Yelp Data Analysis Report

----by Qi Yao qy508

Data Resource:

This dataset is a subset of Yelp's businesses, and I used business, reviews and user data. It was originally put together for the Yelp Dataset Challenge[1]. In the dataset you'll find information about businesses across 11 metropolitan(Las Vegas, Phoenix, Toronto etc) areas in four countries.

Data Preparation:

Table Business(Stored in MySql):

```
CREATE table business(
    business_id varchar(100) not null,
    business_name varchar(100) not null,
    neighborhood varchar(40),
    address varchar(500),
    city varchar(100),
    state varchar(10),
    postal_code varchar(20),
    latitude long,
    longitude long,
    stars double,
    review_count int,
    is_open int,
    categories varchar(500)
);
    LOAD DATA local INFILE
    '/home/2018/spring/nyu/6513/qy508/HW1/data/yelp_business.csv'
    INTO TABLE business
    FIELDS TERMINATED BY ',' #separator
    ENCLOSED BY '"' #Ending
    LINES TERMINATED BY '\n' #Change line
    IGNORE 1 ROWS;
```

Table User(Stored in MySql):

```
CREATE table user(
   user_id varchar(100) not null,
    name varchar(50) not null,
    review_count int,
    yelping_since    datetime,
    friends  text(50000),
    useful   int,
    funny    int,
    cool     int,
    fans     int,
    elite    varchar(100),
    average_stars    double,
    compliment_hot int,
    compliment_more    int,
    compliment_profile  int,
        compliment_cute      int,
        compliment_list int,
        compliment_note      int,
        compliment_plain     int,
        compliment_cool      int,
        compliment_funny     int,
        compliment_writer    int,
        compliment_photos int
);
        LOAD DATA local INFILE
        '/home/2018/spring/nyu/6513/qy508/HW1/data/yelp_user.csv'
       INTO TABLE user
       FIELDS TERMINATED BY ','
       ENCLOSED BY '"'
       LINES TERMINATED BY '\r'
       IGNORE 1 ROWS;
```

Table Review(Read in Rstudio):(Put yelp_review.csv in the working directory)

```
>review<-read_csv("yelp_review.csv")
```

Library Used:

Install these libraries before load them[2][3][4][5]:

```
install.packages('ggplot2')
Install.packages('gridExtra')
Install.packages('leaflet')
Install.packages("readr")
Install.packages("wordcloud")
Install.packages("tidytext")
Install.packages("devtools")
```

Load all packages needed:

```
library(RMySQL) #connect R with Mysql
library(ggplot2) #important drawing tool
library(gridExtra) #use extra function for grid
library(magrittr)
library(wordcloud)
library(stringr)# string manipulation
library(leaflet)#to use map to make use of geospatial data
library(data.table)
library(dplyr)# %>% operator as well as
library(readr) # to use read_csv() which can read review.csv much more quickly
library(tidytext) # NLP manipulation
library(devtools) #to use install_github
install_github("ricardo-bion/ggradar")
library("ggradar") #to draw the radar graph
```

Connect to Mysql database:

```
channel<-dbConnect(MySQL(),user="qy508",password="qy508123",dbname="qy508",host="localhost")
```

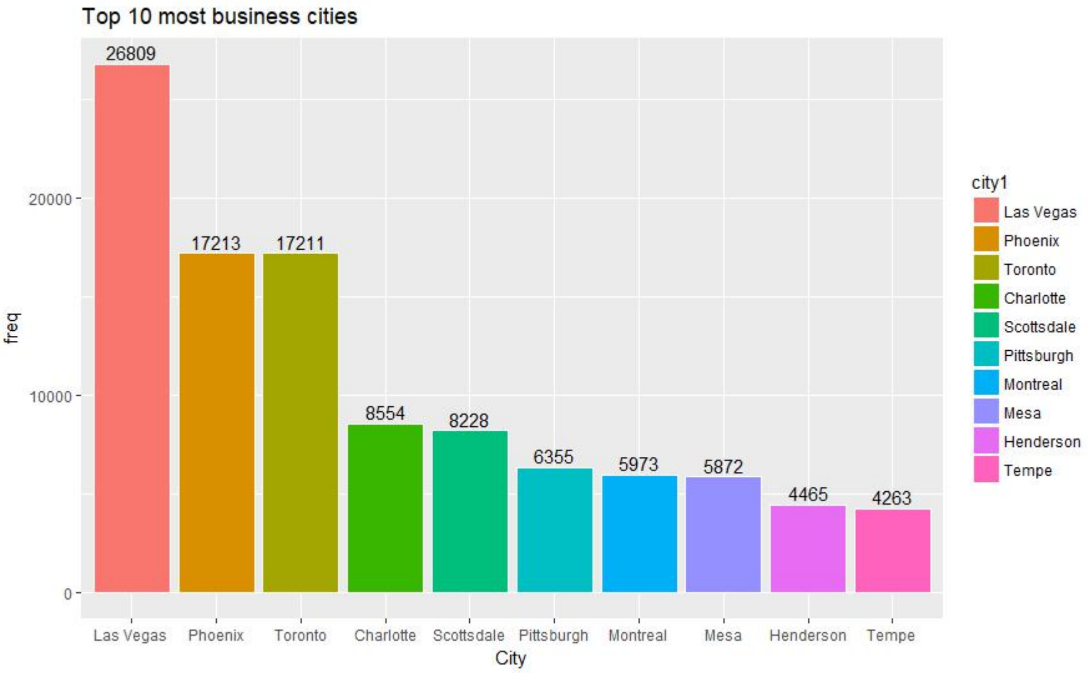
Briefly Introduction:

In this Yelp data analysis, I did a lot of **EDA**(Exploratory Data Analysis), **Geospatial Data Analysis** and tried some **simple NLP analysis**[6] to try to **portrait users** clearly[7].

1. Top 10 Cities with most Business(Amount)[8][9]

Code:

```
res<-dbGetQuery(channel, "select city,count(*)as freq from business group by city Order by freq desc limit 10")
res$city1 <- factor(res$city, levels=unique(res$city))
temp<-ggplot(res,aes(x=city1,y=freq,fill=city1))+geom_bar(stat="identity",col='white')+geom_text(aes(label = freq, vjust = -0.3, hjust = 0.5))+labs(x="City",title="Top 10 most business cities")
show(temp)
```



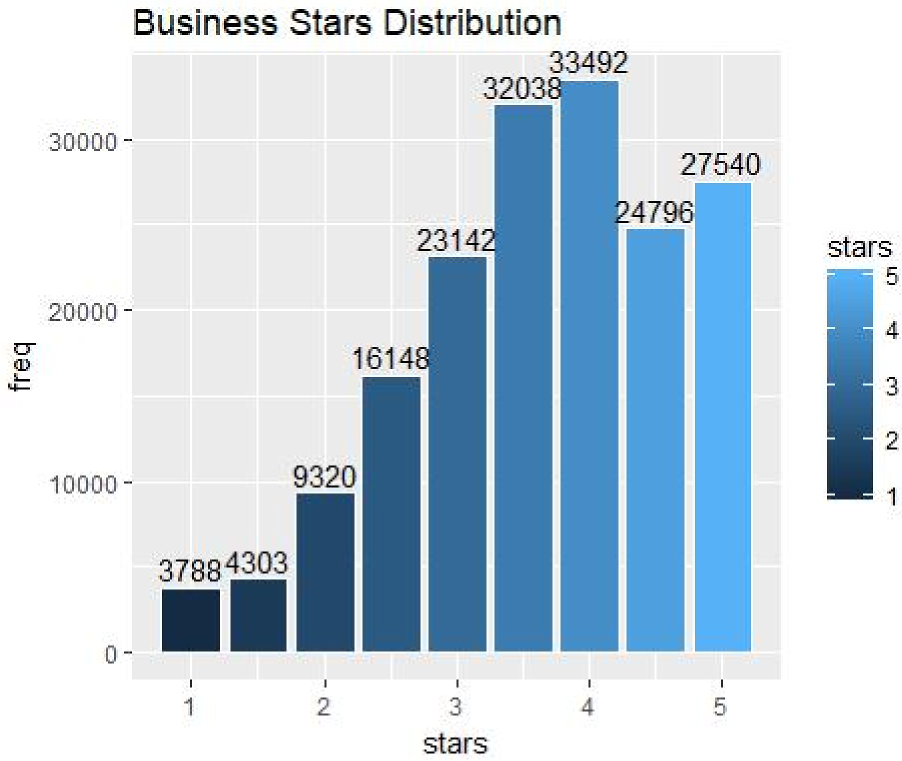
Analysis&Result:

With this graph, we can find the amount of business within these cities. And There are much more business in Las Vegas, Phoenix and Toronto than else cities.

