UCS548 Data Science Foundation
Assignment – 8.1

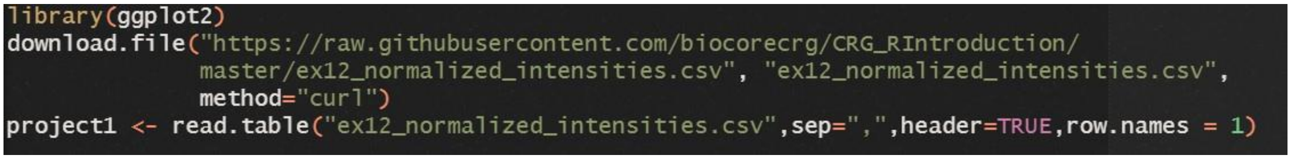
Name: Parth Vohra
Roll No:102016044
Sub Group: 3CS10

Q1.Usethe followingdataset(Project1)toplot thegraphsusingggplot2

`master/ex12_normalized_intensities.csv`

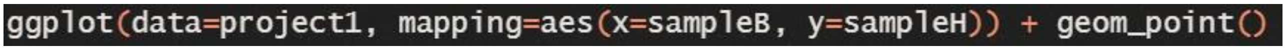
About this file:

- It is comma separated (csv format).
- The first row is the header.
- Take the row names from the first column.

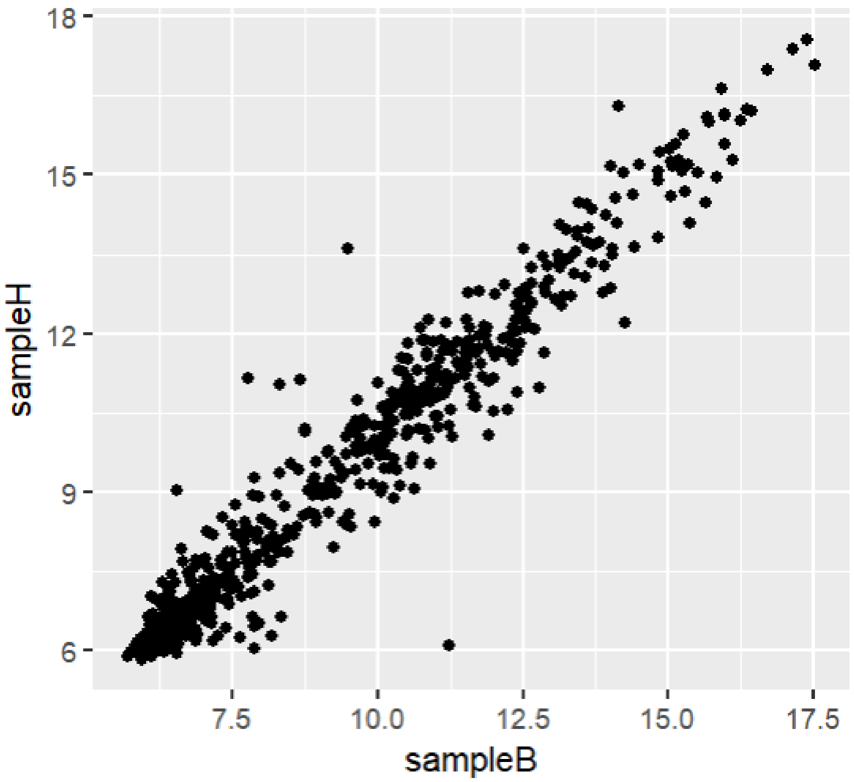


```
library(ggplot2)
download.file("https://raw.githubusercontent.com/biocorecrg/CRG_RIntroduction/
              master/ex12_normalized_intensities.csv", "ex12_normalized_intensities.csv",
              method="curl")
project1 <- read.table("ex12_normalized_intensities.csv",sep=",",header=TRUE,row.names = 1)
```

