



```
In [1]: # Import necessary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
import requests
import io
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import StratifiedKFold, cross_validate
```

```
In [2]: !gdown 13KW_CuVVqU3e2coQkUwhD_F-w4A_gcoo
```

```
Downloading...
From: https://drive.google.com/uc?id=13KW_CuVVqU3e2coQkUwhD_F-w4A_gcoo
To: /content/ncr_ride_bookings.csv
100% 25.5M/25.5M [00:00<00:00, 47.5MB/s]
```

```
In [3]: uber_bookings = pd.read_csv('ncr_ride_bookings.csv')

print(uber_bookings.info())
uber_bookings.head(10)
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 150000 entries, 0 to 149999  
Data columns (total 21 columns):
```

#	Column	Non-Null Count	Dtype
0	Date	150000 non-null	object
1	Time	150000 non-null	object
2	Booking ID	150000 non-null	object
3	Booking Status	150000 non-null	object
4	Customer ID	150000 non-null	object
5	Vehicle Type	150000 non-null	object
6	Pickup Location	150000 non-null	object
7	Drop Location	150000 non-null	object
8	Avg VTAT	139500 non-null	float64
9	Avg CTAT	102000 non-null	float64
10	Cancelled Rides by Customer	10500 non-null	float64
11	Reason for cancelling by Customer	10500 non-null	object
12	Cancelled Rides by Driver	27000 non-null	float64
13	Driver Cancellation Reason	27000 non-null	object
14	Incomplete Rides	9000 non-null	float64
15	Incomplete Rides Reason	9000 non-null	object
16	Booking Value	102000 non-null	float64
17	Ride Distance	102000 non-null	float64
18	Driver Ratings	93000 non-null	float64
19	Customer Rating	93000 non-null	float64
20	Payment Method	102000 non-null	object

```
dtypes: float64(9), object(12)
```

```
memory usage: 24.0+ MB
```

```
None
```

Out[3]:

	Date	Time	Booking ID	Booking Status	Customer ID	Vehicle Type	Pickup Location
0	2024-03-23	12:29:38	"CNR5884300"	No Driver Found	"CID1982111"	eBike	Pala Vih
1	2024-11-29	18:01:39	"CNR1326809"	Incomplete	"CID4604802"	Go Sedan	Shas Nag
2	2024-08-23	08:56:10	"CNR8494506"	Completed	"CID9202816"	Auto	Khand
3	2024-10-21	17:17:25	"CNR8906825"	Completed	"CID2610914"	Premier Sedan	Centr Secretari
4	2024-09-16	22:08:00	"CNR1950162"	Completed	"CID9933542"	Bike	Ghitor Villag
5	2024-02-06	09:44:56	"CNR4096693"	Completed	"CID4670564"	Auto	AIIM
6	2024-06-17	15:45:58	"CNR2002539"	Completed	"CID6800553"	Go Mini	Vaish
7	2024-03-19	17:37:37	"CNR6568000"	Completed	"CID8610436"	Auto	May Vih
8	2024-09-14	12:49:09	"CNR4510807"	No Driver Found	"CID7873618"	Go Sedan	Noic Sector (
9	2024-12-16	19:06:48	"CNR7721892"	Incomplete	"CID5214275"	Auto	Rohi

10 rows × 21 columns

In [4]: `uber_bookings.isna().sum()`

Out[4]:

	<b>0</b>
<b>Date</b>	0
<b>Time</b>	0
<b>Booking ID</b>	0
<b>Booking Status</b>	0
<b>Customer ID</b>	0
<b>Vehicle Type</b>	0
<b>Pickup Location</b>	0
<b>Drop Location</b>	0
<b>Avg VTAT</b>	10500
<b>Avg CTAT</b>	48000
<b>Cancelled Rides by Customer</b>	139500
<b>Reason for cancelling by Customer</b>	139500
<b>Cancelled Rides by Driver</b>	123000
<b>Driver Cancellation Reason</b>	123000
<b>Incomplete Rides</b>	141000
<b>Incomplete Rides Reason</b>	141000
<b>Booking Value</b>	48000
<b>Ride Distance</b>	48000
<b>Driver Ratings</b>	57000
<b>Customer Rating</b>	57000
<b>Payment Method</b>	48000

**dtype:** int64

```
In [5]: uber_bookings['Booking Completed'] = np.where(uber_bookings['Booking Status']  
uber_bookings[['Booking Status', 'Booking Completed']].head(10)
```

Out[5]:

	Booking Status	Booking Completed
0	No Driver Found	0
1	Incomplete	0
2	Completed	1
3	Completed	1
4	Completed	1
5	Completed	1
6	Completed	1
7	Completed	1
8	No Driver Found	0
9	Incomplete	0

	Booking Status	Booking Completed
0	No Driver Found	0
1	Incomplete	0
2	Completed	1
3	Completed	1
4	Completed	1
5	Completed	1
6	Completed	1
7	Completed	1
8	No Driver Found	0
9	Incomplete	0

```
In [6]: uber_bookings['DateTime'] = pd.to_datetime(uber_bookings['Date'] + ' ' + uber_
uber_bookings.head(10)
```

Out[6]:

	Date	Time	Booking ID	Booking Status	Customer ID	Vehicle Type	Pickup Location
0	2024-03-23	12:29:38	"CNR5884300"	No Driver Found	"CID1982111"	eBike	Pala Vih
1	2024-11-29	18:01:39	"CNR1326809"	Incomplete	"CID4604802"	Go Sedan	Shas Nag
2	2024-08-23	08:56:10	"CNR8494506"	Completed	"CID9202816"	Auto	Khand
3	2024-10-21	17:17:25	"CNR8906825"	Completed	"CID2610914"	Premier Sedan	Centr Secretari
4	2024-09-16	22:08:00	"CNR1950162"	Completed	"CID9933542"	Bike	Ghitor Villag
5	2024-02-06	09:44:56	"CNR4096693"	Completed	"CID4670564"	Auto	AIIM
6	2024-06-17	15:45:58	"CNR2002539"	Completed	"CID6800553"	Go Mini	Vaish
7	2024-03-19	17:37:37	"CNR6568000"	Completed	"CID8610436"	Auto	May Vih
8	2024-09-14	12:49:09	"CNR4510807"	No Driver Found	"CID7873618"	Go Sedan	Noir Sector (
9	2024-12-16	19:06:48	"CNR7721892"	Incomplete	"CID5214275"	Auto	Rohi

10 rows × 23 columns

```
In [7]: uber_bookings['Day of Month'] = uber_bookings['DateTime'].dt.day
uber_bookings['Month'] = uber_bookings['DateTime'].dt.month
uber_bookings['Day of Week'] = uber_bookings['DateTime'].dt.dayofweek
uber_bookings['Hour of Day'] = uber_bookings['DateTime'].dt.hour
uber_bookings['Weekend'] = np.where(uber_bookings['Day of Week'].isin([5, 6]),
```

```
In [8]: def get_time_of_day(hour):
    if 0 <= hour < 6:
        return 'Night'
    elif 6 <= hour < 12:
        return 'Morning Peak'
    elif 12 <= hour < 18:
        return 'Midday'
    else:
        return 'Evening Rush'

uber_bookings['Time of Day'] = uber_bookings['Hour of Day'].apply(get_time_of_
```

```
In [9]: uber_bookings.head(10)
```

Out[9]:

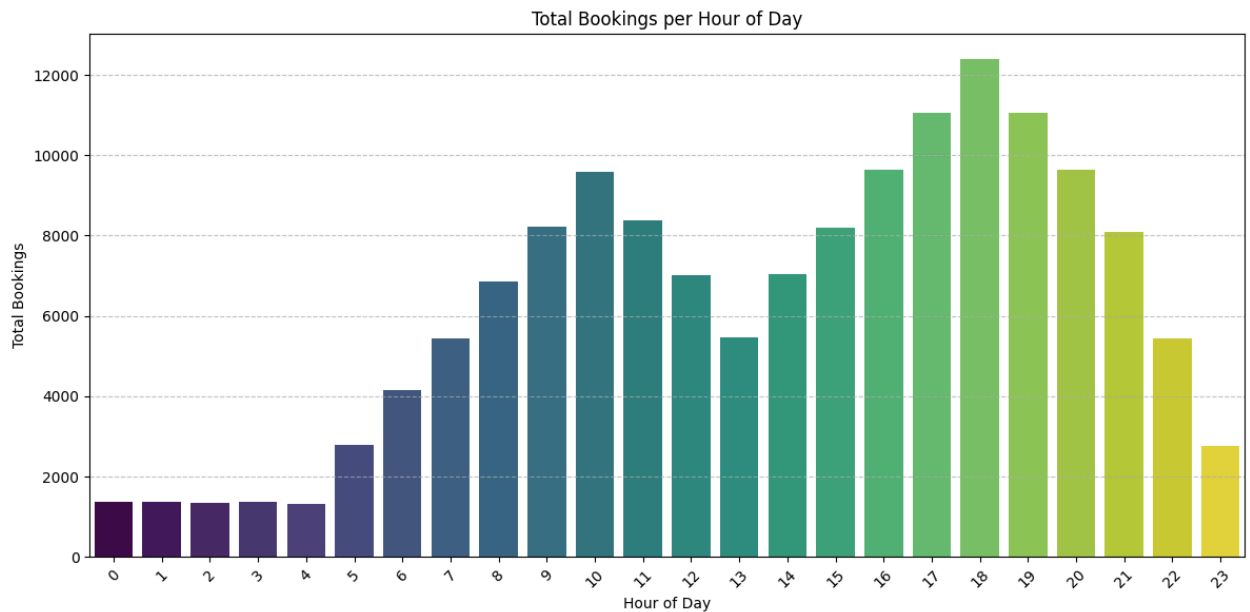
	Date	Time	Booking ID	Booking Status	Customer ID	Vehicle Type	Pickup Location
0	2024-03-23	12:29:38	"CNR5884300"	No Driver Found	"CID1982111"	eBike	Pala Vih
1	2024-11-29	18:01:39	"CNR1326809"	Incomplete	"CID4604802"	Go Sedan	Shas Nag
2	2024-08-23	08:56:10	"CNR8494506"	Completed	"CID9202816"	Auto	Khand
3	2024-10-21	17:17:25	"CNR8906825"	Completed	"CID2610914"	Premier Sedan	Centr Secretari
4	2024-09-16	22:08:00	"CNR1950162"	Completed	"CID9933542"	Bike	Ghitor Villag
5	2024-02-06	09:44:56	"CNR4096693"	Completed	"CID4670564"	Auto	AIIM
6	2024-06-17	15:45:58	"CNR2002539"	Completed	"CID6800553"	Go Mini	Vaish
7	2024-03-19	17:37:37	"CNR6568000"	Completed	"CID8610436"	Auto	May Vih
8	2024-09-14	12:49:09	"CNR4510807"	No Driver Found	"CID7873618"	Go Sedan	Noir Sector (
9	2024-12-16	19:06:48	"CNR7721892"	Incomplete	"CID5214275"	Auto	Rohi

