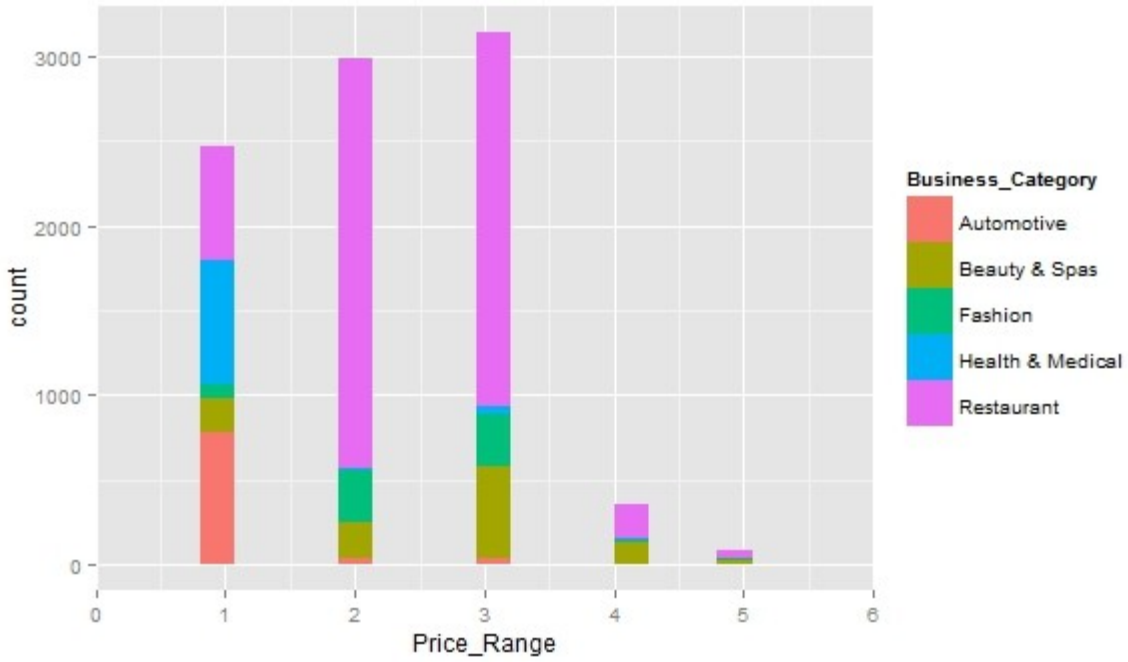
The number of Automotive industries is the lowest in Phoenix AZ as per the graph. Next in rank is the Beauty and Spas followed by Fashion, Health and Medical and leading the other four is the Restaurant industry. According to this website, AZ ranks at number 20 in the automotive manufacturing and production sector and hence this observation is understandable.

http://www.cnn.com/2008/US/12/12/map.us.auto/index.html?eref=rss_latest

As for Beauty and Spas, Fashion – the two industries are correlated , even though to a very slight extent. So the difference between the two is minimum.

The health and Medical Industries are also lower compared to the Restaurant business, which is the most popular in the city of Phoenix, AZ. The number of different cuisines and the ethnicity of the people are a reason for the thriving of the restaurant industry in Phoenix, AZ

The graph of number of categories of business at a particular price range is plotted. This graph is a good measure how expensive each business type is to the customer and also how much revenue each business generates in the city of Phoenix,AZ

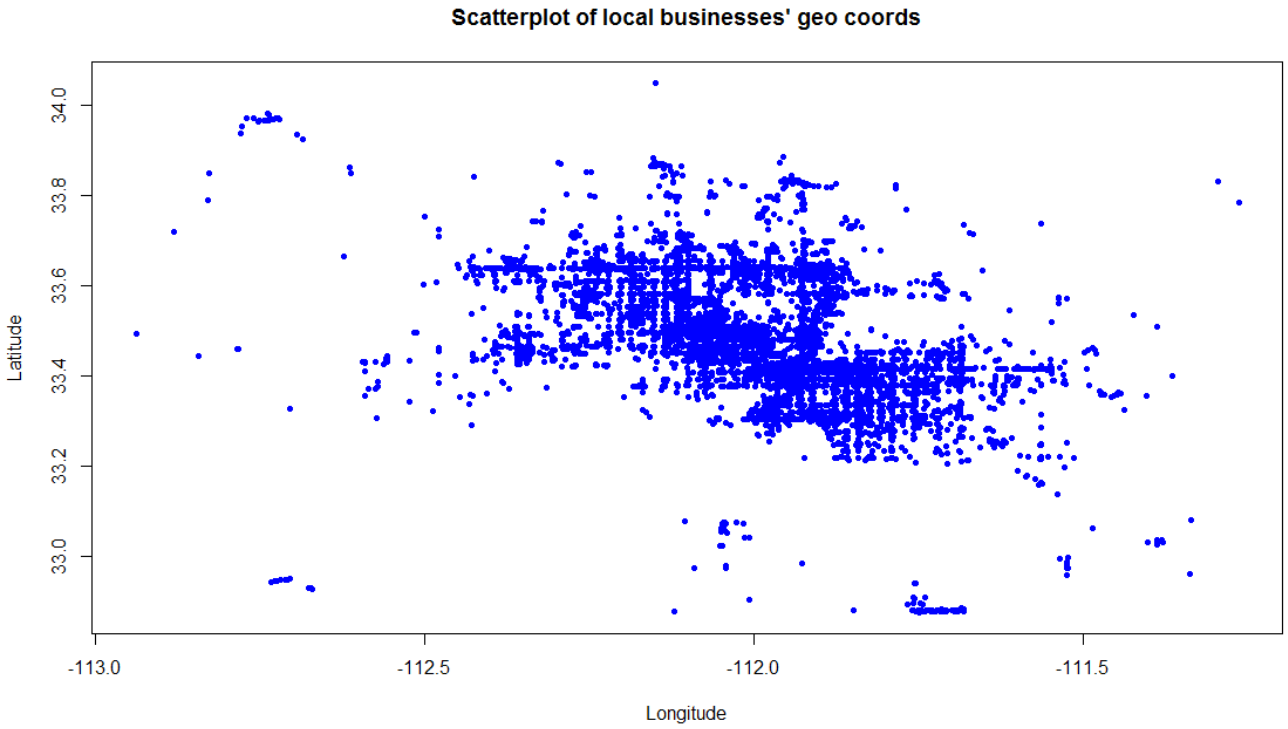


INFERENCE :

We can see that the lower prices ranges correspond to the Automotive industry per year as compared to the rest.