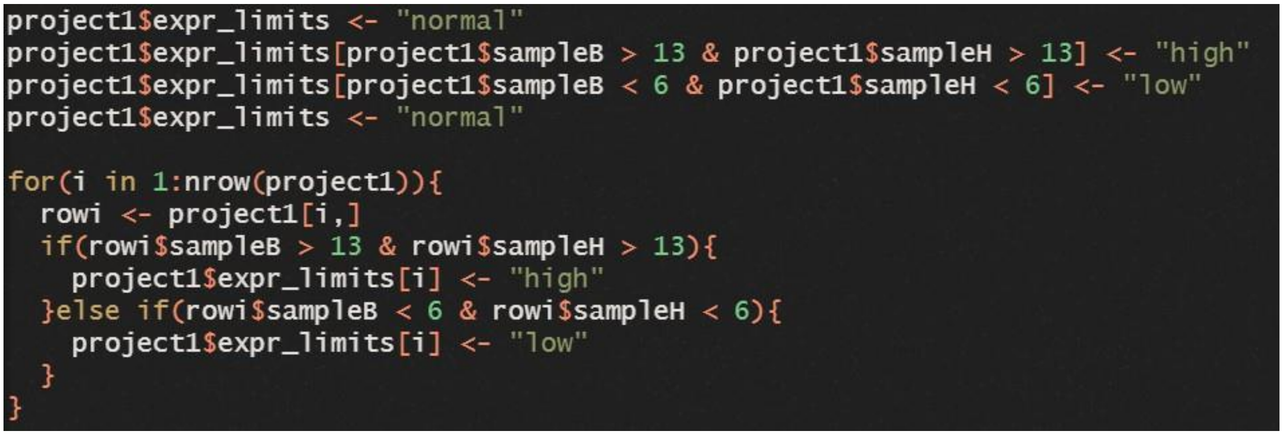
1. Create a simple scatter plot representing gene expression of “sampleB” on the x-axis and “sampleH” on the y-axis.



```
ggplot(data=project1, mapping=aes(x=sampleB, y=sampleH)) + geom_point()
```



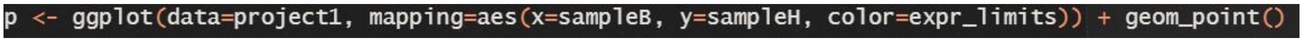
2. Add a column to the data frame “project1” (call this column “expr_limits”), that will be filled the following way:
   i) if the expression of a gene is > 13 in both sampleB and sampleH, set to the value in “expr_limits” to “high”
   ii) if the expression of a gene is < 6 in both sampleB and sampleH, set it to “low”
   iii) if different, set it to “normal”.



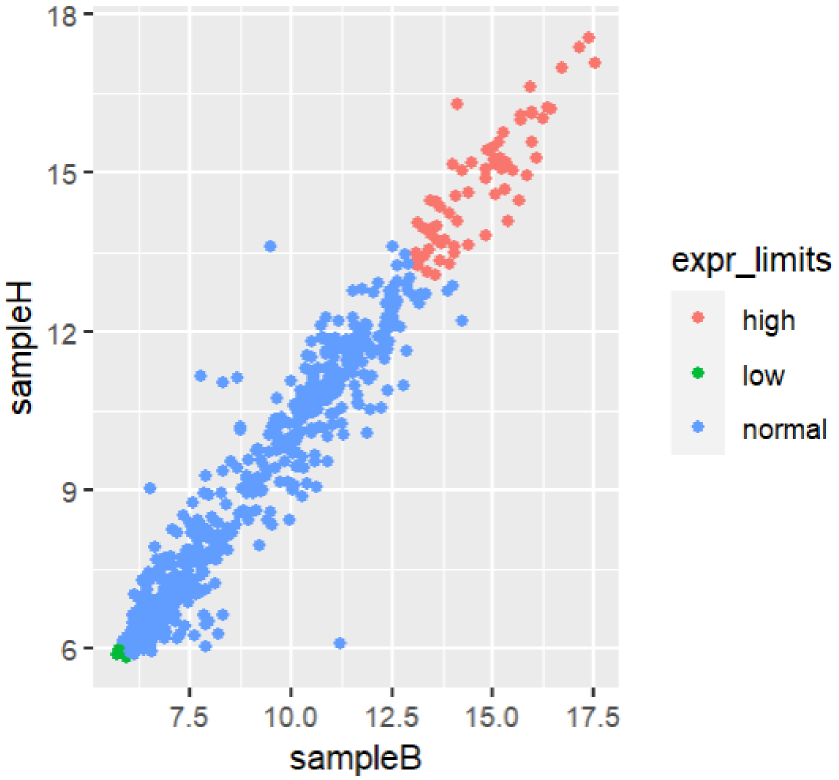
```
project1$expr_limits <- "normal"
project1$expr_limits[project1$sampleB > 13 & project1$sampleH > 13] <- "high"
project1$expr_limits[project1$sampleB < 6 & project1$sampleH < 6] <- "low"
project1$expr_limits <- "normal"

for(i in 1:nrow(project1)){
  rowi <- project1[i,]
  if(rowi$sampleB > 13 & rowi$sampleH > 13){
    project1$expr_limits[i] <- "high"
  }else if(rowi$sampleB < 6 & rowi$sampleH < 6){
    project1$expr_limits[i] <- "low"
  }
}
```

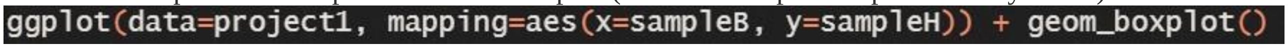
3. Color the points of the scatter plot according to the newly created column “expr_limits”. Save that plot in the object “p”.



