Given the project outline and the datasets provided, the first step will be to conduct a thorough exploratory data analysis (EDA) that will form the basis for deeper investigation and modelling.
Here's how I am going to approach the analysis and what the deliverables will include:

1. Understand and Prepare the Data

4 datasets received.

I'll begin by loading and inspecting each dataset to familiarize myself with the content and structure. I'll look for:

- □ Field names and data types
- □ Relationships across the files
- □ Presence of missing values or obvious data inconsistencies

Here's the plan for loading and initial inspection:

1. Load each dataset into Excel.
2. Inspect the first few rows of each dataset to understand the structure

Inspect the Datasets:

1. ***Cab_Data.csv.***

Data Contents:

'Transaction ID': Unique identifier for each transaction.
'Date of Travel': The date of travel (appears to be in an integer format, likely Julian date).
'Company': The cab company (e.g., "Pink Cab").
'City': The city in which the cab service was used.
'KM Travelled:' The distance travelled during the cab ride.
'Price Charged': The price charged for the trip.
'Cost of Trip': The cost of the trip to the company.

Data Types & Completeness:

- □ The dataset contains 359,392 entries.
- □ All columns have data for each entry (no null values).
- □ 'Date of Travel' was converted to a proper date format for better analysis.

2. *City.csv*

Data Contents:

- City: The city name.
- Population: The total population of the city.
- Users: The number of users who use cab services in that city.

Data Types & Completeness:

- The dataset contains only 20 entries, likely representing major cities covered by the cab services.
- All columns are classified as object type, suggesting that the numbers (Population and Users) might include commas as thousands separators, preventing them from being read as integers or floats.

Next Steps for City.csv:

Converted the Population and Users columns to a numeric format by removing commas(TRIM()) and converting the strings to integers.

3. *Customer_ID.csv*

Data Contents:

- Customer ID: Unique identifier for each customer.
- Gender: The gender of the customer.
- Age: The age of the customer.
- Income (USD/Month): The monthly income of the customer in USD.

Data Types & Completeness:

- The dataset includes 49,171 entries.
- There are no null values in any of the columns.
- Data types are appropriate for each field, with numerical data properly formatted.

4. *Transaction_ID.csv*

Data Contents:

- Transaction ID: Unique identifier for each transaction.
- Customer ID: Link to the customer involved in the transaction.
- Payment Mode: The mode of payment used (e.g., Card, Cash).

Data Types & Completeness:

- The dataset includes 440,098 entries.
- All columns are fully populated with no null values.
- Data types are appropriate: integer for IDs and object for categorical data (payment mode).

Summary:

1. Data Cleaning: For the City.csv, converted Population and Users fields from string to numeric to enable quantitative analysis.
2. Data Transformation: Converted the Date of Travel in the Cab_Data.csv from Julian date to a standard date format and Income converted to Accounting.
3. Data Merging: With identified keys, we can merge these datasets into a master dataset:
    - Transaction ID can link Transaction_ID.csv with Cab_Data.csv.
    - Customer ID can link Customer_ID.csv with the merged transaction data.
    - City information from City.csv can be merged based on city names in the Cab_Data.csv.

With the data now cleaned and transformed, the next steps involve:

1. Merging the Data: Created a master dataset *"Cleaned Dataset.csv"* that combines all the relevant information by joining these datasets on the identified keys.

Analysis Preparation: Once the data is merged, I started creating features, like calculating profit margins, and prepare for the exploratory data analysis (EDA) and hypothesis testing.

Final Merged Data Overview

The datasets have been successfully merged into a comprehensive final dataset *"Cleaned Dataset.csv"* that combines transaction details, customer demographics, and city-specific data. Here's what I have:

- Entries: 359,392 records.
- Columns: 14 columns that include all transaction details, customer information, and city demographics.
- Data Types & Completeness: All columns are populated with no null values, and the data types are appropriate.

This final dataset is ready for exploratory data analysis (EDA) and deeper investigations into the hypotheses. Here are some areas we could explore:

- *Seasonal Patterns*: Analyse cab usage across different seasons or months.
- *Company Performance*: Compare the two cab companies in terms of revenue, number of rides, and profitability.

