group_project

Abhay Sharma, Darshil Dave, Max Kaplan, Kurt Galvez

2024-03-13

Descriptive Analysis

For the analysis of the provided "ObesityDataSet.csv", our team has chosen to delve into the realm of health and wellness, with a particular focus on the factors contributing to obesity. The dataset encompasses data from individuals, including critical details such as gender, age, height, weight, dietary habits, physical activity, and more, spanning from young adults to older individuals across various geographical locations. The primary aim of our analysis is to explore and identify the key factors influencing obesity levels among individuals. Through meticulous data exploration, pre-processing, and analysis, we intend to uncover the relationships between lifestyle choices—such as dietary habits, physical activity, and technology use—and obesity. Our approach involves employing statistical methods and predictive modeling to analyze the dataset comprehensively.

One of the cornerstones of our analysis is the development of a predictive model, possibly through linear regression or a classification approach, to predict an individual's obesity level based on various lifestyle and demographic factors. Moreover, we plan to devise a scoring model that quantifies each individual's risk level of obesity, facilitating a deeper understanding of the impact of lifestyle choices on health.

Our analysis also aims to test several hypotheses to explore intriguing questions, such as the impact of genetic predisposition (family history of overweight) on obesity, the influence of dietary choices (vegetable consumption, snack habits), and physical activity on maintaining a healthy weight, and the role of technology use in sedentary behavior contributing to obesity. Additionally, we are interested in investigating how these factors vary across different demographics and whether specific interventions or lifestyle modifications can significantly impact one's obesity risk.

Current Progress

We first started by comparing different variables with each other. For example, we determined whether weight and height (i.e. BMI) directly relate to obesity level, if smoking affects your weight, if family history has a significant impact on weight, etc. By computing the average BMI for every obesity level followed by creating a box plot, we found that the obesity levels in this data set appeared to be obtained by calculating the BMI. The higher the BMI, the worse the obesity level. Therefore, we will assert that the BMI directly determines the obesity level and BMI will be used for tests where the categorical variable obesity level can not be used. Furthermore...

A correlation matrix between all our continuous variables also showed us that

From analyzing the data and variables we have, we have decided to move forward with finding out which combination of habits would lead to being obese. Several summary statistics and data visualizations lead us to believe that factors such as _____ have a significant ___

Future direction

Data adjustments (only for reference. will not be in the final progress report)

```
#Data initialization

library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.2.3

## Warning: package 'ggplot2' was built under R version 4.2.3

## Warning: package 'tibble' was built under R version 4.2.3

## Warning: package 'tidyr' was built under R version 4.2.3

## Warning: package 'readr' was built under R version 4.2.3

## Warning: package 'purrr' was built under R version 4.2.3

## Warning: package 'dplyr' was built under R version 4.2.3

## Warning: package 'forcats' was built under R version 4.2.3

## Warning: package 'lubridate' was built under R version 4.2.3

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Read data
data = read.csv("~/Downloads/ObesityDataSet_raw_and_data_sinthetic.csv")
head(data)
```

```
##   Gender Age Height Weight family_history_with_overweight FAVC FCVC NCP
## 1 Female  21   1.62   64.0                             yes   no    2   3
## 2 Female  21   1.52   56.0                             yes   no    3   3
## 3   Male  23   1.80   77.0                             yes   no    2   3
## 4   Male  27   1.80   87.0                              no   no    3   3
## 5   Male  22   1.78   89.8                              no   no    2   1
```

```
## 6   Male  29   1.62   53.0                              no  yes    2   3
##        CAEC SMOKE CH2O SCC FAF TUE       CALC                MTRANS
## 1 Sometimes    no    2  no   0   1         no Public_Transportation
## 2 Sometimes   yes    3 yes   3   0  Sometimes Public_Transportation
## 3 Sometimes    no    2  no   2   1 Frequently Public_Transportation
## 4 Sometimes    no    2  no   2   0 Frequently              Walking
## 5 Sometimes    no    2  no   0   0  Sometimes Public_Transportation
## 6 Sometimes    no    2  no   0   0  Sometimes            Automobile
##            NObeyesdad
## 1       Normal_Weight
## 2       Normal_Weight
## 3       Normal_Weight
## 4  Overweight_Level_I
## 5 Overweight_Level_II
## 6       Normal_Weight
```

```
names(data)
```

```
##  [1] "Gender"                          "Age"
##  [3] "Height"                          "Weight"
##  [5] "family_history_with_overweight" "FAVC"
##  [7] "FCVC"                            "NCP"
##  [9] "CAEC"                            "SMOKE"
## [11] "CH2O"                            "SCC"
## [13] "FAF"                             "TUE"
## [15] "CALC"                            "MTRANS"
## [17] "NObeyesdad"
```

```
# Rename some variables for clarity
data <- data %>% rename(Obesity_Level = NObeyesdad) %>%
  rename(Transportation_Use = MTRANS) %>%
  rename(Alcohol_Consump = CALC) %>%
  rename(Tech_Time = TUE) %>%
  rename(Physical_Activ_Amt = FAF) %>%
  rename(Monitor_Calories = SCC) %>%
  rename(Water_Consump = CH2O) %>%
  rename(Does_Smoke = SMOKE) %>%
  rename(Food_bw_Meals = CAEC) %>%
  rename(Main_Meal_Consump = NCP) %>%
  rename(Veggie_Consump = FCVC) %>%
  rename(HiCal_Food_Consump = FAVC) %>%
  rename(Family_History_w_Overweight = family_history_with_overweight)

# Check updated names
names(data)
```

```
##  [1] "Gender"                       "Age"
##  [3] "Height"                       "Weight"
##  [5] "Family_History_w_Overweight" "HiCal_Food_Consump"
##  [7] "Veggie_Consump"               "Main_Meal_Consump"
##  [9] "Food_bw_Meals"                "Does_Smoke"
## [11] "Water_Consump"                "Monitor_Calories"
## [13] "Physical_Activ_Amt"           "Tech_Time"
```

```
## [15] "Alcohol_Consump"              "Transportation_Use"
## [17] "Obesity_Level"
```

