# STAT8010 Mid term Assignment #1

Bernard McNamee

09/11/2020

# Introduction

This R statistical analysis report details the STAT8010 Mid term Assignment #1 with answers to the 11 questions asked. The report is generated using the RMarkdown package as a HTML document using libraries 'kableExtra' and 'gridExtra' for formatting and using libraries 'tidyverse' and 'xlsx' for data manipulation. The dataset is extracted from two sheets of an Excel spreadsheet titled 'assignment1.xlsx' into R. The questions posed require cleaning, transformation and analysis of the data with answers in text and plot/chart output. \

# Questions

**Q1. Using the xlsx or readxl package or otherwise, read each sheet in the "assignment1.xlsx" file into R.**

The package 'read.xlsx' is installed and loaded into R and then using the library, the two sheets are separately read into R.

**Q2. Generate a data frame for each sheet in the file.**

The process is repeated but this time the data is saved into two new dataframes, ie 'Sheet1' and 'Sheet2' are read into dataframes 'Sheet1' and 'Sheet2' respectively.

A peek at the data reveals the structure and example values for each sheet.

```
## 'data.frame':    160 obs. of  11 variables:
##  $ IDNumber       : num  4 8 12 15 23 26 32 36 41 48 ...
##  $ FirstName      : chr  "Marvin" "Glen" "Milan" "Ollie" ...
##  $ Surname        : chr  " Kelly" " O'Sullivan" " Walsh" " Smith" ...
##  $ Age            : num  32.3 27.8 11.3 5.1 15.5 76 9.1 75.5 53.4 50.4 ...
##  $ DoB            : Date, format: "1986-07-08" "1991-01-10" ...
##  $ Height.m.      : num  1.6 1.5 1.9 1.6 2.1 2.1 1.9 1.9 1.6 1.5 ...
##  $ Weight.kg.     : num  205 189 93 131 233 178 67 96 68 131 ...
##  $ Education.Level: chr  "graduate" "doctorate" "secondary" "secondary" ...
##  $ Salary         : num  26425 28301 14745 28268 6905 ...
##  $ CriminalRecord : num  2 2 2 2 1 2 1 2 2 2 ...
##  $ NA.            : logi  NA NA NA NA NA NA ...
```

```
## 'data.frame':    140 obs. of  4 variables:
##  $ IDNumber : num  12 15 23 NA 32 36 41 48 NA 55 ...
##  $ FirstName: chr  "Milan" "Ollie" "Harlan" "Hollis" ...
##  $ Surname  : chr  " Walsh" " Smith" " O'Brien" " Byrne" ...
##  $ Health   : num  1 2 1 2 2 2 2 1 1 1 ...
```

**Q3. 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.**

A new identifier called 'newID' is created and assigned to each row of the dataframe **-** the value of each 'newID' is equal to the row number.

A peek at the data reveals the head of the dataframe 'Sheet1' and the **new variable appears as the last column.**

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Marvin | Kelly | 32.3 | 1986-07-08 | 1.6 | 205 | graduate | 26424.7 | 2 | NA | **1** |
| 8 | Glen | O'Sullivan | 27.8 | 1991-01-10 | 1.5 | 189 | doctorate | 28300.7 | 2 | NA | **2** |
| 12 | Milan | Walsh | 11.3 | 2007-07-10 | 1.9 | 93 | secondary | 14745.3 | 2 | NA | **3** |
| 15 | Ollie | Smith | 5.1 | 2013-08-24 | 1.6 | 131 | secondary | 28268.0 | 2 | NA | **4** |

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 23 | Harlan | O'Brien | 15.5 | 2003-04-12 | 2.1 | 233 | FALSE | 6904.7 | 1 | NA | **5** |
| 26 | Hollis | Byrne | 76.0 | 1942-11-02 | 2.1 | 178 | secondary | 13856.7 | 2 | NA | **6** |

**Q4. 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 subjects first name and then followed by the first letter of each subject's surname. You do not need to do this for Sheet 2.**

A new identifier called 'otherID' is created and assigned to each row of the dataframe - the value of each 'otherID' is equal to 'newID' + 1st letter First Name + 1st letter Surname.

A peek at the data reveals the head of the dataframe 'Sheet1' and the **new variable appears as the last column.**

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID | otherID |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Marvin | Kelly | 32.3 | 1986-07-08 | 1.6 | 205 | graduate | 26424.7 | 2 | NA | 1 | **1MK** |
| 8 | Glen | O'Sullivan | 27.8 | 1991-01-10 | 1.5 | 189 | doctorate | 28300.7 | 2 | NA | 2 | **2GO** |
| 12 | Milan | Walsh | 11.3 | 2007-07-10 | 1.9 | 93 | secondary | 14745.3 | 2 | NA | 3 | **3MW** |
| 15 | Ollie | Smith | 5.1 | 2013-08-24 | 1.6 | 131 | secondary | 28268.0 | 2 | NA | 4 | **4OS** |
| 23 | Harlan | O'Brien | 15.5 | 2003-04-12 | 2.1 | 233 | FALSE | 6904.7 | 1 | NA | 5 | **5HO** |
| 26 | Hollis | Byrne | 76.0 | 1942-11-02 | 2.1 | 178 | secondary | 13856.7 | 2 | NA | 6 | **6HB** |

**Q5. Although the data is not available for most subjects, some data highlighting subjects 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.**

Health values in Sheet2 are found using two nested loops to cycle through each row of each sheet and check for available values and update Sheet1 with a new Health variable for all values found.

A peek at the data reveals the head of the dataframe 'Sheet1' and the **new variable appears as the last column.**

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID | otherID | Health |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Marvin | Kelly | 32.3 | 1986-07-08 | 1.6 | 205 | graduate | 26424.7 | 2 | NA | 1 | 1MK | **NA** |
| 8 | Glen | O'Sullivan | 27.8 | 1991-01-10 | 1.5 | 189 | doctorate | 28300.7 | 2 | NA | 2 | 2GO | **NA** |
| 12 | Milan | Walsh | 11.3 | 2007-07-10 | 1.9 | 93 | secondary | 14745.3 | 2 | NA | 3 | 3MW | **1** |
| 15 | Ollie | Smith | 5.1 | 2013-08-24 | 1.6 | 131 | secondary | 28268.0 | 2 | NA | 4 | 4OS | **2** |
| 23 | Harlan | O'Brien | 15.5 | 2003-04-12 | 2.1 | 233 | FALSE | 6904.7 | 1 | NA | 5 | 5HO | **1** |
| 26 | Hollis | Byrne | 76.0 | 1942-11-02 | 2.1 | 178 | secondary | 13856.7 | 2 | NA | 6 | 6HB | **NA** |

**Q6. 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.**

This is the filtered dataset for 'Sheet 2' with missing ID numbers.

| IDNumber | FirstName | Surname | Health |
|---|---|---|---|
| **NA** | Hollis | Byrne | 2 |
| **NA** | Danilo | O'Reilly | 1 |
| **NA** | Stan | Lynch | 1 |

| IDNumber | FirstName | Surname | Health |
|---|---|---|---|
| **NA** | Hong | McLoughlin | 1 |
| **NA** | Cyrus | O'Connell | 2 |
| **NA** | Alphonso | Dunne | 1 |
| **NA** | Erich | Brennan | 2 |
| **NA** | Abraham | Burke | 2 |

This is the head of the new dataset for 'Sheet 1' following an update to find missing Health values.

