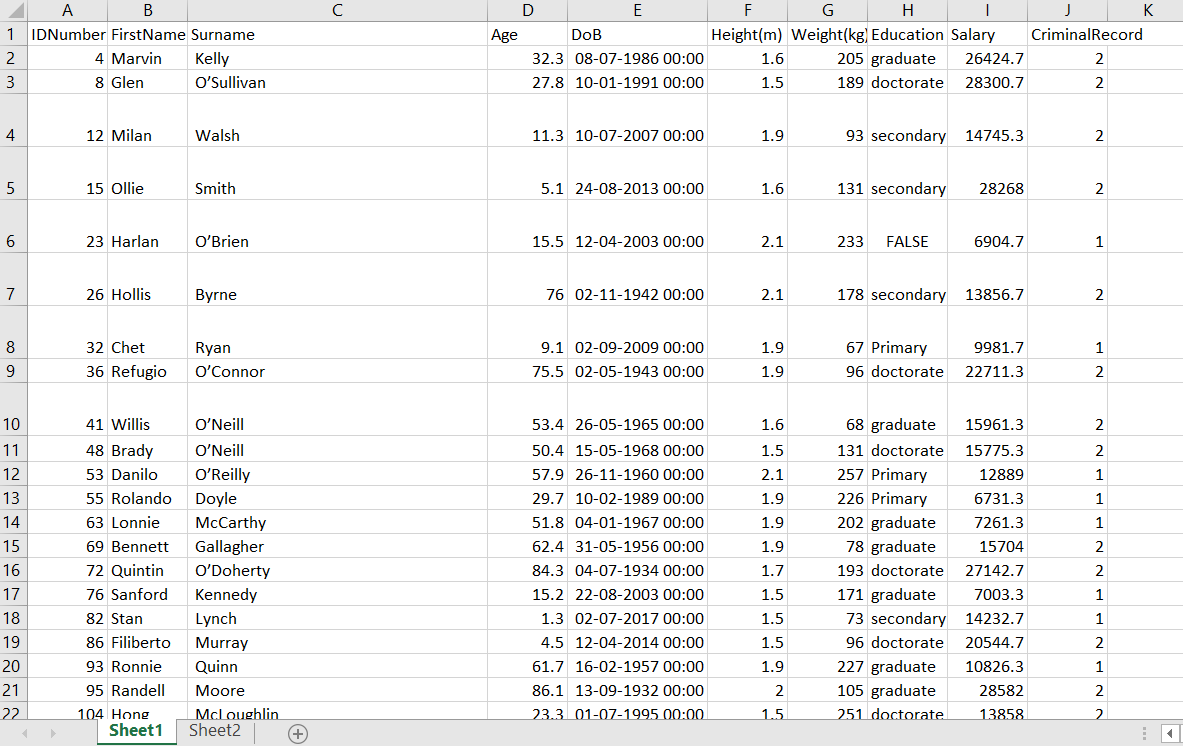
**INTRO TO R FOR DATA SCIENCE ASSIGNMENT-1**

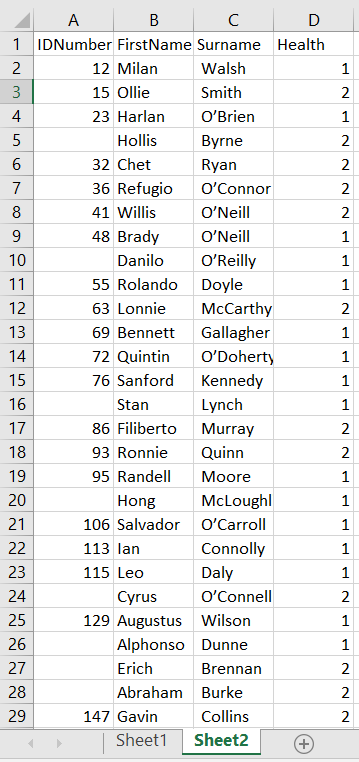
**STUDENT NAME: SHIVAANI KATRAGADDA**

**STUDENT NUMBER: R00183214**

**REPORT**

In this assignment, I was given an assignment1.xlsx file which contains two sheets, which were named Sheet1 and Sheet2. The Sheet1 contains ten columns and 160 rows of data. The Sheet2 contains four columns, 140 rows of data. The ten columns in the sheet1 are named as follows, IDNumber, FirstName, Surname, Age, DOB, Height(m), Weight(kg), Education, Salary, CriminalRecord. In sheet2 there are four columns which are named as follows IDNumber, FirstName, Surname, Health. Sheet2 contains the same columns that are present in sheet1 but there is one column which is different that is Health column. The sheets are shown below.





