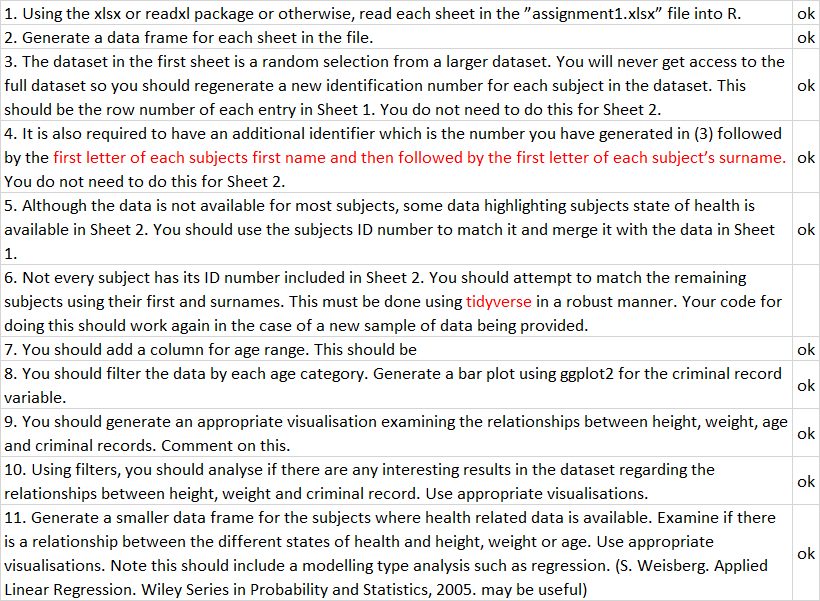
**STAT8010: Intro to R for Data Science**

**Lecturer:** Dr Justin McGuinness

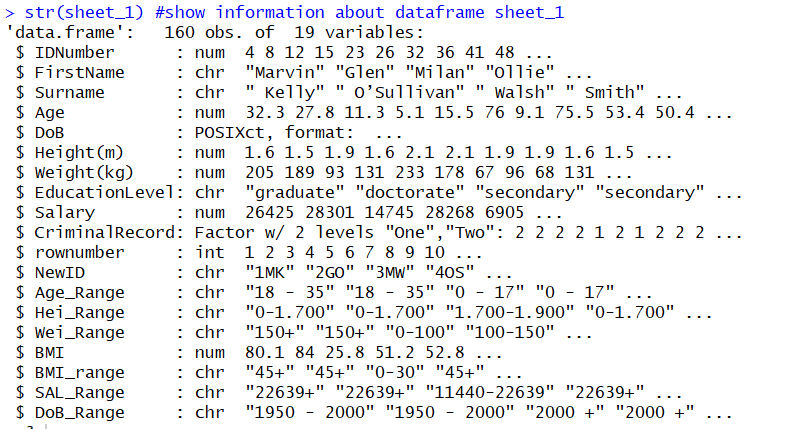
**Student**: Givaldo Francisco da Silva Junior R00182595

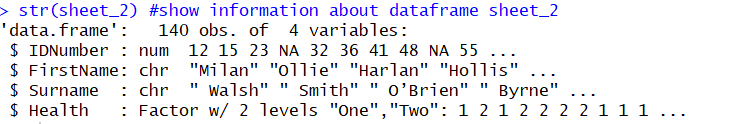
**STAT8010 Assignment 1 - Due Monday 4th November (Week 8) at 9am.**

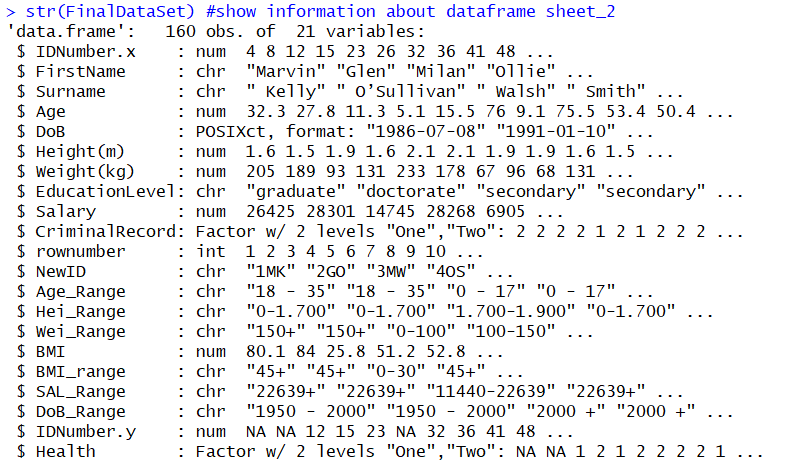
“Three interesting things I learned about R during this assignment”



