DATA2010 Group Project - Progress Report

Abhay Sharma, Darshil Dave, Max Kaplan, Kurt Galvez

2024-03-15

Descriptive Analysis

Our team has chosen to delve into the realm of health and wellness, with a particular focus on the factors contributing to obesity. The data set 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 from the countries of Mexico, Peru and Colombia.

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 data set 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.

Figure 1 below shows a correlation matrix between all our continuous variables. We see that the strongest positive correlation is between BMI and weight. Additionally, we can also see that none of the other continuous variables seems to be strongly correlated with weight which begs the question of whether obesity can be attributed to a single factor/bad-habit, or does it require multiple bad-habits/factors to occur.

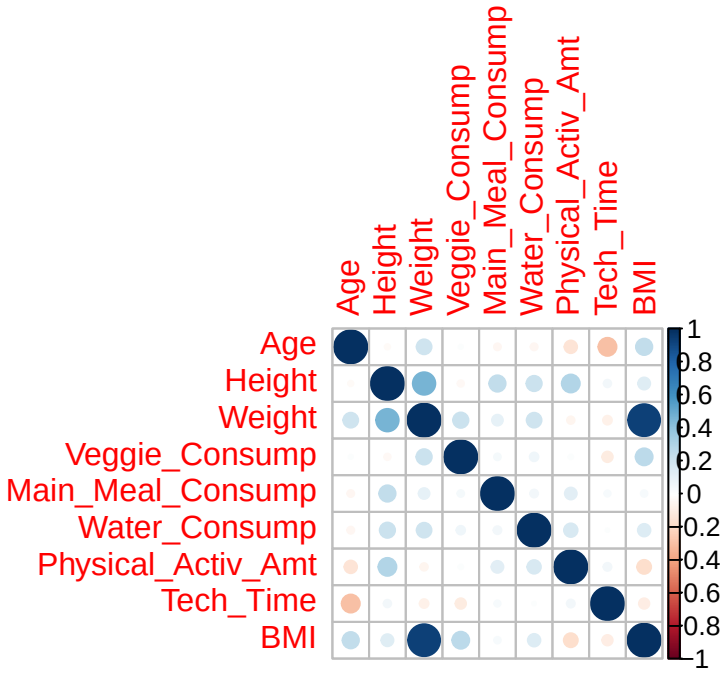

Figure 1, Correlation Matrix

From analyzing the data and variables we have, we have decided to move forward with finding out whether a combination of habits are required in order to become obese. We can express this in terms of a hypothesis:

H_o : Obesity is the result of a single bad habit

H_a : Obesity is the result of multiple bad habits working together in combination

Future direction

To prove this hypothesis, we can perform multiple tests that will give us insights into the true cause of obesity, be it one, or many variables. To start, since we already have to correlations of all numeric variables, we can perform an ANOVA test to determine significant differences in weight for the categorical variables for each individual variable.

Another way to test the correlation of the categorical variables would be a chi-square test with the obesity level variable to determine any associations.

We can also perform t-tests in order to test all the variables using the same test for standardization purposes. This can be done by comparing the mean weight of individuals with and without certain habits.

In order to test the effect of multiple variables on an individuals weight, we can use multiple linear regression analysis using weight as the dependent variable with all the habits as independent variables to understand the combined effect of certain habits on obesity. Another test that can help identify habits that cooperate to cause obesity is factor analysis, which will uncover patterns among collectively occurring variables relating to weight.