EDA of Zomato Bangalore estaurants

Saurabh Parkar

October 4, 2019

About the data

Bangalore city is the largest IT hub of India and so most of the people here depend on restaurant foods. The dataset provides general information like Location, Ratings, Cuisine, Cost etc. of more than 50,000 restaurants in Bangalore which can help in analyzing the various factors that influence the popularity of the restaurants.

Source- Kaggle (https://www.kaggle.com/himanshupoddar/zomato-bangalore-restaurants)

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readr)
df<-read_csv('zomato.csv')
```

```
## Parsed with column specification:
## cols(
##   url = col_character(),
##   address = col_character(),
##   name = col_character(),
##   online_order = col_character(),
##   book_table = col_character(),
##   rate = col_character(),
##   votes = col_double(),
##   phone = col_character(),
##   location = col_character(),
##   rest_type = col_character(),
##   dish_liked = col_character(),
##   cuisines = col_character(),
##   `approx_cost(for two people)` = col_number(),
##   reviews_list = col_character(),
##   menu_item = col_character(),
##   `listed_in(type)` = col_character(),
##   `listed_in(city)` = col_character()
## )
```

Dropping the unwanted columns

We wont be needing columns like url, phone number and complete address of the restaurant in our analysis and so we will drop them

```
df <- select(df, -c(url,reviews_list,phone,address, menu_item))
```

Printing the first 10 observations

```
head(df,10)
```

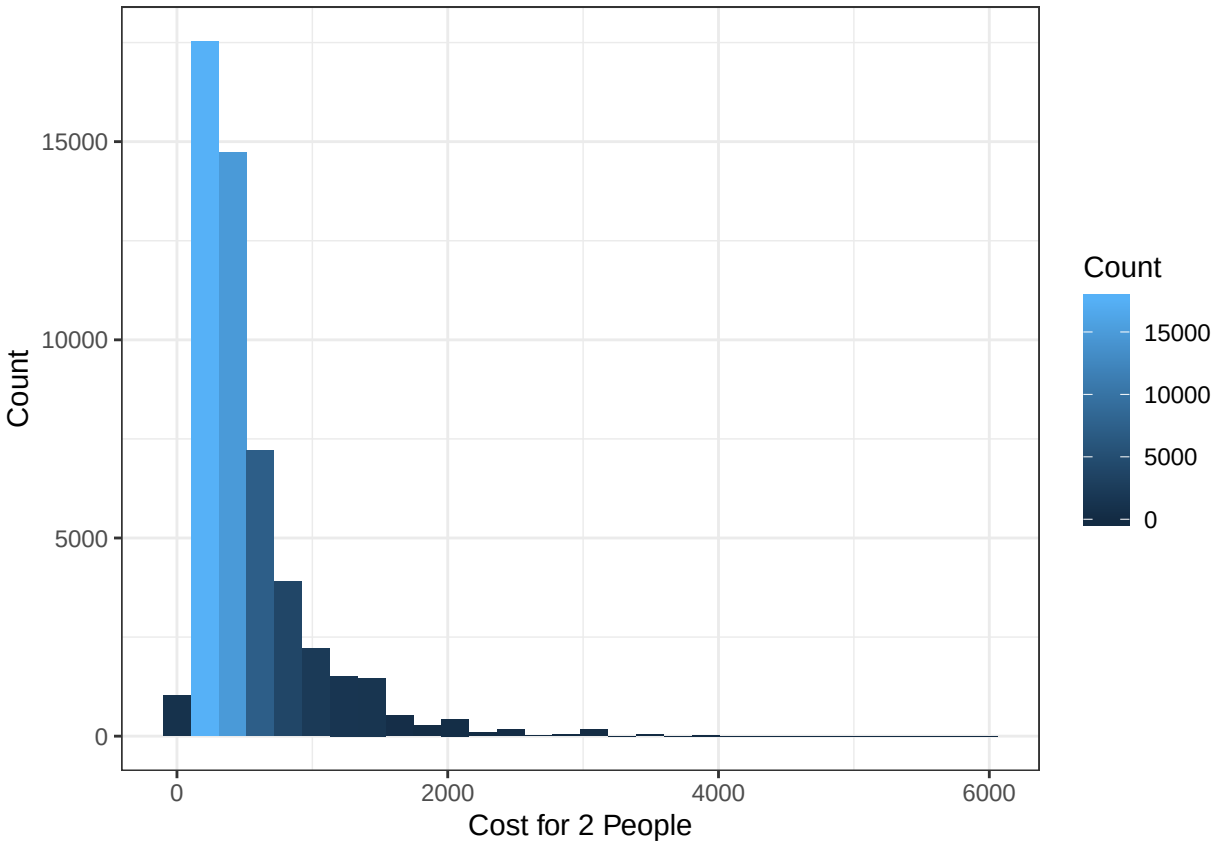
```
## # A tibble: 10 x 12
##    name  online_order book_table rate  votes location rest_type dish_liked
##    <chr> <chr>        <chr>      <chr> <dbl> <chr>    <chr>     <chr>
##  1 Jalsa Yes          Yes        4.1/5   775 Banasha~ Casual D~ Pasta, Lu~
##  2 Spic~ Yes          No         4.1/5   787 Banasha~ Casual D~ Momos, Lu~
##  3 San ~ Yes          No         3.8/5   918 Banasha~ Cafe, Ca~ Churros, ~
##  4 Addh~ No           No         3.7/5    88 Banasha~ Quick Bi~ Masala Do~
##  5 Gran~ No           No         3.8/5   166 Basavan~ Casual D~ Panipuri,~
##  6 Time~ Yes          No         3.8/5   286 Basavan~ Casual D~ Onion Rin~
##  7 Rose~ No           No         3.6/5     8 Mysore ~ Casual D~ <NA>
##  8 Ones~ Yes          Yes        4.6/5  2556 Banasha~ Casual D~ Farmhouse~
##  9 Pent~ Yes          No         4.0/5   324 Banasha~ Cafe      Pizza, Mo~
## 10 Smac~ Yes          No         4.2/5   504 Banasha~ Cafe      Waffles, ~
## # ... with 4 more variables: cuisines <chr>, `approx_cost(for two
## #   people)` <dbl>, `listed_in(type)` <chr>, `listed_in(city)` <chr>
```

Getting Started with Average cost for 2 people

