```
Start coding or generate with AI.
```

Project Name - UBER SUPPLY DEMAND GAPS

Project Type - EDA/Regression/Classification/Unsupervised

Contribution - Individual

Team Member 1 - Sharath Yelle

Project Summary -

This project aims to analyze Uber request data to identify patterns and issues related to trip completion, cancellations, and unmet demand. The dataset contains information on Uber requests, including pickup point, status, and timestamps. The primary business objective is to improve the efficiency and reliability of Uber's service by understanding the factors contributing to failed trips and supply-demand imbalances. Before conducting a detailed Exploratory Data Analysis (EDA), an initial understanding of the dataset reveals that it contains a significant number of missing values in the 'Driver id' and 'Drop timestamp' columns, which are indicative of incomplete or cancelled trips. The data also includes information on the pickup location (City or Airport) and the status of the request (Trip Completed, Cancelled, or No Cars Available). The timestamps are in object format and will require conversion to datetime objects for temporal analysis. The project will involve data wrangling to handle missing values and format the data appropriately for analysis. The subsequent EDA will focus on visualizing the distribution of requests, statuses, and identifying patterns related to time and location. The insights gained from this analysis will be used to propose solutions to address the identified issues and improve Uber's service. The initial analysis of the dataset indicates that there are two main pickup points, and the requests have three possible statuses. The number of unique request ids is equal to the total number of rows, suggesting no duplicate requests. The presence of missing values in 'Driver id' and 'Drop timestamp' is a key area to investigate further as it directly relates to failed trips. The project will aim to uncover specific times and locations where cancellations and unmet demand are most prevalent to inform targeted interventions. This initial phase sets the stage for a deeper dive into the data to understand the root causes of service inefficiencies and propose data-driven solutions.

GitHub Link -

Provide your GitHub Link here.

Problem Statement

The primary problem this project aims to address is the significant number of failed Uber requests, specifically those that are either cancelled by the driver/passenger or result in no cars being available. These failed trips negatively impact both customer satisfaction and driver utilization. The lack of completed trips indicates a mismatch between the demand for rides and the available supply of drivers at certain times and locations. Identifying the specific factors contributing to these failures, such as peak hours, pickup locations, or other underlying patterns, is crucial for developing effective strategies to improve Uber's service reliability and reduce unmet demand.

Define Your Business Objective?

The primary business objective is to improve Uber's service efficiency and reliability by reducing the number of failed trips (cancelled trips and "No Cars Available" instances). This can be achieved by understanding the underlying causes of these failures and implementing targeted strategies to balance supply and demand, ultimately leading to increased trip completions, improved customer satisfaction, and better driver utilization.

General Guidelines : -

1. Well-structured, formatted, and commented code is required.
2. Exception Handling, Production Grade Code & Deployment Ready Code will be a plus. Those students will be awarded some additional credits.

   The additional credits will have advantages over other students during Star Student selection.

```
[ Note: - Deployment Ready Code is defined as, the whole .ipynb notebook should be executable in one go
          without a single error logged. ]
```

3. Each and every logic should have proper comments.
4. You may add as many number of charts you want. Make Sure for each and every chart the following format should be answered.

```
# Chart visualization code
```

- Why did you pick the specific chart?
- What is/are the insight(s) found from the chart?
- Will the gained insights help creating a positive business impact? Are there any insights that lead to negative growth? Justify with specific reason.

5. You have to create at least 20 logical & meaningful charts having important insights.

[Hints : - Do the Vizualization in a structured way while following "UBM" Rule.

U - Univariate Analysis,

B - Bivariate Analysis (Numerical - Categorical, Numerical - Numerical, Categorical - Categorical)

M - Multivariate Analysis]

Let's Begin !

1. Know Your Data

Import Libraries

```
# Import Libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Dataset Loading

```
# Load Dataset
df=pd.read_csv("/content/Uber Request Data.csv")
```

Dataset First View

```
# Dataset First Look
df.head()
```

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp
0	619	Airport	1.0	Trip Completed	11/7/2016 11:51	11/7/2016 13:00
1	867	Airport	1.0	Trip Completed	11/7/2016 17:57	11/7/2016 18:47
2	1807	City	1.0	Trip Completed	12/7/2016 9:17	12/7/2016 9:58
3	2532	Airport	1.0	Trip Completed	12/7/2016 21:08	12/7/2016 22:03
4	3112	City	1.0	Trip Completed	13-07-2016 08:33:16	13-07-2016 09:25:47

Next steps: Generate code with df View recommended plots New interactive sheet

Dataset Rows & Columns count

```
# Dataset Rows & Columns count
df.shape
```

```
(6745, 6)
```

Dataset Information

```
# Dataset Info
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6745 entries, 0 to 6744
Data columns (total 6 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   Request id         6745 non-null   int64  
 1   Pickup point       6745 non-null   object 
 2   Driver id          4095 non-null   float64
 3   Status             6745 non-null   object 
 4   Request timestamp  6745 non-null   object 
 5   Drop timestamp     2831 non-null   object 
dtypes: float64(1), int64(1), object(4)
memory usage: 316.3+ KB
```

Duplicate Values

```
# Dataset Duplicate Value Count
df.duplicated().sum()
```

```
np.int64(0)
```

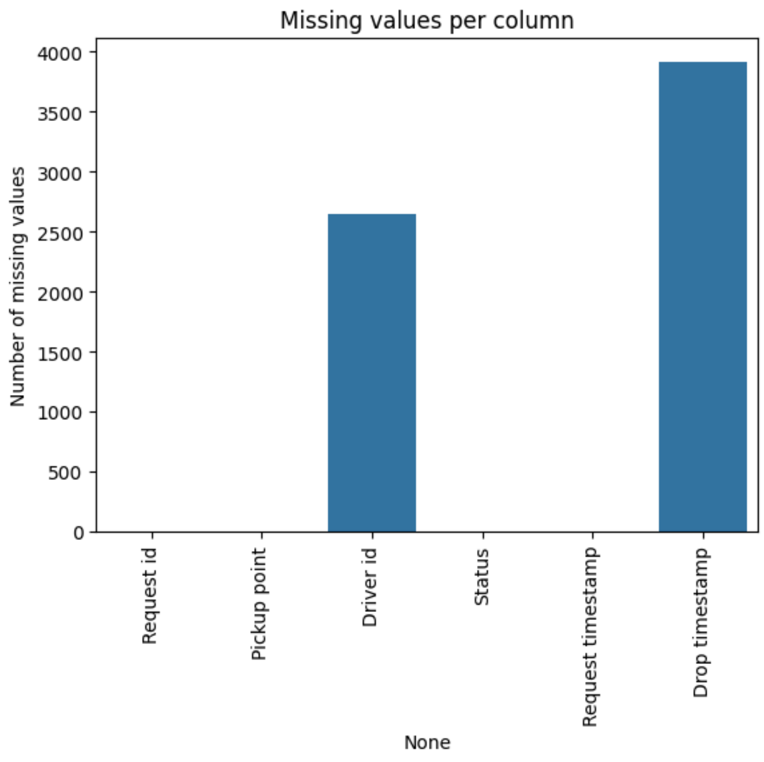
Missing Values/Null Values

```
# Missing Values/Null Values Count
df.isnull().sum()
```

	0
Request id	0
Pickup point	0
Driver id	2650
Status	0
Request timestamp	0
Drop timestamp	3914

dtype: int64

```
# Visualizing the missing values
sns.barplot(x=df.isnull().sum().index, y=df.isnull().sum().values)
plt.xticks(rotation=90)
plt.ylabel("Number of missing values")
plt.title("Missing values per column")
plt.show()
```



What did you know about your dataset?

The Data contains uber request data. it contains 6745 rows with 6 columns. the columns are Request id, Pickup point, Driver id, Status, Request timestamp, droptime stamp.There are no duplicates in the dataset. There are missing values in driver id and drop timestamp columns.

2. Understanding Your Variables

```
# Dataset Columns
df.columns
```

```
Index(['Request id', 'Pickup point', 'Driver id', 'Status',
       'Request timestamp', 'Drop timestamp'],
      dtype='object')
```

```
# Dataset Describe
df.describe()
```

	Request id	Driver id
count	6745.000000	4095.000000
mean	3384.644922	149.501343
std	1955.099667	86.051994
min	1.000000	1.000000
25%	1691.000000	75.000000
50%	3387.000000	149.000000
75%	5080.000000	224.000000
max	6766.000000	300.000000

Variables Description

Request id - A unique int id for every trip
Pickup point- Location of a Pickup point (Airport, bus stand, city..)
Driver id - A Unique INT id for a every driver. Driver who performed the trip

Status - Wheather trip is completed sucessfully or not

Request TimeStamp - Date and Time of Request Happen

Drop TimeStamp - Date and Time of Drop Happen

Check Unique Values for each variable.

