Regression 1 Final Project Code

Data Wrangling

```
# loading data from the source
data_raw <- read.csv('./data/raw_data.csv')

# loading a data dictionary with more readable column names
dic <- openxlsx::read.xlsx("./data/data_dictionary.xlsx")

# cleaning data
data <-
  data_raw |>
  dplyr::mutate(
    dplyr::across(
      dplyr::where(is.character),
      ~factor(stringr::str_to_title(.x))
    ),
    # ordering factors for visualization & intuitive dummy creation
    dplyr::across(
      .cols = c(CAEC, CALC),
      .fns = ~factor(.x, level = c("No","Sometimes","Frequently","Always"))
    ),
    # converting numeric counts to integers (see first paragraph of the results section)
    dplyr::across(
      .cols = c(FCVC, TUE, NCP, CH2O, FAF, Age),
      .fns = as.integer
    ),
    # ordering transit types by their frequency
    MTRANS = forcats::fct_inorder(factor(MTRANS)),
    BMI = Weight/(Height^2)
    ) |>
  # removing unneeded variables
```

```
  dplyr::select(-c(Height, Weight, NObeyesdad))

# converting names to the human readable
names(data) <- dic$Name

# generating a "dirty" copy without integer conversions
data_dirty <-
  data_raw |>
  dplyr::mutate(
    dplyr::across(
      dplyr::where(is.character),
      ~factor(stringr::str_to_title(.x))
    ),
    dplyr::across(
      .cols = c(CAEC, CALC),
      .fns = ~factor(.x, level = c("No","Sometimes","Frequently","Always"))
    ),
    MTRANS = forcats::fct_inorder(factor(MTRANS)),
    BMI = Weight/(Height^2)
  ) |>
  dplyr::select(-c(Height, Weight, NObeyesdad))

names(data_dirty) <- dic$Name
```

Exploratory data analysis

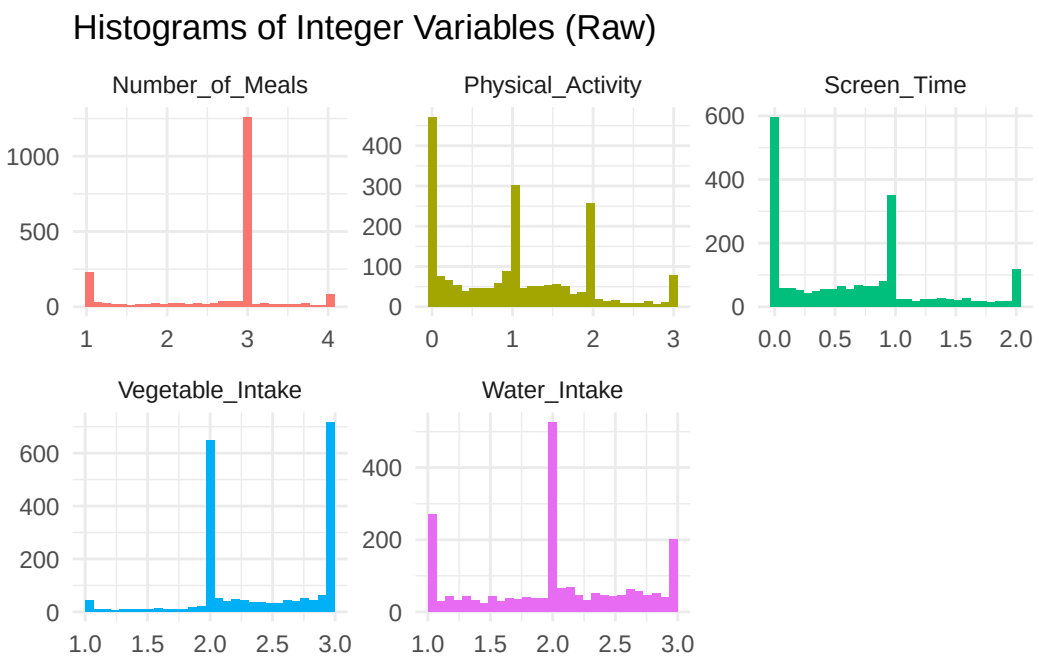
Univariate Analysis

```
psych::describe(data) |>
  dplyr::select(-c(median, trimmed, mad))
```

```
                   vars    n  mean   sd min   max range  skew kurtosis   se
Gender*               1 2111  1.51 0.50   1  2.00  1.00 -0.02    -2.00 0.01
Age                   2 2111 23.97 6.31  14 61.00 47.00  1.56     2.97 0.14
Family_History*       3 2111  1.82 0.39   1  2.00  1.00 -1.64     0.70 0.01
High_Caloric_Food*    4 2111  1.88 0.32   1  2.00  1.00 -2.40     3.74 0.01
Vegetable_Intake      5 2111  2.21 0.60   1  3.00  2.00 -0.12    -0.47 0.01
Number_of_Meals       6 2111  2.52 0.83   1  4.00  3.00 -0.88    -0.46 0.02
Snacking*             7 2111  2.14 0.47   1  4.00  3.00  1.90     5.38 0.01
Smoking*              8 2111  1.02 0.14   1  2.00  1.00  6.70    42.95 0.00
```

