**Assignment 1, Level-3**

**1) #To change the datatype to factor, use as.factor function!**

**as.factor(titanic\_data$Survived)**

**as.factor(titanic\_data$Pclass)**

**as.factor(titanic\_data$Sex)**

**as.factor(titanic\_data$Embarked)**

**2) #A brief description on boxplots -**

**median (Q2/50th Percentile)**: the middle value of the dataset.**first quartile (Q1/25th Percentile)**: the middle number between the smallest number (not the “minimum”) and the median of the dataset.**third quartile (Q3/75th Percentile)**: the middle value between the median and the highest value (not the “maximum”) of the dataset.**interquartile range (IQR)**: 25th to the 75th percentile.**whiskers(Line extending from the IQR mark to the end).Outliers (points outside the whisker).“maximum”**: (Q3 + 1.5\*IQR)**“minimum”**: (Q1 -1.5\*IQR)

A box plot helps to find out about how well the values of the dataset is spread out.

#We have to use ggplot function for all kind of graphical assesments in an easy manner.

install.packages("ggplot2")

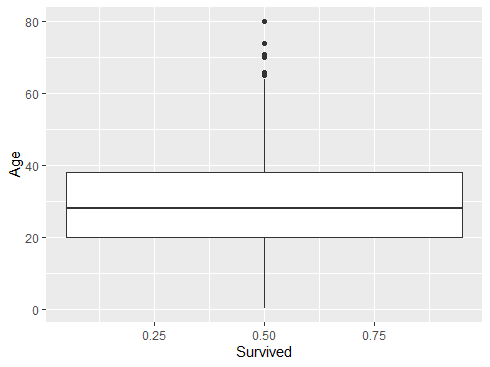
library("ggplot2")

a) ggplot(titanic\_data, aes(x = Survived, y = Age)) + geom\_boxplot()

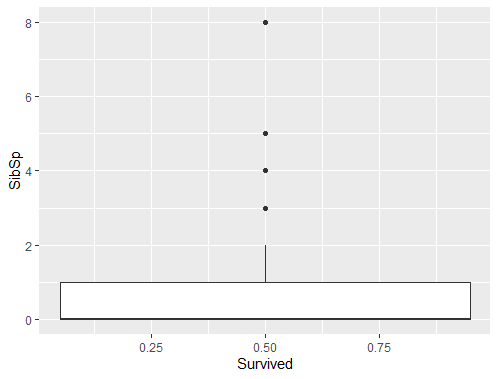
b) ggplot(titanic\_data, aes(x = Survived, y = SibSp)) + geom\_boxplot()

c) ggplot(titanic\_data, aes(x = Survived, y = Parch)) + geom\_boxplot()

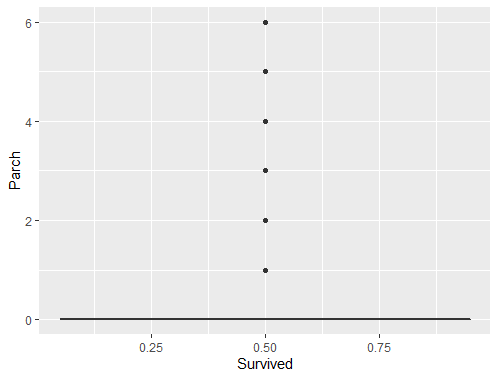
d) ggplot(titanic\_data, aes(x = Survived, y = Fare)) + geom\_boxplot()

(a)

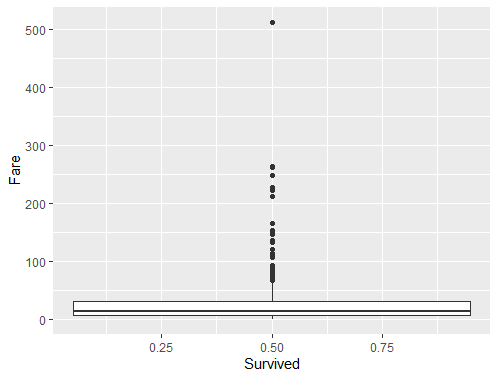
This curve is having median at around age 29, the distribution is quite uniform with the IQR ranging between a closed age range of 18. The min. is at 0, maximum is at around 65(maximum doesn’t denote that more no. of people survived, rather it’s the maximum age in the given dataset). So the IQR contains the highest no. Of persons survived but the highest age that survived is different and the outliers are after 65 age range, which doesn’t really change any data of the plot.