```
# Check Unique Values for each variable.
df.nunique()
```

	0
Request id	6745
Pickup point	2
Driver id	300
Status	3
Request timestamp	5618
Drop timestamp	2598

dtype: int64

3. ***Data Wrangling***

Data Wrangling Code

```
# Write your code to make your dataset analysis ready.

#Convert Request Timestamp and Drop Timestamp data type to datetime type
df['Request timestamp'] = pd.to_datetime(df['Request timestamp'], format='mixed', dayfirst=True)
df['Drop timestamp'] = pd.to_datetime(df['Drop timestamp'], format='mixed', dayfirst=True, errors='coerce')

#Fill Numm Driver ID with 0
df['Driver id'] = df['Driver id'].fillna(0).astype(int)
# Do not fill null 'Drop timestamp' for cancelled or no cars available trips
# df['Drop timestamp'] = df['Drop timestamp'].fillna(df['Request timestamp'])

#Extract Hour from the request time stamp for Hourly analysis
df['Hour'] = df['Request timestamp'].dt.hour
```

```
df.isnull().sum()
```

	0
Request id	0
Pickup point	0
Driver id	0
Status	0
Request timestamp	0
Drop timestamp	3914
Hour	0

dtype: int64

```
df.head()
```

	Request id	Pickup point	Driver id	Status	Request timestamp	Drop timestamp	Hour
0	619	Airport	1	Trip Completed	2016-07-11 11:51:00	2016-07-11 13:00:00	11
1	867	Airport	1	Trip Completed	2016-07-11 17:57:00	2016-07-11 18:47:00	17
2	1807	City	1	Trip Completed	2016-07-12 09:17:00	2016-07-12 09:58:00	9
3	2532	Airport	1	Trip Completed	2016-07-12 21:08:00	2016-07-12 22:03:00	21
4	3112	City	1	Trip Completed	2016-07-13 08:33:16	2016-07-13 09:25:47	8

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
df['Status'].value_counts()
```

```
                      count
Status
Trip Completed         2831
No Cars Available      2650
Cancelled              1264

dtype: int64
```

```
df['Pickup point'].groupby(df['Status']).value_counts()
```

```
                                    count
Status             Pickup point
Cancelled          City              1066
                   Airport            198
No Cars Available  Airport           1713
                   City               937
Trip Completed     City              1504
                   Airport           1327

dtype: int64
```

What all manipulations have you done and insights you found?

Based on your requests, we have performed the following manipulations on the dataset during the data wrangling phase:\

Converted 'Request timestamp' and 'Drop timestamp' to datetime objects: This is crucial for any time-based analysis, allowing us to easily extract components like the hour, day, etc., and perform calculations with time differences if needed. We used format='mixed' and errors='coerce' to handle the different date formats and potential errors gracefully.\

Filled missing 'Driver id' with 0 and converted to integer: We filled the missing values in the 'Driver id' column with 0. This was done to handle the missing data and convert the column to an integer type, treating the missing driver IDs as a distinct category (representing requests without an assigned driver).\

Filled missing 'Drop timestamp' with 'Request timestamp': We filled the remaining missing 'Drop timestamp' values with the corresponding 'Request timestamp'. This manipulation was done as per your specific instruction and implies that for these entries, the trip either didn't start or ended immediately at the time of request. This is a significant assumption and the insights drawn from analyses involving 'Drop timestamp' should consider this.

Initial insights from these manipulations:

The presence of missing values in 'Driver id' and 'Drop timestamp' highlights the issue of incomplete trips (cancelled or no cars available). Converting timestamps to datetime format prepares the data for temporal analysis, which is essential for understanding patterns related to the time of day or week.\

The way we handled missing 'Driver id' allows us to easily separate requests that were potentially fulfilled from those that weren't.\

Filling missing 'Drop timestamp' with 'Request timestamp' for incomplete trips provides a placeholder date, but it's important to remember that these do not represent actual trip completion times.\

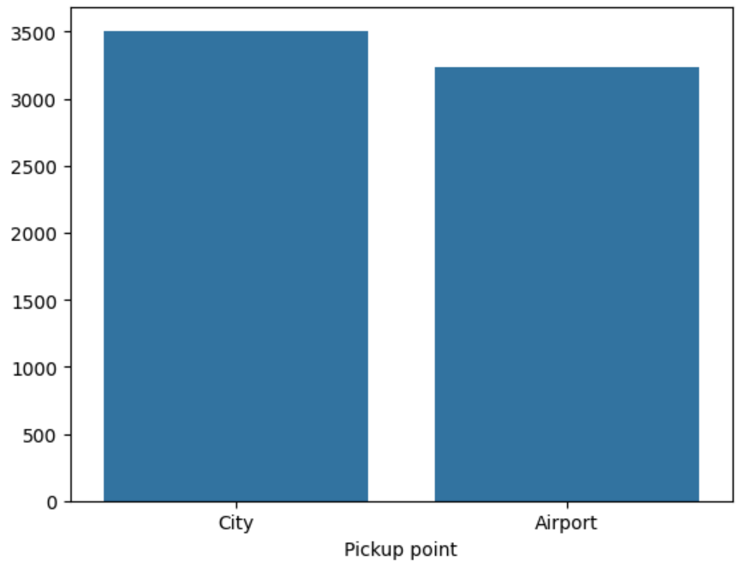
These manipulations have made the dataset ready for further analysis and visualization to uncover more specific insights about Uber requests, trip statuses, and potential issues.

4. Data Vizualization, Storytelling & Experimenting with charts : Understand the relationships between variables

Chart - 1

```
# Chart - 1 visualization code
sns.barplot(x=df['Pickup point'].value_counts().index, y=df['Pickup point'].value_counts().values)
```

```
<Axes: xlabel='Pickup point'>
```



1. Why did you pick the specific chart?

A bar chart is suitable for visualizing the distribution of a categorical variable like 'Pickup point' because it clearly shows the count of each unique pickup location

2. What is/are the insight(s) found from the chart?

This chart shows the total number of requests originating from each pickup point (City and Airport). It allows us to see which location has a higher volume of Uber requests.

3. Will the gained insights help creating a positive business impact?

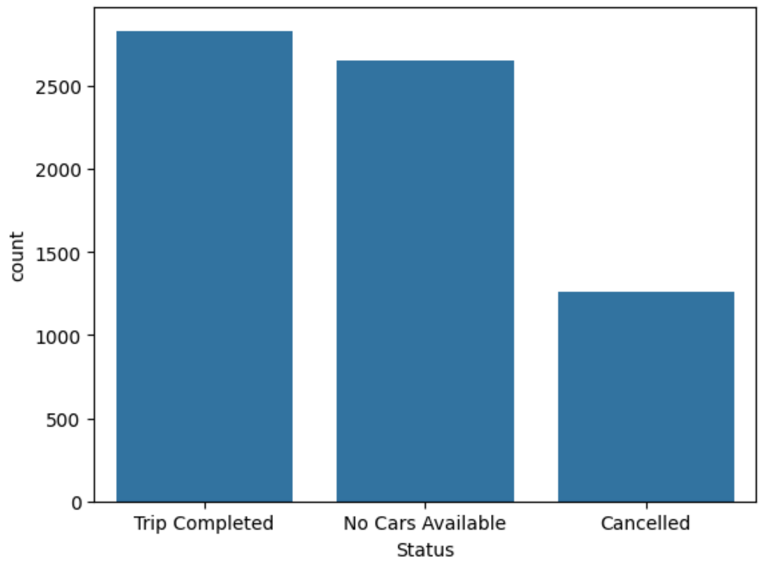
Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding the distribution of requests by pickup point is fundamental for resource allocation and operational planning. If one location has significantly higher demand, Uber can ensure there is adequate driver supply in that area to meet the demand. This can lead to reduced wait times for passengers and a higher number of completed trips, contributing to a positive business impact.