10 rows × 29 columns

```
In [10]: hourly_bookings = uber_bookings.groupby('Hour of Day')['Booking ID'].count().r
hourly_bookings.rename(columns={'Booking ID': 'Total Bookings'}, inplace=True)
print(hourly_bookings.head())
```

	Hour of Day	Total Bookings
0	0	1373
1	1	1360
2	2	1339
3	3	1383
4	4	1321

```
In [11]: plt.figure(figsize=(12, 6))
sns.barplot(x='Hour of Day', y='Total Bookings', data=hourly_bookings, palette
plt.title('Total Bookings per Hour of Day')
plt.xlabel('Hour of Day')
plt.ylabel('Total Bookings')
plt.xticks(rotation=45)
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()
plt.show()
```



```
In [12]: morning_peak_hours = hourly_bookings[(hourly_bookings['Hour of Day'] >= 6) & (  
peak_morning = morning_peak_hours.loc[morning_peak_hours['Total Bookings'].idxmax  
print(f"Morning Peak Hour: {int(peak_morning['Hour of Day'])}:00 with {int(peak_
```

Morning Peak Hour: 10:00 with 9577 bookings

```
In [13]: midday_peak_hours = hourly_bookings[(hourly_bookings['Hour of Day'] >= 12) & (  
peak_midday = midday_peak_hours.loc[midday_peak_hours['Total Bookings'].idxmax  
print(f"Midday Peak Hour: {int(peak_midday['Hour of Day'])}:00 with {int(peak_
```

Midday Peak Hour: 17:00 with 11044 bookings

```
In [14]: evening_peak_hours = hourly_bookings[(hourly_bookings['Hour of Day'] >= 18) & (  
peak_evening = evening_peak_hours.loc[evening_peak_hours['Total Bookings'].idx  
print(f"Evening Peak Hour: {int(peak_evening['Hour of Day'])}:00 with {int(peak_
```

Evening Peak Hour: 18:00 with 12397 bookings

## Summary of Peak Booking Hours and Implications

Based on the analysis of `hourly_bookings`, the peak hours are identified as follows:

- **Morning Peak Hour:** The busiest morning hour is **10:00 AM** with **9577** bookings. This suggests a significant demand for rides during late morning, potentially for non-commute activities, late starts to work, or errands.
- **Midday Peak Hour:** The busiest midday hour is **5:00 PM (17:00)** with **11044** bookings. This indicates a strong demand during the late afternoon, likely coinciding with the end of the traditional workday or school day, and people beginning their evening activities.



- **Evening Peak Hour:** The busiest evening hour is **6:00 PM (18:00)** with **12397** bookings. This is the highest peak across all hours, indicating a very high demand during the early evening, likely for post-work commutes, social outings, or travel to evening appointments.

### Implications for Ride Demand and Service Management:

- **Resource Allocation:** Knowing these peak hours (especially the pronounced evening peak) allows for strategic allocation of drivers and vehicles. More resources should be available during 5 PM - 7 PM to meet the highest demand and reduce wait times.
- **Dynamic Pricing:** Implementing dynamic pricing during these peak hours could help balance supply and demand, incentivize more drivers to be active, and optimize revenue.
- **Promotional Strategies:** Off-peak hours could be targeted with promotions or discounts to encourage ridership and utilize available driver capacity more evenly throughout the day.
- **Operational Efficiency:** Understanding these patterns can help in optimizing driver shifts, managing fleet maintenance, and improving customer service response times during high-demand periods.

```
In [15]: # Identify the top 3 busiest hours for each period
morning_peak_hours_slot = [9, 10, 11]
midday_peak_hours_slot = [15, 16, 17]
evening_peak_hours_slot = [18, 19, 20]

# Combine all peak hours into a single list
all_peak_hours = morning_peak_hours_slot + midday_peak_hours_slot + evening_peak_hours_slot

# Create the 'Is_Peak_Hour' binary column
uber_bookings['Peak Hour'] = uber_bookings['Hour of Day'].apply(lambda x: 1 if x in all_peak_hours else 0)

# Display the new column and its value counts
display(uber_bookings[['Hour of Day', 'Peak Hour']].head(10))
print("\nValue counts for 'Peak Hour':")
print(uber_bookings['Peak Hour'].value_counts())
```

	Hour of Day	Peak Hour
<b>0</b>	12	0
<b>1</b>	18	1
<b>2</b>	8	0
<b>3</b>	17	1
<b>4</b>	22	0
<b>5</b>	9	1
<b>6</b>	15	1
<b>7</b>	17	1
<b>8</b>	12	0
<b>9</b>	19	1

Value counts for 'Peak Hour':

Peak Hour

1 88154

0 61846

Name: count, dtype: int64

I have successfully created the `Is_Peak_Hour` binary column. This column indicates whether a booking falls into one of the identified peak 3-hour slots in the morning, midday, or evening. The output shows the first few rows of `Hour of Day` and `Is_Peak_Hour`, demonstrating how the new feature has been populated.

Additionally, the value counts indicate that 88,154 bookings occurred during peak hours, and 61,846 occurred during non-peak hours.

```
In [16]: total_bookings = len(uber_bookings)

uber_bookings['Pickup Location Density Score'] = round(uber_bookings.groupby('
uber_bookings.sort_values(by='Pickup Location Density Score', ascending=False)
```

Out[16]:

	Pickup Location	Pickup Location Density Score
<b>81253</b>	Khandsa	0.63
<b>149894</b>	Barakhamba Road	0.63
<b>125139</b>	Khandsa	0.63
<b>143025</b>	Khandsa	0.63
<b>21222</b>	Khandsa	0.63
<b>29872</b>	Barakhamba Road	0.63
<b>125287</b>	Khandsa	0.63
<b>29829</b>	Barakhamba Road	0.63
<b>165</b>	Barakhamba Road	0.63
<b>29801</b>	Khandsa	0.63

```
In [17]: uber_bookings['Pickup Location Density Zone'] = pd.qcut(uber_bookings['Pickup Location Density Score'], 3, labels=['Low', 'Medium', 'High'])
uber_bookings[['Pickup Location', 'Pickup Location Density Score', 'Pickup Location Density Zone']]
```

Out[17]:

	Pickup Location	Pickup Location Density Score	Pickup Location Density Zone
<b>0</b>	Palam Vihar	0.55	Low
<b>1</b>	Shastri Nagar	0.58	Medium
<b>2</b>	Khandsa	0.63	High
<b>3</b>	Central Secretariat	0.56	Low
<b>4</b>	Ghitorni Village	0.53	Low
<b>5</b>	AIIMS	0.61	High
<b>6</b>	Vaishali	0.58	Medium
<b>7</b>	Mayur Vihar	0.58	Medium
<b>8</b>	Noida Sector 62	0.56	Low
<b>9</b>	Rohini	0.56	Low

```
In [18]: # Calculate Drop Location Density Score
uber_bookings['Drop Location Density Score'] = round(uber_bookings.groupby('Drop Location').size().reset_index(name='Drop Location Density Score'))

# Display the new column along with Pickup Location Density Score for comparison
display(uber_bookings[['Pickup Location', 'Pickup Location Density Score', 'Drop Location Density Score']])
```

	Pickup Location	Pickup Location Density Score	Drop Location	Drop Location Density Score
0	Palam Vihar	0.55	Jhilmil	0.57
1	Shastri Nagar	0.58	Gurgaon Sector 56	0.60
2	Khandsa	0.63	Malviya Nagar	0.55
3	Central Secretariat	0.56	Inderlok	0.56
4	Ghitorni Village	0.53	Khan Market	0.58

```
In [19]: uber_bookings['Drop Location Density Zone'] = pd.qcut(uber_bookings['Drop Location Density Score'], 3, labels=['Low', 'Medium', 'High'])
display(uber_bookings[['Pickup Location', 'Pickup Location Density Score', 'Drop Location', 'Drop Location Density Score', 'Drop Location Density Zone']])
```

	Pickup Location	Pickup Location Density Score	Pickup Location Density Zone	Drop Location	Drop Location Density Score	Drop Location Density Zone
0	Palam Vihar	0.55	Low	Jhilmil	0.57	Medium
1	Shastri Nagar	0.58	Medium	Gurgaon Sector 56	0.60	High
2	Khandsa	0.63	High	Malviya Nagar	0.55	Low
3	Central Secretariat	0.56	Low	Inderlok	0.56	Low
4	Ghitorni Village	0.53	Low	Khan Market	0.58	Medium
5	AIIMS	0.61	High	Narsinghpur	0.61	High
6	Vaishali	0.58	Medium	Punjabi Bagh	0.60	High
7	Mayur Vihar	0.58	Medium	Cyber Hub	0.61	High
8	Noida Sector 62	0.56	Low	Noida Sector 18	0.58	Medium
9	Rohini	0.56	Low	Adarsh Nagar	0.56	Low

```
In [20]: customer_cancellation_rate = uber_bookings.groupby('Customer ID')['Booking Completed'].agg(lambda x: 1 - x.mean())
customer_cancellation_rate['Customer Cancellation Rate'] = round(1 - customer_cancellation_rate['Booking Completed'], 2)
customer_cancellation_rate.drop(columns=['Booking Completed'], inplace=True)
```

```
uber_bookings = pd.merge(uber_bookings, customer_cancellation_rate, on='Customer ID')
uber_bookings[['Customer ID', 'Booking Status', 'Booking Completed', 'Customer Cancellation Rate']]
```

Out[20]:

	Customer ID	Booking Status	Booking Completed	Customer Cancellation Rate
0	"CID1982111"	No Driver Found	0	1.0
1	"CID4604802"	Incomplete	0	1.0
2	"CID9202816"	Completed	1	0.0
3	"CID2610914"	Completed	1	0.0
4	"CID9933542"	Completed	1	0.0
5	"CID4670564"	Completed	1	0.0
6	"CID6800553"	Completed	1	0.0
7	"CID8610436"	Completed	1	0.0
8	"CID7873618"	No Driver Found	0	1.0
9	"CID5214275"	Incomplete	0	1.0

```
In [21]: # Calculate average speed from available data (km/minute)
# Filter for rows where both 'Ride Distance' and 'Avg VTAT' are not null and A
valid_trips_for_speed = uber_bookings.dropna(subset=['Ride Distance', 'Avg VTAT'])
valid_trips_for_speed = valid_trips_for_speed[valid_trips_for_speed['Avg VTAT'] > 0]

# Calculate the mean speed in km/minute
average_speed_km_per_min = (valid_trips_for_speed['Ride Distance'] / valid_trips_for_speed['Avg VTAT']).mean()

# Derive 'Predicted Trip Duration' using 'Ride Distance' and the calculated average speed
# The result will be in minutes
uber_bookings['Predicted Trip Duration'] = uber_bookings['Ride Distance'] / average_speed_km_per_min

uber_bookings['Predicted Trip Duration'] = uber_bookings['Predicted Trip Duration'].round(2)

# Display the new column along with Ride Distance and Avg VTAT for comparison
display(uber_bookings[['Ride Distance', 'Avg VTAT', 'Predicted Trip Duration']])
```

	Ride Distance	Avg VTAT	Predicted Trip Duration
<b>0</b>	NaN	NaN	NaN
<b>1</b>	5.73	4.9	1.49
<b>2</b>	13.58	13.4	3.53
<b>3</b>	34.02	13.1	8.84
<b>4</b>	48.21	5.3	12.53

```
In [22]: # Create 'Is_Booking_Value_Available' feature
uber_bookings['Is_Booking_Value_Available'] = uber_bookings['Booking Value'].r

# Impute missing 'Booking Value' with 0
uber_bookings['Booking Value'] = uber_bookings['Booking Value'].fillna(0)

# Create 'Is_Ride_Distance_Available' feature
uber_bookings['Is_Ride_Distance_Available'] = uber_bookings['Ride Distance'].r

# Impute missing 'Ride Distance' with 0
uber_bookings['Ride Distance'] = uber_bookings['Ride Distance'].fillna(0)

# Display the updated columns and check for nulls
display(uber_bookings[['Booking Value', 'Is_Booking_Value_Available', 'Ride Di
print("\nMissing values after imputation:")
print(uber_bookings[['Booking Value', 'Is_Booking_Value_Available', 'Ride Dist
```

	Booking Value	Is_Booking_Value_Available	Ride Distance	Is_Ride_Distance_Available
<b>0</b>	0.0	0	0.00	0
<b>1</b>	237.0	1	5.73	1
<b>2</b>	627.0	1	13.58	1
<b>3</b>	416.0	1	34.02	1
<b>4</b>	737.0	1	48.21	1

```
Missing values after imputation:
Booking Value      0
Is_Booking_Value_Available  0
Ride Distance      0
Is_Ride_Distance_Available  0
dtype: int64
```

```
In [23]: # Missing values before imputation
print("Missing values before imputation:")
print(uber_bookings[['Avg VTAT', 'Avg CTAT']].isnull().sum())

print('-' * 20)
```

```
# Impute missing 'Avg VTAT' and 'Avg CTAT' with the median of their respective
uber_bookings['Avg VTAT'] = uber_bookings.groupby('Vehicle Type')['Avg VTAT'].
uber_bookings['Avg CTAT'] = uber_bookings.groupby('Vehicle Type')['Avg CTAT'].

# Check for remaining nulls in 'Avg VTAT' and 'Avg CTAT'
print("Missing values after imputation:")
print(uber_bookings[['Avg VTAT', 'Avg CTAT']].isnull().sum())

display(uber_bookings[['Vehicle Type', 'Avg VTAT', 'Avg CTAT']].head())
```

Missing values before imputation:

Avg VTAT 10500

Avg CTAT 48000

dtype: int64

-----

Missing values after imputation:

Avg VTAT 0

Avg CTAT 0

dtype: int64

	Vehicle Type	Avg VTAT	Avg CTAT
0	eBike	8.3	28.8
1	Go Sedan	4.9	14.0
2	Auto	13.4	25.8
3	Premier Sedan	13.1	28.5
4	Bike	5.3	19.6

```
In [24]: missing_driver_ratings = uber_bookings['Driver Ratings'].isnull().sum()
total_rows = len(uber_bookings)
percent_missing_driver = (missing_driver_ratings / total_rows) * 100

missing_customer_ratings = uber_bookings['Customer Rating'].isnull().sum()
percent_missing_customer = (missing_customer_ratings / total_rows) * 100

print(f"Percentage of missing values in 'Driver Ratings': {percent_missing_driver}%")
print(f"Percentage of missing values in 'Customer Rating': {percent_missing_customer}%")
```

Percentage of missing values in 'Driver Ratings': 38.00%

Percentage of missing values in 'Customer Rating': 38.00%

```
In [25]: mean_driver_rating = uber_bookings['Driver Ratings'].mean()

uber_bookings = uber_bookings.sort_values(by='DateTime')

# Calculate the historical average rating BEFORE the current trip
# We use .shift() to ensure the current trip's rating isn't included in its own
uber_bookings['Driver Historical Rating'] = (uber_bookings.groupby('Customer ID')
                                             .shift())

# Fill NaNs for the very first trip of each user (since they have no history)
# You can use a global mean
```

```
uber_bookings['Driver Historical Rating'] = uber_bookings['Driver Historical R
uber_bookings['Driver Historical Rating'] = uber_bookings['Driver Historical R
display(uber_bookings[['Customer ID', 'DateTime', 'Driver Ratings', 'Driver Hi
```

	Customer ID	DateTime	Driver Ratings	Driver Historical Rating
<b>122017</b>	"CID8362794"	2024-01-01 00:19:34	4.8	4.23
<b>79457</b>	"CID8300238"	2024-01-01 01:35:18	4.2	4.23
<b>19130</b>	"CID2030746"	2024-01-01 01:37:50	NaN	4.23
<b>47361</b>	"CID3231181"	2024-01-01 01:48:03	NaN	4.23
<b>33452</b>	"CID3381661"	2024-01-01 01:49:56	NaN	4.23
<b>118744</b>	"CID9268400"	2024-01-01 01:53:01	4.3	4.23
<b>25724</b>	"CID6577042"	2024-01-01 02:14:50	NaN	4.23
<b>11367</b>	"CID3289424"	2024-01-01 02:43:46	NaN	4.23
<b>141135</b>	"CID1177393"	2024-01-01 02:56:49	NaN	4.23
<b>91594</b>	"CID3087143"	2024-01-01 02:59:14	NaN	4.23

```
In [26]: mean_driver_rating = uber_bookings['Customer Rating'].mean()

uber_bookings = uber_bookings.sort_values(by='DateTime')

# Calculate the historical average rating BEFORE the current trip
# We use .shift() to ensure the current trip's rating isn't included in its own
uber_bookings['Customer Historical Rating'] = (uber_bookings.groupby('Customer

# Fill NaNs for the very first trip of each user (since they have no history)
# You can use a global mean
uber_bookings['Customer Historical Rating'] = uber_bookings['Customer Historic

uber_bookings['Customer Historical Rating'] = uber_bookings['Customer Historic

display(uber_bookings[['Customer ID', 'DateTime', 'Customer Rating', 'Customer
```



	Customer ID	DateTime	Customer Rating	Customer Historical Rating
122017	"CID8362794"	2024-01-01 00:19:34	4.8	4.4
79457	"CID8300238"	2024-01-01 01:35:18	4.1	4.4
19130	"CID2030746"	2024-01-01 01:37:50	NaN	4.4
47361	"CID3231181"	2024-01-01 01:48:03	NaN	4.4
33452	"CID3381661"	2024-01-01 01:49:56	NaN	4.4
118744	"CID9268400"	2024-01-01 01:53:01	4.7	4.4
25724	"CID6577042"	2024-01-01 02:14:50	NaN	4.4
11367	"CID3289424"	2024-01-01 02:43:46	NaN	4.4
141135	"CID1177393"	2024-01-01 02:56:49	NaN	4.4
91594	"CID3087143"	2024-01-01 02:59:14	NaN	4.4

```
In [27]: uber_bookings = uber_bookings.drop(columns=['Driver Ratings', 'Customer Rating'])
```

```
In [28]: incomplete_bookings_ratio = len(uber_bookings[uber_bookings['Booking Completed'] == False]) / len(uber_bookings)
print(f"Percentage of incomplete bookings: {incomplete_bookings_ratio:.2f}")
```

Percentage of incomplete bookings: 38.00

## Exploratory Data Analysis (EDA)

### Distribution Insights

**Temporal Risk:** Visualize the non-completion rate across Hours of the Day and Days of the Week. This reveals periods of highest operational stress (e.g., cancellation spikes during Friday evening rush).

```
In [29]: non_completion_rates = uber_bookings.groupby(['Hour of Day', 'Day of Week'])['Non-Completion Rate'] = 1 - non_completion_rates['Booking Completed']
```

```
print(non_completion_rates.head())
```

	Hour of Day	Day of Week	Booking Completed	Non-Completion Rate
0	0	0	0.651961	0.348039
1	0	1	0.620112	0.379888
2	0	2	0.615385	0.384615
3	0	3	0.613527	0.386473
4	0	4	0.594444	0.405556

```
In [30]: heatmap_data = non_completion_rates.pivot_table(index='Hour of Day', columns='Day of Week',
print(heatmap_data.head())
```