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID | otherID | Health |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Marvin | Kelly | 32.3 | 1986-07-08 | 1.6 | 205 | graduate | 26424.7 | 2 | NA | 1 | 1MK | NA |
| 8 | Glen | O'Sullivan | 27.8 | 1991-01-10 | 1.5 | 189 | doctorate | 28300.7 | 2 | NA | 2 | 2GO | NA |
| 12 | Milan | Walsh | 11.3 | 2007-07-10 | 1.9 | 93 | secondary | 14745.3 | 2 | NA | 3 | 3MW | 1 |
| 15 | Ollie | Smith | 5.1 | 2013-08-24 | 1.6 | 131 | secondary | 28268.0 | 2 | NA | 4 | 4OS | 2 |
| 23 | Harlan | O'Brien | 15.5 | 2003-04-12 | 2.1 | 233 | FALSE | 6904.7 | 1 | NA | 5 | 5HO | 1 |
| **26** | **Hollis** | **Byrne** | **76.0** | **1942-11-02** | **2.1** | **178** | **secondary** | **13856.7** | **2** | **NA** | **6** | **6HB** | **2** |

*Note the Health for the subject on the last row (Hollis Byrne) has now been updated to **'2'** (from **'NA'** in table in previous question).*

**Q7. You should add a column for age range. This should be**

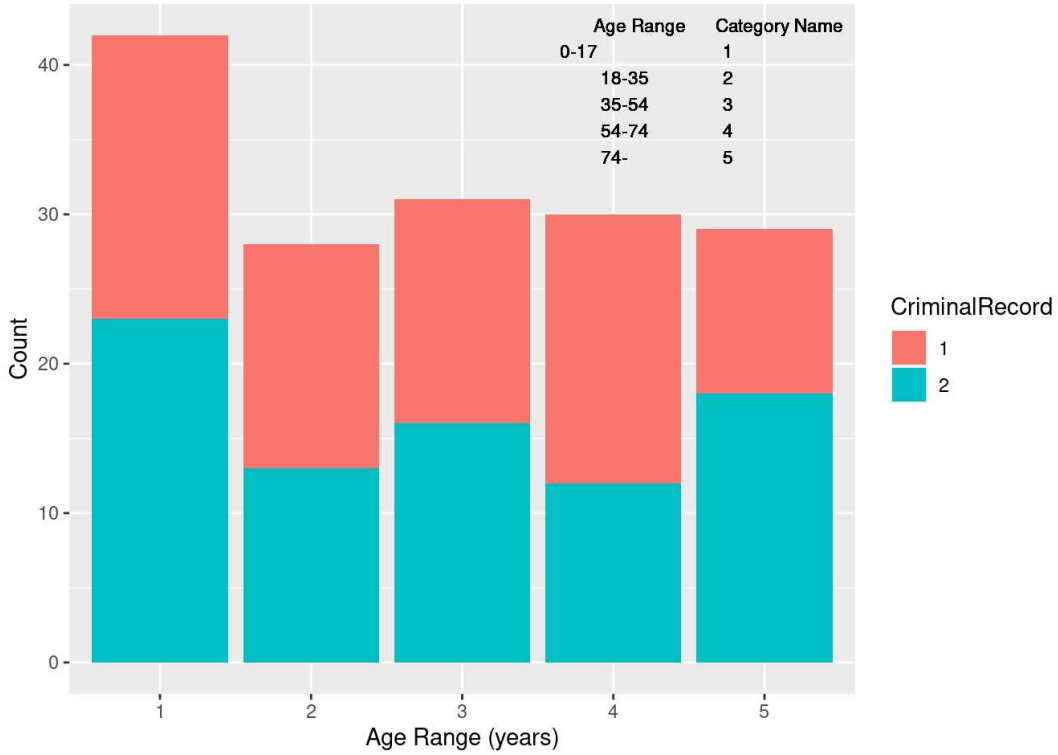
| Age Range | Category Name |
|---|---|
| **0-17** | **1** |
| **18-35** | **2** |
| **35-54** | **3** |
| **54-74** | **4** |
| **74-** | **5** |

A new variable 'Age Range' **appears in the last column** and corresponds to values in the table above.

| IDNumber | FirstName | Surname | Age | DoB | Height.m. | Weight.kg. | Education.Level | Salary | CriminalRecord | NA. | newID | otherID | Health | AgeRange |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 4 | Marvin | Kelly | 32.3 | 1986-07-08 | 1.6 | 205 | graduate | 26424.7 | 2 | NA | 1 | 1MK | NA | **2** |
| 8 | Glen | O'Sullivan | 27.8 | 1991-01-10 | 1.5 | 189 | doctorate | 28300.7 | 2 | NA | 2 | 2GO | NA | **2** |
| 12 | Milan | Walsh | 11.3 | 2007-07-10 | 1.9 | 93 | secondary | 14745.3 | 2 | NA | 3 | 3MW | 1 | **1** |
| 15 | Ollie | Smith | 5.1 | 2013-08-24 | 1.6 | 131 | secondary | 28268.0 | 2 | NA | 4 | 4OS | 2 | **1** |
| 23 | Harlan | O'Brien | 15.5 | 2003-04-12 | 2.1 | 233 | FALSE | 6904.7 | 1 | NA | 5 | 5HO | 1 | **1** |
| 26 | Hollis | Byrne | 76.0 | 1942-11-02 | 2.1 | 178 | secondary | 13856.7 | 2 | NA | 6 | 6HB | 2 | **5** |

