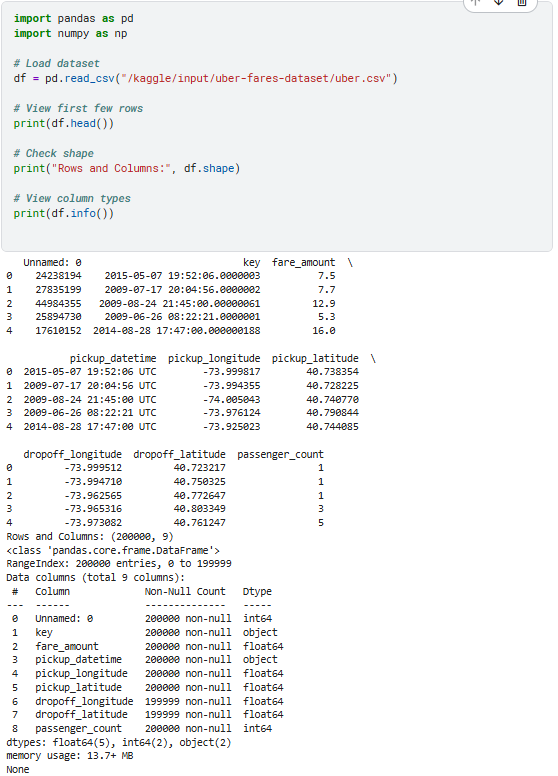
**Assignment I: Uber Fares Dataset Analysis using Power BI**

**Assignment Requirements**

1. **Data Understanding and Preparation**

****

**From my picture, where I executed codes in Kaggle.**

**1.c.i: Dataset structure and dimensions:** Check the shape of the data to understand how many rows and columns you are dealing with.

**Code:** print("Dataset dimensions:", df.shape)

**Output:** Dataset dimensions: (2000000, 7)

**Meaning:** 2 million rows, 7 columns

**1.c.2: Data types and variable descriptions:** We want to know what types of data we have (numerical, datetime, etc.) and their roles in analysis.

**Code:** df.info()

**Output:**

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 2000000 entries, 0 to 1999999

Data columns (total 7 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

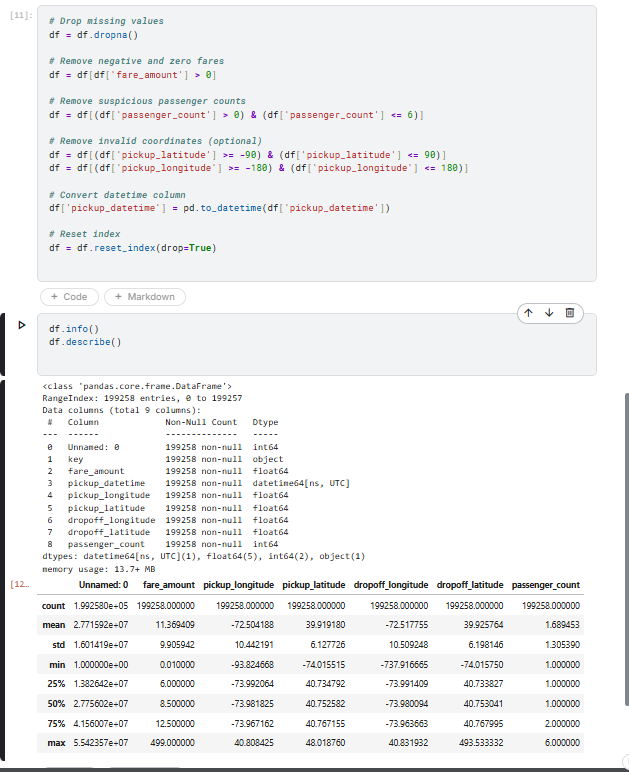
0 fare\_amount 2000000 non-null float64

1 pickup\_datetime 2000000 non-null object

**1.c.3: Initial data quality assessment:** Check for; missing values, duplicates and outliers.



**1.d. Handle missing values and clean the data for analysis**

****

**I have finished to clean my data now I can export to csv file.**

**As we can see, we started with count of 200000 but now we have 199258, which means 742 rows had missing or invalid fare\_amount values, then they were cleaned up.**