```
ggplot(data = df) + geom_histogram(aes(x=`approx_cost(for two people)`, fill=..count..)) + theme_bw() +
        x ="Cost for 2 People", y = "Count", fill='Count')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 346 rows containing non-finite values (stat_bin).
```

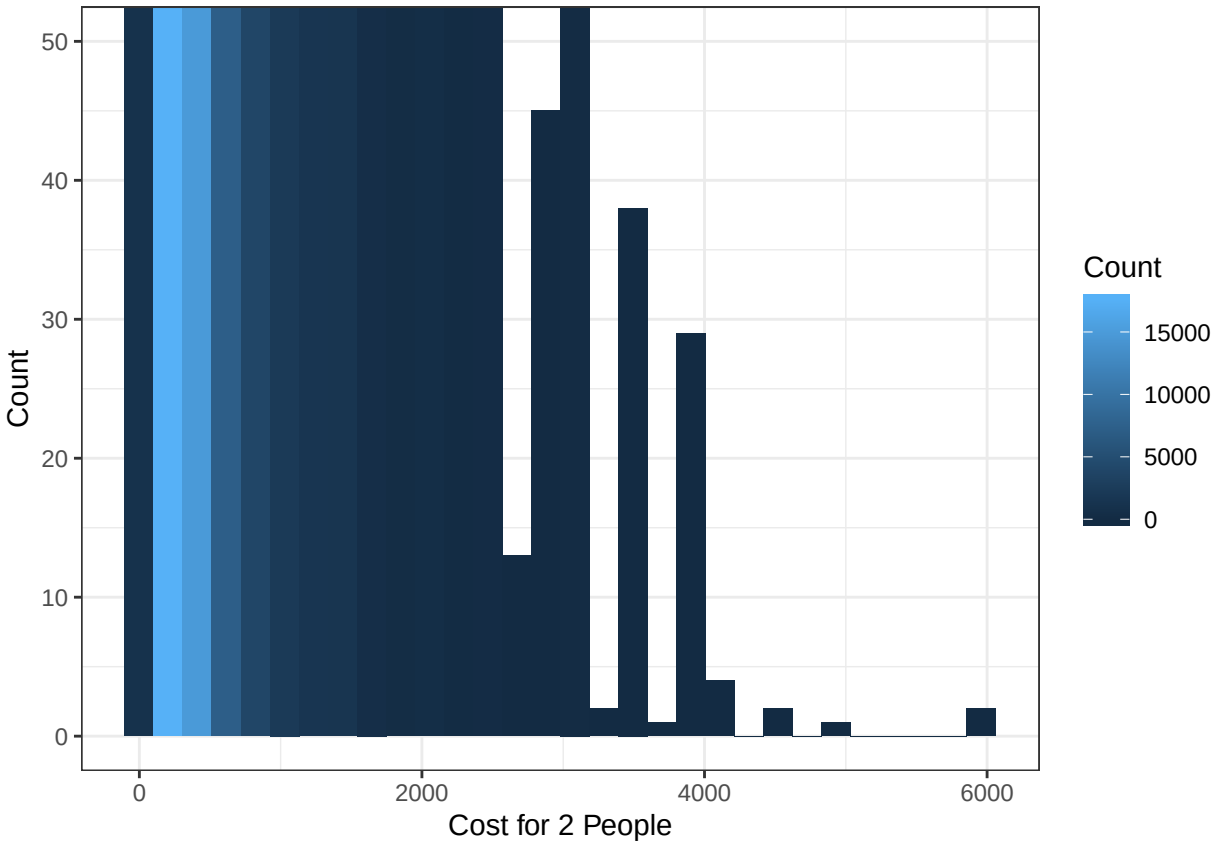


It can be seen that the histogram is right skewed and most of the restaurants fall below ???1000 mark. As they get expensive, their count reduces exponentially.

