Graphical user interface, text

Description automatically generated

# Coversheet for submission of MSc coursework

* DO NOT WRITE YOUR NAME ON YOUR WORK. Instead, please write your student number on this coversheet.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Student No. | 2 | 3 | 2 | 5 | 9 | 8 | 8 |

|  |  |
| --- | --- |
| Date Submitted: | 18/01/2023 |
| Word Count: | 700 |

PLAGIARSIM DECLARATION:

This assignment is entirely my own work, and it adheres to the University of Bristol’s policy on plagiarism *and* academic integrity. Quotations from secondary literature are indicated by the use of inverted commas around ALL such quotations AND by citations in the text or notes to the author concerned. ALL primary and secondary literature used in this piece of work is indicated in the bibliography placed at the end, and dependence upon ANY source used is indicated at the appropriate point in the text. I confirm that no sources have been used other than those stated.

**I understand that plagiarism, collusion, and cheating constitute misconduct and may result in disciplinary action being taken.**

AHDS Assessment2 Repot

2325988

2023-01-16

## data cleaning

The dataset I chose for this assessment is the Body Measures Dataset from the National Health and Nutrition Examination Survey. The data were collected by the health technicians and were used to check how infants and children grow, and to analyse the trends of obesity and overweight in the U.S. The dataset includes body measurement data such as weight, height, head circumference, arm length and so on of infants and children. Data for different factors were collected according to the children’s age.

## SEQN BMDSTATS BMXWT BMIWT BMXRECUM BMIRECUM BMXHEAD BMIHEAD BMXHT BMIHT  
## 1 31127 1 10.2 NA 73.6 NA NA NA NA NA  
## 2 31128 1 40.1 NA NA NA NA NA 151.6 NA  
## 3 31129 1 74.6 NA NA NA NA NA 167.7 NA  
## 4 31130 4 NA NA NA NA NA NA NA NA  
## 5 31131 3 75.2 NA NA NA NA NA 156.0 NA  
## BMXBMI BMXLEG BMILEG BMXCALF BMICALF BMXARML BMIARML BMXARMC BMIARMC BMXWAIST  
## 1 NA NA NA NA NA 16.0 NA 15.5 NA NA  
## 2 17.45 37.6 NA 29.3 NA 34.3 NA 21.7 NA 62.8  
## 3 26.53 42.7 NA 40.6 NA 36.5 NA 32.6 NA 97.8  
## 4 NA NA NA NA NA NA NA NA NA NA  
## 5 30.90 38.0 NA 36.6 NA 35.0 NA 35.8 NA 96.0  
## BMIWAIST BMXTHICR BMITHICR BMXTRI BMITRI BMXSUB BMISUB  
## 1 NA NA NA 12.8 NA 10.0 NA  
## 2 NA 39.5 NA 10.4 NA 8.4 NA  
## 3 NA 55.9 NA 18.8 NA 17.6 NA  
## 4 NA NA NA NA NA NA NA  
## 5 NA 53.7 NA NA 2 NA 1

In the dataset, there are also codes showing the data measure component status and comments noted to identify the problems in the data, which can help us cleaning and analysing the data. After having a look of the raw data, we could find that there are some participants did not have complete data for their age group, also some group did not need to have all the factors tested. Therefore, there are a lot of missing data in the dataset. In some columns, there are more than 9000 missing data among all 9950 children. In order to analyse and visualize the data better, I decided to delete the columns with too many missing data and remove the missing data. After the data cleaning, the dataset has 13 columns and 5885 rows left. Besides, we have the continuous BMI data. We could make better analysis if we put the BMI data into different groups according to the WHO BMI cut-off points. In the end, we have 15 columns and 5885 rows in the dataset when the classification is done.

## data visualization

## X SEQN BMDSTATS BMXWT BMXHT BMXBMI BMXLEG BMXCALF BMXARML BMXARMC BMXWAIST  
## 1 2 31128 1 40.1 151.6 17.45 37.6 29.3 34.3 21.7 62.8  
## 2 3 31129 1 74.6 167.7 26.53 42.7 40.6 36.5 32.6 97.8  
## 3 6 31132 1 69.5 167.6 24.74 40.4 35.6 37.5 31.2 96.5  
## 4 7 31133 1 45.0 163.7 16.79 39.2 31.7 35.8 22.3 62.0  
## 5 8 31134 1 101.9 182.4 30.63 41.5 42.6 42.7 33.0 117.1  
## 6 13 31140 1 42.1 152.3 18.15 40.0 30.9 31.3 24.7 67.3  
## 7 14 31141 1 60.5 180.6 18.55 41.2 34.0 40.0 24.3 69.8  
## 8 16 31143 1 76.4 184.0 22.57 49.1 36.7 41.4 28.6 84.2  
## 9 17 31144 1 69.9 167.1 25.03 40.3 36.7 35.8 33.2 84.6  
## 10 18 31145 1 39.8 151.2 17.41 37.0 30.5 30.7 21.0 64.5  
## BMXTHICR BMXTRI BMXSUB BMI\_cat  
## 1 39.5 10.4 8.4 Underweight  
## 2 55.9 18.8 17.6 Overweight  
## 3 48.0 10.4 22.2 Normal Range  
## 4 41.3 6.8 8.2 Underweight  
## 5 50.5 13.0 16.0 Obese I  
## 6 43.2 15.0 12.0 Underweight  
## 7 46.3 8.0 8.4 Normal Range  
## 8 52.4 12.8 7.4 Normal Range  
## 9 51.7 13.0 17.8 Overweight  
## 10 41.0 11.2 6.0 Underweight

Among the measurements given in the dataset, BMI is the value that can show whether the child is overweight directly, while other values have some relationship with the obesity. So, we want to make some data visualization to estimate the proportion of overweight and obesity, and the correlation between the factors.

