

**Obesity in Jordan:**

**A Comprehensive Study**

**Technical Report**

Course: Data Visualization

Group Members:

* Yazan Hadi Anfoka 20200311
* Faris Al-Faori 20200454

Instructor: Amira Al-Amery

**Contributions:**

* + Yazan Hadi Anfoka:
    - Data Cleaning: Dealt with missing data, outliers, and feature engineering.
    - Exploratory Analysis: Correlation
  + Faris Al-Faori:
    - Data Visualization: Line plots and dot plots.
    - Exploratory Analysis: Correlation

Abstract:

The goal of this study is to find the statistical relations between the main and important variables, mainly the obesity indicator variable. The study is based on an incomplete dataset that has recorded several questions and concluded metrics from the questions. This data is based on a survey that’s been filled out by over 900 Jordanians. Through different analytics, this study will uncover any correlation between obesity and several other variables.

1. Introduction

Obesity is an unnaturally large culmination of fat within a human’s body. It produces a dangerous risk to the individual’s life and its quality. It can lead to the introduction of many diseases, with heart disease being the most prominent.

1 in 8 people in the world are diagnosed with obesity, and because of its danger and widespread influence, it’s important to study its factors and independent variables.

This study aims to do so by studying the main variables and how they correlate to obesity, like PSQI, BRS, K10, family history of overweightness, age, gender, city, mode of transport, and main food sections eaten at a weekly basis.

2. Methodology

There are three main stages that have gone to this study: Data cleaning, data exploratory analysis, statistical tests, and correlation.

2.1 Data Cleaning:

2.1.1 Removing Null Data:

First thing done, we observed the data directly from its source file and found almost half of it being missing.

For example, rows 1001 – 1741 only had 4 out of 47 variables filled, and rows 1742 – 1882 only had one variable filled!

Therefore, it was wise to only take the first 1000 data points into study, like so:

df = df[0:1000, ]

After this, we detected that there were 129 rows with missing data. To resolve the situation, we took the columns that had the missing data and extracted each of their medians:

na\_cols <- names(which(colSums(is.na(df)) > 0))

medians <- apply(df[apply(df, 1, function(x) !any(is.na(x))), na\_cols], 2, median)

After extracting the medians, we replaced the NA values of each column with their respective medians:

i <- 1

for(value in medians) {

df[na\_cols[i]][is.na(df[na\_cols[i]])] <- value

i <- i + 1

}

2.1.2 Renaming Long Column Names:

Due to the long column names, we found it to be unorthodox to use them each time we call for their respective data. To solve this:

colnames(df) <- c("Gender", "Age", "Family\_History\_Overweight", "Food\_Between\_Meals", "Smoke", "Transportaion",

"City", "Physical\_Activity\_Time", "Breakfast\_In\_Week", "Grains\_In\_Week", "Dairy\_In\_Week",

"Caffeine\_In\_Week", "Seafood\_In\_Week", "BRS\_Q1", "BRS\_Q2", "BRS\_Q3", "BRS\_Q4", "BRS\_Q5", "BRS\_Q6",

"K10\_Q1", "K10\_Q2", "K10\_Q3", "K10\_Q4", "K10\_Q5", "K10\_Q6", "K10\_Q7", "K10\_Q8", "K10\_Q9", "K10\_Q10",

"PSQI\_Q1", "PSQI\_Q2", "PSQI\_Q3", "PSQI\_Q4", "PSQI\_Q5A", "PSQI\_Q5B", "PSQI\_Q5C", "PSQI\_Q5D", "PSQI\_Q5E",

"PSQI\_Q5F", "PSQI\_Q5G", "PSQI\_Q5H", "PSQI\_Q5I", "PSQI\_Q6", "PSQI\_Q7", "PSQI\_Q8", "PSQI\_Q9", "Obesity\_Indicator")

This made it significantly easier.

2.1.3 Data Entry Errors:

Two columns of the dataset is plagued with data entry errors. We fixed them accordingly:

df$K10\_Q10[df$K10\_Q10 == 'alfairly often never'] <- 'almost never'

df$K10\_Q9[df$K10\_Q9 == 'alfairly often never'] <- 'almost never'

2.1.4 Outlier Detection & Replacement:

We ran the Grubbs test for each relevant numeric variable to detect outliers and replace them with their variable’s respective median:

This took several stages as we slowly chipped away at each outlier for each variable till the test came back a p < 0.05 for each numeric variable.