**Datasets**



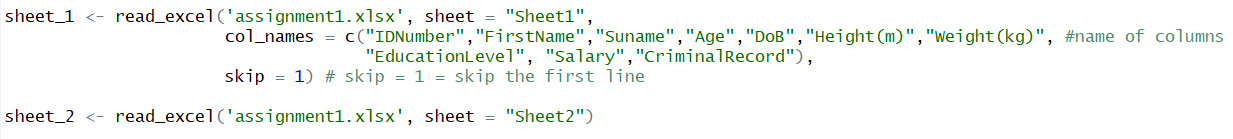


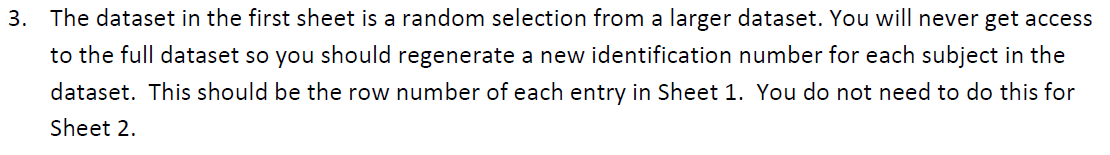


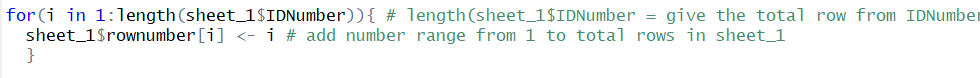


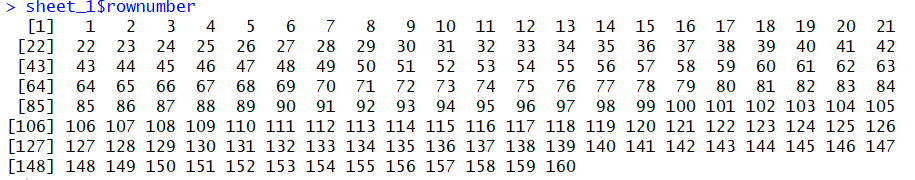


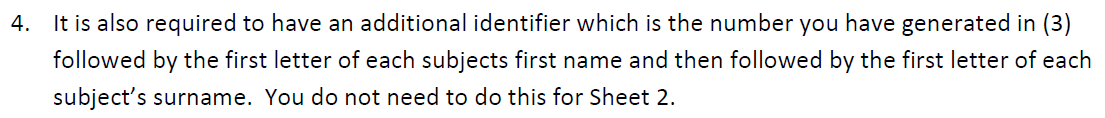


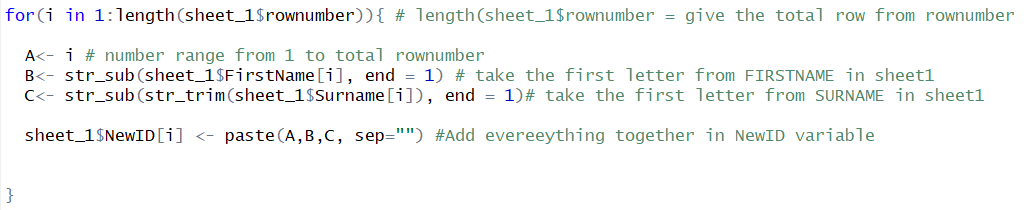


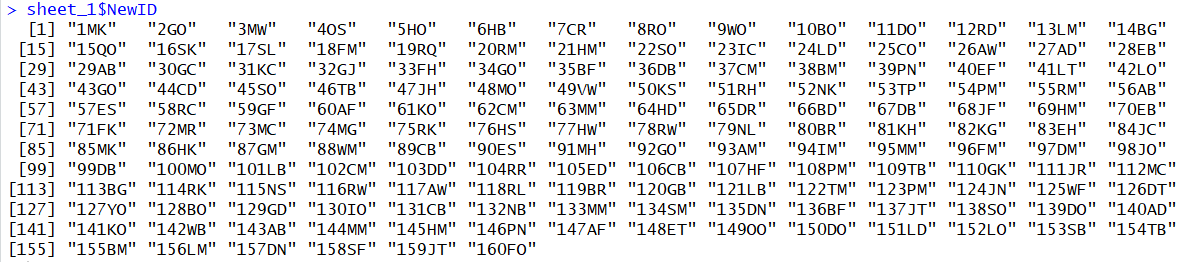


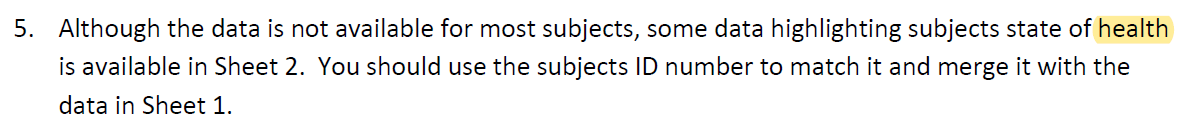




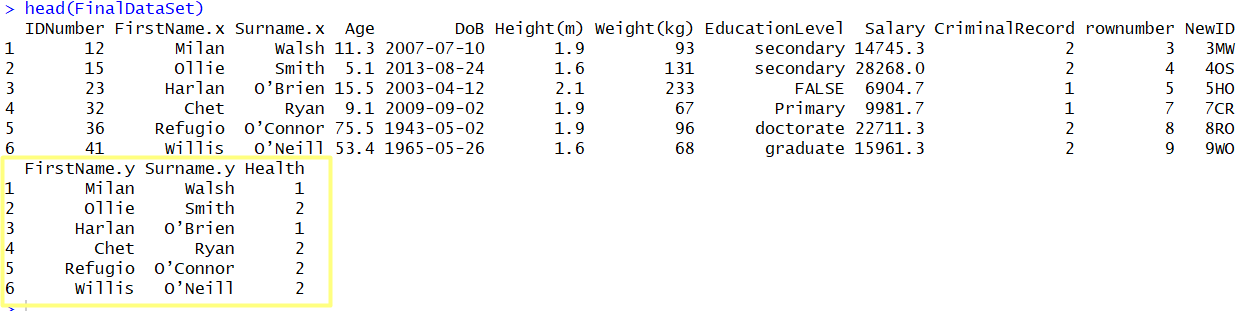


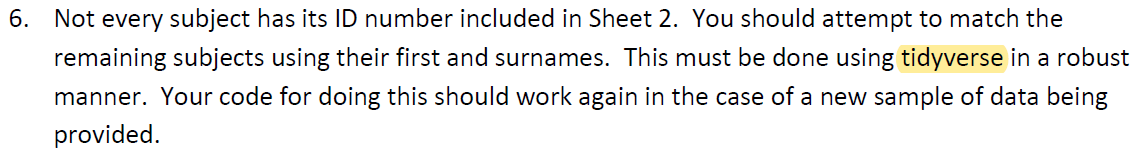




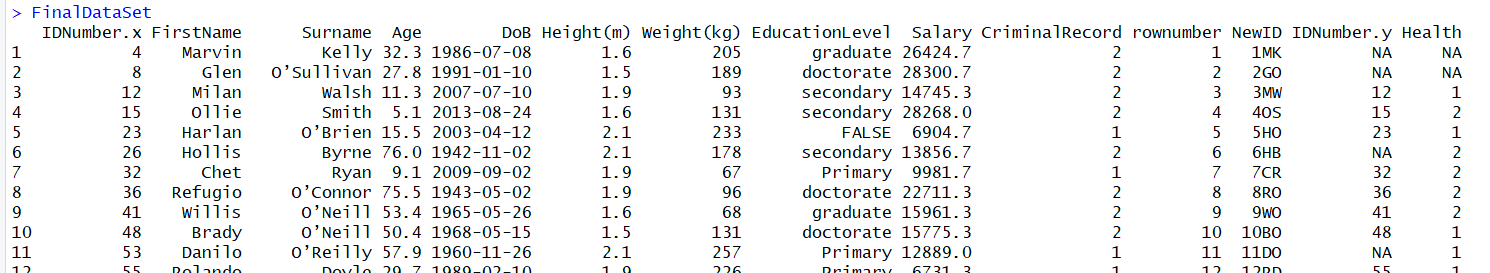


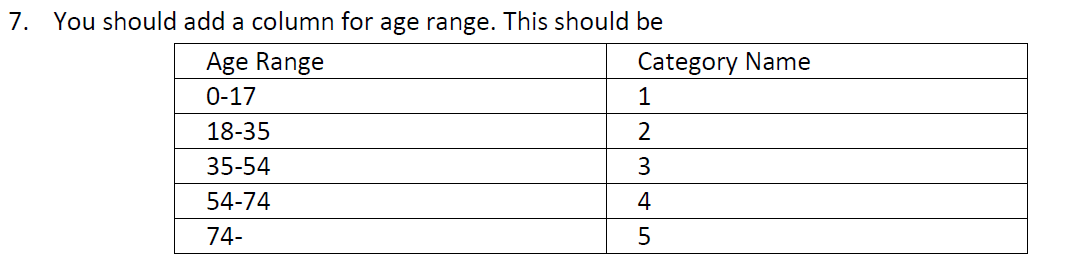


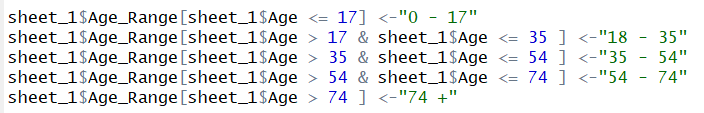




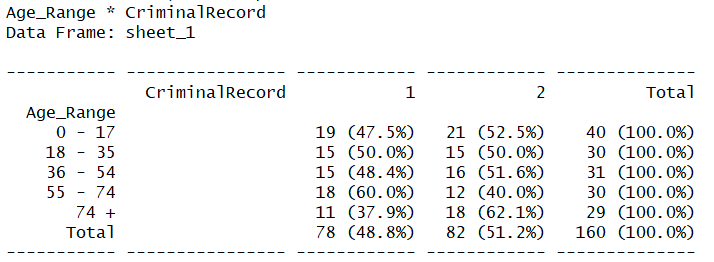


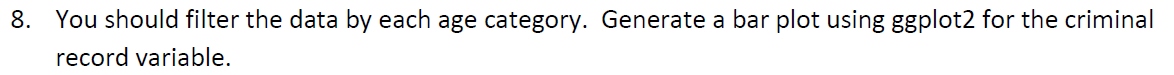






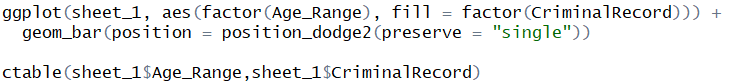


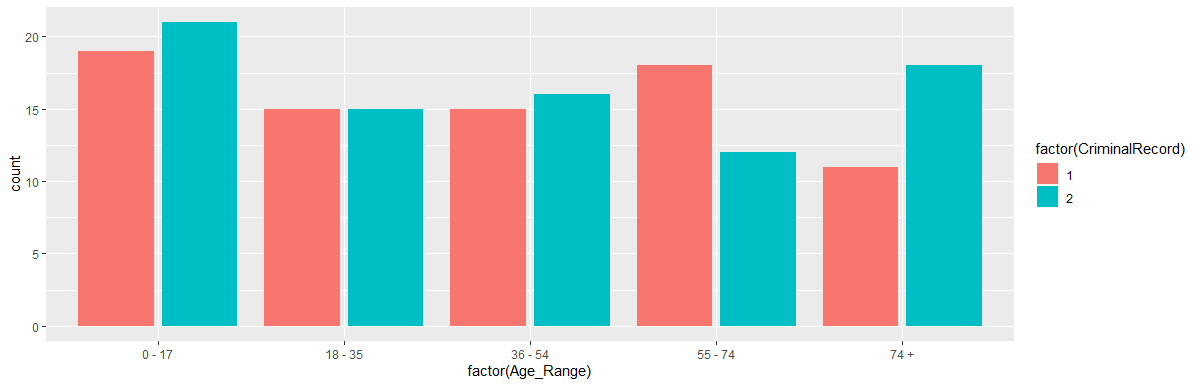




There is a significant difference between the criminal record in age more than 74 and in the opposite direction in age between 54 and 74.

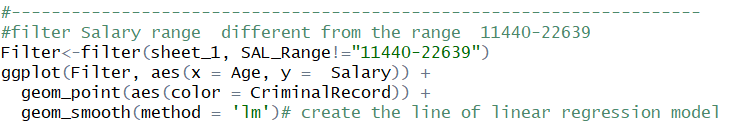
At other ages criminal records are similar

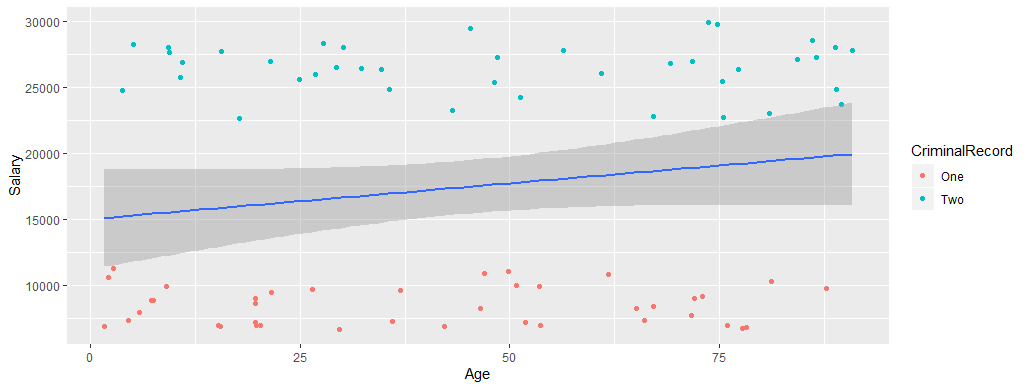




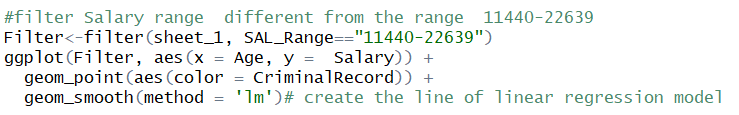
Follow some additional observation with Age compared with some other variable and Criminal record compared with some other variable.