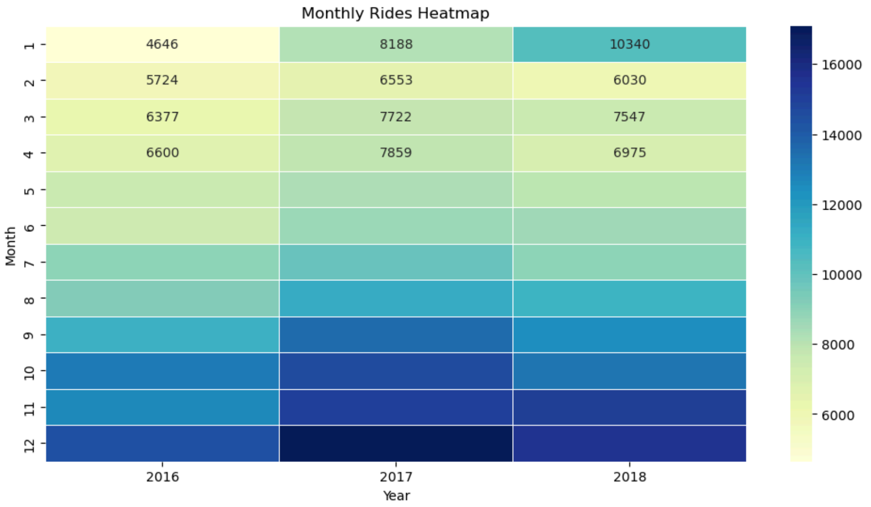
- *Demographic Analysis*: Investigate the relationship between customer demographics (age, income, gender) and cab usage.
- *City Analysis*: Explore how population and the number of cab users in each city correlate with the demand for cabs.
- *Profit Margins*: Calculate and compare the profit margins of the trips for both companies.

Seasonal Patterns

http://localhost:8889/files/1st%20PROJECT/Monthly%20Rides%20Heatmap.ipynb?_xsrf=2%7C3708abf5%7C3adf649b72e7a050b0e6105f1592faef%7C1715764212

Seasonality Effects

I visualized the total number of rides per month to detect any seasonal patterns in the usage of cab services.
I prepared the data for this analysis and create visualizations to explore these trends. Let's start with this analysis.



The heatmap displays the number of cab rides for each month across the years 2016 to 2018. From this visualization, we can observe several points:

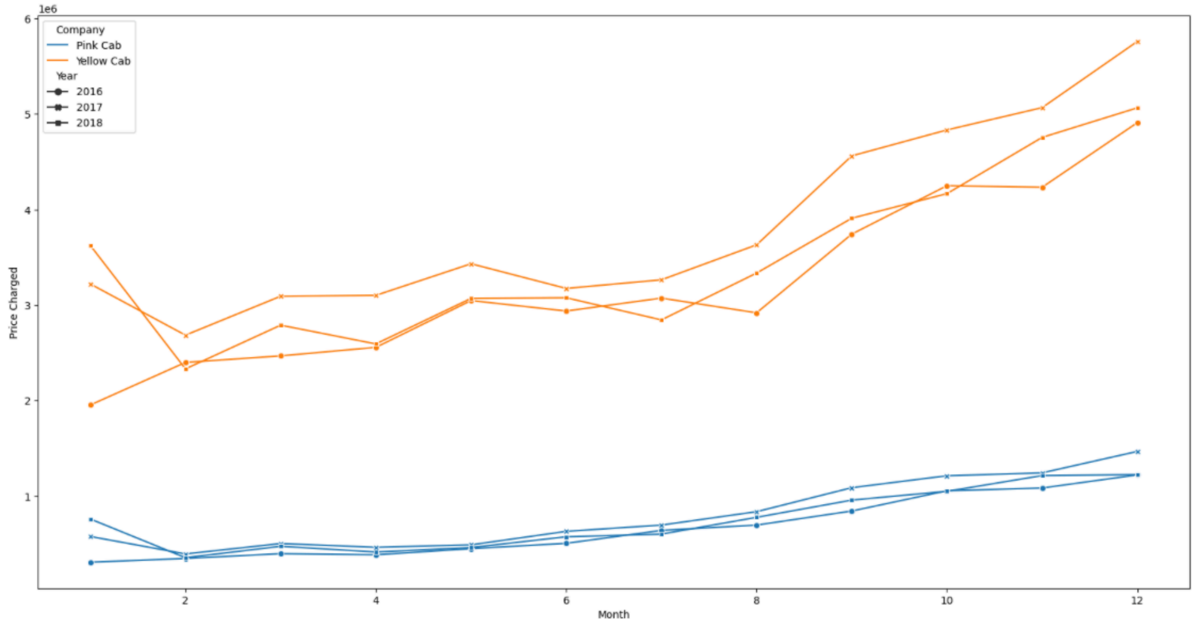
- **Seasonality**: There appears to be an increase in rides towards the end of each year, particularly in December. This could be due to the holiday season and year-end festivities.
- **Yearly Growth**: Each year seems to show a general increase in the number of rides compared to the previous year, indicating growth in the usage of cab services over time.

