Day 3 26.03.2024

1. Analyze a given dataset and perform exploratory data analysis to summarize the data, understand its distribution, and identify outliers. Apply appropriate statistical measures such as mean, variance, covariance, and correlation to quantitatively analyze the dataset. Discuss the results and interpret the findings in terms of data summarization, distribution characteristics, and the presence of outliers. Provide visualizations, statistical calculations, and written explanations to support your analysis.

2.You are provided with a dataset of gene expression levels across different tissue types. Perform the following analysis:

i. Develop a novel measure of asymmetry that accounts for skewness, kurtosis, and tail behavior in the distribution of gene expression levels.

ii. Assess the impact of data transformation on downstream analyses, such as clustering or differential expression analysis, and discuss the benefits and limitations of the selected transformation method.

3. Imagine you are a teacher at a high school, and you have just graded the final exam papers of your students. You have a class of 50 students You have graded the final exam, and the scores range from 40 to 100 points. You record the exam scores of all 50 students. After collecting the exam scores, you decide to create a histogram to visualize the distribution. You find that the histogram of exam scores is roughly bell-shaped, with most students clustering around the mean score of 75. In this case, what you would say that the distribution has no skewness or is approximately symmetric?

4.Your dataset contains scores from a survey conducted among 20 participants The scores are as follows: Scores:

85, 78,92, 89, 67, 76. 94, 82, 91, 88, 75, 81, 96. 90, 79, 83. 77, &5, 98, 72

Calculate the mean, median and quartiles of the scores and create & box pt to visualize the distribution. Additionally, identify potential outliers in the dataset.

5.You are a data analyst working for a social media platform. As part of your analysis, you have a dataset containing user interaction data, including the number of likes received by each post. Your task is to develop a Python program that calculates the frequency distribution of likes among the posts.

6.A weather station wants to know the most common types of weather in their area. They have a list of all the weather conditions that have occurred in the past year, along with the number of times each weather condition has occurred. Write a program that will calculate the frequency distribution of weather conditions and print out the most common weather type.

7. A digital marketing team is keen on understanding the engagement patterns of visitors on their website. The team tracks the time spent on the website (in minutes) for a sample of 20 visitors. This data provides valuable insights into user behavior and helps the team optimize the website's content and features for better user engagement.

Time Spent (minutes): [8, 12, 5, 15, 10, 20, 7, 18, 25, 8, 12, 22, 15, 10, 30, 7, 18, 15, 20, 12]

- Calculate the median of the time spent on the website. The median provides a measure of central tendency and represents the middle value in the dataset.

- Calculate the interquartile range (IQR) to understand the spread of the data. IQR measures the range of values within which the central 50% of the data falls.

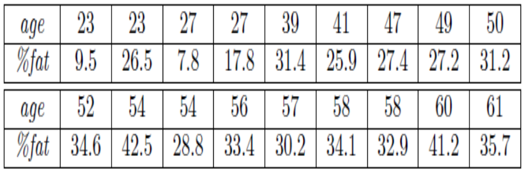
- Create a box plot to visually represent the distribution of time spent on the website. The box plot displays the median, quartiles, and potential outliers.

8. You are a data analyst working for a social media platform. As part of your analysis, you have a dataset containing user interaction data, including the number of likes received by each post. Your task is to develop a Python program that calculates the frequency distribution of likes among the posts.

9. You are a data analyst working for a marketing research company. Your team has collected a large dataset containing customer feedback from various social media platforms. The dataset consists of thousands of text entries, and your task is to develop a Python program to analyze the frequency distribution of words in this dataset. Your program should be able to perform the following tasks:

- Load the dataset from a CSV file (data.csv) containing a single column named "feedback" with each row representing a customer comment.
- Preprocess the text data by removing punctuation, converting all text to lowercase, and eliminating any stop words (common words like "the," "and," "is," etc. that don't carry significant meaning).
- Calculate the frequency distribution of words in the preprocessed dataset.
- Display the top N most frequent words and their corresponding frequencies, where N is provided as user input.
- Plot a bar graph to visualize the top N most frequent words and their frequencies.

10. Suppose a hospital tested the age and body fat data for 18 randomly selected adults with the following result.



age	23	23	27	27	39	41	47	49	50
%fat	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2

age	52	54	54	56	57	58	58	60	61
%fat	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

- Calculate the mean, median and standard deviation of age and %fat using Pandas.
- Draw the boxplots for age and %fat.
- Draw a scatter plot and a q-q plot based on these two variables