Data Science Lab-2

Statistical Techniques for EDA

Objective:

The goal of this lab is to perform exploratory data analysis using statistical techniques to understand the characteristics, distribution, and relationships in NYC yellow taxi trip data. You will compute descriptive statistics, visualize data patterns, and apply inferential statistics to draw conclusions.

Dataset Description

The dataset contains trip-level data of yellow taxi rides in NYC. Each record includes attributes such as pickup/drop-off time, trip distance, fare amount, passenger count, payment type, and more. Refer to the attached data dictionary for field definitions.

Part A: Descriptive Statistics

1. Univariate Analysis

For the following columns, compute and interpret:

- Passenger_count
- Trip_distance
- Fare_amount
- Total_amount
- Tip_amount
- Extra

Tasks:

- Mean, median, mode
- Minimum, maximum
- Standard deviation, variance
- Skewness and kurtosis
- Count and number of missing values

2. Visualizations

Create at least three of the following for selected columns:

- Histogram and Frequency Polygon

- Box Plot and Violin Plot
- Density Plot
- Bar Chart (for categorical columns like Payment_type, RateCodeID)
- Pie Chart (for categorical proportions like VendorID or Store_and_fwd_flag)

Part B: Inferential Statistics

1. Confidence Intervals

Estimate a 95% confidence interval for:

- Mean trip distance
- Mean fare amount
- Mean tip amount

2. Hypothesis Testing

Conduct the following hypothesis tests:

1. **One-sample t-test:**
   H^0: The average tip amount is equal to $2
   H^1: The average tip amount is different from $2
2. **Two-sample t-test:**
   Compare average fare_amount between two Payment_type groups (e.g., credit card vs cash).
3. **Chi-square Test of Independence:**
   Test if Payment_type and RateCodeID are independent.

3. Correlation Analysis

Compute Pearson or Spearman correlation between:

- Trip_distance vs Fare_amount
- Fare_amount vs Tip_amount

Create a correlation matrix heatmap.

Bonus Tasks (Optional for Extra Credit)

- Create a time series plot of trip_count per day or hour.
- Analyze how fare_amount varies with time of day (rush hours).
- Map most common pickup/dropoff zones (if geo-data or zone IDs are mapped).

Deliverables

Submit a detailed **Jupyter Notebook** or **Python/R Script** including:

- Cleaned dataset and handling of missing values
- Descriptive statistical summary with plots
- Inferential statistics results with interpretation
- Conclusion and insights drawn from the data

Tools

Python (Pandas, Matplotlib, Seaborn, Scipy, Statsmodels) or R (ggplot2, dplyr, tidyr, t.test, chisq.test)

References

- NYC Taxi & Limousine Commission: Trip Data (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml)
- Attached Data Dictionary
- EDA and Statistics Lecture Materials