This can be attributed to lot of people in the city being able to afford a car due to ease of access and easy availability of car loans (plausibly)

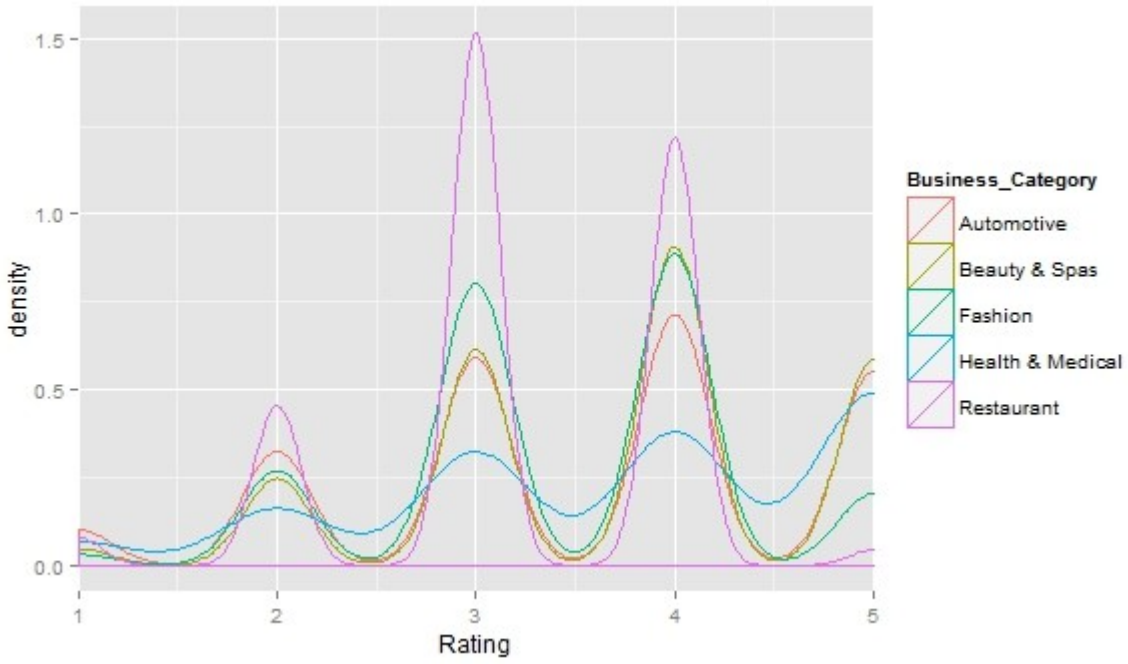
We see that both in number and in price range, the restaurant business is the highest in Phoenix as explained from the previous plot.



The above figure shows the scatter plot of busineess in Phoenix AZ according to Latitude and Longitude. This indicates the Central part of Phoenix AZ has more concentration of busineeses compared to the rest. The concentration is low at the border of the city and at the top and the bottom. Downtown Phoenix AZ has the second concentration of business units after the Central Phoenix. This could indicate :(possibly)

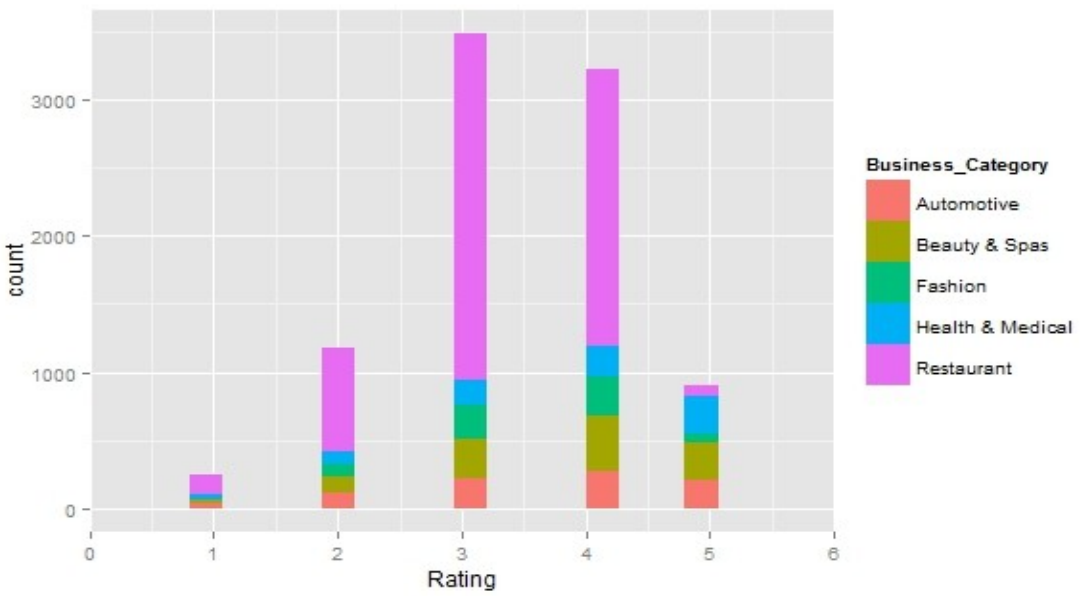
1. Presence of apartment holdings at higher rates due to presence of industry.

2. More restaurants and cuisines next to offices and industrial buildings.



The above figure indicates the user rating of various businesses. It is shown that restaurants have a rating of three and four (many of them) as compared to other holdings.

The factor is rating is constant at the Automotive industry and beauty and Spas as well as Fashion have ratings 2,3 and 4 respectively.

