Derek Graves

07/06/2024

Lab 2: Working with a data set

* Progress 1 - Screen shot of notebook has code and result

A screenshot of a computer

Description automatically generated

* Progress 2 - Analysis from info()
* # data instances: 544
* # features: 4
* feature names: height, weight, age, male
* # missing values: 0
* non-numerical features:  0
* Progress 3 - screenshot of code/result

A screenshot of a computer

Description automatically generated

* Progress 4 – analysis
  1. Are the data instances sorted on any of the attributes? No
  2. What are the units of height? Centimeters
  3. What are the units of weight? Kilograms
  4. What are the minimum, median and max age?
     + Minimum: 0
     + Median: 27
     + Max: 88
  5. What two different features have the highest correlation? Height and Weight.
* Progress 5 - Screen shot of graphs

A screenshot of a computer

Description automatically generated

* Progress 6 – Analysis

1. describe distributions:
   * Height: The height histogram appears to be left skewed, with most data points occurring on the taller side of the data.
   * Weight: Weight distribution appears to be bimodal, as there are two clear peaks of data on the lower and higher ends of the data.
   * Age: Age appears to be right-skewed, with most points lying on the lower-age side of the graph.
2. explanation of age vs weight
   * The age-weight scatter plot shows weight generally increases with age in earlier years showing natural growth and development.
   * It also shows that younger individuals tend to have lower weights, with weights increasing with age. The weight increase seems to stabilize in adulthood once natural development completes, and there appears to be a bit of decrease in weight with the older age groups, possibly attributed to things like loss of muscle mass, health conditions, lifestyle changes that generally come with advanced age, etc.
   * There looks to be a decent amount of variation with weight in each age group, which shows that weight can also be influenced by other factors like genetics, diet, and other health factors.
3. Age histogram
   * The age histogram shows that the majority of individuals in this dataset are young.
4. comparison to modern
   * Modern populations in today’s world either tend to show a more balanced age distribution, or an aging trend, with a higher proportion of the population in the older ranges. This is likely due to things like observed lower birth rates, or increased life expectancy due to advancements in modern medicine.

* Progress 7 - Code/Analysis

A screen shot of a computer screen

Description automatically generated

* + Age at change in character of the data set
    - It appears that the character of the dataset changes at around 18 years old, which correlates with the transition from adolescence to adulthood, where rapid growth in height in the data begins to level off.
* Progress 8 - Screen shot of head

A screenshot of a computer

Description automatically generated

* Progress 9 - Screen shot head and value counts

A screenshot of a computer

Description automatically generated

* Progress 10 - Screen shot of age v bmi plot

A screenshot of a computer screen

Description automatically generated

* Progress 11 - Screen shot of data set print

A screenshot of a computer

Description automatically generated

* Progress 12 - Screen shot of masked plot

A screenshot of a computer

Description automatically generated

* Computation 13 - Ratios computed for Male:Female
  + Adult data frame
  + Training data frame
  + Test data frame

A screenshot of a computer program

Description automatically generated

* Computation 14 - Stratified Ratios computed for Male:Female
  + Adult data frame
  + Training data frame
  + Test data frame

A screenshot of a computer

Description automatically generated