This Dataset is publicly available as part of the Yelp Dataset Challenge. It includes information about local businesses in 10 cities across 4 countries.

We have analysed reviews limited to restaurants only as there are over 80,000 businesses where there are close to 3million reviews of customers. Since, the scope of the analysis cannot be so large, we have done a detailed analyses of the Restaurants mainly, Restaurants in Pittsburgh.

By exploring this data we are trying to answer a couple of interesting questions like- 1. What is the distribution of Average Ratings like over-all? 2. Is there a correlation between the Price-Range that a restaurant falls under and it's Average Rating? 3. Where are the maximum number of 5 star rated restaurants located (within the scope of our data-set)? 4. What are the top categories that most of the 5 star rated restaurants fall under? 5. Which city is the veggie-friendly's paradise?

A few questions based on the Yelp Reviews Dataset that this analysis attempts to answer (reviews limited to Restaurants in Pittsburgh) -1. Can we show an interactive map of restaurants in Pittsburgh with an indication of their Ratings? 2. What are the most frequently occurring phrases in reviews for highly rated restaurants and not-so highly rated restaurants. Is the difference apparent? 3. How do maximum number of ratings for a restaurant compare with the number of high Ratings? Are we making a wise choice by just looking at the Average rating or the number of Ratings? 4. Which neighborhoods house the maximum number of highly rated restaurants?

Methodology Employed

After cleaning the data, some wrangling was performed using Dplyr and Tidyr functions to bring the data into a suitable format before performing Exploratory Data Analysis on it to answer the above questions. Slicing and Dicing was performed on the data using dplyr

Packages Required

```
library(stringr)        #This package is used for string manipulation
functions
library(dplyr)          #This package is used for data manipulation tasks
library(tidyr)          #This package is used for data manipulation tasks
library(data.table)     #This package is used to access the function fread
which is a better/faster way to read large data
library(wordcloud)      #This package is used to generate word clouds
library(tm)             #This is a text mining package used in the process
of generating word clouds
library(RWeka)          # This package is used to generate Bigramws and
Trigrams
library(ggplot2)        #This package is used for visualizations
(chart/graph plotting functions)
library(ggmap)          #This package is used for map functions
library(maps)           #This package is used for map functions
library(leaflet)        #This package is used for plotting maps
library(knitr)
```

Data Preparation

Original Source of the data: https://www.kaggle.com/athoul01/predicting-yelp-ratings-from-review-text/data

The key attributes of the data are as follows:

- 2.7M reviews and 649K tips by 687K users for 86K businesses.
- 566K business attributes, e.g., hours, parking availability, ambience etc.,
- Social network of 687K users for a total of 4.2M social edges
- Aggregated check-ins over time for each of the 86K businesses
- 200,000 pictures from the included businesses.

The dataset consists of six files - business, tips, reviews, users, check-ins and photos. Each file is composed of a single object type, one json-object per-line. For the purposes of this project, we will be working with 3 of these - business, user and reviews. Also, we would be working with the Review Data only for one city - Pittsburgh.

Reading the Data

```
data_business <-
fread("https://s3.amazonaws.com/shreyayelp/yelp_academic_dataset_business.c
sv")
#Dimensions and attribute Names for the Business Data
dim(data_business)
```

```
[1] 85901    98
```

```
DT::datatable(as.data.frame(names(data_business)))
```

```
data_user <-
fread("https://s3.amazonaws.com/shreyayelp/yelp_academic_dataset_user.csv")
```

```
Read 24.8% of 686556 rows
Read 52.4% of 686556 rows
Read 83.0% of 686556 rows
Read 686556 rows and 23 (of 23) columns from 0.152 GB file in 00:00:06
```

```
#Dimensions and attribute Names for the User Data
dim(data_user)
```

```
[1] 686556     23
```

```
DT::datatable(as.data.frame(names(data_user)))
```

```
pa_reviews <-
fread("https://s3.amazonaws.com/shreyayelp/yelp_academic_pittsburgh_restaur
ant_reviews.csv")
#Dimensions and attribute Names for the Reviews Data
dim(pa_reviews)
```

```
[1] 74295    11
```

```
DT::datatable(as.data.frame(names(pa_reviews)))
```

Cleaning up Reviews Data to filter Reviews for Pittsburgh only-

The Reviews Data file was read and filtered for restaurant reviews and further filtered for one city - Pittsburgh. This was done in order to be able to work with a smaller data-set for the scope of this project. Since the original Reviews data is 2.2GB. The following piece of code was used to get to the reviews dataset for restaurants in Pittsburgh from the original reviews dataset.

```
# Reviews Data only for Restaurants
    data_reviews_restaurant <- data_reviews[data_reviews$business_id %in%
restaurants$business_id,]
    write.csv(data_reviews_restaurant,
"data/yelp_academic_restaurant_reviews.csv")
    # Reviews Data only for Restaurants in Pittsburgh
    pa_restaurants <- restaurants %>% filter(city=="Pittsburgh") %>%
select(business_id)
    pa_reviews <-
data_reviews_restaurant[data_reviews_restaurant$business_id %in%
pa_restaurants$business_id,]
    write.csv(pa_reviews,
"data/yelp_academic_pittsburgh_restaurant_reviews.csv")
```

```
# The Final Reviews Data Dimensions and Attribute Names
dim(pa_reviews)
```