2.2 Feature Engineering and Selection:

All string-based data has been turned to either ordered or unordered factors according to logic. For example, binary variables like ‘family\_history\_overweight’ are nominal variable and therefore are unordered. Multi-value string variables like most metric questions that go within the calculations of PSQI, BRS, and K10 are ordinal and are ordered factors.

Numeric variables were left alone as they had no solvable issues. The largest issue was that the data was varied between doubles and integers, which we unfortunately had to deal with.

The highlights of our work within this section are as follows:

* + - Factorizing PSQI questions 1 through 9.
    - Calculating sleep time out of PSQI\_Q1 & PSQI\_Q3, and therefore, calculating sleep efficiency.
    - Calculating the Pittsburgh Sleep Quality Index for each raw by summing the numerated factors.
    - Factorizing the BRS questions 1 through 6.
    - Accounting for the negative nature of 2, 4, 6 BRS and subtracting them from 6 in order to make them positive.
    - Calculating BRS by summing all numerated BRS factors.
    - Factorizing K10 questions.
    - Fixing values of K10 question variables, i.e. 'alfairly often never' to ‘almost never’.
    - Calculating K10 by summing all K10 numerated factors.
    - Categorizing each metric and putting the categorized metric in new columns:

PSQI\_cat <- cut(PSQI, breaks = c(0, 4, 21), labels = c('good', 'poor'), right = T)

BRS\_cat <- cut(BRS, breaks = c(5, 13, 21, 30), labels = c('low resilience', 'normal resilience', 'high resilience'), right = T)

K10\_cat <- cut(K10, breaks = c(9, 19, 29, 50), labels = c('likely to be well', 'likely to have a moderate disorder', 'likely to have a severe disorder'), right = T)

* + - Factorizing ordinal variables: Breakfasts\_In\_Week, Grain\_In\_Week, Dairy\_In\_Week, Caffeine\_In\_Week, Seafood\_In\_Week, Obesity\_Indicator.
    - Factorizing nominal variables: Gender, Family\_History\_Overweight, Food\_Between\_Meals, Smoke, Transportation, City.
    - All numeric variables have been left untouched.

