**Diet Research Midterm Report**

**BTC1855H Coding in R**

**Leen Madani**

**July 26, 2023**

**Dr. Sebnem Kuzulugil**

**Brief summary and description of the dataset:**

The dataset, Diet\_R, was obtained through the stats community project of the University of Sheffield ([www.sheffield.ac.uk](http://www.sheffield.ac.uk/)). The exploratory data analysis (EDA) revealed that the dataset consists of 78 observations for each person following three different diets. It includes 7 columns representing different variables: person (participant number), gender (1 for male, 0 for female), age (in years), height (in cm), pre.weight (weight before the diet in kg), diet (diet type 1, 2, or 3), and weight6weeks (weight after 6 weeks in kg). The variables are stored as integer (int) data type, except for weight6weeks which is stored as double. All variables, except for weight6weeks, are stored as integer <int> data type, reflecting their integer values. Weight6weeks is stored as a double class data type to accommodate numeric values with decimal points, allowing for greater precision.

The summary for each variable is depicted in Figure 1, providing valuable insights into the dataset (Fig. 1). In this visualization, the x-axis represents the values of each variable, while the y-axis represents the count or number of people exhibiting specific characteristics. The dataset consists mostly of numerical variables, except for gender and diet, which are categorical.

The analysis shows that the gender column has 2 missing values, with 43 females and 33 males. The average age of the participants is 39.15 years, with the youngest participant being 16 years old and the oldest participant being 60 years old. The age graph follows more of a normal distribution. The average height is 170.82 cm, with the shortest participant measuring 141 cm and the tallest measuring 201 cm. There is a potential outlier at 141 cm, indicating a significantly shorter height. In addition, the overall distribution of the height graphs is skewed to the right, which means that the higher values shift the mean towards the right, to a higher value.

In terms of weight, the average pre-diet weight is 72.53 kg, ranging from 58 kg to 103 kg. There is a potential outlier at 103 kg, representing a significantly higher weight. The participants were almost evenly distributed across the three diet types, with 24 participants on diet type 1, 27 participants on diet type 2, and 27 participants on diet type 3.

After six weeks on the given diet, the average weight was 68.68 kg, ranging from 53 kg to 103 kg. There is a potential outlier at 103 kg, indicating a significantly higher weight after the diet. The average weight after six weeks (68.68 kg) is lower than the average pre-diet weight (72.53 kg). However, further analysis, such as statistical tests, is necessary to determine if this difference is statistically significant.

To gain more insights, it would be beneficial to compare the three diets and assess if there are significant differences between them. Overall, the basic EDA conducted in R provides valuable information about the dataset, including potential outliers in height and weight, as well as the distribution of participants across the different variables.

A screenshot of a graph

Description automatically generated

**Figure 1. Histograms illustrating the distribution of the seven variables in the Diet\_R dataset.** Each histogram represents the values of the specific variables along the x-axis, with a default bin size of 10. The y-axis denotes the counts or frequency of occurrences. The dataset consists of 78 participants who followed three different diets.

**Preprocessing to the dataset:**

**Missing Values:** To handle missing values in the gender column for person 25 and 26, we will impute "female" or "male" based on a comparison of their characteristics, such as height and weight, with those of the female and male participants in the study (Table 1). First, we calculate the average height and weight for females and males using all available observations. Then, we determine the differences between the height and weight of person 25 and 26 and the average height and weight of females and males, respectively. Comparing the height and weight of person 25 with the averages, the height is closer to the average height of males (175.24cm) while the weight is closer to the average weight of females (67.11kg). Based on this information, person 25 might be more likely to be classified as female. Comparing the height and weight of person 26 with the averages, both the height and weight are closer to the averages of males (175.24cm and 79kg) rather than females. Based on this information, person 26 might be more likely to be classified as male. By using this approach, we impute the missing gender values for persons 25 and 26 with more confidence and reliability.

**Table 1. Observations with missing values entries created using R.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Person | gender | Age | Height | pre.weight | Diet | weight6weeks |
| 25 | NA | 41 | 171 | 60 | 2 | 60 |
| 26 | NA | 32 | 174 | 103 | 2 | 103 |

**Categorical Variables:** As part of the preprocessing to the dataset, the categorical variables were subject to change in the type of class. Firstly, the 'gender' variable was converted into a categorical factor with levels '0' and '1', representing 'Female' and 'Male', respectively. This step ensured a more intuitive representation of gender in the data. Similarly, the 'Diet' variable was also converted into a categorical factor with levels '1', '2', and '3', corresponding to 'Type 1', 'Type 2', and 'Type 3', respectively. By employing these transformations, the data is now more amenable to statistical analysis and facilitates clearer visualization of the results.

**Methodology:**

Please refer to the R-script for the methodology.

**Findings:**

1. **Correlation Plot:**

Based on the findings from the correlation plot, several key observations can be made (Fig. 2). Weight before diet (pre.weight) exhibits a moderately strong positive correlation (around 0.8, dark bluish color) with gender, indicating a relationship. However, as gender is categorical, its interpretation in the correlation plot is limited. There seems to be no correlation (~ 0, white color) between weight before diet (pre.weight) and age, suggesting that age alone does not significantly affect weight. This supports the understanding that weight is influenced by various factors, such as genetics, lifestyle, and environmental elements, rather than age alone. Additionally, weight before diet shows no meaningful correlation (almost white color) with the type of diet provided to participants. Weight change exhibits a very weak negative correlation (light pink color) with weight before diet, implying that those with higher initial weights may face slightly more challenge in losing or gaining weight after the diet. There is no correlation (white color) between age and gender, indicating that the study included participants randomly and in a diverse manner, without any preferential selection based on specific age-gender combinations. However, this finding does not hold significant practical meaning, as the type of diet is controlled by researchers, and age alone does not appear to influence diet selection. Furthermore, weight change displays a weak positive correlation (light blue color, around 0.2) with age, suggesting that older individuals might experience a slight challenge in losing or gaining weight after the diet. Lastly, the correlation between weight change and the type of diet is moderately negative (almost -0.4), indicating varying effects of different diets on the participants’ weights.