Cleaning up Businesses Data to filter only Restaurants Data

```
# Extracting only Restaurant information from the data_business data set
restaurants <- data_business[grepl('Restaurant',data_business$categories),]
#The new data-set restaurants has the following dimensions
dim(restaurants)
```

```
[1] 26729    98
```

```
# Marking missing values as NA
restaurants[restaurants==""] <- NA
sum(is.na(restaurants))
```

```
[1] 1218596
```

```
# Removing columns where all the values are NA
NA_values <- is.na(restaurants)
NA_Count <- apply(NA_values,2,sum)
NA_Count_df <- NA_Count[NA_Count==dim(restaurants)[1]]
#The following columns are irrelevant for restaurants and are completely
empty
```

```
names(NA_Count_df)
```

```
[1] "attributes.Hair Types Specialized In.africanamerican"
[2] "attributes.Hair Types Specialized In.kids"
[3] "attributes.Hair Types Specialized In.straightperms"
[4] "attributes.Hair Types Specialized In.asian"
[5] "attributes.Hair Types Specialized In.coloring"
[6] "attributes.Hair Types Specialized In.extensions"
[7] "attributes.Hair Types Specialized In.perms"
[8] "attributes.Hair Types Specialized In.curly"
```

```
restaurants <- subset(restaurants, select = !(c(names(restaurants)) %in%
c(names(NA_Count_df))))
# Cleaning up the Column Names of the Restaurant Data Set
names(restaurants) <- sapply(list(names(restaurants)),function(x)
str_replace(x,"attributes.",""))
# Final Restaurants Data Dimesnions and Attribute Names
dim(restaurants)
```

```
[1] 26729    90
```

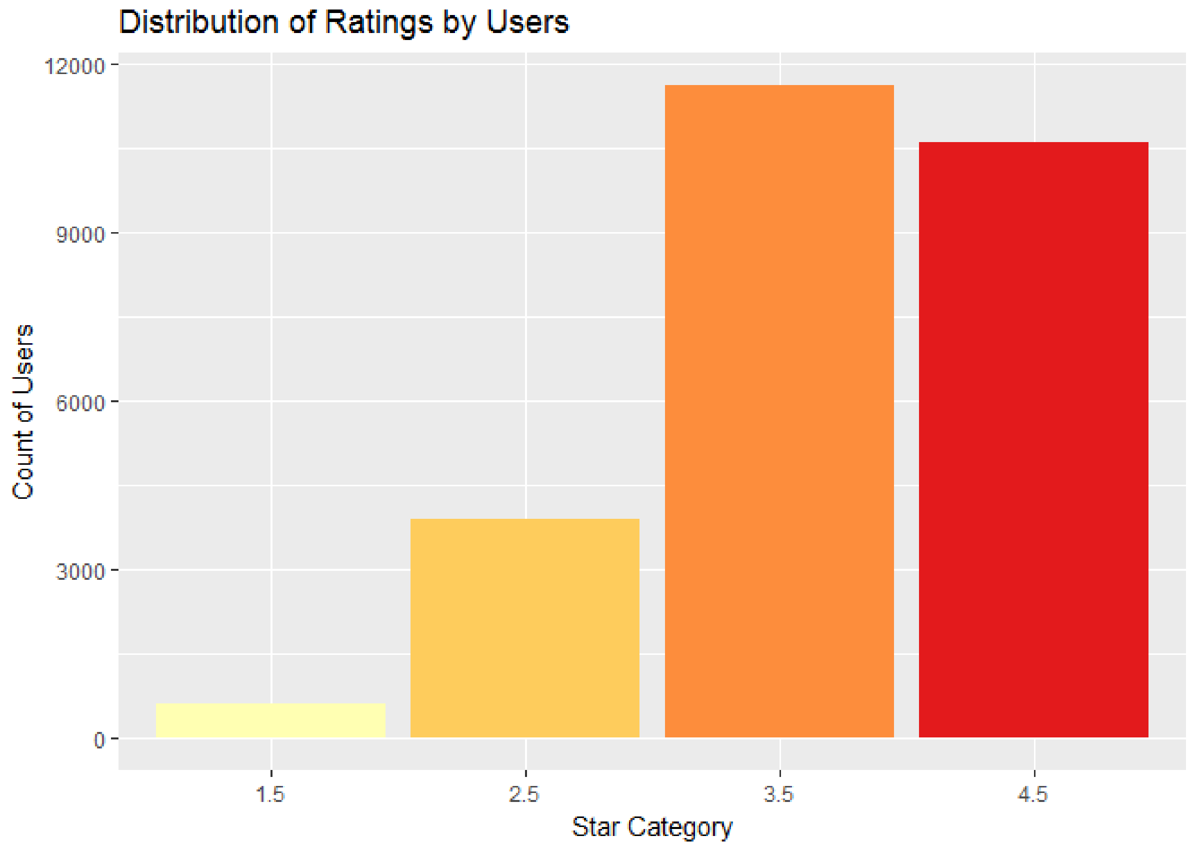
```
DT::datatable(as.data.frame(names(restaurants)))
```

Exploratory Data Analysis

Here, we explore the data of all the Restaurants.