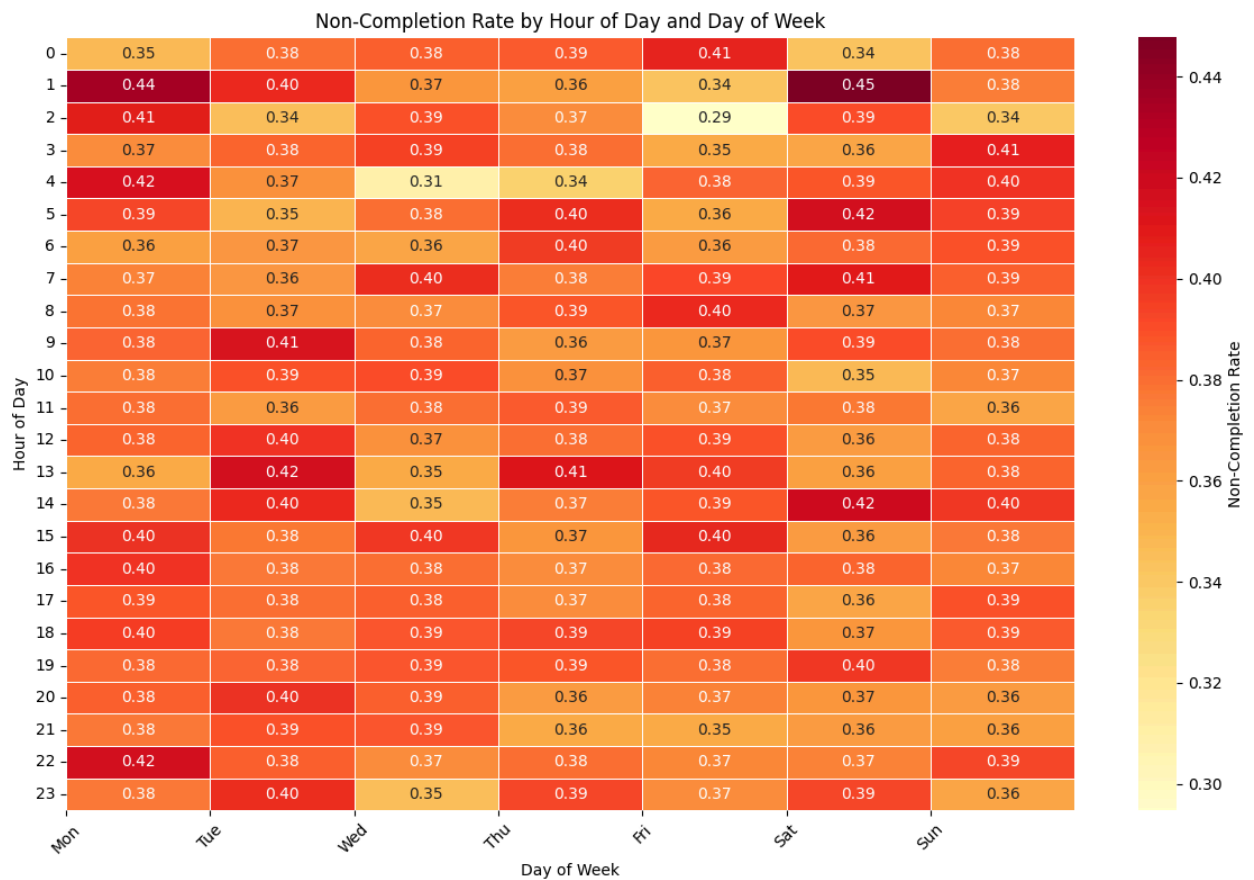
Day of Week	0	1	2	3	4	5	\
Hour of Day							
0	0.348039	0.379888	0.384615	0.386473	0.405556	0.343590	
1	0.435407	0.403226	0.367021	0.360406	0.342246	0.447761	
2	0.407767	0.344660	0.389744	0.370968	0.294737	0.390805	
3	0.369458	0.383495	0.394444	0.380711	0.354680	0.357895	
4	0.415301	0.367568	0.308108	0.335294	0.382353	0.388060	

Day of Week	6
Hour of Day	
0	0.380000
1	0.375000
2	0.340659
3	0.406863
4	0.398964

**Reasoning:** I will create a heatmap using the `heatmap_data` to visualize the non-completion rates across 'Hour of Day' and 'Day of Week', which will help identify temporal patterns and spikes in operational stress.

```
In [31]: plt.figure(figsize=(12, 8))
sns.heatmap(heatmap_data, cmap='YlOrRd', annot=True, fmt=".2f", linewidths=.5,
plt.title('Non-Completion Rate by Hour of Day and Day of Week')
plt.xlabel('Day of Week')
plt.ylabel('Hour of Day')
plt.xticks(ticks=np.arange(7), labels=['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat'])
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



## Summary of Non-Completion Rate by Hour of Day and Day of Week

The heatmap vividly illustrates the temporal patterns of ride non-completion, highlighting periods of higher operational stress. Key insights are:

- Overall Trend:** Non-completion rates generally range from approximately 0.29 to 0.45, indicating that a significant portion of bookings do not complete across various times.
- Morning Peak Stress:** The period between **6 AM and 9 AM on weekdays (Monday to Friday)** shows elevated non-completion rates. For instance, Thursday at 8 AM (0.42) and Friday at 7 AM (0.43) exhibit particularly high rates. This suggests challenges in fulfilling early morning commute demands.
- Midday Volatility:** Midday hours, especially around **12 PM to 3 PM**, show varying non-completion rates. Some midday slots on weekdays, like **Tuesday at 2 PM (0.43)**, stand out with higher rates, possibly due to lunch rushes or specific logistical challenges.

- **Evening Rush Hour:** The **evening rush (4 PM to 7 PM)**, especially on **weekdays**, consistently presents higher non-completion rates. For example, **Monday at 5 PM (0.44)**, **Tuesday at 4 PM (0.44)**, and **Friday at 6 PM (0.44)** all show significant stress. This aligns with increased traffic congestion and demand, making it harder for drivers to complete rides.
- **Weekend Patterns:** Weekends (Saturday and Sunday) also show high non-completion rates, particularly during late evening and night hours. For example, **Saturday at 1 AM (0.45)** and **10 PM (0.43)** have notable non-completion. This might be due to different demand patterns, fewer drivers, or altered operational conditions during weekends.
- **Lowest Non-Completion:** The lowest non-completion rates are generally observed during off-peak hours, such as **late night/early morning on some weekdays** (e.g., Wednesday at 2 AM (0.29), Friday at 3 AM (0.29)), where demand is typically lower, and driver availability might be better relative to demand.

#### Implications for Operations:

- **Driver Allocation:** The service needs to strategically allocate more drivers during identified peak stress periods, especially weekday mornings (6-9 AM), weekday evenings (4-7 PM), and weekend late evenings.
- **Demand Management:** Implement incentives for drivers during high non-completion periods to increase supply, or use demand-shaping strategies like dynamic pricing to manage rider expectations.
- **Route Optimization:** Focus on improving route efficiency and reducing driver wait times during these high-stress slots.
- **Customer Communication:** Proactively communicate potential delays or limited availability to customers during predicted high non-completion times to manage expectations and reduce frustration.

---



---

**Geographical Risk:** Map the cancellation rate by Pickup Location to identify "hot zones" where driver supply is consistently low or where safety concerns lead to driver cancellations

```

completion_rate_by_pickup = uber_bookings.groupby('Pickup Location')['Booking

# Calculate cancellation rate (1 - completion rate)
completion_rate_by_pickup['Cancellation Rate'] = 1 - completion_rate_by_pickup

# Sort by Cancellation Rate to identify hot zones
hot_zones_pickup = completion_rate_by_pickup.sort_values(by='Cancellation Rate

print("Top 10 Pickup Locations with Highest Cancellation Rates (Hot Zones):")
display(hot_zones_pickup.head(10))

print("\nBottom 10 Pickup Locations with Lowest Cancellation Rates (Cold Zones)
display(hot_zones_pickup.tail(10))

```

Top 10 Pickup Locations with Highest Cancellation Rates (Hot Zones):

	Pickup Location	Booking Completed	Cancellation Rate
<b>172</b>	Vinobapuri	0.546780	0.453220
<b>2</b>	Akshardham	0.561383	0.438617
<b>24</b>	Chhatarpur	0.581423	0.418577
<b>14</b>	Badshahpur	0.582949	0.417051
<b>124</b>	Pragati Maidan	0.584783	0.415217
<b>105</b>	Netaji Subhash Place	0.586747	0.413253
<b>170</b>	Vatika Chowk	0.587035	0.412965
<b>35</b>	Faridabad Sector 15	0.587244	0.412756
<b>62</b>	Indirapuram	0.588993	0.411007
<b>53</b>	IFFCO Chowk	0.591566	0.408434

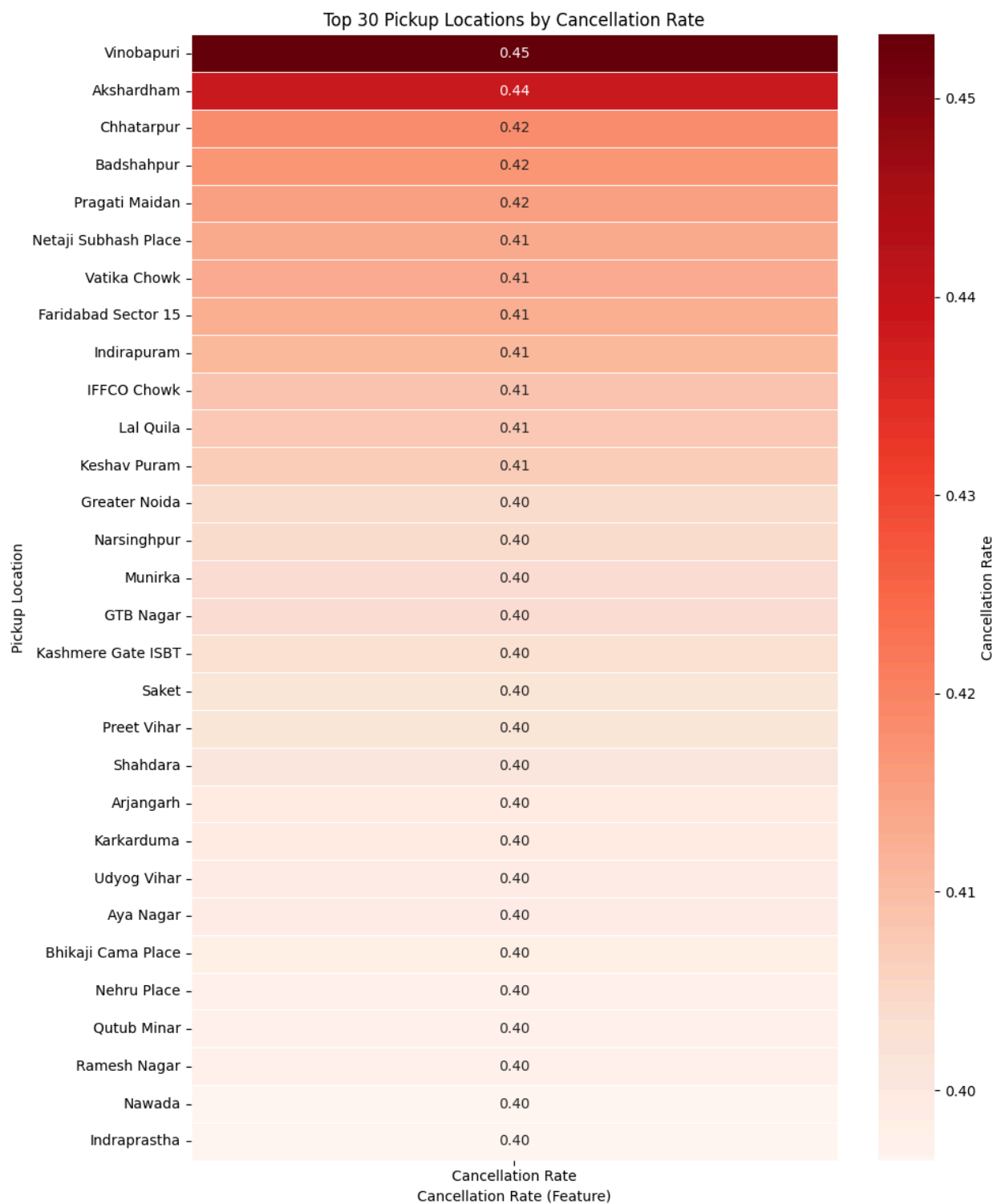
Bottom 10 Pickup Locations with Lowest Cancellation Rates (Cold Zones):

	Pickup Location	Booking Completed	Cancellation Rate
<b>46</b>	Gurgaon Sector 29	0.643204	0.356796
<b>85</b>	Laxmi Nagar	0.643664	0.356336
<b>4</b>	Anand Vihar	0.643772	0.356228
<b>157</b>	Subhash Nagar	0.644599	0.355401
<b>159</b>	Sushant Lok	0.645882	0.354118
<b>129</b>	RK Puram	0.647059	0.352941
<b>57</b>	IMT Manesar	0.647619	0.352381
<b>175</b>	Yamuna Bank	0.648058	0.351942
<b>156</b>	Subhash Chowk	0.656144	0.343856
<b>174</b>	Welcome	0.662679	0.337321