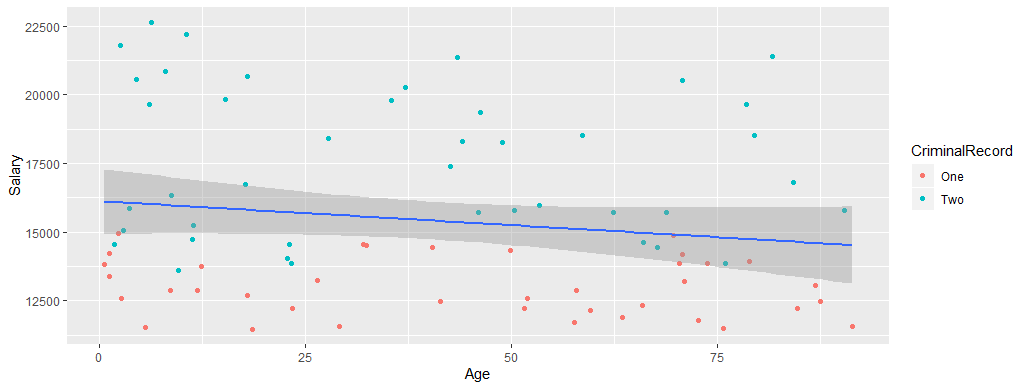
Comparing the Criminal record with the salary above and below the range (11440-22639) we can see a different linear correlation. In the first graph, there is a positive correlation e the sample are very separated.





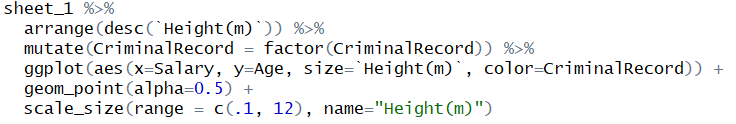
In the graph below, the mean salary, there is a negative correlation and the samples are lightly mixed, but still having a division between One and Two.

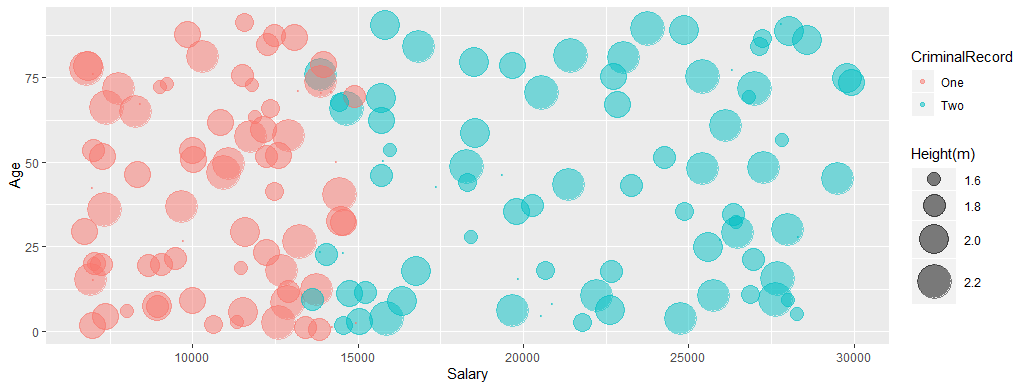


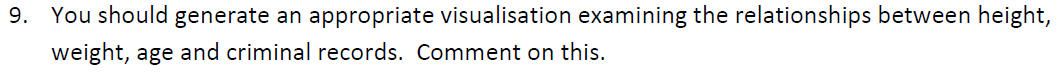


In the graph below there is a clear division by salary versus Criminal record.

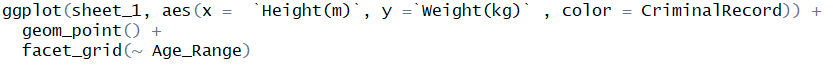
The lowest salary is grouped by criminal record = 1 and on the opposite side the highest salaries are grouped by criminal record = 2

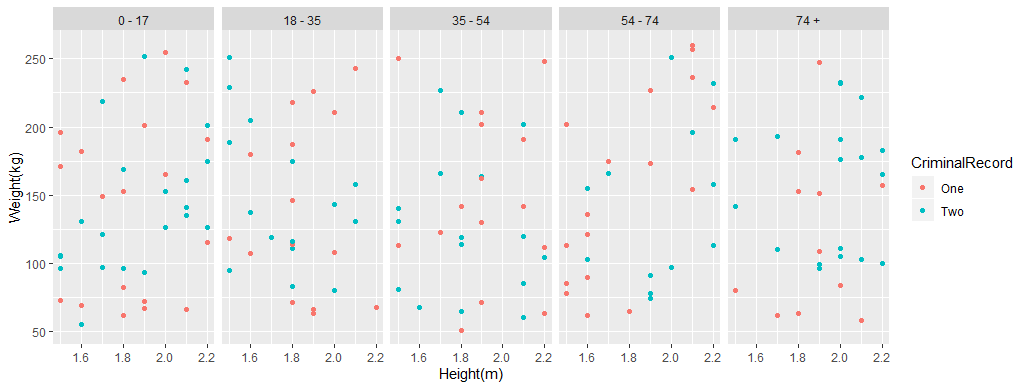




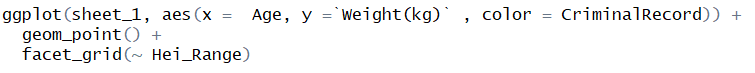


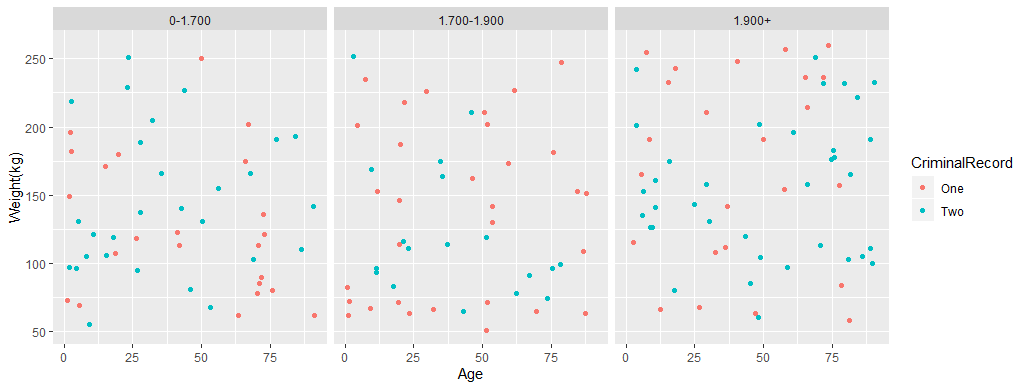
This first graph comparing the variable Height, Weight, Age, and Criminal record, all together, does not present so much information, we cannot see clearly any grouping of samples related to the criminal record.



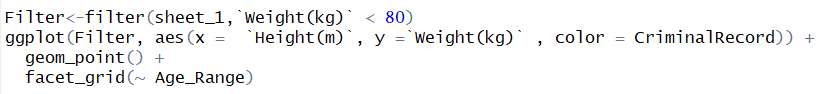


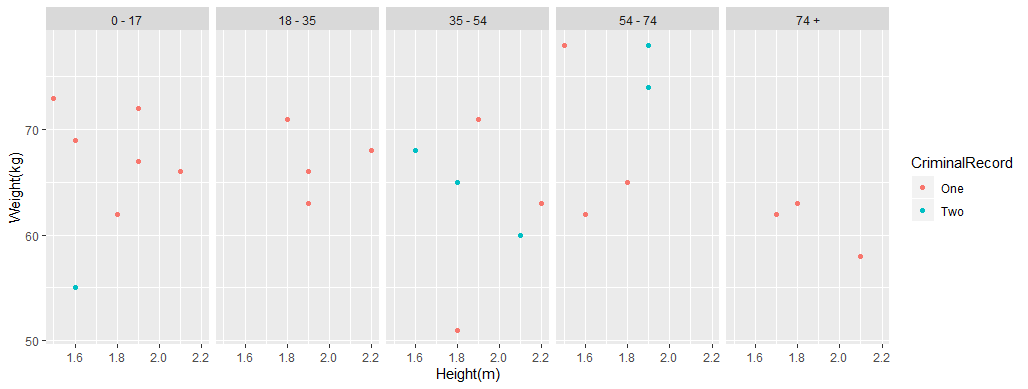
In the next graph it is possible to notice there is an accumulation of criminal record “One” on the samples heavier than 200kg and less than 80 kg approximately.



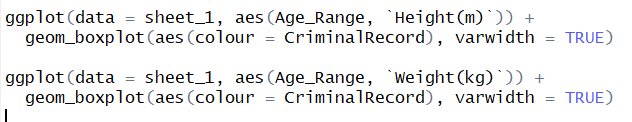


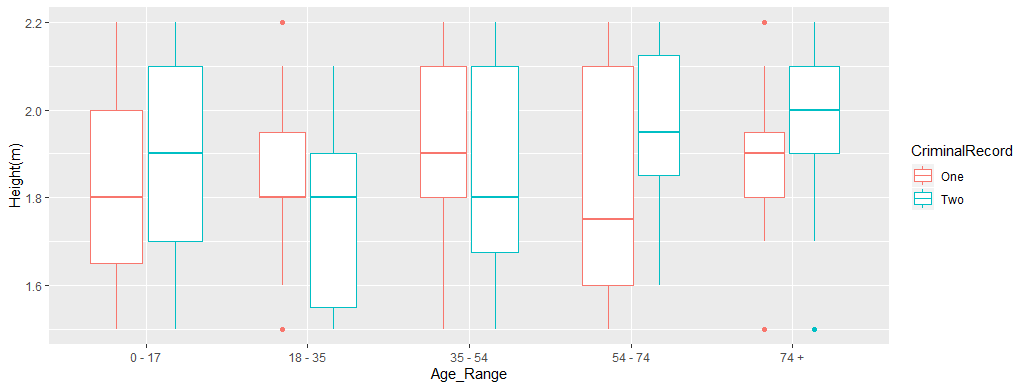
In this slice of the data most samples with Weight below 80 kg are in Criminal record one.

