Indian Institute of Technology Dharwad
CS214: Artificial Intelligence Laboratory

Lab 1: Data Visualization and Statistics from Data

Note

There are two problems that need to be worked out in the lab. **Problem Statement 1** is compulsory. **Problem Statement 2** is a bonus.

Problem Statement 1

A dataset related to Indian diabetes, containing medical attributes, is provided in a CSV file ('pima-indians-diabetes.csv'). The dataset includes various features that provide insights into the conditions leading to diabetes. These features are: `pregs`, `plas`, `pres`, `skin`, `test`, `BMI`, `pedi`, `Age`, and `class`.

Write a Python program to perform the following tasks:

1. Display the first 10 tuples of the given dataset using the 'head()' function.

2. Display the structure of the data to provide details about the number of entries, data types, and memory usage of the dataset using the 'info()' function.

3. Generate the descriptive statistics for each numerical attribute using the 'describe()' function to provide the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum of the columns.

4. Calculate and display the following statistical measures for each attribute using the respective functions: mean using ‘mean()‘, median using ‘median()‘, mode using ‘mode()‘, minimum using ‘min()‘, maximum using ‘max()‘, and standard deviation using ‘std()‘. Additionally, compute the quartiles (Q1, Q2, and Q3) using ‘quantile()‘ and the Interquartile Range (IQR) by subtracting Q1 from Q3.

5. Generate scatter plots to explore the relationships between `skin`, `Age`, `BMI`, `pregs`, and `plas`.

6. Plot histograms for all numerical attributes and the kernel density estimation (KDE) curve.

7. Group the data according to the attribute ‘class‘ and bar plots to analyze the distribution of the attributes ‘BMI‘, ‘Age‘, ‘plas‘, and ‘pres‘ for each value of the ‘class‘ variable using the ‘groupby()‘ function.

8. Create boxplots for all attributes to identify outliers and compare distributions.

Problem Statement 2

A dataset related to red variants of the Portuguese "Vinho Verde" wine is given as a CSV file (winequality-red.csv). This dataset contains the values of different physicochemical tests from each sample of red wine [1]. The original goal of the dataset is to model wine quality based on physicochemical tests. The attributes of the dataset based on physicochemical tests are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH value, sulphates, alcohol content, and the last attribute is quality. Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

Write a Python program to perform the following tasks:

1. Convert the given CSV file to a DataFrame and display the last 10 tuples using the ‘tail()‘ function.

2. Display the structure of the data to provide details about the number of entries, data types, and memory usage of the dataset using the ‘info()‘ function.

3. Generate the descriptive statistics for each numerical attribute using the ‘describe()‘ function to provide the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile or Q2), 75th percentile (Q3), and maximum of the columns.

4. Calculate and display the following statistical measures for each attribute using the respective functions: mean using ‘mean()‘, median using ‘median()‘, mode using ‘mode()‘, minimum using ‘min()‘, maximum using ‘max()‘, and standard deviation using ‘std()‘. Additionally, compute the quartiles (Q1, Q2, and Q3) using ‘quantile()‘ and the Interquartile Range (IQR) by subtracting Q1 from Q3.

5. Generate scatter plots to explore the relationships between each attribute `citric acid`, `residual sugar`, `chlorides`, `pH`, `sulphates`, `alcohol`.

6. Plot histograms for all numerical attributes and the kernel density estimation (KDE) curve.

7. Group the dataset by the variable `quality` and bar plots to analyze the distribution of the attributes `citric acid`, `residual sugar`, `free sulfur dioxide`, and `total sulfur dioxide` for each quality group. Use the ‘groupby()‘ function to achieve this and visualize the results.

8. Create boxplots for all attributes to identify outliers and compare distributions.

Note

Kindly upload the completed Jupyter Notebook for each of the problem statements separately to Moodle for evaluation.

Deliverables

- Python scripts containing the implementation of all the specified tasks.
- Visualizations generated for scatter plots, histograms, bar plots, and boxplots.
- A summary of findings from the statistical analysis and visualizations.

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, "Modeling wine preferences by data mining from physicochemical properties," *Decision Support Systems*, vol. 47, no. 4, pp. 547–553, 2009.