202380-Fall 2023-ITCS-3162-051-Introduction to Data Mining: Project 1

Meelad Doroodchi
09/12/2023

Kaggle Dataset Link:
https://www.kaggle.com/datasets/rukenmissonnier/age-weight-height-bmi-analysis

Introduction to Data:

The dataset I am using is named "Age, Weight, Height, BMI Analysis" which has 741 rows of individual records of patients and has 5 columns that include key variables which are: Age, Height(in meters), Weight(in kilograms), Body Mass Index, and Body Mass Index class which includes: "Obese Class 1," "Obese Class 2,", “Obese Class 3” "Overweight," "Underweight," “Normal”. I am especially interested in this dataset because I believe it is very open-ended but the features allow for the dataset to be visualized and understood in many ways that can be useful in society.

Introduction to Questions:

Some of the questions I would like to solve using the variables would be to understand the relationship between age and body mass index to get an understanding of how age has an impact on someone's body mass index. Along with this, I would like to look at the correlation between weight and body mass index class to be able to get an understanding of the age differences in each BMI class. Looking at different trends with specific ages will be able to help me come to show trends between age groups. The correlation between how much someone weighs and which specific class they are in. Those are just a few of the questions or problems that I would begin to look at with this dataset, I believe that height and body mass index have somewhat of a correlation which allows me to look at a few different relationships to be able to understand the significance of these features on someone's health. A few more questions that I could possibly look at along with these is the distribution of each individual record to show the specific BMI class, BMI index, and the age of each person to be able to look at the overall trend of the dataset.

Additionally, a good thing to look at which BMI in this dataset is the most common and to look at the maximum and minimum to get an understanding of how weight is equivalent to BMI class. This dataset allows me to be able to explore and get an understanding of different trends and connections!

Adv. & Disadv. of Questions:

A few positive benefits that would come from this data is that people would be able to get an understanding of their body types and gain more knowledge knowing who they are. I believe this dataset allows for visualizations and relationships to be shown between age, weight, height, and a person's body and body mass index. This dataset looks at a controlled age range and is a very clean dataset with many open-ended ways to make connections and show that certain relationships can help people gain more understanding!

A few negatives of this dataset include that the age range is from 15-61 which can be both good and bad since it is controlled, but there are many people outside of this age range. As well as this it does not speak about certain health issues that these 741 patients could possibly have that could impact certain things like weight which would ultimately have an impact on body mass index and body mass index class.

Pre-Processing the Data:

Before getting into any visualizations, the data has to be pre-processed. Data cleaning and preprocessing are essential because they ensure data accuracy, improve model performance, handle missing values, outliers, and inconsistencies, standardize data formats, and prepare categorical data for analysis. These tasks create a reliable foundation for data mining, enhancing the quality and effectiveness of the analyses and modeling. In data cleaning a few things that need to be cleaned include the quality of the data. There are many times the data has missing values or the dataset is not fully ready to be used, the programmer has to go back and clean the data before being able to pull any analyses from it. In this project, the dataset that was chosen was quite clean and the implemented data was easy to understand, but to make it simple I changed the weight from kilograms to pounds and the height from meters to feet. I left both of the columns in the dataset so that everyone can be able to view the numbers in both pounds/kilograms and meters/feet.

Data Visualization/Story:

The first thing I needed to do before creating any visualizations was to implement any packages I needed into **RStudio**. I downloaded the dplyr and the ggplot2 packages into my console. Using methods in RStudio to create a bar chart, two scatterplots, and a histogram. After looking at some of the questions from above visualizations began to be created to look at relationships. Look at some of the visualizations and connections made below.