From the two sheets, it is very clear that there is some data missing in the sheet2 IDNumber column. So In order to tidy up the data the two data sheets should be matched by using IDNumber, FirstName, Surname columns and merge the two sheets and make it into one sheet.

In order to do this task, there are some built-in packages in R which are required to complete the task. The packages that I have used in the task are as follows

1.TidyVerse

2.Readxl

3.Dplyr

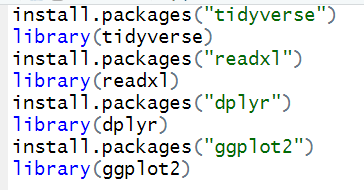
4.GGplot2

Tidyverse: The Tidyverse package is a powerful package that is used for transforming and visualizing the data.

Readxl: The readxl package is useful to get the data from excel to R easily.

Dplyr: Dplyr is a part of the tidyverse package, which is useful for manipulating datasets in R very effectively.

GGplot2: This package is particularly useful for visualizing the data, it is a part of the tidyverse package.



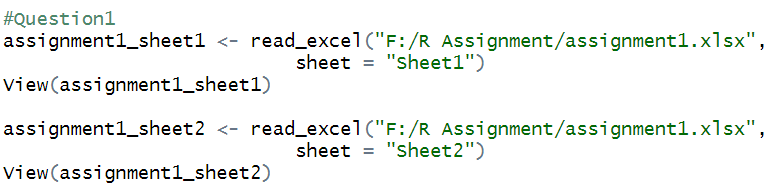
Install.packages(“ “) Is used for installing the packages in R

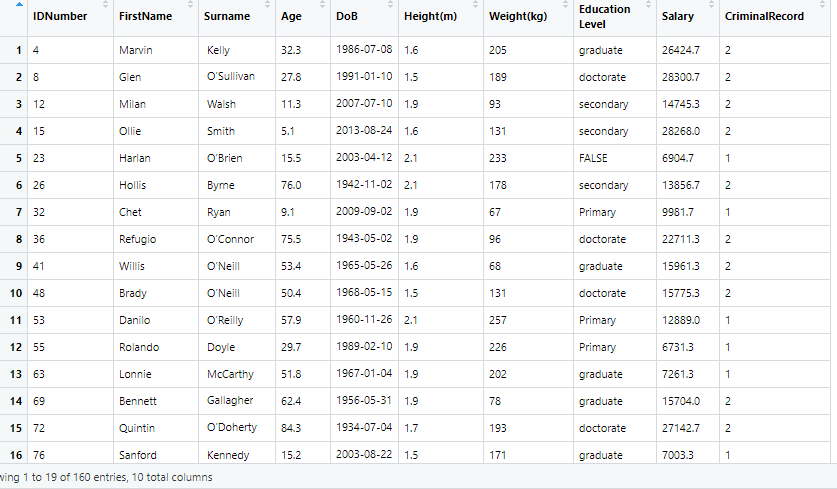
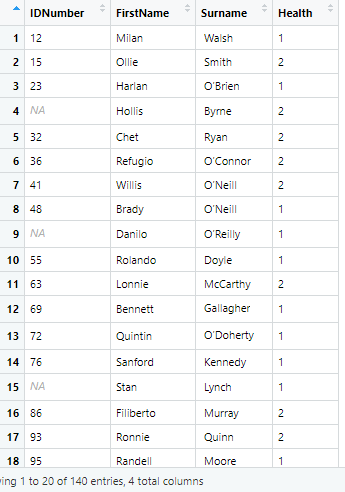
library() loads and attaches the add-on packages.

--------------------------------------------------------------------------------------------------------------------------

**TASK1**. Using the xlsx or readxl package or otherwise, read each sheet in the ”assignment1.xlsx” file into R.

In this task, I have to read both the sheets. So I used readxl package to read each sheet from the given file “assignment1.xlsx”



****

I have read both the sheets with the help of readxl and stored sheet1 in assignment\_sheet1 and sheet2 in assignment\_sheet2, after reading the data will be viewed as shown in the figure in Rstudio.