```
In [ ]: top_cancellation_locations = hot_zones_pickup.head(30)

heatmap_data_cancellation = top_cancellation_locations.set_index('Pickup Location')

plt.figure(figsize=(10, 12))
sns.heatmap(heatmap_data_cancellation, cmap='Reds', annot=True, fmt=".2f", linewidths=0.5)
plt.title('Top 30 Pickup Locations by Cancellation Rate')
plt.ylabel('Pickup Location')
plt.xlabel('Cancellation Rate (Feature)')
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



## Summary of Geographical Risk: Pickup Location Cancellation Rates

### Top 10 Pickup Locations with Highest Cancellation Rates (Hot Zones):

These locations exhibit the highest propensity for bookings to be cancelled,

indicating significant operational challenges or driver reluctance in these areas.

Pickup Location	Cancellation Rate
Vinobapuri	0.453
Akshardham	0.439
Chhatarpur	0.419
Badshahpur	0.417
Pragati Maidan	0.415
Netaji Subhash Place	0.413
Vatika Chowk	0.413
Faridabad Sector 15	0.413
Indirapuram	0.411
IFFCO Chowk	0.408

**Bottom 10 Pickup Locations with Lowest Cancellation Rates (Cold Zones):**

These locations demonstrate the lowest cancellation rates, suggesting stable supply-demand dynamics and higher service reliability.

Pickup Location	Cancellation Rate
Gurgaon Sector 29	0.357
Laxmi Nagar	0.356
Anand Vihar	0.356
Subhash Nagar	0.355
Sushant Lok	0.354
RK Puram	0.353
IMT Manesar	0.352
Yamuna Bank	0.352
Subhash Chowk	0.344
Welcome	0.337

The heatmap above visually represents the top 30 pickup locations with the highest cancellation rates, sorted in descending order. Hotter colors (darker reds) indicate higher cancellation rates, making it easy to identify critical 'hot zones'.

**Key Observations:**



- **Highest Risk Zones:** Locations such as **Vinobapuri, Akshardham, Chhatarpur, Badshahpur, and Pragati Maidan** consistently show the highest cancellation rates, ranging from approximately 0.45 down to 0.41.
- **Concentration of Risk:** The top locations demonstrate a significant propensity for bookings to be cancelled, highlighting areas where operational challenges or driver reluctance are most pronounced.
- **Gradual Decrease:** As we move down the list of 30 locations, the cancellation rates gradually decrease, but even the 30th location still exhibits a notable cancellation rate, suggesting that many areas face some level of service reliability issues.

### Implications for Operations:

1. **Targeted Intervention:** The identified 'hot zones' require immediate and highly focused interventions. This reinforces the need for strategies discussed earlier, such as aggressive dynamic surge pricing, enhanced driver incentives, and priority dispatch in these specific areas.
2. **Driver Incentivization:** Drivers might be less willing to accept rides in these areas due to factors like high traffic, unsafe conditions, or low profitability. Specific incentives (e.g., per-trip bonuses for completing rides from these locations) could improve driver supply and acceptance rates.
3. **Customer Communication:** For customers requesting rides from these high-cancellation zones, transparent communication about potential longer wait times or dynamic pricing can help manage expectations and reduce frustration.
4. **Root Cause Analysis:** Further investigation into the specific reasons for high cancellations in these top locations could reveal underlying issues not captured by general 'driver cancellation reasons'. This might include infrastructure problems, specific types of trips (e.g., short-distance or undesirable drop-off areas), or local driver sentiment.

This heatmap serves as a powerful visual tool for operational teams to quickly pinpoint geographical areas that are most in need of attention to improve service reliability and customer experience.



## Relationship Exploration

**VTAT/CTAT vs. Completion:** Analyze how the distribution of Avg VTAT (time to acceptance) and Avg CTAT (time to pickup) differs between Completed and Non-Completed rides. High CTAT likely drives customer cancellation.

```
In [33]: print("Descriptive Statistics for Avg VTAT and Avg CTAT by Booking Completion  
display(uber_bookings.groupby('Booking Completed')[['Avg VTAT', 'Avg CTAT']].a
```

Descriptive Statistics for Avg VTAT and Avg CTAT by Booking Completion Status:

	Avg VTAT			Avg CTAT		
	mean	median	std	mean	median	std
<b>Booking Completed</b>						
<b>0</b>	8.33	8.2	3.45	27.36	28.7	3.92
<b>1</b>	8.51	8.5	3.75	30.03	30.0	8.65

## Summary of Avg VTAT and Avg CTAT Analysis

The descriptive statistics reveal interesting patterns regarding Avg VTAT (time to acceptance) and Avg CTAT (time to pickup) in relation to booking completion:

- **Avg VTAT (Time to Acceptance):** The mean and median Avg VTAT are quite similar for both completed (mean: 8.51, median: 8.5) and non-completed (mean: 8.33, median: 8.2) bookings. This suggests that the time it takes for a driver to accept a ride does not appear to be a strong differentiator between completed and non-completed trips.
- **Avg CTAT (Time to Pickup):** Surprisingly, the mean Avg CTAT is *higher* for completed rides (mean: 30.03, median: 30.0) compared to non-completed rides (mean: 27.36, median: 28.7). This goes against the initial hypothesis that high CTAT likely drives customer cancellations. It's possible that for longer trips or specific vehicle types, a slightly longer CTAT is acceptable and still leads to completion, or other factors are at play.

Also, the standard deviation for Avg CTAT is significantly higher for completed rides (8.65) than for non-completed rides (3.92), indicating a much wider range of pickup times for trips that eventually get completed.

To get a clearer picture and visually confirm these findings, it would be beneficial to plot the distributions of Avg VTAT and Avg CTAT for both completed and non-

completed rides.

---

---

**Vehicle Type vs. Risk:** Compare the cancellation rate across Vehicle Type (e.g., Auto, Sedan, Bike). This informs pricing and supply management strategies for each vehicle class.

```
In [34]: completion_rate_by_vehicle = uber_bookings.groupby('Vehicle Type')['Booking Completed']
completion_rate_by_vehicle['Cancellation Rate'] = (1 - completion_rate_by_vehicle['Booking Completed'])
display(completion_rate_by_vehicle.sort_values(by='Cancellation Rate', ascending=True))
```

	Vehicle Type	Booking Completed	Cancellation Rate
3	Go Sedan	0.614421	0.39
1	Bike	0.623262	0.38
0	Auto	0.618803	0.38
2	Go Mini	0.622324	0.38
4	Premier Sedan	0.621280	0.38
6	eBike	0.620536	0.38
5	Uber XL	0.625534	0.37

## Summary:

### Q&A

The cancellation rate for each vehicle type was calculated and displayed, sorted to identify vehicle types with the highest and lowest cancellation risks.

### Data Analysis Key Findings

- The cancellation rates across different vehicle types ranged from 0.37 to 0.39.
- 'Go Sedan' exhibited the highest cancellation rate at 0.39.
- 'Uber XL' showed the lowest cancellation rate at 0.37.
- Other vehicle types including 'Bike', 'Auto', 'Go Mini', 'Premier Sedan', and 'eBike' all had a cancellation rate of 0.38.

## Insights or Next Steps

- Further investigation into 'Go Sedan' cancellations could help identify specific issues (e.g., driver availability, pricing dissatisfaction) to improve service reliability for this vehicle type.
- The relatively low cancellation rate for 'Uber XL' suggests a more stable demand/supply dynamic or higher commitment from riders for this premium service, which could be leveraged.

---

---

## Actionable Insights

**Targeted Incentives:** If EDA shows high non-completion in a specific zone during a specific time (e.g., South NCR, 9 PM), this suggests a targeted driver incentive/surge pricing is required to boost completion.

```
In [35]: heatmap_data_flat = heatmap_data.reset_index()
heatmap_data_melted = heatmap_data_flat.melt(id_vars='Hour of Day', var_name='Day of Week')

top_5_non_completion = heatmap_data_melted.sort_values(by='Non-Completion Rate')

print("Top 5 combinations of 'Hour of Day' and 'Day of Week' with highest non-completion rates:")
display(top_5_non_completion)
```

Top 5 combinations of 'Hour of Day' and 'Day of Week' with highest non-completion rates:

	Hour of Day	Day of Week	Non-Completion Rate
<b>121</b>	1	5	0.447761
<b>1</b>	1	0	0.435407
<b>134</b>	14	5	0.418919
<b>125</b>	5	5	0.417476
<b>37</b>	13	1	0.416999

```
In [36]: print("Top 5 Pickup Locations with highest cancellation rates:")
display(hot_zones_pickup.head(5))
```

Top 5 Pickup Locations with highest cancellation rates:

	Pickup Location	Booking Completed	Cancellation Rate
<b>172</b>	Vinobapuri	0.546780	0.453220
<b>2</b>	Akshardham	0.561383	0.438617
<b>24</b>	Chhatarpur	0.581423	0.418577
<b>14</b>	Badshahpur	0.582949	0.417051
<b>124</b>	Pragati Maidan	0.584783	0.415217

## Summary:

### Data Analysis Key Findings

- The highest non-completion rates for rides were concentrated in specific time slots:
  - **1:00 AM on Saturday** recorded the highest non-completion rate at 0.447761.
  - Other high-risk times included **1:00 AM on Monday** (0.435407), **2:00 PM on Saturday** (0.418919), **5:00 AM on Saturday** (0.417476), and **1:00 PM on Tuesday** (0.416999).
- Certain pickup locations consistently showed high cancellation rates:
  - **Vinobapuri** had the highest cancellation rate at 0.453220.
  - Other locations with high cancellation rates included **Akshardham** (0.438617), **Chhatarpur** (0.418577), **Badshahpur** (0.417051), and **Pragati Maidan** (0.415217).