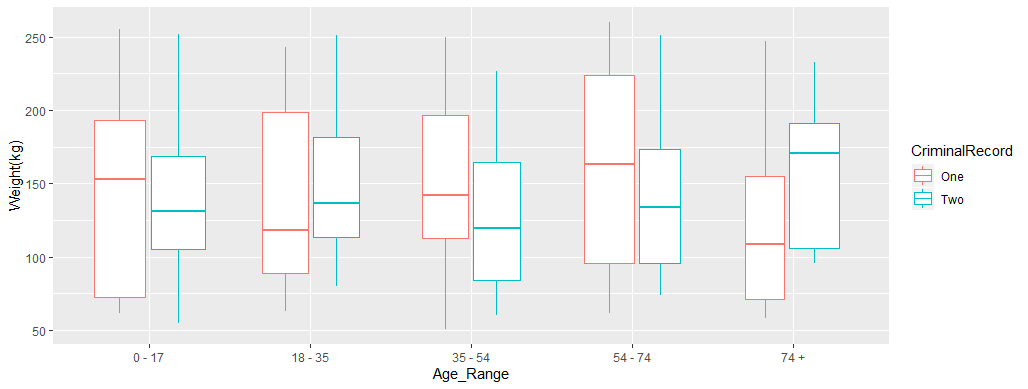




In the next two graphs it is possible to see a tendency in Height and Weight only in the criminal record “Two”, as much older the samples "Two" get, heavier and taller they are.

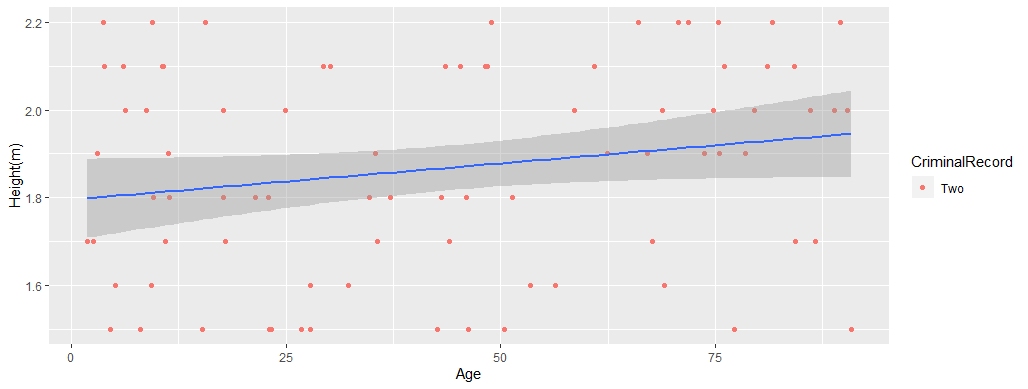




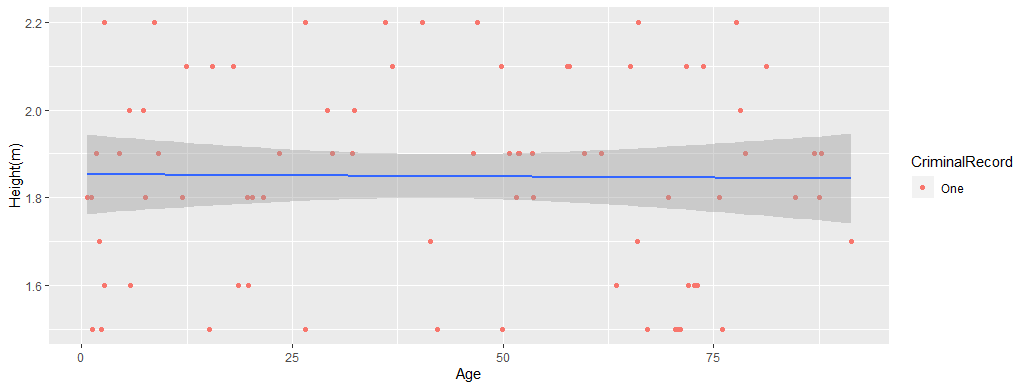


In the next two graphs there is a comparison between criminal records “One” and “Two”.

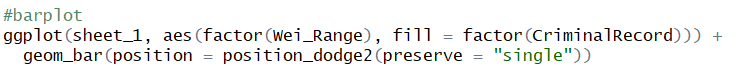
That is not an expressive difference, but Criminal record – “Two” is apparently taller than samples with criminal record = “One” as the previous graphs showed.

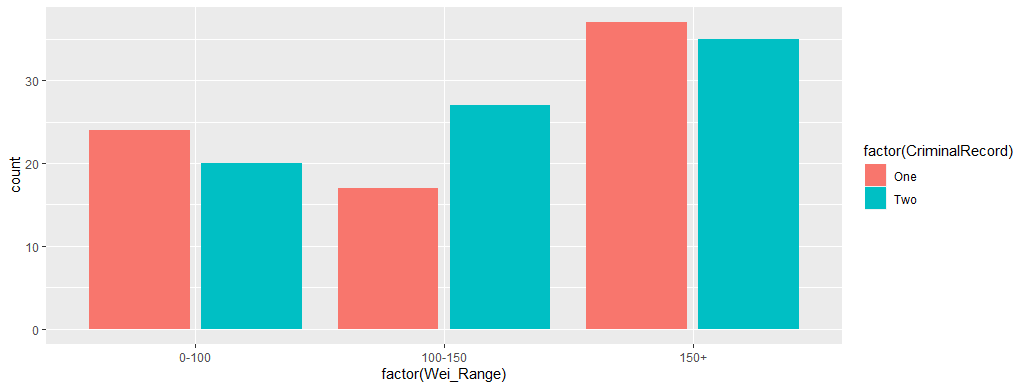


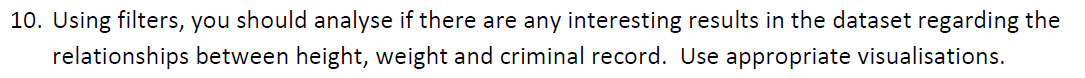
The criminal record “One” does not show the same positive correlation, it seems there is no correlation.



Comparing criminal records and groups of weight, the biggest difference is in the group “100-150”.

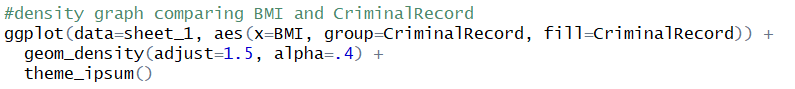


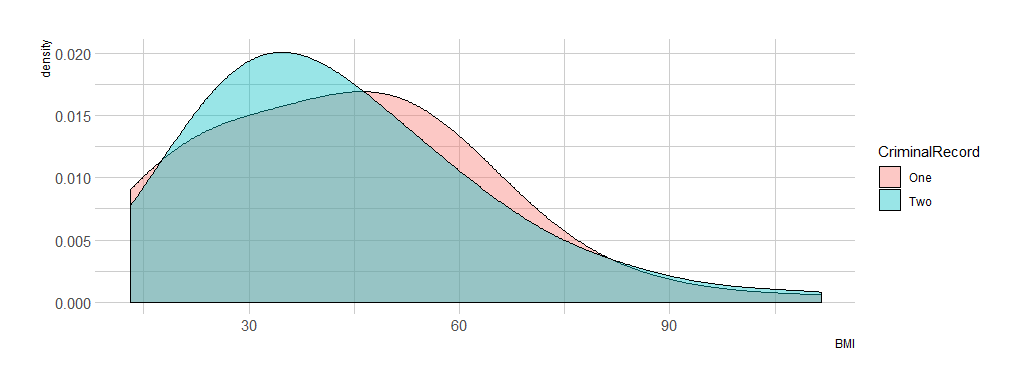




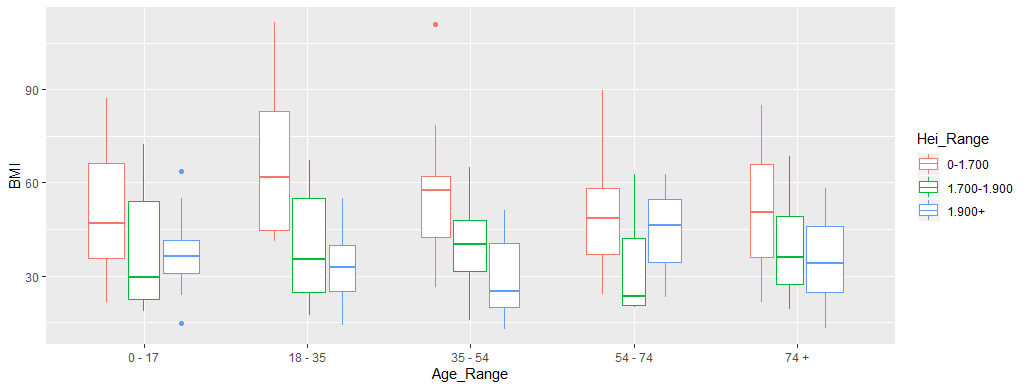
# Body Mass Index (BMI) is a person’s weight in kilograms divided by the square of height in meters, that gives you your ideal weight.

Comparing Criminal record vs BMI, it shows a right-skewed distribution, most of criminal record “two” are below 50 BMI level and criminal record “one” a little higher BMI level.



