Assignment 3

Vivian, 300525329 Saumya Sajwan, 300529034 Samanalie Perera, 300486075

2022-08-30

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(psych)
```

Background and Data

Datasets

Target Disease: Depression

NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Mental Health - Depression Screener datasets.

Predictors

Demographics NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Demographic Variables & Sample Weights datasets.

Sleep Disorders NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Sleep Disorders datasets.

Alcohol Use NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Alcohol Use datasets.

Smoking - Cigarette Use NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Smoking - Cigarette Use datasets.

Weight History NHANES (National Health and Nutrition Examination Survey) 2005 - March 2020 Weight History datasets.

Sources

https://wwwn.cdc.gov/nchs/nhanes/Search/DataPage.aspx?Component=Questionnaire

https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics

Interest

We are interested in creating a machine learning model that finds and uses key features to identify individuals who are potentially suffering from depression.

Structure of the dataset

There are 29 variables in our dataset: 20 categorical variables and 9 numerical variables. Therefore, we will use all variables excluding id and target output as our predictors; each is a question the respondents need to answer to help us identify whether a person has depression.

Completeness of the dataset

- The dataset consists of data from 2005 to March of 2020. We were going to add in later data from 2004-1999, however the datasets for them were completely different. They had more questions and didn't match the 2005-2020 questions, so we decided to not add them in.
- For classifying whether or not an individual has depression, we calculate the score from each variable in the depression dataset. Therefore, we deleted all the NA values for more accurate output. This meant that our dataset went down from 43,928 to 26,627. 17, 301 values were taken out.
- For predictor datasets, we converted the "Refused" and "Don't know" categories to NA values for more accessible future data preprocessing.

Steps to integrate the datasets

We decided to use binary classification to predict our target disease, meaning 1 has depression and 0 is not. To get the target output column, here are the steps we did:

- Loaded the NHANES package on R. Only the datasets from 2005 to 2014 were part of the package. So for 2015 - March 2020, we had to read them using an R function (foreign::read.xport), which reads the data from a SAS XPORT file and returns it in a data frame.
- Adding more columns: filename, cycle, start year and end year, so the data from the newer datasets match the older ones.
- Added all the datasets together into one large dataset using rbind.
- To find out the total score each individual had, we added each of the columns together and added them to a new column called Score.

- We added another column for classifying whether or not an individual has depression. For that, we said that if the Score was 10 or higher, that person has depression. If not, they don't. We used this website for the score: https://bmcrheumatol.biomedcentral.com/articles/10.1186/s41927-021-00236-w#:~:text=Adults%20(%E2%89%A5%2018%20years)%20with,9%20(PHQ%2D9)

For generating the predictor variables, here is how we did it:

- First, using the same method we did for downloading the depression dataset. We decided to use the variables from Demographic Variables & Sample Weights, Sleep Disorders, Alcohol Use, Smoking - Cigarette Use and Weight History datasets.
- Each dataset has different variables every year. So for each dataset, we only use the variables that appear every year and then combine them as one dataset.
- Then we will look at the categorical variables. For example, the category "Refused" is 777 and "Don't know" is 999 in some datasets. We convert them to NA values, as the number doesn't have any meaning, which may affect our prediction accuracy in the future.
- Finally, we will combine all the datasets by their id numbers and add the target output column as our final dataset. It's to be used in later building the model.

Ethics, Privacy and Security

Ethical considerations

There are a number of ethical concerns that need to be considered when developing and deploying a machine learning algorithm that predicts if an individual is depressed. First, it is important to consider the potential impact of false positives and false negatives. If the algorithm incorrectly predicts that an individual is depressed, this could lead to them being unnecessarily treated or stigmatized. On the other hand, if the algorithm fails to predict that an individual is depressed, this could result in them not receiving the help they need.

Privacy concerns

Access and use

Whether we got permission and notified every respondent before using their information is essential, especially under legislation in different states. Also, our model requires access to many respondents' data, and we need to prevent the data from being used in different ways over time.

Re-identification

Another concern with healthcare data is whether we can protect patients' information. A lot of research shows that people can use different techniques to re-identify individuals in the data. But on the other hand, too much de-identification may diminish the clinical utility of the data, but too little de-identification may lead to a breach of privacy. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2826964/)

Security concerns

If the algorithm is made public via a data leak, then anyone could use it to find out which individuals are more likely to be depressed. This information could be used to target those individuals with ads or content that exploits their vulnerabilities. For example, an advertiser could show ads for antidepressant medications to someone who is predicted to be depressed.

Exploratory Data Analysis