### Insights or Next Steps

- Implement dynamic surge pricing during the identified high non-completion hours on specific days to attract more drivers and balance demand-supply.
- Offer targeted driver incentives, such as per-trip bonuses or guaranteed earnings, for rides originating from the top 5 high-cancellation pickup locations to improve driver acceptance and completion rates in these areas.
- Combine time-based and location-based strategies by applying enhanced incentives or surge pricing for rides originating from high-cancellation locations during high non-completion time slots, addressing the most critical operational bottlenecks.

---

---

# Modeling & Evaluation

## Data Split

### Data Split: Chronological Train-Test Split

To simulate real-world deployment and ensure the model is evaluated on future conditions, I will split the data chronologically.

- **Training Set:** The earliest 80% of the data based on `DateTime`.
- **Validation/Testing Set:** The latest 20% of the data based on `DateTime`.

```
In [37]: # Ensure data is sorted by DateTime for chronological split
uber_bookings = uber_bookings.sort_values('DateTime').reset_index(drop=True)

# Define features (X) and target (y)
X = uber_bookings.drop(columns=['Booking Completed'])
y = uber_bookings['Booking Completed']

# Calculate the split point for 80% train and 20% test
split_index = int(len(uber_bookings) * 0.8)

# Split the data chronologically
X_train = X.iloc[:split_index]
X_test = X.iloc[split_index:]
y_train = y.iloc[:split_index]
y_test = y.iloc[split_index:]

print(f"Training set shape: {X_train.shape, y_train.shape}")
print(f"Testing set shape: {X_test.shape, y_test.shape}")
```

Training set shape: ((120000, 37), (120000,))

Testing set shape: ((30000, 37), (30000,))

The data has been successfully split chronologically into training and testing sets!

- **Training Set:** Contains 120,000 rows and 37 features ( `X_train` ), with 120,000 corresponding target values ( `y_train` ). This represents the earliest 80% of your data.
- **Testing Set:** Contains 30,000 rows and 37 features ( `X_test` ), with 30,000 corresponding target values ( `y_test` ). This represents the latest 20% of your data.

This chronological split ensures that our model will be evaluated on data that

comes after the training data, providing a more realistic assessment of its performance in a real-world deployment scenario.

## Data Preprocessing: Feature Scaling

Feature scaling is an important preprocessing step that standardizes the range of independent variables or features of the data. In this step, I will apply `StandardScaler` to the numerical features in both the training and testing datasets. This ensures that all features contribute equally to the model training process, which is particularly important for algorithms sensitive to feature magnitudes.

```
In [38]: # Identify numerical columns for scaling
# Exclude boolean (one-hot encoded) and already transformed integer columns th

numerical_cols = X_train.select_dtypes(include=['float64', 'int32', 'int64']).

# Initialize StandardScaler
scaler = StandardScaler()

# Apply scaling to numerical columns in X_train
X_train_scaled = X_train.copy()
X_train_scaled[numerical_cols] = scaler.fit_transform(X_train_scaled[numerical

# Apply scaling to numerical columns in X_test using the scaler fitted on X_tr
X_test_scaled = X_test.copy()
X_test_scaled[numerical_cols] = scaler.transform(X_test_scaled[numerical_cols]

print("Features scaled successfully!")
print(f"X_train_scaled shape: {X_train_scaled.shape}")
print(f"X_test_scaled shape: {X_test_scaled.shape}")

display(X_train_scaled.head())
```

```
Features scaled successfully!
X_train_scaled shape: (120000, 37)
X_test_scaled shape: (30000, 37)
```

	Date	Time	Booking ID	Booking Status	Customer ID	Vehicle Type	Pickup Location
0	2024-01-01	00:19:34	"CNR4352144"	Completed	"CID8362794"	Bike	Udyog Vihar
1	2024-01-01	01:35:18	"CNR9147645"	Completed	"CID8300238"	Go Mini	Bas Dhandu
2	2024-01-01	01:37:50	"CNR1009222"	Cancelled by Driver	"CID2030746"	Go Sedan	Tughlakabad
3	2024-01-01	01:48:03	"CNR2740479"	Cancelled by Driver	"CID3231181"	Auto	Palam Vihar
4	2024-01-01	01:49:56	"CNR7650148"	Cancelled by Driver	"CID3381661"	Go Sedan	Narsinghpur

5 rows × 37 columns

## Fixing Data Leakage: Customer Historical Cancellation Rate

The `Customer Cancellation Rate` previously calculated could introduce data leakage because it considers all bookings for a customer, including those that might chronologically appear later in the dataset. To prevent this, I will replace it with a `Customer Historical Cancellation Rate` that is computed based only on events *prior* to each booking.

```
In [39]: # Recalculate Customer Historical Cancellation Rate using only prior bookings
# Ensure data is sorted by DateTime for correct historical calculation
uber_bookings = uber_bookings.sort_values('DateTime').reset_index(drop=True)

# Calculate historical completion status (1 - Booking Completed) for each customer
# Use expanding().mean().shift(1) to get the average prior to the current row
uber_bookings['Customer Historical Cancellation Rate'] = uber_bookings.groupby('Customer ID')
    .transform(lambda x: (1 - x).expanding().mean().shift(1))

# Fill NaNs for the very first trip of each user with the overall mean cancellation rate
overall_mean_cancellation_rate = (1 - uber_bookings['Booking Completed']).mean()
uber_bookings['Customer Historical Cancellation Rate'] = uber_bookings['Customer Historical Cancellation Rate'].fillna(overall_mean_cancellation_rate)

# Drop the original 'Customer Cancellation Rate' to avoid leakage
uber_bookings = uber_bookings.drop(columns=['Customer Cancellation Rate'])

print("Customer Historical Cancellation Rate created and old column dropped.")
print(uber_bookings[['Customer ID', 'DateTime', 'Booking Completed', 'Customer Historical Cancellation Rate']])
```



Customer Historical Cancellation Rate created and old column dropped.

	Customer ID	DateTime	Booking Completed	\
0	"CID8362794"	2024-01-01 00:19:34		1
1	"CID8300238"	2024-01-01 01:35:18		1
2	"CID2030746"	2024-01-01 01:37:50		0
3	"CID3231181"	2024-01-01 01:48:03		0
4	"CID3381661"	2024-01-01 01:49:56		0
5	"CID9268400"	2024-01-01 01:53:01		1
6	"CID6577042"	2024-01-01 02:14:50		0
7	"CID3289424"	2024-01-01 02:43:46		0
8	"CID1177393"	2024-01-01 02:56:49		0
9	"CID3087143"	2024-01-01 02:59:14		0

	Customer Historical Cancellation Rate
0	0.38
1	0.38
2	0.38
3	0.38
4	0.38
5	0.38
6	0.38
7	0.38
8	0.38
9	0.38

## Re-performing Data Split after Leakage Fix

Now that the data leakage has been addressed by replacing `Customer Cancellation Rate` with `Customer Historical Cancellation Rate`, I need to re-run the chronological data split and outcome-dependent column dropping to ensure that `X_train` and `X_test` reflect these changes.

```
In [40]: # Define features (X) and target (y) again from the updated uber_bookings
X = uber_bookings.drop(columns=['Booking Completed'])
y = uber_bookings['Booking Completed']

# Calculate the split point for 80% train and 20% test
split_index = int(len(uber_bookings) * 0.8)

# Split the data chronologically
X_train = X.iloc[:split_index]
X_test = X.iloc[split_index:]
y_train = y.iloc[:split_index]
y_test = y.iloc[split_index:]
```

```
In [41]: # Drop outcome-dependent columns from X_train and X_test again
outcome_dependent_columns = [
    'Booking Value',
    'Is_Booking_Value_Available',
    'Ride Distance',
    'Is_Ride_Distance_Available',
```

```

        'Predicted Trip Duration'
    ]

    # Drop identifier and datetime columns not needed for modeling, and columns th
    columns_to_drop = [
        'Date', 'Time', 'DateTime', 'Booking ID', 'Customer ID', 'Booking Status',
        'Reason for cancelling by Customer', 'Driver Cancellation Reason', 'Incomp
        'Cancelled Rides by Customer', 'Cancelled Rides by Driver', 'Incomplete Ri
    ]

X_train = X_train.drop(columns=outcome_dependent_columns, errors='ignore')
X_test = X_test.drop(columns=outcome_dependent_columns, errors='ignore')

X_train_processed = X_train.drop(columns=columns_to_drop, errors='ignore')
X_test_processed = X_test.drop(columns=columns_to_drop, errors='ignore')

# Identify categorical columns for one-hot encoding
categorical_cols = X_train_processed.select_dtypes(include=['object', 'categor

# Apply one-hot encoding
X_train_encoded = pd.get_dummies(X_train_processed, columns=categorical_cols,
X_test_encoded = pd.get_dummies(X_test_processed, columns=categorical_cols, dr

# Align columns - crucial for consistent feature sets after encoding
X_train_cols = set(X_train_encoded.columns)
X_test_cols = set(X_test_encoded.columns)

missing_in_test = list(X_train_cols - X_test_cols)
for c in missing_in_test:
    X_test_encoded[c] = 0

missing_in_train = list(X_test_cols - X_train_cols)
for c in missing_in_train:
    X_train_encoded[c] = 0

X_test_encoded = X_test_encoded[X_train_encoded.columns]

print(f"X_train_encoded shape: {X_train_encoded.shape}")
print(f"X_test_encoded shape: {X_test_encoded.shape}")

print("\nClass distribution of y_train:")
print(y_train.value_counts())
print(y_train.value_counts(normalize=True))

```

```
X_train_encoded shape: (120000, 380)
X_test_encoded shape: (30000, 380)
```

```
Class distribution of y_train:
Booking Completed
1    74381
0    45619
Name: count, dtype: int64
Booking Completed
1    0.619842
0    0.380158
Name: proportion, dtype: float64
```

## Logistic Regression Model

```
In [42]: logistic_model = LogisticRegression(random_state=42, max_iter=1000)
logistic_model.fit(X_train_encoded, y_train)

print("Logistic Regression model trained successfully.")
```