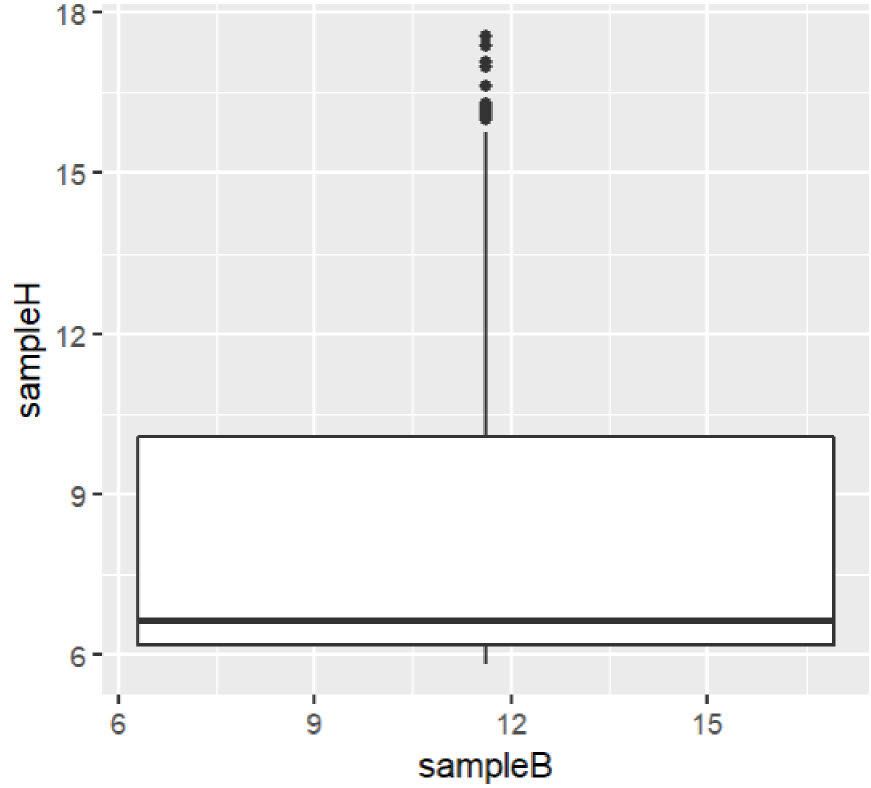
```
p <- ggplot(data=project1, mapping=aes(x=sampleB, y=sampleH, color=expr_limits)) + geom_point()
```



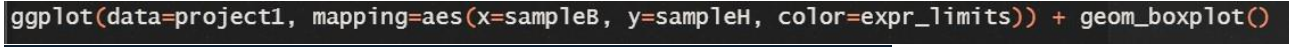
4. Produce a boxplot of the expression of all samples (i.e. each sample is represented by a box).



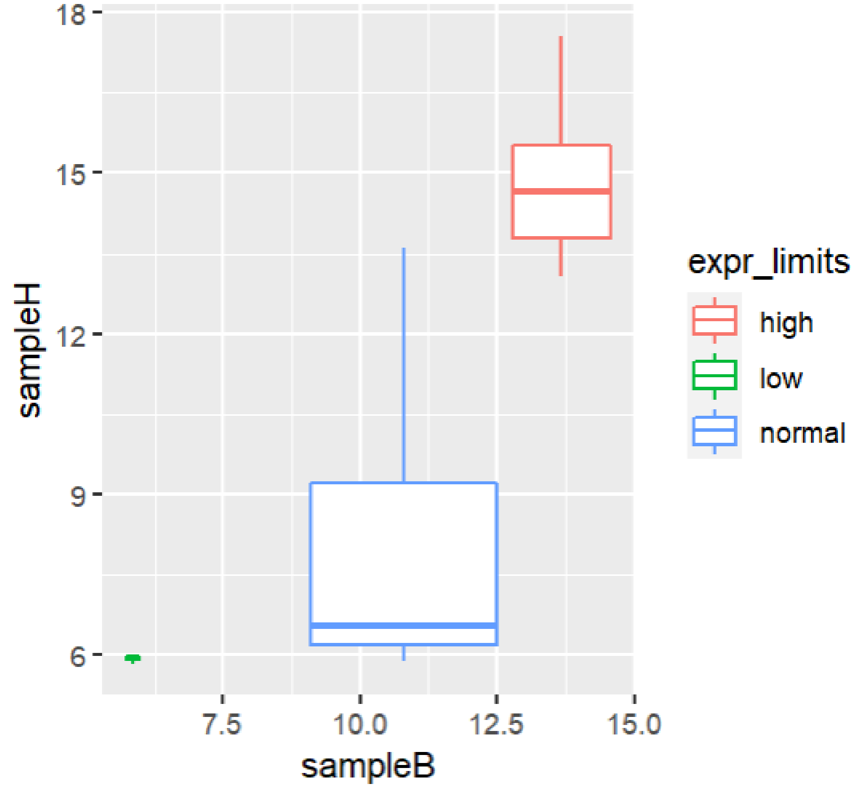
```
ggplot(data=project1, mapping=aes(x=sampleB, y=sampleH)) + geom_boxplot()
```



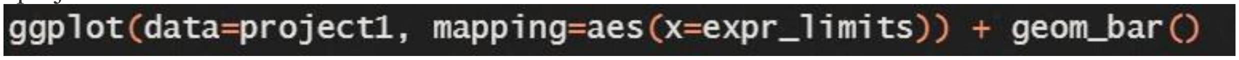
5. Modify the previous boxplot so as to obtain 3 “sub-boxplots” per sample, each representing the expression of either “low”, “normal” or “high” genes.