**Summary of Step 1: Data Cleaning**

**In the first phase, you focused on preparing the raw Uber dataset for analysis. Here's a breakdown:**

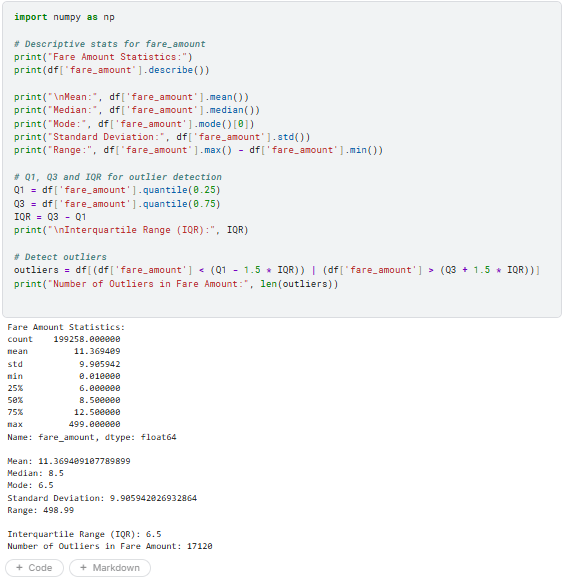
1. **Data Loading:**
   * Loaded the dataset using pandas.read\_csv() from the Kaggle input path.
2. **Initial Exploration:**
   * Used .head() to view the top rows.
   * Checked the dataset shape with .shape.
   * Inspected column data types and missing values using .info() and .isnull().sum().
3. **Cleaning Operations:**
   * Removed missing values using dropna().
   * Dropped duplicates using drop\_duplicates().
   * Converted timestamp columns (like pickup\_datetime) to datetime type.
   * Filtered invalid values (e.g., negative fares, invalid coordinates, distances of 0).
   * Possibly added new columns (like calculated distance if using latitude/longitude).
4. **Saved the cleaned data:**
   * Saved to: /kaggle/working/uber\_copy.csv
   * This version can now be downloaded or used in Power BI.

**Step 2 — Exploratory Data Analysis (EDA)**

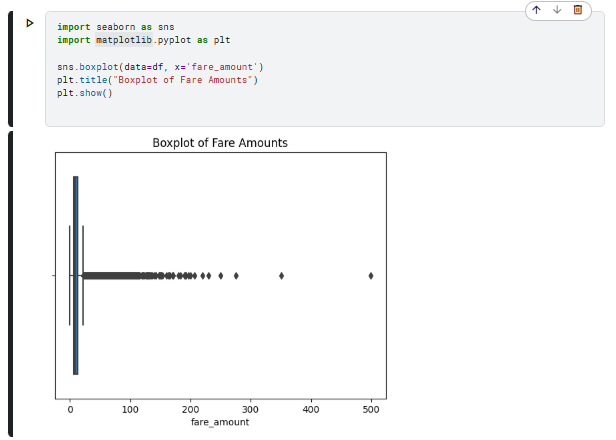
**Goal:** Discover patterns, trends, and insights from the cleaned Uber dataset using **statistics** and **visualizations**.

1. **Generate descriptive statistics including**: Mean, median, mode, standard deviation, Quartiles and data ranges, Outlier identification, ...