--------------------------------------------------------------------------------------------------------------------------

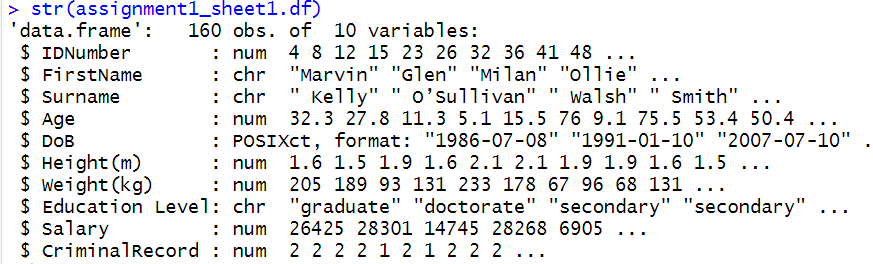
**TASK2**: Generate a data frame for each sheet in the file.

I have to generate data frame for each sheet i.e for assignment\_sheet1,assignment\_sheet2

assignment1\_sheet1.df <- as.data.frame(assignment1\_sheet1)

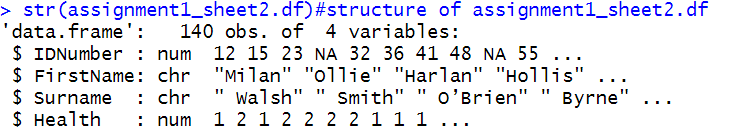
str(assignment1\_sheet1.df)

The structure for assignment1\_sheet1.df is as follows



assignment1\_sheet2.df <- as.data.frame(assignment1\_sheet2)

str(assignment1\_sheet2.df)



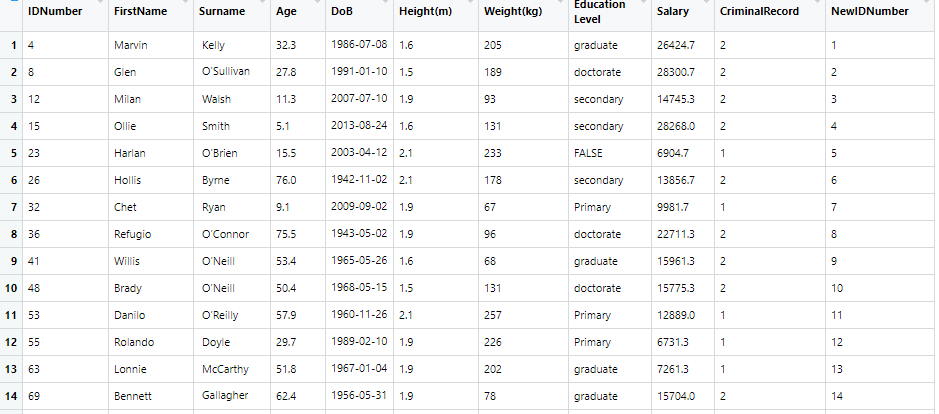
The as.data.frame will convert the given sheets to data frames, the sheet1 data frame is named as assignment1\_sheet1.df and for sheet2 data frame is named as assignment1\_sheet2.df

--------------------------------------------------------------------------------------------------------------------------

**TASK3:** The dataset in the first sheet is a random selection from a larger dataset. You will never get access to the full dataset so you should regenerate a new identification number for each subject in the dataset. This should be the row number of each entry in Sheet 1. You do not need to do this for Sheet 2.

I have to generate a new subject i.e new column which will contain a unique identification number for each subject in the dataset for sheet1 and that column is named as NewIDNumber

assignment1\_sheet1.df$NewIDNumber<-c(seq(1,160))



The new column NewIDNumber is added for sheet1 which contains a unique identification number for all 1 to 160 rows in sequence. The seq(Start, End) will help to generate the numbers in sequence order.

--------------------------------------------------------------------------------------------------------------------------

**TASK4:** It is also required to have an additional identifier which is the number you have generated in (3) followed by the first letter of each subject's first name and then followed by the first letter of each subject’s surname. You do not need to do this for Sheet 2.

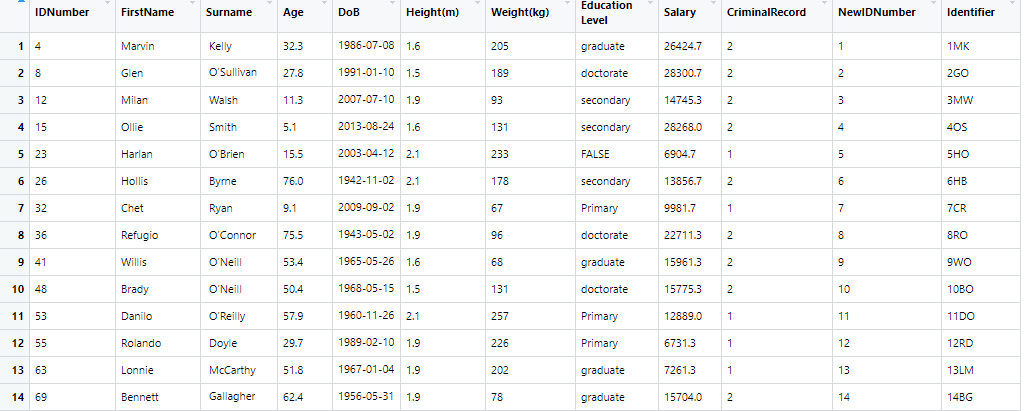
In this task, I have to generate a new column that contains the unique identification number followed by the first letter in full name and surname and that column is named as an Identifier. For this, I used substr() function which helps to take the desired position of the letter in the string and the gsub() function helps to remove spaces in the start, end and in between the string. The paste() function helps to combine the data from different columns.