```
ggplot(data = df) + geom_histogram(aes(x=`approx_cost(for two people)`, fill=..count..)) + theme_bw() +
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 346 rows containing non-finite values (stat_bin).
```



After zooming in, we can also see few restaurants that cost over ???4000 (Very Expensive). And 2-3 restaurants that cost ???6000 for 2 people.

Fact: In India, restaurants that cost ???1000 and above (for 2 people) are considered moderately expensive and those above ???2000 are considered expensive. For bars and buffets, ???2000 can be considered normal and those above it as expensive.

15 Most popular cuisines in Bangalore

Handling untidy data and visualising popular cuisines

Many of the columns like rest_type and cuisines have multiple values in a single column

```
df[c(1,3,8,15),c(4:7,9)]
```

```
## # A tibble: 4 x 5
##   rate  votes location     rest_type           cuisines
##   <chr> <dbl> <chr>        <chr>               <chr>
## 1 4.1/5   775 Banashankari Casual Dining       North Indian, Mughlai, Chin~
## 2 3.8/5   918 Banashankari Cafe, Casual Dining Cafe, Mexican, Italian
## 3 4.6/5  2556 Banashankari Casual Dining, Cafe Pizza, Cafe, Italian
## 4 3.8/5   918 Banashankari Cafe, Casual Dining Cafe, Mexican, Italian
```

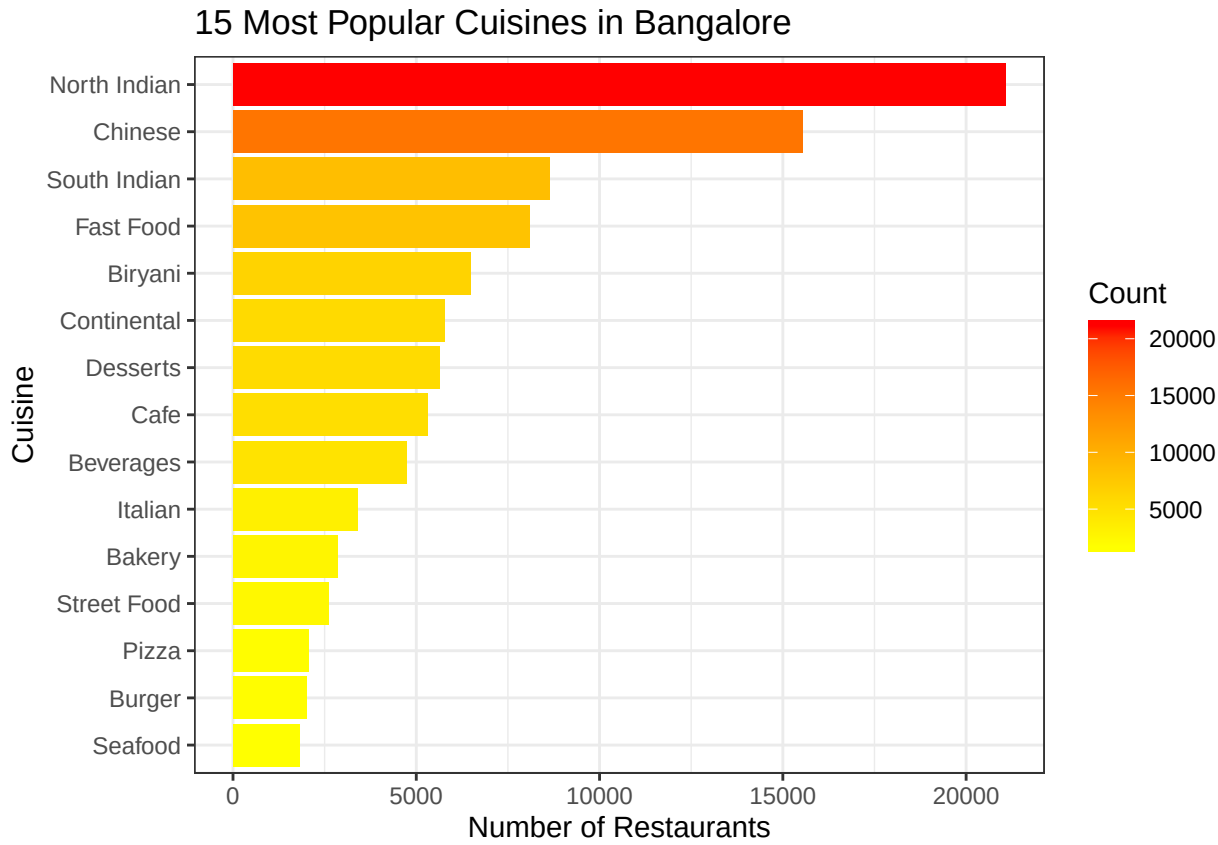
In order to obtain useful information from them, the values in each column need to be separated first and then summarized by taking taking their sum.

