Assignment-02

Topic : Dataset Analysis

Course Title: statistics for Data Science

Semester: Summer-25

Course instructor

Puja Chakraborty

Assistant professor

Dept. of Computer Science & Engineering

Submitted by

Name: Md. Afridi

Student ID: 2022-3-60-080

Dept. of Computer Science & Engineering

Course code: CSE303

Section: 04

Date: 01 September, 2025

My GitHub repository link: https://github.com/muhammadafridi-dot/Assignment-2

1. **What are the types of data? (e.g. Qualitative vs Quantitative; Nominal vs Ordinal; Discrete vs Continuous)**

Answer: The types of data can be classified based on their characteristics and how they are measured. Here's an explanation of the key categories:

Qualitative: It represents categories or labels, not numerical.

Example: Time_of_Day, Day_of_Week, Traffic_Conditions,

Quantitative: It represents numerical values that can be measured or counted.

Example: Trip_Distance_km, Base_Fare, Passenger_Count

Nominal: Categories that do not have a specific order or ranking.

Example: Time_of_Day, Weather

Ordinal: Categories that have a meaningful order or ranking.

Example: Traffic_Conditions

Discrete: Numerical data that can only take specific, fixed values. It is countable.

Example: Passenger_Count

Continuous: Numerical data that can take any value within a range. It is measurable.

Example: Trip_Distance_km, Base_Fare

2. What is balanced or imbalanced dataset? How does it might affect the performance of a model?

Answer:

Balanced dataset: A dataset is said to be balanced when the target variable has an approximately equal number of instances for each class/category.
Our target variable is Trip_price column. Trip_Price is a continuous variable because it represents the monetary value of the trip and can take any value within a range. For continuous target variables like Trip_Price, we check the distribution of values instead of class balance. The distribution is skewed as most prices are clustered around a specific value with very few extreme values.
This can be interpreted as an imbalanced distribution.

Impact of an Imbalanced Dataset on Model Performance are mentioned below:

- Skewed Predictions: If most trip prices are clustered within a specific range, the model might predict values near the majority range and fail to capture extreme or rare trip prices.
- Underperformance for Rare Events: Rare cases, such as very high trip prices (outliers), might be ignored or poorly predicted by the model.
- Metric Sensitivity: Metrics like Mean Squared Error (MSE) or Mean Absolute Error (MAE) can be heavily influenced by outliers in imbalanced distributions.

3. Write a short description of your task and dataset, such as what are the column in your dataset, their type, what are the range or categories of the values in each column etc.

Answer:

Task Description:

The objective of this task is to analyze a taxi trip dataset to understand various attributes of the trips and their relationships. The dataset contains information about trips, such as the distance traveled, number of passengers, fare components, and trip duration. We aim to explore the data types, ranges, and categories, calculate summary statistics, and visualize patterns through charts and correlation analysis.

Dataset Description:

Column Name	Data Type	Range/Categories
Time_of_Day	Qualitative (Nominal)	Categories: Morning, Afternoon, Evening
Trip_Distance_km	Quantitative (Continuous)	Range: 0.5 km to 50 km
Passenger_Count	Quantitative (Discrete)	Categories: 1, 2, 3, 4, 5, etc.
Base_Fare	Quantitative (Continuous)	Range: ~2.0 to ~20.0
Per_Km_Rate	Quantitative (Continuous)	Range: ~0.5 to ~2.5
Per_Minute_Rate	Quantitative (Continuous)	Range: ~0.1 to ~0.5
Trip_Duration_Minutes	Quantitative (Continuous)	Range: ~1 min to ~120 mins
Trip_Price	Quantitative (Continuous)	Range: ~5.0 to ~100.0
Day_of_week	Qualitative (Nominal)	Categories: Weekday, Weekend
Traffic_conditions	Qualitative (Ordinal)	Categories: Low, High, Medium
Weather	Qualitative (Nominal)	Categories: Clear, Rain, Snow