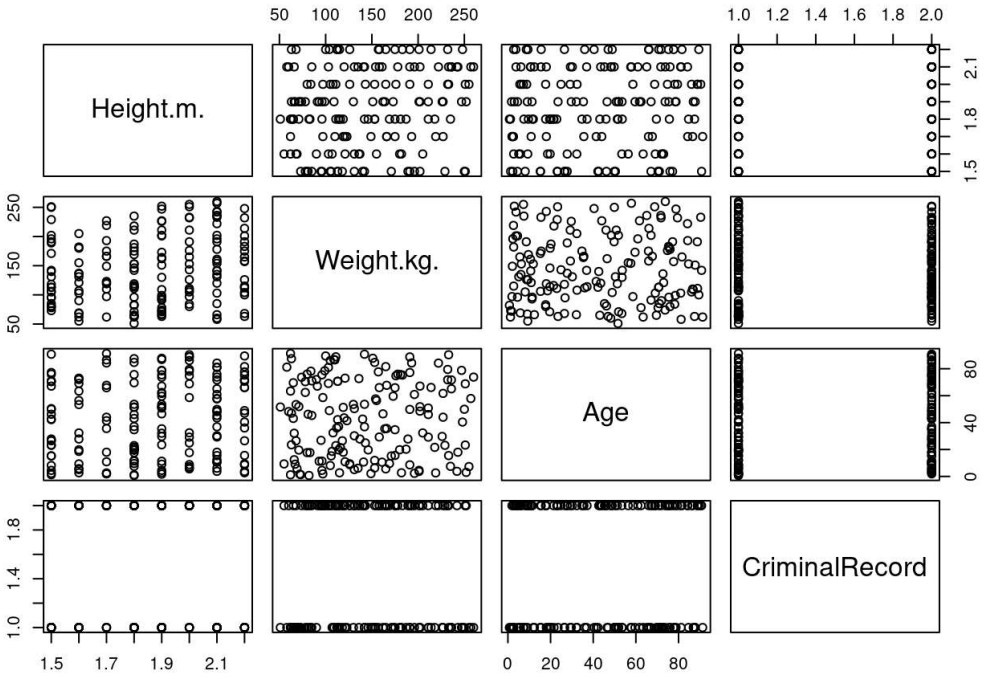
**Q8. You should filter the data by each age category. Generate a bar plot using ggplot2 for the criminal record variable.**

The two variables 'Age Range' and 'Criminal Record' are filtered into a dataframe, then a table created to record the frequency of 'Age Range' values for each level of 'Criminal Record', and finally a ggplot2 function histogram plot is created.




**Q9. You should generate an appropriate visualisation examining the relationships between height, weight, age and criminal records. Comment on this.**

The three variables are filtered into a new dataframe and each plotted against the other. These charts show the relationship between each variable and the other and appears to show a **null or no relationship** between any of the variables for this dataset.

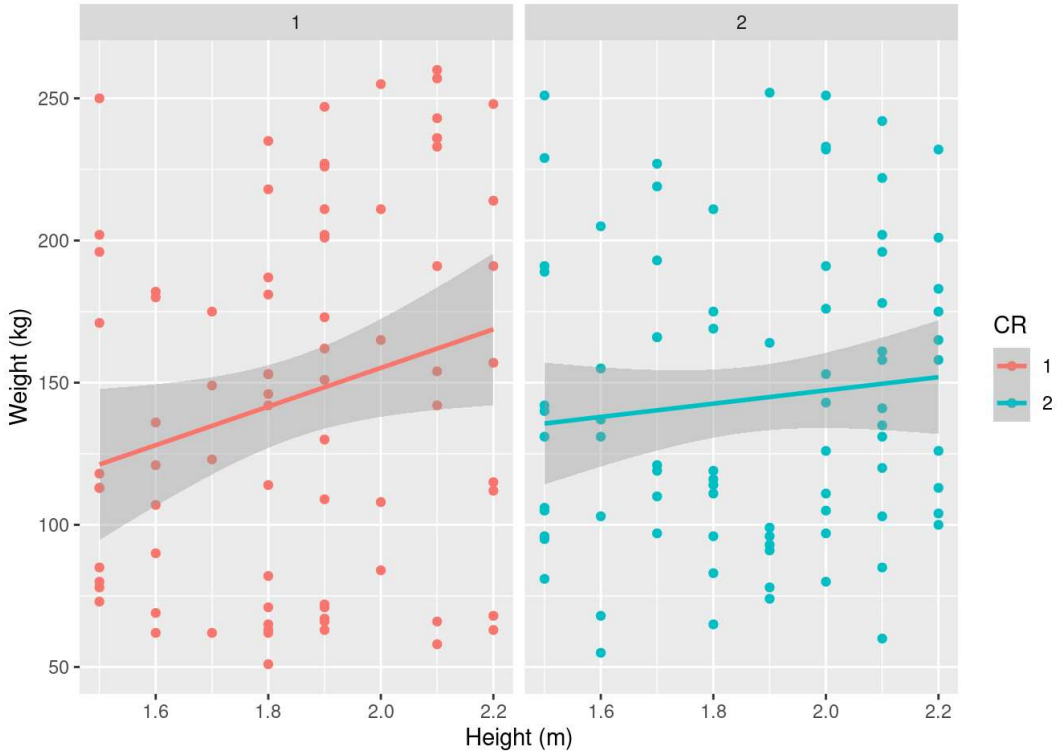



However, if we look more closely at each pair of variables, we see a different story.

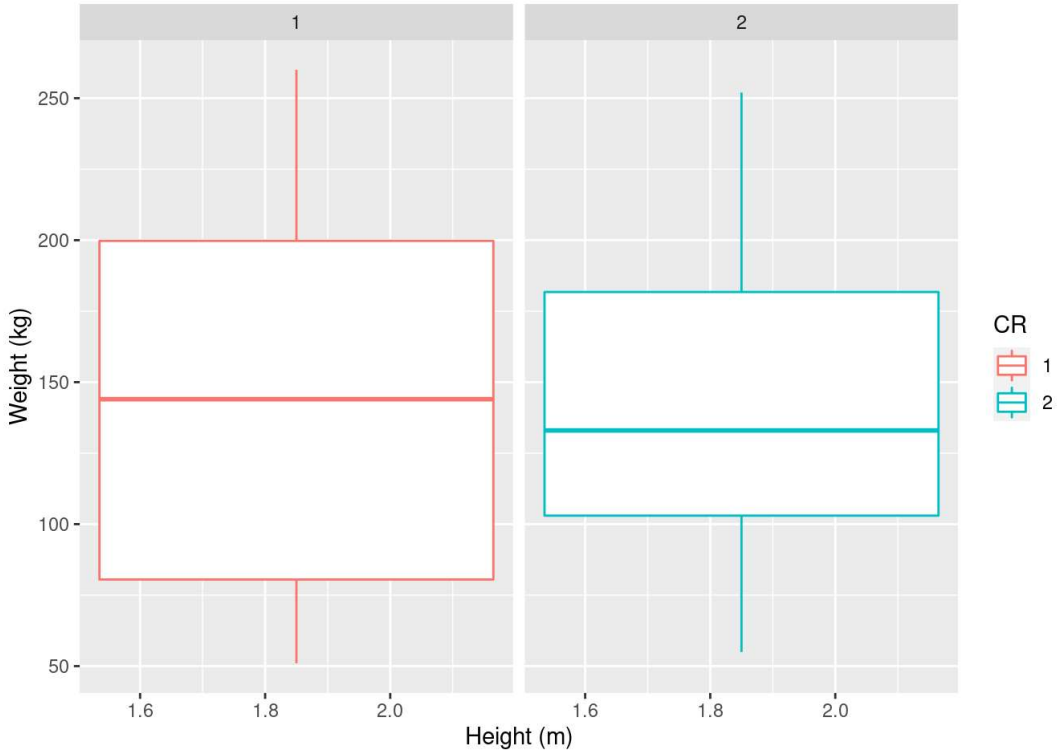
A **scatterplot with linear regression smoothing of 'Weight' v 'Height'** variables for each category of 'CriminalRecord (CR)' shows there is

