Scripting for Data Analysis
IST-652 - Mini Project 2

SUID:202315111
Sai Akash Addala
Email: saddala@syr.edu

The data consists of sales tax collection from each state. I've retrieved the data from kaggle.

Data Exploration steps:

- To comprehend the dataset's structure, including the quantity of rows and columns, a preliminary overview was conducted.
- To find outliers and abnormalities, descriptive statistics were computed for numerical variables, such as mean, median, and mode.
- To find trends, data distributions were shown using box plots, scatter plots, and histograms.

Data Cleaning steps:

- Missing Data: Various techniques, such as imputation and removal, were used to manage missing data points when missing values in the dataset were found.
- Outliers: Statistical techniques were used to identify outliers, and depending on how they affected the study, choices were made on whether to keep, modify, or eliminate them.
- Data Type Conversion: To maintain consistency, data types were changed where necessary. For instance, date variables were formatted consistently.
- Data Integrity: To make sure that the numbers fell under reasonable and allowed ranges, the data integrity was examined.

Comparison Questions:
Question 1:

- Unit of Analysis: Numeric month and month.
- Comparison Values: Sales tax per month.
- Computation: We calculate the sales tax in each month for the given months while summarizing and plotting.

Question 2:

- Unit of Analysis: Year
- Comparison Values: Sales tax in terms of months.
- Computation: Categorizing the dates into each month and summarizing the values with visualization.

Question 3:

- Unit of Analysis: State and month.
- Comparison Values: We compare state and month with sales tax.
- Computation: The two collections are state and month which are compared to sales tax and visualized.

Program Description:
Using the NumPy, Matplotlib, and Pandas modules, the study was conducted using the Python programming language. Google Colab was used to complete the code presentation and documentation. These instruments were applied to data cleaning, exploration, and statistical analysis. If there are any missing values in the dataset, we locate them by using a variable that reads and includes the dataset file. Additionally, we look for anomalies.
One strategy for filling in the missing values in a dataset is to take the mean of all the values in that particular column and fill it in. Additionally, we may inspect the dataset's initial six rows and obtain information about it. The dataset's columns are also visible to us as graphs.

Result of Analysis:
For Question 1, the analysis revealed the sales tax for each month which gives the highest tax paying on that month. So that we can see why the tax is paid that much in that month.

For Question 2, the analysis provides the sales tax sales tax which are summarized into each month, so that it is easy to calculate per month and the visualization makes it seem easy.

For Question 3, the analysis gives the comparisons between state and sales tax summaries. And month and sales tax summaries. It also shows the visualizations of the data summary.

Sales tax department can use this information to make strategic decisions and optimize operations thanks to its clear data, actionable insights, and thorough visualizations.

In summary, this paper presents the results of the sales tax dataset analysis along with an overview of the data exploration process, data cleaning procedures, two comparative questions, and analytical tools. The aim of this analysis is to improve overall performance and assist the Sales tax department in making more informed decisions.