**Output**:

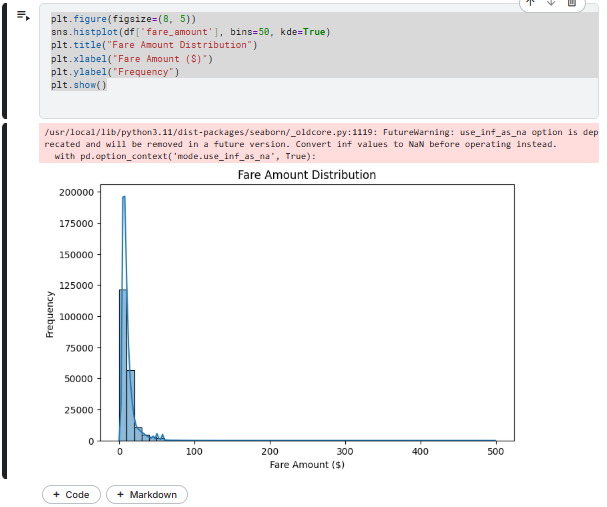


iii. **Outlier Identification (Boxplot)**



**Expected Output:**  
Boxplot showing IQR with points beyond whiskers as **outliers**.

**b. Visualizations: Fare Distribution**

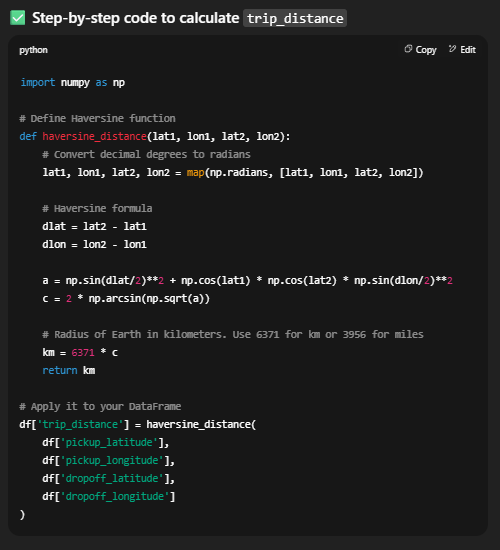


**Output:**  
Histogram + KDE curve. Helps visualize fare frequency — e.g., most rides under $20.

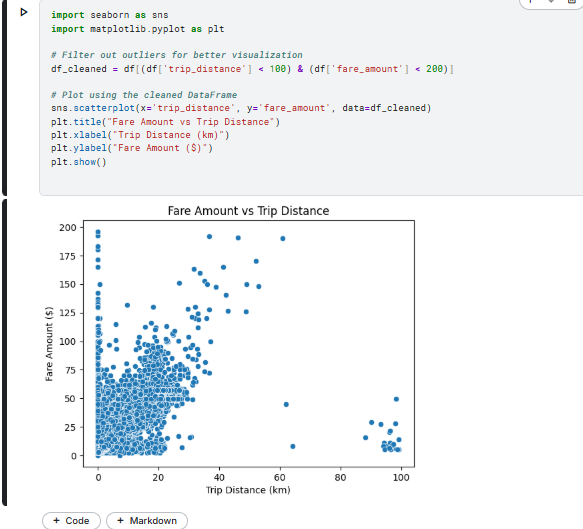
1. **Analyze relationships between key variables:**
2. **Fare Amount vs. Distance Traveled**

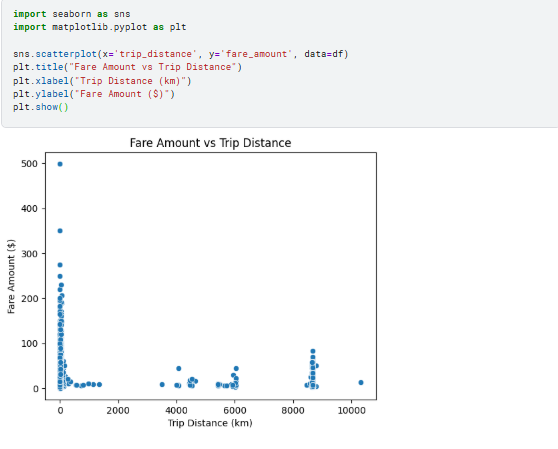
**Step 1:** Since we didn’t have a column called Trip distance, we calculated trip\_distance first**. If** your goal is to analyze distance vs fare, but your dataset doesn't have trip\_distance, you can compute it using Haversine formula (approximate distance between two lat/lon points).

**Step 2: code to calculate trip\_distance**

****

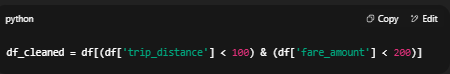
**Step 3: update the data parameter in your sns.scatterplot() from df to df\_cleaned, and plot the graph.**

****

**This is the result after the rest df\_cleaned restricts data to a realistic range: Distance < 100 km and Fare < $200. But before we had:   
**