```
lc = list()

for (i in df$cuisines){
  a <- unlist(strsplit(i, ", "))
  for (j in a){
    if (j %in% names(lc)){
      lc[[j]] <- lc[[j]] + 1
    } else {
      lc[[j]] <- 1
    }
  }
}
```

Arranging the results in descending order of their counts and plotting a bar chart of first 15 rows helps us visualize the 15 most popular cuisines in Bangalore.

```
countc <- as.numeric(paste(unlist(lc)))
lc <- lc[-51]
df_lc <- data.frame("Type" = names(lc), "Tot" = countc)
df_lco <- arrange(df_lc, desc(Tot))
df_lco_samp <- df_lco[1:15,]
ggplot(df_lco_samp) + geom_col(mapping=aes(x=reorder(Type,Tot), y= Tot, fill=Tot )) + coord_flip() + sca
        x ="Cuisine", y = "Number of Restaurants",fill = "Count")
```



Interesting results

This plot is particularly interesting because Bangalore is considered to be a "South Indian" city. But ironically, the most popular cuisines in Bangalore are North Indian and Chinese and then followed by South Indian. Not only these 2 cuisines are popular but also beat the other cuisines by a wide mark. Rest of the cuisines have gradually decreasing popularity.

The restaurants that offer these 2 cuisines are more in number and beat other cuisines by a wide mark. Rest of the cuisines have a gradually decreasing popularity.

Fact: North Indian Cuisine consists of famous dishes like butter chicken, naan, biryani and lots of other varieties of sabzis, which are not very expensive and extremely popular all over India. These also happen to be the typical go-to lunch/dinner order and hence must be responsible for their high popularity.

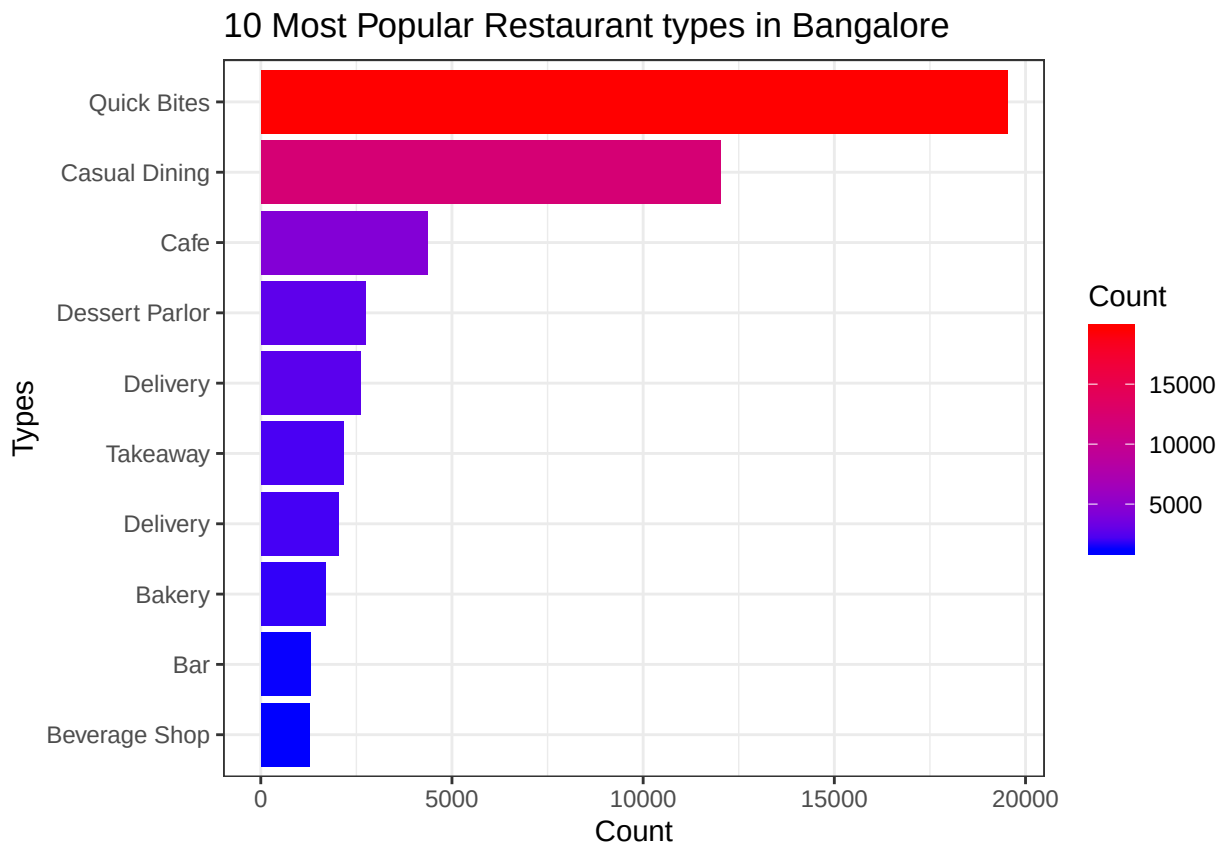
10 Most popular restaurant types in Bangalore

Doing the same process as above-

