Part 3: Advanced Analytics

7. Statistical Analysis

Problem Statement: Perform a statistical analysis on a given dataset to identify significant trends and correlations. Provide a summary of your findings.

Summary of Statistical Analysis

Dataset Overview: The dataset contains information about various car attributes including cylinders, displacement, horsepower, weight, acceleration, year, origin, and mpg.

1. Data Exploration and Correlation Analysis:

- **Correlation Matrix:** Calculated and visualized the correlation between continuous variables such as cylinders, displacement, horsepower, weight, acceleration, and mpg.
- **Heatmap:** Used a heatmap to display correlations, indicating how strongly different variables are related.

2. Distribution of Continuous Variables:

- **Histograms:**
    - **MPG:** Displayed the distribution of mpg with a histogram and KDE plot.
    - **Horsepower:** Analysed the distribution of horsepower similarly.

3. Scatter Plots for Relationship Analysis:

- **MPG vs. Weight:** Visualized the relationship between mpg and weight.
- **MPG vs. Displacement:** Visualized the relationship between mpg and displacement.

4. Analysis by Categorical Variables:

- **Frequency Counts:**
    - **Cylinders:** Showed the frequency of different numbers of cylinders.
    - **Origin:** Showed the frequency of car origins.
- **Average MPG:**
    - **By Cylinders:** Calculated and visualized the average mpg by the number of cylinders.
    - **By Origin:** Calculated and visualized the average mpg by car origin.

5. Hypothesis Testing:

- **T-Test:** Compared the mean mpg between cars with 4 cylinders and 6 cylinders, finding whether the difference is statistically significant.
- **ANOVA:** Tested for significant differences in mpg across different numbers of cylinders, assessing if at least one group mean is significantly different.

Conclusion: The analysis provides insights into the relationships between various car attributes and mpg, and evaluates statistical significance in differences observed across different groups of cylinders. The visualizations and statistical tests help in understanding patterns and trends within the dataset.

8. Predictive Analytics

Problem Statement: Build a predictive model to forecast sales for the next quarter using

historical sales data. Explain the steps taken and the rationale behind your model choice.

Summary of Predictive Analytics Code

Objective: Building a predictive model to forecast sales for the next quarter using historical sales data.

1. **Load and Inspect Data:**
   - **Data Loading:** Reading the dataset from a CSV file.
   - **Initial Inspection:** Checking the first few rows, shape, and columns of the dataset.
   - **Check for Missing Values:** Calculating the percentage of missing values for each column.
2. **Date Parsing and Pre-processing:**
   - **Check and Convert Dates:** Inferring and converting dates to `datetime` format for `Order Date` and `Ship Date`.
   - **Check Missing Dates:** Verifying any missing values in the date columns.
3. **Feature Engineering:**
   - **Extract Temporal Features:** Creating new columns for `Year`, `Month`, and `Quarter` from `Order Date`.
   - **Aggregate Monthly Sales:** Summarizing total revenue by month and year.
   - **Create Lag Features:** Adding lag features to capture previous months' sales for forecasting.
4. **Model Building:**
   - **Define Features and Target:** Setting up features (`Lag_1`, `Lag_2`) and target variable (`Total Revenue`).
   - **Train-Test Split:** Splitting the data into training and test sets without shuffling (time series data).
5. **Model Training and Evaluation:**
   - **Train the Model:** Using `RandomForestRegressor` for forecasting.
   - **Predict and Evaluate:** Predicting on the test set and evaluate the model using Mean Absolute Error (MAE) and Mean Squared Error (MSE).
6. **Forecast Future Sales:**
   - **Predict Future Sales:** Using the trained model to forecast sales for the next 3 months, updating lag features iteratively.

Key Points:

- **Data Pre-processing:** Ensured correct date parsing and handled missing values.
- **Feature Engineering:** Created temporal features and lag variables to aid forecasting.
- **Model Building:** Employed a Random Forest model for prediction and evaluated its performance.
- **Forecasting:** Generated sales forecasts for the next quarter based on historical data.