2. The distribution of the business with different stars

Code:

```
# select the stars distribution according to the stars from Mysql

res<-dbGetQuery(channel, "select stars,count(*)as freq from business group by stars")

#draw the histogram with ggplot2

temp<-ggplot(res,aes(x=stars,y=freq,fill=stars))+geom_bar(stat="identity",col='white')+geom_text(aes(label = freq, vjust = -0.3, hjust = 0.5))+labs(title="Business Stars Distribution")
show(temp)
```



Analysis:

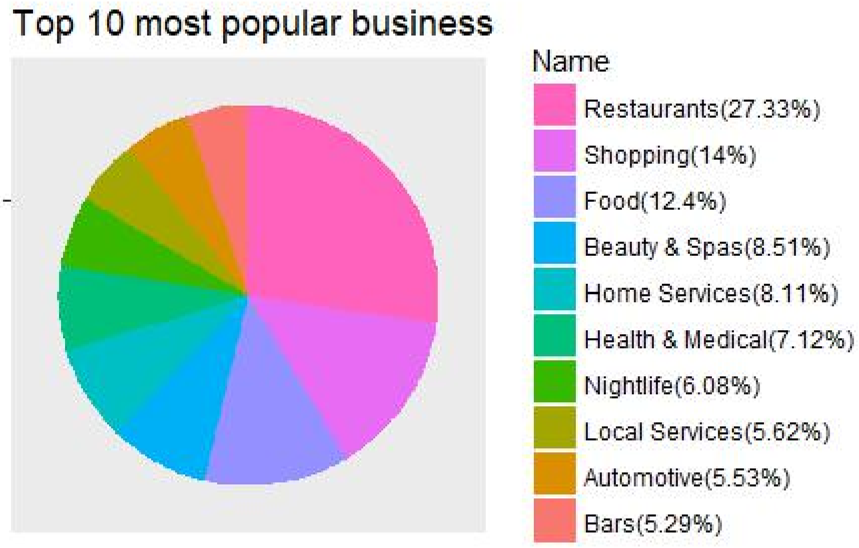
In this part, we can clearly see the distribution of the stars. We can find that nearly 80% business are over 3 stars. While there are still some business which are not quite satisfying. So yelp can improve itself to get higher ratio of high quality business.

.

3. Top 10 Most Popular Business[10]

Code:

```
categories = str_split(str_replace_all(business$categories,"\r",""),";")
categories = as.data.frame(unlist(categories))
colnames(categories) = c("Name")

temp<-(categories %>%
group_by(Name) %>%
summarise(Count = n()) %>%
arrange(desc(Count)) %>%
ungroup() %>%
mutate(Name = reorder(Name,Count)) %>%
head(10))

ggplot(temp,aes(x = "", y = Count, fill = Name)) +
geom_bar(stat = "identity",width=1) +
coord_polar(theta = "y")+
Theme(axis.text.x = element_blank()) +scale_fill_discrete( breaks=temp$Name,labels =
paste( as.vector(temp$Name),"(", round(temp$Count/ sum(temp$Count) * 100, 2), "%)", sep =
""))+
theme(panel.grid=element_blank())+
labs(x="",y="",title="Top 10 most popular business")
```



Analysis:

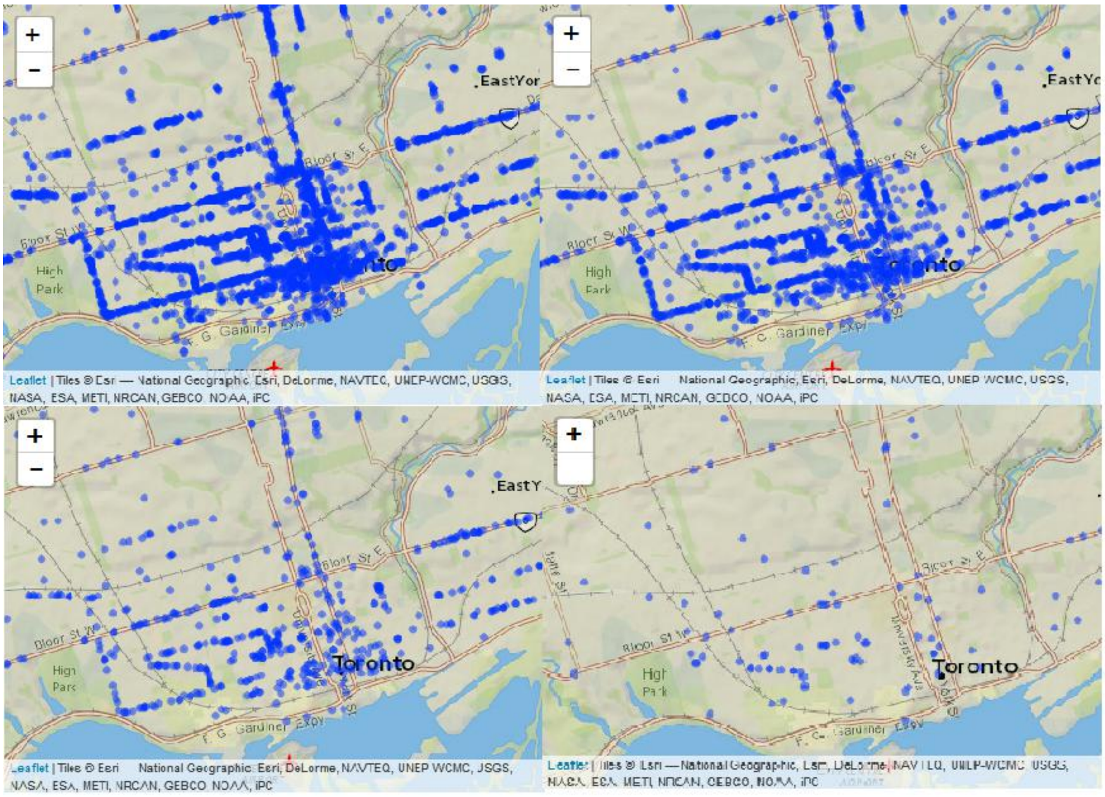
We can see the clearly distribution of business categories in yelp. I try to find out the top 10 business categories from all these business. We can find that yelp actually offers a quite great range of business. And Restaurant is the main goal of yelp.

4. Restaurants Map in Toronto(From 3.5 Stars to 5 Stars)[11]

Toronto is the capital of the Canadian province of Ontario. It is located within the Golden Horseshoe in Southern Ontario on the northern shore of Lake Ontario. With 2,731,571 residents in 2016, it is the largest city in Canada and fourth-largest city in North America by population.

