Assignment No. 6.1

**QUESTION NO.1 : Process the names of the passengers and visualize the titles of the families**

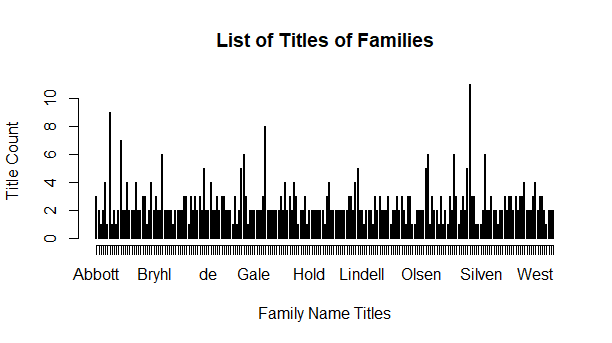
> titanic3\_1\_$FamilyName <- sapply(strsplit(titanic3\_1\_$name, " "), head, 1)

> titanic3\_1\_$FamilyName <- gsub(",","",titanic3\_1\_$FamilyName)

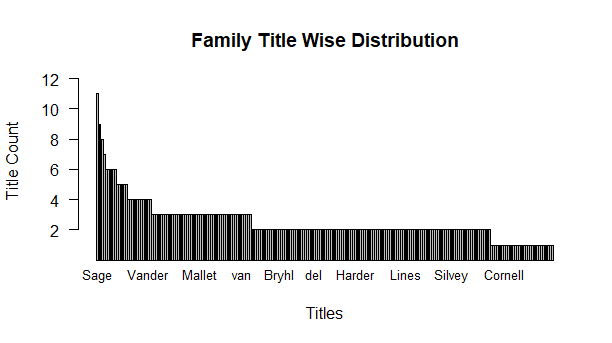
> titanic3\_1\_ <- transform(titanic3\_1\_, Family = ifelse(as.numeric(as.character(sibsp))>0,"F",ifelse(as.numeric(as.character(parch))>0,"F","S")))

> familiesdf <- sqldf('SELECT \* FROM titanic3\_1\_ WHERE Family ="F" ')

> plot(table(familiesdf$FamilyName),xlab ="Family Name Titles", ylab = "Title Count", main ="List of Titles of Families")



> barplot(sort$Freq, names.arg = sort$FamilyNames, horiz = FALSE, las = 1, cex.names = 0.8, xlab = "Titles", ylab = "Title Count", main = "Family Title Wise Distribution", ylim = c(1,12))



**QUESTION NO.2 : Proportion of people who survived from a family size**

> library(sqldf)

> survivors <- sqldf('SELECT survived, FamilyName FROM familiesdf')

> FamSize <- as.data.frame(survivors %>% group\_by(FamilyName) %>% count(FamilyName))

> head(FamSize)

FamilyName n

1 Abbott 3

2 Abelson 2

3 Ahlin 1

4 Aks 2

5 Allison 4

6 Andersen-Jensen 1

> FamSurvive <- as.data.frame(survivors %>% group\_by(FamilyName) %>% summarise(survived = sum(survived)))

> head(FamSurvive)

FamilyName survived

1 Abbott 1

2 Abelson 1

3 Ahlin 0

4 Aks 2

5 Allison 1

6 Andersen-Jensen 1

> SizeSurvive <- cbind(FamSize, FamSurvive)

> head(SizeSurvive)

FamilyName n FamilyName survived

1 Abbott 3 Abbott 1

2 Abelson 2 Abelson 1

3 Ahlin 1 Ahlin 0

4 Aks 2 Aks 2

5 Allison 4 Allison 1

6 Andersen-Jensen 1 Andersen-Jensen 1

> colnames(SizeSurvive)[2] <- "Size"

> SizeSurvive <- SizeSurvive[-3]

> SizeSurvive %>% group\_by(Size) %>% count(Size)

