**Trend of new users joining yelp**

**SQL Query:**

select count(\*), yelping\_since

from test\_user

group by year(yelping\_since), month(yelping\_since)

order by year(yelping\_since), month(yelping\_since);

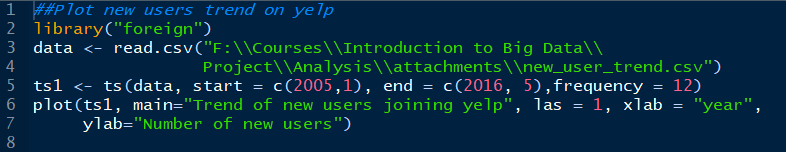
**Visualizing with R:**

library("foreign")

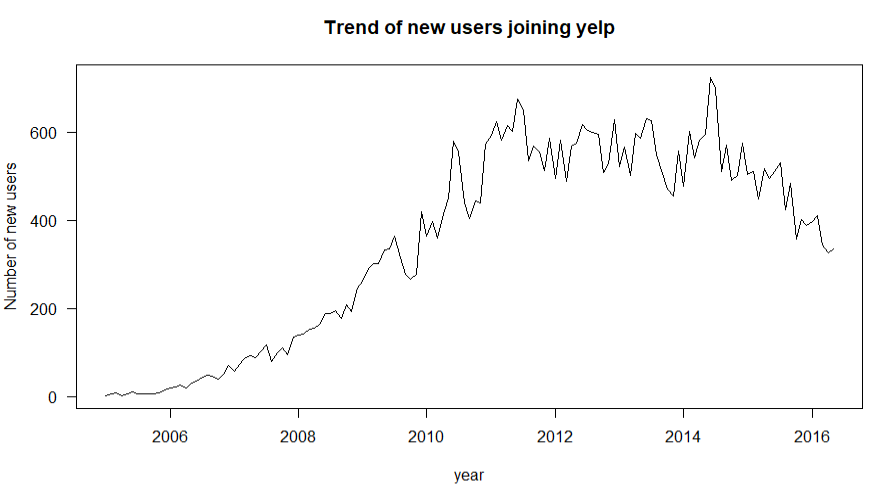
data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\Analysis\\attachments\\new\_user\_trend.csv")

ts1 <- ts(data, start = c(2005,1), end = c(2016, 5),frequency = 12)

plot(ts1, main="Trend of new users joining yelp", las = 1, xlab = "year", ylab="Number of new users")



Results:



**Top 25 cities based on presence of restaurants on Yelp:**

**SQL Query:**

select city, count(\*)

from yelp\_business

group by city

order by count(\*) desc

limit 25;

**Visualizing in R:**

library("RColorBrewer")

data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\

Analysis\\attachments\\Largest\_presence.csv")

cities <- data$city

max\_count <- max(data$count)

