Jupyter Notebook Homework 1.1:

Creating Data Visualizations

Due: 04th March 2025

Report

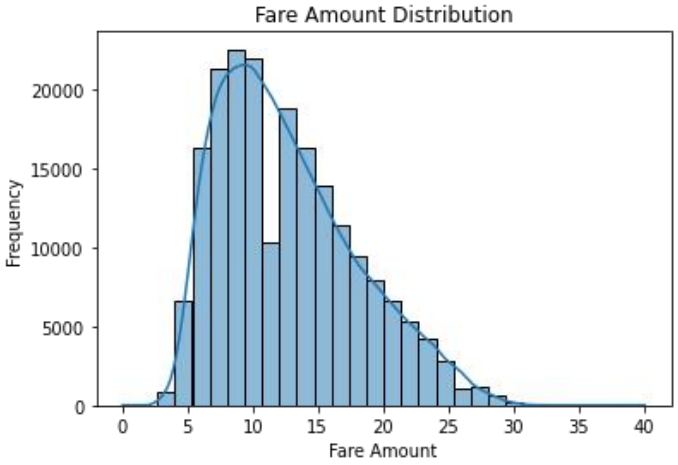
Individual Contribution			
CWID	**Name**	**Contribution(description)**	**Percent Contribution**
A20563467	Zongkun Qin	Visualization	33.3%
A20567177	Qingyu Yao	Report	33.3%
A20564181	Yifan Chen	Collating and Update	33.3%

1. Describe the dataset

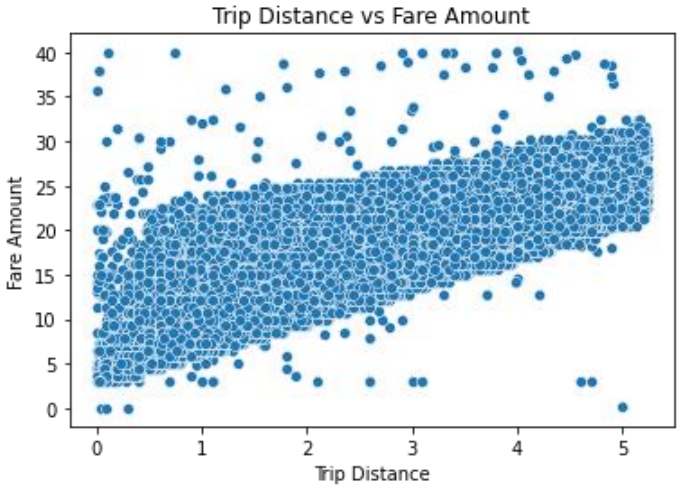
This dataset contains detailed information about taxi trips, including the number of passengers, travel distance, payment type, cost amount, and travel duration.

2. Present your visualization(s) – at least 4 different visualizations

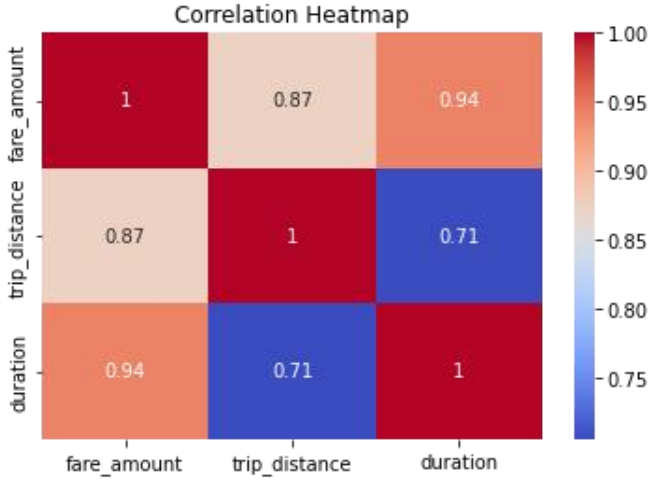
(1)



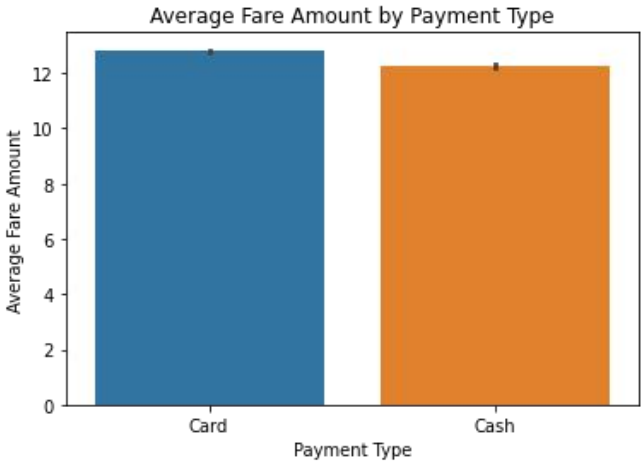
(2)