z<-substr(assignment1\_sheet1.df$FirstName,1,1)

x <- gsub("\\s+", "",assignment1\_sheet1.df$Surname)

z1<-substr(x,1,1)

assignment1\_sheet1.df$Identifier<paste(assignment1\_sheet1.df$NewIDNumber,z,z1,sep=")

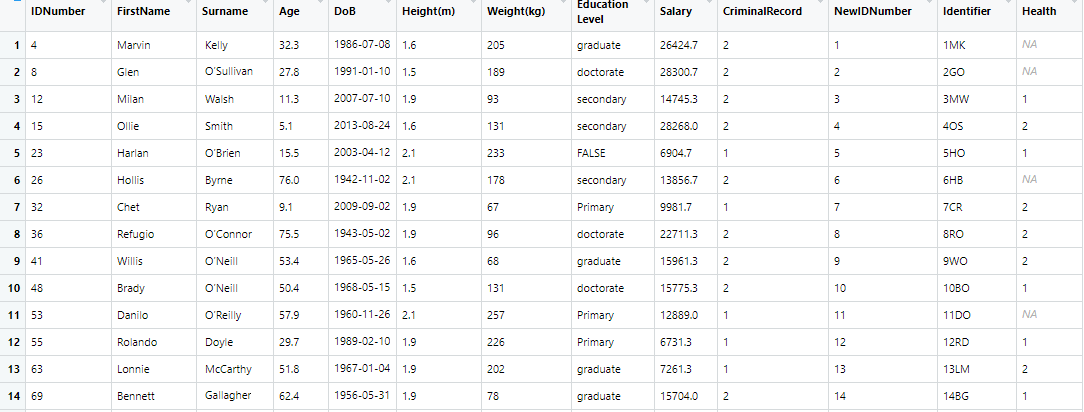


--------------------------------------------------------------------------------------------------------------------------

**TASK5:** Although the data is not available for most subjects, some data highlighting the subject's state of health is available in Sheet 2. You should use the subjects' ID number to match it and merge it with the data in Sheet 1.

In this task, I have to merge the two sheets by using the IDNumber column. So I have used Left join here to merge both the sheets so that all the rows in the sheet1 will get the data from sheet2. If the data is not present in sheet2 for that particular row NA(missing value) will appear.

merge\_data<-left\_join(assignment1\_sheet1.df,assignment1\_sheet2.df[c("IDNumber","Health")], by="IDNumber")



I have used the left join for merging both the data frames by matching it with IDNumber, so both the data frames will be merged.

--------------------------------------------------------------------------------------------------------------------------

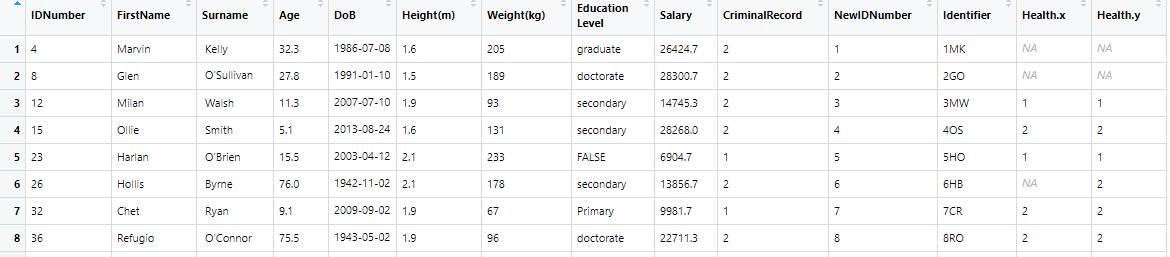
**TASK6:** Not every subject has its ID number included in Sheet 2. You should attempt to match the remaining subjects using their first and surnames. This must be done using tidyverse in a robust manner. Your code for doing this should work again in the case of a new sample of data being provided.

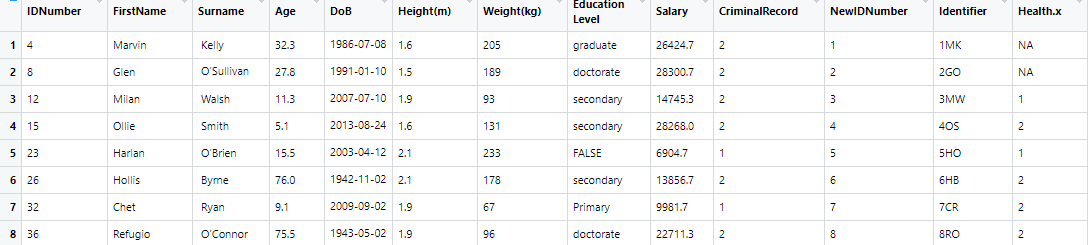
In this task, i have to fill the data in rows by using FirstName and surname columns so I used left\_join and piped it with the merge\_data and stored that data in merge\_data1, in this we get two Health columns i.e Health.x,Health.y. I have to use unite() function to unite both the Health.x, Health.y columns inorder to get the whole data in one column and assigned it to r variable. The data will be united and will be stored in Health.x, and I have Renamed health.x as Health assigned to final\_df variable.

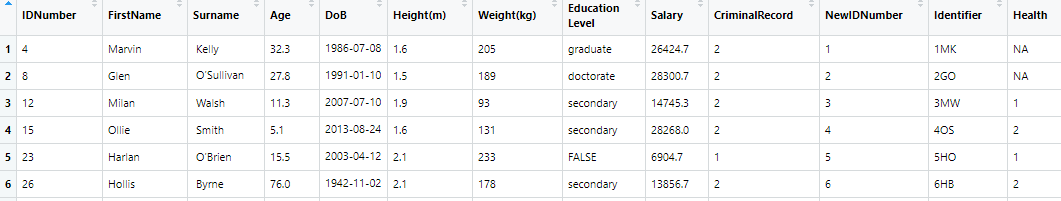
merge\_data1<-merge\_data%>% left\_join(assignment1\_sheet2.df[c("FirstName","Surname","Health")], by=c("FirstName","Surname"))

r=unite(merge\_data1, Health.x,Health.y)

final\_df= r%>% rename(Health = Health.x)







--------------------------------------------------------------------------------------------------------------------------

**TASK7:** You should add a column for age range. This should be Age Range Category Name 0-17->1,18-35-> 2,35-54->3,54-74->4,74+-> 5

Adding the new column CategoryName which contains the category number of the age range. Category Name is 1 if the age is between 0-18,2 if the age is between 18-35,3 if the age is between 35-54,4 if the age is between 54-74,5 if the age 74 and above.

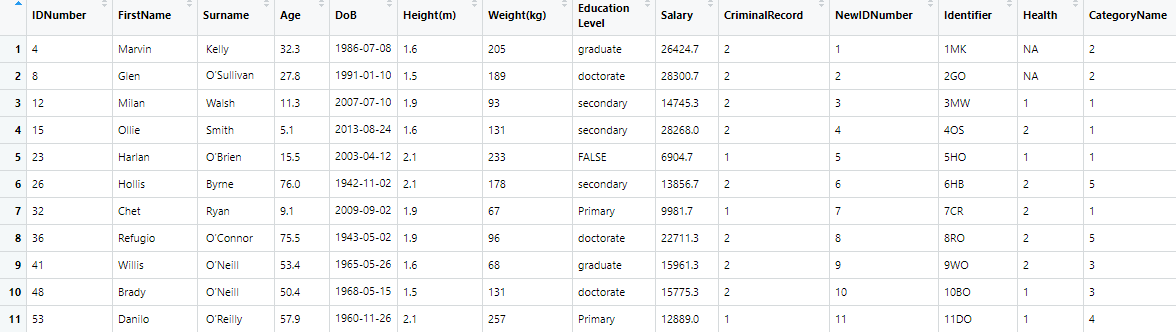
final\_df$CategoryName[final\_df$Age >= 0 & final\_df$Age <= 18] <- "1"

final\_df$CategoryName[final\_df$Age >= 18 & final\_df$Age < 35] <- "2"

final\_df$CategoryName[final\_df$Age >= 35 & final\_df$Age < 54] <- "3"

final\_df$CategoryName[final\_df$Age >= 54 & final\_df$Age < 74] <- "4"

final\_df$CategoryName[final\_df$Age >= 74] <- "5"