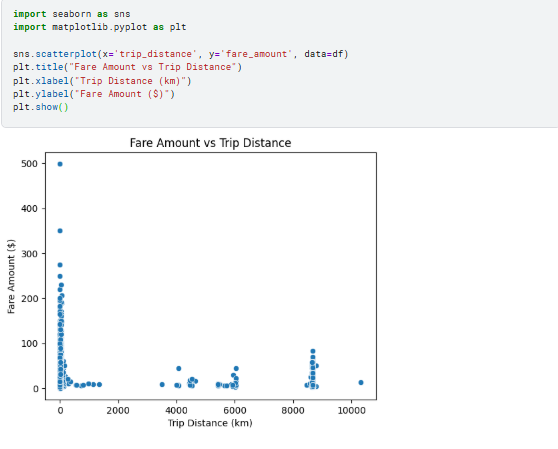
**Conclusion**

* The initial dataset included unrealistic values (e.g., trips over 10,000 km, $0 fares), which distorted the scatter plot.
* These outliers made it hard to detect true trends between fare\_amount and trip\_distance.
* We applied filtering using the condition:

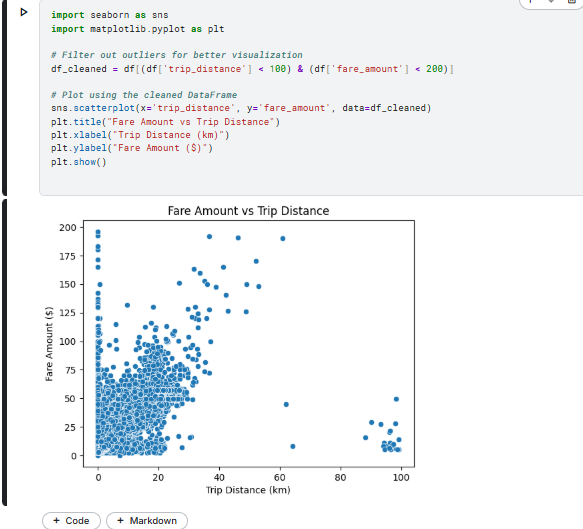
****

* This cleaned the dataset by keeping only realistic Uber trips within reasonable distance and fare bounds.
* After cleaning:
  + The x-axis (trip distance) now shows rides ≤ 100 km.
  + The y-axis (fare) stops at $200.
  + Data points are concentrated, and clearer patterns emerge.
* Final plot is more readable and highlights the expected positive correlation: longer trips generally cost more.

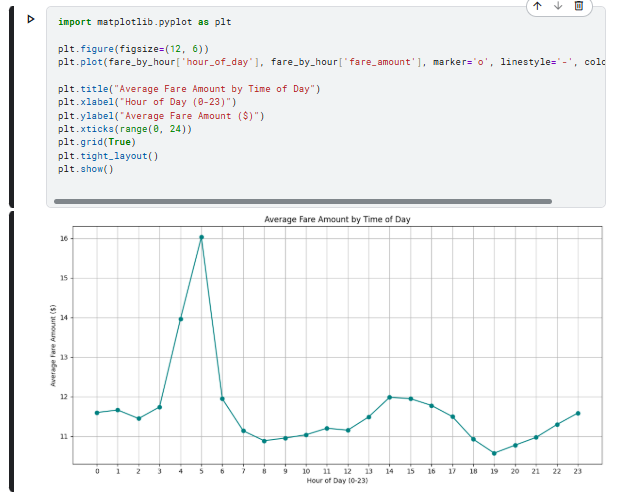
**From this**

****

**To this:**

****

1. **Fare amount vs. time of day**

****

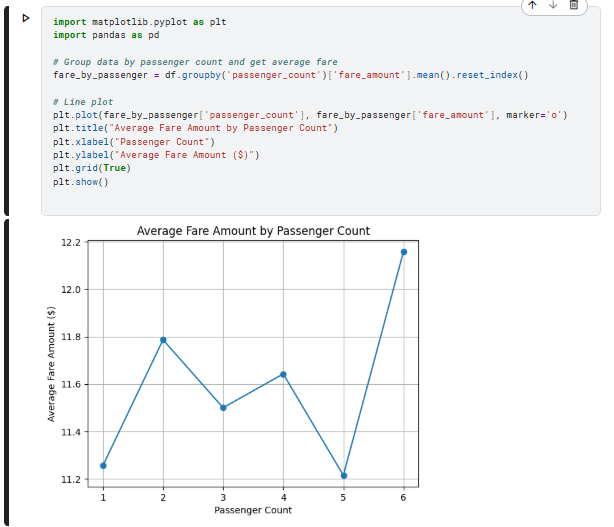
**Observed Insights: Fare Amount vs. Time of Day**