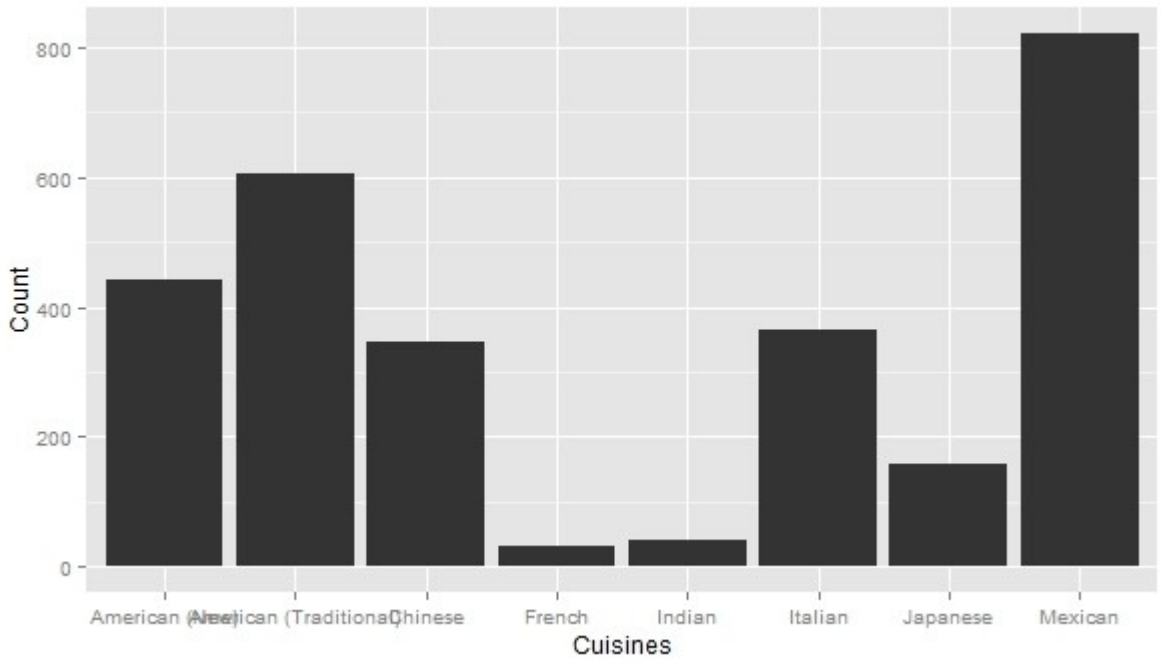


The above figure shows the same trend as shown in the above graph – Restaurants leading in the maximum amount of business holdings based on Rating.

Hence due to the popularity and no of business holdings in the Restaurant Business, it is chosen for in depth analysis.

We present here various facts about restaurants that could be useful as a review to the user – such a s type of cuisine, price range, ratings and so on.

These attributes are plotted on both the map of Phoenix, AZ and also on various plots like Bar graph and ggplot.



The above graph shows the analysis of various types of Cuisines in the city. Some inferences are shown below.

The various cuisine counts are shown above. The above graph indicates the following :

1 . The number of native American restaurants is at a count of 400 which us nearly equal to the number of Chinese restaurants. This indicates lot of Chinese population at this point. It also indicates the popularity of Chinese food across Americans and also people of other races. Italian food is also of a significant number.

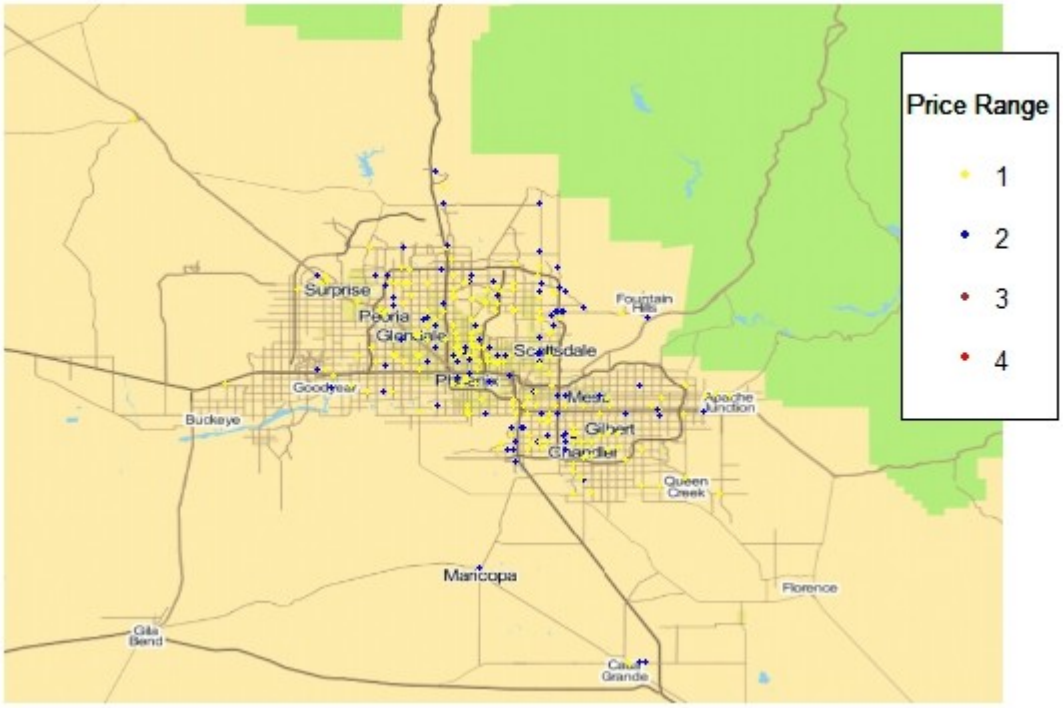
2. The trend in the Mexican restaurants in the highest ever in Phoenz AZ indicating its popularity, affordability etc. The number of Indian restaurants are extremely now and are nearly equal to French Cuisine.

3. This also indicates that Mexiacan restaurants generate largest revenue.

4. This trend can be correlated to the user reviews, Businesses id's across other JSON files to indicate the overall picture.

CUSINE VERSUS PRICE , RATINGS ...

1. The map shown below indicates the price range of the Chinese restaurants distributed across a map.

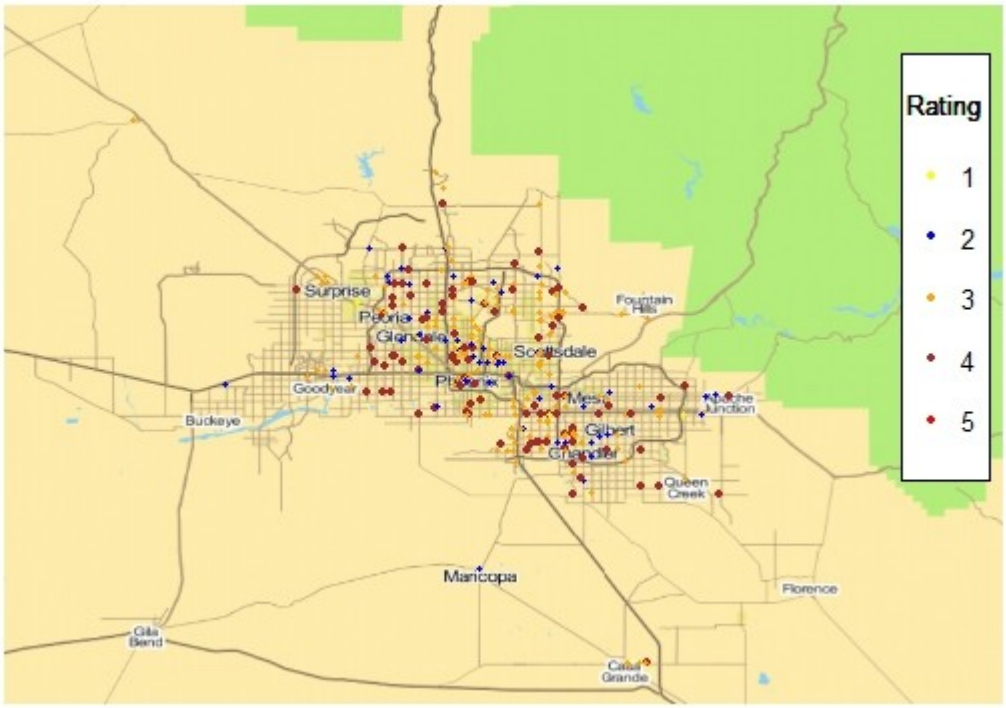


The map shows more concentration of cheaper restaurants all over Phoenix AZ and lesser concentration of average price all over town.