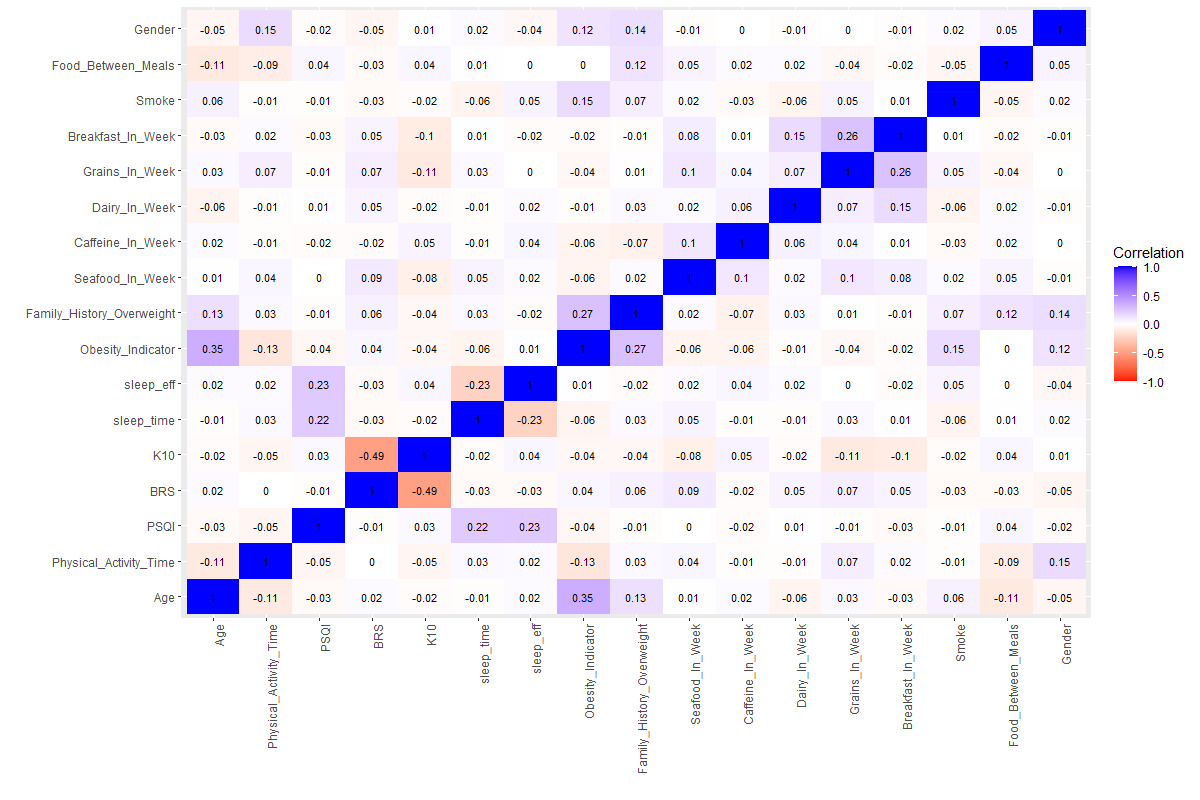
2.3 Exploratory Data Analysis

2.3.1 Correlation

For almost all correlations, we used ‘spearman’ correlation due to its wide use and flippant attitude towards non-normality. Yes, even with some nominal variables, simply because all nominal variables are binary.

First, we determined whether the most important variables have a significant correlation with obesity (P < 0.05), otherwise this dataset does not provide sufficient evidence to prove correlation.

This is the resulting heat map:



2.3.1.1 Correlation Heatmap

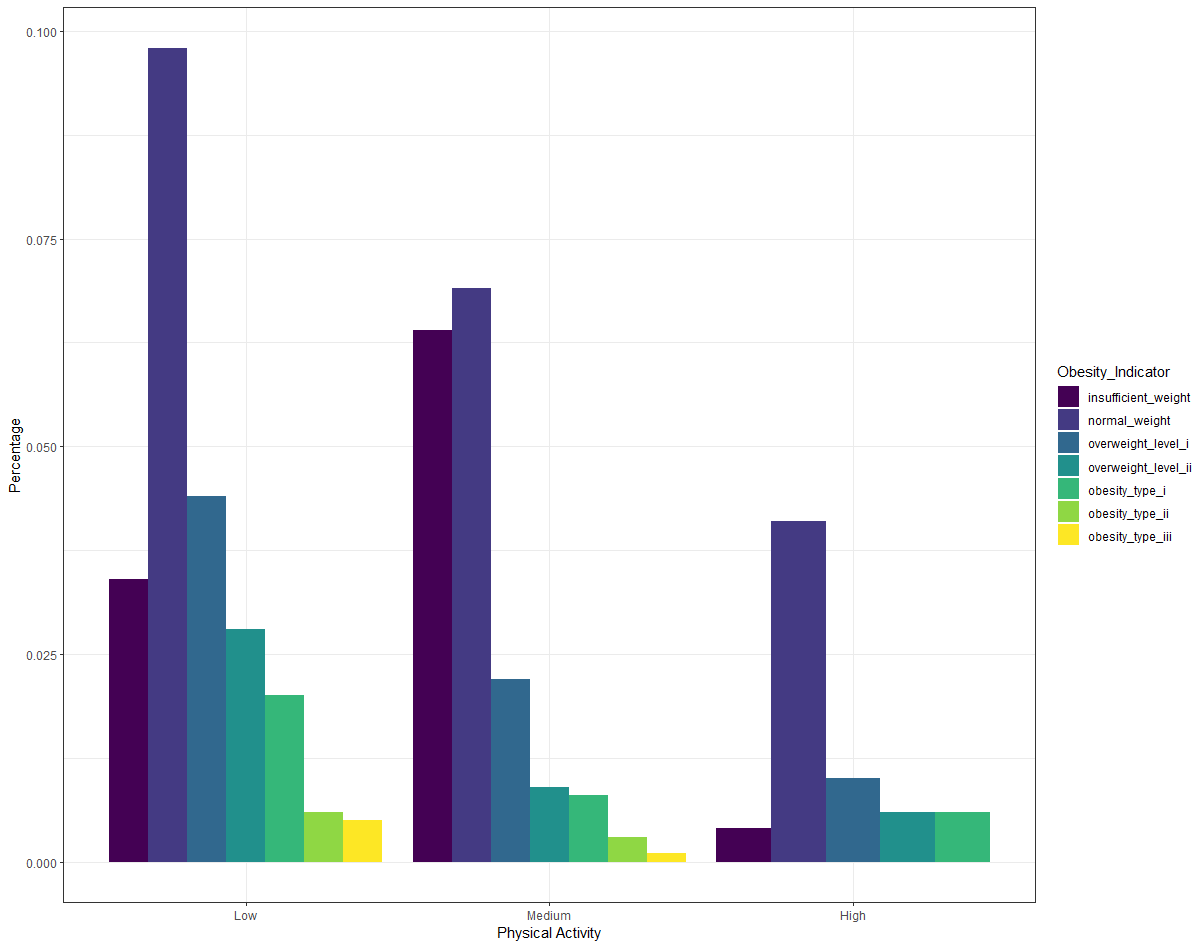
The most unique correlations to obesity were as follows:

2.3.1.2 Correlation between variables and obesity

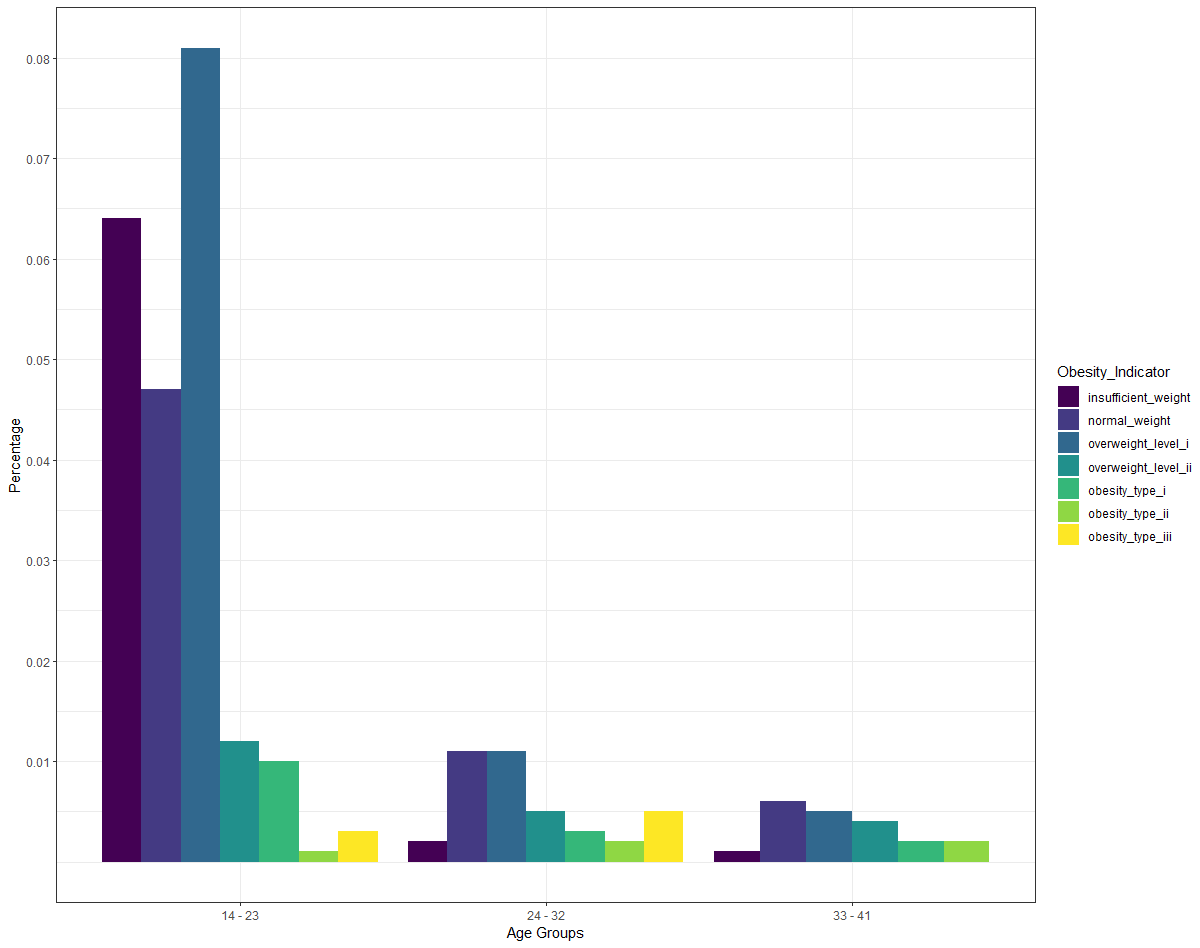
|  |  |  |
| --- | --- | --- |
| Variable | Corr % | P-val |
| Age | 35% | 2.2e-16 (0) |
| Physicial\_Activity\_Time | -14% | 4.591-e6 |
| Family\_History\_Overweight | 27% | 2.2e-16 (0) |
| Smoke | 15% | 0.0001905 |
| Gender | 12% | 8.291-e6 |