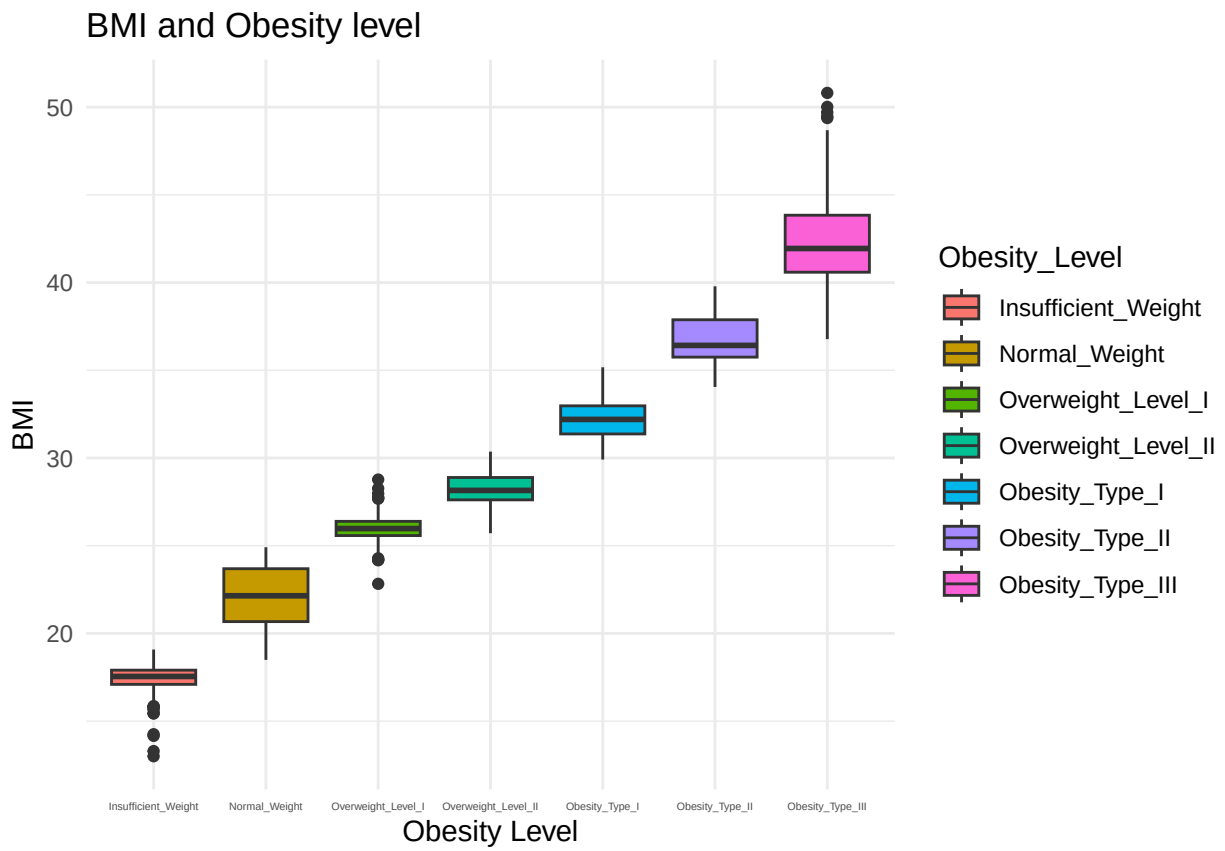
```r
# Add BMI variable
data <- data %>%
  mutate(BMI = Weight/(Height^2))
```