There are very few costly restaurants of price range 4 . There are very few restaurants in the Down South Area of Phoenix, AZ.

Lower concenterations are found in Fountain Hill and Maricopa.

These details are very useful when it comes to planning an event or business cum lunch event at a certain part of town.

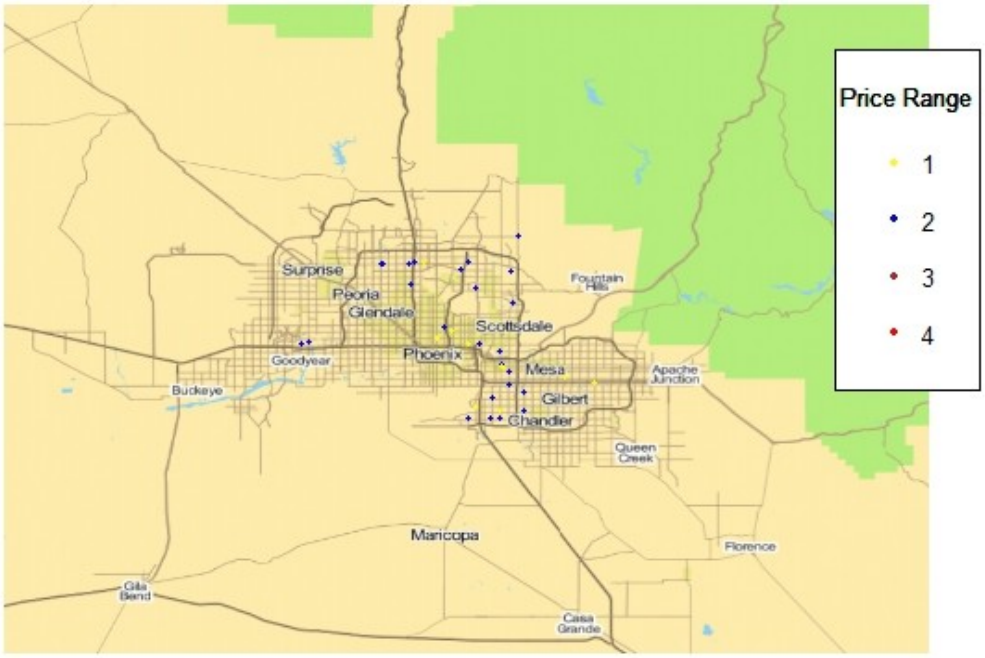


The above figure shows the rating of Chinese Restaurants in Phoenix AZ. It shows that there are good rated Chinese restaurants all across Phoenix except in the East.

There are more concentration of average 3.0 ratings of Chinese restaurants in Scars dale and Peoria.

INDIAN RESTAURANTS :

The maps shown below indicate the statistics of Indian Restaurants – Price Range, Rating and other factors influencing choice.

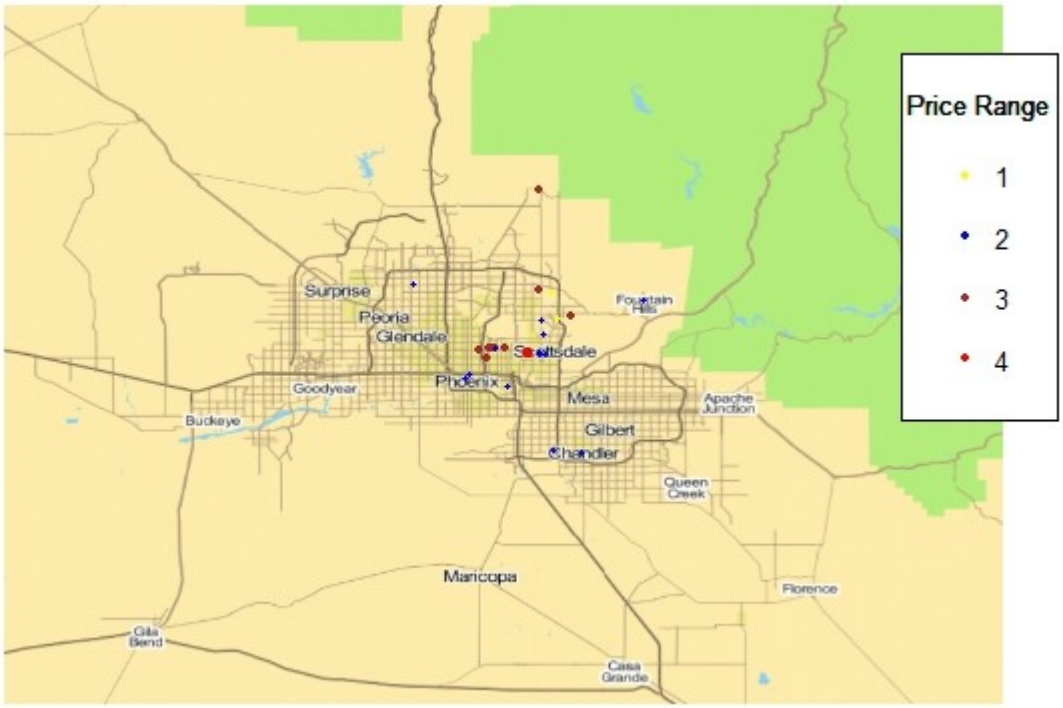


This shows that there are very very few Indian restaurants in Phoenix and most of them are cheap priced restaurants.

More Concentrations of average rating 2 is found in Good Year.