Code:

```
    res1<-dbGetQuery(channel, "select latitude,longitude from business where city='Toronto' and
stars>=3.5 and categories LIKE '%restaurant%';")        #can change the stars here
    res1$latitude<-as.numeric(res1$latitude)
    res1$longitude<-as.numeric(res1$longitude)
    center_lon = median(res1$longitude,na.rm = TRUE)
    center_lat = median(res1$latitude,na.rm = TRUE)

    leaflet(res1) %>% a
      ddProviderTiles("Esri.NatGeoWorldMap") %>%
      addCircles(lng = ~longitude, lat = ~latitude,radius = 1)    %>%

      # controls
      setView(lng=center_lon, lat=center_lat,zoom = 12)
```


EastYork
Bloor St E
Bloor St W
High Park
Toronto
F. G. Gardiner Expy
Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

(Upper left: >=3.5 stars, Upper Right: >=4 stars, Lower Left:>= 4.5 Stars, Lower Right: >=5 stars)

Analysis:

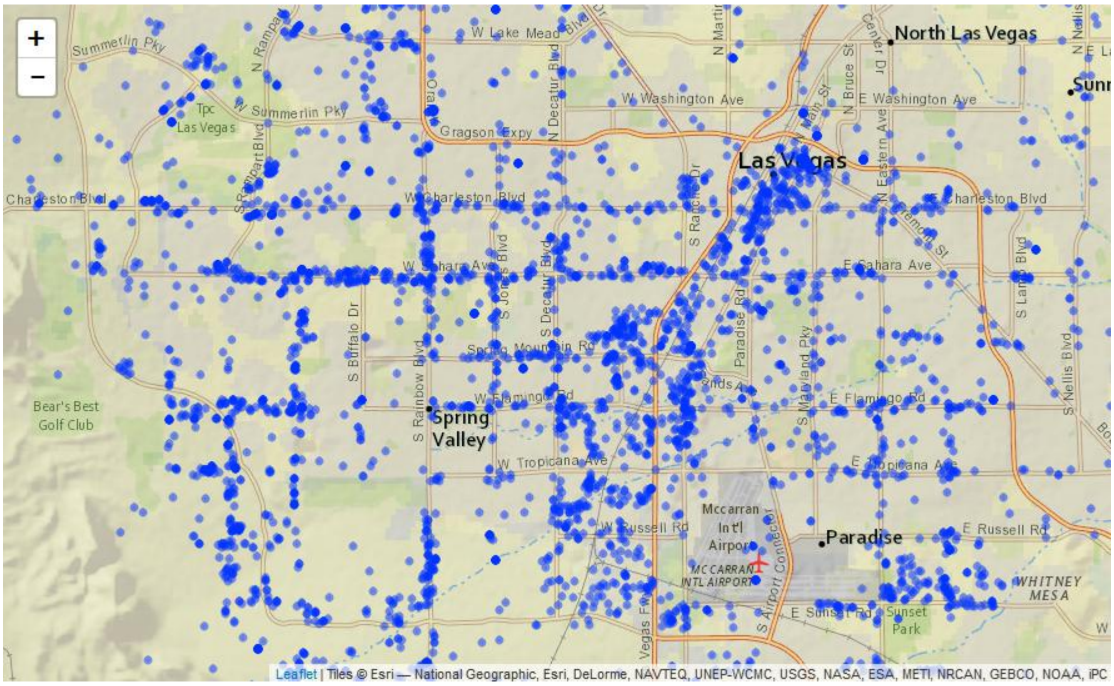
From Wiki(Toronto) we can find that there is a highly concentrate in area of business in Old Toronto, York and East York which is the downtown of the Toronto. In some way it shows the density of population. And high quality business has a more sparse distribution, from downtown to suburban area.

5. Best Business places in Las Vegas(5 stars)

The Las Vegas Strip is a stretch of South Las Vegas Boulevard in Clark County, Nevada that is known for its concentration of resort hotels and casinos. The Strip is approximately 4.2 miles (6.8 km) in length,[1] located immediately south of the Las Vegas city limits in the unincorporated towns of Paradise and Winchester.

Code:

```
res1<-dbGetQuery(channel, "select latitude,longitude from business where city='Las Vegas' and
stars>=5;")
res1$latitude<-as.numeric(res1$latitude)
res1$longitude<-as.numeric(res1$longitude)
center_lon = median(res1$longitude,na.rm = TRUE)
center_lat = median(res1$latitude,na.rm = TRUE)

leaflet(res1) %>%
addProviderTiles("Esri.NatGeoWorldMap") %>%
addCircles(lng = ~longitude, lat = ~latitude,radius = 1)    %>%

# controls
setView(lng=center_lon, lat=center_lat,zoom = 12)
```



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

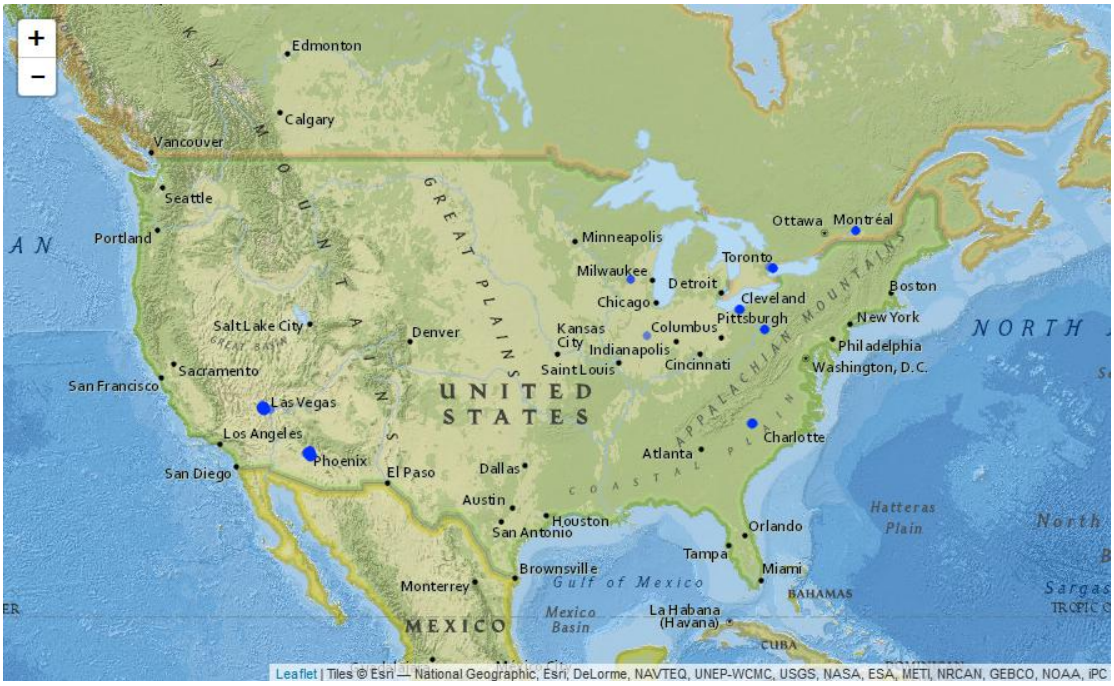
Analysis:

The high quality business is along with main road. And have a sparse distribution of entire city.