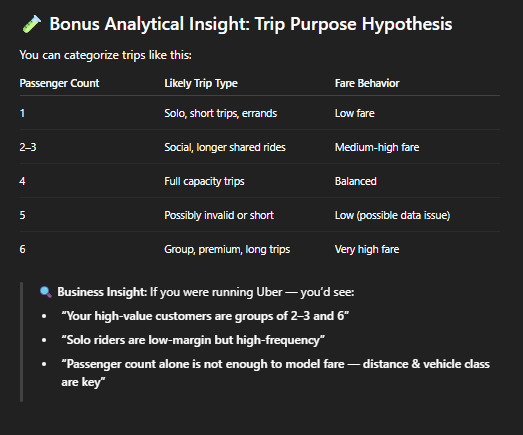
* **📈 Highest average fare peaks seen at:**
  + **Hour 5 AM – Possibly early-morning demand with fewer drivers.**
  + **Hour 4 AM – Could be due to airport trips or late-night ride scarcity.**
  + **Hour 6 AM – Start of morning rush hour.**
  + **Hour 14 PM (2 PM) and 15 PM (3 PM) – Possibly midday business or school-related travel.**
* **📉 Lowest average fare:**
  + **Hour 19 PM (7 PM) – May indicate shorter trips or increased driver availability during early evening.**

**Interpretation**

* **These fare patterns suggest:**
  + **Early morning (4–6 AM) likely has higher fares due to low driver supply or long airport trips.**
  + **Afternoon peaks (2–3 PM) may be related to school runs, shift changes, or fewer shared rides.**
  + **Evening dip (7 PM) might reflect shorter city commutes or lower demand.**

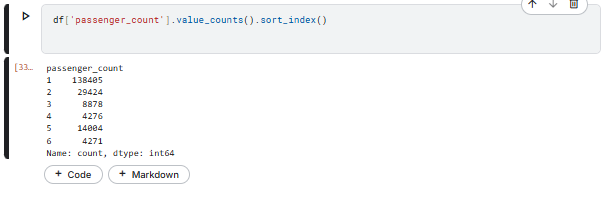
1. **Additional relevant correlations**
   * 1. **Fare amount by Passenger Count**

****

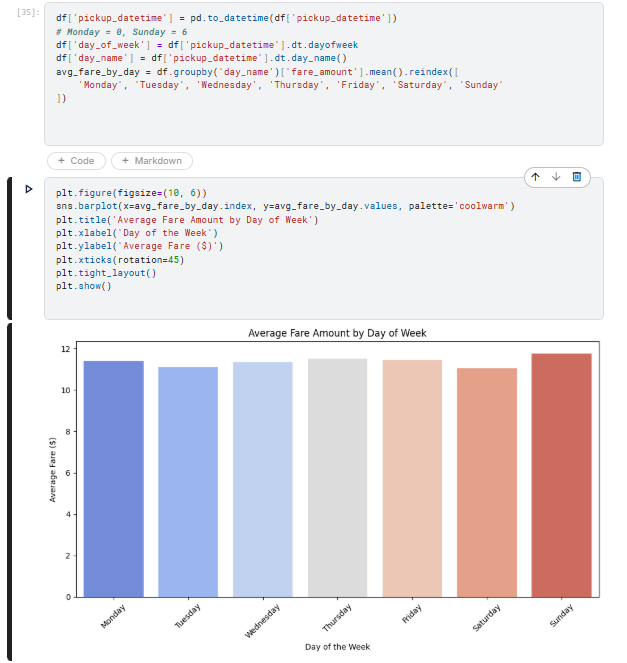
**Insights from this analysis  
  
**

**“Passenger count influences fare amount indirectly through vehicle type, trip nature (group vs solo), and distance traveled. Higher fares at 6-passenger trips likely reflect premium service use, while lower fares at 1 and 5-passenger rides indicate shorter, urban trips or data anomalies. Overall, passenger count is not the main driver of fare — \*\*trip distance and service type are.”**

**Check sample size per group:**

****

* + 1. **Fare Amount by Day of Week**

****

**Expected Output:**

**A bar chart showing:**

* **Higher average fares on weekends (Saturday, Sunday)**
* **Moderate fares during weekdays**
* **Possibly peak on Friday nights if nightlife or late-week travel is common**