Company Performance

http://localhost:8889/files/1st%20PROJECT/Company%20performace.ipynb?_xsrf=2%7C3708abf5%7C3adf649b72e7a050b0e6105f1592faef%7C1715764212

I analysed and compare the monthly revenues of the two companies to see which one has been more successful in terms of revenue generation over the specified period.

I prepared and visualized the monthly revenue data for both companies.



The line graph illustrates the monthly revenue trends for the two cab companies across different years:

- **Revenue Trends**: Both companies show a clear cyclical pattern with peaks generally towards the end of each year, likely driven by increased holiday travel.
- **Company Comparison**: The revenue for Yellow Cab is consistently higher than that for Pink Cab across almost all months and years. This suggests that Yellow Cab has a larger share of the market in terms of revenue.

This analysis supports the hypothesis that there is a seasonal pattern in cab usage and indicates that Yellow Cab might be the better performing company in terms of revenue.

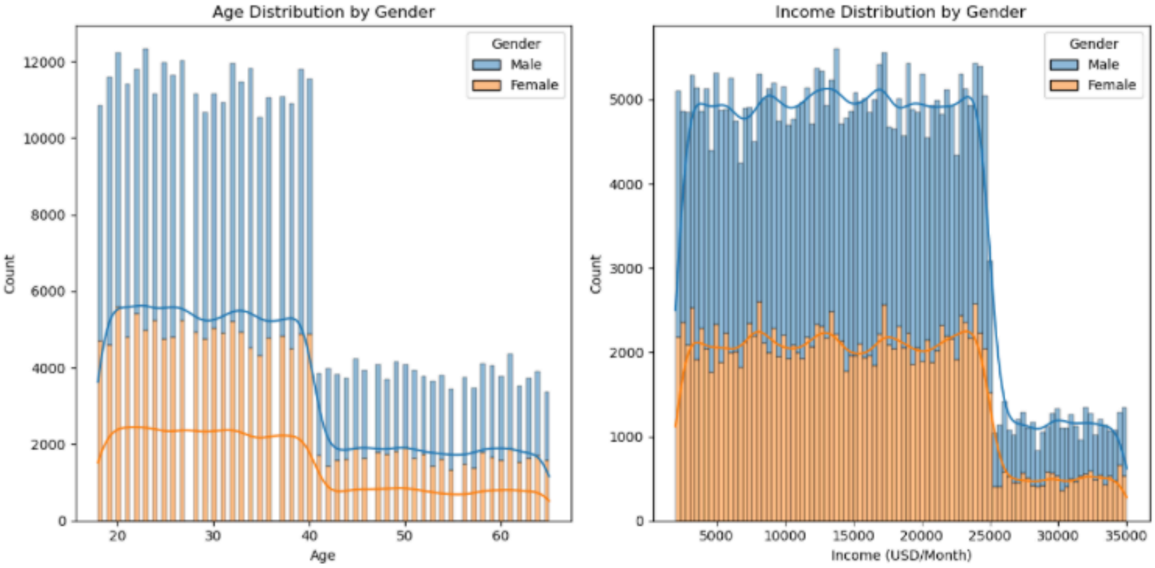
We can further explore other hypotheses such as:

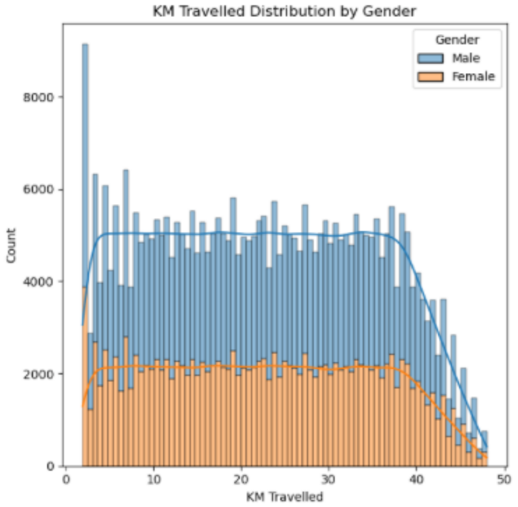
- **Customer Segmentation**: Identify characteristics of the most profitable customer segments.
- **Profitability Analysis**: Examine the relationship between the number of users and profit margins.
- **Impact of City Demographics**: Analyse how demographic factors like population and user density impact cab usage.

Demographic Analysis

http://localhost:8889/files/1st%20PROJECT/BY%20Gender.ipynb?_xsrf=2%7C3708abf5%7C3adf649b72e7a050b0e6105f1592faef%7C1715764212

I investigated the relationship between customer demographics (age, income, gender) and cab usage:





Observations:

1. **Age Distribution by Gender:**
   - The majority of users are within the 20-40 age range for both genders.
   - Male users appear to have a slightly broader age distribution compared to female users.

