Data 2010 Project

Lucas Berzuk, Ethan Robson, Hugo Ng, Rodell Salonga

March 15, 2024

Data Anaylsis

This data comes from a study performed in Mexico, Peru and Colombia. It has 17 attributes and 2111 data points. The 17 attributes comprise 4 different variable types: 6 continuous, 2 ordinal, 5 categorical, and 4 binary. This data consists of individual's behavioural patterns such as eating habits and physical condition, as well as their level of obesity. It is important to note that up to 77% of this data has been synthetically generated because of a greatly unbalanced number of samples between the different obesity categories in the sample data. The purpose of the data generation was to create data points within the obesity categories with lower counts. This lead to an overall data set with uniform counts for the different obesity levels as displayed in Figure 1. This was an online survey that was 16 questions with a variety of responses. It was accessible for 30 days for users to complete. The researchers use the equation Body Mass Index (BMI) = Weight / (Height^2) to come up with the body mass index of every individual and then they compared it to data that was provided by the WHO to come up with the obesity level categories described in Table 1.

Table 1: Categorizing Obesity Levels

Obesity Level	Insufficient Weight	Normal Weight	Overweight Level I	Overweight Level II	Obesity Type I	Obesity Type II	Obesity Type III
BMI	<18.5	18.5-22.9	23.0-25.9	26.0-29.9	30.0-34.9	35.0-39.9	>40

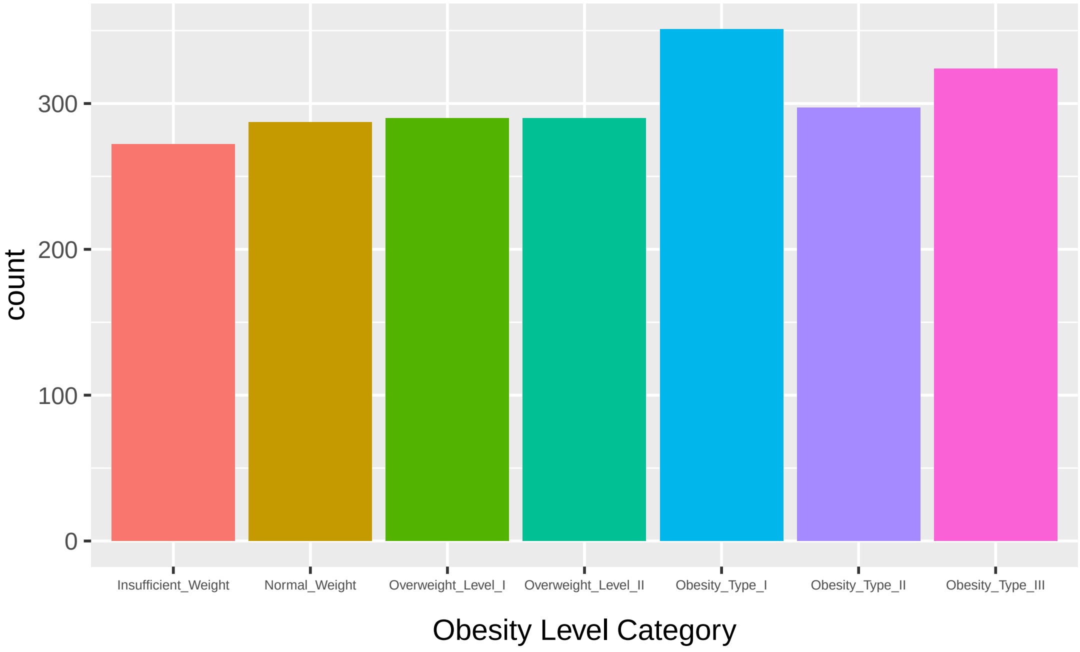

Figure 1: Obesity Level Counts

Current Progress

At this stage in our data analysis, we are observing the correlation between the obesity level and the other variables of the data. The given data has classified individuals into categories given their BMI as described above. However, this grouping of individuals based on BMI values may hide information so we used the weight and height given in the data set to calculate the exact BMI value of each individual. We will use these BMI values to assess the correlation between the other variables of the data.

To determine the correct method to calculate correlation, we used a Kolmogrov-Smirnov Test to test the normality of the calculated BMI variable. We rejected the null hypothesis at a 5% level of significance, which told us that the BMI of the individuals is not normally distributed and so we will use **Spearman's** method to assess correlation. Since this method requires ordinal or continuous data we will determine the correlation of each continuous and ordinal variable with the BMI values. From Table 2 below, we observe that some variables have a positive correlation while others have a negative correlation with BMI. It is also clear that some of these variables have stronger relationships than others.

Table 2: Correlation

Variable	Age	Vegetable Consumption	Meals Quantity	Water Consumption	Physical Activity	Screen Time
Correlation	0.4006	0.2605	-0.0526	0.1593	-0.1686	-0.0752

For the categorical and binary variables, we cannot apply Pearson's or Spearman's method to determine correlation. For each variable we will determine if it is independent with BMI. To do this we will apply **Pearson's Chi-Squared Test of Independence**. Since two categorical variables are required in this hypothesis test, we will use the obesity level categories created by partitioning BMI values as described above. We have conducted this hypothesis test with the categorical variable describing alcohol consumption and obtained an extremely small p-value. This indicates that obesity level and alcohol consumption are dependent and motivates our interest in evaluating their relationship closer and looking at the other categorical variables.

Future Direction

In our current exploration of the data, we have determined that relationships exist between the variables of the data and BMI values of individuals. This directs our interest to explore a scoring function that uses the various variables to produce a score that could predict an individual's obesity level. An important note is that these scores would not include weight and height since these are directly used in the calculation of BMI. We plan to use BMI as the *Gold Standard* and validate our scoring values against the individual's BMI value. This scoring function would be useful for when the weight and height of individuals is missing or difficult to measure but their level of obesity is desired.

In the process of creating this model we will also determine the variables that have the strongest positive and negative correlation with obesity. Since the variables of the data mainly describe behaviours of individuals, we will assess which of these are possibly associated with excess body fat or with healthy weight levels. This will allow us to suggest to the reader behaviours to pursue and behaviours to avoid to reduce their risk of obesity. We acknowledge the growth of obesity in recent years and the associated health consequences, and we hope to discover insightful and actionable information in our future analysis.