Statistics for Data Science
UE21CS241A
PES University, Bangalore

Case Study for the Datathon

Teaching Assistant: Anushka Hebbar
anushkahebbar@pesu.pes.edu

Chapter 1

Introduction

The two sections below, `Background` and `The Case Study` provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset, analyze the features using descriptive statistics such as summary statistics, tables, and graphs, and spin up a suitable predictive model. Happy coding!

Background

In medicine, the term *pulse rate* refers to the number of pulses generated by the heart's activity in the blood vessels during a specific unit of time (usually one minute). As a rule, the pulse rate coincides with the heart rate. Various factors, such as exercise, stress, or certain diseases, can influence the pulse rate in the short or long term.

A normal resting heart rate for adults ranges from 60 to 100 beats per minute. Generally, a lower heart rate at rest implies more efficient heart function and better cardiovascular fitness. For example, a well-trained athlete might have a normal resting heart rate of closer to 40 beats per minute. Although there's a wide range of normal, an unusually high or low heart rate may indicate an underlying problem.

The Case Study

A medical researcher wants to study the effects of various factors on pulse rates. The researcher records the height, weight, gender, smoking preference, activity level, and resting pulse rate of a set of undergraduate students. The students are made to complete three 'laps' walking up and down a nearby set of stairs. Then, the researcher records the students' pulse rates again. We will use the dataset thus obtained to discover insights about the set of students surveyed.

Data Description

Collect the dataset from the GitHub repository provided. The dataset contains data collected from students with the following 7 variables:

1. `Active`: Pulse rate (beats per minute) after exercise
2. `Rest`: Resting pulse rate (beats per minute)
3. `Smoke`: Indicator variable for smoking preference (`1=smoker` or `0=nonsmoker`)
4. `Gender`: Indicator variable for the gender (`1=female` or `0=male`)
5. `Exercise`: Typical hours of exercise (per week)
6. `Hgt`: Height of the student (in inches)
7. `Wgt`: Weight of the student (in pounds)

Chapter 2

Problem Set

1. Explore the high-level characteristics of the dataset using standard functions in the Pandas Python library. How many students have been surveyed for the collection of the dataset? Analyze the set of attributes available in the dataset. List out the type of each attribute: is it categorical or numerical? Is it nominal, ordinal, discrete, or continuous?
2. A summary statistic quantitatively summarizes the data in a particular feature using a single number. There are two popular types of summary statistics: measures of central tendency and measures of dispersion. For each attribute in the dataset, figure out the most effective measure of central tendency and state its value. Also, determine the standard deviation and range of values in each column.
   Hint: Think about whether the mean, median, or mode is a more insightful measure of central tendency for a feature depending on its type (categorical or numerical).
3. Dealing with missing data is a crucial step in the process of data analysis. It can have a significant impact on the conclusions drawn from the data. Determine the number of missing records in each column of the dataset. Figure out the most effective way to deal with missing data in each column. Transform each column and make sure none of the columns have any missing values.
4. Determine the number of classes of every categorical variable in the dataset. Plot bar graphs for all the categorical attributes to determine the frequency of the classes in each attribute. For example, plot a bar chart to determine the number of smokers and non-smokers. Do so for all categorical variables. What inferences about the survey can you make based on the visualizations?
5. Plot histograms for all the numerical attributes to determine the distribution of the values in each attribute. Discuss the modality and skewness of each distribution. What inferences about the survey can you make based on the visualizations?
6. Visualize the active and resting pulse rates of all the participants using box plots. What is the interquartile range of the active pulse rates? How does this compare to the IQR of the resting pulse rates? Now, visualize the active and resting pulse rates when grouped by gender. What inferences can you make from the graph? Repeat the same exercise for both the pulse rates when grouped by the amount of exercise and the smoking preference, separately. Draw inferences from all the graphs obtained.
7. Explore the relationship between every pair of numerical variables by constructing a scatterplot matrix. Create a separate scatterplot matrix for each categorical variable: this will allow you to color code the scatterplots using the categorical variable. By pure visualization alone, which pair of numerical variables depict the highest linear relationship?
   Now, validate these insights by using Python to calculate the Pearson Correlation Coefficients for each pair of variables in the datasets. Check if these values follow the pattern in the scatterplots created earlier.
8. Run a simple linear regression model for `Active` (the active pulse rate) against `Rest` (the resting pulse rate). Make sure to create proper training and test sets and evaluate the goodness of the trained model. What is the R^2 score of your model? State the fitted model equation using the slope and intercept values.