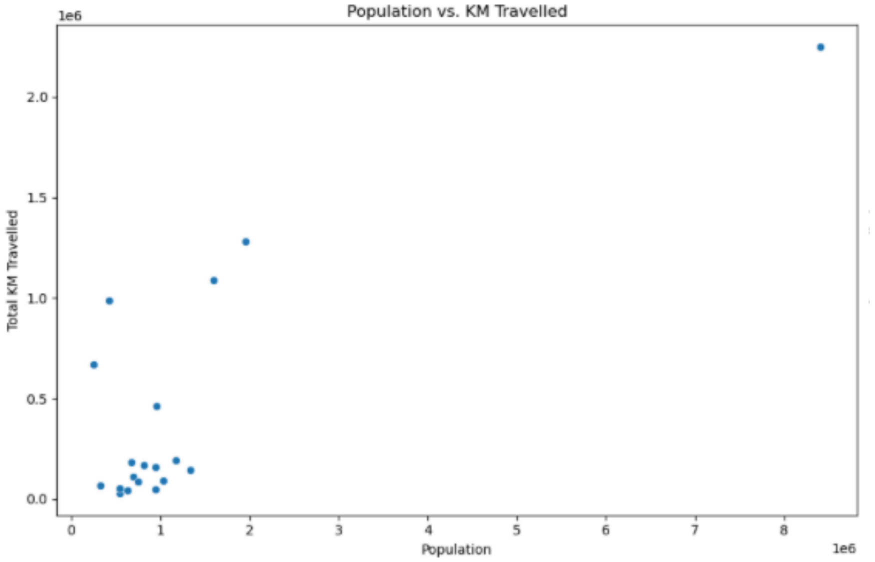
2. **Income Distribution by Gender:**
   - □ Income distribution for both genders shows a peak around the lower income ranges (less than $10,000/month).
   - □ Male users have a broader range of income compared to female users.
3. **KM Travelled Distribution by Gender:**
   - □ Both genders have similar distribution patterns, with most trips being less than 40 KM.
   - □ There are some outliers with longer trips, but these are relatively few.

City Analysis

http://localhost:8889/files/1st%20PROJECT/City%20Analysis.ipynb?_xsrf=2%7C3708abf5%7C3adf649b72e7a050b0e6105f1592faef%7C1715764212

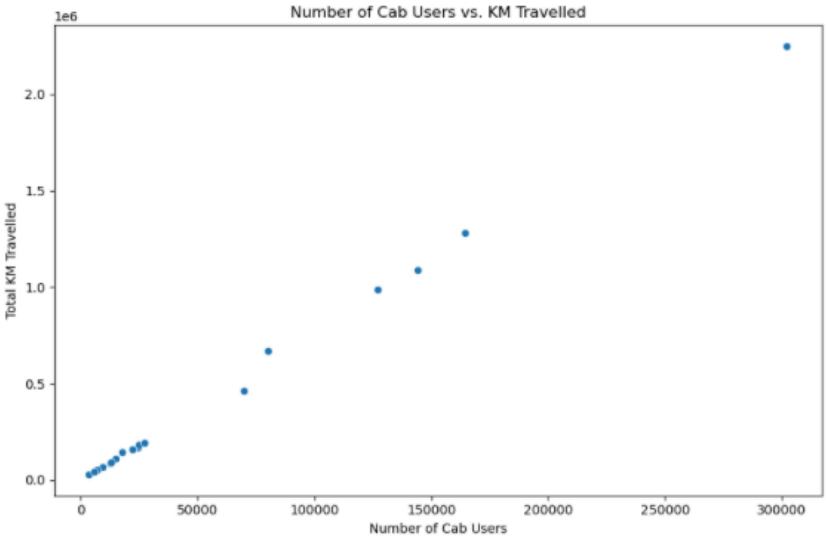
We'll explore how population and the number of cab users in each city correlate with the demand for cabs.

perform the city analysis by exploring how population and the number of cab users in each city correlate with the demand for cabs.



Profit Margins

We'll calculate and compare the profit margins of the trips for both companies. Number of Cab Users vs. KM Travelled:



There is a strong positive correlation between the number of cab users and total kilometers travelled.

Cities with more cab users tend to have higher total kilometres travelled, indicating higher demand.

Profit margin is calculated as:

$$\text{Profit Margin} = \frac{\text{Price Charged} - \text{Cost of Trip}}{\text{Price Charged}} \times 100$$

I computed the profit margins for "Pink Cab"

- trips from "Pink Cab," and the average profit margin for these trips is approximately 24.78%.