**-** a **strong positive relationship** between 'Height' and 'Weight' for 'CR' = 1

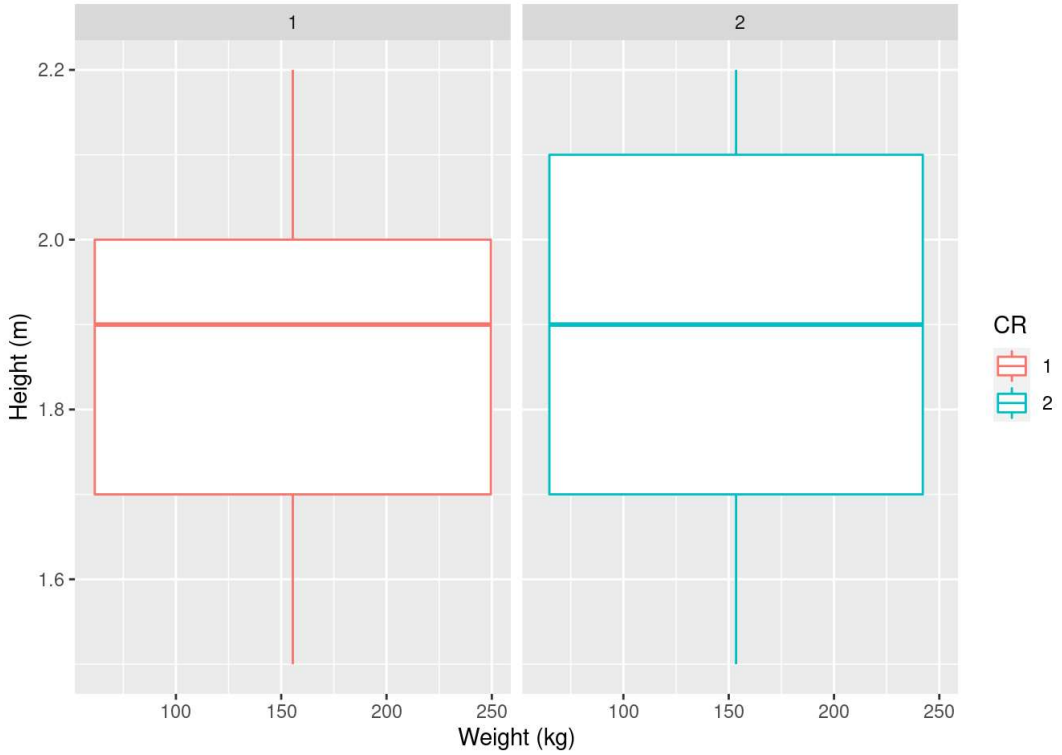
**-** a **weak positive relationship** between 'Height' and 'Weight' for 'CR' = 2



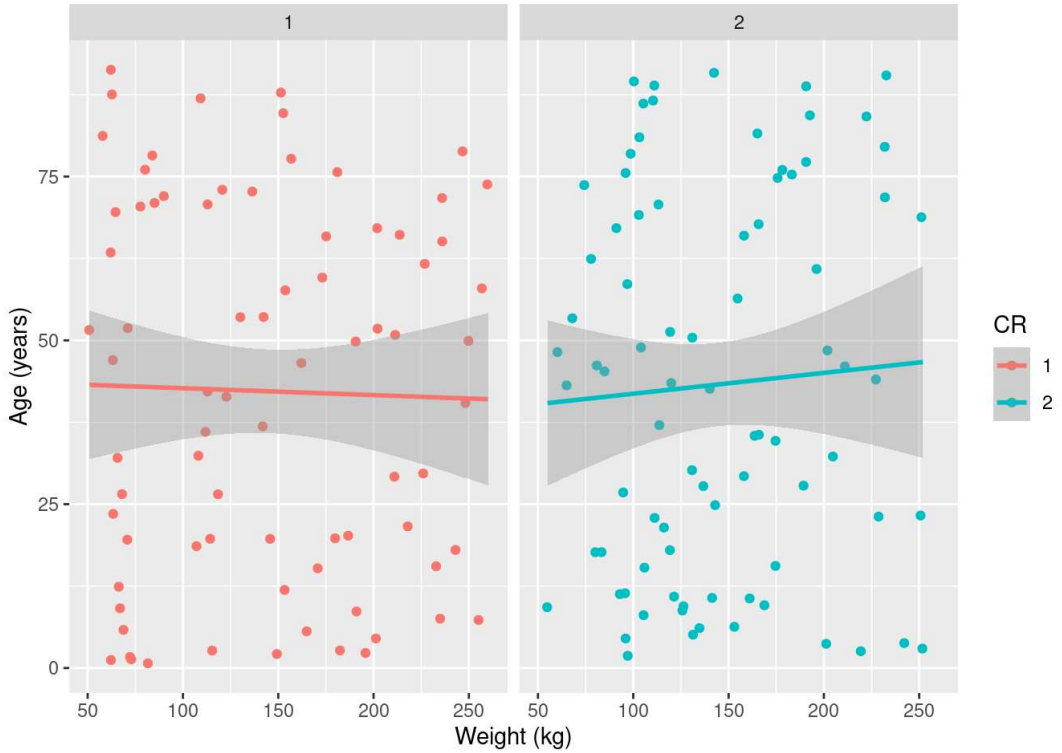

For a different view of the data (and same variables), a **boxplot** shows
- the **'Weight' overall range, IQR range median is greater for CR=1** than CR=2