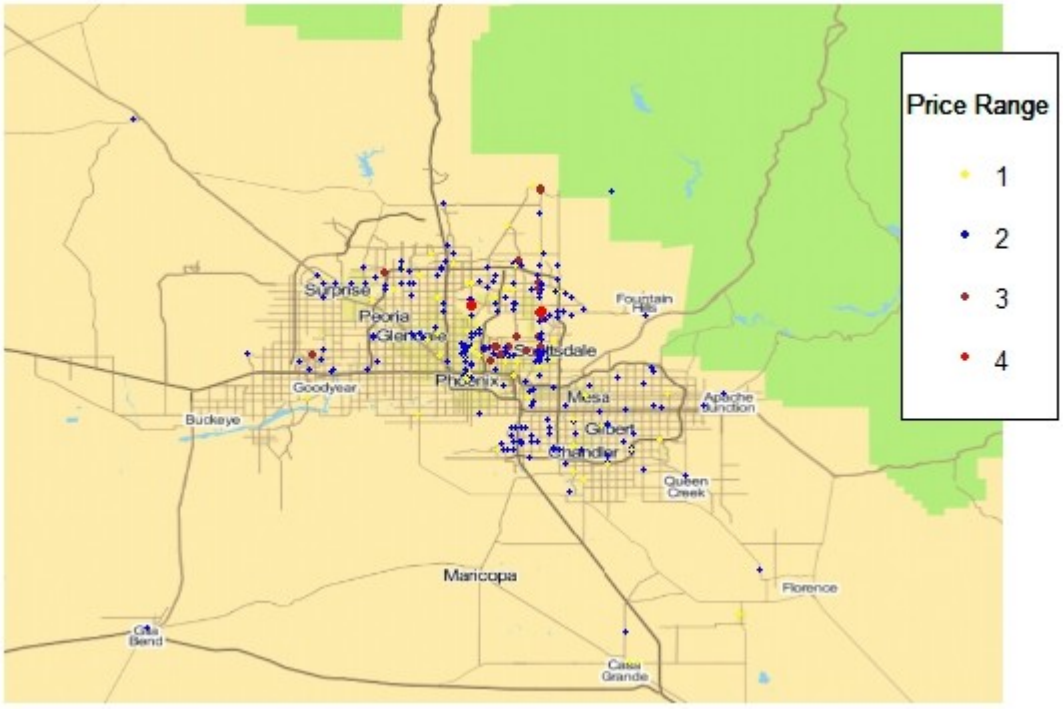
The above map shows the rating of Indian restaurants. It shows more high rated restaurants in Gilbert and near Scarsdale and Good Year counties.

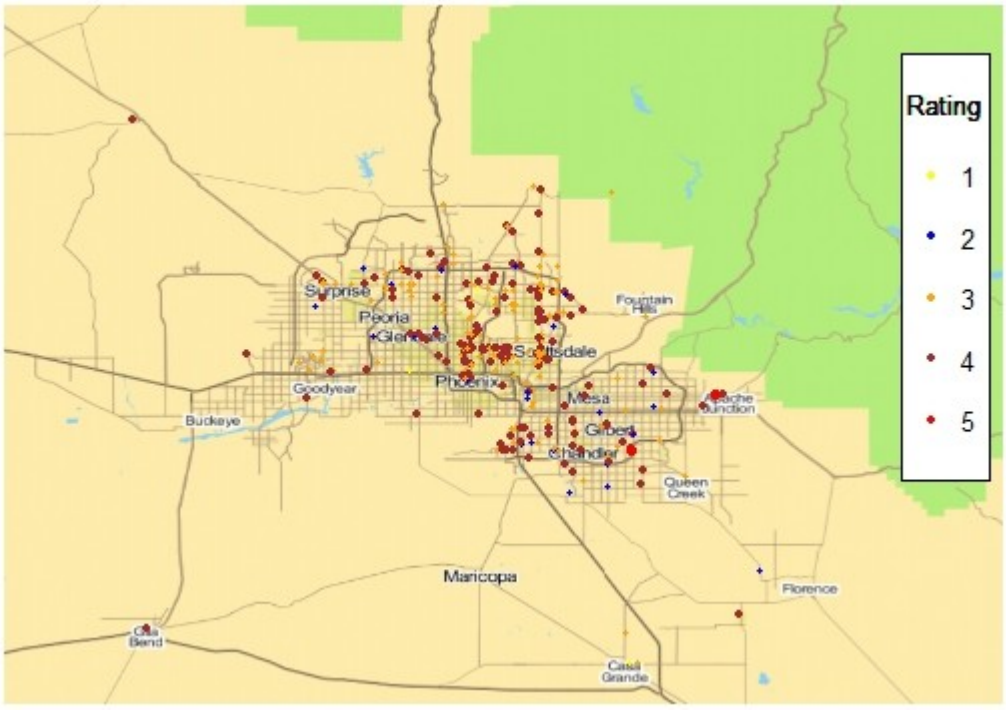
FRENCH RESTAURANTS


Price Range
1
2
3
4
Surprise
Peoria
Glendale
Fountain Hills
Goodyear
Phoenix
Mesa
Apache Junction
Buckeye
Gilbert
Chandler
Queen Creek
Maricopa
Florence
Gila Bend
Casa Grande

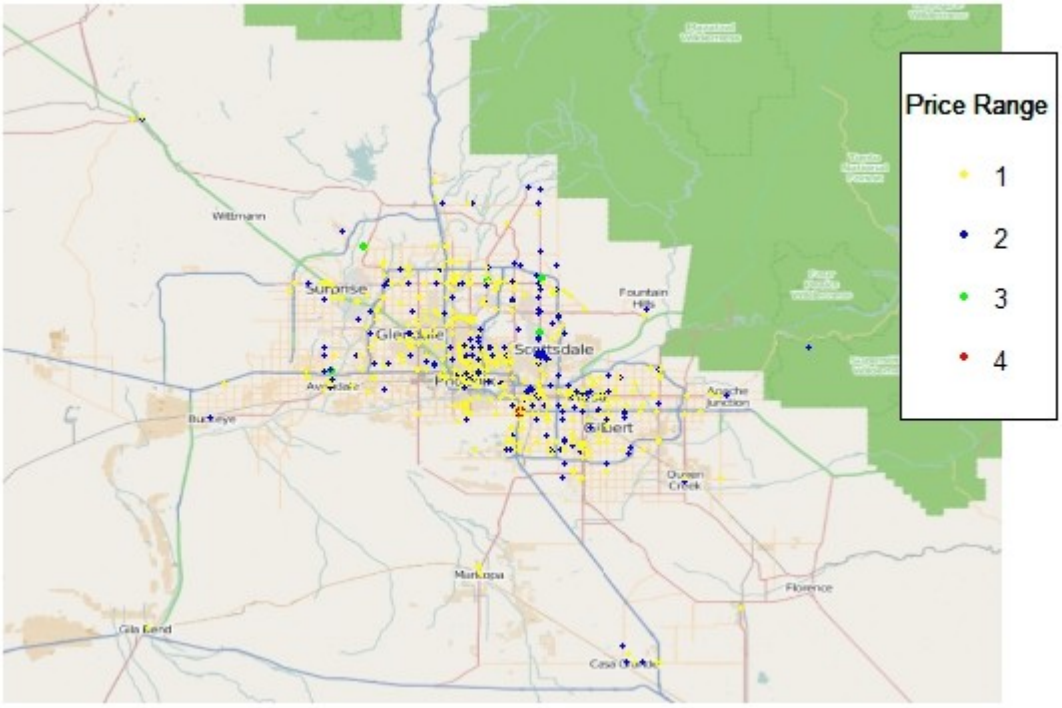

Rating
1
2
3
4
5
Surprise
Peoria
Glendale
Fountain Hills
Goodyear
Phoenix
Mesa
Apache Junction
Buckeye
Gilbert
Chandler
Queen Creek
Maricopa
Florence
Gila Bend
Casa Grande