1. Distribution of Ratings

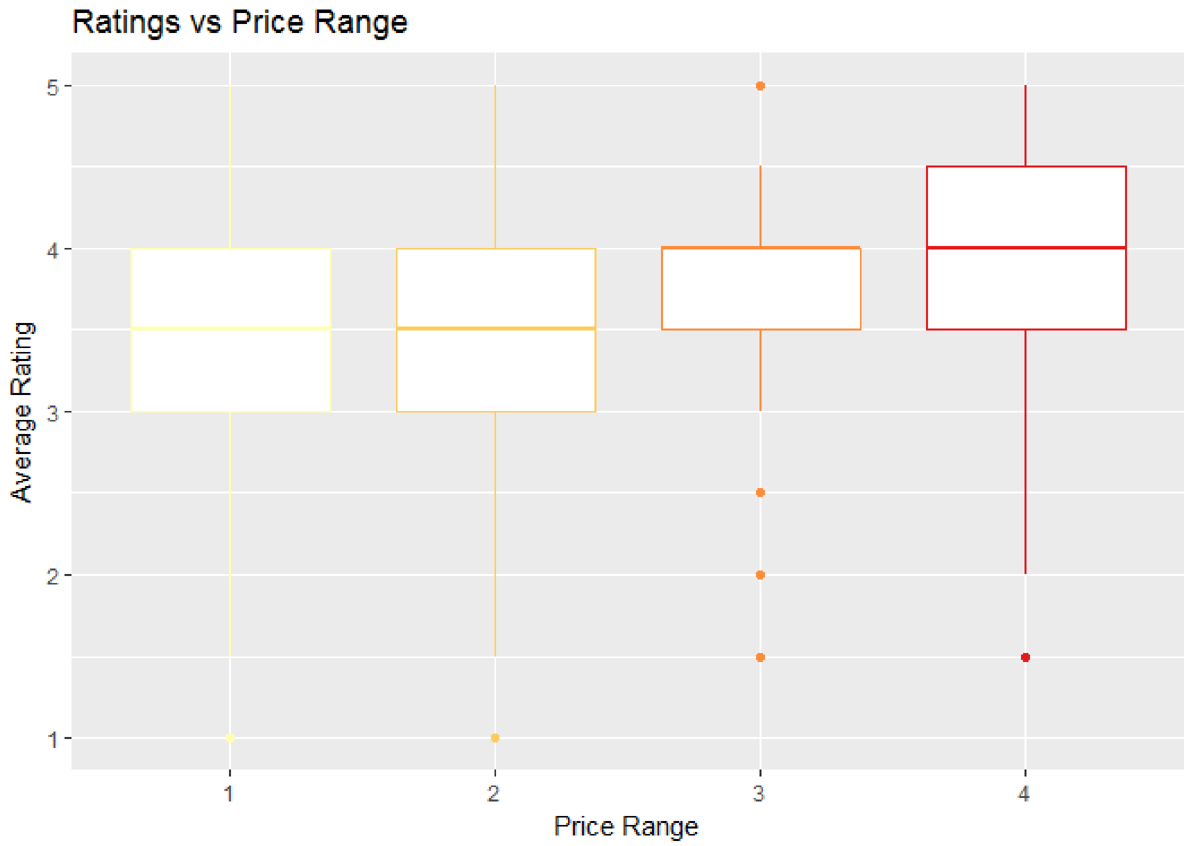
```
ratings_and_users <- restaurants %>% group_by(stars) %>% count()
ggplot(data=ratings_and_users, aes(x=cut(stars,c(0,0.9,1.9,2.9,3.9,5)),
y=n,fill=cut(stars, c(0,0.9,1.9,2.9,3.9,5)))) +
  geom_bar(stat="identity") + scale_fill_brewer(palette = "YlOrRd") +
labs(title = "Distribution of Ratings by Users", y = "Count of Users", x =
"Star Category", fill = "Star Category") + theme(legend.position="none") +
  scale_x_discrete(labels=c("1.5","2.5","3.5","4.5"))
```



The distribution of ratings is considerably skewed, there are a lot more 4 and 5 star ratings than 1,2 and 3. we could say that people seem to review things they like. In general, people seem to be more likely to write a review for a positive experience than a negative one.

2. <u>Correlation of Price on Ratings</u>

```
ggplot(data=restaurants[!is.na(restaurants$`Price Range`)],
aes(x=factor(`Price Range`), y=stars, color=factor(`Price Range`))) +
  geom_boxplot()+ labs(title = "Ratings vs Price Range", y = "Average
Rating", x = "Price Range", fill = "Price Range") +
scale_color_brewer(palette = "YlOrRd") + theme(legend.position="none")
```