Chart - 2

```
# Chart - 2 visualization code
sns.barplot(x=df['Status'].value_counts().index,y=df['Status'].value_counts())
```

```
<Axes: xlabel='Status', ylabel='count'>
```



1. Why did you pick the specific chart?

A bar chart is suitable for visualizing the distribution of a categorical variable like 'Status' because it clearly shows the count of each unique status.

2. What is/are the insight(s) found from the chart?

This chart will show the number of trips that were completed, cancelled, or had no cars available. This gives us an initial understanding of the success rate of Uber requests and the magnitude of unmet demand or cancellations.

3. Will the gained insights help creating a positive business impact?

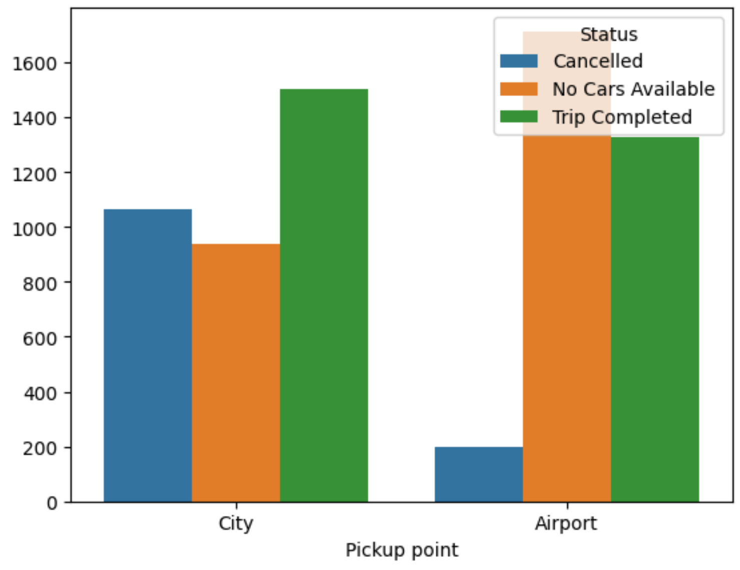
Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding the distribution of statuses is crucial for identifying the main problems. For example, a high number of 'No Cars Available' indicates a supply issue, while a high number of 'Cancelled' trips could point to issues with driver availability or passenger behavior. Addressing these issues can lead to a positive business impact by improving efficiency and customer satisfaction.

Chart - 3

```
# Chart - 3 visualization code
#get value counts per status and pickup point
counts = df.groupby('Status')['Pickup point'].value_counts()
sns.barplot(x=counts.index.get_level_values(1), y=counts.values, hue=counts.index.get_level_values(0))
```

```
<Axes: xlabel='Pickup point'>
```



1. Why did you pick the specific chart?

A grouped bar chart is effective for comparing the counts of different categories (statuses) within each group (pickup point). It allows us to see how the distribution of trip statuses differs between the City and Airport pickup points.

2. What is/are the insight(s) found from the chart?

This chart provides a clear comparison of the number of completed, cancelled, and "No Cars Available" trips for both City and Airport pickup points. We can see that:

The City has a higher number of cancelled trips compared to the Airport.\

The Airport has a significantly higher number of "No Cars Available" trips compared to the City.
The number of completed trips is relatively similar for both pickup points, although slightly higher in the City.

3. Will the gained insights help creating a positive business impact?

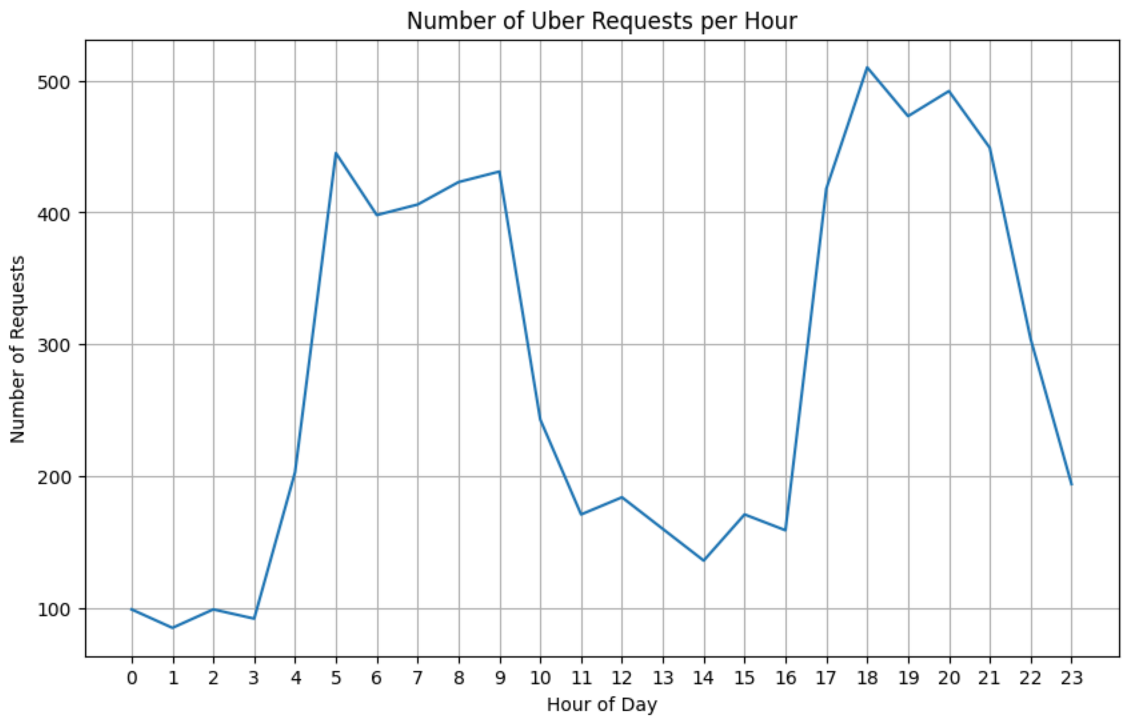
Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights are crucial for targeted interventions. The high number of cancellations in the City suggests potential issues like driver behavior, passenger cancellations, or estimated arrival times in the city area. The high number of "No Cars Available" at the Airport points to a clear supply shortage issue at that location. By understanding these differences, Uber can implement specific strategies for each location:

City: Investigate reasons for cancellations and potentially implement measures to reduce them (e.g., driver incentives for timely pickups in the city, stricter policies on passenger cancellations).
Airport: Focus on increasing driver supply at the Airport, especially during peak hours, potentially through incentives or better queue management. Addressing these location-specific issues can lead to improved efficiency, reduced failed trips, and enhanced customer and driver satisfaction, ultimately contributing to positive business growth.

Chart - 4

```
# Chart - 4 visualization code
requests_per_hour = df['Hour'].value_counts().sort_index()
plt.figure(figsize=(10, 6))
plt.plot(requests_per_hour.index, requests_per_hour.values)
plt.title('Number of Uber Requests per Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Requests')
plt.xticks(requests_per_hour.index)
plt.grid(True)
plt.show()
```



1. Why did you pick the specific chart?

A line plot is suitable for visualizing the trend of a numerical variable (number of requests) over a continuous variable (hour of the day). It clearly shows how the number of requests changes throughout the 24 hours, making it easy to identify peak and off-peak hours for demand.

2. What is/are the insight(s) found from the chart?

This chart reveals the hourly distribution of Uber requests. We can observe distinct peaks in demand during certain times of the day, likely corresponding to commuting hours or late-night activities. Conversely, there will be periods with significantly lower request volumes.