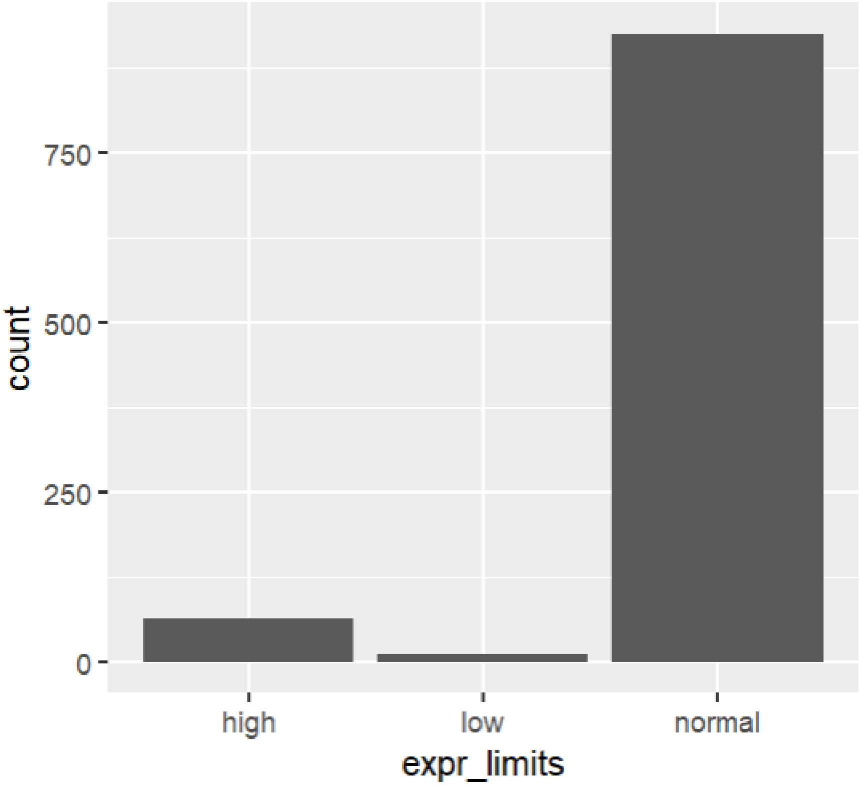
```
ggplot(data=project1, mapping=aes(x=sampleB, y=sampleH, color=expr_limits)) + geom_boxplot()
```



6. Produce a bar plot of how many low/normal/high genes are in the column “expr_limits” of “project1”.

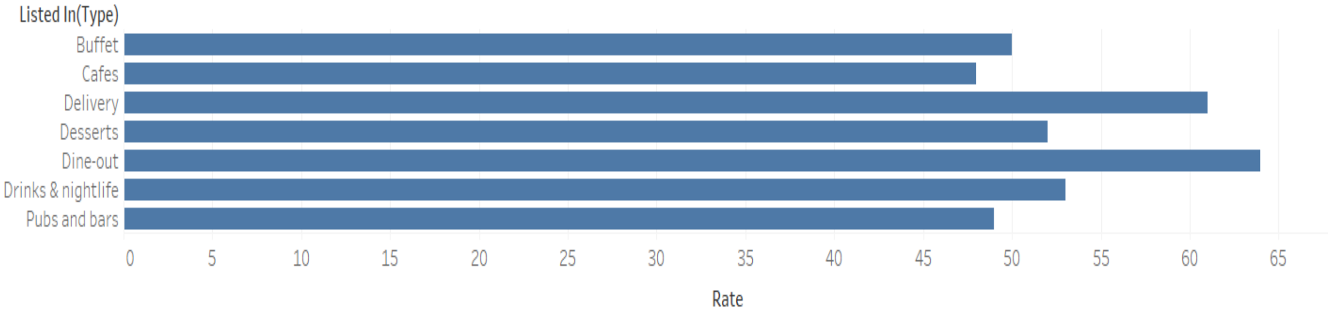


```
ggplot(data=project1, mapping=aes(x=expr_limits)) + geom_bar()
```