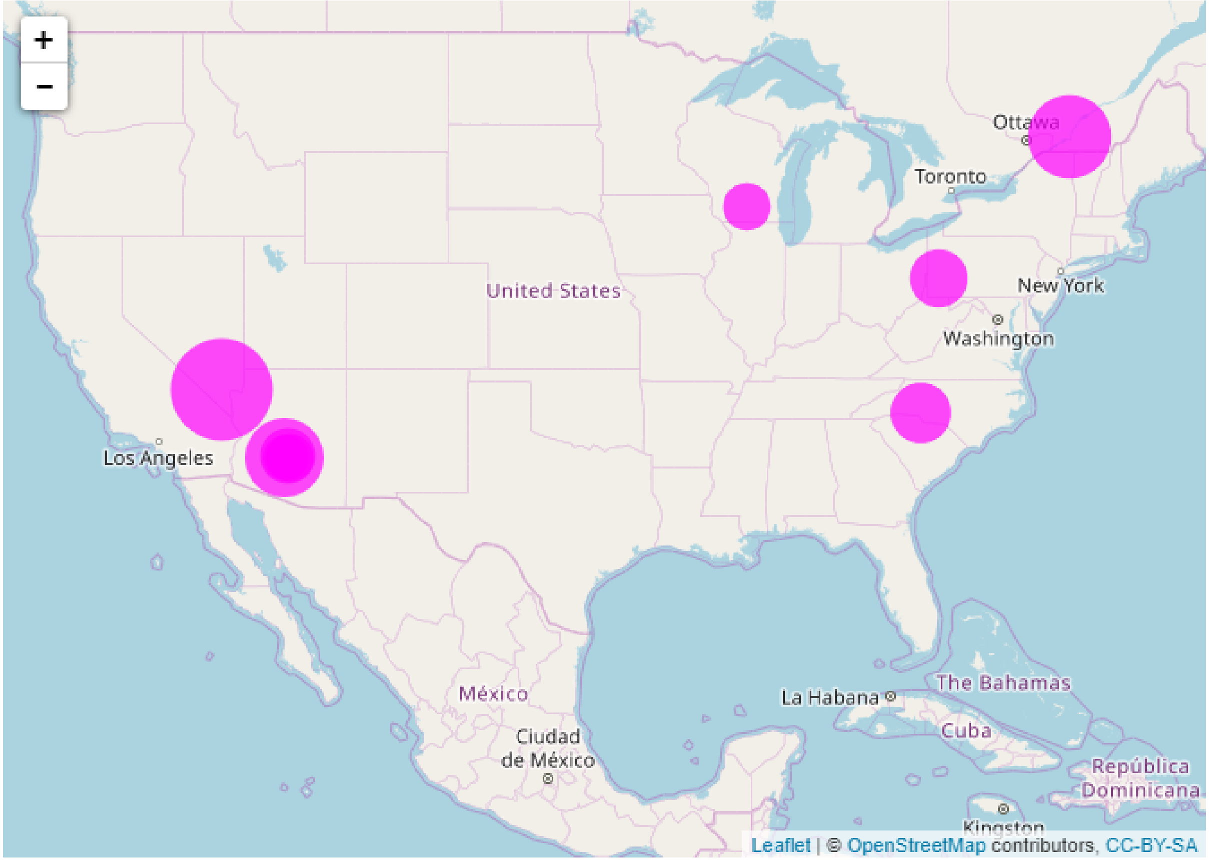


It appears that there is some evidence of this hypothesis from the data. Especially, the restaurants in the Price Bracket of 4 (highest) have a higher average Rating.

Where are all the 5-star restaurants located?

```
top_cities <- restaurants %>% filter(stars>=4) %>% group_by(city) %>%
summarize(n=n()) %>% arrange(desc(n)) %>% head(10)
top_cities$longitude <- rep(0,length(top_cities$city))
top_cities$latitude <- rep(0,length(top_cities$city))
long_lat <- geocode(as.character(top_cities$city))
top_cities$longitude <- long_lat[,1]
top_cities$latitude <- long_lat[,2]

map1 <- leaflet(top_cities) %>% addTiles() %>% setView(lng = -96.503906,
lat = 38.68551, zoom = 4) %>%
  addCircleMarkers(lng = ~longitude, lat = ~latitude, weight = 0,
radius=~n/100+10, fillOpacity = 0.7 , color="Magenta" , popup = ~city)
map1
```



Screenshot of the map generated.

What are the top Categories that 5-star rated restaurants are listed under?

```
top_rated_categories <- restaurants %>% filter(stars==5) %>%
select(name,categories)
docs <- Corpus(VectorSource(top_rated_categories$categories))
#Converting to lower case, removing stopwords, punctuation and numbers
docs <- tm_map(docs, removePunctuation)
#docs <- tm_map(docs, removeNumbers)
docs <- tm_map(docs, tolower)
docs <- tm_map(docs, removeWords, c(stopwords("english"),"s","ve"))
docs <- tm_map(docs, PlainTextDocument)
#Term Document Matrix
Ngrams <- function(x) NGramTokenizer(x, Weka_control(min = 1, max = 3))
tdm <- TermDocumentMatrix(docs, control = list(tokenize = Ngrams))
freq = sort(rowSums(as.matrix(tdm)),decreasing = TRUE)
freq.df = data.frame(word=names(freq), freq=freq)
DT::datatable(as.data.frame(freq.df[3:18,]$word))
```

1	bars
2	mexican
3	sandwiches
4	cafes
5	mexican restaurants
6	american
7	pizza

8	sandwiches restaurants
9	italian
10	cafes restaurants

Which city is the veggie-friendly’s paradise?

```
veggie_friendly_city <- restaurants[city!='Edinburgh',] %>% filter(`Dietary
Restrictions.vegetarian`=="TRUE") %>%  group_by(city) %>%
summarise(count=n()) %>% arrange(desc(count)) %>% subset(city!='Karlsruhe')
DT::datatable(veggie_friendly_city)
```

S.No	City	Count
1	Phoenix	19
2	Las Vegas	9
3	Scottsdale	6
5	Tempe	4
4	Chandler	4
9	Urbana	2
8	Pittsburgh	2
7	Gilbert	2
6	Charlotte	2
11	Henderson	1