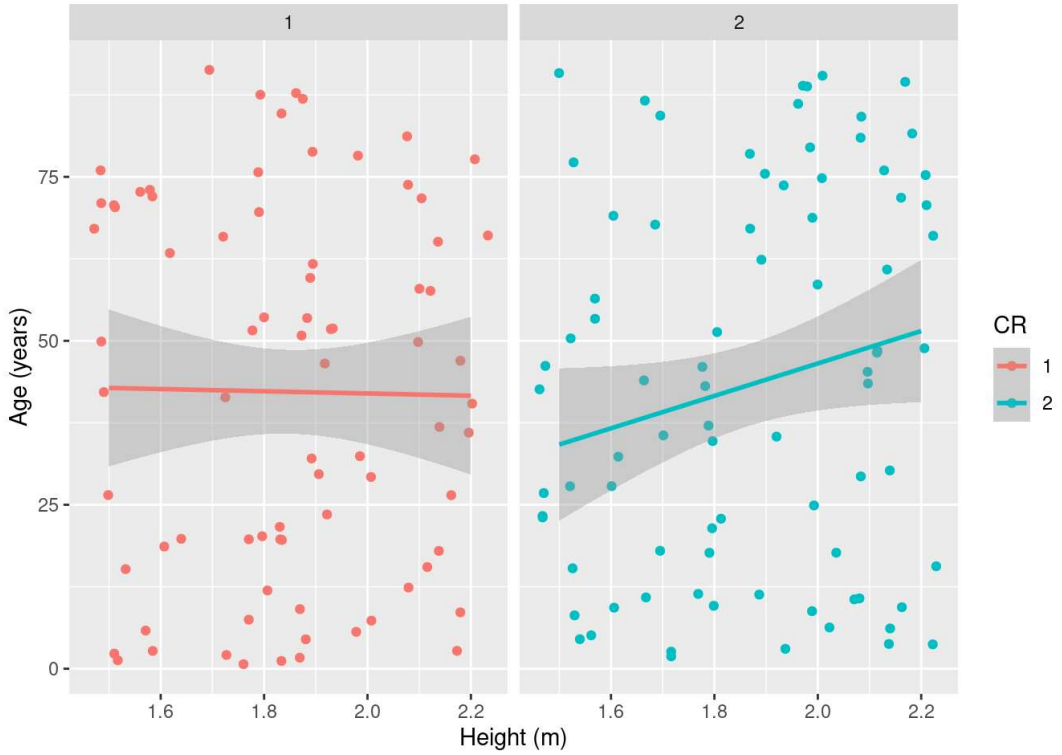

Also, a **boxplot of 'Height' v 'Weight'** variables for each category of 'CriminalRecord (CR)' shows
- the **'Height' overall range, IQR range median is greater for CR=1** than CR=2




A scatterplot of **'Age' v 'Weight'** variables for each category of 'CriminalRecord (CR)' shows
- a **weak negative relationship** between 'Weight' and 'Age' when 'CR' = 1
- a **weak positive relationship** between 'Weight' and 'Age' when 'CR' = 2



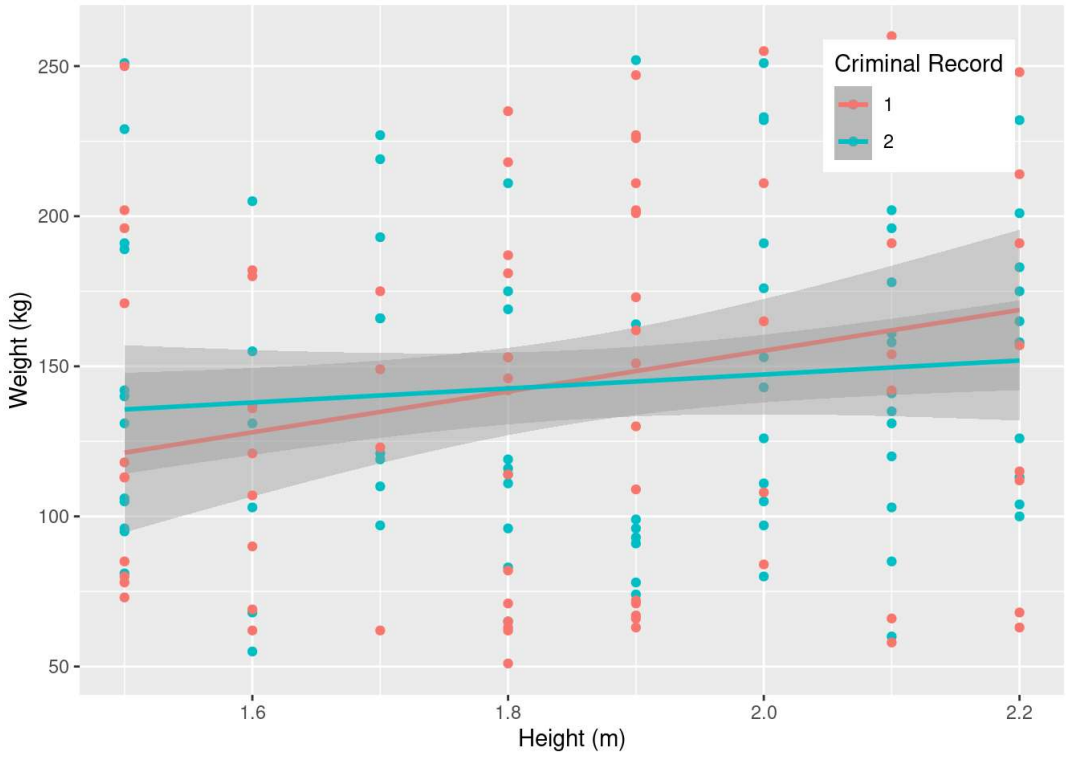

And a scatterplot of **'Age' v 'Height'** variables for each category of 'CriminalRecord (CR)' shows
- a **weak negative relationship** between 'Height' and 'Age' when 'CR' = 1
- a **strong positive relationship** between 'Height' and 'Age' when 'CR' = 2




**Q10. Using filters, you should analyse if there are any interesting results in the dataset regarding the relationships between height, weight and criminal record. Use appropriate visualisations.**