--------------------------------------------------------------------------------------------------------------------------

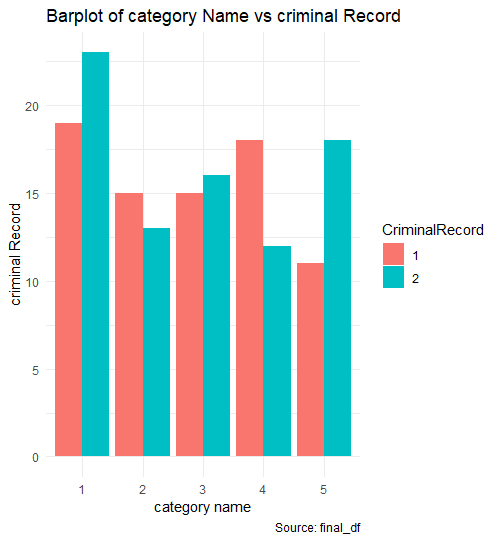
**TASK8:** You should filter the data by each age category. Generate a bar plot using ggplot2 for the criminal record variable.

I have filtered the data by each age category and generated bar plot between category name and criminal record using ggplot. Before generating the plot I have converted the Criminal Record to categorical data because the age category is categorical data.

final\_df$CriminalRecord<-as.factor(final\_df$CriminalRecord)

ggplot(final\_df,aes(x=CategoryName,fill=CriminalRecord)) + labs(caption = "Source: final\_df") +geom\_bar(position="dodge")+ ggtitle(label="Barplot of category Name vs criminal Record")+xlab("category name")+ylab("criminal Record")+theme\_minimal()

This barplot is between category Name and Criminal Record. In x axis category Name is considered and on y axis Criminal Record is taken depending up on the age category and criminalRecord the graph will be plotted .And this graph main source is final\_df dataframe.



--------------------------------------------------------------------------------------------------------------------------

**TASK9**: You should generate an appropriate visualization examining the relationships between height, weight, age and criminal records. Comment on this.

In this task I have taken Height(m), CriminalRecords as categorical data and Age, Weight(kg) as numerical data.

final\_df$`Height(m)`<-as.factor(final\_df$`Height(m)`)

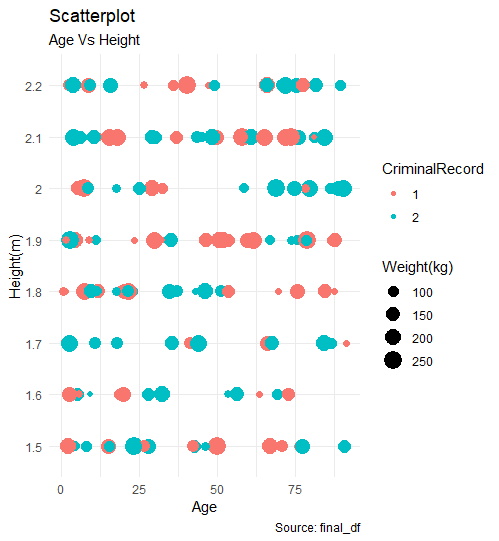
final\_df$CriminalRecord<-as.factor(final\_df$CriminalRecord)

After that plotting the graph using Age,Height(m),Weight(kg),CriminalRecord. I have plotted three types of plots they are scatterplot,Barplot and Bubble plot

I have considered Age,criminalRecord on x axis and Height(m),Weight(kg) on y axis and plotted graph using ggplot

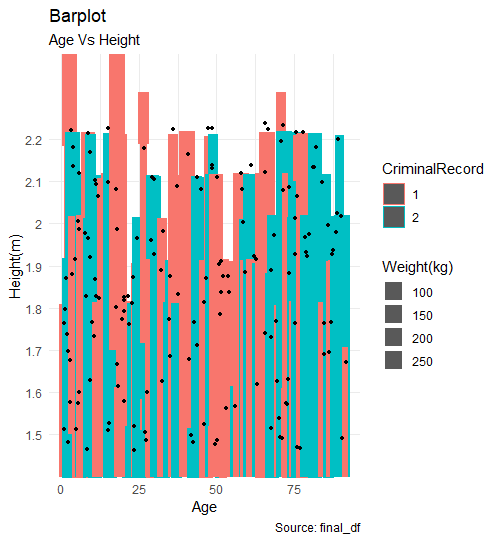
1.Scatterplot

ggplot(final\_df, aes(x=Age, y=`Height(m)`)) + geom\_point(aes(col=CriminalRecord, size=`Weight(kg)`))+ theme\_minimal()+labs(subtitle="Age Vs Height", x="Age",y="Height(m)",title="Scatterplot",caption = "Source: final\_df")