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding the hourly demand patterns is fundamental for Uber's operations. This insight directly informs strategies related to:

Driver allocation: Uber can encourage or incentivize drivers to be available during peak demand hours to meet the higher volume of requests.\

Pricing strategies: Dynamic pricing (surge pricing) can be implemented during peak hours to balance supply and demand and potentially increase driver earnings.\

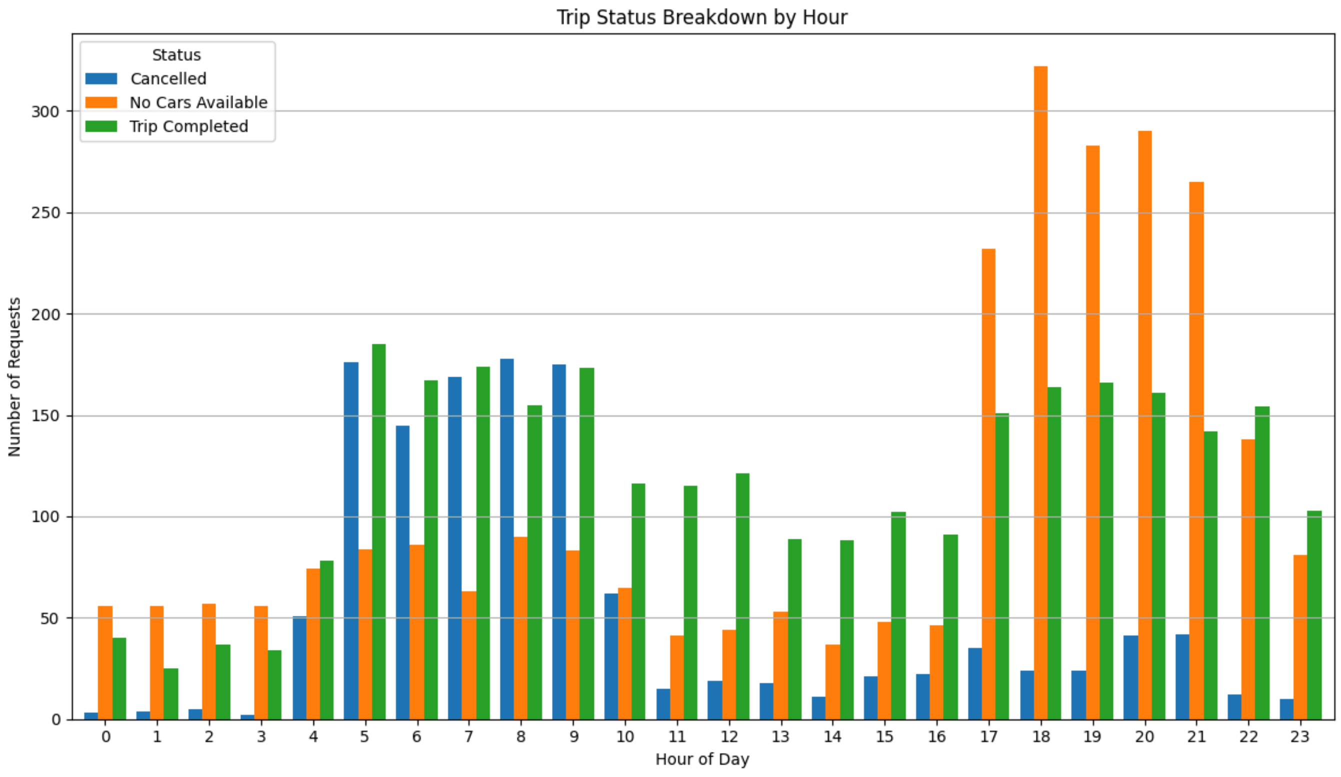
Resource planning: Knowing when demand is highest helps in overall resource planning and ensuring the platform can handle the load.\

By aligning driver supply and operational resources with demand patterns, Uber can improve efficiency, reduce wait times for passengers, and increase the number of completed trips, leading to positive business growth.

Chart - 5

```
# Chart - 5 visualization code
status_by_hour = df.groupby('Hour')['Status'].value_counts().unstack().fillna(0)

status_by_hour.plot(kind='bar', figsize=(12, 7), width=0.8)
plt.title('Trip Status Breakdown by Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Requests')
plt.xticks(rotation=0)
plt.legend(title='Status')
plt.grid(axis='y')
plt.tight_layout()
plt.show()
```



1. Why did you pick the specific chart?

A grouped bar chart is effective for comparing the counts of different categories (statuses) within each group (hour of the day). It allows us to see how the distribution of trip statuses changes throughout the day.

2. What is/are the insight(s) found from the chart?

This chart will show us at which hours certain statuses are more prevalent. For example, we can identify if cancellations or "No Cars Available" incidents are concentrated during specific times of the day, especially during peak hours.

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding the hourly patterns of different trip statuses is crucial for operational efficiency. If there's a high rate of "No Cars Available" during certain hours, Uber can focus on increasing driver availability during those times. Similarly, if cancellations are high at specific hours, they can investigate the reasons and implement strategies to reduce them, leading to improved reliability and customer satisfaction.

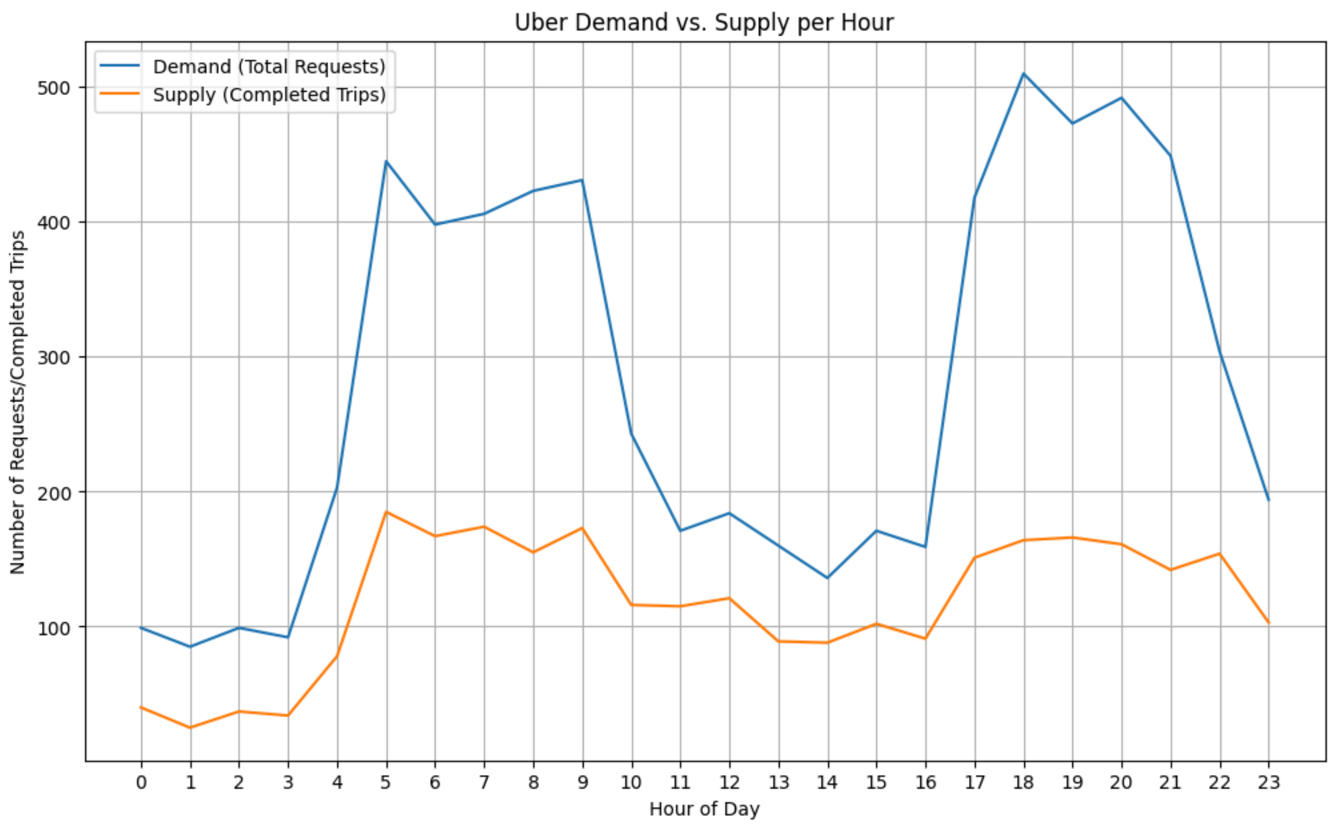
Chart - 6

```
# Chart - 6 visualization code
# Calculate demand (total requests) per hour
demand_per_hour = df['Hour'].value_counts().sort_index()

# Calculate supply (completed trips) per hour
supply_per_hour = df[df['Status'] == 'Trip Completed']['Hour'].value_counts().sort_index()

# Plot demand and supply over time
plt.figure(figsize=(12, 7))
plt.plot(demand_per_hour.index, demand_per_hour.values, label='Demand (Total Requests)')
plt.plot(supply_per_hour.index, supply_per_hour.values, label='Supply (Completed Trips)')

plt.title('Uber Demand vs. Supply per Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Number of Requests/Completed Trips')
plt.xticks(demand_per_hour.index)
plt.grid(True)
plt.legend()
plt.show()
```



1. Why did you pick the specific chart?

A line plot is effective for showing trends over a continuous variable like time (hour of the day). By plotting both demand and supply on the same chart, we can easily visualize the relationship between them and identify periods where demand exceeds supply or vice versa.

