CS215 Assignment 5

Exploratory Data Analysis (EDA)

Prof. Sunita Sarawagi

Due: November 7, 2024

Important Instructions

1. **Sample Format:** A sample notebook (`Q1.ipynb`) will be provided as a reference for attempting all questions. Please use it as a guide to structure your solutions.

2. **Submission Format:** Answer each question in a separate Jupyter Notebook file named after the question (e.g., `Q1.ipynb`, `Q2.ipynb`, etc.). Each analysis must include clear explanations in Markdown cells alongside the code for clarity.

3. **Directory Structure:** After completing the assignment, organize all notebooks in a folder named in the following format:

   ```
   A5-RollNumber1-RollNumber2-RollNumber3
   ```

   Include all notebooks, additional code, and any other files in this folder. Compress (zip) the folder before uploading.

4. **File Naming Convention:** Name each file carefully to correspond to each question, avoiding ambiguity during grading.

5. **Single Submission per Group:** Only one member per group should submit the zip file on Moodle. The group will receive a shared grade.

6. **Deadlines and Late Submission Policy:** Submit the file on Moodle **before 11:59 pm on the due date**. Late submissions may result in penalties or be marked as incomplete.

7. **Data Retention:** Retain a copy of all work submitted until the end of the semester for reference.

EDA Instructions

In this assignment, you are expected to perform exploratory data analysis (EDA) using various plots and summary statistics. Please adhere to the following guidelines for each question to ensure a thorough and insightful analysis:

- **Summary for Each Plot:** For every plot, write a brief summary explaining the key insights. Describe what the data shows visually and any significant patterns or trends observed.
- **Gist of Analysis:** Provide a short, concise analysis for each table and plot. Discuss any notable trends, correlations, or unusual findings, especially those that might inform further analysis or model development.
- **Outlier Identification:** Where applicable, identify and briefly discuss outliers or anomalies. Explain why these may exist and how they might affect the data interpretation.
- **Comparative Insights:** For any comparisons between variables, such as fare versus distance, highlight any relationships or lack thereof. Explain the practical implications of these relationships.
- **Data Quality Remarks:** For questions involving missing data or data type issues, comment on data quality and any recommended preprocessing steps.
- **Judgment of Visualization Quality:** Ensure that all visualizations are well-labeled and easy to interpret, and include units of measurement where applicable. Mention any limitations in the visualization.
- **Flexibility in Tools and Methods:** You are free to use any tools, libraries, or methods you find suitable, as the primary goal of this assignment is to gain a thorough understanding of the dataset rather than to adhere to specific tools or techniques.

Contents

1 Data Icebreaker: *Let's break the ice with some data!* (3 Marks) 4

2 Rush Hour Rush: *A race against time and distance* (15 Marks) 5

3 Fare Frenzy: *Understanding fare, tips, and customer behavior* (15 Marks) 6

4 Data Mayhem: *Conjuring insights from the data* (9 Marks) 7

1 Data Icebreaker: *Let's break the ice with some data!* (3 Marks)

Initial Dataset Exploration(3 Marks)

Description: Familiarize yourself with the **TaxiData.csv** dataset and perform an initial data check and cleaning.

1. Categorize columns: classify columns as categorical, numerical, or mixed data types.
2. Type Conversion: Convert data types where necessary to facilitate analysis, noting why certain columns are converted.
3. Identify Missing Values: Check for missing values in each column and visualize them using a heatmap or bar chart. Discuss which columns have significant missing values.
4. Handling Missing Values: Propose methods to handle missing data based on the column's type (e.g., mean imputation for numerical,) and apply them where ever appropriate. Mention the method used and the column name on which it is applied.

2 Rush Hour Rush: *A race against time and distance* (15 Marks)

Trip Patterns Over Time and Distance

Description: Analyze taxi trip data to understand variations in trip frequency by time, distance, and seasonal patterns. You will need to interpret these visualizations in relation to time-based trends and how they impact taxi usage.