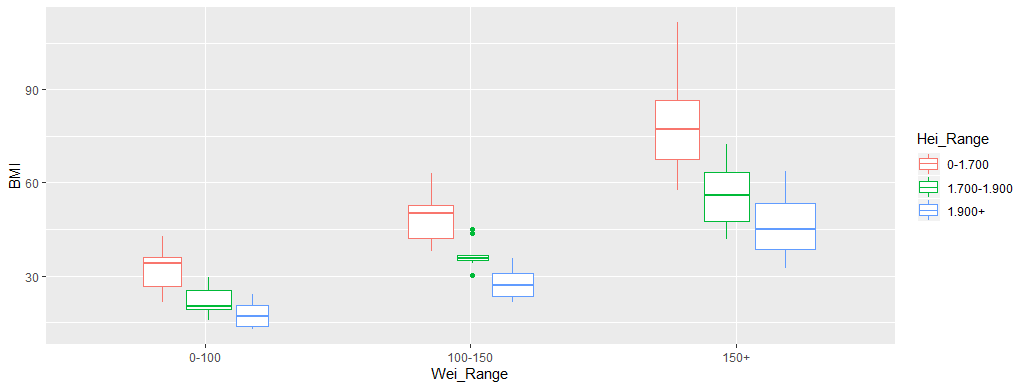


Samples smaller than 1700cm have the highest level of BMI, this makes sense because if the sample is small and heavy, it will result in a high level of BMI

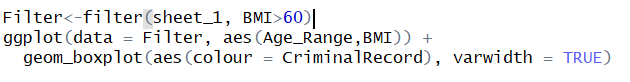


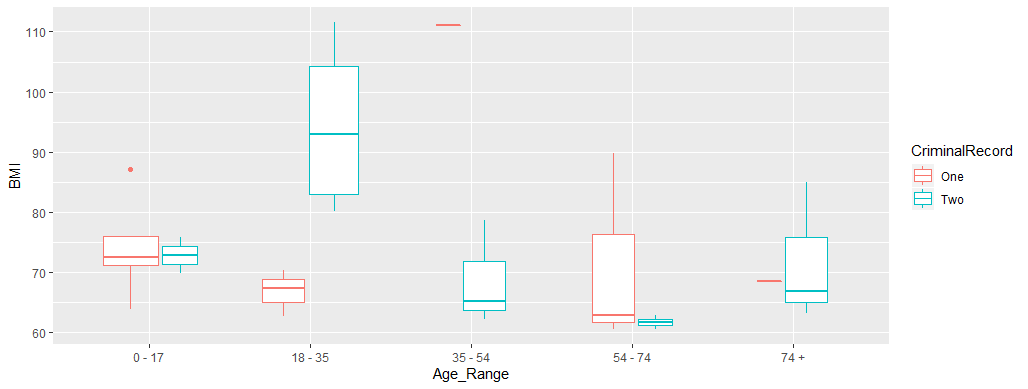
It is possible to observe that the highest level of BMI is concentrated in samples smaller than 1700 cm and heavier than 150kg, this is considered morbid obesity.



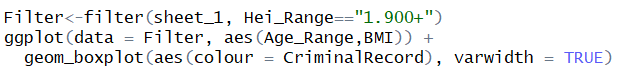


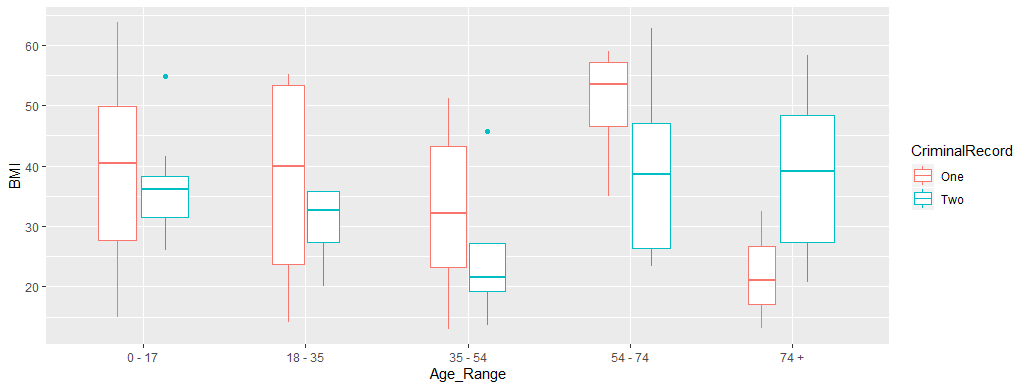
Looking at this specific group with BMI level higher than 60, it is possible to see samples in age between 18 to 35 having criminal record two, have the highest level of BMI.



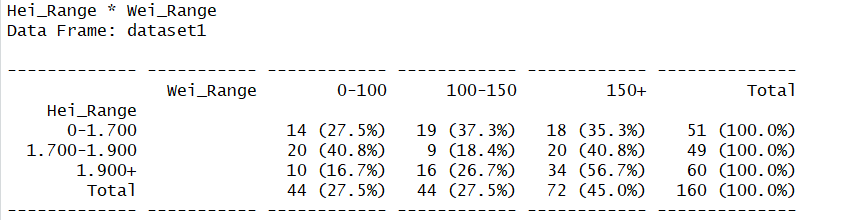


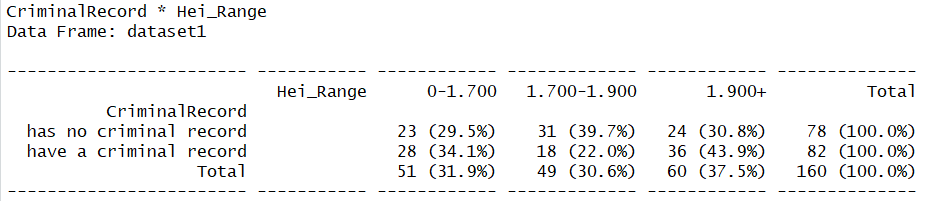
Collecting only samples from Height range more than 1900 cm it is possible to notice that criminal record “one” has the highest level of BMI in almost all ages, only samples older than 74 that level is lower compared with a criminal record “two”.

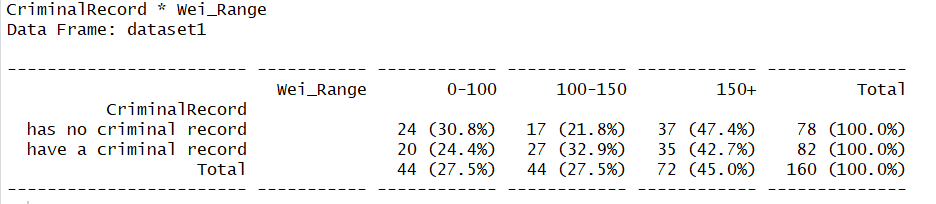




Almost half population (45%) have more than 150 kg

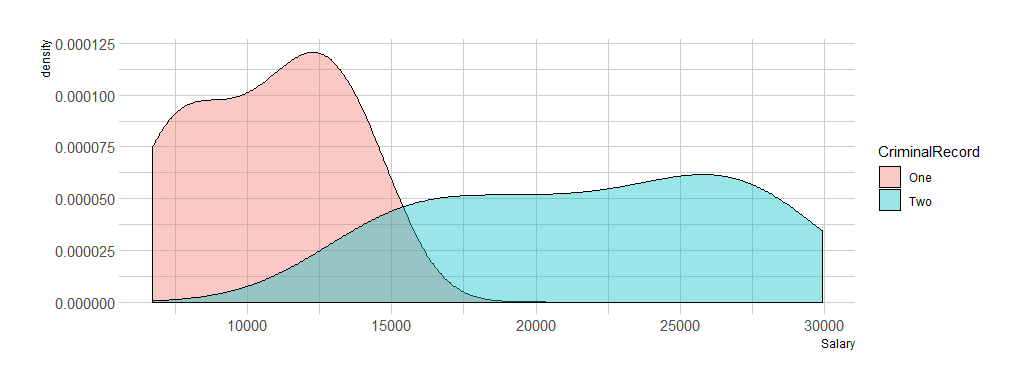




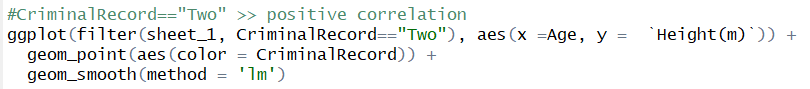


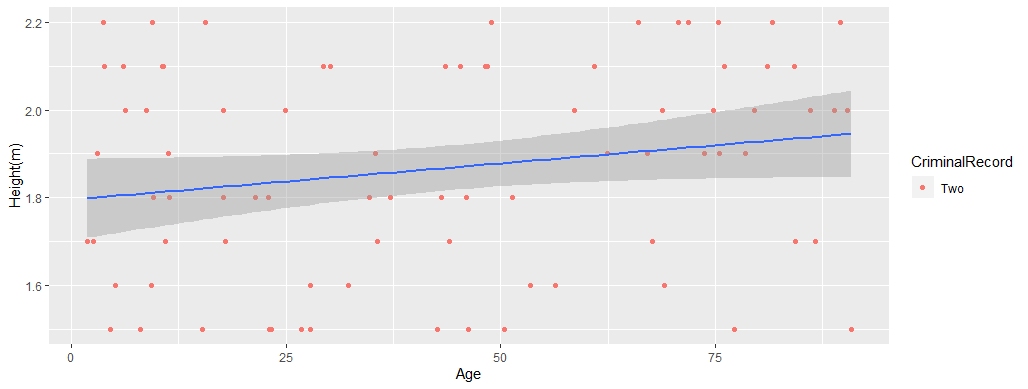
Comparing criminal record vs Salary the density graph presents clearly how criminal record “one” and “two” are separated by Salary.