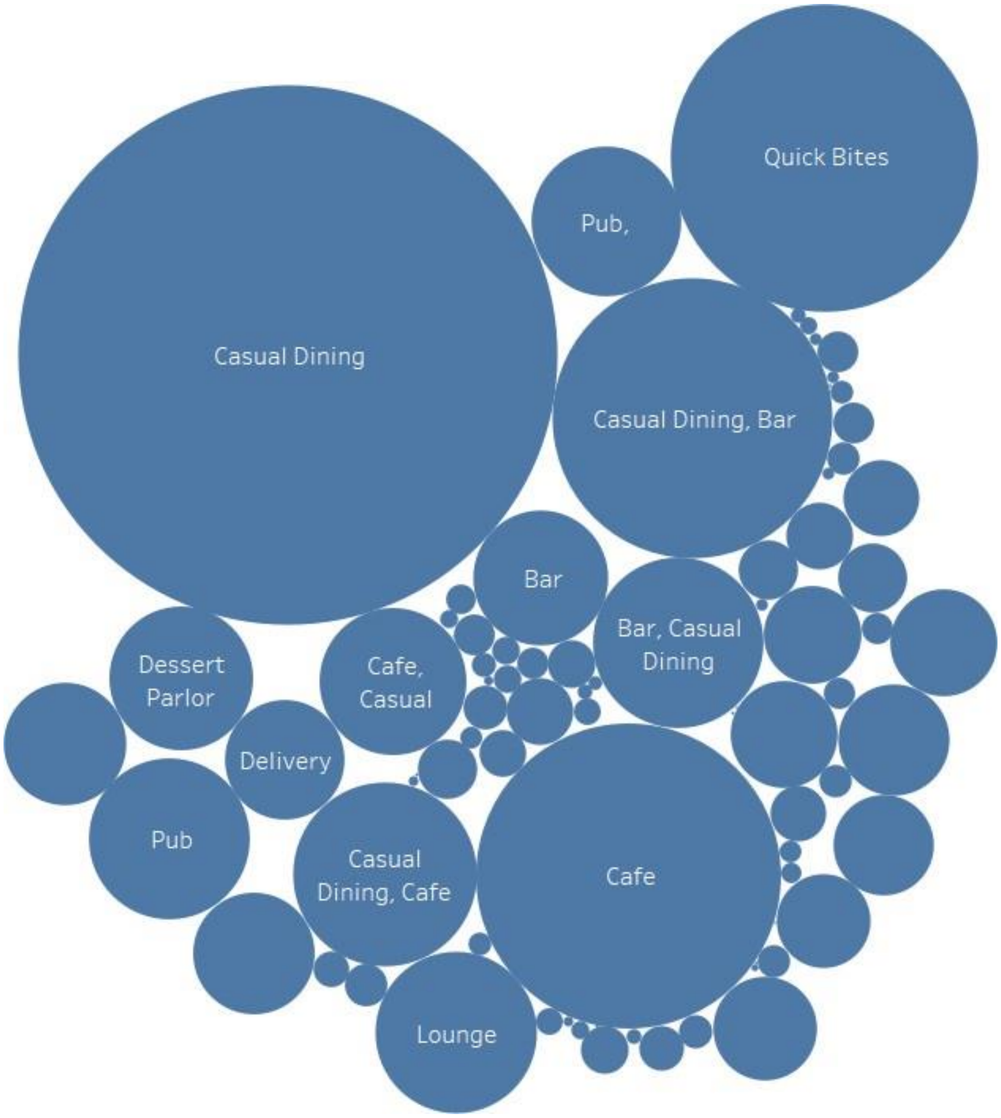


Q2. Use following dataset of Zomato for exploratory data analysis using Tableau: https://www.kaggle.com/datasets/pranavuikey/zomato-eda