2.a Trip Frequency by Time of Day, Month, and Day (3 Marks):

(a) Create bar charts to visualize trip distributions for:
- Hours of the Day (to capture daily patterns and peak hours)
- Days of the Week (to identify weekday vs. weekend patterns)
- Months of the Year (to see seasonal trends, if any)

(b) Describe notable patterns, like peak hours, high-demand days and peak months. Discuss how these patterns could be relevant for managing taxi demand.

2.b Peak Hours and Distance Analysis (3 Marks):

(a) Identify Peak Hours: Create a heatmap showing trip frequency by the hour to identify the busiest hours for taxi trips.

(b) Distance Analysis: Use a histogram to analyze the distribution of trip distances during peak hours.

(c) Interpretation: Discuss how peak hours combined with distance trends might affect taxi service efficiency (e.g., congestion, longer wait times).

2.c Trip Duration and Distance Distribution (3 Marks):

(a) Plot histograms for trip duration and trip distance to display their distributions. What can you infer from each histogram?

(b) Correlation Analysis: Use a scatter plot to show the relationship between trip duration and distance.

(c) Interpretation: Comment on any noticeable outliers or trends, especially patterns that could influence route planning.

2.d Seasonal Trends in Taxi Usage and Distance (3 Marks):

(a) Use line charts to track the number of trips and average trip distance across months to identify seasonal peaks and lows.

(b) Interpretation: Explain how these trends could influence resource allocation during high-demand periods (e.g., holiday season).

2.e Distance vs. Time of Day (3 Marks):

(a) Plot the variation in trip distances across hours of the day to capture typical distance trends for peak and non-peak hours.

(b) Interpretation: Explain your findings, including any insights into typical distance trends during peak and non-peak hours.

3 Fare Frenzy: *Understanding fare, tips, and customer behavior* (15 Marks)

Fare and Tips Overview

Description: Examine the relationship between fare, tips, and payment behavior, analyzing any patterns or anomalies. This section should help understand customer behavior and fare distribution.

3.a Fare and Tip Distribution (4 Marks):

(a) Create histograms and box plots for both fare and tip amounts.

(b) Identify the range, outliers, and any clusters in the data.

(c) Interpretation: Provide a statistical summary(mean, median, standard deviation,), and analyze any patterns in fare and tipping behavior, such as outliers or clusters in high or low values.

3.b Payment Method Insights (3 Marks):

(a) Visualize the frequency of each payment method using bar charts or pie charts.

(b) Interpretation: Identify any dominant payment methods and discuss how this may impact service providers

3.c Tips by Payment Method (3 Marks):

(a) Create a box plot to compare tips across different payment methods.

(b) Interpretation: Summarize the average tipping amounts for each method and analyze variations.

3.d Fare vs. Distance (5 Marks):

(a) Plot a scatter plot showing the relationship between trip distance and fare.

(b) Calculate and interpret the correlation coefficient between the two variables.

(c) Interpretation: Explain any findings on the fare-distance relationship and outline any potential implications for fare pricing.

4 Data Mayhem: *Conjuring insights from the data* (9 Marks)

Advanced Data Analysis and Prediction

Description: This section will involve investigating unusual data patterns, building predictive models, and reducing dimensionality to gain deeper insights.

4.a Outliers in Fare and Tips (4 Marks):

(a) Use box plots to identify outliers in fare and tip values.

(b) Interpretation: Discuss how these outliers might skew average calculations.

(c) Suggest methods for handling these outliers in predictive models (e.g., removal or adjustment).

4.b Predicting Fare (5 Marks):

(a) Implement a regression model using distance, trip duration, and location as predictors to estimate fares.

(b) Model Evaluation: Present metrics such as RMSE to assess model accuracy.

(c) Discuss the effectiveness of the model, noting any potential limitations or areas for improvement.