6. Business Popularity analysis

Code:

```
res1<-dbGetQuery(channel, "select latitude,longitude from business order by review_count desc limit 500;")
res1$latitude<-as.numeric(res1$latitude)
res1$longitude<-as.numeric(res1$longitude)

eaflet(res1) %>% addProviderTiles("Esri.NatGeoWorldMap") %>%
addCircles(lng = ~longitude, lat = ~latitude,radius = 1)
```



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

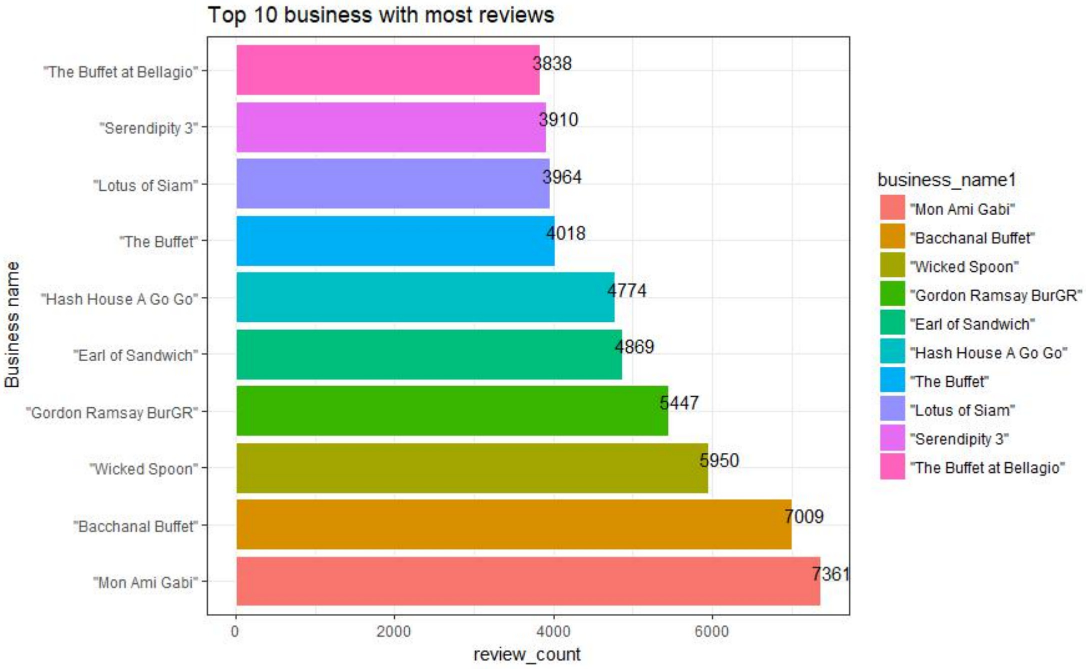
Analysis:

I list the first 500 business with most reviews. And I take a big map of these business distribution within the whole US. It in some way shows the people activity distribution.

7. Top 10 business with most reviews

Code:

```
        # select the number of reviews according to the stars from Mysql
        res<-dbGetQuery(channel, "select business_name,review_count from business order by
review_count desc limit 10;")

        #draw the histogram with ggplot2
        res$business_name1 <- factor(res$business_name, levels=unique(res$business_name))
        temp<-ggplot(res,aes(x=business_name1,y=review_count,fill=business_name1))+
        geom_bar(stat="identity",col='white')+
        geom_text(aes(label = review_count, vjust = -0.1, hjust = 0.2))+
        labs(x="Business name",title="Top 10 business with most
reviews",legend.title=element_blank())+
        coord_flip()+
        theme_bw()
        show(temp)
```



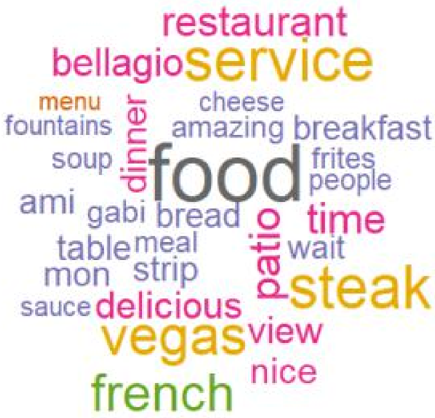
Nine of these ten these businesses are in Las Vegas.

7.1 Some analysis of Mon Ami Gabi[12]:

7.1.1Word Cloud of Reviews:

Code:

```
createWordCloud = function(train)
{
   train %>%
      unnest_tokens(word, text) %>%
      filter(!word %in% stop_words$word) %>%
      count(word,sort = TRUE) %>%
      ungroup()    %>%
      head(30) %>%
    with(wordcloud(word, n, max.words = 30,colors=brewer.pal(8, "Dark2")))
}

createWordCloud(review %>%
                      filter(business_id == "4JNXUYY8wbaaDmk3BPzlWw"))
```

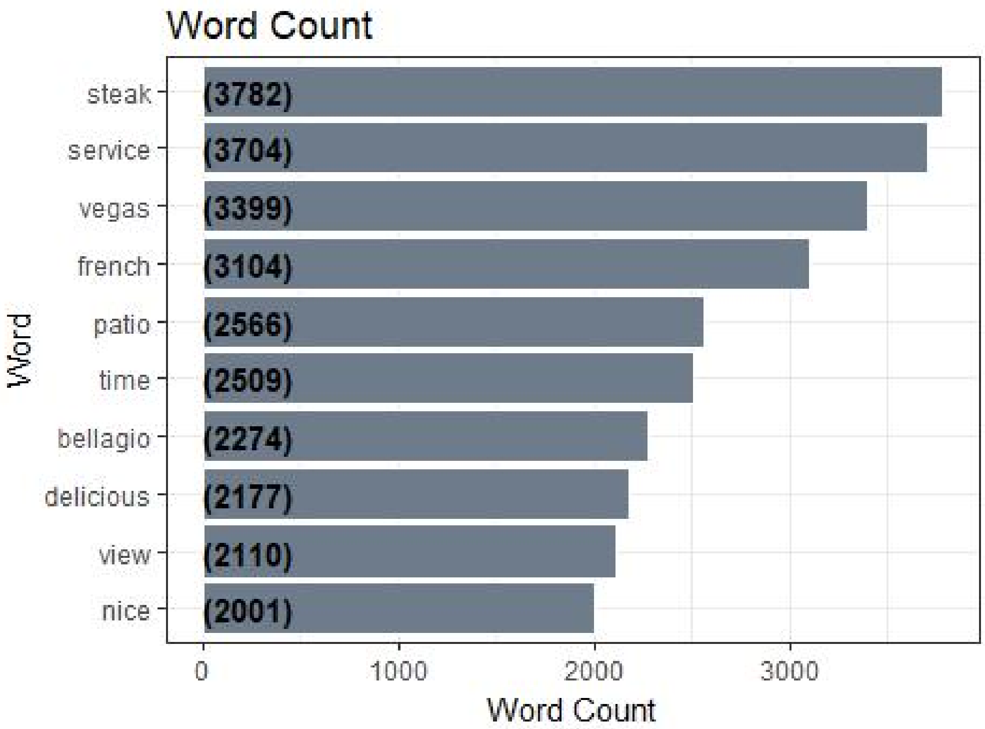


A word cloud is a graphical representation of frequently used words in the text. It can show a clearly and straightforward understanding of the frequency of the words appeared in all the reviews. Meanwhile, the height of each word in this picture is an indication of frequency of occurrence of the word in the entire text. We can find that the words steak, service, vegas, french, patio, bellagio, delicious, dinner. the words which have been used very frequently in the reviews.

