**Assignment**

# Step 1: Identify business problem

Question 1 – What are the problems that we are trying to solve in this assignment?

Question 2 – What are the algorithms that could be used to solve each problem?

***Step 2: Identify data sources and acquire data***

# Step 3: Process/Clean data

Question 3 – Find the number of missing data for each feature.

#To know if there is missing data is.na(df)

#To know the sum of missing data

sum(is.na(df)) mean(is.na(df))

# list rows of data that have missing values df[!complete.cases(df),]

Question 4 – What is the best way to deal with missing data (delete, calculate the average or the median, etc.)?

#delete NA value

row.has.na <- apply(df, 1, function(x){any(is.na(x))})

final.filtered <- df[!row.has.na,]

View(final.filtered) df = final.filtered sum(is.na(df))

Question 5 – Using an outlier detection technique (such as scatter, Z-score, or Box Plot), find if there are any outliers in the data (give these outliers).

-Remark – Remove any existing outliers.

#Dectecting outlier

continuous <-select\_if(df, is.numeric)

summary(continuous)

#Removing existing outliers

#Age outliers boxplot(continuous$age) continuous<-continuous[continuous$age<21,] boxplot(continuous$age) #Medu outliers boxplot(continuous$Medu) boxplot(continuous$Medu) #Fedu outliers boxplot(continuous$Fedu) continuous<-continuous[continuous$Fedu>0,] boxplot(continuous$Fedu) #traveltime outliers boxplot(continuous$traveltime) continuous<-continuous[continuous$traveltime<4,] boxplot(continuous$traveltime) #studytime outliers boxplot(continuous$studytime) continuous<-continuous[continuous$studytime<4,] boxplot(continuous$traveltime) #famrel outliers boxplot(continuous$famrel) continuous<-continuous[continuous$famrel<=2,] boxplot(continuous$famrel) #freetime outliers boxplot(continuous$freetime) boxplot(continuous$freetime) #goout outliers boxplot(continuous$goout) #health outliers boxplot(continuous$health) #absences outliers boxplot(continuous$absences)

#G1 outliers boxplot(continuous$G1) #G2 outliers boxplot(continuous$G2) #G3 outliers

boxplot(continuous$G3)

**Step 4**: **Perform exploratory analysis**

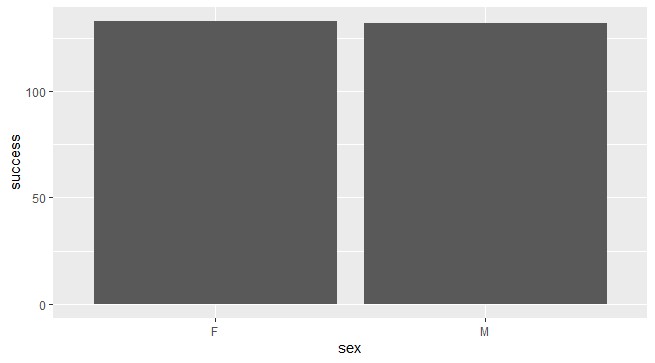
#Female and Male success and fail

library(xts) library(magrittr) result2Sex<-df%>%

mutate(success=ifelse(G3>=10,1,0), fail= ifelse(G3<10,1,0))%>% filter(sex=="F"|sex=="M")%>% group\_by(sex)%>% summarise(success=sum(success), fail=sum(fail))

View(result2Sex) result2Sex%>% ggplot(aes(x=sex,y=success))+

geom\_bar(stat="identity")



Output:

