Mathematics For Computer Science Engineers
UE23MA242A

Teaching Assistants : Archishman VB, Suchir M Velpanur, Neha Bhaskar

Mobility Services Case Study

Introduction
The two sections below, Background and Case Study provide context for the data science hackathon. This exercise will allow you to test your skills in using the Python programming language to effectively explore the characteristics of a dataset and analyze the features using descriptive statistics such as summary statistics, tables, and graphs. Happy coding!

Background
The given dataset is crucial for understanding mobility services and optimizing transportation services. By analyzing variables such as average fare, rides completed, and driver availability, insights can be gained into pricing strategies and demand patterns. The inclusion of weather conditions and traffic indices allows for a comprehensive assessment of external factors influencing ride frequency and pricing fluctuations. Additionally, features like vehicle type and surge multipliers provide valuable information for resource allocation and driver incentives. Overall, this dataset serves as a foundation for predictive modeling, enabling companies to enhance user experience, improve operational efficiency, and develop targeted marketing strategies.

Case Study
A researcher, having acquired a certain dataset, would like to view it from the lens of statistical analysis to analyze various factors affecting each of average fare, surge multiplier, driver availability and other factors given in the dataset

Dataset Description
The different variables involved in the dataset include :

1) timestamp: This represents the date and time when the data was recorded.
2) average_fare: This is the average cost of rides during the specified time period.
3) rides_completed: This indicates the total number of rides completed within the specified time frame.
4) driver_availability: This metric shows the percentage of drivers are available for rides, indicating how many drivers are on the road and ready to accept ride requests.

5) surge_multiplier: This value indicates a multiplier applied to the fare during high-demand periods (surge pricing).
6) vehicle_type: This describes the category of the vehicle used for the rides (e.g., bike, car, auto).
7) weather: This feature indicates the weather conditions during the specified time period (e.g., Clear, Cloudy, Stormy).
8) traffic_index: This is a numerical representation of the traffic conditions at the time of data collection.
9) special_event: This binary feature indicates whether a special event (e.g., concerts, festivals) is occurring during the specified time.

Problem Set

Unit 1

1) Classify the features in the dataset into their appropriate data types (ordinal, nominal, interval, or ratio). Provide a rationale for each classification.
2) Identify and describe any data quality issues or inconsistencies within the dataset. What steps would you take to clean and preprocess the data to ensure its accuracy and
reliability for further analysis?
3) A summary statistic provides a numerical summary of a specific feature within the dataset. There are two commonly used categories of summary statistics: those that indicate the central tendency and those that indicate the spread of the data. Identify the most appropriate measure of central tendency for each attribute in the dataset and state its corresponding value. Additionally, calculate the standard deviation and range of values for each column.
4) Using a histogram and box plot, assess the presence of outliers in the 'average_fare' and 'traffic_index' variables. Describe the visualizations and identify what distribution do both of these variables belong to.
5) Count and identify the number of outliers in ‘average_fare’ and ‘traffic_index’ columns. Explain the method of identification and steps to resolve outliers
6) Examine the normal probability plot (Q-Q plot) for the 'rides_completed'' variable in the dataset. Based on the shape and trend of the plot, what conclusions can be drawn? Provide a rationale for your conclusions.
7) Calculate the correlations between all numerical variables. From this, identify variable which has highest correlation with 'driver_availability'

8) Generare a pairplot of ‘average_fare’, ‘driver_availability’, and ‘vehicle_type” with ‘weather’ as hue in the dataset. What insights can be gained from the pair plot, and how does it help in visualizing the relationships between these variables?

Unit 2

9) Use hypothesis testing to answer the following :
Define a null and alternative hypothesis to investigate whether there is no significant difference in the average fare paid between bike and car. Use T - test to analyze the relationship between these two variables. Plot a histogram to analyze your hypothesis and its results. Assume significance level as 0.05.
10) Calculate the margin of error to quantify the precision of the analysis done previously and what you can infer from the results.

Unit 3

11) Perform a linear regression to predict 'rides_completed' using 'average_fare', 'driver_availability', 'surge_multiplier', 'traffic_index' and 'special_event'. Plot the predicted versus actual values of 'rides_completed'. Calculate MSE, RMSE and R^2 values, and explain their significance.
12) In analyzing the dataset, we aim to understand factors influencing average fare and ride completion. What additional features could be aggregated from the existing features to enhance the predictive modeling for average fare or rides completed (any 2 feature aggregations) ?