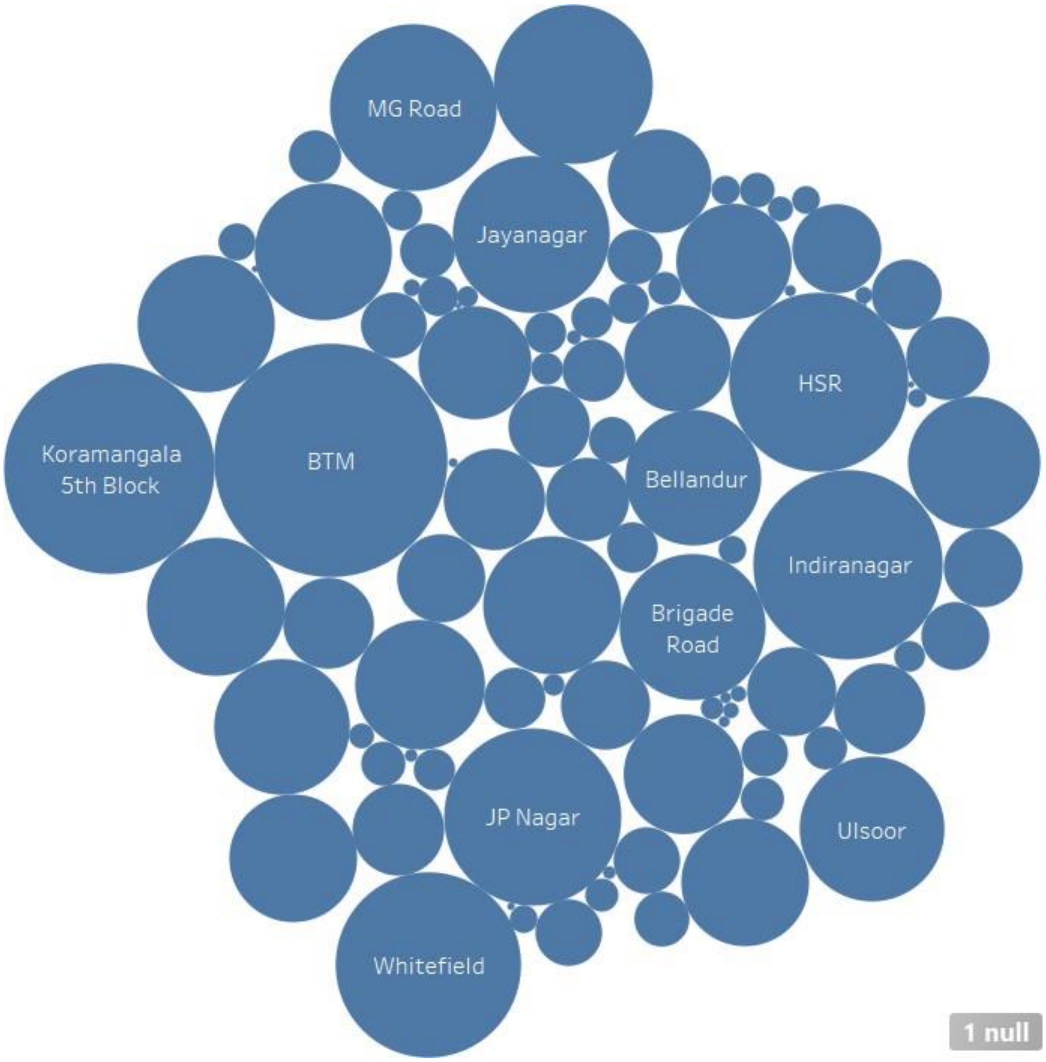
1. Find the highest rating (use rate attribute) for the type of service using listed In(Type) information through bar graph.



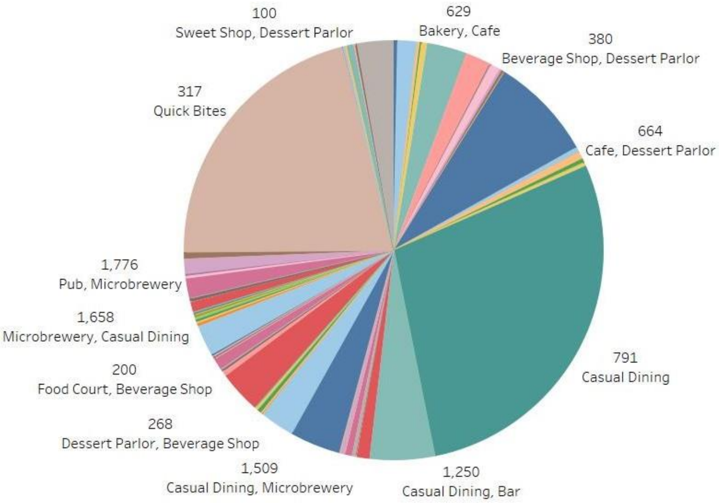
2. With the help of packed bubbles visualization
    a. Identify the types of restaurant getting highest votes.