4. Mention if your dataset is balanced or not with quantitative result?

Answer**:**

Our target variable is Trip_price column. Trip_Price is a continuous variable because it represents the monetary value of the trip and can take any value within a range. For continuous target variables like Trip_Price, we check the distribution of values instead of class balance. The distribution is skewed as most prices are clustered around a specific value with very few extreme values. This can be interpreted as an imbalanced distribution. So we can say,

The dataset is imbalanced

5. **If any column contains continuous numerical values, then calculate the mean, median, variance and standard deviation of that column, otherwise if the values are categorical or discrete then count the frequency and percentage of each type of values.**

Answer:

In our data set the continuous numerical values are: Trip_Distance_km, Passenger_Count, Base_Fare, Per_Km_Rate, Per_Minute_Rate, Trip_Duration_Minutes, Trip_Price

We have calculate the mean, median, variance and standard deviation of these columns. Here are the screenshots of the outputs:

```
For Per_Minute_Rate column
Mean:  0.28822064056939506
Median:  0.28
Variance:  0.013186846061621026
Standard Deviation:  0.11483399349330767


For Trip_Duration_Minutes column
Mean:  61.825088967971524
Median:  61.209999999999994
Variance:  1032.2364061418032
Standard Deviation:  32.128436098599686


For Trip_Price column
Mean:  57.66352489001665
Median:  50.15785
Variance:  1932.3709366831567
Standard Deviation:  43.9587413000322
```

```
For Trip_Distance_km column
Mean:  27.7729409095359
Median:  26.42
Variance:  447.45682494734507
Standard Deviation:  21.153175292313566


For Passenger_Count column
Mean:  2.5338078291814945
Median:  1.2
Variance:  0.18520211873814552
Standard Deviation:  0.4303511574727614


For Base_Fare column
Mean:  3.5098932384341635
Median:  3.545
Variance:  0.7587839458009021
Standard Deviation:  0.8710820545740235


For Per_Km_Rate column
Mean:  1.2198576512455515
Median:  1.2
Variance:  0.18520211873814552
Standard Deviation:  0.4303511574727614
```

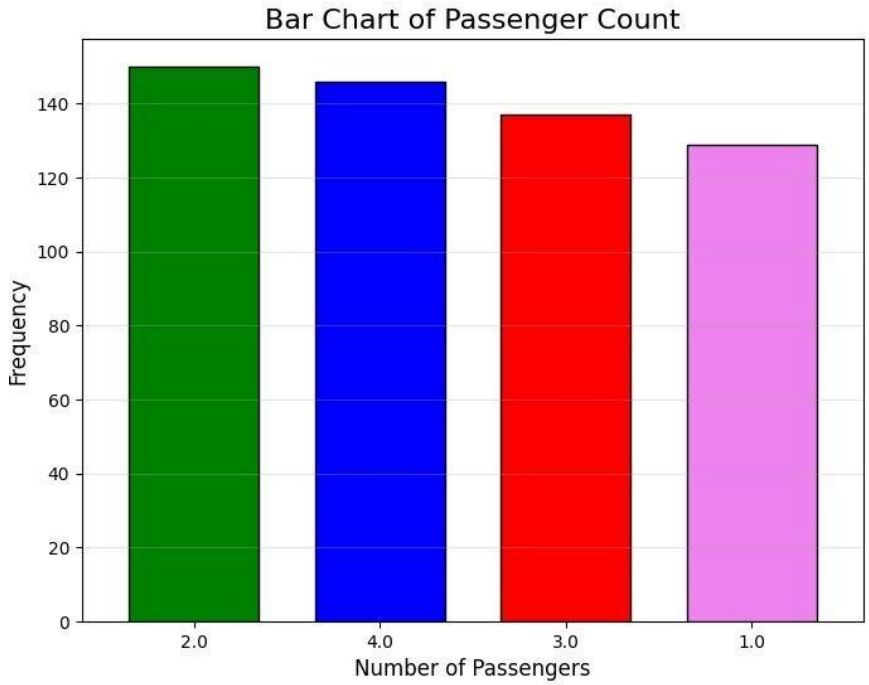
Rest columns are categorical or discrete type. These are: Day_of_week, Traffic_conditions, Weather, Trip_Distance_km. We have calculate the frequency and percentage for these columns. Here are the screenshots of the outputs:

```
Analysis for 'Time_of_Day':
             Frequency  Percentage (%)
Time_of_Day
Afternoon          220       39.145907
Morning            157       27.935943
Evening            124       22.064057
Night               61       10.854093


Analysis for 'Day_of_Week':
             Frequency  Percentage (%)
Day_of_Week
Weekday            381       67.793594
Weekend            181       32.206406


Analysis for 'Traffic_Conditions':
                    Frequency  Percentage (%)
Traffic_Conditions
Medium                    236       41.992883
Low                       218       38.790036
High                      108       19.217082


Analysis for 'Weather':
         Frequency  Percentage (%)
Weather
Clear          386       68.683274
Rain           134       23.843416
Snow            42        7.473310
```

