Problem Statement

Porter, India's largest marketplace for intra-city logistics, works with a wide range of restaurants to deliver their items directly to customers. The company wants to estimate the delivery time for each order based on various features, such as the items ordered, the restaurant, and the availability of delivery partners. An accurate estimation of delivery time will enhance customer satisfaction and optimize the delivery process.

Dataset Description and Dictionary

The dataset includes the following information for each order:

1. **market_id**: Integer ID for the market where the restaurant is located
2. **created_at**: Timestamp at which the order was placed
3. **actual_delivery_time**: Timestamp when the order was delivered
4. **store_primary_category**: Category of the restaurant
5. **order_protocol**: Integer code value for the order protocol (e.g., through Porter,call to restaurant, pre-booked, third-party, etc.)
6. **total_items**: Total number of items in the order
7. **subtotal**: Final price of the order
8. **num_distinct_items**: Number of distinct items in the order
9. **min_item_price**: Price of the cheapest item in the order
10. **max_item_price**: Price of the most expensive item in the order
11. **total_onshift_partners**: Number of delivery partners on duty when the order was placed
12. **total_busy_partners**: Number of delivery partners attending to other tasks
13. **total_outstanding_orders**: Total number of orders to be fulfilled at that moment

You can download the dataset from
https://drive.google.com/file/d/1SzTmeY9FZEnfLchM819C8JD_tI-LwIYx/view?usp=sharing

Concepts Tested

- Exploratory Data Analysis (EDA)
- Data preprocessing and feature engineering
- Handling missing values and encoding categorical data
- Data visualization and cleaning
- Outlier detection and removal

1. Defining the Problem Statement, Importing Data, and Data Structure Analysis (10 points)

- Clearly define the problem.
- Import the dataset and understand its structure.
    - Dataset shape
    - Data types
    - Missing values
    - Statistical summary

2. Data Preprocessing and Feature Engineering (30 points)

- Data cleaning
- Handling missing values
- Creating the target column (time taken for delivery) from order timestamp and delivery timestamp
- Extracting hour and day of the week from timestamps
- Encoding categorical columns

3. Data Visualization and Cleaning (20 points)

- Visualize various columns for better understanding (e.g., count plots, scatter plots)
- Check if the data contains outliers
- Remove outliers using appropriate methods
- Plot the data again to see improvements

4. Insights and Recommendations (40 points)

- Provide actionable insights and recommendations based on the analysis.

Basic Questions

1. **Data Structure and Overview**
    - What is the shape of the dataset (number of rows and columns)?
    - What are the data types of each column?
    - Are there any missing values in the dataset? If so, how many and in which columns?
2. **Descriptive Statistics**
    - What are the basic statistical summaries (mean, median, standard deviation) for the numerical features?
    - What is the distribution of the categorical variables like `store_primary_category` and `order_protocol`?
3. **Datetime Features**
    - How many orders were placed each day/week/month?
    - What is the distribution of order times throughout the day?

Intermediate Questions

4. **Feature Engineering**
    - How can we create a new feature for the time taken for each delivery?
    - How can we extract additional features from the datetime columns, such as the hour of the day or the day of the week?
5. **Exploratory Data Analysis (EDA)**
    - What are the distribution plots for continuous variables like `total_items`, `subtotal`, `min_item_price`, and `max_item_price`?
    - What are the count plots for categorical variables like `store_primary_category` and `order_protocol`?
6. **Missing Values Handling**
    - How can we handle missing values in the dataset, especially for important columns like `store_primary_category`?

7. **Correlation Analysis**
    - What are the Pearson and Spearman correlation coefficients between numerical features (e.g., `total_items`, `subtotal`, `min_item_price`, `max_item_price`)? What do these correlations suggest?
8. **Multivariate Analysis**
    - How do multiple factors (e.g., `market_id`, `store_primary_category`, `order_protocol`) together influence the `subtotal` or delivery time?
9. **Outlier Analysis**
    - Are there any outliers in the dataset? Which method can be used to identify and handle these outliers?
    - How do the data distributions change after removing outliers?
10. **Categorical Feature Encoding**
    - How can we encode categorical variables like `store_primary_category` and `order_protocol` for further analysis?
11. **Advanced Feature Engineering**
    - Can we create a feature based on the availability of delivery partners, such as a ratio of `total_busy_partners` to `total_onshift_partners`?
    - How do engineered features like order time of day or week enhance the predictive power or insights of the analysis?
12. **Advanced Visualization**
    - Use advanced visualization techniques (e.g., heatmaps, pair plots) to explore relationships between multiple variables simultaneously.
    - How do interactions between categorical variables (e.g., `store_primary_category` * `order_protocol`) affect the delivery time?
13. **Statistical Tests**
    - Perform statistical tests to determine if there are significant differences in delivery times between different groups (e.g., different restaurant categories or order protocols).

Deliverables:

- **EDA Report**: A comprehensive report of the exploratory data analysis.
- **Data Preprocessing Report**: Documentation of the preprocessing steps, including handling missing values, feature engineering, encoding, and outlier removal.
- **Insights and Recommendations**: Provide actionable insights and recommendations based on the analysis.

Optional:

- Convert the Jupyter notebook into a PDF and upload it.
- Optionally, add images/graphs in the text editor.