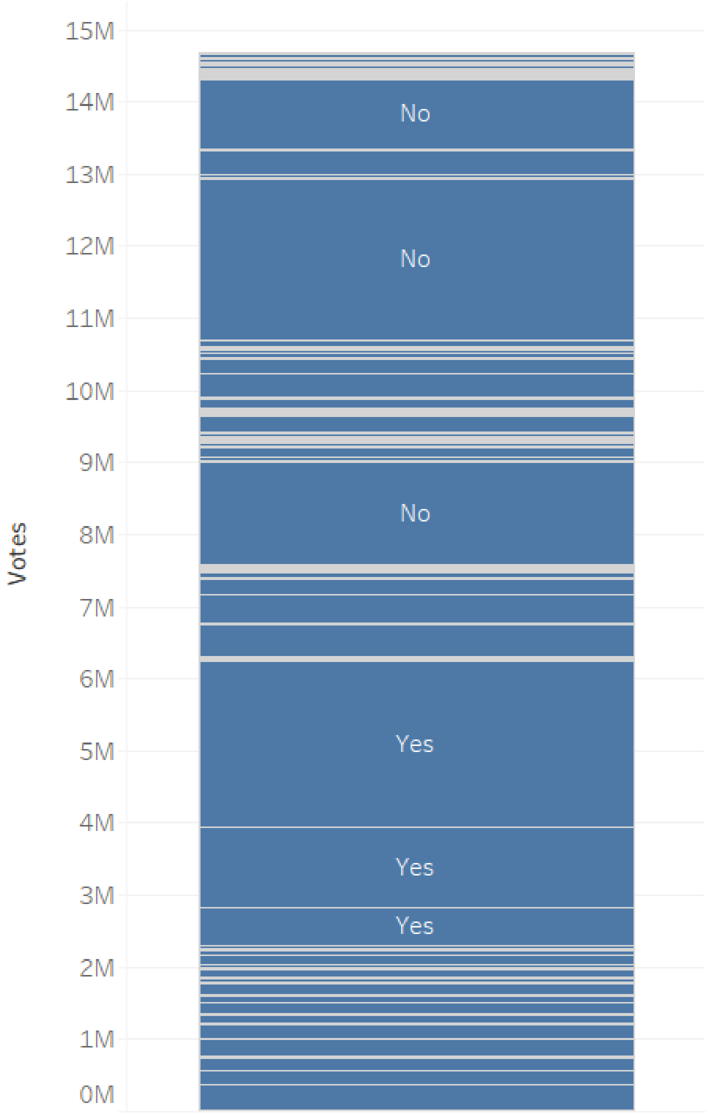
b. Find the top three locations having highest approx. cost per two people.



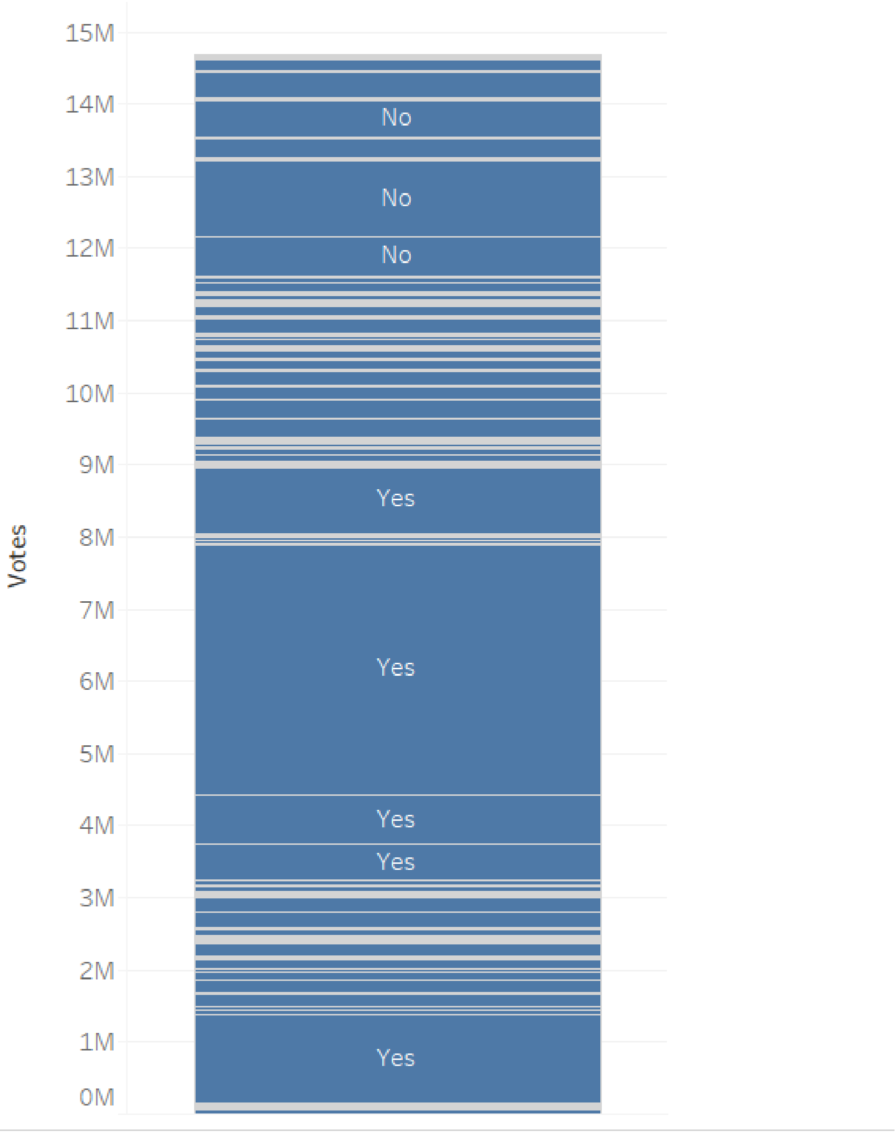
3. Find the costly type of restaurant (consider approx. cost per two people) with the help of pie chart.