ITALIAN RESTAURANTS :

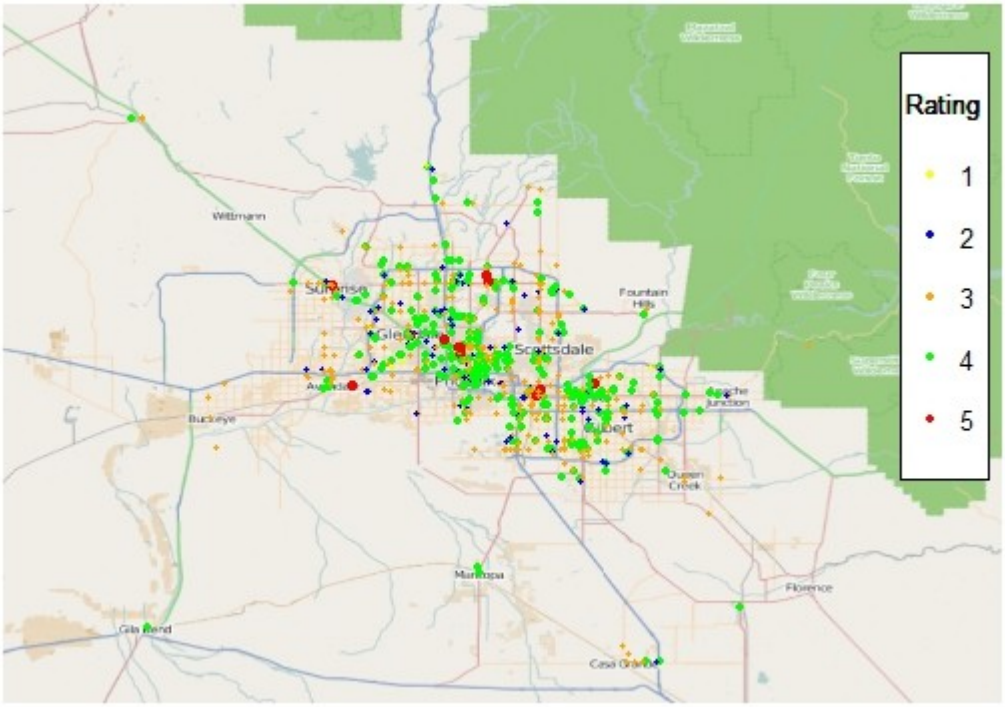



Rating
1
2
3
4
5
Surprise
Peoria
Glendale
Scottsdale
Fountain Hills
Phoenix
Goodyear
Buckeye
Mesa
Apache Junction
Gilbert
Chandler
Queen Creek
Maricopa
Florence
Casa Grande
Gila Bend

The figure shown below shows the rating and price of Mexican restaurants. High Priced Classy restaurants are found in the down south.



The figure below shows the rating of Mexican restaurants.

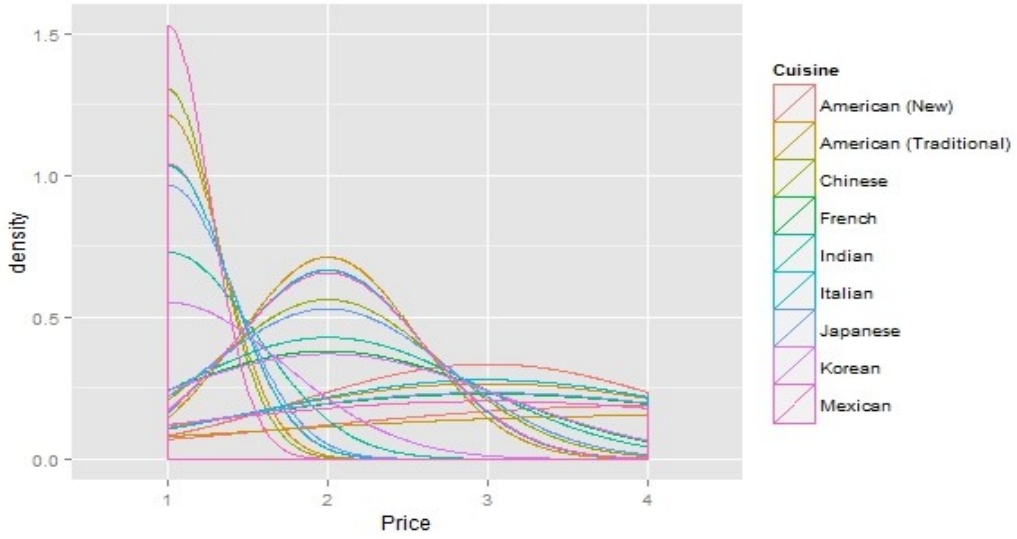


This shows that Most mexican restaurants are HIGHLY rated which indicates their popularity. They are also spread all over town so that there is ease of access.

They are also at strategically good locations across the city. There is a small amount of low rated restaurants.

PRICE VERSUS CUISINE :

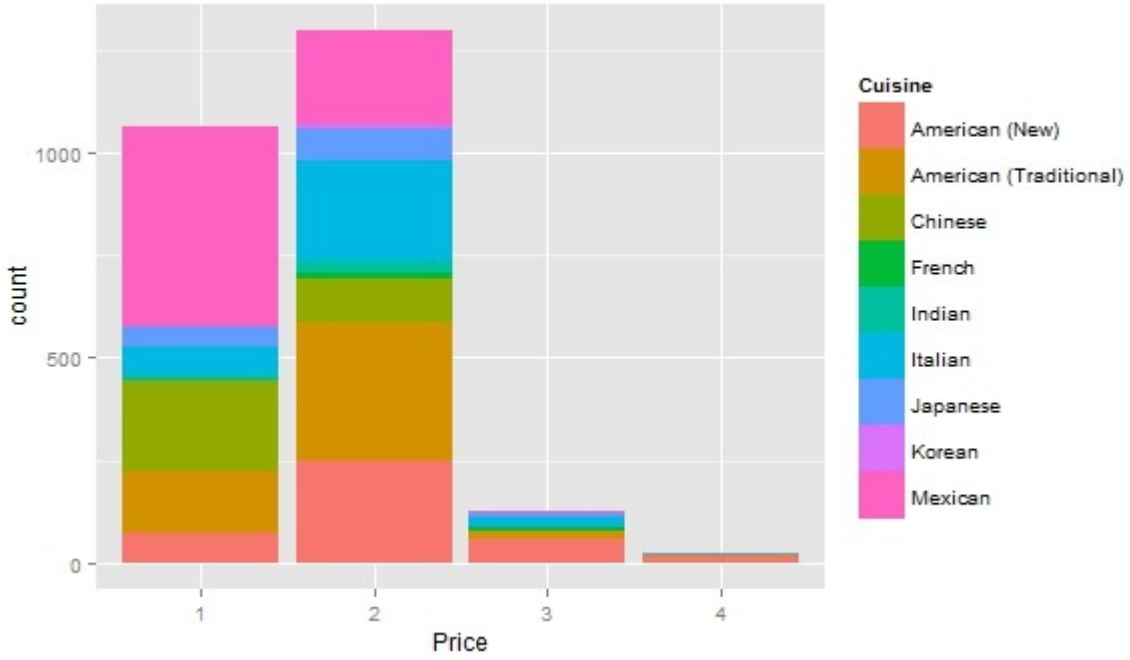
Another important analysis that we must know is the Price versus Cuisine. This is shown in the density plot shown below :