(b)

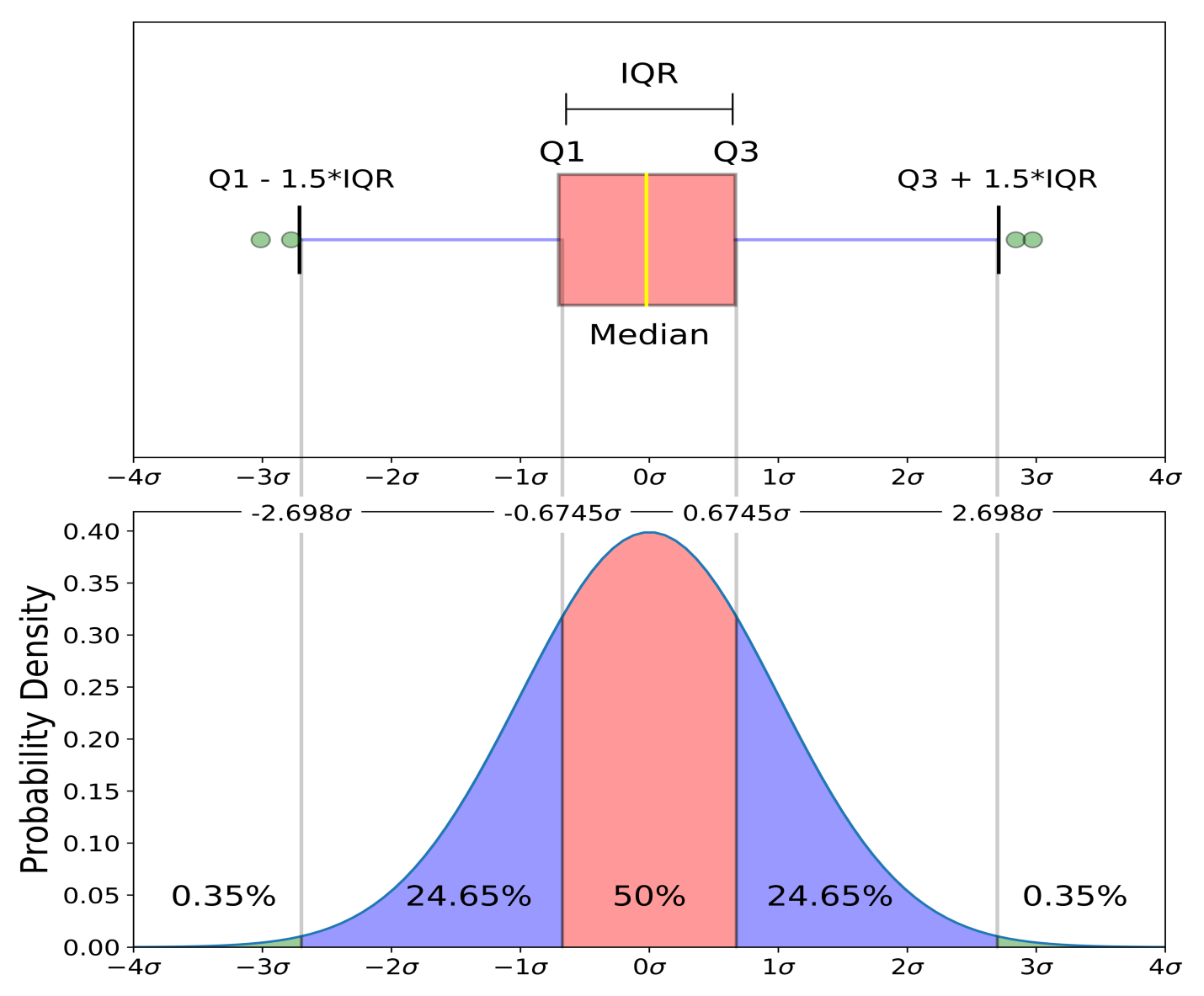
Im not quite sure but this plot does have its median at 0 and the value range of 2 siblings is having the maximum survival group. The maximum SibSp value at which a person survived is 2 after which outliers are there.