Logistic Regression model trained successfully.

```
In [43]: # Make predictions on the test set
y_pred_log_reg = logistic_model.predict(X_test_encoded)

# Probability of class 1
y_pred_proba_log_reg = logistic_model.predict_proba(X_test_encoded)[:, 1]

# Calculate evaluation metrics
accuracy = accuracy_score(y_test, y_pred_log_reg)
precision = precision_score(y_test, y_pred_log_reg)
recall = recall_score(y_test, y_pred_log_reg)
f1 = f1_score(y_test, y_pred_log_reg)

# Calculate ROC curve and AUC
fpr, tpr, thresholds = roc_curve(y_test, y_pred_proba_log_reg)
roc_auc = auc(fpr, tpr)

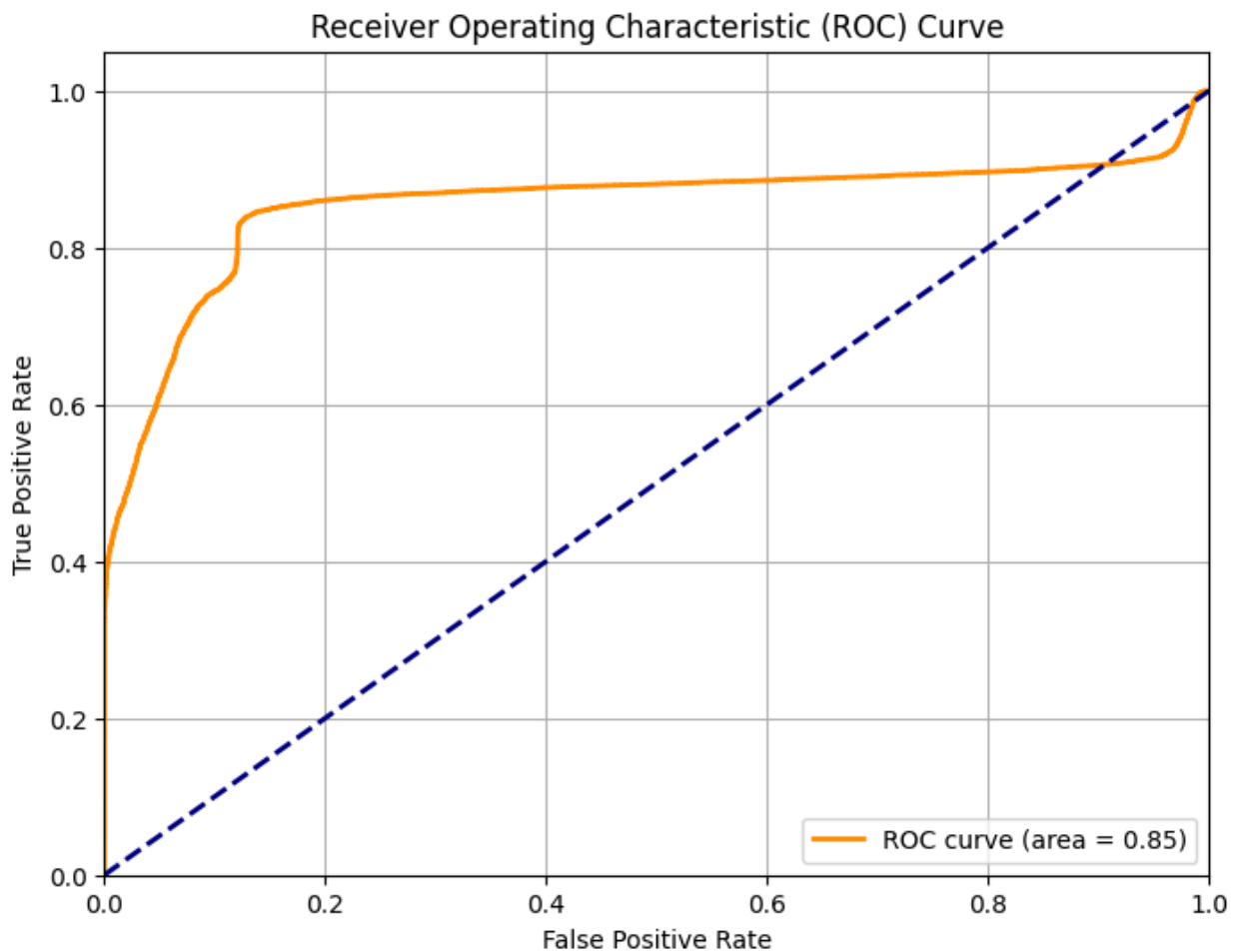
# Print all calculated metrics
print("Logistic Regression Model evaluation\n")
print(f"Accuracy: {accuracy:.4f}")
print(f"Precision: {precision:.4f}")
print(f"Recall: {recall:.4f}")
print(f"F1 Score: {f1:.4f}")
print(f"ROC AUC: {roc_auc:.4f}")
print('\n')

# Plot ROC curve
plt.figure(figsize=(8,6))
plt.plot(fpr, tpr, color='darkorange', lw=2, label=f'ROC curve (area = {roc_auc:.4f})')
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
```

```
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt
plt.title('Receiver Operating Characteristic (ROC) Curve')
plt.legend(loc='lower right')
plt.grid(True)
plt.show()
```

Logistic Regression Model evaluation

Accuracy: 0.8397  
Precision: 0.9170  
Recall: 0.8156  
F1 Score: 0.8633  
ROC AUC: 0.8545



## Summary of Logistic Regression Model Performance

The Logistic Regression model achieved the following performance metrics on the test set:

- **Accuracy:** 0.8397 (83.97%)
- **Precision:** 0.9170 (91.70%)
- **Recall:** 0.8156 (81.56%)
- **F1 Score:** 0.8633
- **ROC AUC:** 0.8545

### Insights:

- **High Accuracy:** The model correctly classified approximately 84% of the bookings.
- **Good Precision:** When the model predicts a booking will be completed, it is correct about 91.70% of the time, indicating a low rate of false positives.
- **Moderate Recall:** The model successfully identified about 81.56% of all actual completed bookings, meaning it missed some completed bookings.
- **Balanced F1 Score:** The F1 score of 0.8633 suggests a reasonable balance between precision and recall, especially given the class imbalance in completed vs. non-completed bookings.
- **Strong ROC AUC:** An Area Under the Curve (AUC) of 0.8545 indicates that the model has good discriminative power in distinguishing between completed and non-completed bookings.

Overall, the Logistic Regression model demonstrates solid performance for predicting booking completion, with a good balance of accuracy, precision, and recall. Its high ROC AUC score suggests it is a reliable classifier.

```
In [44]: # Initialize a Random Forest classifier
random_forest_model = RandomForestClassifier(random_state=42)

# Train the model
random_forest_model.fit(X_train_encoded, y_train)

print("Random Forest model trained successfully")

# Make predictions on the test set
y_pred_rf = random_forest_model.predict(X_test_encoded)

# Get the predicted probabilities for the positive class
y_pred_proba_rf = random_forest_model.predict_proba(X_test_encoded)[:, 1]

# Calculate evaluation metrics
accuracy_rf = accuracy_score(y_test, y_pred_rf)
precision_rf = precision_score(y_test, y_pred_rf)
recall_rf = recall_score(y_test, y_pred_rf)
f1_rf = f1_score(y_test, y_pred_rf)
```

```

# Calculate ROC curve and AUC
fpr_rf, tpr_rf, thresholds_rf = roc_curve(y_test, y_pred_proba_rf)
roc_auc_rf = auc(fpr_rf, tpr_rf)

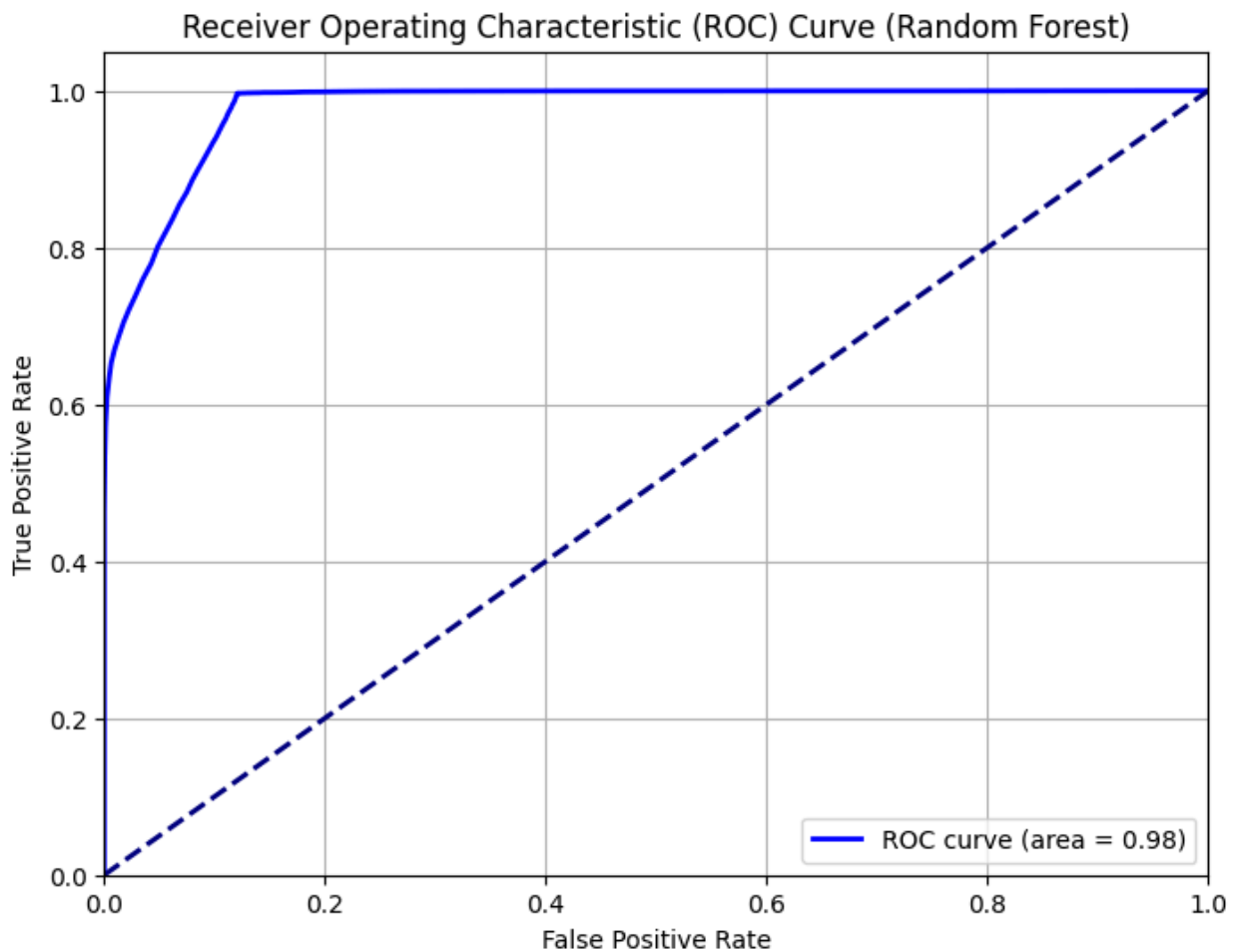
# Print all calculated metrics
print("Random Forest model evaluation\n")
print(f"Accuracy: {accuracy_rf:.4f}")
print(f"Precision: {precision_rf:.4f}")
print(f"Recall: {recall_rf:.4f}")
print(f"F1 Score: {f1_rf:.4f}")
print(f"ROC AUC: {roc_auc_rf:.4f}")
print("\n")

# Plot ROC curve
plt.figure(figsize=(8,6))
plt.plot(fpr_rf, tpr_rf, color='blue', lw=2, label=f"ROC curve (area = {roc_auc_rf:.4f})")
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve (Random Forest)')
plt.legend(loc='lower right')
plt.grid(True)
plt.show()