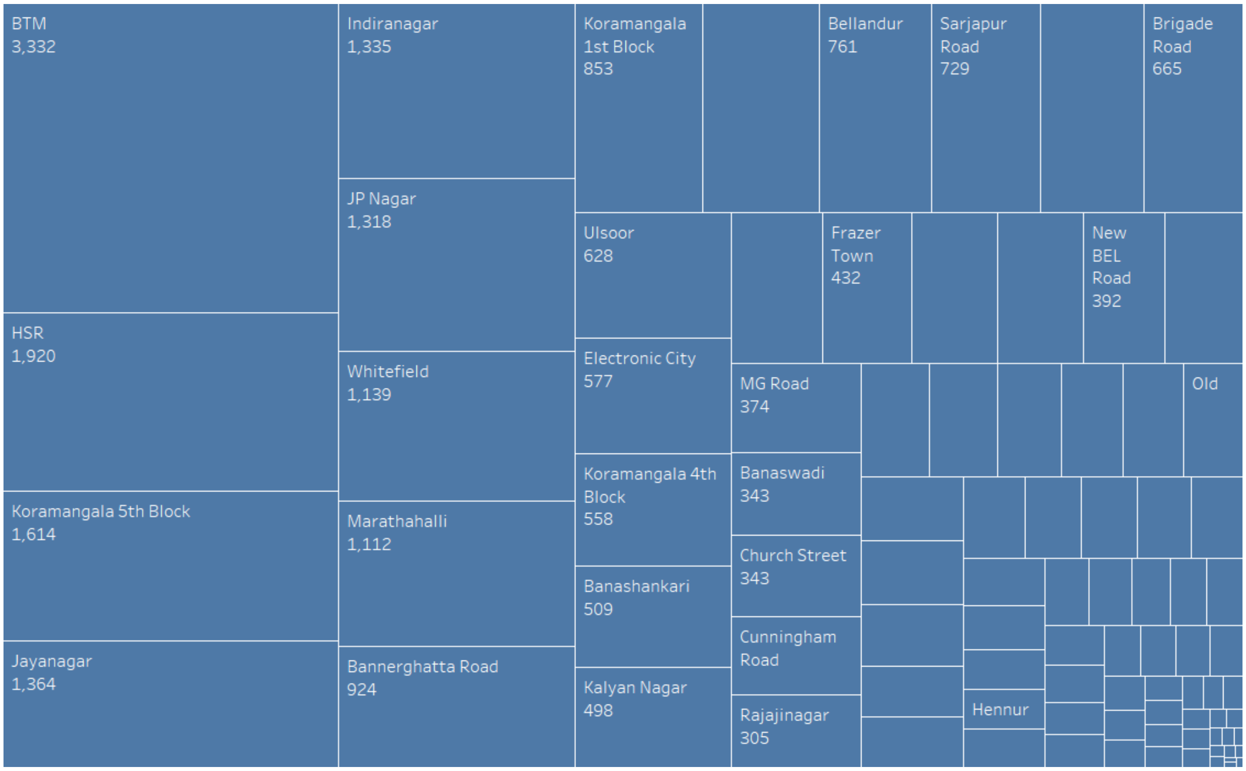
4. Utilize stacked bars to check type of restaurants with highest votes
    a. Providing table booking facility or not.



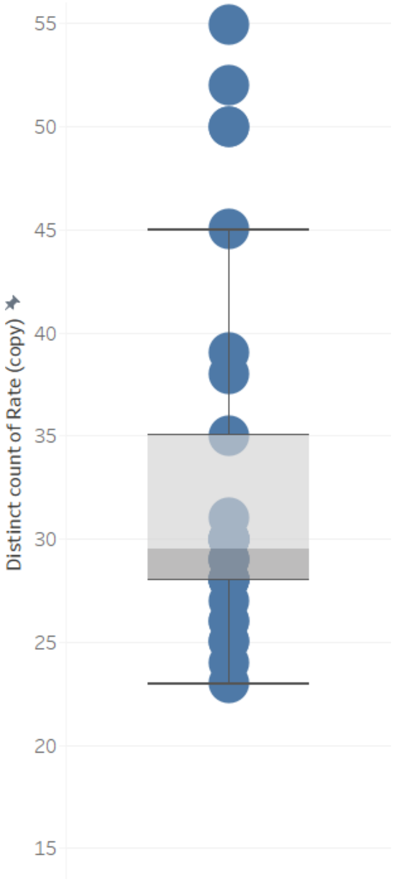
b. Accepting online orders or not.



5. Identify the location from where maximum online orders places using tree maps. Also find the count of online orders for “Sarjapur Road” location.



6. Find the city names getting maximum and minimum rating by using rate and Listed In(City) attributes to plot boxplot.


Distinct count of Rate (copy)