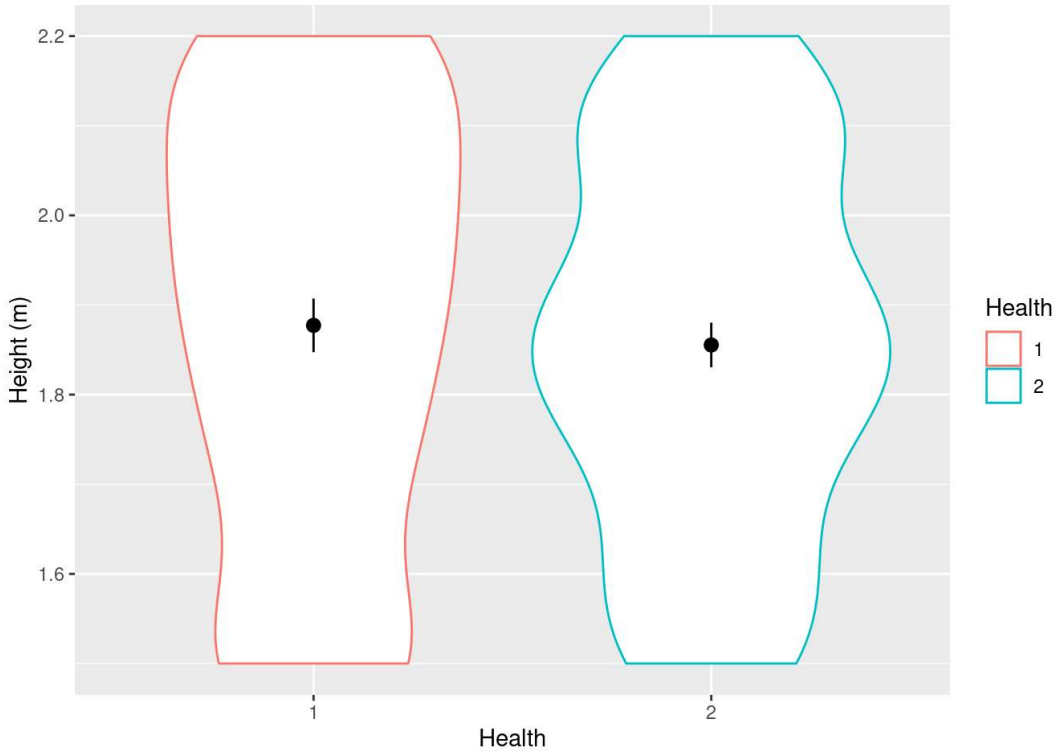
The dataset was filtered on 'Criminal Record' into two separate dataframes and then plotted as a scatterplot with linear regression and this plot shows

- a **weak positive relationship** between 'Weight' and 'Height' when 'CR' = 1
- a **strong positive relationship** between 'Weight' and 'Height' when 'CR' = 2

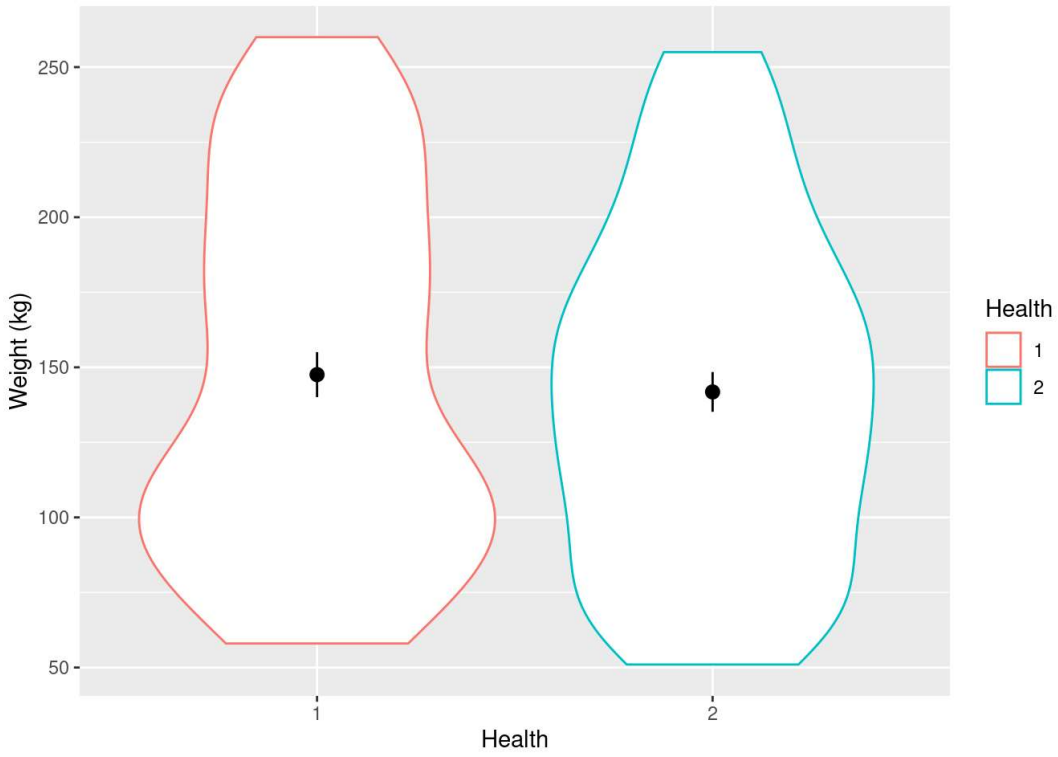



**Q11. Generate a smaller data frame for the subjects where health related data is available. Examine if there is a relationship between the different states of health and height, weight or age. Use appropriate visualisations. Note this should include a modelling type analysis such as regression. (S.Weisberg. Applied Linear Regression. Wiley Series in Probability and Statistics, 2005. may be useful)**

