The purpose of this project is to demonstrate various techniques for exploring and analyzing data, including descriptive and correlation analysis, creating regression models, identifying key predictor variables, and manipulating existing variables through transformations or additional analysis.

Introduction:

Cab Taxi is a thriving transportation business in many parts of the world. It facilitates easy movement of people, goods, and services from one location to another. The aim of this data analysis project is to analyze the Yellow Medallion Taxi cabs dataset: the famous New York City (NYC) yellow taxis that provide transportation exclusively through street hails (i.e., the pickups are not prearranged). Passengers stand by the street and hail on an available taxi with their hand.

Data:

The dataset was sourced from the NYC Taxi & Limousine Commission (TLC) official website. The dataset contains several explanatory variables used to assess a completed trip such as pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts.

A subsample of the original data is provided to use for the tasks itemized in the following sections. The trip records are divided into two files Main Sample and New Sample, respectively.

In this analysis, I will answer some business questions and perform regression analysis to predict the total amount paid by the passengers after a given trip.

Description of Tasks:

Task A:

Knowing some statistics about the daily cab activities can help to improve the transportation business. Thus, analyzing the dataset provided in the Main Sample file I answered the following business questions:

i. What is the average demand for the taxis in the days of the week (i.e., daily trend). Which of the days has the highest and which lowest demand?
ii. Which time of the day (morning, afternoon, evening, and night) is likely be a peak period for the taxi's operation from the data?
iii. On average, how much revenue was generated in the weekdays and weekends for the business for the period covered in the dataset?

Task B:

Creating a regression model to predict the total amount paid for taxi ride, given the trip information in the dataset:

- Sequentially split the data in the Main Sample file into two sets, such that the first 80% of the records in the file is used for fitting the regression model. While the last 20% is used for testing the model and reporting the prediction errors (e.g., RMSE) and R2 scores respectively.

- Provide the equation for the finalized model in the report.

- Once the model is finalized, predict the total amount paid on a trip for the trip records shown in New Sample file and tabulate the predicated values in the report, in the order the records are arranged in the file.

The report will have the following Chapters:

1. **Introduction:** This includes the goals of the analysis and a brief description of the data. It will also include EDA and transformation I had to perform on the data.
2. **Data Analysis:** This section covers the discussion and answers to the question from Task A.
3. **Regression Analysis:** This section includes discussion about the activities in Task B including the modelling process.
4. **Discussion:** This includes relevant predictions and/or conclusions drawn from the model.
5. **Conclusion:** In this section, I summarized my analysis and highlighted final points from the analysis.
6. **Reference:** In this section, I provided the references to the source of data.