**Limitations:** It's essential to acknowledge that since gender and type of diet are categorical variables, their interpretation in a correlation plot is limited. For example, while the plot shows correlations, it does not provide information on whether being male or female, or following a specific type of diet (type 1, type 2, or type 3), is positively or negatively associated with weight change. Additionally, the correlation plot does not provide information on the direction of weight change (i.e., weight loss or gain) for the participants. Moreover, the correlation plot might not capture more complex relationships among variables. Additional analysis, such as using box plots or advanced statistical techniques, may be necessary to gain a more comprehensive understanding of the relationships between variables.

A diagram of different colored squares

Description automatically generated

**Figure 2. Correlation plot of selected variables from the Diet\_R dataset.** Relationships between the variables pre.weight (weight before diet), gender, age, diet, and weight\_change (difference between weight after diet and before diet) are illustrated. Correlations are represented on a scale from -1 to 1, with negative correlations shown in red tones and positive correlations displayed in blue tones. The plot was created using ggplot2 in R.

1. **Scatterplot**

Based on the findings from the scatterplot, several observations can be made (Fig. 3). First, the study seems to have a relatively balanced distribution of female and male participants, indicated by the similar number of data points in both the top (female) and bottom (male) plots. Additionally, the vast majority of participants, except for four individuals (three female and three male), experienced weight loss after following any type of diet. The individuals who lost weight are evident from the blue data points clustered below the dashed line, indicating a negative weight change. Furthermore, two participants (one female and one male) were able to maintain their weight, as shown by data points exactly on the dashed line. In contrast, four participants (two female and two male) gained weight, represented by the presence of red data points with a positive weight change. The maximum weight loss observed in both female and male participants was around 8.75 kg, while the maximum weight gain was approximately 2.5 kg.

Although it is challenging to determine the most effective diet for weight loss from the scatterplot alone, upon closer examination of the female plot, it appears that type 3 diet has a notable concentration of data points with high values of negative weight change, signifying that those who adopted type 3 diet experienced the greatest weight loss.

Overall, the observed trends of weight loss among the majority of participants suggest that the diets employed in the study might have a positive impact on weight management. However, to draw more definitive conclusions, further analysis and additional information about the specific diets are necessary. Nonetheless, the scatterplot serves as a valuable visualization for these variables, guiding further investigation into the relationships between diet, gender, and the dependent variable weight change in the study participants.

A screenshot of a graph

Description automatically generated

**Figure 3. Scatterplot displaying the relationship between Type of Diet and Weight Change (kg) by gender.** Each point on the plot represents an individual participant's weight change after following a specific type of diet. The x-axis represents the different types of diets (Type 1, Type 2, and Type 3), while the y-axis represents the corresponding weight change in kilograms. Data points above the dashed gray line indicate weight gain, and those below the line indicate weight loss. The plot was created using ggplot2 in R**.**

1. **Box plot**

Based on the findings from the boxplot, several observations can be made (Fig. 4). First, a general trend is observed where males tend to have higher initial weight values compared to females, evident by the yellow panels situated at higher weight before diet values in the plot. Additionally, the boxplot reveals notable insights about the distribution of weight change among the different types of diets. Specifically, type 1 diet exhibits the most considerable number of outliers, representing individuals experiencing drastic weight changes.

Furthermore, type 2 diet appears to be administered to individuals with the highest initial weight, particularly in the male group. Upon examining the medians for each diet type, it is observed that type 1 and type 2 diets demonstrate comparable median weight loss for females (around 3), but males show a higher median weight loss (around 3.75) for type 1 and even more substantial median weight loss (around 4.2) for type 2. As for type 3 diet, it stands out with the highest median weight loss for females (approximately 7), while the median weight loss for males remains lower (around 3.75). The median is also in the center of most boxes except for females given type 3, where the median is situated right at the bottom of the box. In addition, the median generally appears well-balanced within most categories, such as males and females given type 1 and type 2 diets. This suggests a relatively even distribution of weight change values between the upper and lower halves of the boxes for these groups. However, for females given type 3 diet, the median is located right at the bottom of the box. This observation indicates that the lower half of the weight change data for this group is more concentrated, while the upper half may contain some outliers or extreme values that are pulling the median downwards.

These insights provide valuable information regarding the relationship between initial weight, gender, and the effect of different diet types on weight change. It is important to consider these observations when evaluating the potential impact of various diets on weight management in different groups.

Top of Form

A chart with different colored lines

Description automatically generated

**Figure 4. Box plot displaying the relationship between Weight Before Diet (kg) and Weight Change (kg) stratified by gender.** Each box plot represents the distribution of weight change for specific participants, grouped according to their initial weight before the diet. Outliers are depicted in pink. The data is further split into different diet types (Type 1, Type 2, and Type 3), displayed in separate panels for easy comparison. The x-axis represents the weight before the diet, while the y-axis represents the weight change. The plot was created using ggplot2 in R.