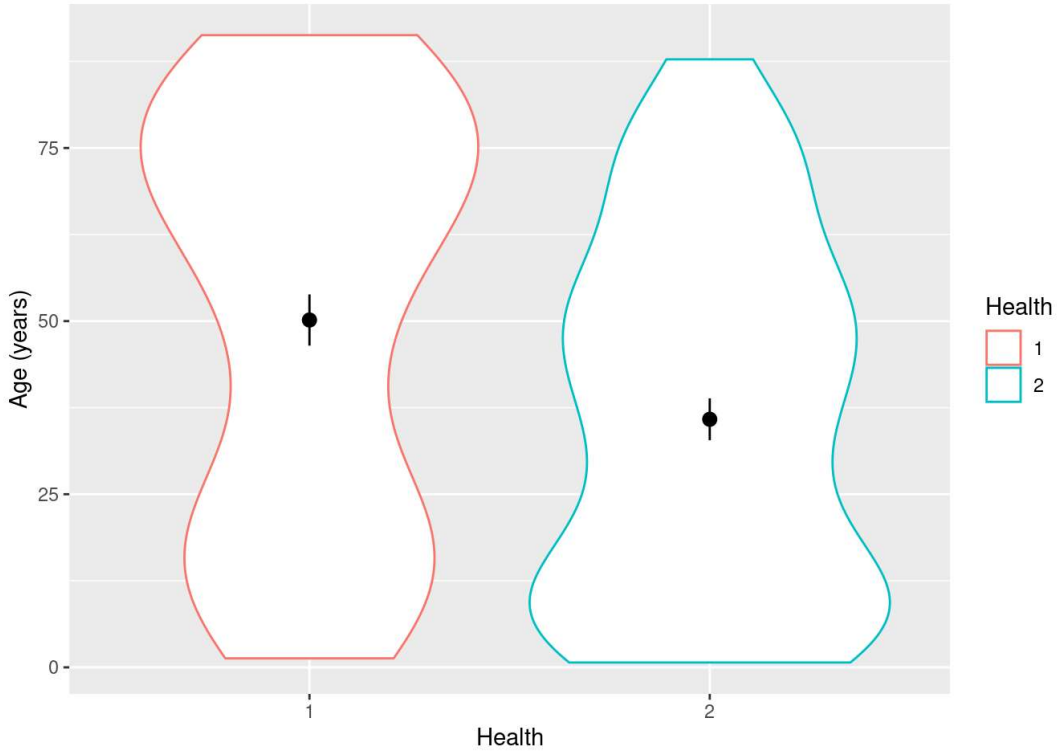
A **violin plot of 'Health' v 'Height'** variables shows
- an uneven distribution of data points for both Health categories
- weak relationship; more tall (in range 1.8-2.2m) subjects with Health = 1
- weak relationship; more subjects with mid height than tall or short with Health = 2



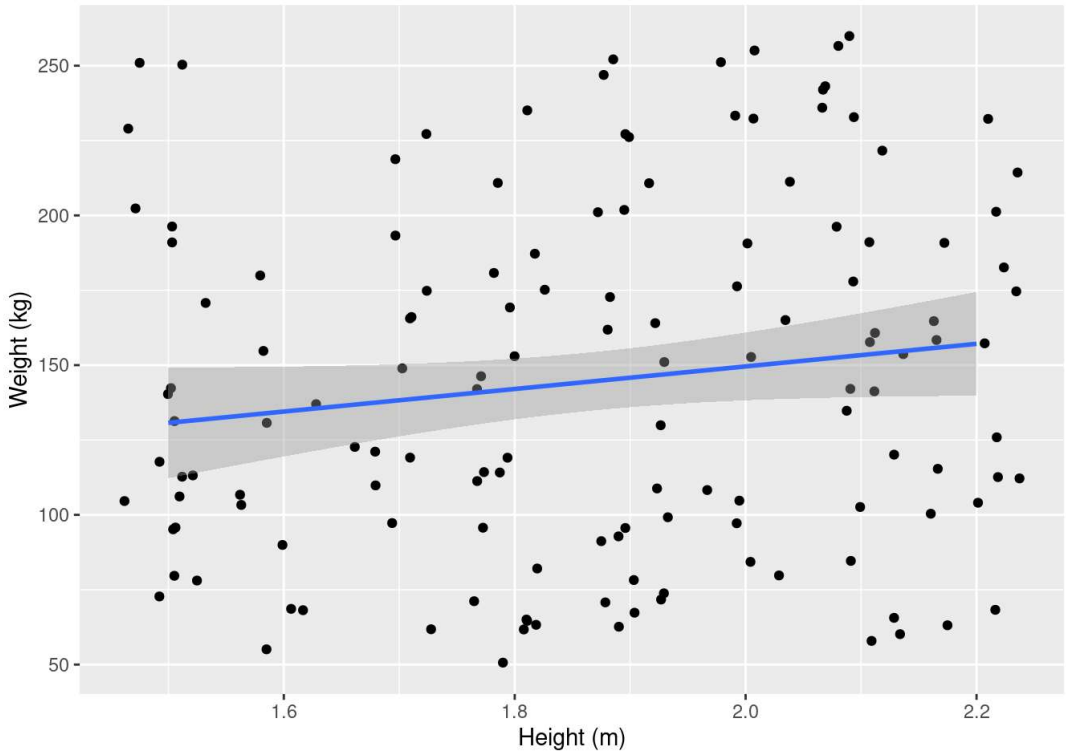

A **violin plot of 'Health' v 'Weight'** variables shows
- an uneven distribution of data points for both Health categories
- weak relationship; more light (in range 60-75kg) subjects with Health = 1
- weak relationship; more subjects with mid weight than heavy or light with Health = 2



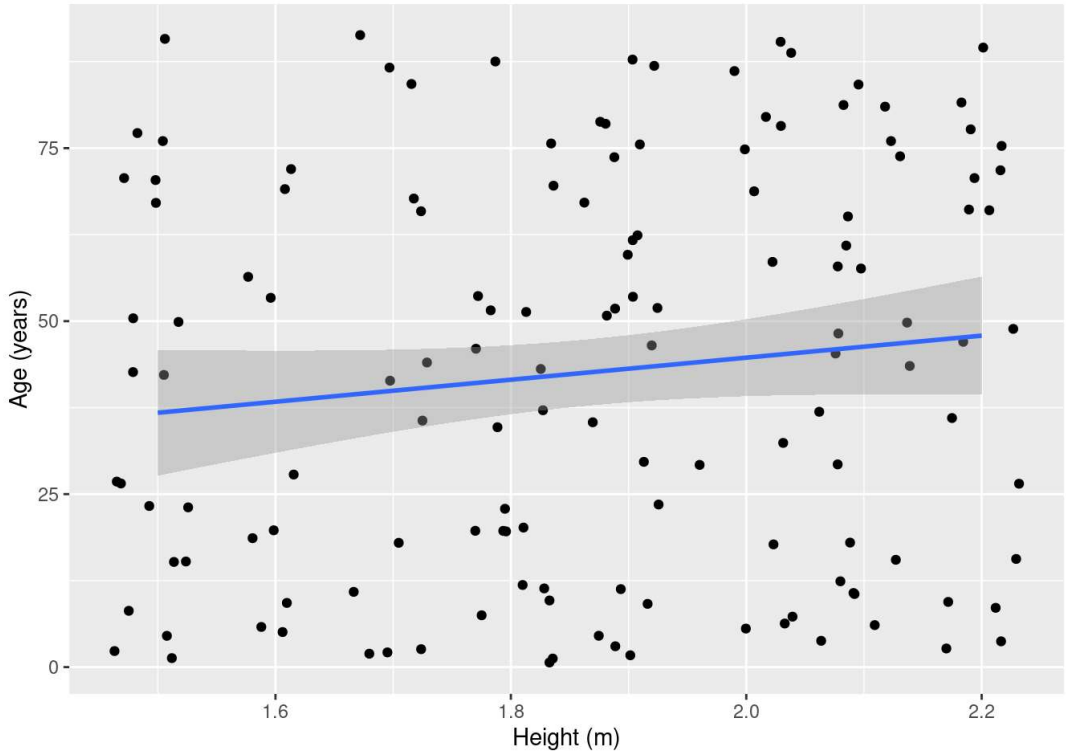

A **violin plot of 'Health' v 'Age'** variables shows
- an uneven distribution of data points for both Health categories
- positive linear relationship; more old (in range 50-75years) subjects with Health = 1
- negative linear relationship; more young (in range 0-25) subjects with Health = 2