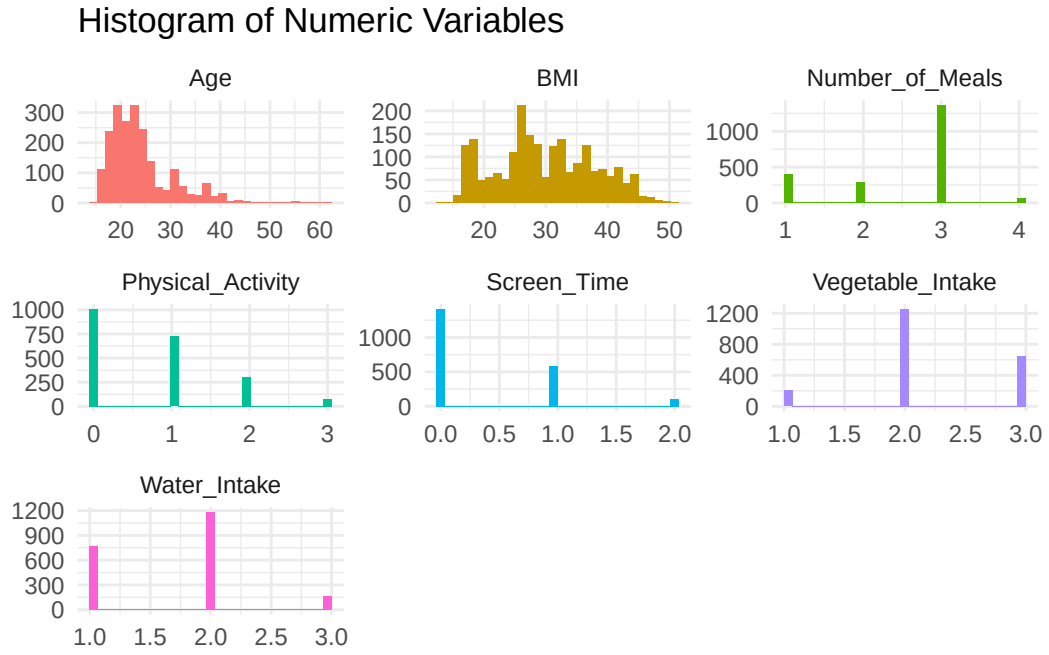
```
Water_Intake             9 2111  1.71 0.60   1  3.00  2.00  0.21    -0.60 0.01
Calorie_Monitoring*     10 2111  1.05 0.21   1  2.00  1.00  4.36    17.02 0.00
Physical_Activity       11 2111  0.73 0.83   0  3.00  3.00  0.90     0.00 0.02
Screen_Time             12 2111  0.38 0.58   0  2.00  2.00  1.25     0.55 0.01
Alcohol_Consumption*    13 2111  1.73 0.52   1  4.00  3.00 -0.24    -0.33 0.01
Transportation_Type*    14 2111  1.49 0.87   1  5.00  4.00  1.36     0.32 0.02
BMI                     15 2111 29.70 8.01  13 50.81 37.81  0.15    -0.81 0.17
```

```r
# figure B
data_dirty |>
  tidyr::pivot_longer(cols = dplyr::where(is.numeric)) |>
  dplyr::filter(!(name %in% c("Age","BMI"))) |>
  ggplot2::ggplot(ggplot2::aes(value, fill = name)) +
  ggplot2::geom_histogram() +
  ggplot2::facet_wrap(~name, scales = "free") +
  ggplot2::labs(title = "Histograms of Integer Variables (Raw)") +
  ggplot2::theme(legend.position = "none",
                 axis.title = element_blank())
```




```r
# CH2O, FAF, FCVC, NCP, TUE are discrete
# Age, BMI, Height, Weight are continuous, normal or log normal distributed
```

```r
# figure A
data |>
  tidyr::pivot_longer(cols = dplyr::where(is.numeric)) |>
  ggplot2::ggplot(ggplot2::aes(value, fill = name)) +
  ggplot2::geom_histogram() +
  ggplot2::facet_wrap(~name, scales = "free") +
  labs(title = "Histogram of Numeric Variables") +
  ggplot2::theme(legend.position = "none",
                 axis.title = element_blank())
```




```r
factors <- c("Alcohol_Consumption","Transportation_Type",
             "Calorie_Monitoring", "Snacking","Smoking",
             "Family_History","High_Caloric_Food", "Gender")

# bivariate frequency table (part 1 of Table A)
frequencies <-
  purrr::map_df(factors, \(i){
    f <- data |>
      dplyr::pull(var = i) |>
      table() |>
      t() |>
      data.frame() |>
```

```
      dplyr::mutate(
        Question = i,
        Total = sum(Freq),
        Proportion = round(Freq/sum(Freq), digits = 2)
      )
    mean <- data |>
      dplyr::summarise(
        Mean_BMI = mean(BMI),
        .by = i
      ) |>
      tidyr::pivot_longer(i, names_to = "Question", values_to = "Var2")
    dplyr::left_join(f, mean)
  }) |>
  dplyr::select(Question, Var2, Freq, Proportion, Mean_BMI)
```