7.1.2 Word Count for review of Mon Ami Gabi:

Code:

```
review %>%
filter(business_id == "4JNXUYY8wbaaDmk3BPzlWw") %>%
unnest_tokens(word, text) %>%
filter(!word %in% stop_words$word) %>%
filter(!word %in% c('food','restaurant')) %>%
  count(word,sort = TRUE) %>%
  ungroup() %>%
mutate(word = factor(word, levels = rev(unique(word)))) %>%
  head(10) %>%

  ggplot(aes(x = word,y = n)) +
  geom_bar(stat='identity',colour="white", fill ="lightsteelblue4") +
  geom_text(aes(x = word, y = 1, label = paste0("(",n,")",sep="")),
            hjust=0, vjust=.5, size = 4, colour = 'black',
            fontface = 'bold') +
  labs(x = 'Word', y = 'Word Count',
         title = 'Word Count') +
coord_flip()+
theme_bw()
```



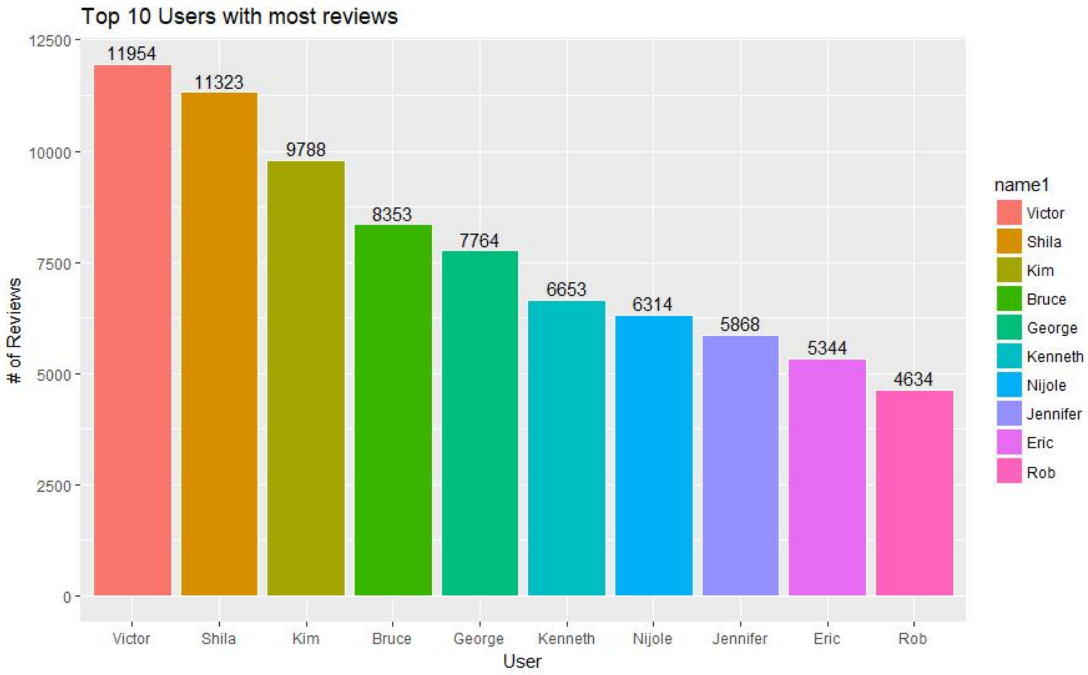
And this graph gives the frequencies of words in quantity. We can have very direct understanding of Mon Ami Gabi. It mainly offers delicious steak, has great service, patio with great view. It generally gets mainly great comment. So all these analysis gives us a great understanding of this restaurant.

8. User Analysis:

Top 10 Users made most reviews

Code:

```
# select the number of res according to the stars from Mysql

res<-dbGetQuery(channel, "select name,review_count from user order by review_count desc limit
10;")

#draw the histogram with ggplot2
res$name1 <- factor(res$name, levels=unique(res$name))
temp<-ggplot(res,aes(x=name1,y=review_count,fill=name1))+
geom_bar(stat="identity",col='white')+
geom_text(aes(label = review_count, vjust = -0.3, hjust = 0.5))+
labs(x="User",y="# of Reviews"title="Top 10 Users with most reviews")
show(temp)
```



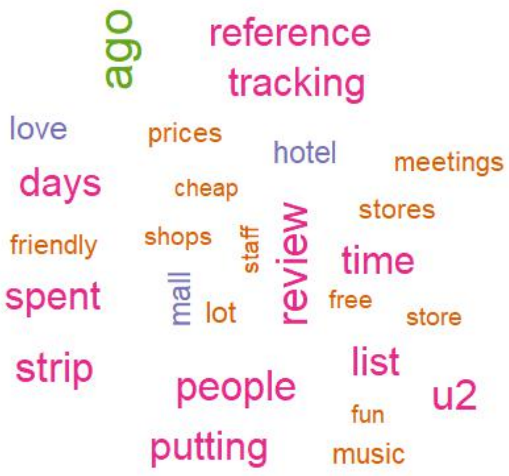
This shows the most ten active ten users with their amount of history reviews .

8.1 Some deep analysis of top users(User portrait):

8.1.1 Wordcloud of Shila's reviews:

Code:

```
createWordCloud = function(train)
{
   train %>%
      unnest_tokens(word, text) %>%
      filter(!word %in% stop_words$word) %>%
      count(word,sort = TRUE) %>%
      ungroup()    %>%
      head(30) %>%

      with(wordcloud(word, n, max.words = 30,colors=brewer.pal(8, "Dark2")))
}

createWordCloud(review %>%
                              filter(user_id =="RtGqdDBvvBCjcu5dUqwfzA"))
```

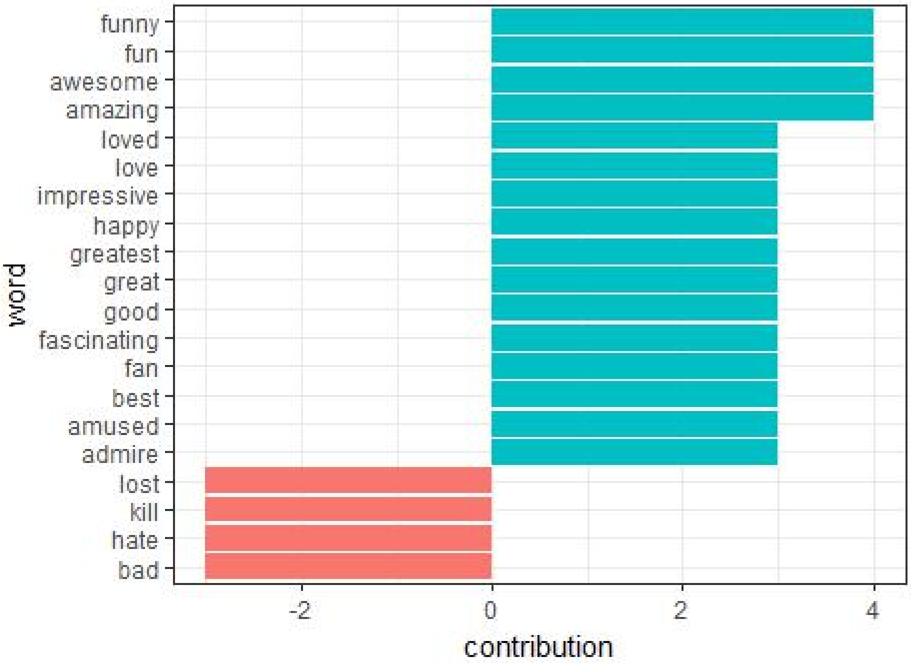


Also. I do the wordcloud for Shila’s reviews. And we can find that the most frequent words in her comments are review, reference, meeting, shops, stores etc. In this way, we can speculate that she is probably a business woman who have to go business trips frequently. Therefore, yelp can make use of this to deliver some related advertisement(like sales in shop, coupon of hotel) on purpose.

