Exploratory Data Analysis Lab Report

Course: Data Science Lab (AI)

Roll Number: **U23AI114**

Name: **Prakriti**

August 2025

Contents

Chapter 1

Objective

The goal of this lab is to perform exploratory data analysis (EDA) using NYC yellow taxi trip data. We compute descriptive statistics, visualize data distributions, and apply inferential statistics to draw insights.

Chapter 2

Dataset Description

The dataset contains trip-level data of yellow taxi rides in NYC with attributes such as pickup/dropoff time, trip distance, fare amount, passenger count, payment type, and more.

Chapter 3

Part A: Descriptive Statistics

3.1 Univariate Analysis

Below, we compute summary statistics (mean, median, mode, min, max, std, variance, skewness, kurtosis, missing values) for selected columns.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy import stats
import seaborn as sns

%matplotlib inline

df = pd.read_csv('yellow_tripdata_sample.csv',
                 parse_dates=['tpep_pickup_datetime','
                    tpep_dropoff_datetime'])
df.head()
```

```
cols = ['passenger_count','trip_distance','fare_amount',
        'total_amount','tip_amount','extra']

stats_summary = []
for c in cols:
    series = df[c].dropna()
    stats_summary.append({
        'feature': c,
        'count': series.count(),
        'missing': df[c].isna().sum(),
        'mean': series.mean(),
        'median': series.median(),
        'mode': series.mode().iloc[0] if not series.mode().empty else
           np.nan,
        'min': series.min(),
        'max': series.max(),
        'std': series.std(),
        'variance': series.var(),
        'skewness': series.skew(),
        'kurtosis': series.kurt()
    })

pd.DataFrame(stats_summary).set_index('feature').T
```

3.2 Visualizations

```
# Histogram & Density for trip_distance
plt.figure(figsize=(8,4))
plt.hist(df['trip_distance'].dropna(), bins=30, alpha=0.6, edgecolor='
    black')
plt.twinx()
sns.kdeplot(df['trip_distance'].dropna(), bw_adjust=0.5)
plt.title('Trip Distance Distribution')
plt.xlabel('Distance (miles)')
plt.ylabel('Density')
plt.show()
```

```
# Box Plot & Violin Plot for fare_amount
plt.figure(figsize=(10,4))
plt.subplot(1,2,1)
sns.boxplot(x=df['fare_amount'])
plt.title('Fare Amount Box Plot')
plt.subplot(1,2,2)
sns.violinplot(x=df['fare_amount'])
plt.title('Fare Amount Violin Plot')
plt.tight_layout()
plt.show()
```

```
# Bar Chart for Payment Type
payment_counts = df['payment_type'].value_counts()
plt.figure(figsize=(6,4))
payment_counts.plot(kind='bar')
plt.title('Payment Type Counts')
plt.xlabel('Payment Type')
plt.ylabel('Number of Trips')
plt.show()
```

Chapter 4

Part B: Inferential Statistics

4.1 Confidence Intervals

```
from scipy import stats

# 95% CI for trip_distance, fare_amount, tip_amount
for col in ['trip_distance','fare_amount','tip_amount']:
    data = df[col].dropna()
    ci = stats.t.interval(0.95, len(data)-1,
                          loc=np.mean(data),
                          scale=stats.sem(data))
    print(col, ci)
```

4.2 Hypothesis Testing

4.2.1 One-sample t-test: Tip amount = 2?

```
stats.ttest_1samp(df['tip_amount'].dropna(), 2)
```

4.2.2 Two-sample t-test: Fare amount (credit vs cash)

```
fare_credit = df[df['payment_type']==1]['fare_amount']
fare_cash   = df[df['payment_type']==2]['fare_amount']
stats.ttest_ind(fare_credit.dropna(), fare_cash.dropna())
```

4.2.3 Chi-square Test of Independence

```
contingency = pd.crosstab(df['payment_type'], df['RatecodeID'])
chi2, p, dof, expected = stats.chi2_contingency(contingency)
chi2, p
```

4.3 Correlation Analysis

```
# Pearson & Spearman correlations
df[['trip_distance','fare_amount','tip_amount']].corr(method='pearson')
df[['trip_distance','fare_amount','tip_amount']].corr(method='spearman'
    )

# Heatmap
plt.figure(figsize=(6,4))
sns.heatmap(df[['trip_distance','fare_amount','tip_amount']].corr(),
            annot=True, cmap='coolwarm')
plt.title("Correlation Matrix Heatmap")
plt.show()
```

Chapter 5

Conclusion

In this lab, we performed EDA, computed descriptive statistics, visualized patterns, and applied hypothesis tests and correlation analysis. This provided insights into how fare, tips, and distances are distributed and interrelated in NYC taxi data.