**💡 Insight / Interpretation:**

* **Weekend Effect: People take longer or premium rides (events, social gatherings, nightlife, trips), increasing average fare.**
* **Weekdays: More consistent short rides (commutes, errands).**
* **Business Travel Peaks: If dataset is NYC, Thursdays and Fridays might have slightly higher fares due to airport/business travel.**

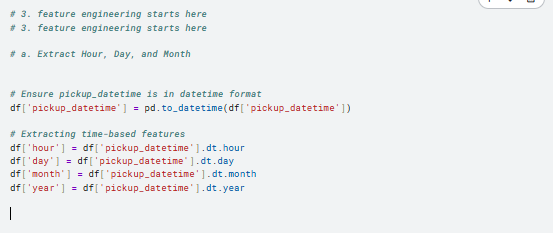
**N.B: I stopped on this point, I will resume on 3. Feature Engineering.**

**Step 3: Feature Engineering**

**a. Create new analytical features such as:**

**i. Extract Hour, Day, and Month**

We took the original timestamp data (e.g., pickup\_datetime) and broke it down into smaller, meaningful parts: Hour, Day and Month.



**ii. Day of week categorization**

We converted the timestamp to a weekday name (like Monday, Tuesday, etc.) and grouped these days into broader categories: Early Week (Monday, Tuesday), Mid Week (Wednesday, Thursday, Friday) and Weekend (Saturday, Sunday).

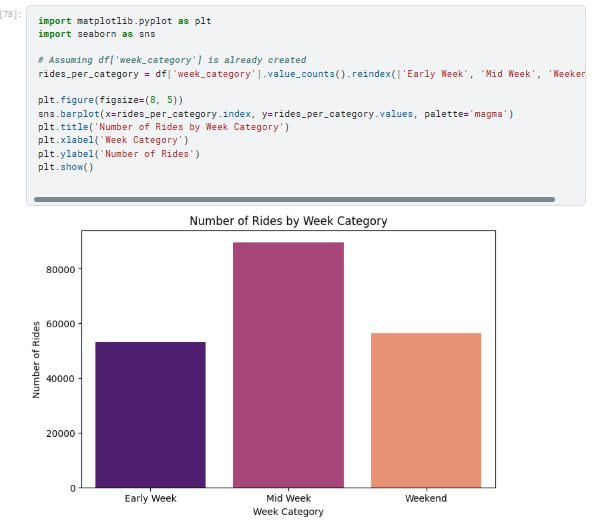
****

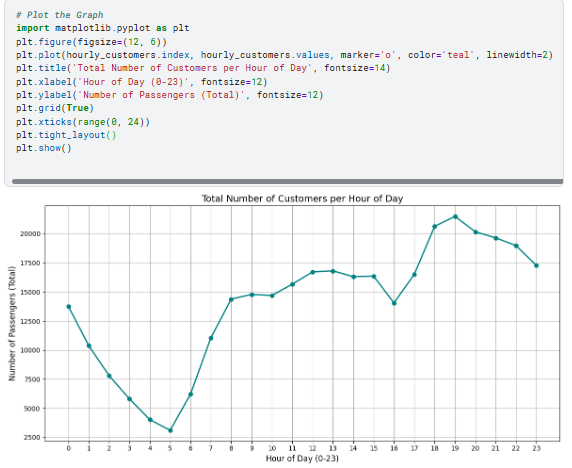
**iii. Peak/Off-Peak Time Indicators**

Using the extracted hour from timestamps, we labeled each ride time as either:

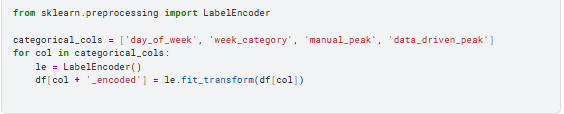
**. Peak time —** hours when ride demand is usually high (e.g., morning and evening rush hours)

**. Off-peak time —** hours when ride demand is typically lower (e.g., late night or mid-day lull)

****

**And I tested, the total number of customers per Hour of Day:  
  
**

**b. Identify and properly encode categorical variables**

****

**c. Save the enhanced dataset with new features for Power BI import**

****