8.1.2 Sentiment Analysis for Shila:

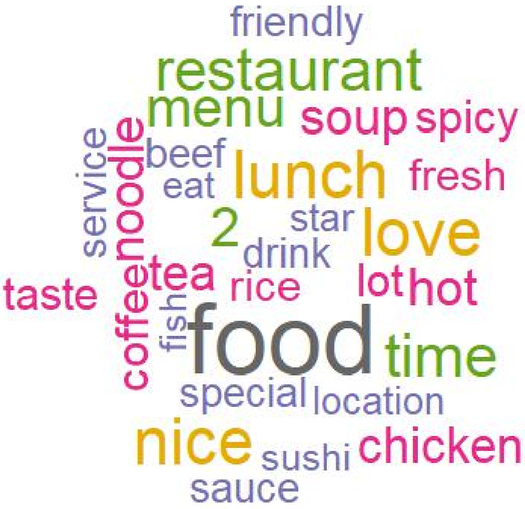
Code:

```
positiveWordsBarGraph <- function(SC) {
  contributions <- SC %>%
    unnest_tokens(word, text) %>%
    count(word,sort = TRUE) %>%
    ungroup() %>%

    inner_join(get_sentiments("afinn"), by = "word") %>%
    group_by(word) %>%
    summarize(occurences = n(),
              contribution = sum(score))

  contributions %>%
    top_n(20, abs(contribution)) %>%
    mutate(word = reorder(word, contribution)) %>%
    head(20) %>%
    ggplot(aes(word, contribution, fill = contribution > 0)) +
    geom_col(show.legend = FALSE) +
    coord_flip() + theme_bw()
}

positiveWordsBarGraph(review %>%
                         filter(user_id == "RtGqdDBvvBCjcu5dUqwfzA"))
```



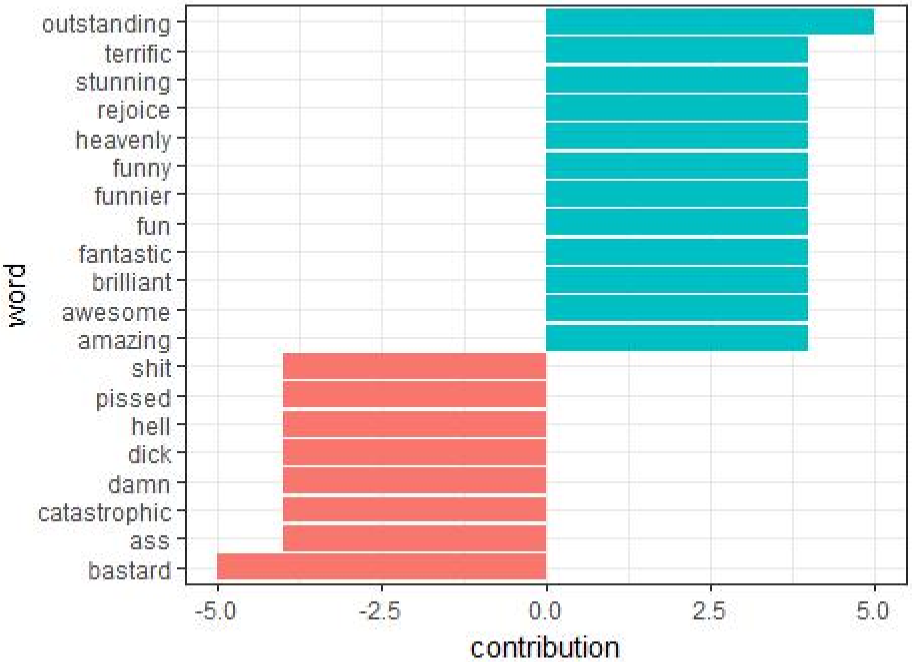
In this sentiment analysis I adapt **AFINN sentiment lexicon**[13], which provides numeric positivity scores for each word and try to visualize them with the bar plot. And we can find that, in Shila’s review, she mainly used fun,awesome, amazing,impressive, loved,happy,great positive words. And relatively few negative words in comment. Therefore we can further assume shila is quite educated.

8.1.3 Wordcloud for Jennifer’s review:



This time we can take a look at the wordcloud of Jennifer. We can find the key words are food, restaurant, chicken, menu, coffee, taste, sushi, spicy, hot etc. All these words are related with food. So we can find that Jennifer is definitely a foodie.

8.1.4 Sentiment analysis for Jennifer:



We can find that positive words in Jennifer's reviews including outstanding, heavenly, funny,fantastic, awesome, brilliant etc. While the negative words are more and much worse than the previous girl. So we can speculate Jennifer as a student.

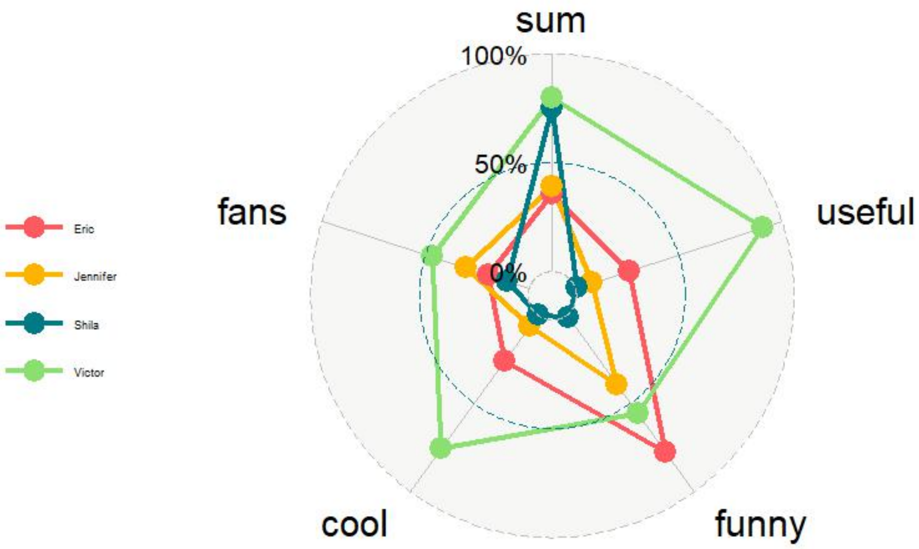
8.1.5 User review comment analysis[14]:

Here I choose Eric, Jennifer, Shila as well as Victor as the sample users and try to figure out their user portrait.

Code:

```
res0<-dbGetQuery(channel, "select name,review_count as sum, useful,funny,cool,fans
from user where user_id LIKE'%CxDOIDnH8gp9KXzpBHJYXw%' or user_id LIKE
'%RtGqdDBvvBCjcu5dUqwfzA%' or user_id LIKE '%HFECrzYDpgbS5EmTBtj2zQ%' or user_id LIKE
'%8k3aO-mPeyhbR5HUucA5aA%';")
```

```
res0 %>%
        mutate(sum=sum/15000) %>%
        mutate( useful= useful/15000) %>%
        mutate( funny=funny/5000) %>%
        mutate(cool=cool/15000) %>%
        mutate(fans=fans/2000)%>%
      ggradar()
```