Bivariate Analysis

```
# Part 2 of Table A
tests <-
  purrr::map_df(factors, \(i){
    q <- colnames(data[,i])
    bmi <- aov(
      formula = as.formula(paste("BMI ~ ", i)),
      data = data
    )
    tibble::tibble(
      Question = i,
      P_Value = c(summary(bmi)[[1]]["Pr(>F)"][1])
    )
  })

analysis <-
  dplyr::left_join(
    x = frequencies,
    y = tests
  ) |>
  dplyr::mutate(dplyr::across(c(4:6), ~round(.x, digits = 2)))

analysis |> gt::gt()
```

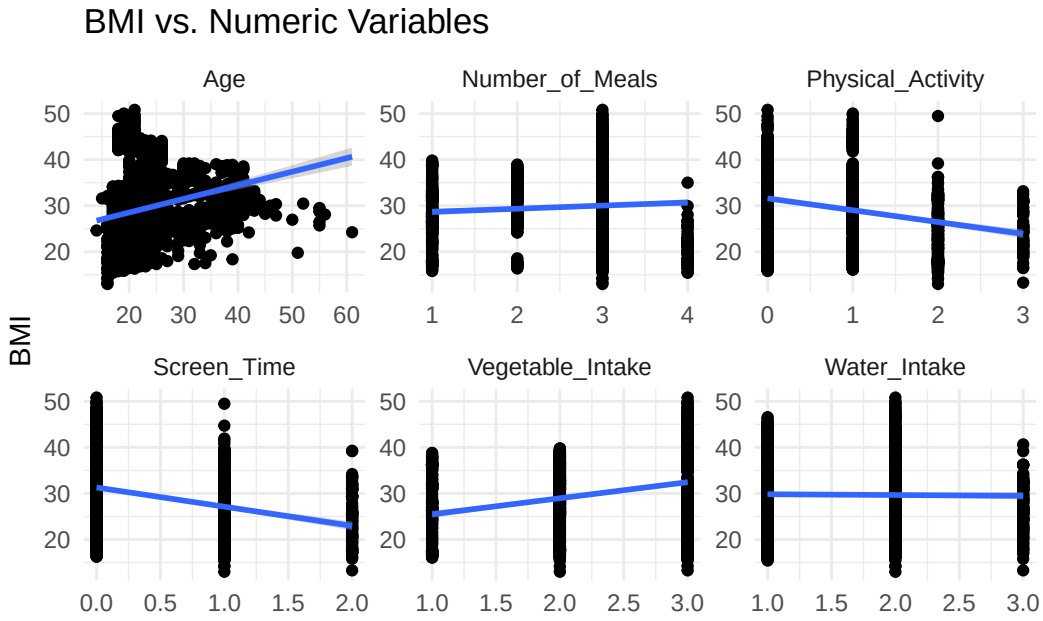
Question	Var2	Freq	Proportion	Mean_BMI	P_Value
Alcohol_Consumption	No	639	0.30	27.06	0.00
Alcohol_Consumption	Sometimes	1401	0.66	31.04	0.00
Alcohol_Consumption	Frequently	70	0.03	26.98	0.00
Alcohol_Consumption	Always	1	0.00	22.49	0.00
Transportation_Type	Public_transportation	1580	0.75	30.11	0.00
Transportation_Type	Walking	56	0.03	23.66	0.00
Transportation_Type	Automobile	457	0.22	29.19	0.00
Transportation_Type	Motorbike	11	0.01	25.76	0.00
Transportation_Type	Bike	7	0.00	25.17	0.00
Calorie_Monitoring	No	2015	0.95	30.02	0.00
Calorie_Monitoring	Yes	96	0.05	22.94	0.00
Snacking	No	51	0.02	25.43	0.00
Snacking	Sometimes	1765	0.84	31.19	0.00
Snacking	Frequently	242	0.11	20.90	0.00
Snacking	Always	53	0.03	24.32	0.00
Smoking	No	2067	0.98	29.70	0.97
Smoking	Yes	44	0.02	29.66	0.97
Family_History	No	385	0.18	21.50	0.00
Family_History	Yes	1726	0.82	31.53	0.00
High_Caloric_Food	No	245	0.12	24.26	0.00
High_Caloric_Food	Yes	1866	0.88	30.41	0.00
Gender	Female	1043	0.49	30.13	0.01
Gender	Male	1068	0.51	29.28	0.01

```r
# testing diff between bikes and motorbikes to finalize the merge
transit <- data |>
  dplyr::filter(Transportation_Type %in% c("Motorbike", "Bike"))

t.test(transit$BMI ~ transit$Transportation_Type) # 0.8402
```