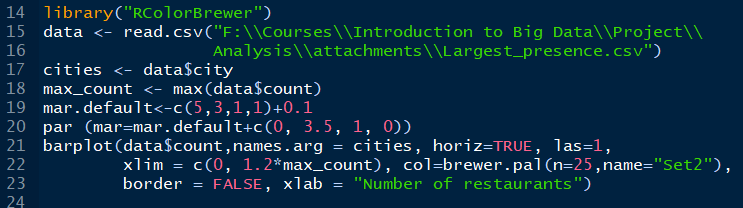
mar.default<-c(5,3,1,1)+0.1

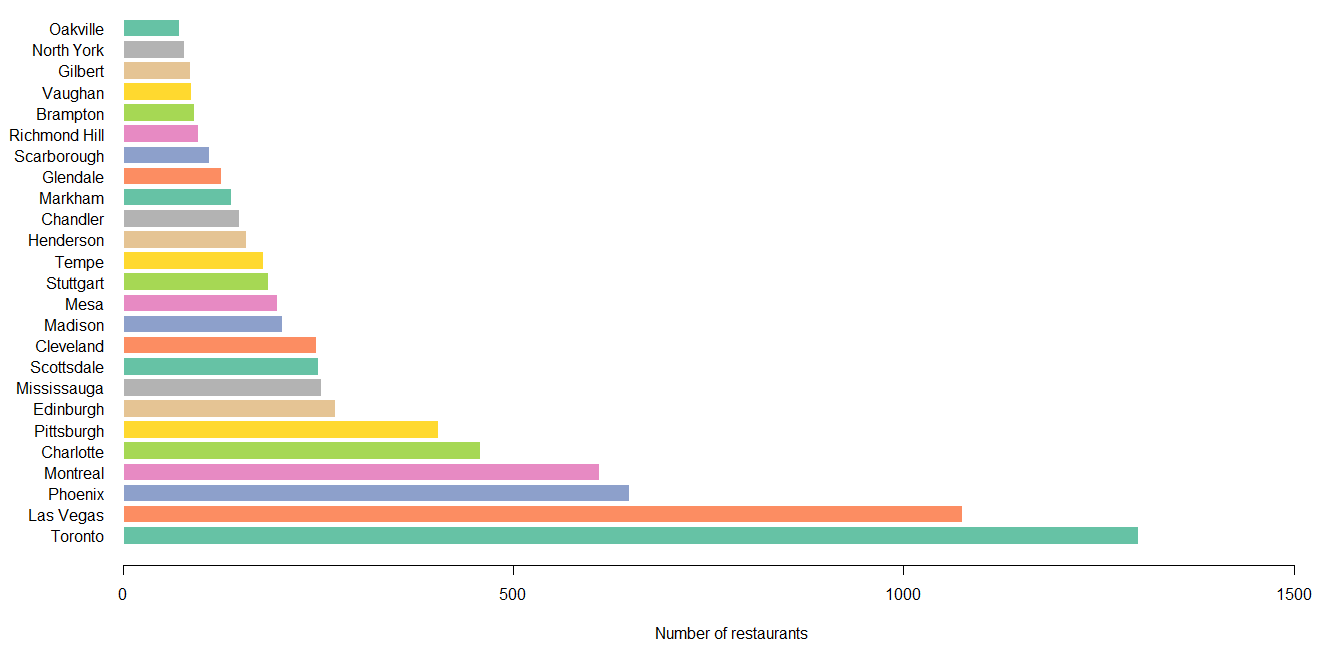
par (mar=mar.default+c(0, 3.5, 1, 0))

barplot(data$count,names.arg = cities, horiz=TRUE, las=1,

xlim = c(0, 1.2\*max\_count), col=brewer.pal(n=25,name="Set2"),

border = FALSE, xlab = "Number of restaurants")





**Find out number of businesses in each top 8 restaurant categories:**

**SQL Query:**

select category, count(\*)

from categories

group by category

order by count(\*) desc

limit 8;

**Visualizing in R:**

library("RColorBrewer")

data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\

Analysis\\attachments\\top\_categories.csv")

mar.default<-c(5,6,1,1)+0.1

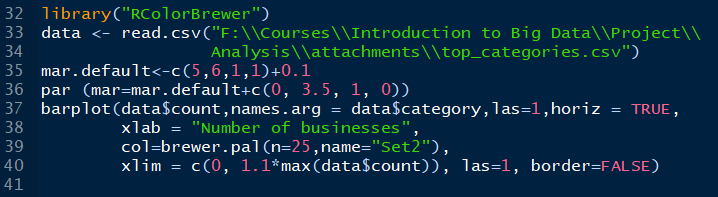
par (mar=mar.default+c(0, 3.5, 1, 0))