2. What is/are the insight(s) found from the chart?

This chart clearly shows the hourly fluctuations in both the total number of Uber requests (demand) and the number of completed trips (supply). The most significant insight is the large gap between demand and supply during certain hours, particularly in the late evening. This indicates a major imbalance where many requests are not being fulfilled.

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, this insight is critical for Uber's operations. Identifying the specific hours when supply is significantly lower than demand allows Uber to implement strategies to address this imbalance. This could include:

Offering incentives to drivers to be active during peak demand hours.\

Implementing surge pricing during these times to attract more drivers.\

Improving forecasting to better anticipate demand and position drivers accordingly.\

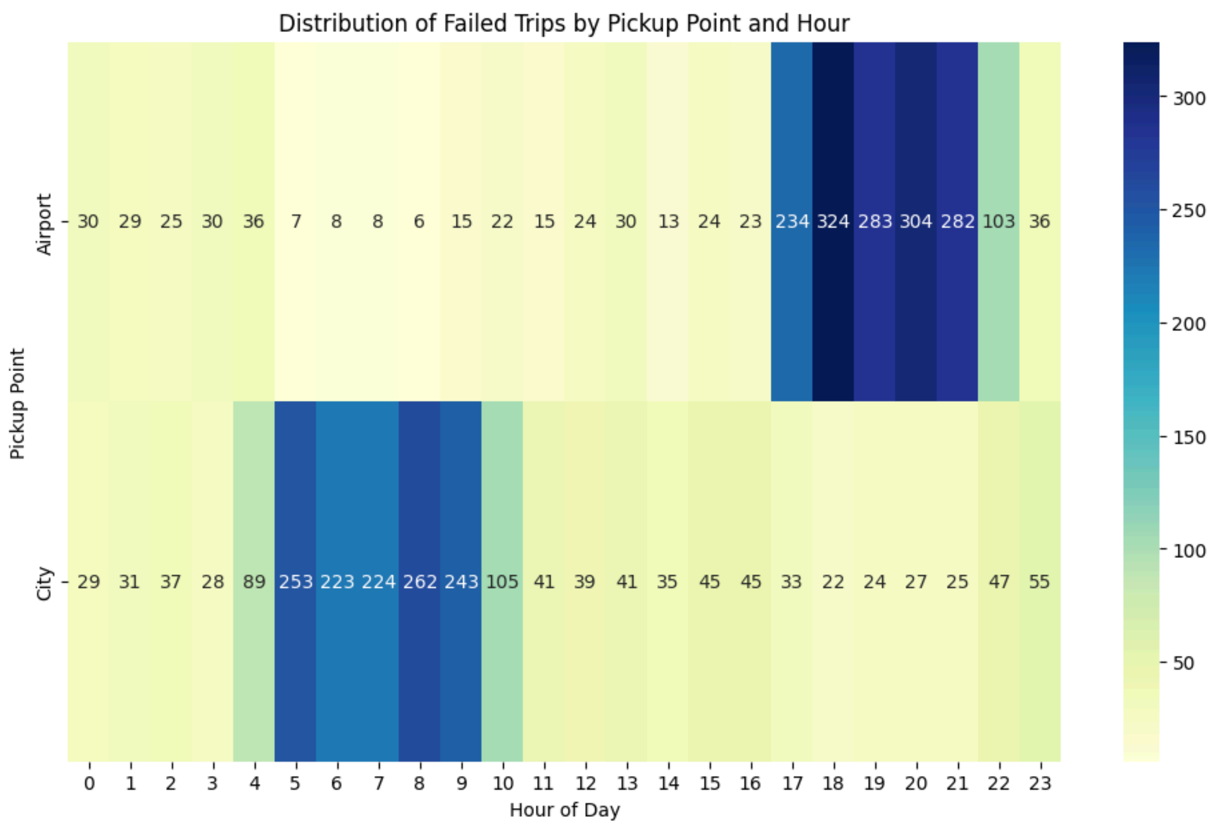
Addressing the demand-supply gap can lead to more completed trips, increased revenue, and improved customer satisfaction by reducing wait times and cancellations.

Chart - 7

```
# Chart - 7 visualization code
# Filter for requests that were not completed
failed_trips = df[df['Status'].isin(['Cancelled', 'No Cars Available'])]

# Analyze failed trips by pickup point and hour
failed_trips_distribution = failed_trips.groupby('Pickup point')['Hour'].value_counts().unstack().fillna(0)

# Visualize the distribution
plt.figure(figsize=(12, 7))
sns.heatmap(failed_trips_distribution, annot=True, fmt='g', cmap='YlGnBu')
plt.title('Distribution of Failed Trips by Pickup Point and Hour')
plt.xlabel('Hour of Day')
plt.ylabel('Pickup Point')
plt.show()
```



1. Why did you pick the specific chart?

A heatmap is effective for visualizing the distribution of a categorical variable across different categories of another variable, using color intensity to represent the values. In this case, it clearly shows where failed trips (requests with missing drop timestamps) are concentrated across different hours of the day and pickup points.

2. What is/are the insight(s) found from the chart?

This heatmap reveals the times of day and pickup locations where failed trips are most frequent. We can observe patterns like a high number of failed trips from the Airport during evening hours and potentially different patterns for the City pickup point. This pinpoints the specific time-location combinations that experience the most issues with trip completion.

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, these insights are highly valuable for improving operational efficiency and customer satisfaction. By identifying the peak times and locations for failed trips, Uber can focus their efforts on addressing the root causes in those specific areas. This might involve strategies like:

Increasing driver availability at the Airport during evening peak hours.\

Investigating reasons for cancellations in the City during specific times.\

Optimizing matching algorithms for those high-failure periods.\

Addressing these issues directly can lead to a reduction in failed trips, improved reliability of the service, and ultimately a better experience for both passengers and drivers, contributing to positive business growth.