```
l = list()

for (i in df$rest_type){
  a <- unlist(strsplit(i, ","))
  for (j in a){
    if (j %in% names(l)){
      l[[j]] <- l[[j]] + 1
    } else {
      l[[j]] <- 1
    }
  }
}
l <- l[-29]
lp<-l

countp <- as.numeric(paste(unlist(lp)))
df_lp <- data.frame("Type" = names(lp), "Tot" = countp)
df_lpo <- arrange(df_lp, desc(Tot))
df_lpo_samp <- df_lpo[1:10,]
ggplot(df_lpo_samp) + geom_col(mapping=aes(x=reorder(Type,Tot), y= Tot, fill=Tot )) + coord_flip() + sc
        x ="Types", y = "Count",fill = "Count")
```



Interesting Results

Quick bites and fast foods are the most popular type of restaurants in Bangalore. Bangalore is the largest IT hub of India and hence lot of people in the area depend on restaurants for food. Fast foods are relatively cheaper and wholesome and hence might be the most popular choice of the people. These type of restaurants usually offer free deliveries too and hence becomes convenient for most people. Casual dining are another popular type but still the count of restaurants are almost half of the quick bite types. Other types of restaurants in Bangalore like cafes and dessert parlor are far too less as compared to other types.

Does Higher Prices mean Better Ratings?

The Rate column consists of character type values e.g. 4.2/5. In order to be able to compare the values we first need to remove the '/5' string and then convert the entire column into numeric type.

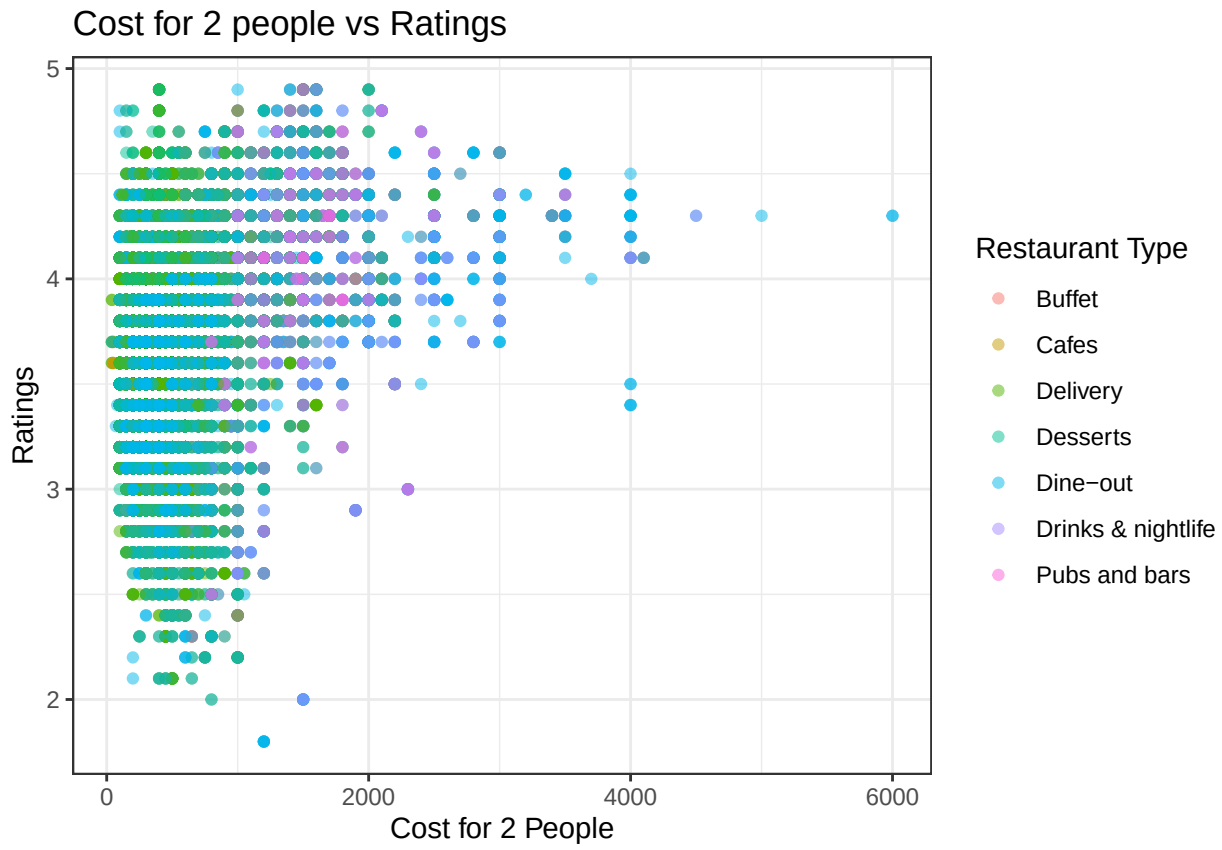
```
df_copy <- df
df_copy$rate <- sapply(strsplit(df_copy$rate,"/"), `[`, 1)
df_copy$rate <- as.numeric(df_copy$rate)
```