# A tibble: 10 x 2

# Groups: Size [10]

Size n

<int> <int>

1 1 29

2 2 109

3 3 46

4 4 11

5 5 5

6 6 5

7 7 1

8 8 1

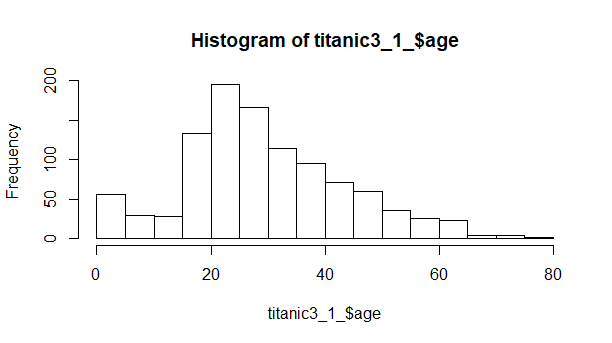
9 9 1

10 11 1

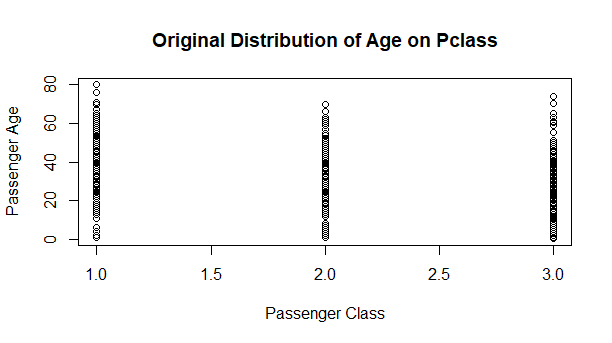
|  |
| --- |
| > SizeFamCount <- as.data.frame(SizeSurvive %>% group\_by(Size) %>% count(Size))  > SizeSurvCount <- as.data.frame(SizeSurvive) %>% group\_by(Size) %>% summarise(survived = sum(survived))  > head(SizeFamCount)  Size n  1 1 29  2 2 109  3 3 46  4 4 11  5 5 5  6 6 5  > head(SizeSurvCount)  # A tibble: 6 x 2  Size survived  <int> <dbl>  1 1 23  2 2 110  3 3 83  4 4 27  5 5 6  6 6 8 |
|  |
| |  | | --- | | > SizeSurvivalAnalaysis <- cbind(SizeFamCount, SizeSurvCount)  > head(SizeSurvivalAnalaysis)  Size n Size survived  1 1 29 1 23  2 2 109 2 110  3 3 46 3 83  4 4 11 4 27  5 5 5 5 6  6 6 5 6 8  > SizeSurvivalAnalaysis <- SizeSurvivalAnalaysis[-3]  > attach(SizeSurvivalAnalaysis)  The following object is masked from titanicgenclass:  n  > SizeSurvivalAnalaysis$PropSurv <- round((survived / (Size \* n)),2)\*100  > detach(SizeSurvivalAnalaysis)  > head(SizeSurvivalAnalaysis)  Size n survived PropSurv  1 1 29 23 79  2 2 109 110 50  3 3 46 83 60  4 4 11 27 61  5 5 5 6 24  6 6 5 8 27  > plot(x = SizeSurvivalAnalaysis$Size, y = SizeSurvivalAnalaysis$PropSurv, xlab = "Family Size", ylab = "Proportion Survived", main = "Family Size Wise Survival Analysis") | |

**QUESTION NO.3: Imputing the missing values in Age and comparing the before and after imputation distribution of age**

> hist(titanic3\_1\_$age) # before the imputation of missing values



> xyplot(titanic3\_1\_, age~pclass, pch=18,cex=1, main = "Distribution of Original and Imputed Age on Pclass")



> summary(titanic3\_1\_$age)