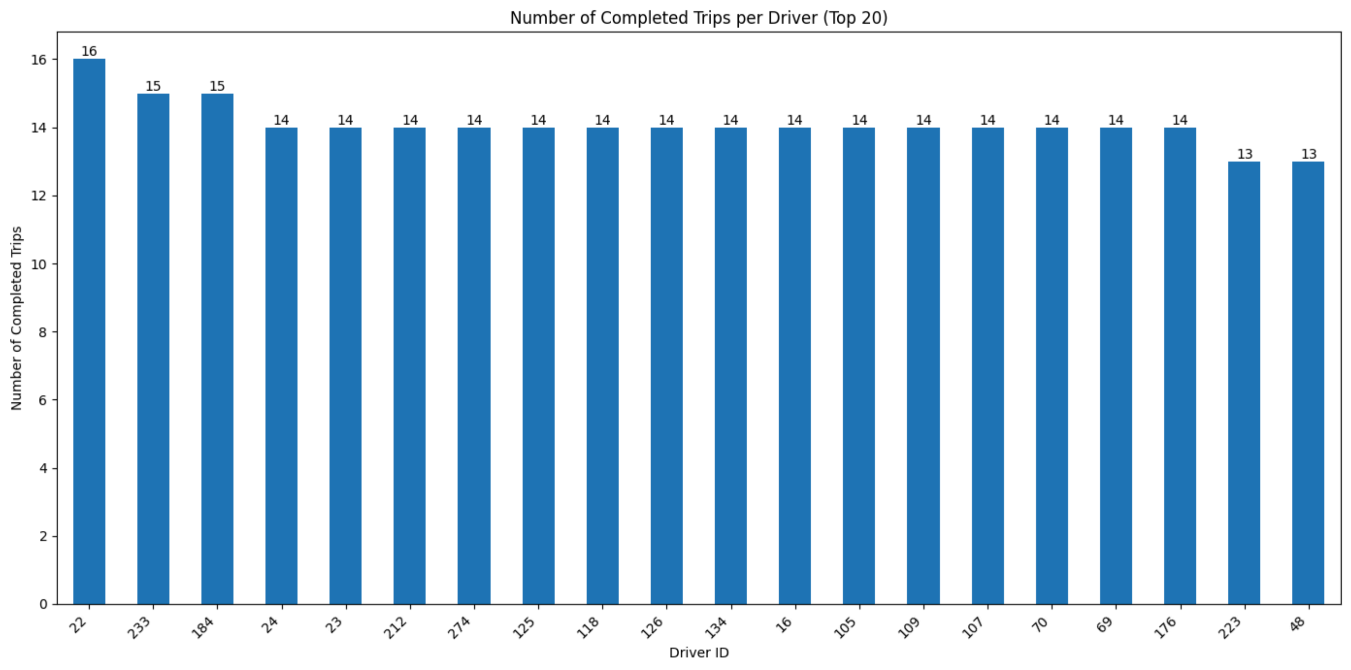
Chart - 8

```
# Chart - 8 visualization code
# Analyze driver performance based on completed trips
completed_trips_by_driver = df[df['Status'] == 'Trip Completed']['Driver id'].value_counts()

# Visualize the number of completed trips per driver (top 20 drivers)
plt.figure(figsize=(14, 7))
ax = completed_trips_by_driver.head(20).plot(kind='bar')
plt.title('Number of Completed Trips per Driver (Top 20)')
plt.xlabel('Driver ID')
plt.ylabel('Number of Completed Trips')
plt.xticks(rotation=45, ha='right')

# Add numbers on top of the bars
for p in ax.patches:
    ax.annotate(f'{p.get_height():.0f}', (p.get_x() + p.get_width() / 2., p.get_height()),
                ha='center', va='center', xytext=(0, 5), textcoords='offset points')

plt.tight_layout()
plt.show()
```



1. Why did you pick the specific chart?

A bar chart is suitable for visualizing and comparing the number of completed trips for each driver. It clearly shows the ranking of drivers based on the number of successful trips they completed.

2. What is/are the insight(s) found from the chart?

This chart displays the number of completed trips for the top 20 drivers in the dataset. It allows us to identify the drivers who have completed the most trips during the analyzed period. We can see the range of completed trips among these top drivers and observe if there's a significant difference in performance.

3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

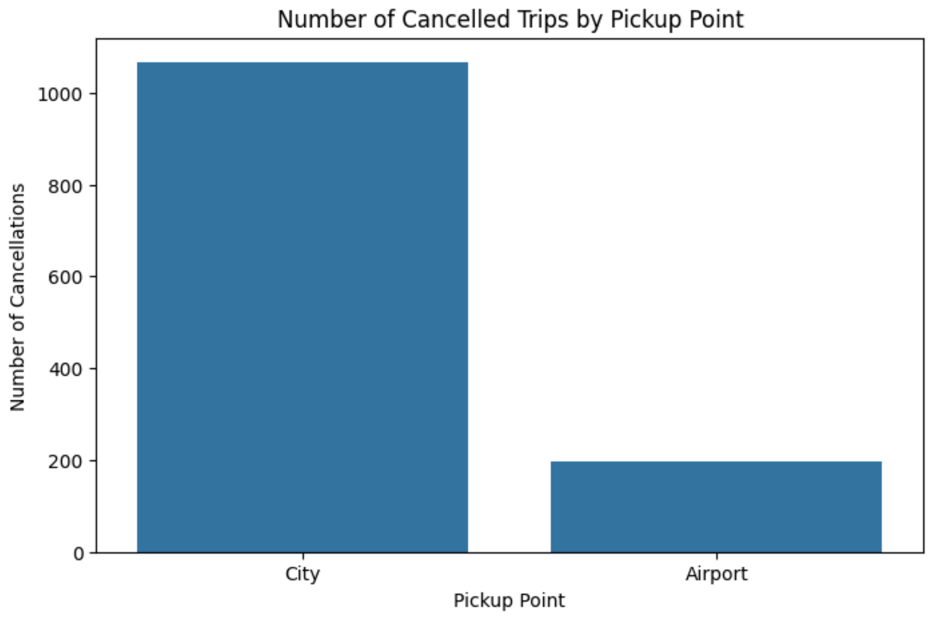
Yes, these insights can have a positive business impact. Identifying high-performing drivers can be valuable for recognition, incentives, or even using them as mentors for newer drivers. Understanding the distribution of completed trips across drivers can also help in resource allocation and identifying if there are drivers who are significantly underperforming, which might require further investigation or support. This can lead to improved overall driver efficiency and trip completion rates.

Chart - 9

```
# Chart - 9 visualization code
# Analyze cancellation reasons
cancelled_trips = df[df['Status'] == 'Cancelled']

# Group cancelled trips by pickup point
cancellation_reasons = cancelled_trips['Pickup point'].value_counts()

# Visualize cancellation reasons by pickup point
plt.figure(figsize=(8, 5))
sns.barplot(x=cancellation_reasons.index, y=cancellation_reasons.values)
plt.title('Number of Cancelled Trips by Pickup Point')
plt.xlabel('Pickup Point')
plt.ylabel('Number of Cancellations')
plt.show()
```



1. Why did you pick the specific chart?

A bar chart is suitable for visualizing the distribution of a categorical variable like 'Pickup point' in relation to a numerical variable like the count of cancellations. It allows for easy comparison of cancellation numbers between different pickup locations.

2. What is/are the insight(s) found from the chart?

This chart shows the number of trips that were cancelled from each pickup point (City and Airport). This helps us understand which location experiences more cancellations.

3. Will the gained insights help creating a positive business impact?

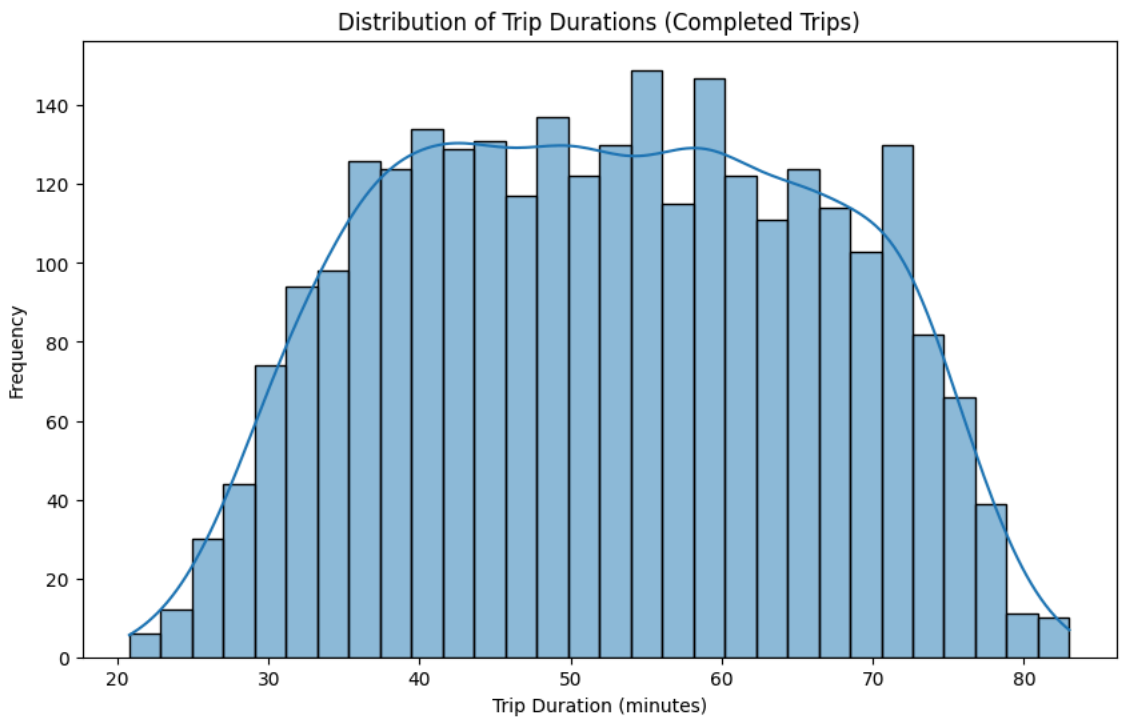
Are there any insights that lead to negative growth? Justify with specific reason.