Visualization #1: Bar chart of Distribution of BMI classes by Age: Story

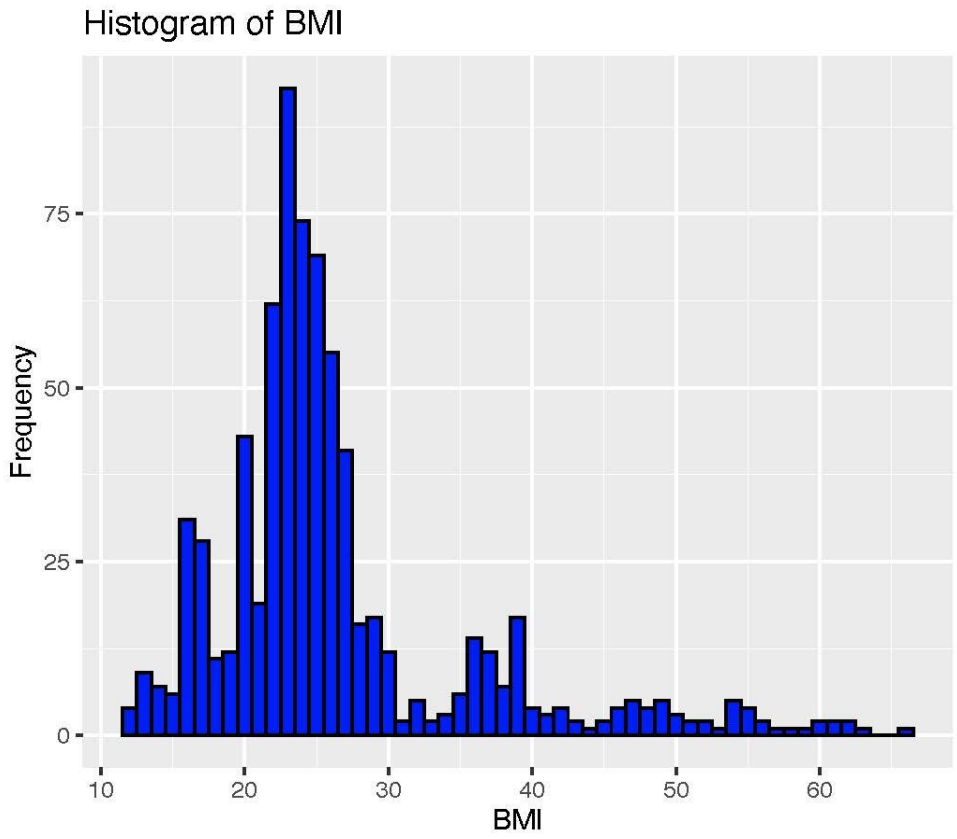
- This visualization below looks at the distribution of BMI classes by age and the number of people in the dataset pertaining to each BMI class. This visualization was a bar chart created in RStudio and this distribution looks at the dataset as a whole and looks at the distribution of age with the BMI class. After looking at the plot some conclusions that were made include the majority of normal weighted people in this dataset ranged from ages 15-50 and not any after the age of 50 which is interesting to see. This looks at the question of age with body mass index which shows that many people over the age of 50 regardless of other health issues will most likely find themselves not at 'normal weight'. This plot shows the BMI class distribution by age and the number of individuals in this dataset in that specific class and age group by color.

Visualization #2: Scatterplot of Age vs. BMI and classifying by BMI Class: Story

This visualization below is a scatter plot of Age vs. BMI while classifying the individual records in this dataset by BMI class as well which answers the following question of what is the specific BMI that is matched with each BMI class and age trends with BMI class and BMI itself. After looking at this plot some of the answers that were concluded include that the BMI for 'Obese Class 3' is 40+ as it is the heaviest, the BMI for 'Obese Class 2' is 35-39.5, the BMI for 'Obese Class 1' is 30-34.5, the BMI for 'Overweight' is 25-29.5, the BMI for 'Normal Weight' is 19-24.5, and the BMI for 'Underweight' is less than 19. This graph displays different trends to show that this dataset was split in terms of types of patients and records. There are many different chunks of each group significant enough to see on this graph which was interesting to see. It also goes to show that there were not many people past the age of 55 in 'Obese Class 3' which could be due to health.