Data visualizations (only for reference. will not be in the final progress report)

```r
# Box plot for BMI ~ Obesity Level

# Re-organize the levels for the box plot
obesity_levels <- c("Insufficient_Weight","Normal_Weight",
                    "Overweight_Level_I","Overweight_Level_II",
                    "Obesity_Type_I","Obesity_Type_II","Obesity_Type_III")
data$Obesity_Level <- data$Obesity_Level %>%
  factor(levels=obesity_levels)
```

```r
# Create the box plot
data %>%
  ggplot(aes(x = Obesity_Level, y = BMI, fill = Obesity_Level)) +
  geom_boxplot() +
  labs(title = "BMI and Obesity level",
       x = "Obesity Level",
       y = "BMI") +
  theme_minimal() +
  theme(axis.text.x = element_text(size = 4))
```




```r
# Selecting only numerical variables for correlation analysis
numerical_vars <- data[, sapply(data, is.numeric)]

# Calculating the correlation matrix
correlation_matrix <- cor(numerical_vars)

# Displaying the correlation matrix
print(correlation_matrix)
```

```
##                                  Age      Height      Weight Veggie_Consump
## Age                       1.00000000 -0.02595813  0.20256010     0.01629089
## Height                   -0.02595813  1.00000000  0.46313612    -0.03812106
## Weight                    0.20256010  0.46313612  1.00000000     0.21612471
## Veggie_Consump            0.01629089 -0.03812106  0.21612471     1.00000000
## Main_Meal_Consump        -0.04394373  0.24367173  0.10746899     0.04221630
## Water_Consump            -0.04530386  0.21337592  0.20057539     0.06846147
## Physical_Activ_Amt       -0.14493833  0.29470900 -0.05143627     0.01993940
## Tech_Time                -0.29693059  0.05191167 -0.07156136    -0.10113485
## BMI                       0.24416312  0.13178454  0.93480575     0.26365084
##                          Main_Meal_Consump Water_Consump Physical_Activ_Amt
## Age                            -0.04394373   -0.04530386        -0.14493833
## Height                          0.24367173    0.21337592         0.29470900
## Weight                          0.10746899    0.20057539        -0.05143627
## Veggie_Consump                  0.04221630    0.06846147         0.01993940
## Main_Meal_Consump               1.00000000    0.05708800         0.12950431
## Water_Consump                   0.05708800    1.00000000         0.16723649
```

```
## Physical_Activ_Amt        0.12950431    0.16723649         1.00000000
## Tech_Time                 0.03632557    0.01196534         0.05856207
## BMI                       0.03996928    0.14420028        -0.17753732
##                         Tech_Time         BMI
## Age                   -0.29693059  0.24416312
## Height                 0.05191167  0.13178454
## Weight                -0.07156136  0.93480575
## Veggie_Consump        -0.10113485  0.26365084
## Main_Meal_Consump      0.03632557  0.03996928
## Water_Consump          0.01196534  0.14420028
## Physical_Activ_Amt     0.05856207 -0.17753732
## Tech_Time              1.00000000 -0.09972039
## BMI                   -0.09972039  1.00000000
```

```
# If you want to visualize the correlation matrix
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```
corrplot(correlation_matrix, method = "circle")
```