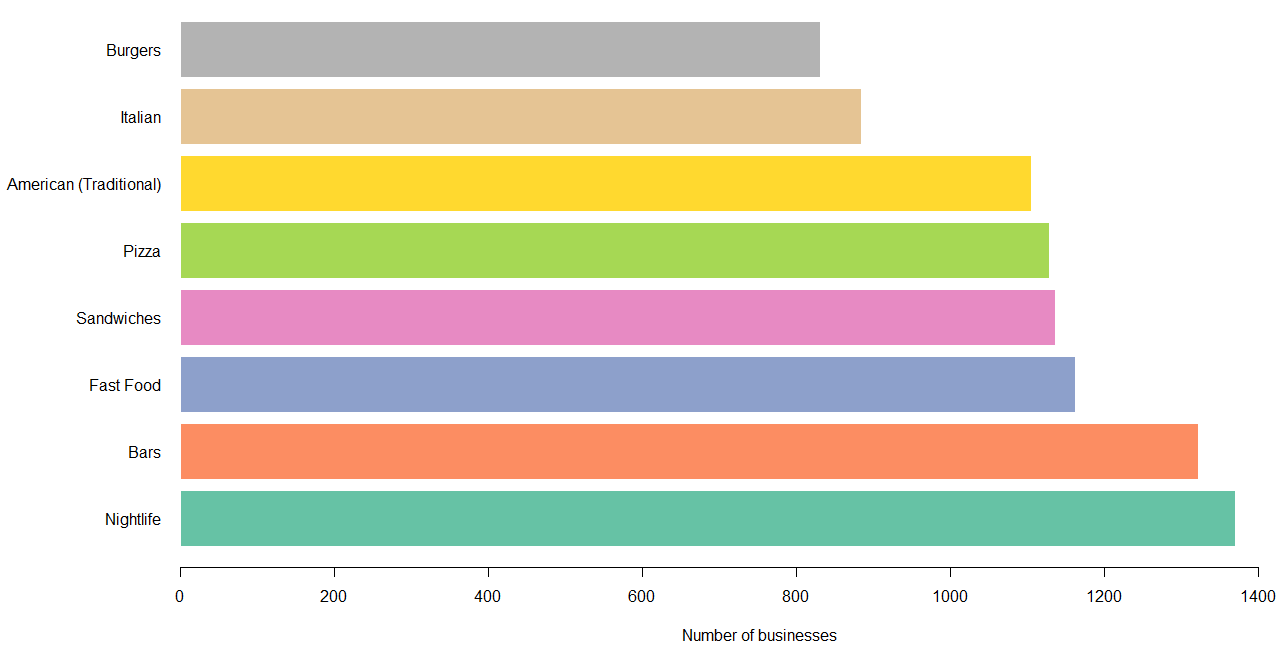
barplot(data$count,names.arg = data$category,las=1,horiz = TRUE,

xlab = "Number of businesses",

col=brewer.pal(n=25,name="Set2"),

xlim = c(0, 1.1\*max(data$count)), las=1, border=FALSE)





**Usual busy hours at restaurants:**

**SQL Query:**

select time,count(\*) from checkedin group by time order by count(\*);

**R Visualization:**

**library("RColorBrewer")**

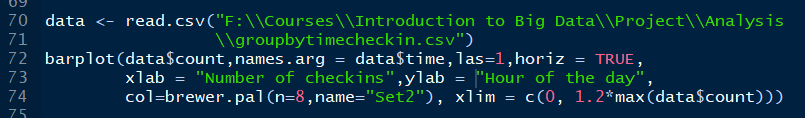
data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\Analysis\\groupbytimecheckin.csv")

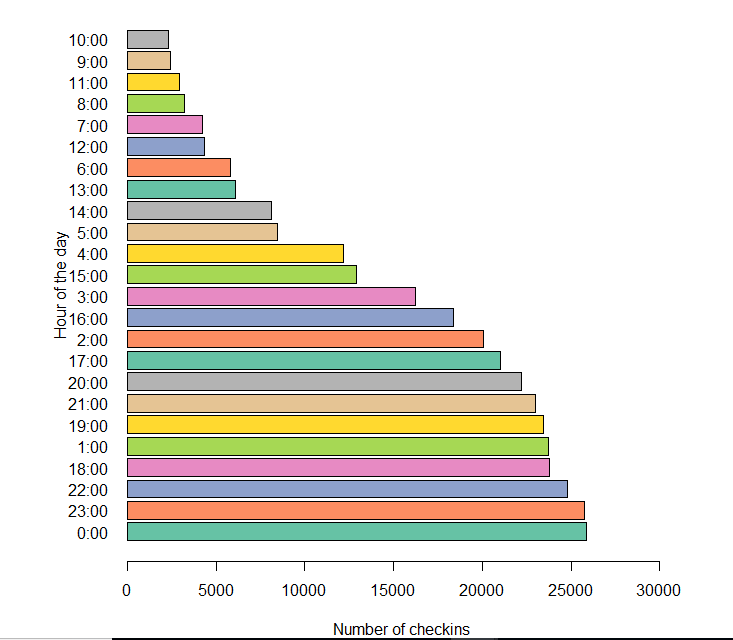
data

barplot(data$count,names.arg = data$time,las=1,horiz = TRUE,

xlab = "Number of checkins",ylab = "Hour of the day",

col=brewer.pal(n=8,name="Set2"), xlim = c(0, 1.1\*max(data$count)))