Which are the best rated restaurants for vegetarians?

```
veggie_friendly_restaurants <- restaurants %>% filter(`Dietary Restrictions.vegetarian` == "TRUE") %>%
filter(city == "Phoenix") %>% arrange(desc(stars)) %>% select(name, stars, categories)
DT::datatable(veggie_friendly_restaurants)
```

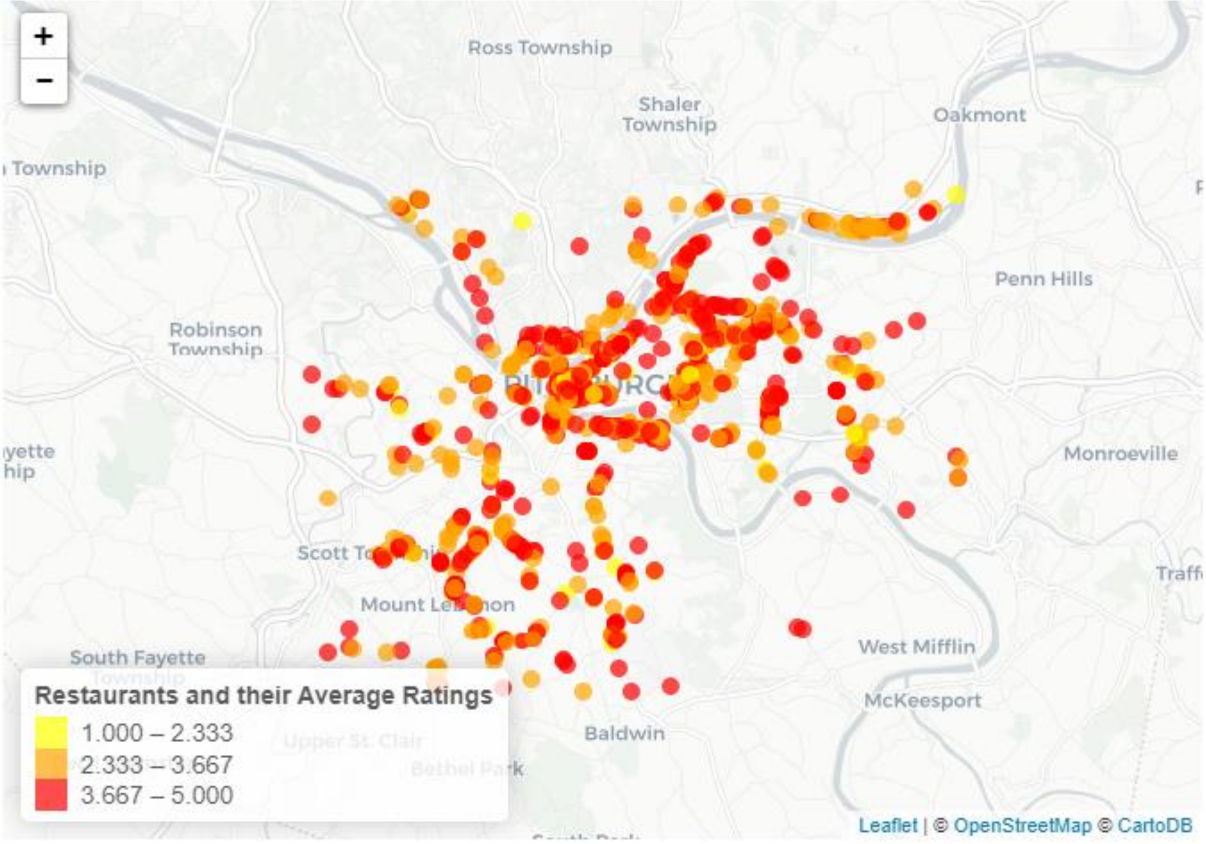
Name of the restaurant	Star rating	categories
Postino Arcadia	4.5	['Bars', 'Breakfast & Brunch', 'Wine Bars', 'Nightlife', 'Italian', 'Restaurants']
Loving Hut	4	['Vegetarian', 'Vietnamese', 'Vegan', 'Restaurants']
Persian Garden Cafe	4	['Vegetarian', 'Persian/Iranian', 'Mediterranean', 'Restaurants']
Pink Pepper Thai Cuisine	4	['Thai', 'Restaurants']
Thai Elephant	4	['Thai', 'Restaurants']
Hana Japanese Eatery	4	['Sushi Bars', 'Japanese', 'Restaurants']
Cherryblossom Noodle Cafe	4	['Noodles', 'Ramen', 'Japanese', 'Restaurants']
El NorteÃ±o	4	['Mexican', 'Restaurants']

Phoenix comes out on top for maximum number of vegetarian-friendly restaurants and Postino- Arcadia in particular enjoys the best average rating.