```
df = read.csv("project_data.csv")
target = "result"

## Dropping redundant features
df = subset(df, select = -c(X, id))
```

Number of Features & Observations

```
dim(df)
```

```
## [1] 26627    28
```

```
sum(unlist(lapply(df, is.numeric)))
```

```
## [1] 28
```

There are 26627 observations and 28 features in our dataset. Out of those 28 features, 27 are explanatory while 1 is our target variable. Due to the way the NHANES data was encoded, all 28 features are currently considered to be numeric by R even though some of these variables are actually categorical. This will need to be changed before we start our modeling.

The actual numerical variables are; “age”, “sleep_hours”, “SMD641”, “WHD010”, “WHD020”, “WHD050”, “WHD110”, “WHD120”, “WHD140”

Summary of all variables

```
summary(df)
```

```
##      result             age            gender          race
##  Min.   :0.0000   Min.   :18.00   Min.   :1.00   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:30.50   1st Qu.:1.00   1st Qu.:2.000
##  Median :0.0000   Median :46.00   Median :2.00   Median :3.000
##  Mean   :0.1498   Mean   :47.08   Mean   :1.55   Mean   :3.023
##  3rd Qu.:0.0000   3rd Qu.:62.00   3rd Qu.:2.00   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :85.00   Max.   :2.00   Max.   :5.000
##
##  marital_status    family_PIR    education_level_adults    language
##  Min.   :1.000   Min.   :0.000   Min.   :1.000          Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.030   1st Qu.:3.000          1st Qu.:1.000
##  Median :2.000   Median :1.950   Median :4.000          Median :1.000
##  Mean   :2.456   Mean   :2.399   Mean   :3.418          Mean   :1.084
##  3rd Qu.:4.000   3rd Qu.:3.810   3rd Qu.:4.000          3rd Qu.:1.000
##  Max.   :6.000   Max.   :5.000   Max.   :9.000          Max.   :2.000
##  NA's   :1339    NA's   :2300    NA's   :1684           NA's   :582
##   sleep_hours     trouble_sleeping_history drinks_per_occasion     SMQ020
```

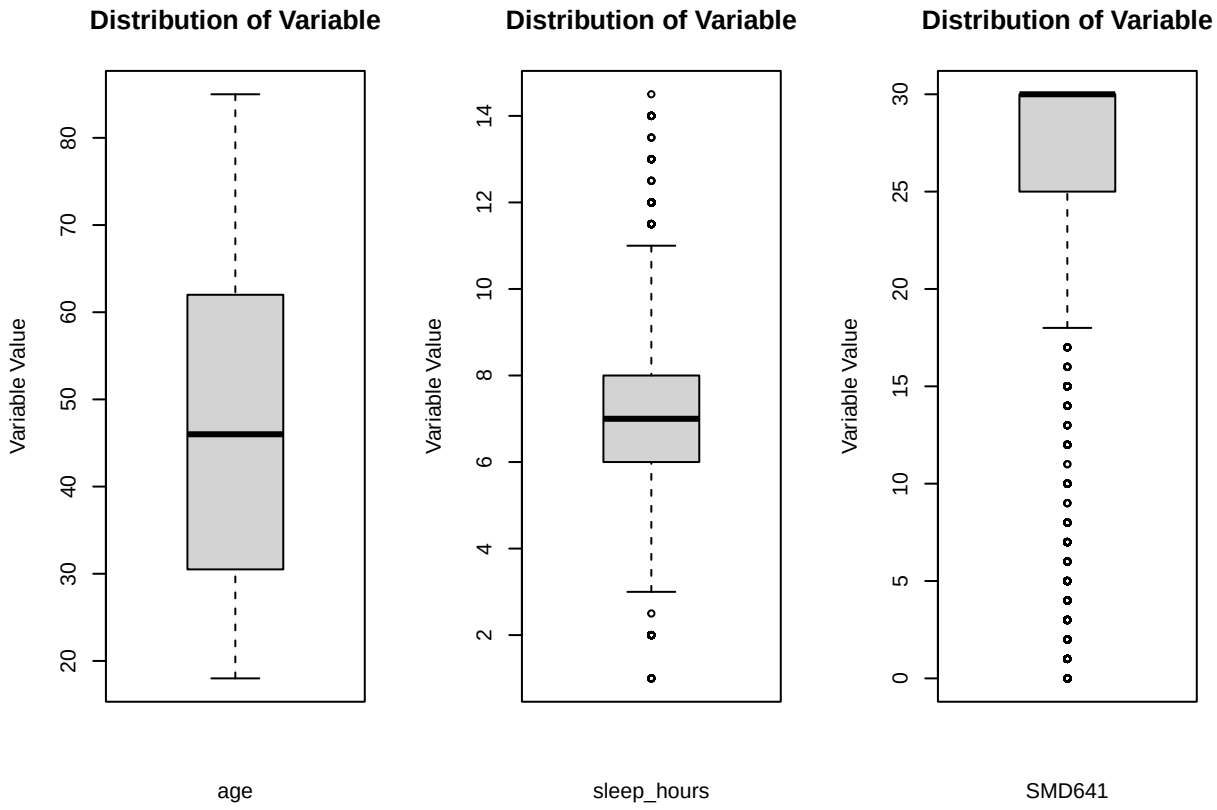
```
##  Min.   : 1.000   Min.   :1.00            Min.   : 1.000      Min.   :1.00
##  1st Qu.: 6.000   1st Qu.:1.00            1st Qu.: 1.000      1st Qu.:1.00
##  Median : 7.000   Median :2.00            Median : 2.000      Median :2.00
##  Mean   : 7.048   Mean   :1.68            Mean   : 2.869      Mean   :1.54
##  3rd Qu.: 8.000   3rd Qu.:2.00            3rd Qu.: 3.000      3rd Qu.:2.00
##  Max.   :14.500   Max.   :9.00            Max.   :83.000      Max.   :2.00
##  NA's   :108                              NA's   :9139        NA's   :998
##      SMD030          SMQ040          SMD641          SMD650
##  Min.   : 0.00   Min.   :1.000   Min.   : 0.00   Min.   : 1.00
##  1st Qu.:15.00   1st Qu.:1.000   1st Qu.:25.00   1st Qu.: 4.00
##  Median :17.00   Median :3.000   Median :30.00   Median :10.00
##  Mean   :17.42   Mean   :2.123   Mean   :24.87   Mean   :12.04
##  3rd Qu.:19.00   3rd Qu.:3.000   3rd Qu.:30.00   3rd Qu.:20.00
##  Max.   :76.00   Max.   :3.000   Max.   :30.00   Max.   :95.00
##  NA's   :14885   NA's   :14844   NA's   :20476   NA's   :20684
##      SMD630          SMQ670          WHD010          WHD020          WHQ030
##  Min.   : 6.00   Min.   :1.000   Min.   :41.00   Min.   : 70.0   Min.   :1.00
##  1st Qu.:14.00   1st Qu.:1.000   1st Qu.:63.00   1st Qu.:145.0   1st Qu.:1.00
##  Median :16.00   Median :1.000   Median :66.00   Median :173.0   Median :1.00
##  Mean   :21.94   Mean   :1.475   Mean   :66.29   Mean   :179.7   Mean   :1.83
##  3rd Qu.:18.00   3rd Qu.:2.000   3rd Qu.:69.00   3rd Qu.:205.0   3rd Qu.:3.00
##  Max.   :55.00   Max.   :2.000   Max.   :83.00   Max.   :600.0   Max.   :3.00
##  NA's   :26051   NA's   :23577   NA's   :517     NA's   :448     NA's   :58
##      WHQ040          WHD050          WHQ070          WHD110
##  Min.   :1.000   Min.   : 60.0   Min.   :1.000   Min.   : 70.0
##  1st Qu.:2.000   1st Qu.:145.0   1st Qu.:1.000   1st Qu.:140.0
##  Median :2.000   Median :172.0   Median :2.000   Median :165.0
##  Mean   :2.174   Mean   :179.8   Mean   :1.612   Mean   :174.1
##  3rd Qu.:3.000   3rd Qu.:206.0   3rd Qu.:2.000   3rd Qu.:198.0
##  Max.   :3.000   Max.   :618.0   Max.   :2.000   Max.   :700.0
##  NA's   :30      NA's   :504     NA's   :3441    NA's   :9372
##      WHD120          WHD140          WHQ150
##  Min.   : 50.0   Min.   : 73.0   Min.   : 7.00
##  1st Qu.:125.0   1st Qu.:160.0   1st Qu.:25.00
##  Median :145.0   Median :187.0   Median :37.00
##  Mean   :153.5   Mean   :197.2   Mean   :39.65
##  3rd Qu.:175.0   3rd Qu.:225.0   3rd Qu.:52.00
##  Max.   :530.0   Max.   :700.0   Max.   :85.00
##  NA's   :5872    NA's   :425     NA's   :523
```

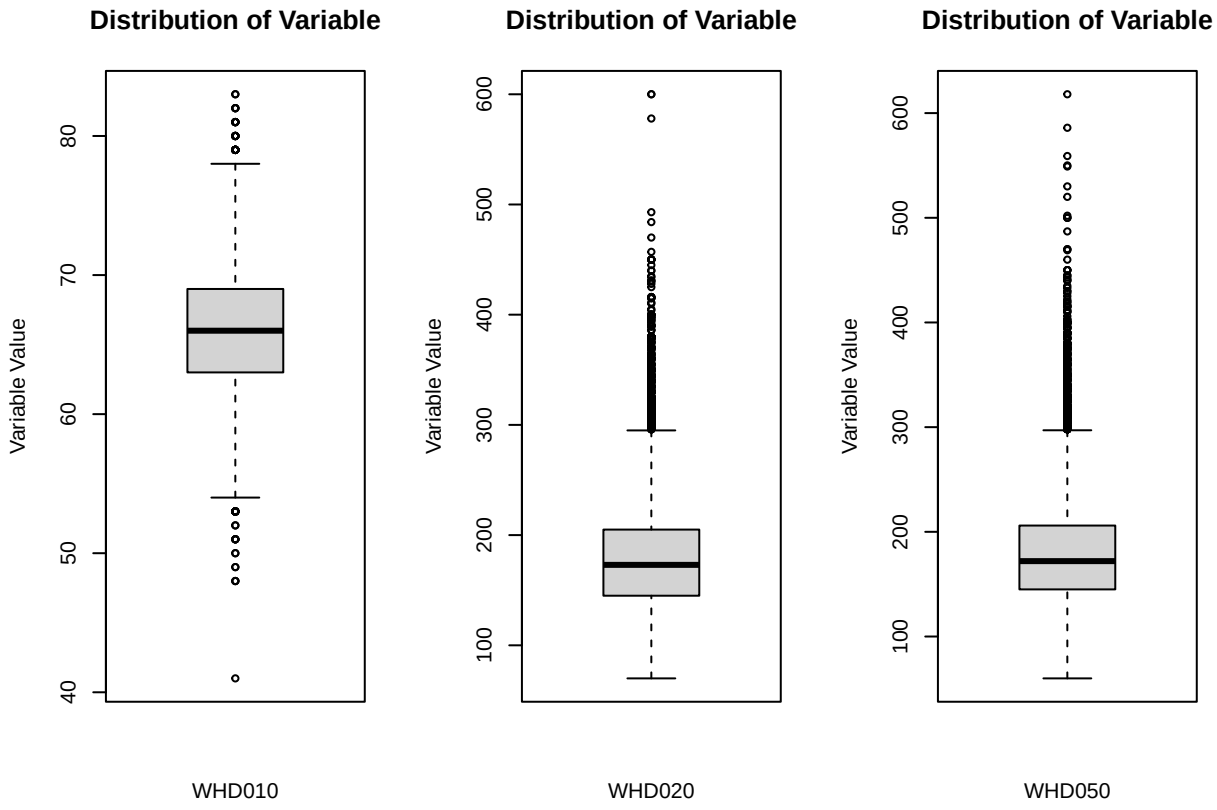
All the variables apart from result, age, gender, and race have missing values. Imputation of some form will be used to deal with this before we start modelling.

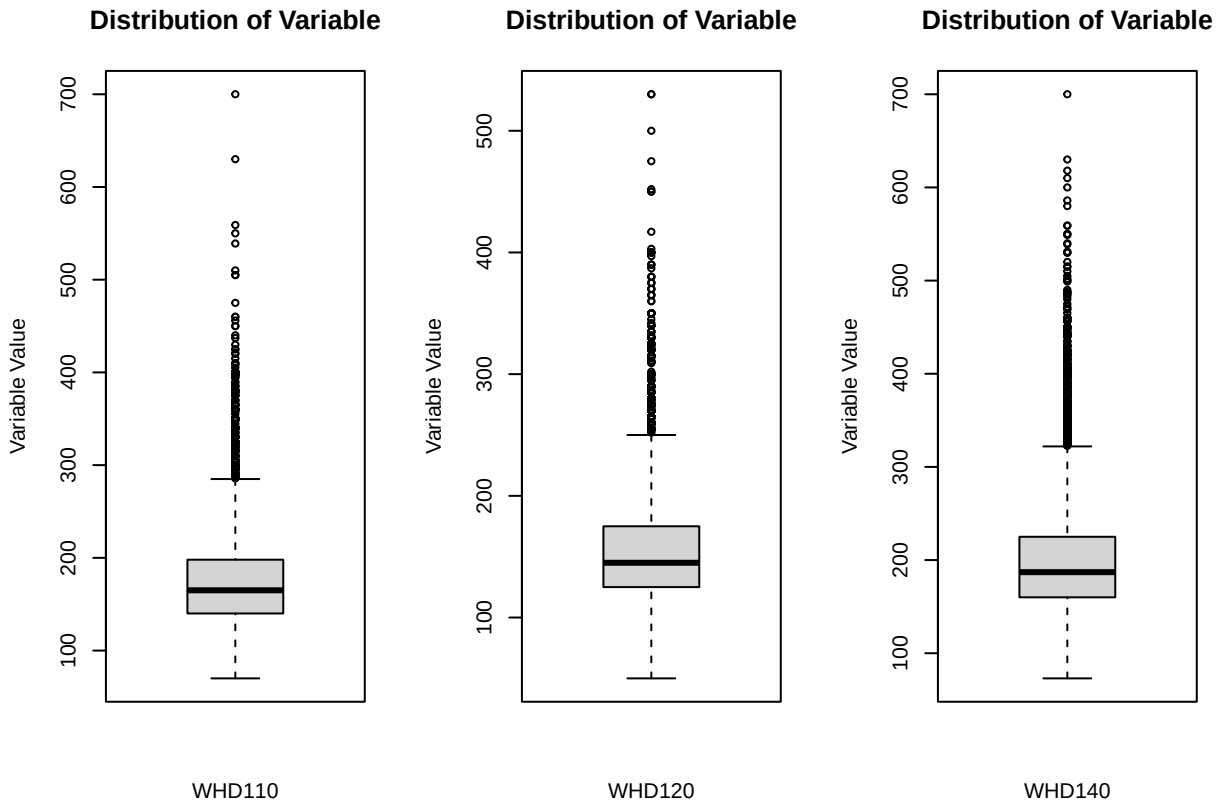
Distribution of Numerical Variables