2.Barplot

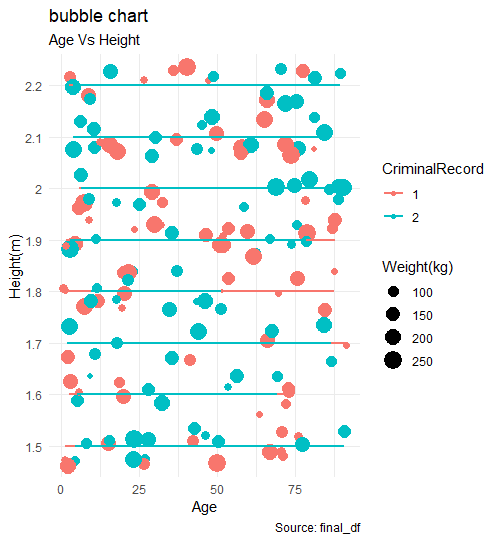
ggplot(final\_df, aes(x=Age, y=`Height(m)`)) + geom\_bar(stat="identity",aes(col=CriminalRecord, size=`Weight(kg)`))+geom\_jitter(width = .5, size=1)+ theme\_minimal()+labs(subtitle="Age Vs Height", x="Age",y="Height(m)",title="Barplot",caption = "Source: final\_df")



3.Bubbleplot

g <- ggplot(final\_df, aes(x = Age,y=`Height(m)`)) + theme\_minimal()+labs(subtitle="Age Vs Height", x="Age",y="Height(m)",title="Scatterplot",caption = "Source: final\_df")+geom\_jitter(aes(col=CriminalRecord, size=`Weight(kg)`))+ geom\_smooth(aes(col=CriminalRecord), method="lm", se=F)

g



From the plots we can say that the criminal record 1 is normal lessly scattered and it is not showing any effective changes when height and age increasing.When compare to both criminal record 2 is showing much progress and densly populated with respect to increase in the height and weight.

--------------------------------------------------------------------------------------------------------------------------

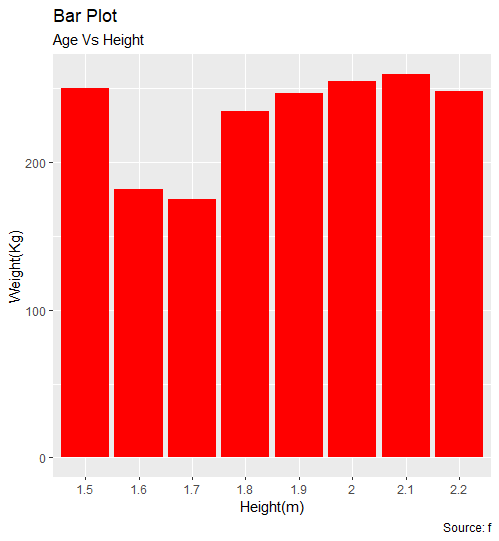
**TASK10**: Using filters, you should analyze if there are any interesting results in the dataset regarding the relationships between height, weight and criminal record. Use appropriate visualizations.

In this task, I have filtered the data by taking the criminal record. At first, I filtered the data by taking a criminal record equal to one and plotted the barplot. And then I filtered the data by taking a criminal record equal to two and plotted the barplot.

f<-filter(final\_df,CriminalRecord==1)

Barplot with respect to Height and Weight for Criminal Record 1

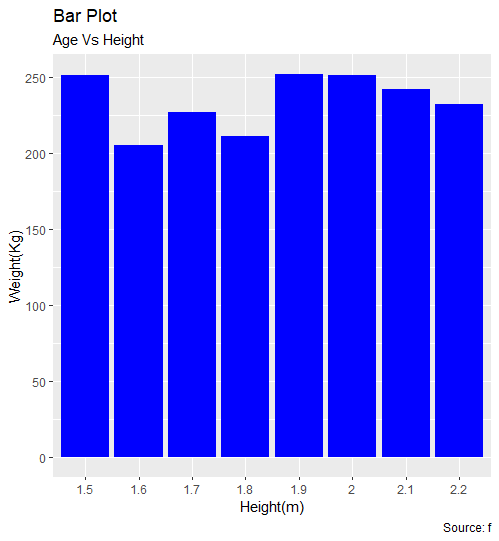
ggplot(f,aes(x=`Height(m)`, y=`Weight(kg)`)) + geom\_bar(stat="identity",position="dodge",fill="red")+ labs(subtitle="Age Vs Height", x="Height(m)",y="Weight(Kg)",title="Bar Plot",caption = "Source: f")



f1<-filter(final\_df,CriminalRecord==2)