The rest of the variables have either P > 0.05 or are nominal with multiple variables (City, Transportation)

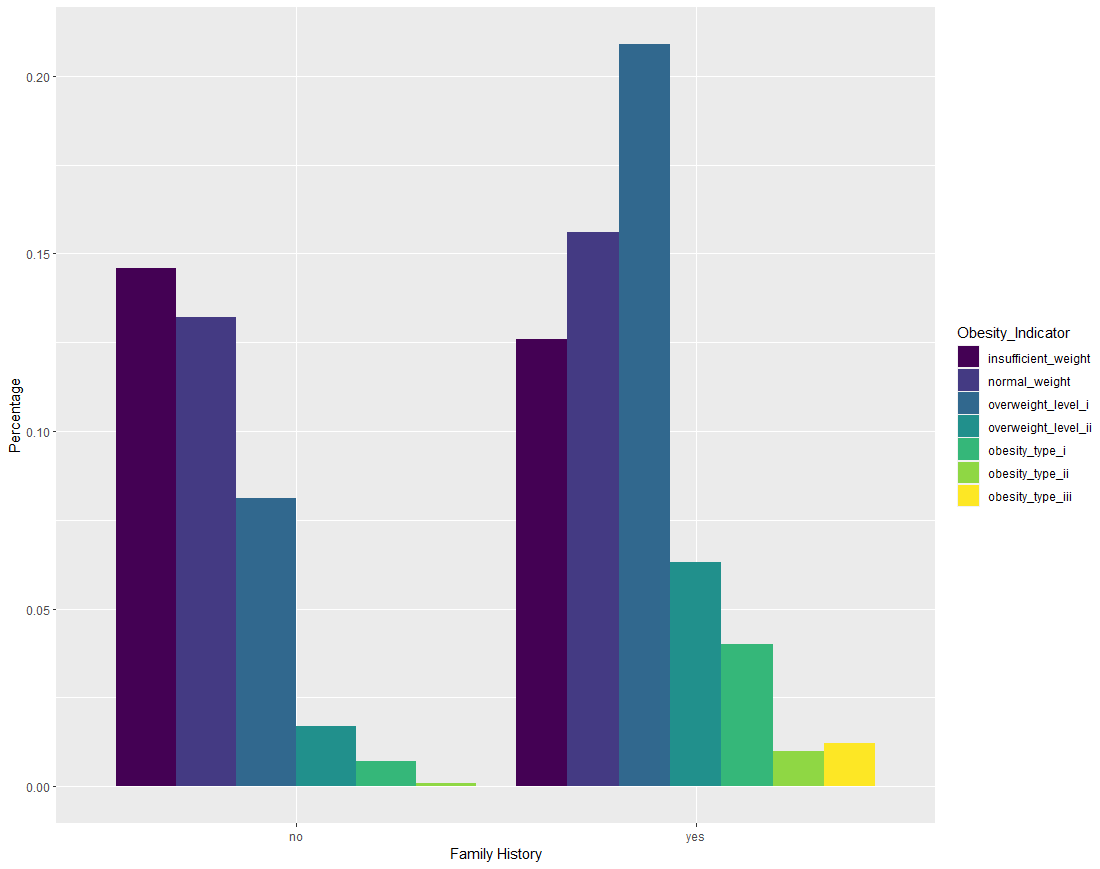
The visualization of the variables is as follows:



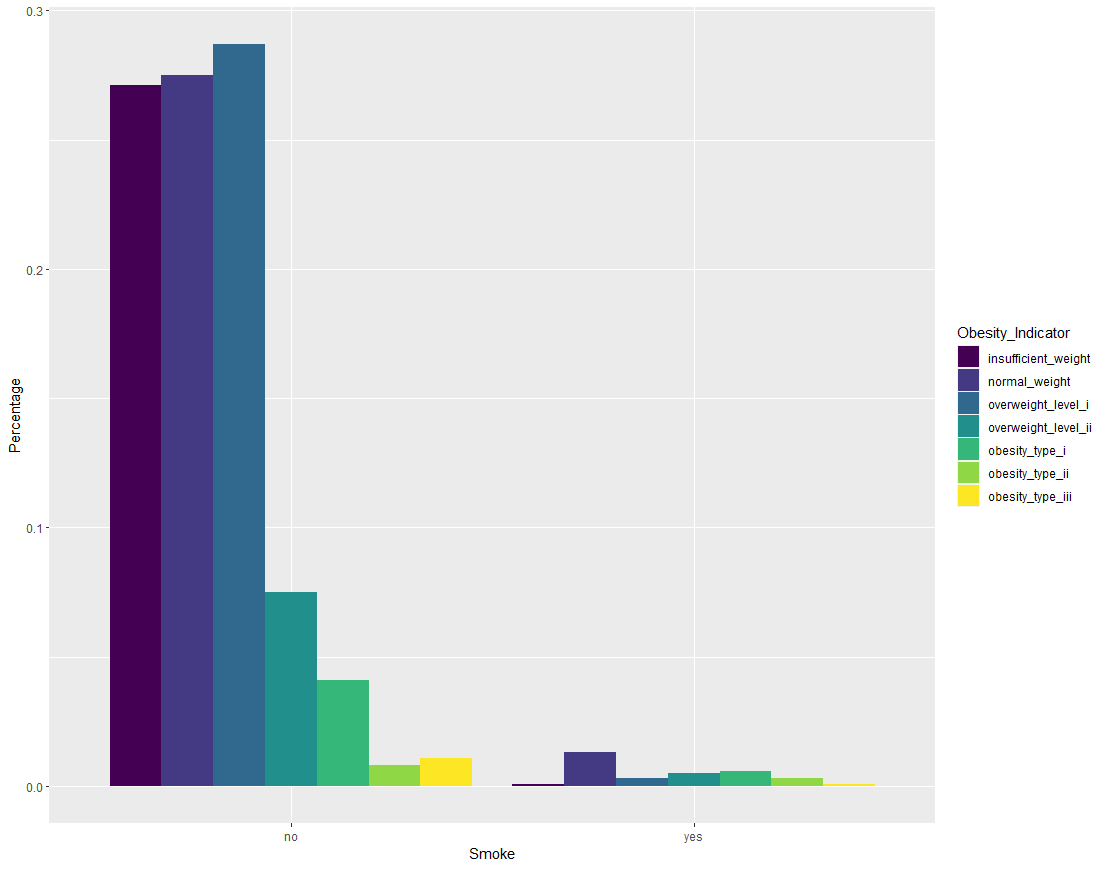
2.3.1.3 Histogram between Physical Activity and Obesity



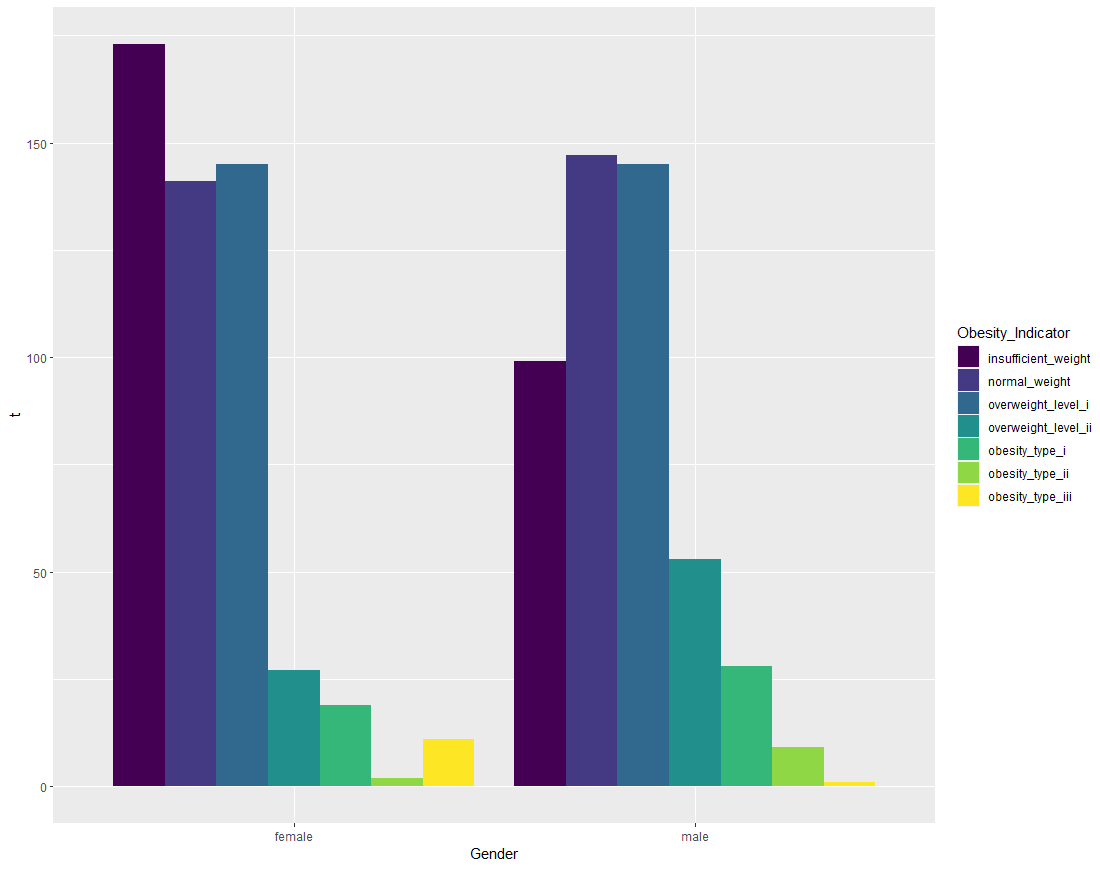
2.3.1.4 Histogram between Age and Obesity