```
## Warning: NAs introduced by coercion
```

A scatter plot of Cost vs Ratings using color for Restaurant types

```
ggplot(data= df_copy) + geom_point(aes(x=`approx_cost(for two people)`, y=rate, color=`listed_in(type)`
        x ="Cost for 2 People", y = "Ratings", color = "Restaurant Type")
```

```
## Warning: Removed 10299 rows containing missing values (geom_point).
```



Interesting Results

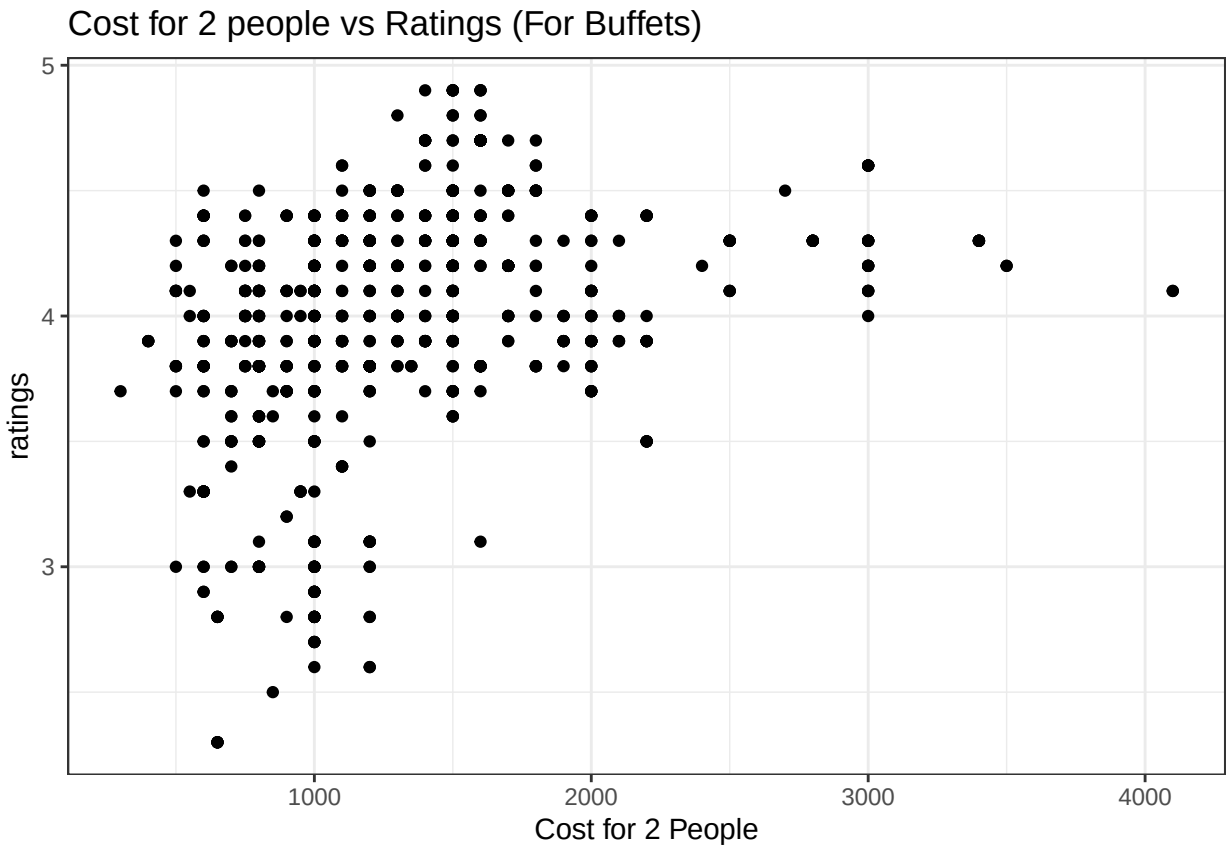
For cost uptoaround Rs.1000, the ratings seem to be uniformly distributed. After that, as the cost increases, the ratings increase too, meaning higher cost equals better experience..It is surprising to see the ratings of many of the cheaper restaurants (<1000), above the more expensive ones. (The number of votes may be responsible for this bias)

The expensive restaurants mostly include the Dine-out types (actual "Five Star" Restaurants) and the more expensive nightclubs. Delivery, dessert and other dine out restaurants fall under the cheaper range. Pubs and bars are moderately expensive.

Buffets are not clearly visible due to their small number.