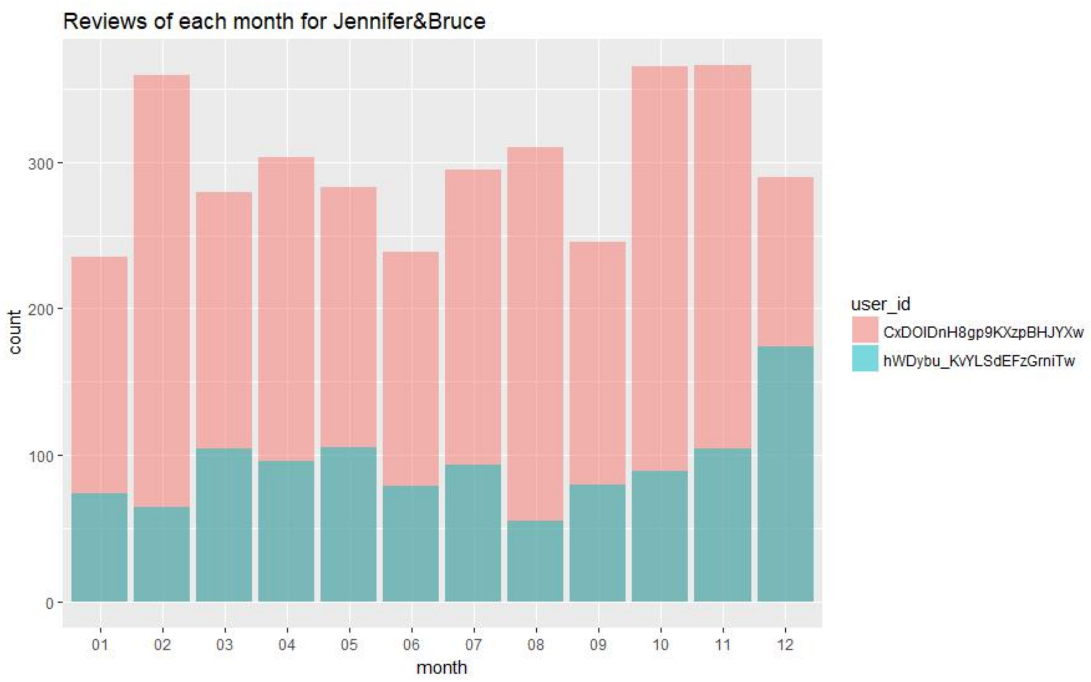
In this radar graph we can clearly find the category of different user. There is no doubt that these four people are all active users. But they are totally different kinds of users. For example, Victor is a kinda a perfect user. He made the most reviews as well as in some other fields. While Shila, made comments as many as Victor did. But her reviews are not quite accepted by other users. So maybe most her reviews won't include any details maybe quite subjective opinions like "It's awesome!""Thats great"etc. So she can be regraded as a general active user. While another example Eric who made only half of the amount of Shila did. But he is the most funny guy, he might be humorous and likes to sharing. So different users can be tagged with different labels which is quite helpful for the friend recommendation and information collection.

8.1.6 # of User reviews related with Month:

Code:

```
jennifer<-(review %>%
filter(user_id =="CxDOIDnH8gp9KXzpBHJYXw"))
month<-format(jennifer$date,format="%m")
jennifer<-cbind(jennifer,month)
bruce<-(review %>%
filter(user_id =="hWDybu_KvYLSdEFzGrniTw"))
month<-format(bruce$date,format="%m")
bruce<-cbind(bruce,month)
together<-rbind(jennifer,bruce)
ggplot(together, aes(x=month,fill=user_id)) +
geom_histogram(stat="count",alpha=.5,position = "identity")+
labs(title="Reviews of each month for Jennifer&Bruce")
```

Result:



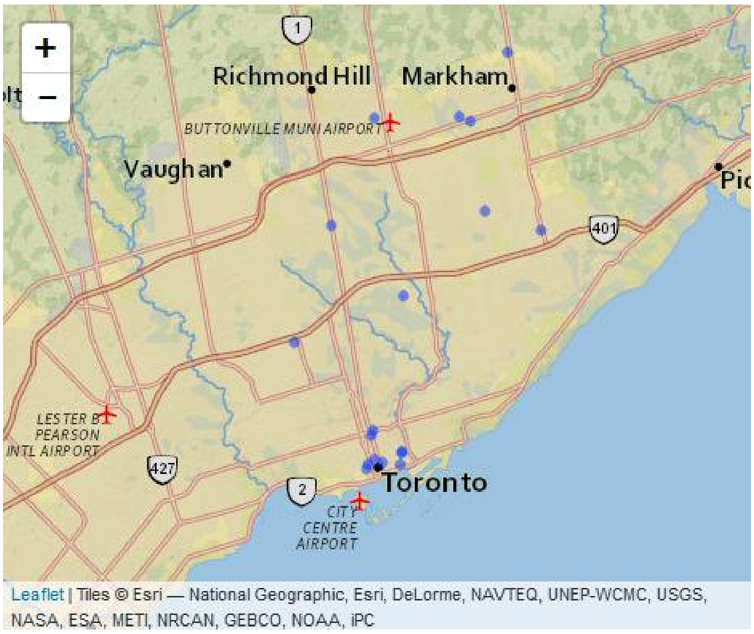
In this graph, we can easily see the relationship between months and # of reviews which further can be regarded as consumption. And here red represents Jennifer and green represents Bruce. We can see that Jennifer keeps a general high consumption all over the year while Bruce is kinda low. In terms of each person, we can find in which months they prefer to consume while in other months they not. E.g for Jennifer, she prefers to go out in Sep and Feb while not much in Jan, Jun and Dec. Maybe it's the cause of weather. While Bruce behaves different, he has a sharply increase in Dec, it's the Christmas break. Maybe he doesn't have time to consume in other times. From which we can aim the ad to different consumers according to their regular activity period.

8.1.7 Stalk Users most recent activity area

In this part, I choose Jennifer as the sample user and stalk her most recent 20 activities she went and gave reviews.

Code:

```
business<-dbGetQuery(channel, "select * from business;")
jennifer<-(review %>%
                filter(user_id ==

                           "CxDOIDnH8gp9KXzpBHJYXw"))
recent_review<-head(jennifer[order(jennifer$date,decreasing = T),],20)
recentactivity<-(business%>%
                       filter(business$business_id %in%
                         ((recent_review %>%
                         filter(user_id ==
                                  "CxDOIDnH8gp9KXzpBHJYXw"))$business_id)
                       )
                   )
recentactivity$latitude<-as.numeric(recentactivity$latitude)
recentactivity$longitude<-as.numeric(recentactivity$longitude)
center_lon = median(recentactivity$longitude,na.rm = TRUE)
center_lat = median(recentactivity$latitude,na.rm = TRUE)

leaflet(recentactivity) %>% addProviderTiles("Esri.NatGeoWorldMap") %>%
  addCircles(lng = ~longitude, lat = ~latitude,radius = 1)    %>%

  # controls
  setView(lng=center_lon, lat=center_lat,zoom = 10)
```



Leaflet | Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC

In this analysis we can easily find the activity area of Jennifer. She went most to the downtown of Toronto and has a sparse activities of suburban of Toronto. So this is a quite useful information for yelp, which can help to send related information much much more precisely.