2.3.1.5 Histogram between Family History Overweight and Obesity



2.3.1.6 Histogram between Smoke and Obesity



2.3.1.7 Histogram between Gender and Obesity

2.3.2 Visualizing Nominal Multi-valued Variables

To see the relationship between obesity and nominal multi-valued variables, we can only use the chi-square test and Cramer’s V:

2.3.2.1

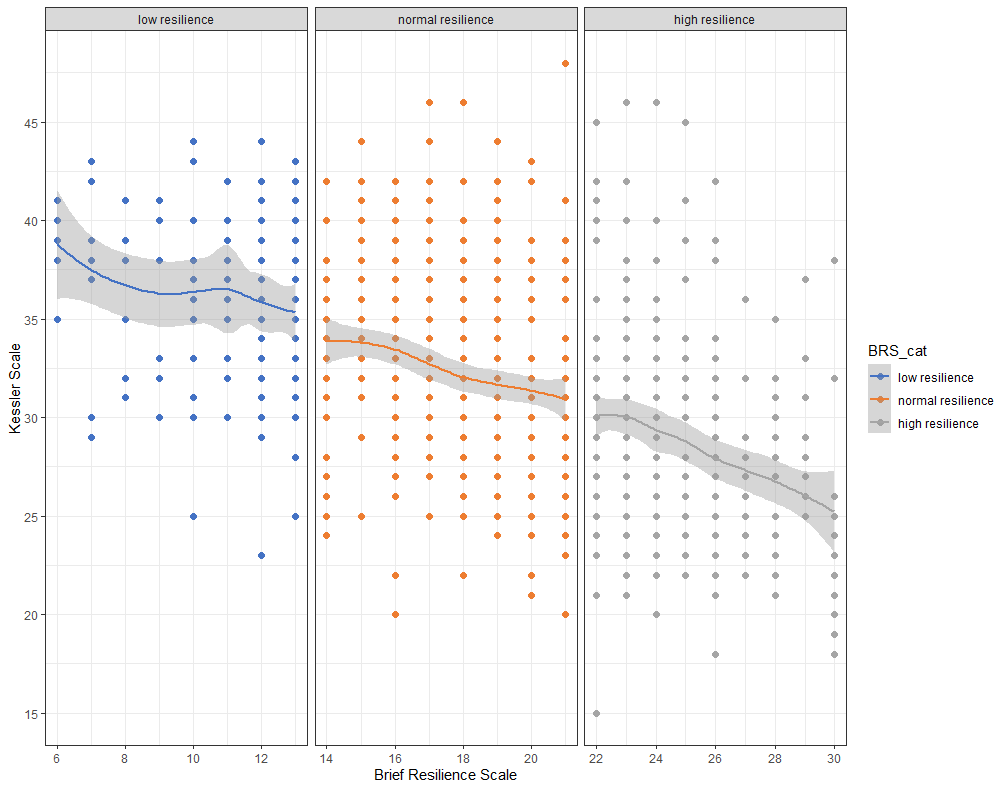
|  |  |  |
| --- | --- | --- |
| Variable | Corr % | P-Val |
| Transportation | 14% | 3.336-e8 |
| City | 6% | 0.7472 |

