Data Science project: Parking Lot Analysis

Project overview

You are provided with a dataset that simulates the operations of various parking lots across different states. This dataset is divided into three CSV files, each representing different aspects of the parking lot operations: parking_lots.csv, events.csv, and transactions.csv. Your task is to perform a comprehensive analysis using Python in a Jupyter Notebook. This is an open-book assignment, and you are allowed to use any resources you find helpful.

Dataset overview

The dataset you'll be working with simulates operations across various parking lots, designed to mimic real-world scenarios that a parking lot management system might encounter. In these scenarios a camera is placed at the entrance of the parking lot, recording the incoming and outgoing traffic. Parking lot users must pay for the usage of the parking lot by sending their plate numbers and required amount of money for their parking. This transaction is linked to the parking session based on the time of the transaction and the plate number. The parking lot users that did not buy their ticket for their sessions (no matching transaction) will be fined.

This simulated dataset is divided into three CSV files, each focusing on different aspects of the parking lot operations. Understanding the structure and content of these files is crucial for your analysis.

parking_lots.csv:

This file contains information about each parking lot, including its unique identifier, geographic location (latitude and longitude), the number of maximal parking spaces, and the state and timezone which it is located in.

Timezones:

GMT - 5	GMT - 6	GMT - 7

parking_sessions.csv:

This file tracks individual parking sessions, detailing when cars enter or exit the lots' entrance. Each session is tied to a specific parking lot and includes information on the car (make, colour, and plate number) and the sessions' timing (entry and exit date) *(in the given lot's time zone)*.

transactions.csv:

This file logs transactions related to parking sessions. Each transaction includes the transaction amount and transaction date and the cars' plate number, which was given by the parking lot user. The dataset simulates various transaction scenarios, including zero or negative amounts to represent failed transactions and some transactions not linked to any parking session, mimicking erroneous charges. The transaction time should be always between the parking session start and end time.

Additional Complexities and Considerations

- **Time Zone Handling**: With parking lots across different time zones, you'll need to standardize time data to a single time zone for consistent analysis.
- **Data Anomalies:** Look out for subtle patterns and anomalies, such as peak parking times or frequent misreads of plate numbers.

Tasks

1. **Read CSV Files:** Start by loading the three CSV files into separate Pandas DataFrames.

2. **Exploratory Data Analysis (EDA)**: Conduct an exploratory analysis to understand the datasets' characteristics, identify patterns, anomalies, and relationships between the columns. Write your findings in a bullet point structure under in a text cell. Visualize some characteristics you found important.
   - **Question:** What insights can you gather about the parking lot operations from the distributions and summaries of different variables? How do the time zones affect the data?

3. **Feature engineering:** Create a new dataframe by merging the existing ones, where you match the sessions to the transactions. If there is no transaction with the matching plate number, the parking lot user might not pay for the parking, or there might be an error. If the payment is valid that should be shown on a field as well.
   - **Question**: Are there any discrepancies during merging?
   - **Question**: Think of features that could be created that could be helpful for modelling or provide deeper insights into the parking lot operations.

4. **Understand parking personas:**
   - **Question:** What type of parking lot users can you differentiate based on the length of their parking?

5. **Explain the Results**: Document your findings and explain the significance of your analysis, visualizations results.

6. **Bonus task:** Not all events can be paired to another based on the cars' plate numbers, because there are character mismatches. These are sometimes because the parking lot user mistyped a character in the license plate number. Match the sessions and transactions together, where there is only one character difference between the two plate numbers, namely in the last character! (The last character's Levenshtein distance is not more than 1)! (Match them by filling the transactions data frames the missing session ids.)

Guidelines

- Complete your analysis in a Jupyter Notebook. Ensure that your code is well-commented to explain your logic and steps.

- Alongside your code, include explanations of your findings and the reasoning behind your analysis.

- The notebook should be structured logically, with clear headings for each section of the analysis.

Evaluation Criteria

- **Analytical Rigor:** Show depth in your EDA and feature engineering process.
- **Insightfulness:** Demonstrate your ability to derive meaningful insights from the data and clearly communicate these findings.