Analysis on Reviews

An interactive map of restaurants in Pittsburgh with their average ratings indicated by color.

```
binpal <- colorBin(c("Yellow", "Orange", "Red"), pa_restaurants$stars, 3,
pretty = FALSE)
map2 <- leaflet(pa_restaurants) %>% addProviderTiles("CartoDB.Positron")
%>%
  addCircleMarkers(lng = ~longitude, lat = ~latitude, weight = 0, radius=5,
fillOpacity = 0.7 , color = ~binpal(stars), popup = ~name) %>%
  addLegend("bottomleft", pal=binpal, values=stars, title = "Restaurants
and their Average Ratings", labels=~binpal(stars),labFormat =
labelFormat(prefix = ""), opacity = 0.7)
map2
```



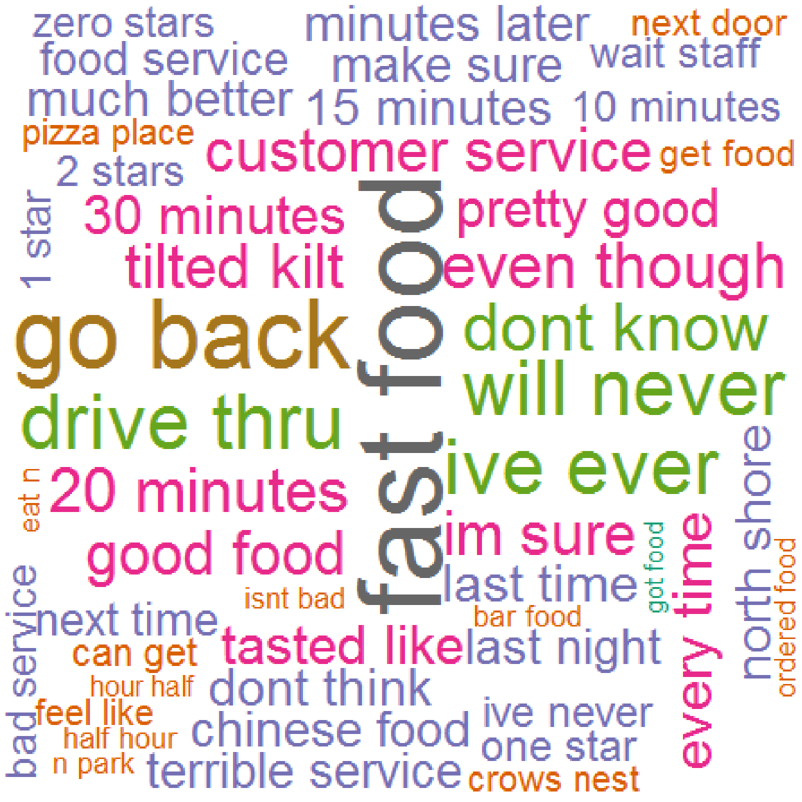
We built Word Clouds. This function takes a single parameter that indicates the Rating Range (higher or lower) and builds word clouds for words occuring with the highest frequencies in reviews for these restaurants.

```
build_word_cloud <- function(range)
{
  if(range=="lower")
    wordcloud_restaurants <- pa_restaurants %>% filter(stars<2.5) %>%
select(business_id,name)
  else
    wordcloud_restaurants <- pa_restaurants %>% filter(stars>4.5) %>%
select(business_id,name)
```

```
  wordcloud_restaurants_reviews <- pa_reviews %>% filter(business_id %in%
wordcloud_restaurants$business_id)
  dir(wordcloud_restaurants_reviews$text)
  docs <- Corpus(VectorSource(wordcloud_restaurants_reviews$text))
  #Converting to lower case, removing stopwords, punctuation and numbers
  docs <- tm_map(docs, removePunctuation)
  #docs <- tm_map(docs, removeNumbers)
  docs <- tm_map(docs, tolower)
  docs <- tm_map(docs, removeWords, c(stopwords("english"),"s","ve"))
  docs <- tm_map(docs, PlainTextDocument)
  #Term Document Matrix
  Ngrams <- function(x) NGramTokenizer(x, Weka_control(min = 2, max = 3))
  tdm <- TermDocumentMatrix(docs, control = list(tokenize = Ngrams))
  freq = sort(rowSums(as.matrix(tdm)),decreasing = TRUE)
  freq.df = data.frame(word=names(freq), freq=freq)
  wordcloud(freq.df$word,freq.df$freq,min.freq=5,max.words=150,random.order
= F, colors=brewer.pal(8, "Dark2"))
}
```

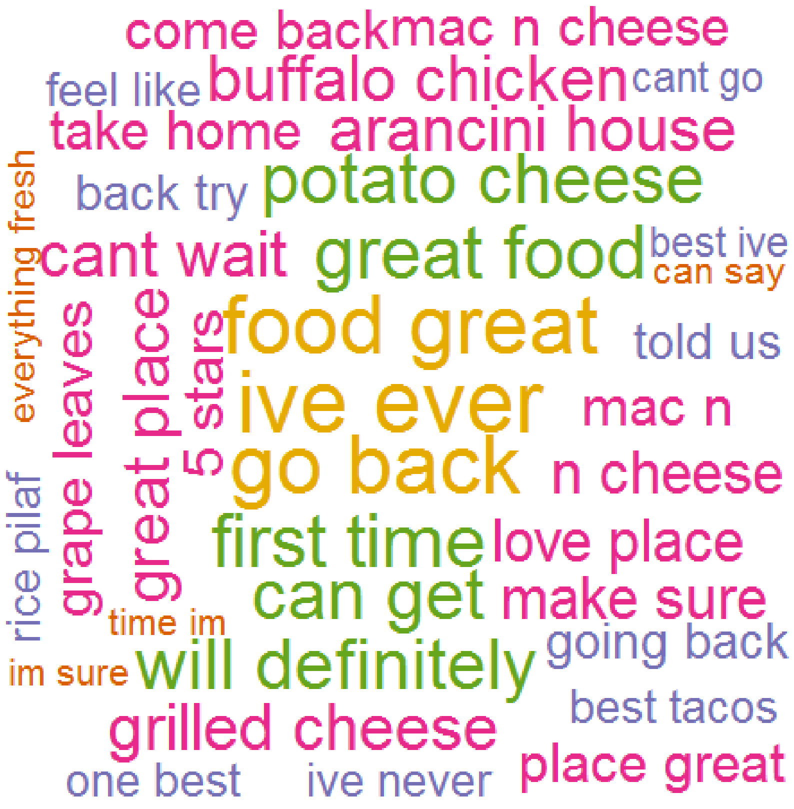
Word cloud for Reviews of Restaurants with an average rating of below 2.5

```
build_word_cloud("higher")
```