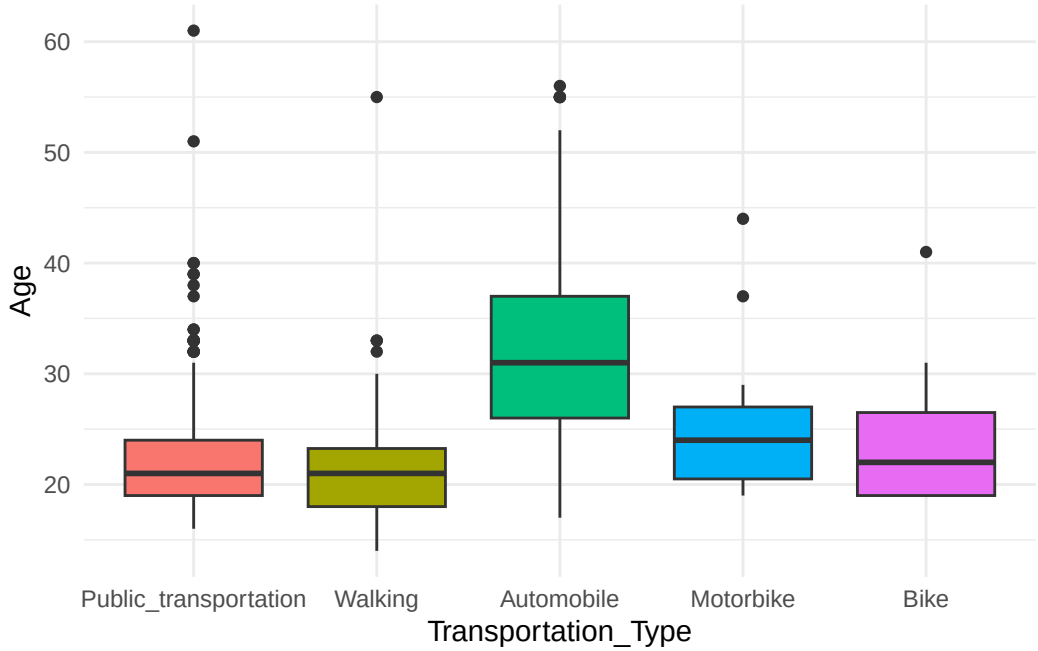
```
    Welch Two Sample t-test

data:  transit$BMI by transit$Transportation_Type
t = 0.20697, df = 10.064, p-value = 0.8402
alternative hypothesis: true difference in means between group Motorbike and group Bike is no
95 percent confidence interval:
 -5.790771  6.977841
sample estimates:
```

```
mean in group Motorbike      mean in group Bike
               25.76255                25.16902
```