**Usual busy days for restaurants:**

**SQL Query:**

select day,count() from checkedin group by day order by count() desc;

**Visualization in R:**

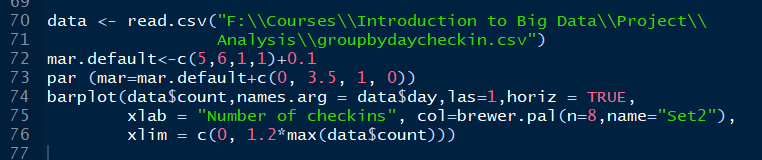
data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\Analysis\\groupbydaycheckin.csv")

mar.default<-c(5,6,1,1)+0.1

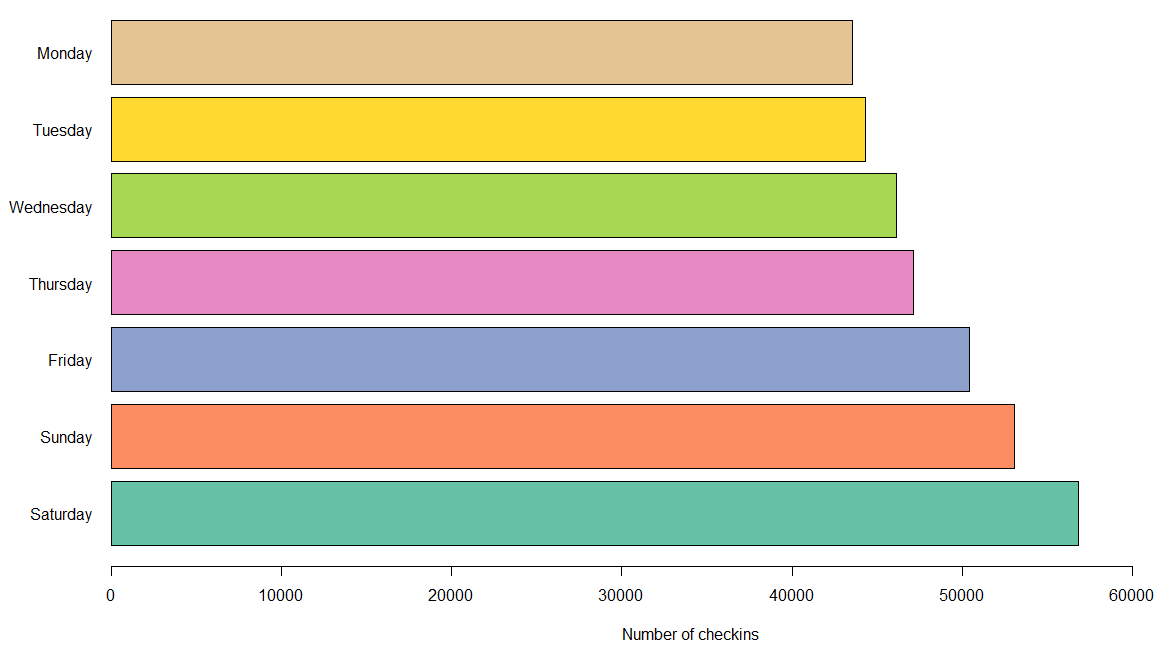
par (mar=mar.default+c(0, 3.5, 1, 0))

barplot(data$count,names.arg = data$day,las=1,horiz = TRUE,

xlab = "Number of checkins", col=brewer.pal(n=8,name="Set2"), xlim = c(0, 1.2\*max(data$count)))



Results:



**Trend of number of reviews being posted on Yelp:**

**SQL Query:**

select year(date), month(date),count(\*)

from review

group by year(date), month(date)

order by year(date), month(date);

**Visualizing in R:**

data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\

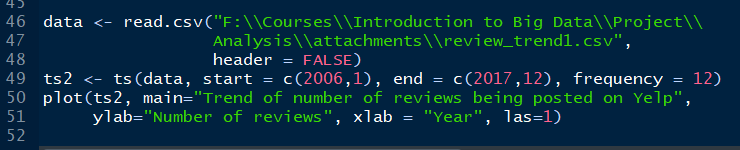
Analysis\\attachments\\review\_trend1.csv",

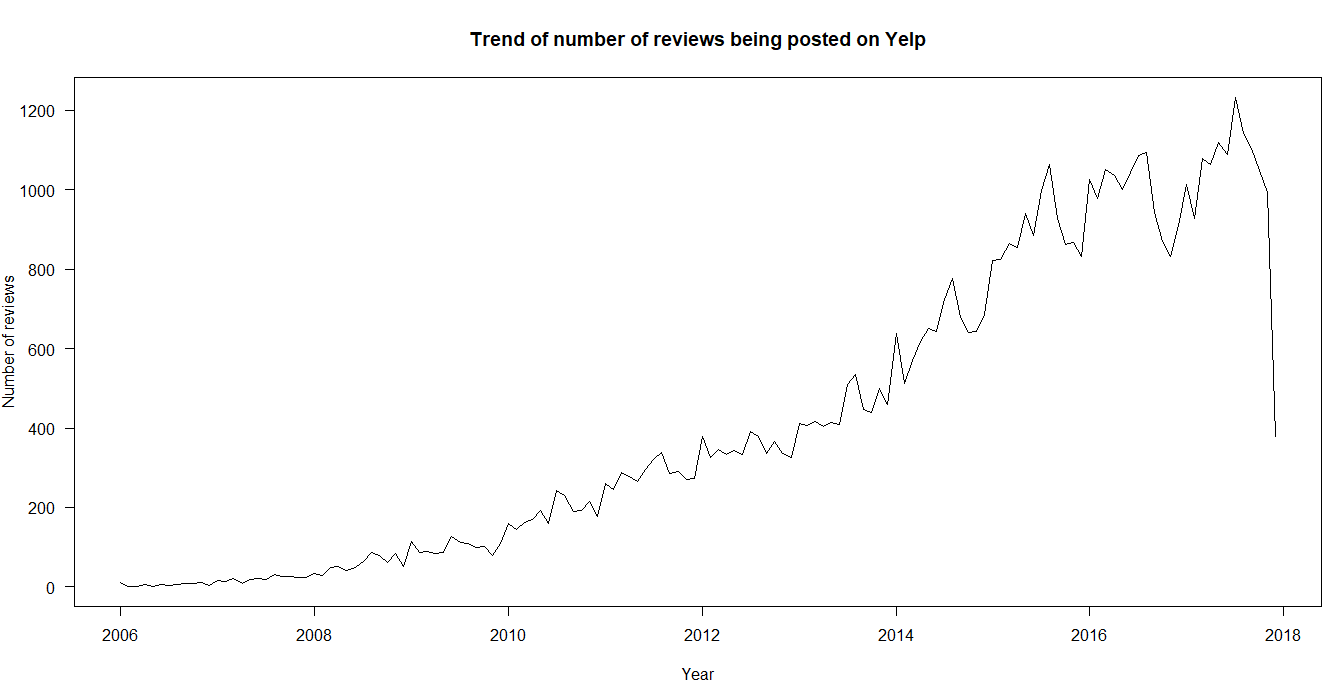
header = FALSE)

ts2 <- ts(data, start = c(2006,1), end = c(2017,12), frequency = 12)

plot(ts2, main="Trend of number of reviews being posted on Yelp",

ylab="Number of reviews", xlab = "Year", las=1)





**Overall percentage distribution of reviews according to rating**

**Query:**

data <- read.csv("F:\\Courses\\Introduction to Big Data\\Project\\Analysis\\all.csv")

slices <- c(data$count)

labels <- c(data$stars)