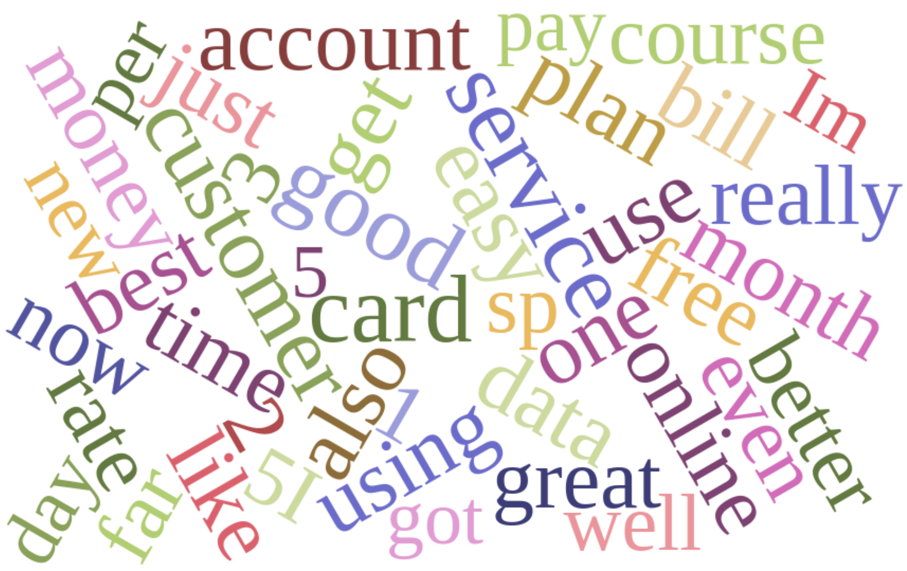


Word cloud for Reviews of Restaurants with an average rating of above 4.5

```
build_word_cloud("higher")
```



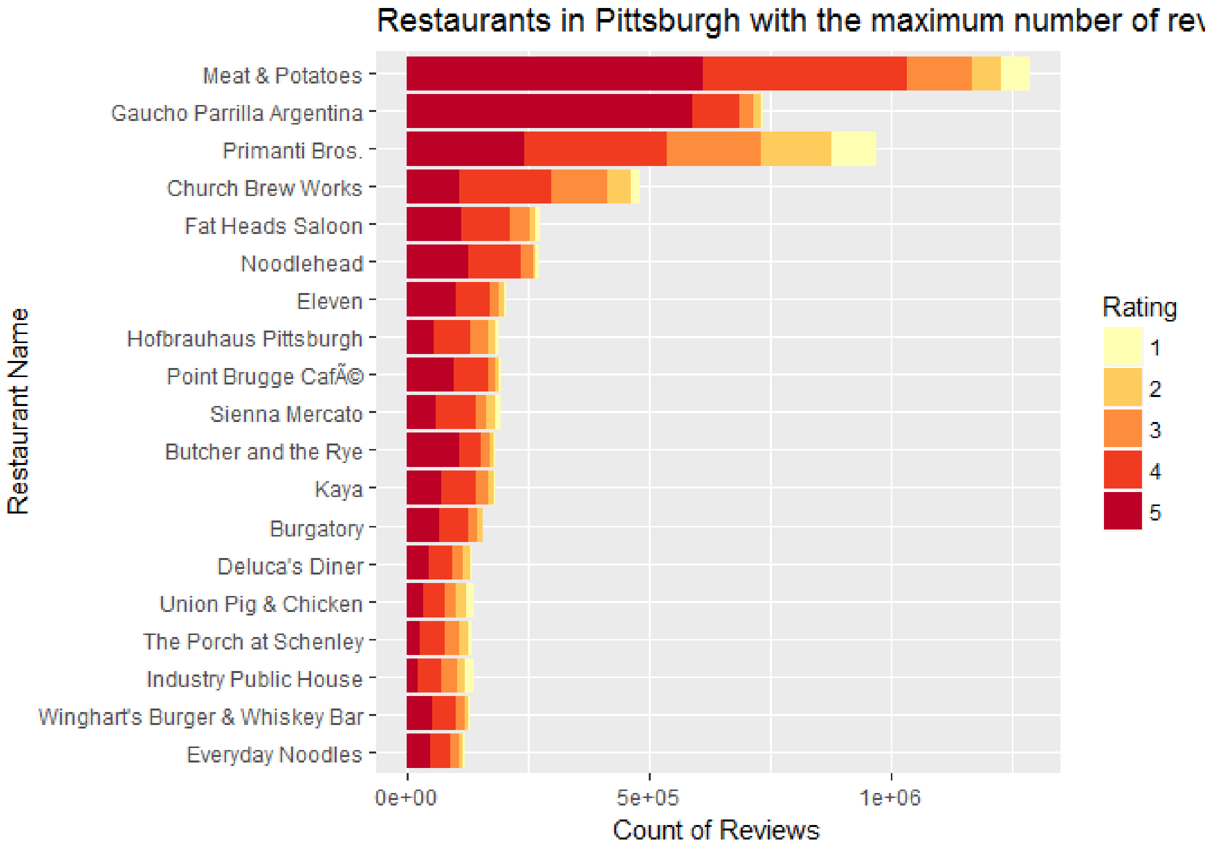
Word cloud for Reviews on the Seedly dataset. The dataset is small and hence we made the cloud for all the customer reviews on Seedly, not just the restaurant reviews.