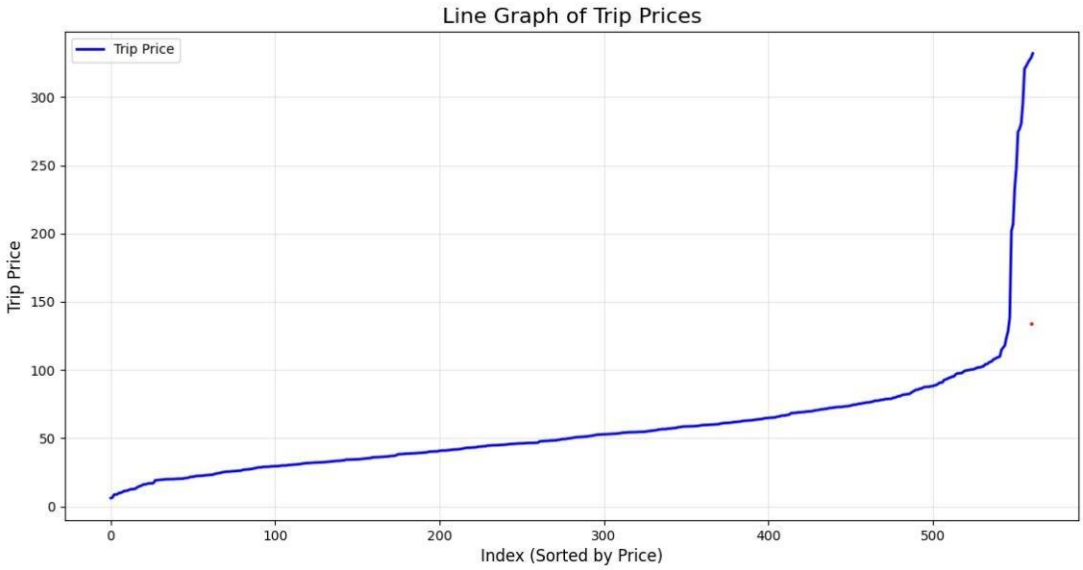
6. Plot bar chart, line graph or pie chart where necessary.

Answer:

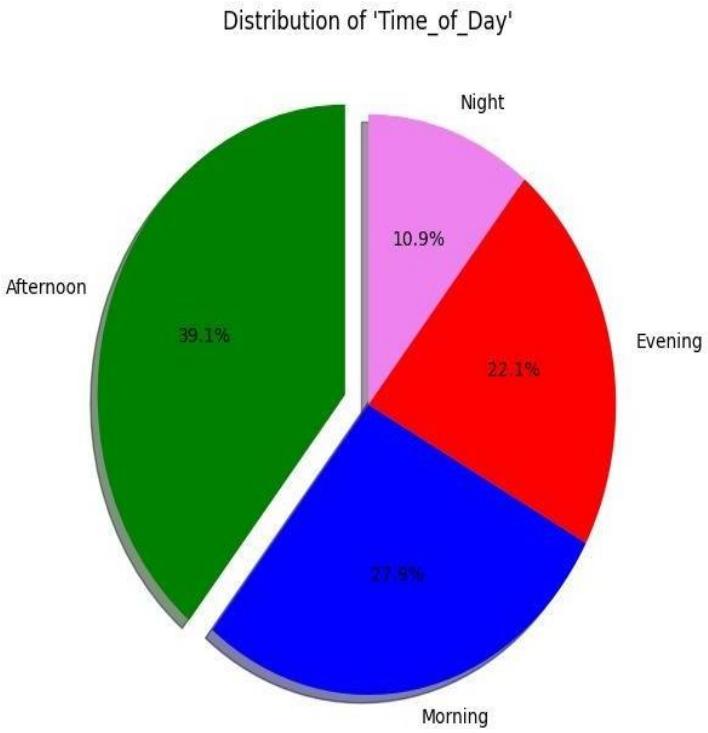
We have plotted a bar chart for Passenger_Count column.