3. Results:

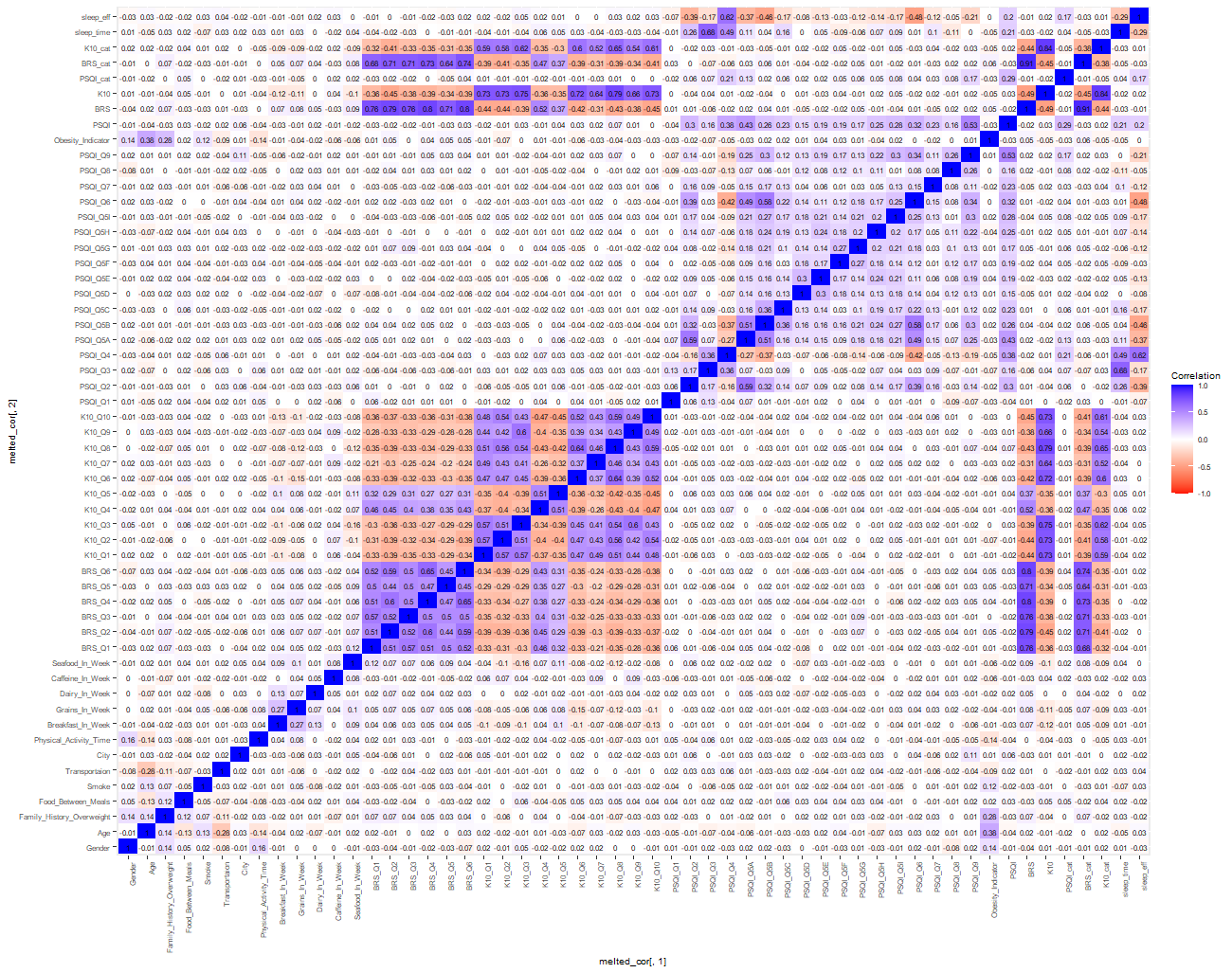
Concludingly, this study has found no sufficient evidence to prove correlation between Obesity and PSQI, BRS, or K10 metrics. However, it found correlations for several other variables, both numerical and categorical, like: Physical Activity Time, Gender, Smoke, Transportation, etc…

We find it strange that we found no evidence of at least one of the major metrics, i.e. PSQI, BRS, and K10, to have had a correlation with obesity, though interestedly, if we go back to Figure 2.3.1.1, we find a significant relationship between both BRS and K10 metrics, suggesting an intent to investigate the relation between both of the variables, but that’s for another study.

4. Appendix



4.1 Scatterplot & LOESS of K10 & BRS



4.2 Correlations of Everything (Warning: Some correlations are incompatible!)