Visualization #3: Histogram of BMI: Story

The visualization below is a Histogram of BMI with frequency or amount of records on this dataset to show the maximum and minimum BMI of this dataset along with the most common BMI. After looking at this dataset some conclusions that were drawn include that the minimum BMI in this dataset is 12 and the maximum is 66 with the most frequent or mode being 23 which is 'Normal Weight' which proves the fact that many individuals in this dataset were of a healthy and normal weight but there were still many different records of people that were severely overweight or underweight according to their BMI's and the conclusions drawn from the previous



Visualization #4: Scatterplot of Age vs. Height vs. Weight: Story

This visualization below is a scatterplot to show the relationship in this dataset between age, weight, and height. The colors on the graph depict the individuals in this dataset's weight, the brighter red is the heavier. Some conclusions that can be drawn include that many of the people that are heavier are taller than 6', at least the people that are towards 'Obese Class 3'. This visual along with some of the other ones shows that many people between the ages of 25 and 40 are not underweight but it's the majority of the ages around it which was interesting to see. This along with the other visualizations was able to create a story and relationships that answered some of the problems at hand.

Impact:

The impact of the visualizations above can be both positive and negative. This dataset with these visualizations is positive because it shows the trends and differences from this dataset which can allow health professionals and even humans themselves to understand their bodies and where they fit into. These four visualizations above show many things including the bar chart that highlights different age groups and shows how many people past the age of 50 in this dataset were no longer 'Normal Weight' shows that Age can be a deviation in BMI. Including this, these visualizations were able to show the scale of the BMI classes and the threshold for the BMI in each class. With the visualization it is simple to understand where someone can fall under. Some negative impacts of this dataset can be generalization, since this dataset is only 741 but over a broad range it can not be generalized for everybody because as stated previously, many different

people have underlying health conditions that can effect many different things pertaining to BMI, BMI class, and weight which is not spoken about in this dataset. This study is only looked about from an outer perspective of each person and not certain things each person may have. Overall, these visualizations collectively tell a compelling story about this dataset's characteristics and relationships, and addressing and showing key patterns and trends within the data.

References:

Missonnier, Ruken. "Age, Weight, Height, BMI Analysis." *Kaggle*, 1 Sept. 2023, www.kaggle.com/datasets/rukenmissonnier/age-weight-height-bmi-analysis.

Code: RStudio

Code for Pre-Processing Steps

```
> bmi$Weight_in_Pounds <- bmi$Weight * 2.205
> bmi$Height_in_Inches <- bmi$Height * 3.28084
> colnames(bmi)[7] ="Height_in_Feet"
```

Code for Bar chart of Distribution of BMI classes by Age:

```
> ggplot(data = bmi, aes(x = Age, fill = BmiClass)) +
+     geom_bar() +  # Create a bar chart
+     labs(title = "Distribution of BMI Classes by Age", x = "Age", y = "Count") +
+     scale_fill_brewer(palette = "Set3")  # Custom color palette (optional)
```

Code for Scatterplot of Age vs. BMI and classifying by BMI Class

```
> ggplot(data = bmi, aes(x = Age, y = Bmi, color = BmiClass)) +
+     geom_point() +
+     labs(title = "Scatter Plot of Age vs. BMI", x = "Age", y = "BMI") +
+     scale_color_brewer(palette = "Set3")  # Choose a color palette
```

Code for Histogram of BMI

```
> ggplot(data = bmi, aes(x = Bmi)) +
+   geom_histogram(binwidth = 1, fill = "blue", color = "black") +
+  labs(title = "Histogram of BMI", x = "BMI", y = "Frequency")
```

Code for Scatterplot of Age vs. Height vs. Weight

```
> ggplot(data = bmi, aes(x = Age, y = Height_in_Feet, color = Weight_in_Pounds)) +
+ geom_point(size = 3) +
+ labs(title = "Scatter Plot of Age vs. Height vs. Weight",
+ x = "Age", y = "Height (in feet)") +
+ scale_color_gradient(low = "blue", high = "red")  # Choose a color scale
```