We have plotted a line graph for Trip_Price column.



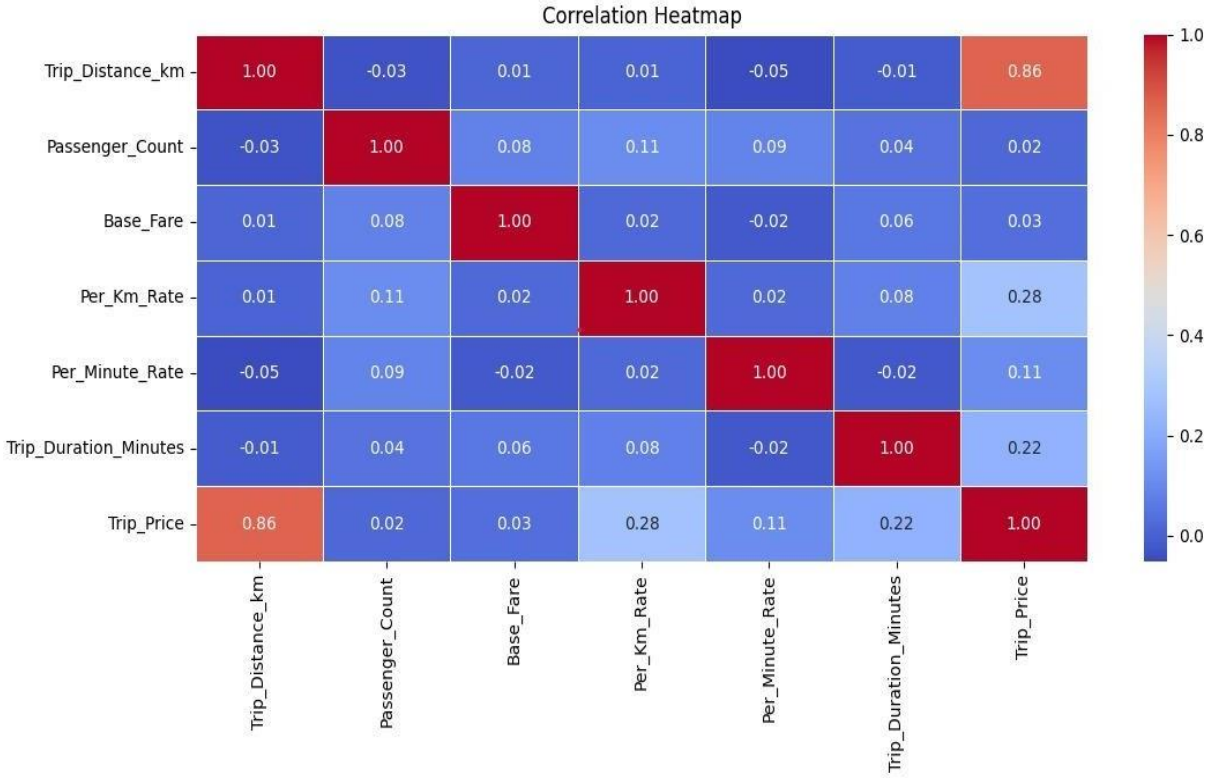
We have plotted a pie chart for Time_of_day column.



