Descriptive Statistics

This assignment aims to develop students' skills in data analysis and statistical methods using Python. By working with real-world datasets, students will learn how to identify and handle outliers, deal with missing data, analyze measures of central tendency and variability, and create meaningful visualizations. Additionally, students will gain an understanding of potential sources of bias and confounding variables, further enhancing their critical thinking and decision-making abilities in data analysis.

Through a combination of hands-on Python coding and thoughtful interpretation of results, this assignment fosters a comprehensive understanding of data analysis techniques and their practical applications in different scenarios.

Tasks:

1. **Outlier Identification and Handling:** In this task, you will work with a real-world dataset to identify and handle outliers. Choose a dataset (e.g., from Kaggle, UCI Machine Learning Repository) from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, that exhibits outliers or extreme values. Write a Python script that identifies and handles the outliers using at least two methods (e.g., z-score, interquartile range). Use visualization techniques to demonstrate the impact of the outliers on measures of central tendency and variability.
2. **Bias and Confounding Variables Identification:** Identify potential sources of bias or confounding variables in the dataset selected in Task 1 above and discuss how they might impact the analysis.
3. **Handling Missing Data:** Develop and justify an appropriate statistical method to handle missing data in the dataset selected in Task 1.
4. **Analysis of Mean and Median Values:** In this task, you will analyze a dataset to understand the difference between mean and median values. Choose a dataset from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, where the mean and median values differ significantly. Write a Python script to calculate and visualize the mean and median values of the dataset. Interpret the results and provide insights into what the difference means for the dataset. Propose solutions to handle this discrepancy and implement them using Python.
   Compare and contrast the effectiveness of four different measures of central tendency and variability in capturing the characteristics of the data.
5. **Data Visualization**: In this task, you will use Python to create visualizations that effectively communicate data distribution. Choose a dataset from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, and create basic plots to visualize the data distribution (e.g., histogram, boxplot). Analyze the plots to gain insights into the data distribution and interpret the results.
6. **Measures of Central Tendency and Variability:** In this task, you will calculate and interpret measures of central tendency and variability using Python. Choose a dataset from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, and write a Python script to calculate the mean, median, mode, range, variance, and standard deviation of the dataset. Interpret the results and discuss how the measures of central tendency and variability relate to the data distribution.

7. **Data Cleaning**: In this task, you will use Python to clean a dataset and prepare it for analysis. Choose a messy dataset (e.g., missing values, inconsistent formatting) from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, and write a Python script to clean the dataset. Use appropriate methods to handle missing values, remove duplicates, and convert data types. Visualize the cleaned dataset to demonstrate the impact of the cleaning process.
8. **Group Analysis**: In this task, you will use Python to conduct group analysis on a dataset. Choose a dataset from the list of "Repositories for Finding Suitable Datasets," located in Class Resources, and write a Python script to group the data by a categorical variable (e.g., gender, age group). Calculate measures of central tendency and variability for each group and visualize the results using appropriate plots. Interpret the results and discuss any differences between the groups.

DataSet:

https://www.kaggle.com/datasets/rumanaamin/comprehensive-property-rental-listings-of-dhaka

https://www.kaggle.com/datasets/elmoallistair/population-by-age-group-2021

https://www.kaggle.com/datasets/jamiewelsh2/nba-player-salaries-2022-23-season

https://www.kaggle.com/datasets/krishnaraj30/finance-loan-approval-prediction-data

References:

Goyal, C. (2021). Why you shouldn't just delete outliers. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/why-you-shouldnt-just-delete-outliers/

Codebasic.(2020). Outlier detection and removal: z score, standard deviation | Feature engineering tutorial python. https://www.youtube.com/watch?v=KFuEAGR3HS4

Coursera. https://www.coursera.org/learn/foundations-of-data-science/lecture/2vmAG/welcome-to-module-2

Mirko, S.(2019). Python Statistics Fundamentals: How to Describe Your Data. https://realpython.com/python-statistics/

Clinfo. Stuck in the middle – mean vs. median. https://www.clinfo.eu/mean-median/

Jim, F. Statistics By Ji. Mean, Median, and Mode: Measures of Central Tendency. https://statisticsbyjim.com/basics/measures-central-tendency-mean-median-mode/

Lumen. Introduction to Statistics. Skewness and the Mean, Median, and Mode. https://courses.lumenlearning.com/introstats1/chapter/skewness-and-the-mean-median-and-mode/