1 . The
above plots show that there is more density of American restaurants at a lower price

2. It also shows few Indian restaurants which are of average price.

3. The traditional American restaurants are of average price rating of about 2.5

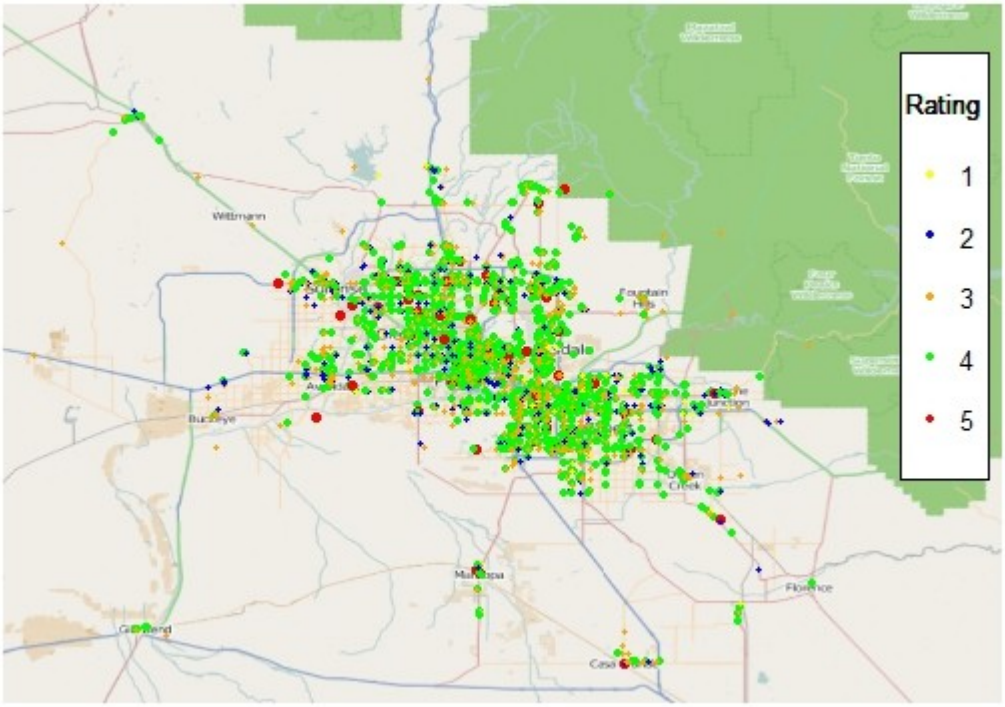


The above figure shows that lowest price and maximum number is among Mexican Restaurants. The lowest amount and the highest price is the American new Cuisine. The traditional cuisine is at an average price of 2.0

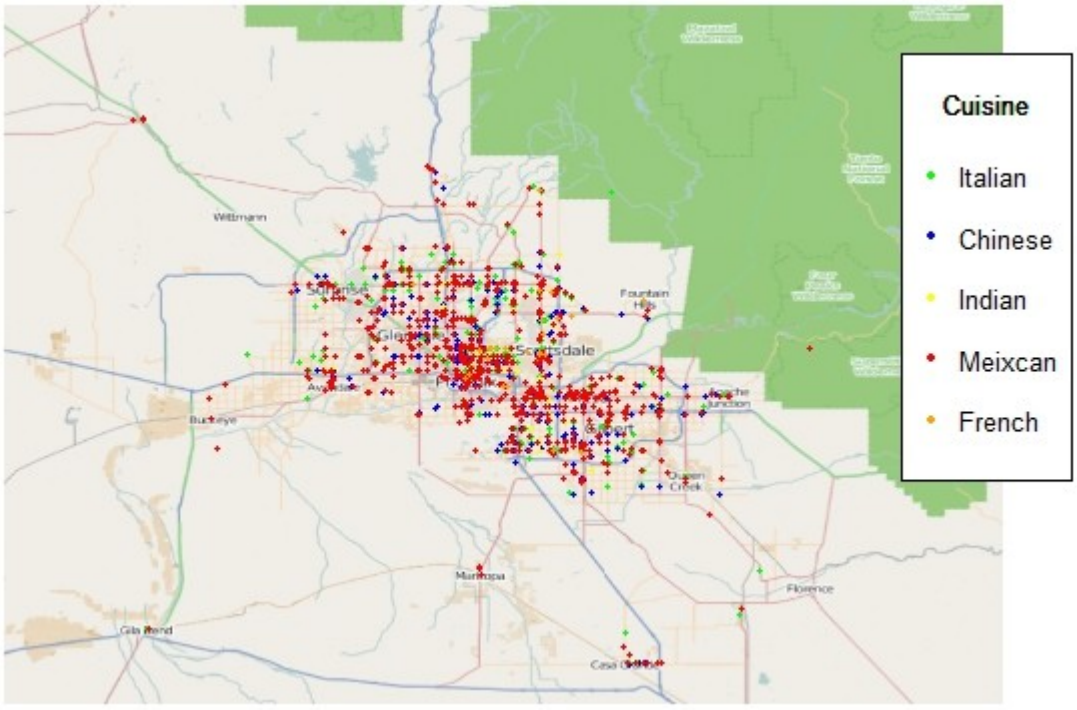
Mexican restaurants also have a high rating and are moderately priced making it the best cuisine in Phoenix AZ.

RESTAURANTS – GENERAL RATING :

The figure shown below indicates that most Restaurants in Central Phoenix have an average rating of 3.0. The highest rated restaurants are Downtown.