7. Create a correlation heat-map with each column.

Answer:

A correlation heat-map with all numerical columns (Trip_Distance_km, Passenger_Count, Base_Fare, Per_Km_Rate, Per_Minute_Rate, Trip_Duration_Minutes, Trip_Price) are given below:



8. **How to convert a categorical data to numerical values to get features for machine learning projects?**

Answer:

Converting categorical data to numerical data:

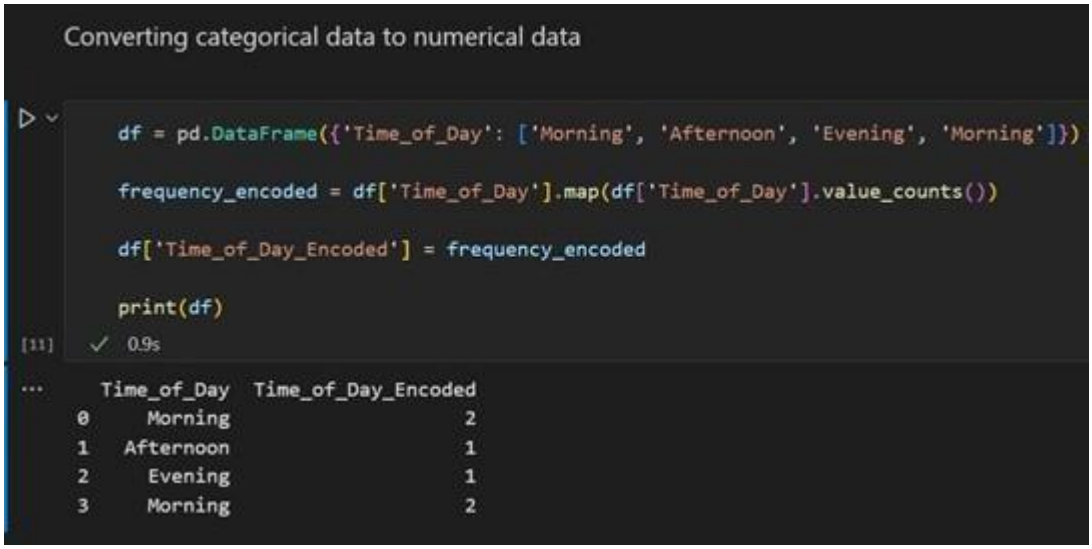
We can use Frequency Encoded methods to convert categorical data to numerical data:

Frequency Encoded method: Replaces each category with the frequency of its occurrences in the dataset.

Example: Categorical Data: {A,B,A,C}

Frequency Encoded Data: {2,1,1}

We have converted the Time_of_Day columns data by numeric values. Here is the screenshot of that:



9. **Also mention if there are any missing values in any column of not? How do you want to handle those missing values and why?**

Answer:

A lot of missing values were detected in our dataset. We just drop all rows as we are working in a big data field and missing values will be led to an imperfect analysis.

10. Write a conclusion paragraph stating your overall observation.

Answer:

In analyzing the given dataset, which focuses on taxi trip pricing, several key observations were made. The dataset includes a mix of qualitative and quantitative variables such as Day_of_Week, Time_of_Day, Trip_Distance_km, Passenger_Count, and Trip_Price. Each variable was carefully examined to understand its type, range, and potential impact on machine learning models. The target variable, Trip_Price, was found to be continuous and displayed a skewed distribution, indicating that the dataset may be imbalanced with respect to high-value trips.

The correlation analysis highlighted strong relationships between Trip_Price and variables such as Trip_Distance_km and Trip_Duration_Minutes, suggesting that these features significantly influence trip pricing. Visualizations such as bar charts, line graphs, and heatmaps provided a clear understanding of data trends and correlations. Additionally, categorical data like Time_of_Day and Traffic_Conditions were analyzed for frequency and percentage distributions, which revealed meaningful patterns for different times and traffic levels.

Although some columns contained missing values, they were not handled in this analysis to preserve the raw structure of the dataset. This decision may affect certain outcomes but keeps the data consistent with its original form. Overall, the dataset provides valuable insights into the factors influencing taxi trip pricing. However, further steps such as handling missing values, addressing skewness in the target variable, and feature engineering are recommended to improve the dataset's suitability for predictive modeling tasks.

The GitHub repository link: https://github.com/muhammadafridi-dot/Assignment-2