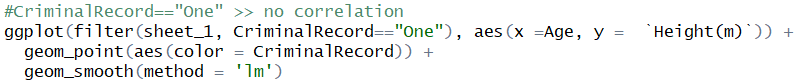


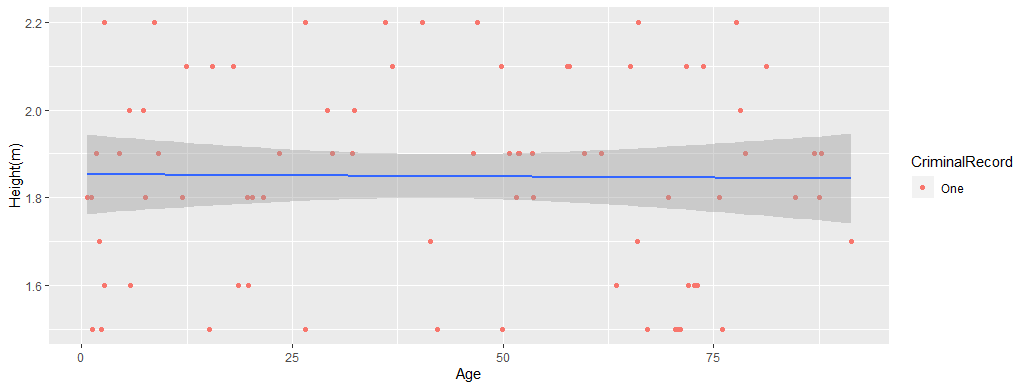
Filter samples that only have a criminal record “Two”, there is a positive correlation.



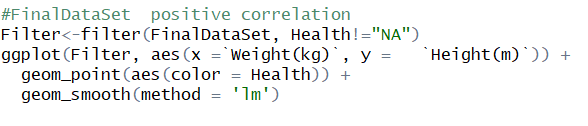


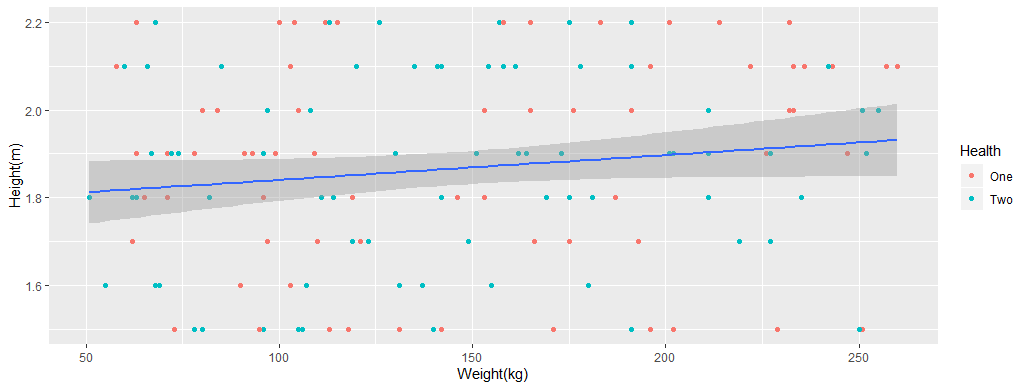
Filter samples that only have a criminal record “One”, there is no correlation.

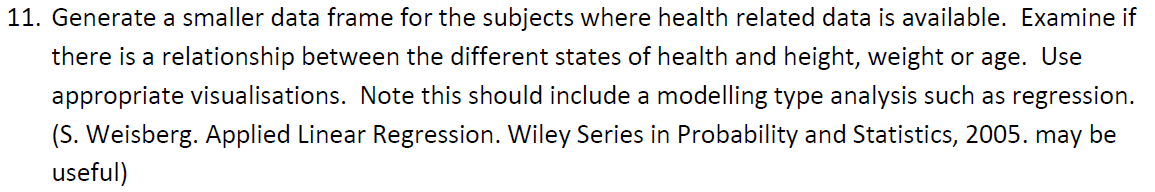


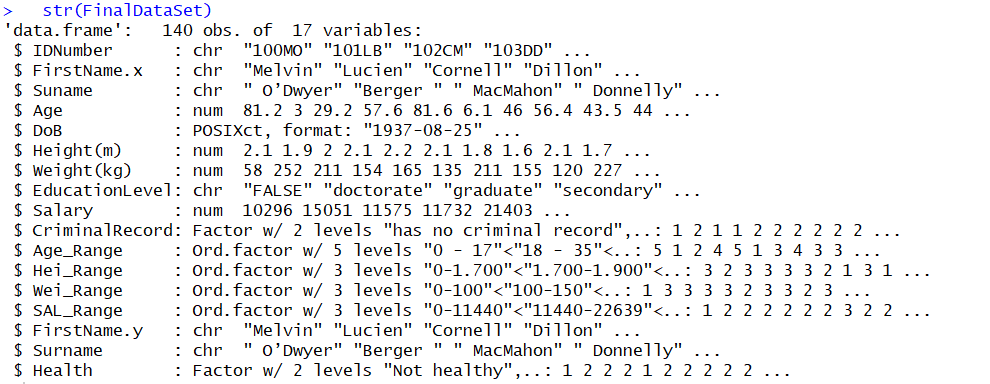


Filter samples by Health show a positive correlation between Height and Weight in Health One and Two.

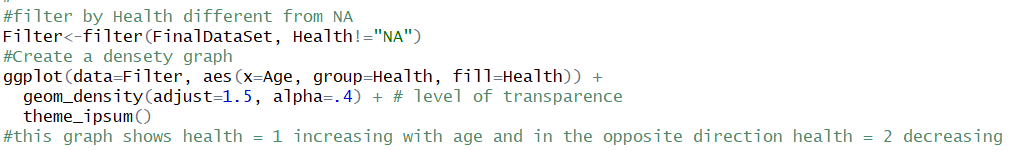


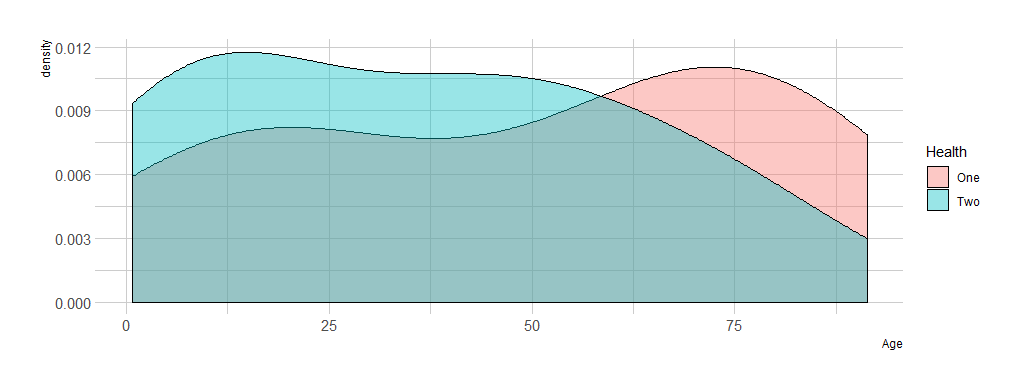




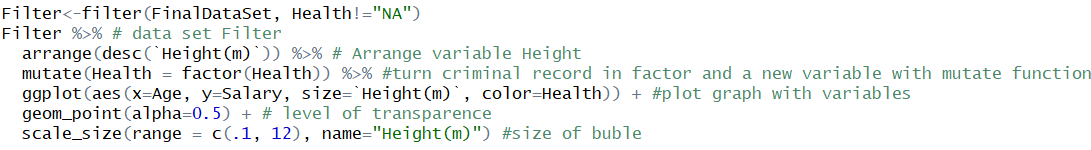


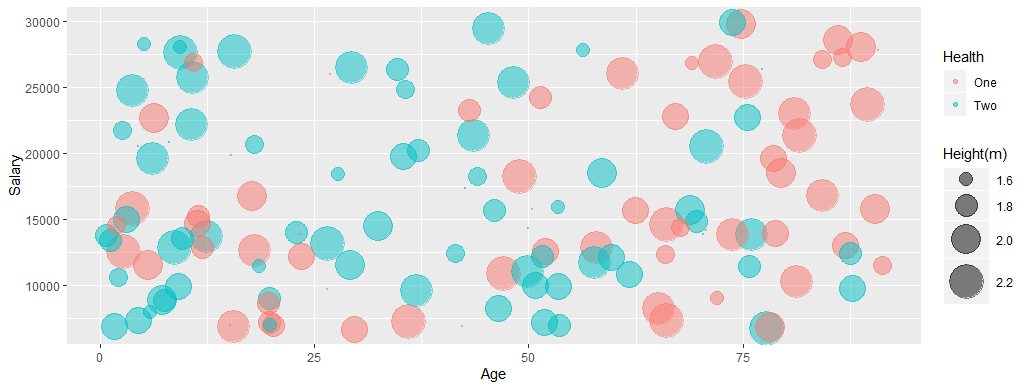
Variable Health “one” increases as the samples get older and in the opposite direction health “Two” decreased.