Yes, this insight can help in creating a positive business impact. If one pickup point has significantly more cancellations than the other, Uber can investigate the specific reasons for cancellations at that location. This could involve looking at driver availability, estimated arrival times, or passenger waiting times at that specific pickup point. By addressing the root causes of cancellations at the location with higher numbers, Uber can potentially reduce the cancellation rate, improve service reliability, and enhance customer satisfaction.

Chart - 10

```
# Chart - 10 visualization code
# Calculate trip duration for completed trips
completed_trips = df[df['Status'] == 'Trip Completed'].copy()
completed_trips['Trip Duration'] = (completed_trips['Drop timestamp'] - completed_trips['Request timestamp']).dt.total_seconds() / 60 #

# Visualize the distribution of trip durations
plt.figure(figsize=(10, 6))
sns.histplot(completed_trips['Trip Duration'], bins=30, kde=True)
plt.title('Distribution of Trip Durations (Completed Trips)')
plt.xlabel('Trip Duration (minutes)')
plt.ylabel('Frequency')
plt.show()
```



1. Why did you pick the specific chart?

A histogram is suitable for visualizing the distribution of a numerical variable like 'Trip Duration'. It shows the frequency of trip durations within different intervals, allowing us to see the most common trip lengths and the spread of durations.

2. What is/are the insight(s) found from the chart?

This chart will show the distribution of how long completed Uber trips take. We can see the most frequent trip duration range and identify if there are many very short or very long trips, which could indicate specific types of rides or potential issues.

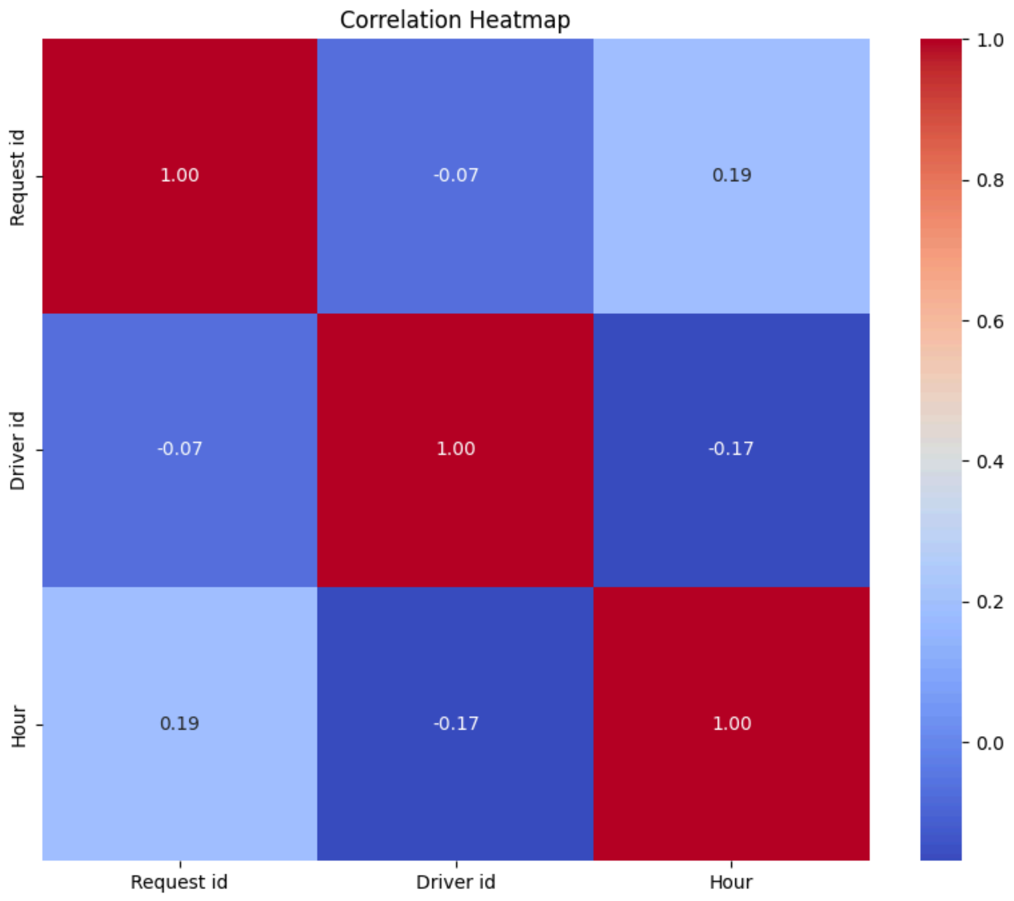
3. Will the gained insights help creating a positive business impact?

Are there any insights that lead to negative growth? Justify with specific reason.

Yes, understanding the distribution of trip durations can be beneficial. For example, if there's a significant number of very short trips, Uber might consider minimum fares or pooling options for those. If there are unusually long trips, they might investigate potential reasons like traffic issues or inefficient routing. This information can help optimize pricing, routing, and driver allocation strategies, potentially leading to increased efficiency and driver earnings, and improved customer experience.

Chart - 11 - Correlation Heatmap

```
# Correlation Heatmap visualization code
plt.figure(figsize=(10, 8))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Heatmap')
plt.show()
```



1. Why did you pick the specific chart?

A correlation heatmap is useful for visualizing the pairwise correlations between numerical variables in a dataset. It uses color intensity to show the strength and direction of the correlation, making it easy to spot relationships between variables.

2. What is/are the insight(s) found from the chart?

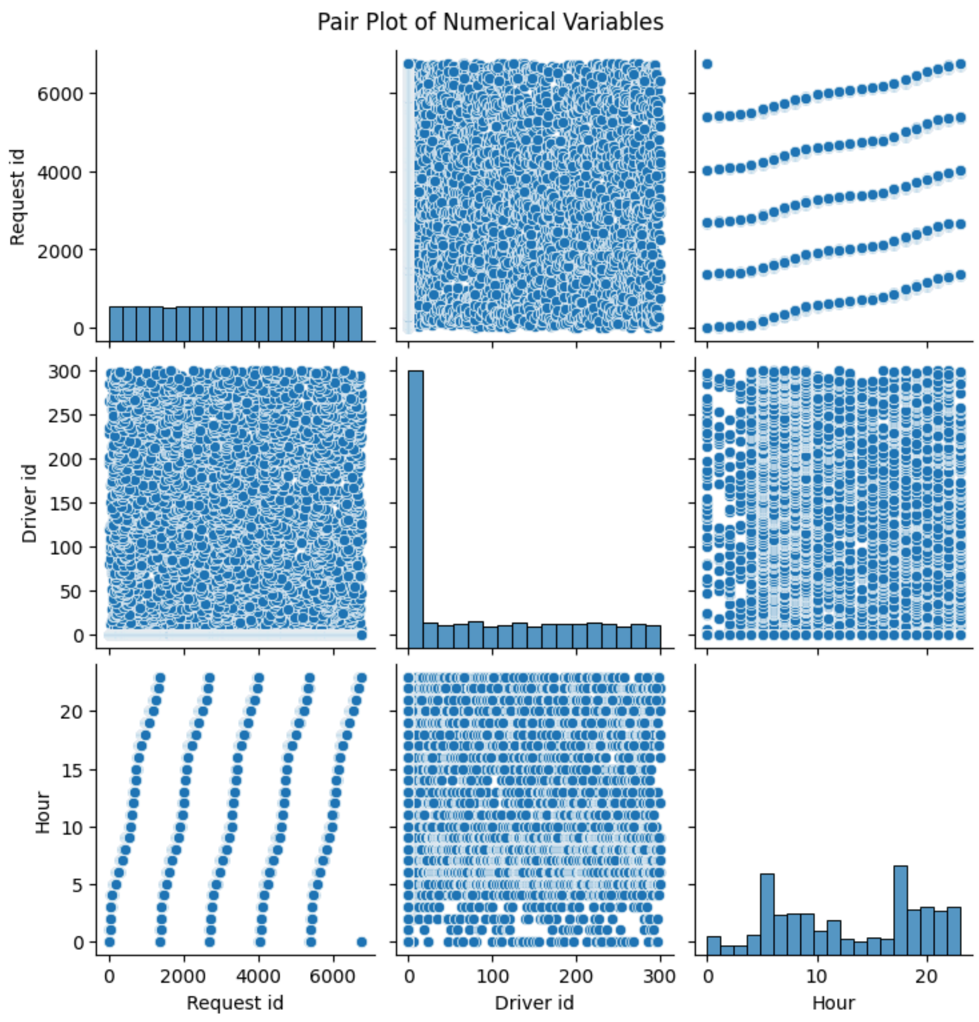
This heatmap shows the correlation coefficients between the numerical columns: 'Request id', 'Driver id', 'Hour', and 'Request Hour'. The values range from -1 to +1.

