HW4
Report

This report describes the analysis of the dataset from the CSV file "2023 June Unemployment Rate by County (Percent).csv." It generates various graphs to examine and analyze data, identifies probable outliers, and provides insights into the dataset.

Data Reading and Preparation:

It separates the feature variables (x) and the target variables (y) before extracting the data for visualization.

Box Plot by Region:

The code makes a box plot of unemployment rates categorized by 'Region Name'. Each box reflects the regional distribution of unemployment rates.
The plot's title is "Box Plot of Unemployment Rates by Region."

Scatter Plot:

The code removes non-numeric values and converts the remainder to floats.
It then plots each data point in a scatter plot of unemployment rates.
The plot's name is "Unemployment Rate Scatter Plot."

Quantile-Quantile (QQ) Plot:
The code computes quantiles and generates a QQ plot for the data. This figure is used to determine whether the data has a normal distribution. Theoretical quantiles are displayed against sample quantiles, and a normal distribution is represented by a diagonal red dashed line.

Plot of Density with Outliers:
This plot helps visualize the distribution of unemployment rates. The title is "Unemployment Rate Density Plot (With Outliers).

Bar Plot:
The code calculates the total unemployment rates for each region by summing the rates of all data points in the region.
It then creates a bar plot showing the total unemployment rates for each region.

Histogram:

A histogram of the unemployment rates is generated with 20 bins to show the frequency of rate values.

Simpson Paradox:

I can't find Simpson Paradox in my dataset because Simpson's Paradox can be subtle and may not be present in every dataset. It frequently occurs when there are confounding variables or hidden factors influencing the connections between variables within subgroups.

Discuss possible problems you plan to investigate for future studies. For future studies, I plan to investigate:

Exploratory Data Analysis (EDA): To acquire deeper insights, try undertaking more complete EDA.

Regression (Supervised Learning)

• Classification (Supervised Learning) • Text Mining

• Clustering (Unsupervised Learning) • Neural Network and DeepLearning

Finally, this report highlights the critical procedures involved in preparing a dataset comprising unemployment rates by county. This function prepares the data for future analysis, visualization, or modeling by importing the data, investigating its features, addressing missing values, and exporting the cleaned dataset.