```
ggplot(data= df_copy[df_copy$`listed_in(type)`=='Buffet',]) + geom_point(aes(x=`approx_cost(for two peo
        x ="Cost for 2 People", y = "ratings")
```

```
## Warning: Removed 34 rows containing missing values (geom_point).
```

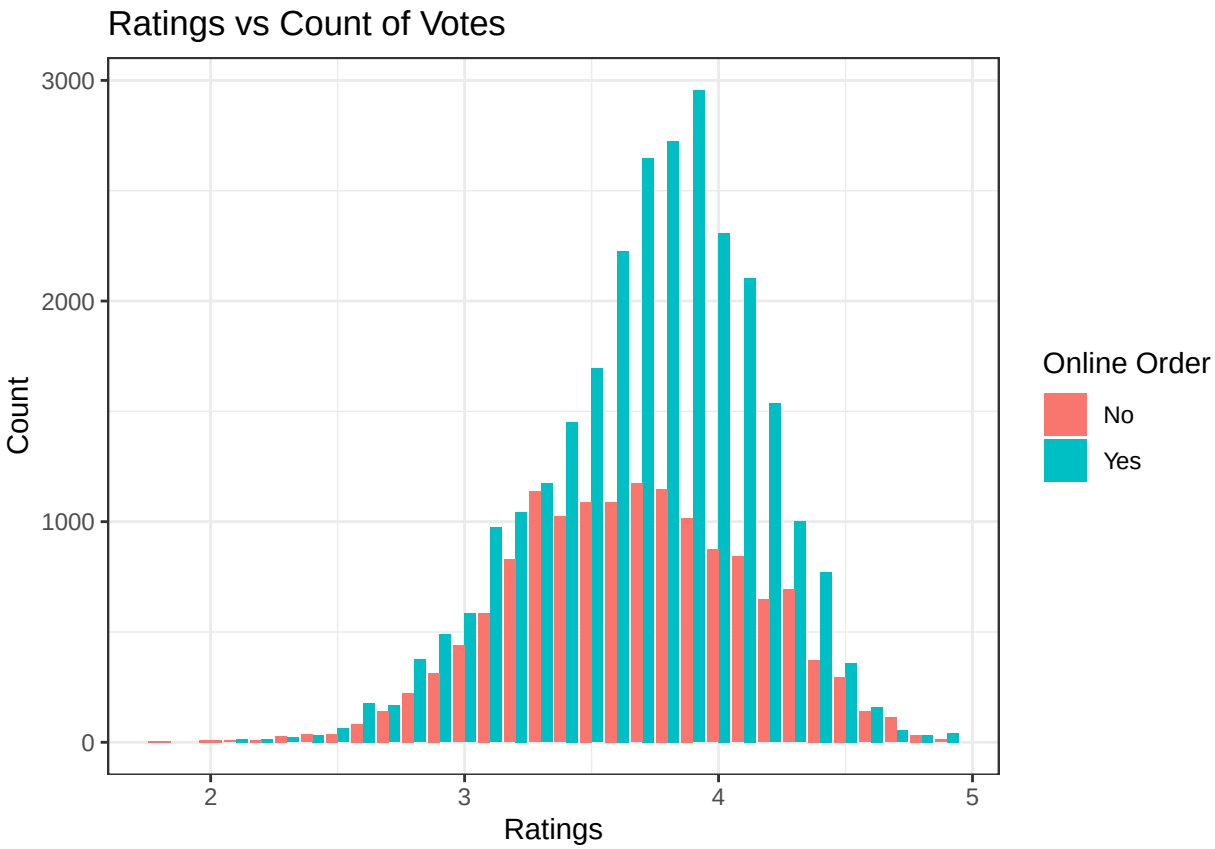


Since the data points for buffets were not clearly visible (due to their small number and overlapping), I plotted them separately and the results are quite surprising. The number of buffet restaurants with cost less than 1000 is lot more than I expected with decent ratings too.

Ratings and Count of Votes(Comparison with online orders)

```
ggplot(data = df_copy) + geom_bar(aes(x=rate, fill=online_order), position='dodge') + theme_bw() + labs
        x ="Ratings", y = "Count",fill = "Online Order")
```