(C)

This curve neither contains an IQR nor a whisker showing the maximum point. Survived column is quite independent relating to the Parch data.

(d)

This curve is having the IQR very narrow comparing to the data range. Median is at around 10, Q1 value is less in range to Q3 value and the IQR is 25, so as per the rule maximum Fare value at which a person will survive would be 1.5\*25+30 = 67.5. And there after the outliers will continue.



#This image will make us understand the concept more well.

3) A bar plot is a lot more easier to understand than the box plot, It shows a categorical variable with is frequency or we can compare two categorical. For categorical comparision, we can have either grouped or stacked bar plots.

#Here we will use stacked bar plots as we are asked to.

a) ggplot(titanic\_data, aes(x = Survived, fill = Pclass)) +

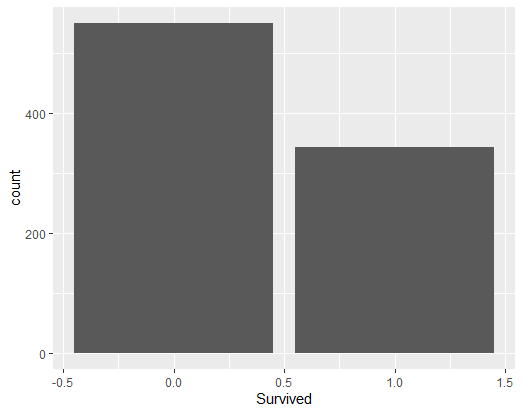
geom\_bar(position = "stack")

b) ggplot(titanic\_data, aes(x = Survived, fill = Sex)) +

geom\_bar(position = "stack")

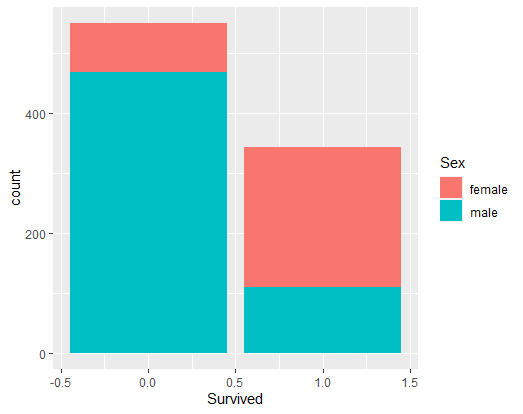
c) ggplot(titanic\_data, aes(x = Survived, fill = Embarked)) +

geom\_bar(position = "stack")

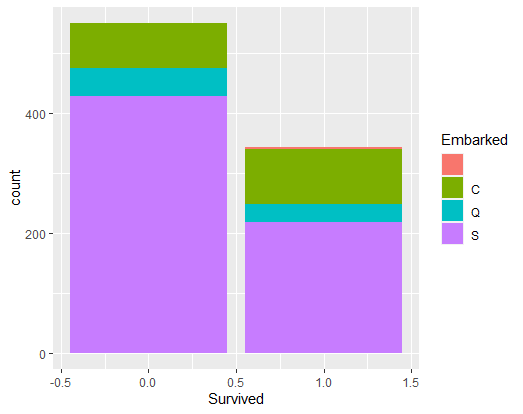
(a)

As Pclass is not a categorical data, we only have the comparision of the numbers of parents with survived column. (Otherwise there would’ve been division of the block into colors of different categories of y axis ex- Male/Female).

We can clearly see the values(count of all the range of people in the dataset) along with the survival rate.

(b)

This is a lot more simpler example where we see clearly about how categorical data (here male/female) is related with the survived column in the full ranged age group. As the survival rate is increasing, females are surviving in a greater ratio than the males.

(c)

In this curve, we see that the ship embarked from port Southampton has the population ratio which survives the most and simultaneously it’s increasing for Cherbourg port as the total population is decreasing.