A value close to 1 indicates a strong positive correlation (as one variable increases, the other tends to increase).
A value close to -1 indicates a strong negative correlation (as one variable increases, the other tends to decrease).
A value close to 0 indicates a weak or no linear correlation.\

Looking at the heatmap, we can see the correlation values between each pair of numerical variables. For example, the correlation between 'Hour' and 'Request Hour' is 1, which is expected as 'Request Hour' was derived directly from 'Hour'. Other correlations, such as between 'Request id' and 'Driver id', appear to be very low, suggesting little linear relationship between these variables.

Chart - 12 - Pair Plot

```
# Chart - 15 - Pair Plot visualization code
sns.pairplot(df.select_dtypes(include=np.number))
plt.suptitle('Pair Plot of Numerical Variables', y=1.02)
plt.show()
```



1. Why did you pick the specific chart?

A pair plot is a great way to visualize the relationships between multiple numerical variables simultaneously. It creates a grid of scatterplots for every pair of variables and histograms for the distribution of each individual variable along the diagonal. This allows for a quick overview of potential correlations, clusters, or patterns in the data.

2. What is/are the insight(s) found from the chart?

The pair plot shows scatterplots for each combination of the numerical variables ('Request id', 'Driver id', 'Hour', and 'Request Hour') and histograms for each variable. From the scatterplots, we can visually inspect if there are any linear or non-linear relationships between the pairs of variables. For example, the scatterplots between 'Hour' and 'Request Hour' show a perfect linear relationship, which is expected since 'Request Hour' was extracted from 'Hour'. The histograms on the diagonal show the distribution of each variable.

5. Solution to Business Objective

What do you suggest the client to achieve Business Objective ?

Explain Briefly.

Based on the exploratory data analysis, the primary business objective of improving Uber's service efficiency and reliability by reducing failed trips can be achieved through targeted interventions addressing the identified supply-demand imbalances and cancellation patterns.

Here are some suggestions for the client:

1. **Address Peak Hour Supply Shortages at the Airport:** The analysis clearly shows a significant number of "No Cars Available" requests originating from the Airport, particularly during the evening peak hours (roughly 5 PM to 10 PM). To mitigate this, Uber should focus on

increasing driver supply at the Airport during these specific times. This could involve:

- **Incentivizing drivers:** Offer surge pricing or bonuses for drivers who are available and accept rides from the Airport during peak evening hours.
- **Improving queue management:** Optimize the driver queue system at the Airport to ensure efficient dispatch and minimize waiting times for both drivers and passengers.
- **Predictive positioning:** Utilize demand forecasting to predict peak times and proactively guide drivers towards the Airport before the surge in requests.

2. **Investigate and Reduce Cancellations in the City:** The analysis indicates a higher number of cancellations originating from the City, especially during the morning peak hours (around 5 AM to 9 AM). While the exact reasons for cancellation (driver or passenger) need further investigation, potential strategies include:

   - **Analyzing cancellation reasons:** If possible, collect more granular data on *why* trips are cancelled (e.g., driver couldn't find the location, passenger no longer needed the ride, estimated time of arrival too long).
   - **Optimizing driver-passenger matching:** Improve the algorithm to match passengers with the closest available drivers to reduce estimated arrival times in the city, which could be a reason for passenger cancellations.
   - **Driver incentives for timely pickups:** Encourage drivers to accept and complete rides in the city during peak morning hours through incentives.
   - **Passenger education/policies:** Communicate estimated arrival times clearly to passengers and consider implementing policies or fees for frequent passenger cancellations.

3. **Balance Supply and Demand Throughout the Day:** The overall demand vs. supply chart highlights periods of significant imbalance. Beyond the peak hours at specific locations, Uber should continuously monitor real-time demand and available supply across the city. Dynamic pricing and targeted driver incentives can be used to encourage drivers to move to areas with high demand and low supply.
4. **Analyze Driver Behavior:** While the top-performing drivers were identified, further analysis into the behavior of drivers with high cancellation rates or low trip completion rates could be beneficial. Understanding the challenges they face can inform training programs or operational adjustments.

By implementing these data-driven strategies, Uber can directly address the root causes of failed trips, improve the reliability of their service, enhance the experience for both passengers and drivers, and ultimately achieve their business objective of increased efficiency and reliability.

Conclusion

The exploratory data analysis of the Uber request data has revealed significant insights into the patterns of failed trips, which are a major impediment to Uber's service efficiency and customer satisfaction. The analysis highlighted a critical imbalance between demand and supply, particularly during peak hours. Specifically, the Airport experiences a substantial shortage of available cars during evening peak hours, leading to a high number of "No Cars Available" requests. Conversely, the City pickup point shows a higher incidence of cancellations, especially during morning peak hours, suggesting potential issues related to driver availability or passenger behavior in that area. The demand-supply gap across the day further emphasizes the need for better resource allocation.

To address these challenges and achieve the business objective of reducing failed trips and improving reliability, the suggested data-driven strategies focus on targeted interventions. These include implementing incentives and optimizing operations to increase driver supply at the Airport during evening peaks, investigating and mitigating cancellation reasons in the City during morning peaks, and utilizing dynamic pricing and predictive analytics to balance supply and demand throughout the day. Furthermore, analyzing individual driver performance can help identify areas for improvement and recognize high-performing drivers.

In conclusion, by leveraging the insights gained from this analysis and implementing the proposed strategies, Uber can significantly reduce the number of failed trips, enhance the reliability and efficiency of their service, improve both passenger and driver experience, and ultimately drive positive business growth.

Summary of Key Findings

Based on the visualizations and analysis performed, here are some key insights:

Peak Hours for Cancellations and "No Cars Available":

- The hourly breakdown of trip statuses (Chart 5 and Chart 7) clearly shows that the late evening hours (around 5 PM to 10 PM) experience a significant spike in "No Cars Available" statuses, particularly at the Airport pickup point. This indicates a major supply-demand gap during this period for airport pickups.
- Cancellations are also noticeable during certain hours, especially in the morning from the City pickup point (around 5 AM to 9 AM), as seen in Chart 7.

Locations with Persistent Car Shortages:

- The analysis of failed trips by pickup point and hour (Chart 7) highlights the Airport as a location with persistent car shortages during the evening peak hours. The high numbers of "No Cars Available" requests originating from the Airport during these times point to a consistent issue with driver availability to meet the demand at the Airport.

Periods of Greatest Unmet Demand:

- The comparison of demand and supply per hour (Chart 6) visually demonstrates the periods of greatest unmet demand. The largest gap between the total number of requests (demand) and the number of completed trips (supply) is evident during the evening peak hours (again, roughly 5 PM to 10 PM). This significant difference represents the requests that could not be fulfilled due to cancellations or no cars being available.