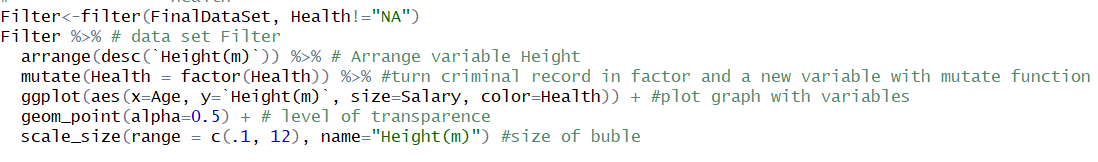


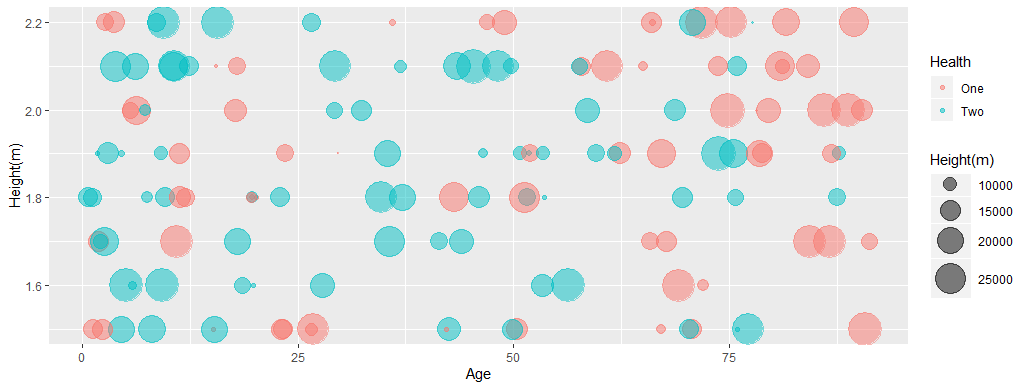


It is clear there is a higher proportion of samples Health = “one” grouped in ages more than 60 approximately.

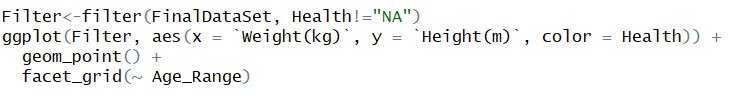


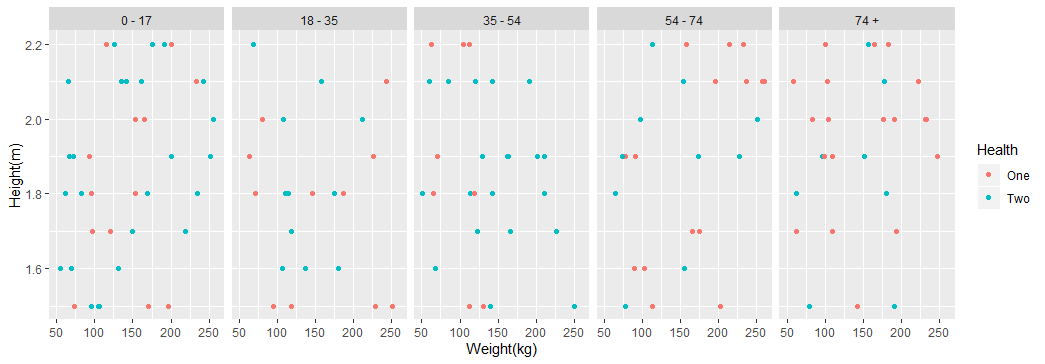




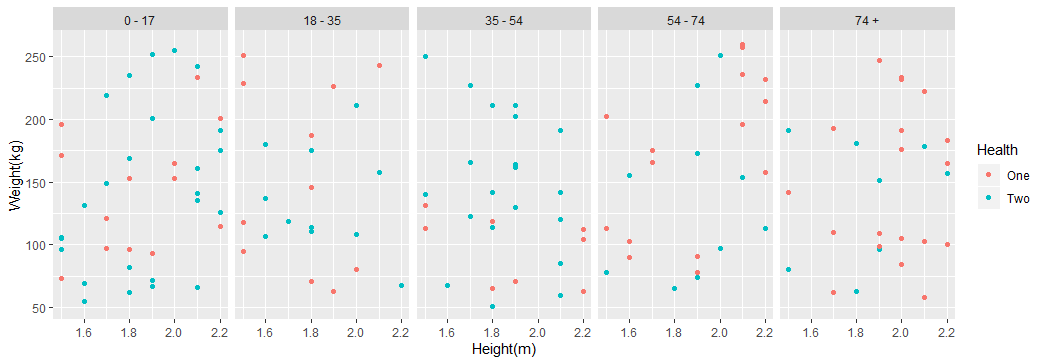
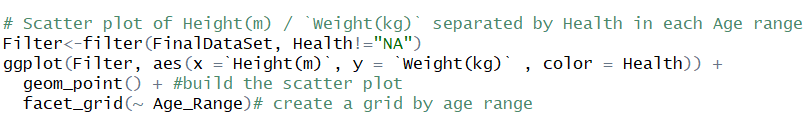


People between 36 and 54 years old are grouped in Health “two”.

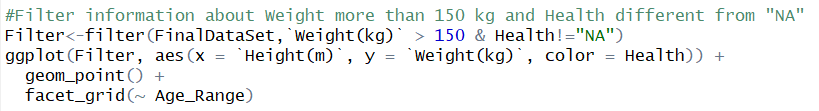


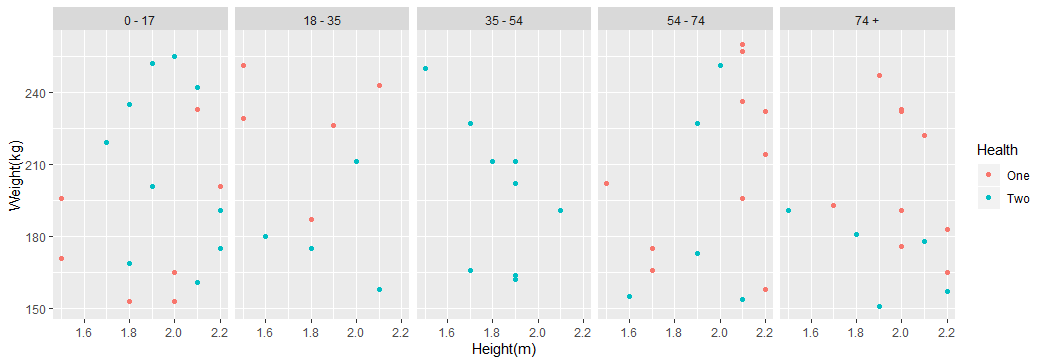


There is a group of health “two” in samples above 150 kg and in ages between 35 and 54.

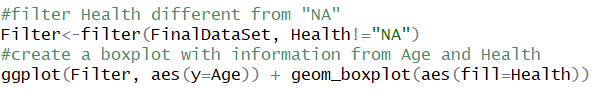


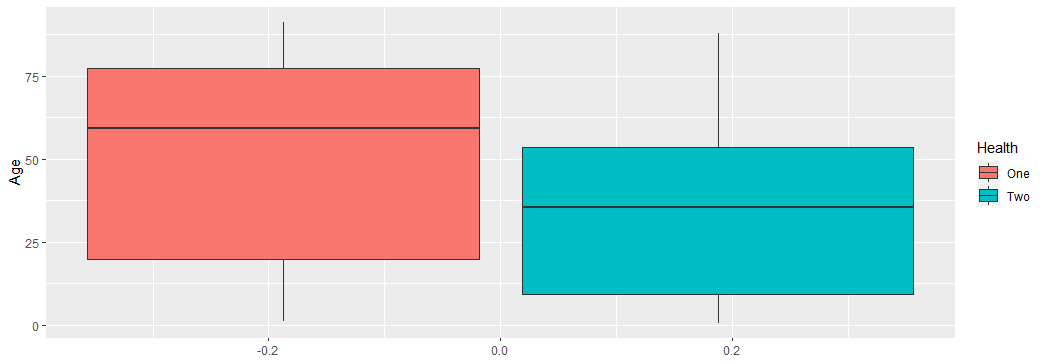
If that specific group is selected (samples above 150 kg) it is possible to verify that between ages 35 and 54 there are just Health “two” samples.



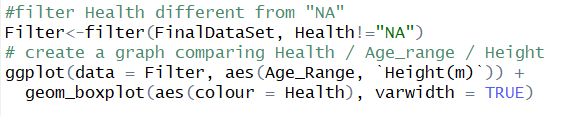


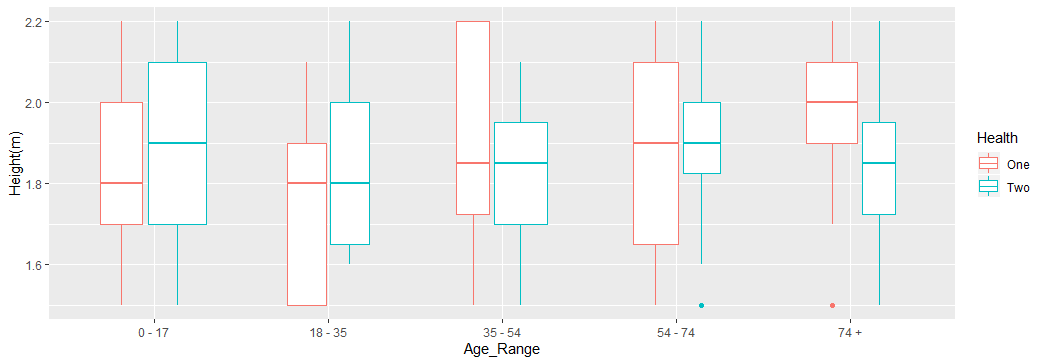
This graph shows a relative difference between Health “one” and “two” compared with the Age variable.