Barplot with respect to Height and Weight for Criminal Record 2

ggplot(f1,aes(x=`Height(m)`, y=`Weight(kg)`)) + geom\_bar(stat="identity",position="dodge",fill="blue")+ labs(subtitle="Age Vs Height", x="Height(m)",y="Weight(Kg)",title="Bar Plot",caption = "Source: f")



From the two graphs based on criminalRecord we can say that for criminal record 1 the graph is constantly increasing from certain point and there is no deep raise or fall from the particular point with respect to weight and height. But in case of criminal Record2 the graph is constantly changing from one point to another point with respect to height and weight

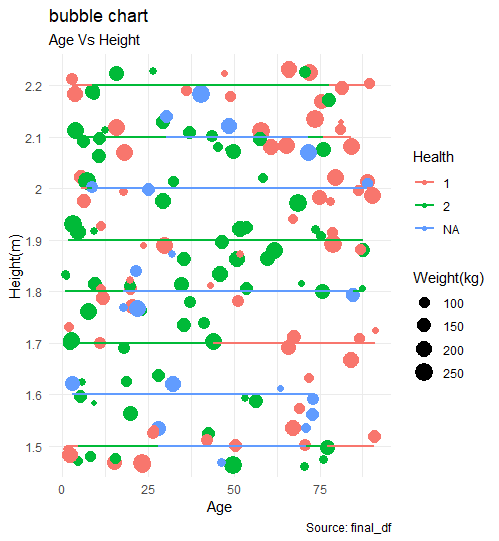
--------------------------------------------------------------------------------------------------------------------------

**TASK11**: Generate a smaller data frame for the subjects where health related data is available. Examine if there is a relationship between the different states of health and height, weight or age. Use appropriate visualisations. Note this should include a modelling type analysis such as regression.

In this task first I am subsetting the data from the health column, I am considering all the rows of data where health not equal to missing values and assigned to ss variable.

ss<-subset(final\_df,Health!="NA")

I plotted graph a bubble chart between Age, Height(m), Weight(kg) and Health



From the graph, we can say that Health 2 value is densely populated with respect to Age, Height(m), Weight(kg).

Now I am using Multiple logistic Regression

Firstly, I have to install a nnet package which is used for neutral networks. After that have to convert health data into categorical data. Later Assign ss$Health to variable Health, ss$Age to Age, ss$Height(m) to Height, ss$Weight(kg) weight. And then there are 2 values in health data so we need reference level, therefore, relevel() is used and considered reference 1 i.e ref=1 and assigning to ss$out variable. Now create the model by using the three columns i.e Age, Height, Weight and have to look at the summary of the model.

Now the model is built and predicts the model followed by misclassification error. And then perform the testing method.

#installing nnet package which is used for neutral networks

install.packages("nnet")

library(nnet)#loads and attaches the add-on packages.

#converting health data in to categorical data

ss$Health <- as.factor(ss$Health)

#Assigning ss$Health to variable Health

Health<-ss$Health

#Assigining ss$Age to variable Age

Age<-ss$Age

#Assigining ss$Height(m) to variable Height

Height<-ss$`Height(m)`

#Assigining ss$Weight(kg) to variable weight

Weight<-ss$`Weight(kg)`

#there are 2 values in health data so we need reference level therefore relevel() is used and considered reference 1 i.e ref=1 and assigning to ss$out variable

ss$out<-relevel (ss$Health, ref="1")

#creating the model by using the three columns i.e Age, Height, Weight

mymodel<-multinom(out~Age+Height+Weight, data=ss)

#summary of the model

summary(mymodel)

#predict

predict(mymodel,ss) #predicting the model

predict(mymodel,ss,type="prob") #predicting the model with probability

#missclassification error

cm<-table(predict(mymodel),ss$Health)#preparing confusion matrix and named as cm

cm #printing the output of confusion matrix

1-sum(diag(cm))/sum(cm)#checking the misclassification error percentage by formula

#2-tailed z test

z <- summary(mymodel)$coefficients/summary(mymodel)$standard.errors#performing 2-tailed z test

p <- (1 - pnorm(abs(z), 0, 1)) \* 2#performing 2-tailed z test

p#printing the output of p

**CONCLUSION:**

The given data is cleaned and performed visualization for the multiple columns, I observed that among all the columns in the dataset the health, criminal Record columns have only two values 1 and 2, for both the columns the value 2 is mostly populated and scattered.