Chart, radar chart, sunburst chart

Description automatically generated

First, I have the rose plot of the BMI Categories, showing the amount of people in each group. Among the 5885 children in the cleaned dataset, most children (n=2561) had the normal BMI value, while 740 of them were underweight and 1703 of them were overweight. And 881 of them had different levels of obesity, group obese I, II, III had 686, 152, 43 children respectively. 43.5% of the children in this study had normal BMI, however, 12.6% were underweight and 43.9% were overweight or obese. The proportion of unhealthy kids is high in this research, showing a worrisome trend in the American children.

## `geom\_smooth()` using formula = 'y ~ x'

Chart, scatter chart

Description automatically generated

Also, I make the scatter plot of weight and standing height, and fit a linear regression module of their relationship. It shows that there is a corelation between. If the weight is 0, the expected height is 130.9448cm. And when the weight increases 1 kg, height increases 0.4918 cm averagely.

Chart

Description automatically generated

Chart, line chart, histogram

Description automatically generated

Chart

Description automatically generated

There are also box plots and density plots, which show the distribution of the data, and the distribution of different values in each BMI categories. From the plots, we can see that upper leg length, maximal calf circumference, upper arm length, arm circumference, and thigh circumference had narrower distribution, while the distribution of standing height and waist circumference were wider. Moreover, when looking at the density plot of BMX data grouped by BMI categories, we can tell that children from different group did not have much difference in standing height, upper arm length and upper leg length. On the other hand, the gap of arm circumference, maximal calf circumference, thigh circumference and waist circumference between each category are much bigger. That means the degree of obesity has more influence on the circumference of the body parts than the height of the length of the body parts. From the visualization, we can conclude that 43.5% of the children in the study had the normal BMI, 12.6% of them were underweight and 43.9% of them had different degrees of overweight or obesity, which shows a worrying trend among the U.S. children. Besides, the circumference of body part can indicate the obesity better than the length. Later, a snakemake was created according to the codes, and all the files were uploaded to the github for the version control. Here is computer environment and packages used in this study.

R and package versions

─ Session info ───────────────────────────────────────────────────────────────  
 setting value  
 version R version 4.1.1 (2021-08-10)  
 os macOS 13.1  
 system aarch64, darwin20  
 ui X11  
 language (EN)  
 collate en\_US.UTF-8  
 ctype en\_US.UTF-8  
 tz Europe/London  
 date 2023-01-18  
 pandoc 2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (via rmarkdown)  
  
