Process of my project

- I started by importing the necessary libraries: pandas, matplotlib.pyplot, and seaborn. These are essential for data manipulation and visualization tasks.
- Then, I loaded our train data from a CSV file called 'train.csv' into a DataFrame named df_train. This dataset contains information about taxi trips.
- After loading the data, I wanted to get a quick overview, so I checked the data types of each column using the dtypes attribute. It helps ensure that the data is correctly interpreted.
- Next, I looked for any missing values in the dataset. I used the isnull().sum() method to sum up the null values across all columns. This step is crucial for data cleaning and preprocessing.
- To understand the size of our dataset, I used the shape attribute, which gives the total number of rows and columns.
- After that, I explored the summary statistics of the train dataset using the describe() method. It provides valuable insights into the distribution and central tendencies of numerical variables.
- Moving on to visualization, I wanted to understand the proportion of trips by passenger count. So, I created a pie chart using the plt.pie() function. I also defined a function called autopct_format() to format the percentage labels nicely. This chart gives us a quick snapshot of how passengers are distributed across different trip counts
- Then, I wanted to see how the average trip duration varies by day of the week. To do this, I converted the 'pickup_datetime' column to datetime format and extracted the day of the week. Then, I calculated the average trip duration for each day of the week using the groupby() and mean() functions. Finally, I plotted a line chart to visualize these averages over the weekdays.
- Lastly, I created a scatter plot to visualize the geographical relationship between dropoff longitude and latitude. This scatter plot helps us understand the geographical distribution of dropoff points and identify any patterns or clusters.
- Each step in the code serves a specific purpose, from data loading and exploration to visualization and interpretation, providing valuable insights into various aspects of the dataset.