```
numeric_columns = c("age", "sleep_hours", "SMD641", "WHD010", "WHD020", "WHD050", "WHD110", "WHD120", "W


par(mfcol=c(1, 3))
for (i in numeric_columns) {
        boxplot(df[,i], main = "Distribution of Variable", xlab = names(df[i]), ylab = "Variable Value")
}
```


Distribution of Variable
Variable Value
20
30
40
50
60
70
80
age
Distribution of Variable
Variable Value
2
4
6
8
10
12
14
sleep_hours
Distribution of Variable
Variable Value
0
5
10
15
20
25
30
SMD641

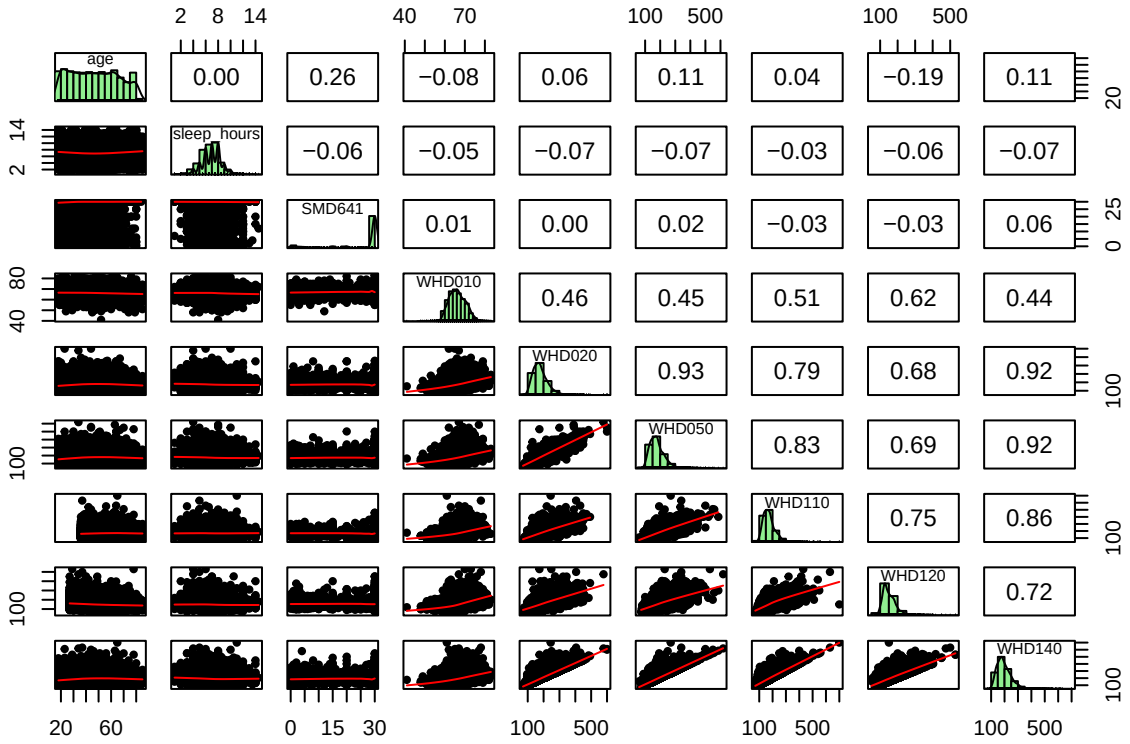

Distribution of Variable
Variable Value
40
50
60
70
80
WHD010
Distribution of Variable
Variable Value
100
200
300
400
500
600
WHD020
Distribution of Variable
Variable Value
100
200
300
400
500
600
WHD050



All the actual numerical variables, have outliers except 'WHQ150'. The outliers in these variables will need to be checked to determine whether the observation is reasonable (in the context of the variable) or requires removal. Variables "sleep_hours", "WHD010" and "WHQ150" all appear to be symmetrical while the other variables appear to be skewed in some way.

Checking for Multi-Collinearity of Numerical Variables

```
df[, numeric_columns]%>%
  pairs.panels(method = "spearman", # correlation method
               hist.col = "lightgreen", # histogram color
               density = TRUE,  # show density plots
               ellipses = FALSE # do not show correlation ellipses
               )
```



From this plot we can see that the numerical variable "sleep hours" is the only variable that appears to be approximately normally distributed. All the other numeric variables appear do not have a bell-shape or are skewed in some way.

There also appears to be strong linear relationships between the following paris of variables; "WHD020" and "WHD050", "WHD050" and "WHD110" , "WHD020" and "WHD140", "WHD050" and "WHD140", "WHD050" and "WHD140", "WHD110" and "WHD140" as they all have pearson coefficients greater than 0.8. This is expected as all of these variables are related to the individuals weight and/or height, which are known to have a linear relationship between them.

This suggests that multicollinearity may be present, meaning some of these variables may need to be removed before modelling, depending on the type of model being created.

Individual Contributions

Vivian, 300525329

Saumya Sajwan, 300529034

Samanalie Perera, 300486075