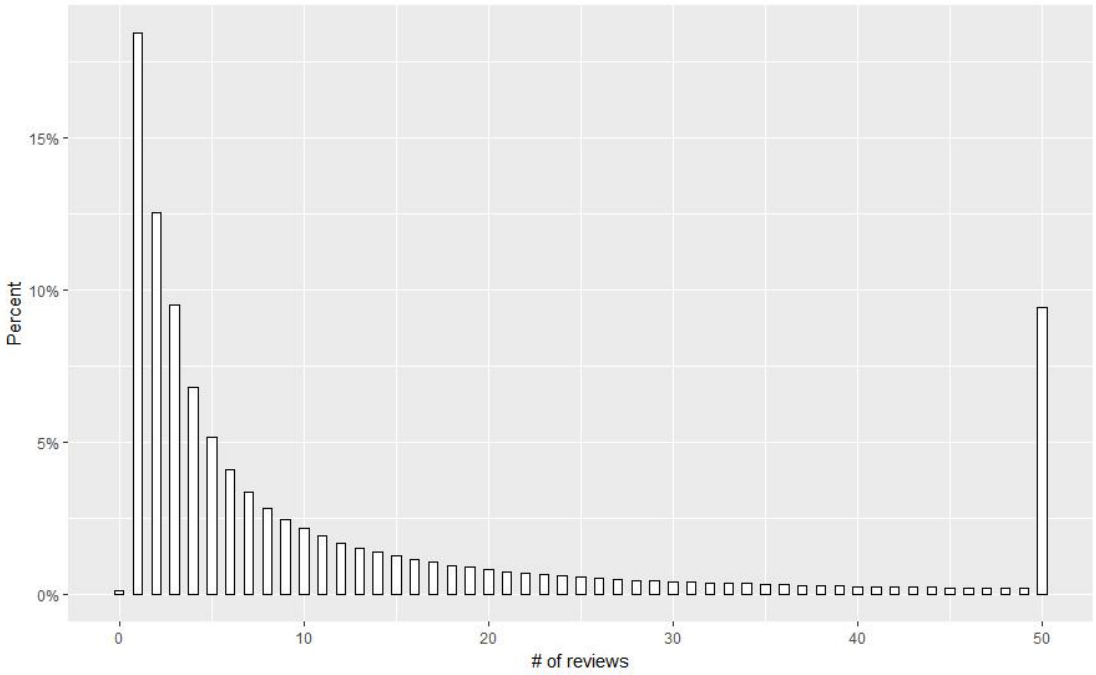
8.2 Further Study of General User Information:

8.2.1General Analysis of the # of reviews users takes:

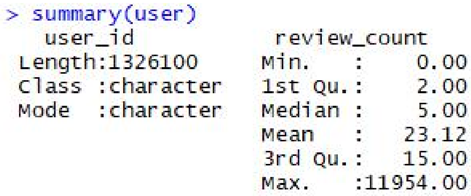
I do a preparation of the data here I put all the user made ever more than 50 reviews together into 50

Code:

```
res<-dbGetQuery(channel, "select review_count,count(*)as freq from user where review_count<50 group by review_count;")
row<-dbGetQuery(channel, "select 50 as review_count,count(*) as freq from user where review_count>=50;")
res1<-rbind(res,row)
res2 <- res1 %>% mutate(f=freq/sum(freq))
ggplot(res2,aes(x=review_count,y=f))+
  geom_bar(stat='identity',width = 0.5,col="black",fill="white")+
  scale_y_continuous(labels = scales::percent)+
  labs(x="# of reviews",y="Percent")
```



General Statistic



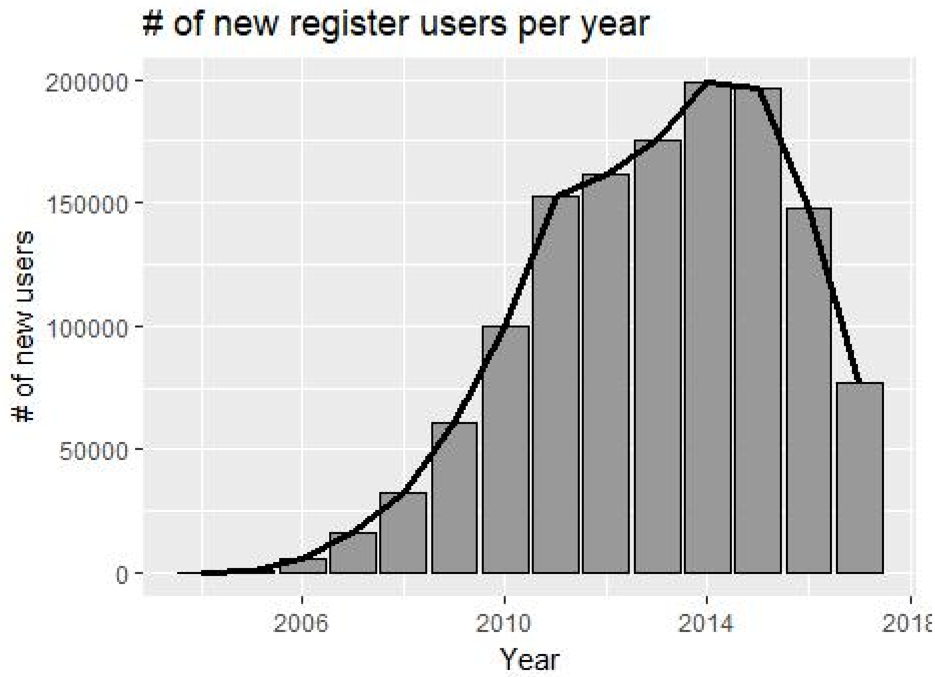
```
> summary(user)
   user_id           review_count
 Length:1326100     Min.   :    0.00
 Class :character   1st Qu.:    2.00
 Mode  :character   Median :    5.00
                    Mean   :   23.12
                    3rd Qu.:   15.00
                    Max.   :11954.00
```

We can find that the mean is 23. Thus, if we regard those users who made reviews less than 20 as inactive users. Then ~80% users are inactive users. Nearly 70% users made less than 10 reviews ever. Most users only ever made 1-5 reviews.

8.2.2New Register User:

Code:

```
res<-dbGetQuery(channel, "select year(yelping_since) as year, count(*)as freq from user group by year(yelping_since);")

temp<-ggplot(res,aes(x=year,y=freq,group=1))+
   geom_bar(stat='identity',col='black',fill='gray60')+
   geom_line(size=1.2)+
   labs(x="Year",y="# of new users",title="# of new register users per year")
temp
```



We all know that yelp was set up in 2004. So the first few years the amount of users are relatively small. But within the following 5-6 years, the number of user has a sharply increase. In some way it's getting better and much more helpful in our life and from the other side it might be help of the popularity of intelligent phones. While, we can also find that there is a sharply decrease since 2015 which may the result of the conflict with other similar apps.

Reference

[1]https://www.yelp.com/dataset/challenge
[2]http://ggplot2.org/
[3]https://rstudio.github.io/leaflet/
[4]http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know
[5]https://www.tidytextmining.com/tidytext.html
[6]https://www.r-bloggers.com/natural-language-processing-tutorial/
[7]Liang S S, Yang J, Wang C, et al. Location based user behavior analysis and applications: U.S. Patent 8,725,569[P]. 2014-5-13.
[8]http://lilibei.net/2017/04/08/R%E8%AF%AD%E8%A8%80%E7%BB%98%E5%88%B6%E9%A2%91%E7%8E%87%E7%9B%B4%E6%96%B9%E5%9B%BE/
[9]https://www.plob.org/article/7264.html
[10]http://www.sthda.com/english/wiki/ggplot2-pie-chart-quick-start-guide-r-software-and-data-visualization
[11]https://rstudio.github.io/leaflet/markers.html
[12]https://www.r-bloggers.com/building-wordclouds-in-r/
[13]Nielsen F Å. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs[J]. arXiv preprint arXiv:1103.2903, 2011.
[14]https://github.com/ricardo-bion/ggradar

Acknowledgement

The author wish to thank New York University, Dr. Raman Kannan, for providing necessary academical assistance to Qi Yao to finish this project successfully. The author also like to thank the continuous support and help from TA, Om Narayan. The author would also like to thank IBM sincerely for sharing IBM Cloud Facility as academical initials! Finally, the author also would like to thank Yelp company for sharing the data to public.