─ Packages ───────────────────────────────────────────────────────────────────  
 package \* version date (UTC) lib source  
 assertthat 0.2.1 2019-03-21 [1] CRAN (R 4.1.0)  
 bitops 1.0-7 2021-04-24 [1] CRAN (R 4.1.0)  
 broom.helpers 1.9.0 2022-09-23 [1] CRAN (R 4.1.1)  
 caTools 1.18.2 2021-03-28 [1] CRAN (R 4.1.0)  
 cli 3.4.1 2022-09-23 [1] CRAN (R 4.1.1)  
 clipr 0.8.0 2022-02-22 [1] CRAN (R 4.1.1)  
 colorspace 2.0-3 2022-02-21 [1] CRAN (R 4.1.1)  
 DBI 1.1.3 2022-06-18 [1] CRAN (R 4.1.1)  
 desc 1.4.2 2022-09-08 [1] CRAN (R 4.1.1)  
 details 0.3.0 2022-03-27 [1] CRAN (R 4.1.1)  
 digest 0.6.30 2022-10-18 [1] CRAN (R 4.1.1)  
 dplyr 1.0.10 2022-09-01 [1] CRAN (R 4.1.1)  
 evaluate 0.18 2022-11-07 [1] CRAN (R 4.1.1)  
 fansi 1.0.3 2022-03-24 [1] CRAN (R 4.1.1)  
 farver 2.1.1 2022-07-06 [1] CRAN (R 4.1.1)  
 fastmap 1.1.0 2021-01-25 [1] CRAN (R 4.1.0)  
 generics 0.1.3 2022-07-05 [1] CRAN (R 4.1.1)  
 ggiraph \* 0.8.5 2022-12-03 [1] CRAN (R 4.1.1)  
 ggiraphExtra \* 0.3.0 2020-10-06 [1] CRAN (R 4.1.1)  
 ggplot2 \* 3.4.0 2022-11-04 [1] CRAN (R 4.1.1)  
 glue 1.6.2 2022-02-24 [1] CRAN (R 4.1.1)  
 gplots \* 3.1.3 2022-04-25 [1] CRAN (R 4.1.1)  
 gt 0.8.0 2022-11-16 [1] CRAN (R 4.1.1)  
 gtable 0.3.1 2022-09-01 [1] CRAN (R 4.1.1)  
 gtools 3.9.3 2022-07-11 [1] CRAN (R 4.1.1)  
 gtsummary \* 1.6.2 2022-09-30 [1] CRAN (R 4.1.1)  
 highr 0.9 2021-04-16 [1] CRAN (R 4.1.0)  
 htmltools 0.5.3 2022-07-18 [1] CRAN (R 4.1.1)  
 htmlwidgets 1.5.4 2021-09-08 [1] CRAN (R 4.1.1)  
 httr 1.4.4 2022-08-17 [1] CRAN (R 4.1.1)  
 insight 0.18.7 2022-11-18 [1] CRAN (R 4.1.1)  
 KernSmooth 2.23-20 2021-05-03 [1] CRAN (R 4.1.1)  
 knitr 1.41 2022-11-18 [1] CRAN (R 4.1.1)  
 labeling 0.4.2 2020-10-20 [1] CRAN (R 4.1.0)  
 lattice 0.20-45 2021-09-22 [1] CRAN (R 4.1.1)  
 lifecycle 1.0.3 2022-10-07 [1] CRAN (R 4.1.1)  
 magrittr 2.0.3 2022-03-30 [1] CRAN (R 4.1.1)  
 MASS 7.3-58.1 2022-08-03 [1] CRAN (R 4.1.1)  
 Matrix 1.5-1 2022-09-13 [1] CRAN (R 4.1.1)  
 mgcv 1.8-41 2022-10-21 [1] CRAN (R 4.1.1)  
 munsell 0.5.0 2018-06-12 [1] CRAN (R 4.1.0)  
 mycor 0.1.1 2018-04-10 [1] CRAN (R 4.1.1)  
 nlme 3.1-160 2022-10-10 [1] CRAN (R 4.1.1)  
 pillar 1.8.1 2022-08-19 [1] CRAN (R 4.1.1)  
 pkgconfig 2.0.3 2019-09-22 [1] CRAN (R 4.1.0)  
 plyr 1.8.8 2022-11-11 [1] CRAN (R 4.1.1)  
 png 0.1-7 2013-12-03 [1] CRAN (R 4.1.0)  
 ppcor 1.1 2015-12-03 [1] CRAN (R 4.1.0)  
 purrr 0.3.5 2022-10-06 [1] CRAN (R 4.1.1)  
 R6 2.5.1 2021-08-19 [1] CRAN (R 4.1.1)  
 RColorBrewer 1.1-3 2022-04-03 [1] CRAN (R 4.1.1)  
 Rcpp 1.0.9 2022-07-08 [1] CRAN (R 4.1.1)  
 reshape2 1.4.4 2020-04-09 [1] CRAN (R 4.1.0)  
 rlang 1.0.6 2022-09-24 [1] CRAN (R 4.1.1)  
 rmarkdown 2.18 2022-11-09 [1] CRAN (R 4.1.1)  
 rprojroot 2.0.3 2022-04-02 [1] CRAN (R 4.1.1)  
 rstudioapi 0.14 2022-08-22 [1] CRAN (R 4.1.1)  
 scales 1.2.1 2022-08-20 [1] CRAN (R 4.1.1)  
 sessioninfo 1.2.2 2021-12-06 [1] CRAN (R 4.1.1)  
 sjlabelled 1.2.0 2022-04-10 [1] CRAN (R 4.1.1)  
 sjmisc 2.8.9 2021-12-03 [1] CRAN (R 4.1.1)  
 stringi 1.7.8 2022-07-11 [1] CRAN (R 4.1.1)  
 stringr 1.4.1 2022-08-20 [1] CRAN (R 4.1.1)  
 systemfonts 1.0.4 2022-02-11 [1] CRAN (R 4.1.1)  
 tibble 3.1.8 2022-07-22 [1] CRAN (R 4.1.1)  
 tidyr 1.2.1 2022-09-08 [1] CRAN (R 4.1.1)  
 tidyselect 1.2.0 2022-10-10 [1] CRAN (R 4.1.1)  
 utf8 1.2.2 2021-07-24 [1] CRAN (R 4.1.0)  
 uuid 1.1-0 2022-04-19 [1] CRAN (R 4.1.1)  
 vctrs 0.5.1 2022-11-16 [1] CRAN (R 4.1.1)  
 viridisLite 0.4.1 2022-08-22 [1] CRAN (R 4.1.1)  
 withr 2.5.0 2022-03-03 [1] CRAN (R 4.1.1)  
 xfun 0.35 2022-11-16 [1] CRAN (R 4.1.1)  
 xml2 1.3.3 2021-11-30 [1] CRAN (R 4.1.1)  
 yaml 2.3.6 2022-10-18 [1] CRAN (R 4.1.1)  
  
 [1] /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/library  
  
──────────────────────────────────────────────────────────────────────────────