Trip Distance vs Fare Amount
Fare Amount
Trip Distance
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5

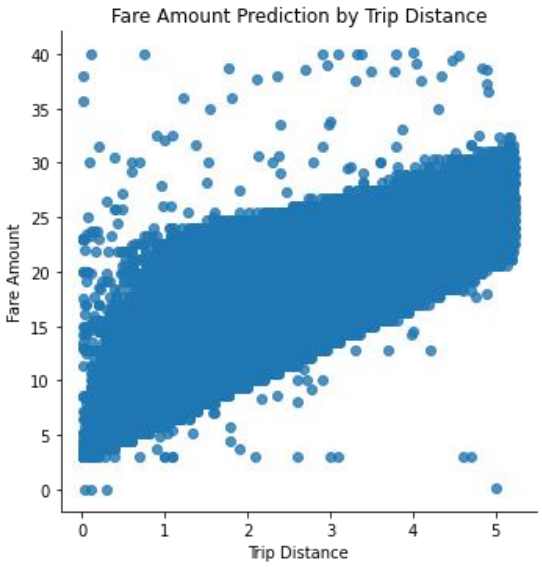
(3)


Correlation Heatmap
fare_amount
trip_distance
duration
1
0.87
0.94
0.87
1
0.71
0.94
0.71
1
1.00
0.95
0.90
0.85
0.80
0.75

(4)


Average Fare Amount by Payment Type
Average Fare Amount
Payment Type
Card
Cash
0
2
4
6
8
10
12

(5)



3. Explain the method you used to create the visualization

(1) sns.histplot is a function in the Seaborn library used to draw histograms.

df ['fare-amount '] extract data from the' fare_amount 'column in the data box df.

bins=30 specifies dividing the data into 30 bins, which is the number of bars in the histogram.

kde=True means overlaying a kernel density estimation (KDE) curve on the histogram to present the distribution trend of the data more smoothly.

(2) sns.scatterplot is a function in the Seaborn library uesd to draw scatterplot.

X=df ['tip_istance '] takes the trip_istance column in the data box df as the X-axis data.

Y=df ['fare_ amount '] takes the far_ amount column in the data box df as the Y-axis data.

(3) df extracts data from the data box df for the columns' fare_amount ',' trip-distance ', and' duration '.

. corr() calculates the correlation coefficient matrix between these three columns of data. The correlation coefficient is a value between -1 and 1, indicating a linear relationship between two variables. 1 represents complete positive correlation, -1 represents complete negative correlation, and 0 represents no linear relationship.Annot=True indicates the specific value of the correlation coefficient displayed in each cell of the heatmap. Cmap='coolwarm' Set the color mapping of the heatmap to coolwarm.

(4) sns.barplot is used to draw bar charts.X='payment_type' takes the payment_type column in the data box df as the X-axis data,Y='fare_amount' takes the fare_amount column in data box df as the Y-axis data.

(5) sns.lmplot is used to plot scatter plots with regression lines.

4. Share what libraries you have used

(1). pandas as pd

(2). os

(3). matplotlib.pyplot as plt

(4). seaborn as sns

5. Describe the results you found and what they tell you about the dataset you chose

The histogram visually displays the frequency distribution of different fare amounts through the height of the bars, while the KDE curve provides a smooth way to estimate the probability density function of the data, helping to understand the distribution pattern of the data. This visualization method provides a more detailed display of the central tendency, dispersion, and distribution shape of the data than a simple bar chart.

A scatter plot can reveal the linear or nonlinear relationship between travel distance and fare amount. As the travel distance increases, the fare amount also increases, which may reflect the pricing mechanism of taxi fares. By observing the distribution of data points in the scatter plot, it can be discovered whether there are outliers or outliers.

A heatmap visually displays the linear relationship between variables, helping us quickly identify which variables have strong correlations. For example, if the correlation coefficient between fare_amount and trip_istance is 1, it indicates a strong positive correlation between the two.

A bar chart can visually compare the differences in average fare amounts for different payment types. The highest bar for card payment types may indicate that passengers using cash for payment typically incur higher fares.

The regression line in the scatter plot can help us understand how travel distance affects the fare amount. The slope of the regression line is positive and significant, indicating that travel distance is an important factor affecting the amount of fare. This may reflect the pricing mechanism of taxi fares, where the longer the distance, the higher the fare.

END