```
## Warning: Removed 10052 rows containing non-finite values (stat_count).
```



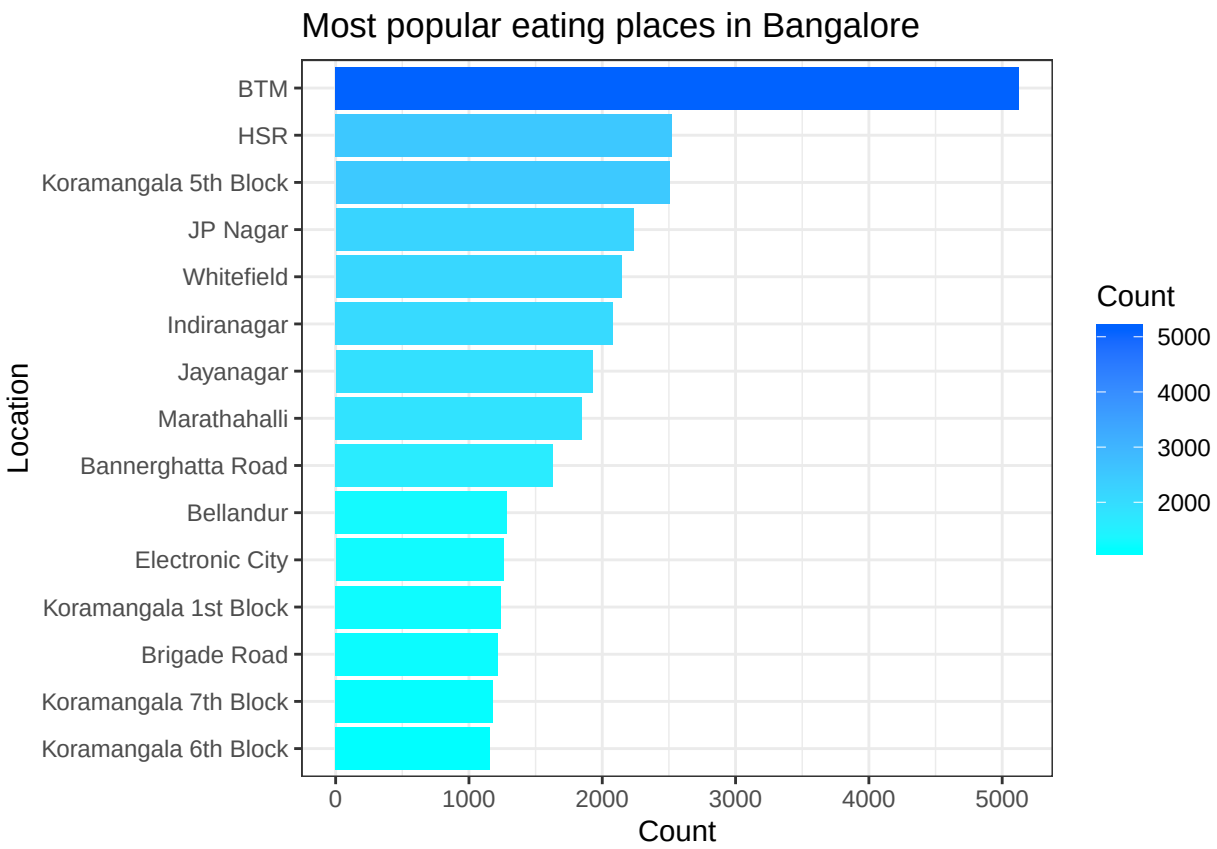
Interesting Results

The ratings that are very high and very low have comparatively lower vote counts and hence there is greater chance that the ratings are biased. More votes usually means greater popularity and most of the popular restaurants have ratings around 3.5 and 4.2, which is quite decent..Although both are similarly distributed, the count of votes for online orders is much higher than those of traditional orders. This can be attributed to the fact that the online ordering apps request users to rate the restaurants after every order resulting in higher number of votes.

Fact: People prefer to order online since various offers are available and delivery fee is nominal in India. Bangalore being an IT hub, most of the people there rely on restaurant foods and online ordering becomes a convenient option.

Most popular Locations

```
pop <- summarise(group_by(df_copy,location),count= n()) %>% arrange(desc(count))
pop15 <- pop[1:15,]
ggplot(pop15) + geom_col(mapping=aes(x=reorder(location,count), y= count, fill=count )) + coord_flip() +
        x ="Location", y = "Count",fill = "Count")
```



15 Largest Food chains in Bangalore

```
pop_rest <- summarise(group_by(df_copy,name),count= n()) %>% arrange(desc(count))
pop_rest[1:15,]
```

```
## # A tibble: 15 x 2
##    name              count
##    <chr>             <int>
##  1 Cafe Coffee Day      96
##  2 Onesta               85
##  3 Just Bake            73
##  4 Empire Restaurant    71
##  5 Five Star Chicken    70
##  6 Kanti Sweets         68
##  7 Petoo                66
##  8 Polar Bear           65
##  9 Baskin Robbins       64
## 10 Chef Baker's         62
## 11 Pizza Hut            62
## 12 Beijing Bites        60
## 13 Domino's Pizza       60
## 14 KFC                  60
## 15 Subway               60
```