Min. 1st Qu. Median Mean 3rd Qu. Max. NA's

0.1667 21.0000 28.0000 29.8811 39.0000 80.0000 263

> skewness(titanic3\_1\_$age, na.rm = TRUE)

[1] 0.407087

> kurtosis(titanic3\_1\_$age, na.rm = TRUE)

[1] 3.140518

> library(mice)

> tempdata <- mice(titanic3\_1\_, m=5, method = 'pmm', maxit = 50, seed = 500)

summary(tempdata)

Multiply imputed data set

Call:

mice(data = titanic3\_1\_, m = 5, method = "pmm", maxit = 50, seed = 500)

Number of multiple imputations: 5

Missing cells per column:

pclass survived name sex age sibsp parch

0 0 0 0 263 0 0

VisitSequence:

age fare body

5 9 13

> completetitanic <- complete(tempdata,5)

> hist(completetitanic$age)

> summary(completetitanic$age)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.1667 20.0000 28.0000 28.9580 38.0000 80.0000

> mean(completetitanic$age, na.rm = FALSE)

[1] 28.95798

> mean(completetitanic$age, na.rm = TRUE)

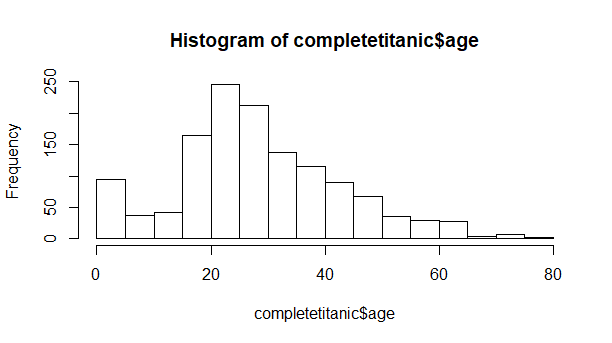
[1] 28.95798

> skewness(completetitanic$age, na.rm = TRUE)

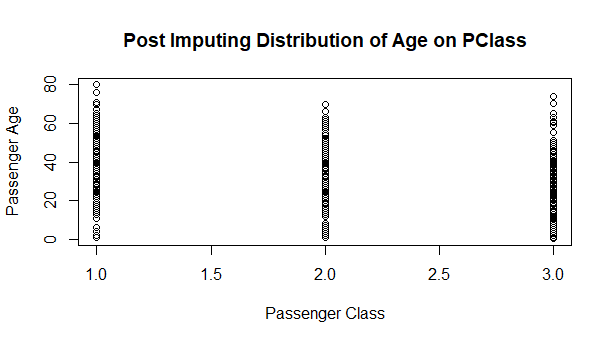
[1] 0.3913394

> kurtosis(completetitanic$age, na.rm = TRUE)

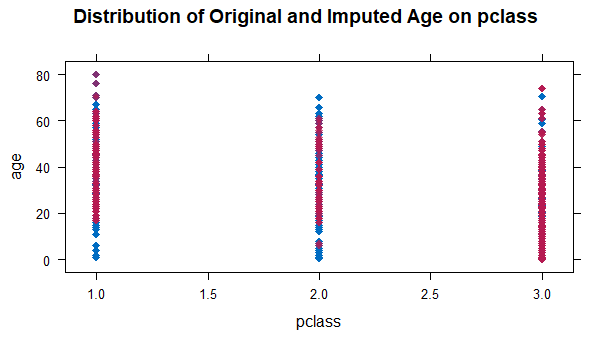
[1] 3.24207



> plot(x = completetitanic$pclass, y = completetitanic$age, xlab = "Passenger Class", ylab = "Passenger Age", main = "Post Imputing Distribution of Age on PClass")



> xyplot(tempdata, age~pclass, pch=18,cex=1, main = "Distribution of Original and Imputed Age on pclass")



> xyplot(tempdata, age~fare, pch=18,cex=1, main = "Distribution of Original and Imputed Age on fare")