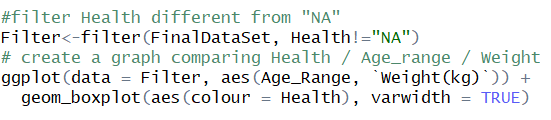


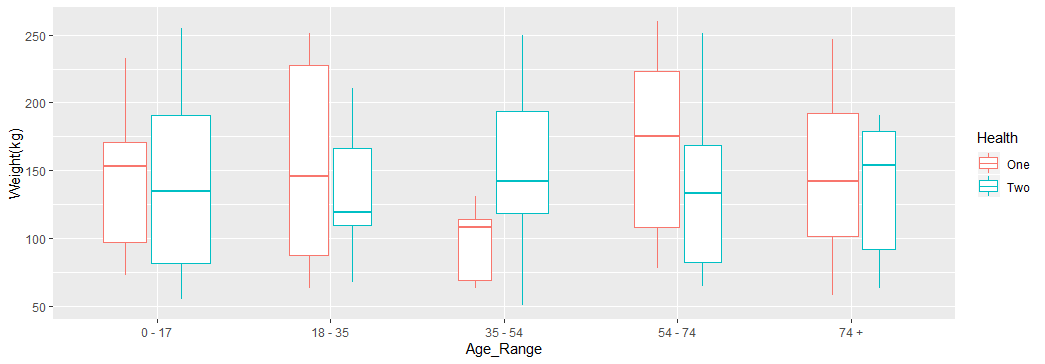
Interesting situation in the graph below, three ages range have the same mean of Height.



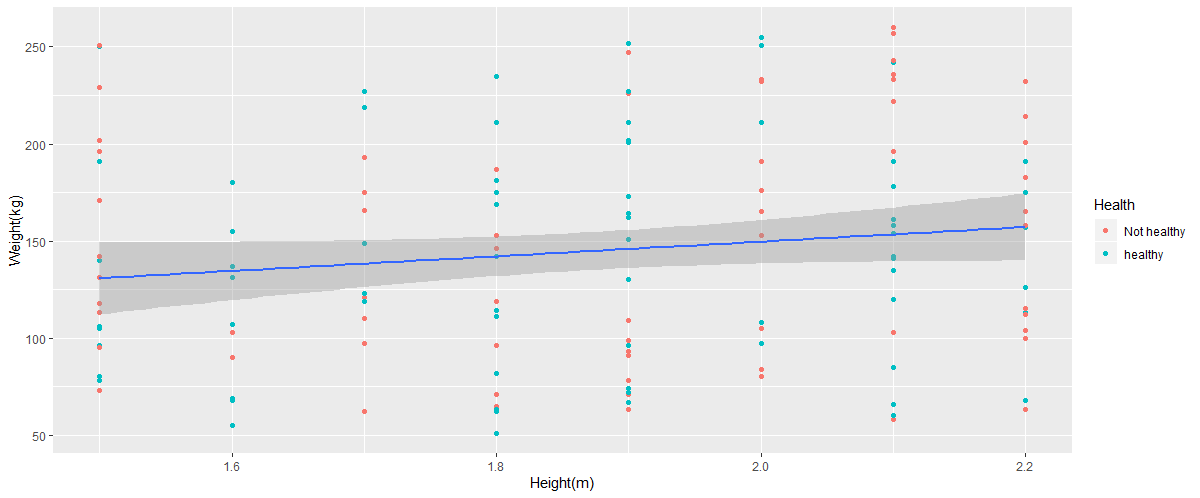


In the graph below samples of Health = “one” are heavier than other ages in Health = “two”



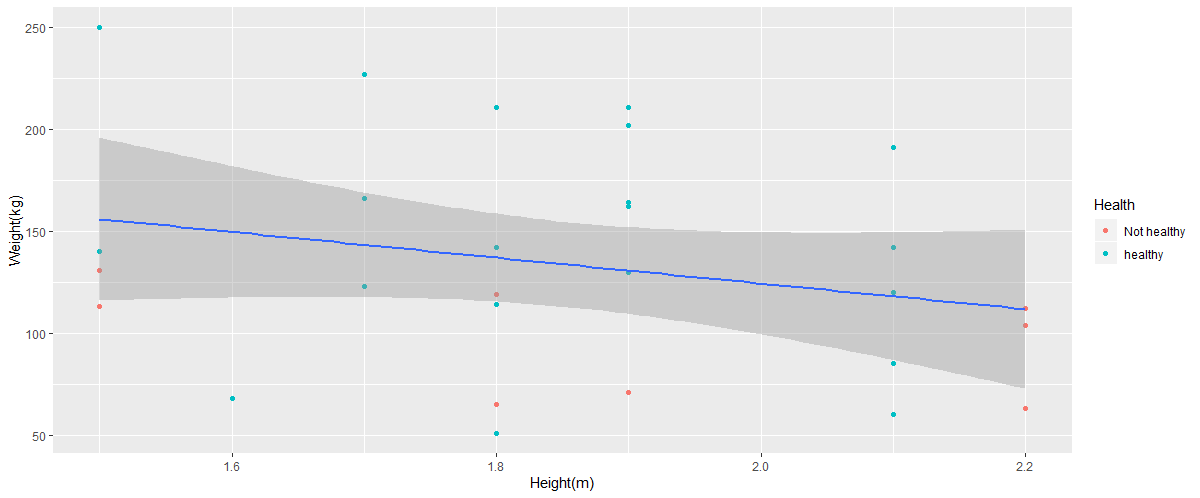


There is a weak positive correlation between Weight and Height



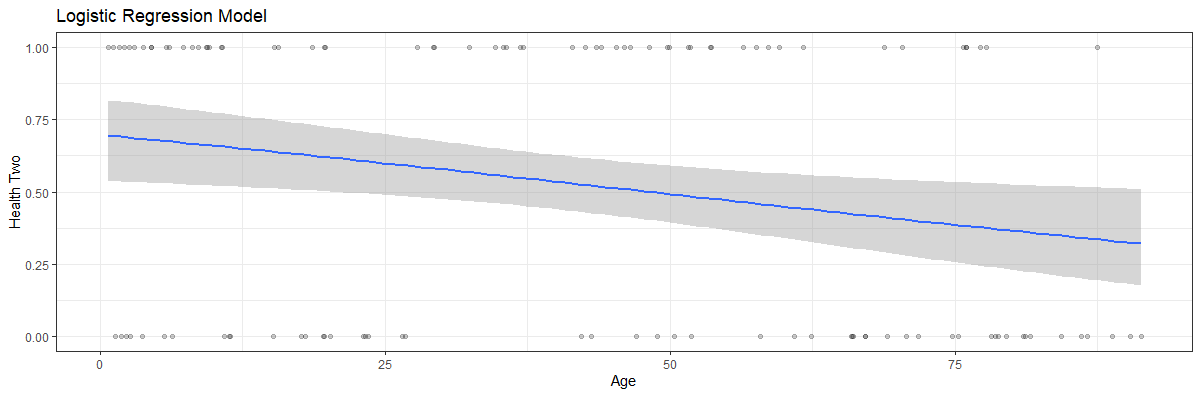
When It is sliced just by Age between 36 to 54 I got a different result.

It seems to be a negative correlation and a relationship between Height with Weight

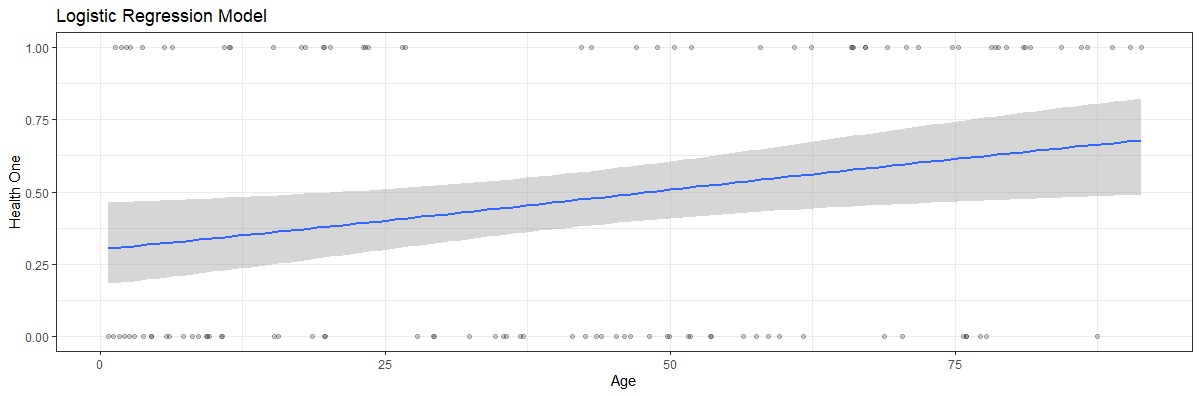


Logistic regression considering Age and Health as target

As older the samples are, closer to health = “two”

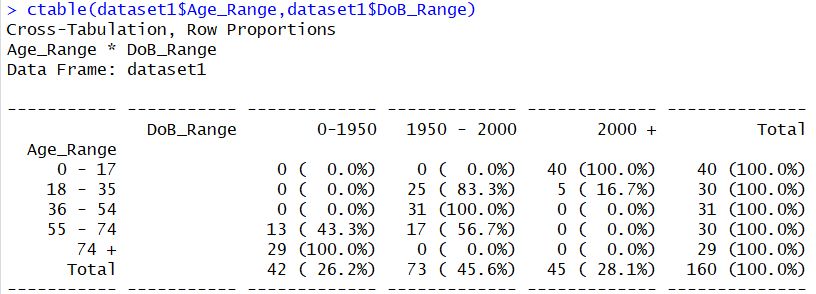


In an opposite direction, as older the samples are, closer to health = “One”



Variables Weight and Height did not present a significant difference

DoB – Date of birthday ?



**References**

STAT8010: Intro to R for Data Science

CIT, Dr Justin McGuinness

<https://www.r-graph-gallery.com/index.html>

[Micro Master Data Science](https://www.edx.org/micromasters/data-science) / San Diego University

[Big Data Analytics with R and Microsoft Azure M.L.](https://www.datascienceacademy.com.br/path-player?courseid=analise-de-dados-com-r&unit=5c71a6365e4cdebb5e8b456aUnit)

DataScience Academy

<http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/>