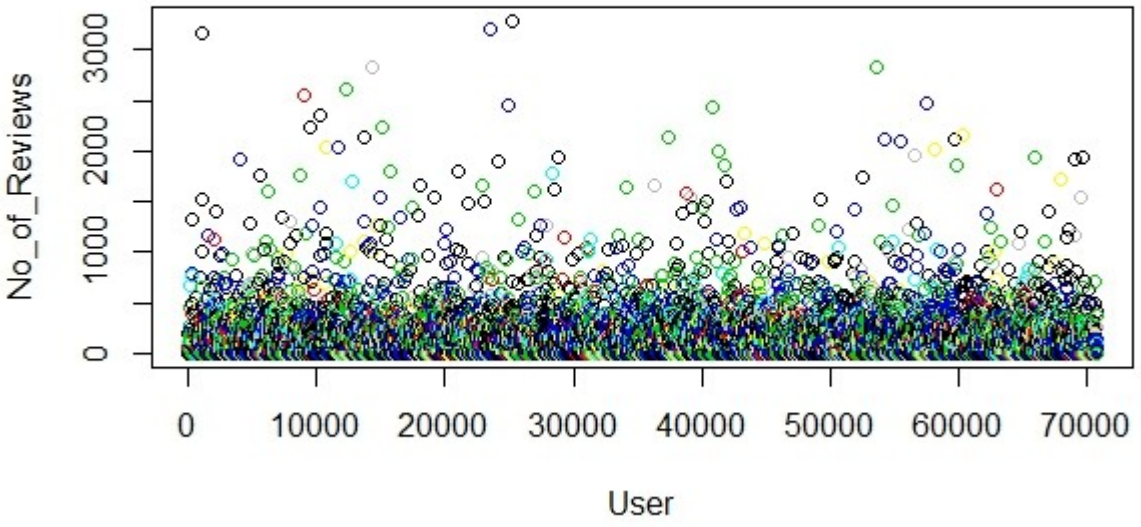


The map below shows the distribution of various cuisines at Phoenix AZ. It shows more concentrations of Mexican restaurants:



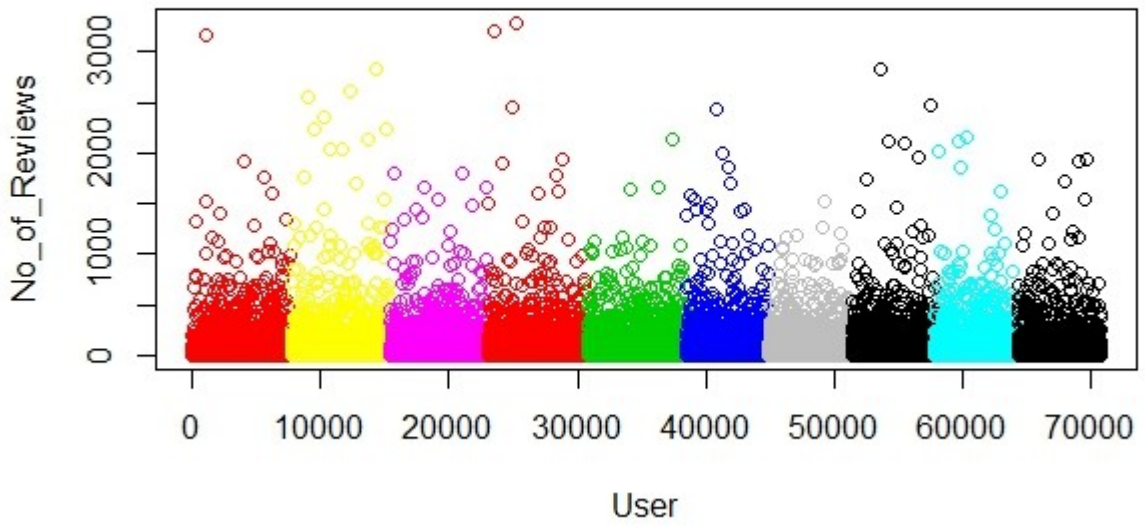
Other Datasets Analysis – Review and User :

KMEANS CLUSTERING

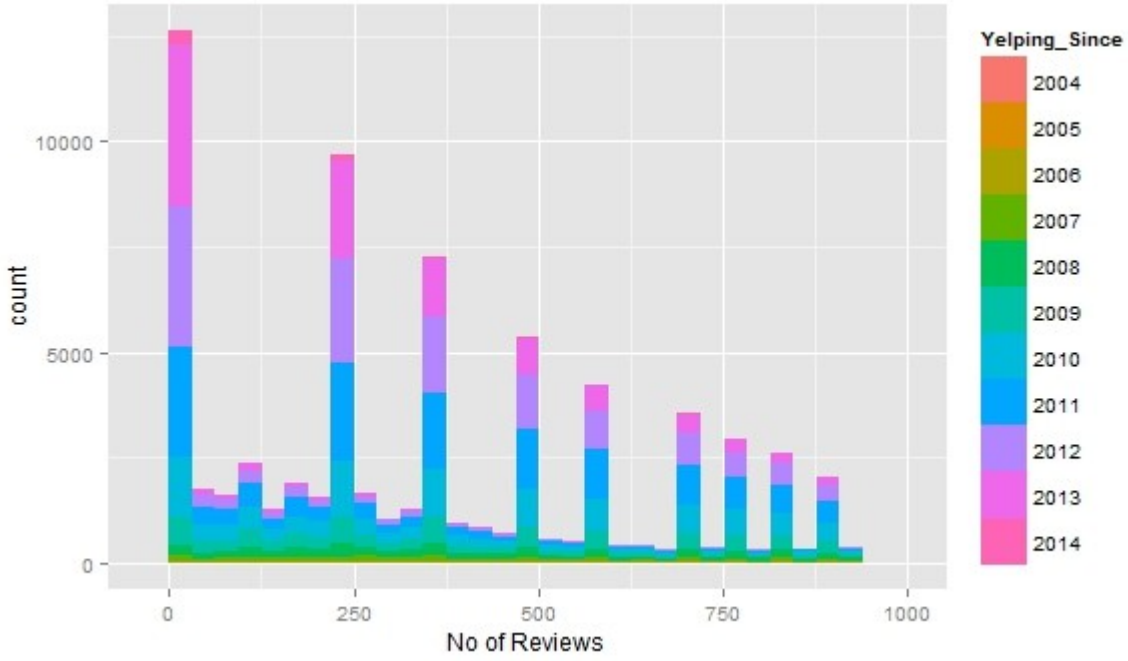


This shows the K Means Clustering plot of the number of reviews versus the Number of Users.

This assumes ten clusters. The plot shows the haphazard distribution of the clusters across all users.



The above plot shows the categorized clusters in a sorted manner. This is the plot of No of Reviews and the number of users, same as the above case.



This above plot shows no of reviews versus count and time by which they were yelping.