The word clouds indicate with great clarity the difference between reviews for restaurants with low ratings versus those with higher ratings. We can see that phrases like ‘will never’, ‘go back’, ‘one star’, ‘waste time’, ‘2 stars’, ‘dont think’ etc are used frequently for lower-rated restaurants. Whereas phrases like ‘really good’, ‘great food’, ‘will definitely’, ‘great place’, ‘best’, ‘can’t wait’ etc are used with the highest frequency in reviews for highly rated restaurants.

A Stacked Bar Chart of 20 Restaurants in Pittsburgh with the highest number of reviews and a segmentation by Rating.

```
restaurants_maximum_ratings <- pa_restaurants %>%
arrange(desc(review_count)) %>% head(20) %>% select(business_id, name,
review_count, Average_Rating=stars, neighborhoods)
restaurants_maximum_ratings_data <- merge(restaurants_maximum_ratings,
pa_reviews, by="business_id") %>% select(business_id, name, review_count,
Average_Rating, stars, neighborhoods)
ggplot(data = restaurants_maximum_ratings_data, aes(x = reorder(name,
review_count), y = review_count, fill=as.factor(stars))) + coord_flip() +
  geom_bar(stat="identity") + labs(title = "Restaurants in Pittsburgh with
the maximum number of reviews",
                                    y = "Count of Reviews", x = "Restaurant
Name", fill = "Rating") + scale_fill_brewer(palette = "YlOrRd")
```



```
restaurants_maximum_ratings_percent <- restaurants_maximum_ratings_data %>%
select(-business_id) %>% group_by(name, review_count, stars) %>%
  summarise(count=n()) %>% spread(stars,count) %>%
arrange(desc(review_count)) %>%
mutate(Star5_Percent=round(`5`/review_count*100,0),
Star1_Percent=round(`1`/review_count*100,0))
DT::datatable(restaurants_maximum_ratings_percent)
```

S.No	Restaurant name	Review count	1 star	2 star	3 star	4 star	5 star	Star5 %	Star1 %
1	Meat & Potatoes	1175	50	52	114	357	523	45	4
2	Primanti Bros.	927	80	124	167	242	223	24	9
3	Gaucho Parrilla Argentina	893	7	16	32	108	663	74	1
4	Church Brew Works	739	24	69	154	259	146	20	3
5	Fat Heads Saloon	547	18	23	71	191	202	37	3
6	Noodlehead	543	13	9	48	199	234	43	2
7	Eleven	464	7	29	40	149	219	47	2
8	Hofbrauhaus Pittsburgh	463	17	26	84	162	121	26	4
9	Primanti Bros.	463	41	70	85	149	80	17	9
10	Point Brugge Cafe	462	5	18	32	155	210	45	1

According to this analysis, it becomes clear that only the number of reviews does not give us sufficient and reliable information. As we can see, that although Meat & Potatoes has the highest number of reviews, Gaucho Parrilla Argentina actually enjoys the highest percentage of 5-star reviews.

Revelations and Insights from the Analysis -

1. Since the distribution of the ratings is considerably skewed, we could say that people seem to review things they like. In general, people seem to be more likely to write a review for a positive experience than a negative one.
2. Also, it appears that there is some evidence for the hypothesis that there is a correlation between Price Range and Rating. Especially, the restaurants in the Price Bracket of 4 (highest) have a higher average Rating. Las Vegas is a foodie's heaven with highest number of restaurants that have an average Rating of 5.
3. The restaurants with maximum number of 5-star ratings mostly serve Italian, Mexican and American cuisines.
4. The word clouds indicate with great clarity the difference between reviews for restaurants with low ratings versus those with higher ratings. We can see that phrases like 'will never', 'go back', 'one star', 'waste time', '2 stars', 'dont think' etc are used frequently for lower-rated restaurants. Whereas phrases like 'really good', 'great food', 'will definitely', 'great place', 'best', 'can't wait' etc are used with the highest frequency in reviews for highly rated restaurants.
5. If you are ever in Pittsburgh, be sure to check out Gaucho Parrilla Argentina which boasts the highest percentage of 5-star reviews (amongst it's total number of reviews) as compared to other restaurants.
6. Phoenix is the city of paradise for Vegetarians! And the best restaurant is Postino Arcadia.

**Note:* This Analysis is strictly confined to based on the reviews on the Yelp dataset. So, the results are not meant to be generalized for the whole population.