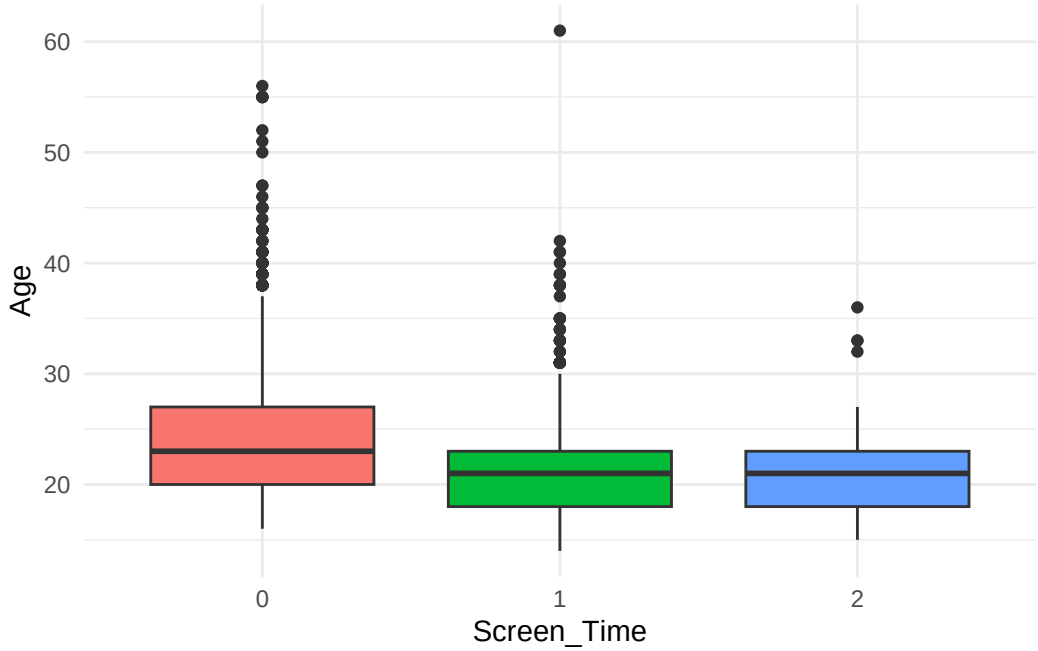
```
# figure C
data |>
  tidyr::pivot_longer(cols = c(2, 5,6,9,11,12)) |>
  ggplot(aes(value, BMI)) +
  geom_point() +
  facet_wrap(~name, scales = "free") +
  geom_smooth(method = "lm") +
  labs(x = "",
       title = "BMI vs. Numeric Variables")
```



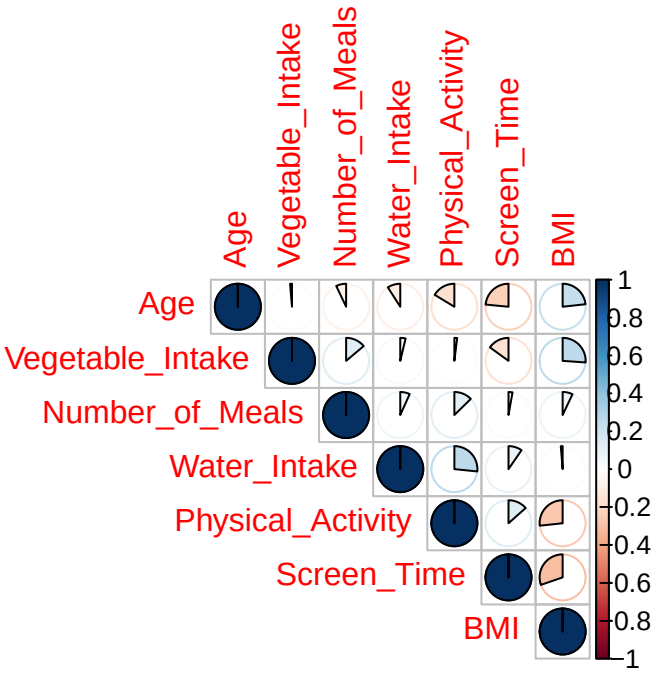

```
# looking for additional patterns
data |>
  ggplot(aes(Transportation_Type, Age, fill = Transportation_Type)) +
  geom_boxplot() +
  theme(legend.position = "none")
```




```
data |>
  ggplot(aes(factor(Screen_Time), Age, fill = factor(Screen_Time))) +
  geom_boxplot() +
  labs(x = "Screen_Time") +
  theme(legend.position = "none")
```




```
corrplot::corrplot(
  corr = cor(data |> dplyr::select(dplyr::where(is.numeric))),
  method = "pie",
  type = "upper"
  )
```


Age
Vegetable_Intake
Number_of_Meals
Water_Intake
Physical_Activity
Screen_Time
BMI
1
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1