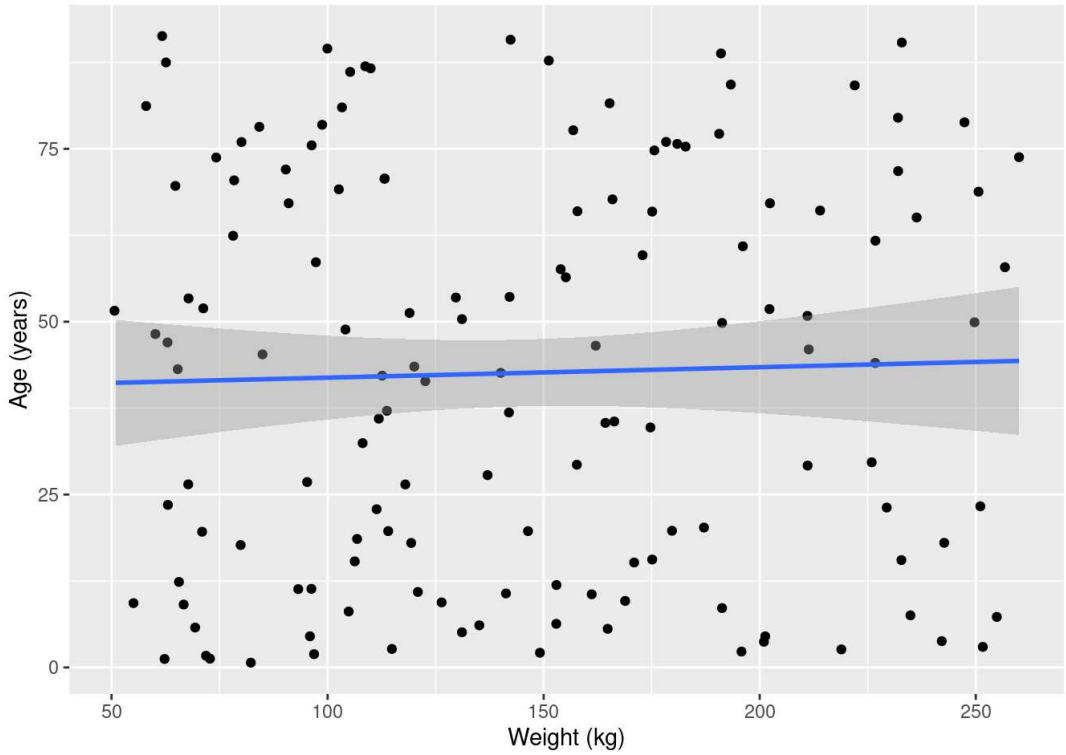

A **scatterplot of 'Height' v 'Weight'** variables with linear regression shows
- moderate positive linear relationship; ie taller subjects are heavier (no surprise there)




A **scatterplot of 'Height' v 'Age'** variables with linear regression shows
- moderate positive linear relationship; ie taller subjects are older

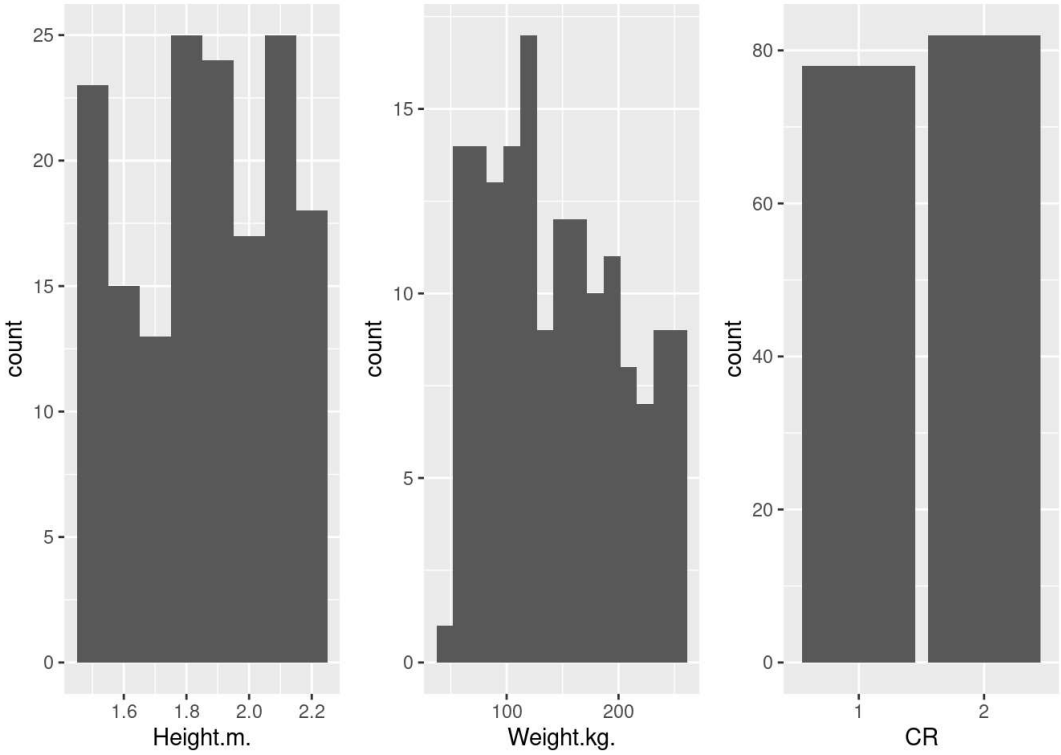



A **scatterplot of 'Weight' v 'Age'** variables with linear regression shows
- null or no relationship




# Additional Interesting Findings

A histogram for each of the three variables shows the data is not uniform and may not reflect the general population at large.




A histogram for 'each of the three variables'Education Level' shows the data that may not reflect the general population at large. The assumption that most male (dataset has men only) criminals come from a lower socio-econonomic background with lower education levels **does not appear to be true for this dataset.**