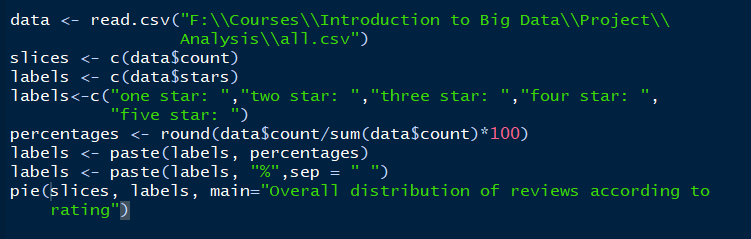
labels<-c("one star: ","two star: ","three star: ","four star: ","five star: ")

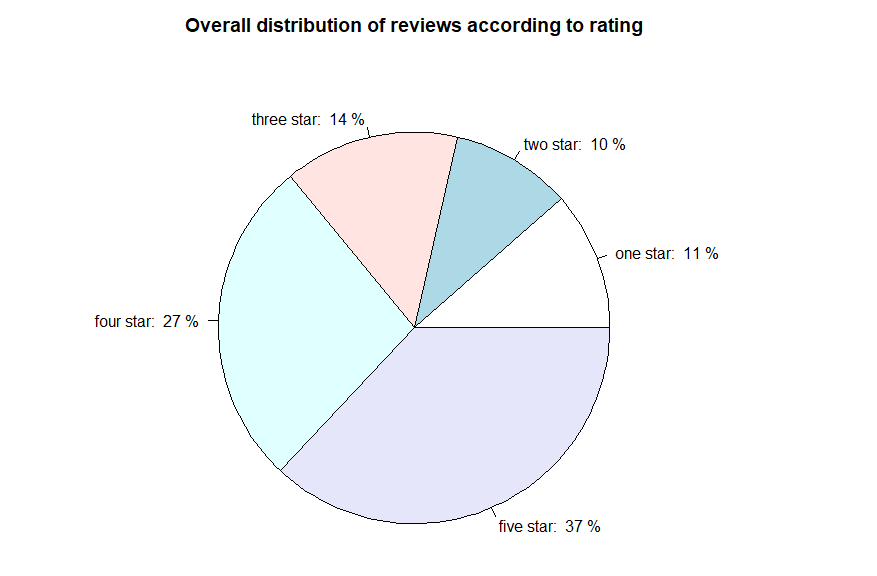
percentages <- round(data$count/sum(data$count)\*100)

labels <- paste(labels, percentages)

labels <- paste(labels, "%",sep = " ")

pie(slices, labels, main="Overall distribution of reviews according to rating")





**Finding out most common words occurring in reviews accompanying 5-star, 1-star or 3-star ratings:**

**Code:**

import nltk as nl;

from nltk.corpus import stopwords;

from collections import Counter;

def **main**(input\_file):

stopword\_set = set(stopwords.words(*"english"*))

frequent\_words = []

all\_words = set()

with open(input\_file) as tf:

next(tf)

data = tf.read()

data = data.lower()

without\_stopwords = []

data = nl.word\_tokenize(data)

for word in data:

if word not in stopword\_set and word.isalpha():

without\_stopwords.append(word)

print(Counter(without\_stopwords).most\_common(5))

if \_\_name\_\_ == *"\_\_main\_\_"*:

main(*".csv"*)

**Results:**

**Most common 10 words in reviews with 5-star rating:**

[('food', 14631), ('great', 13068), ('place', 12840), ('good', 10378), ('service', 7735), ('best', 6250), ('time', 5879), ('one', 5799), ('delicious', 5750), ('back', 5686)]

**Most common 10 words in reviews with 3-star rating:**

[('good', 7951), ('food', 7676), ('place', 5266), ('like', 4605), ('would', 3980), ('service', 3679), ('really', 3320), ('one', 3188), ('get', 3084), ('time', 3064)]

**Most common 10 words in reviews with 1-star rating:**

[('food', 6316), ('place', 3749), ('service', 3372), ('us', 3210), ('would', 3029), ('like', 2987), ('back', 2848), ('one', 2812), ('order', 2735), ('get', 2727)]