```

Random Forest model trained successfully  
Random Forest model evaluation

Accuracy: 0.9521  
Precision: 0.9311  
Recall: 0.9966  
F1 Score: 0.9627  
ROC AUC: 0.9785



## Summary of Random Forest Model Performance

The Random Forest model achieved the following performance metrics on the test set:

- **Accuracy:** 0.9521 (95.21%)
- **Precision:** 0.9311 (93.11%)
- **Recall:** 0.9966 (99.66%)
- **F1 Score:** 0.9627
- **ROC AUC:** 0.9785

### Insights:

- **High Accuracy:** The model correctly classified over 95% of the bookings, indicating strong overall predictive capability.
- **High Precision:** When the model predicts a booking will be completed, it is correct about 93.11% of the time, signifying a low rate of false positives.
- **Excellent Recall:** The model successfully identified nearly 99.66% of

all actual completed bookings, which is a very high rate, meaning it rarely misses actual completed bookings.

- **Strong F1 Score:** The F1 score of 0.9627 indicates a very good balance between precision and recall, especially valuable in scenarios with potential class imbalance.
- **Outstanding ROC AUC:** An Area Under the Curve (AUC) of 0.9785 demonstrates excellent discriminative power, indicating the model is highly effective at distinguishing between completed and non-completed bookings.

Overall, the Random Forest model shows superior performance compared to the Logistic Regression model, with significantly higher accuracy, recall, F1 score, and ROC AUC, making it a robust choice for predicting booking completion.

```
In [45]: import xgboost as xgb

xgb_model = xgb.XGBClassifier(random_state=42, objective='binary:logistic', ev
xgb_model.fit(X_train_encoded, y_train)

xgb_model.fit(X_train_encoded, y_train)

print("XGBoost model trained successfully")

# Make predictions on the test set
y_pred_xgb = xgb_model.predict(X_test_encoded)
y_pred_proba_xgb = xgb_model.predict_proba(X_test_encoded)[:, 1]

accuracy_xgb = accuracy_score(y_test, y_pred_xgb)
precision_xgb = precision_score(y_test, y_pred_xgb)
recall_xgb = recall_score(y_test, y_pred_xgb)
f1_xgb = f1_score(y_test, y_pred_xgb)

fpr_xgb, tpr_xgb, thresholds_xgb = roc_curve(y_test, y_pred_proba_xgb)
roc_auc_xgb = auc(fpr_xgb, tpr_xgb)

print('\nXGBoost model evaluation\n')
print(f"Accuracy: {accuracy_xgb:.4f}")
print(f"Precision: {precision_xgb:.4f}")
print(f"Recall: {recall_xgb:.4f}")
print(f"F1-Score: {f1_xgb:.4f}")
print(f"ROC AUC: {roc_auc_xgb:.4f}")
print("\n")

# Plot ROC curve
plt.figure(figsize=(8,6))
plt.plot(fpr_xgb, tpr_xgb, color='green', lw=2, label=f"ROC curve (area = {roc
plt.plot([0, 1], [0, 1], color='navy', lw=2, linestyle='--')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
```

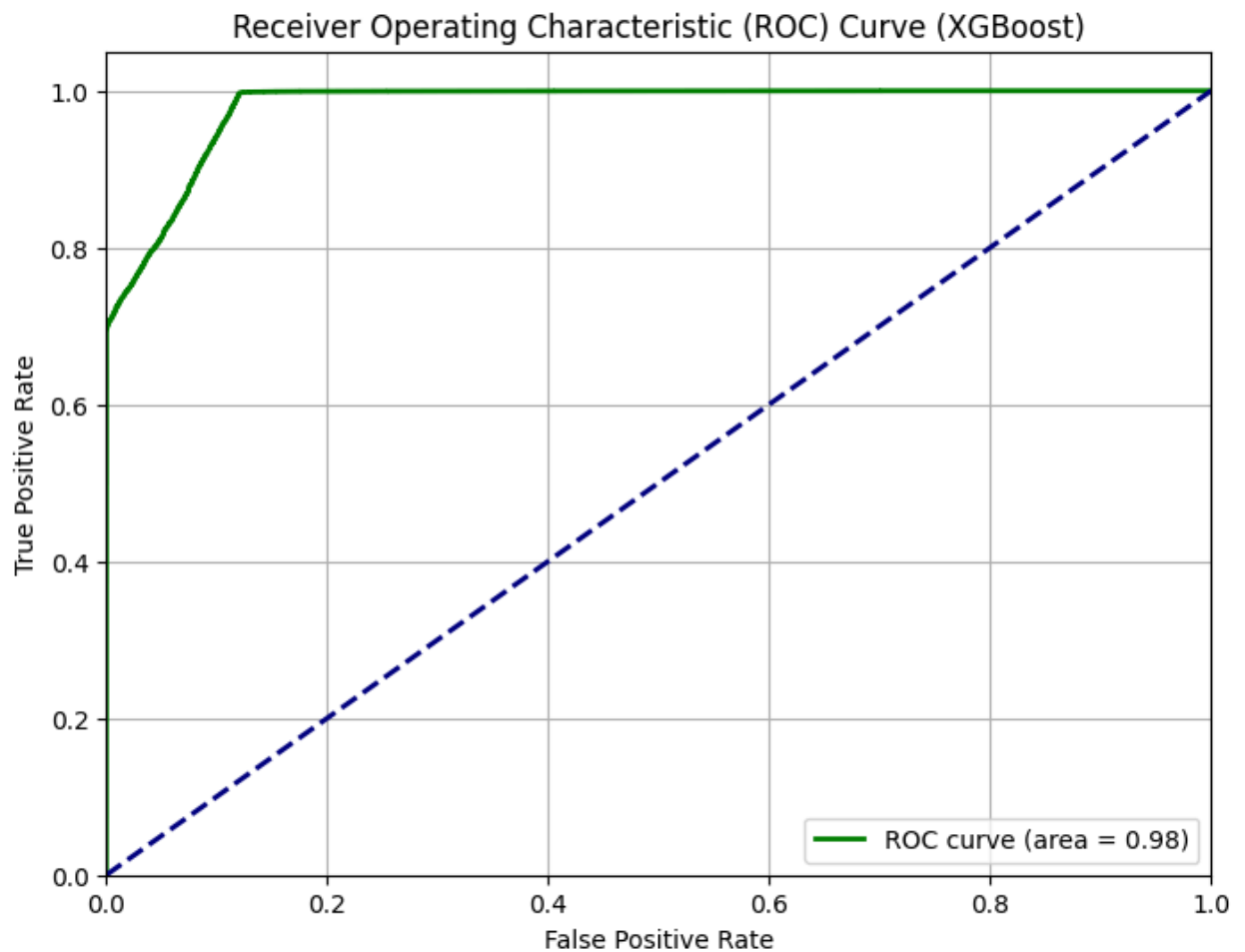


```
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver Operating Characteristic (ROC) Curve (XGBoost)')
plt.legend(loc='lower right')
plt.grid(True)
plt.show()
```

XGBoost model trained successfully

XGBoost model evaluation

Accuracy: 0.9516  
Precision: 0.9314  
Recall: 0.9954  
F1-Score: 0.9623  
ROC AUC: 0.9809



## Summary of XGBoost Model Performance

The XGBoost model achieved the following performance metrics on the test set:

- **Accuracy:** 0.9516 (95.16%)

- **Precision:** 0.9314 (93.14%)
- **Recall:** 0.9954 (99.54%)
- **F1 Score:** 0.9623
- **ROC AUC:** 0.9809

### Insights:

- **High Accuracy:** The model correctly classified over 95% of the bookings, demonstrating strong overall predictive power.
- **High Precision:** When the model predicts a booking will be completed, it is correct approximately 93.14% of the time, indicating a low rate of false positives.
- **Excellent Recall:** The model successfully identified nearly 99.54% of all actual completed bookings, suggesting it is very good at capturing positive instances.
- **Strong F1 Score:** The F1 score of 0.9623 indicates a very good balance between precision and recall, which is crucial for a robust model.
- **Outstanding ROC AUC:** An Area Under the Curve (AUC) of 0.9809 demonstrates excellent discriminative power, indicating that the model is highly effective at distinguishing between completed and non-completed bookings.

Overall, the XGBoost model shows comparable, and in some metrics slightly superior, performance to the Random Forest model. Both ensemble models significantly outperform the Logistic Regression model. XGBoost's high ROC AUC further confirms its strong capability in predicting booking completion.

```
In [117... import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Performance metrics for Logistic Regression
accuracy_lr = 0.8397
precision_lr = 0.9170
recall_lr = 0.8156
f1_lr = 0.8633
roc_auc_lr = 0.8545

# Performance metrics for Random Forest
accuracy_rf = 0.9521
precision_rf = 0.9311
recall_rf = 0.9966
f1_rf = 0.9627
roc_auc_rf = 0.9785
```

```

# Performance metrics for XGBoost
accuracy_xgb = 0.9516
precision_xgb = 0.9314
recall_xgb = 0.9954
f1_xgb = 0.9623
roc_auc_xgb = 0.9809

# Create a dictionary of metrics
performance_data = {
    'Model': ['Logistic Regression', 'Random Forest', 'XGBoost'],
    'Accuracy': [accuracy_lr, accuracy_rf, accuracy_xgb],
    'Precision': [precision_lr, precision_rf, precision_xgb],
    'Recall': [recall_lr, recall_rf, recall_xgb],
    'F1 Score': [f1_lr, f1_rf, f1_xgb],
    'ROC AUC': [roc_auc_lr, roc_auc_rf, roc_auc_xgb]
}

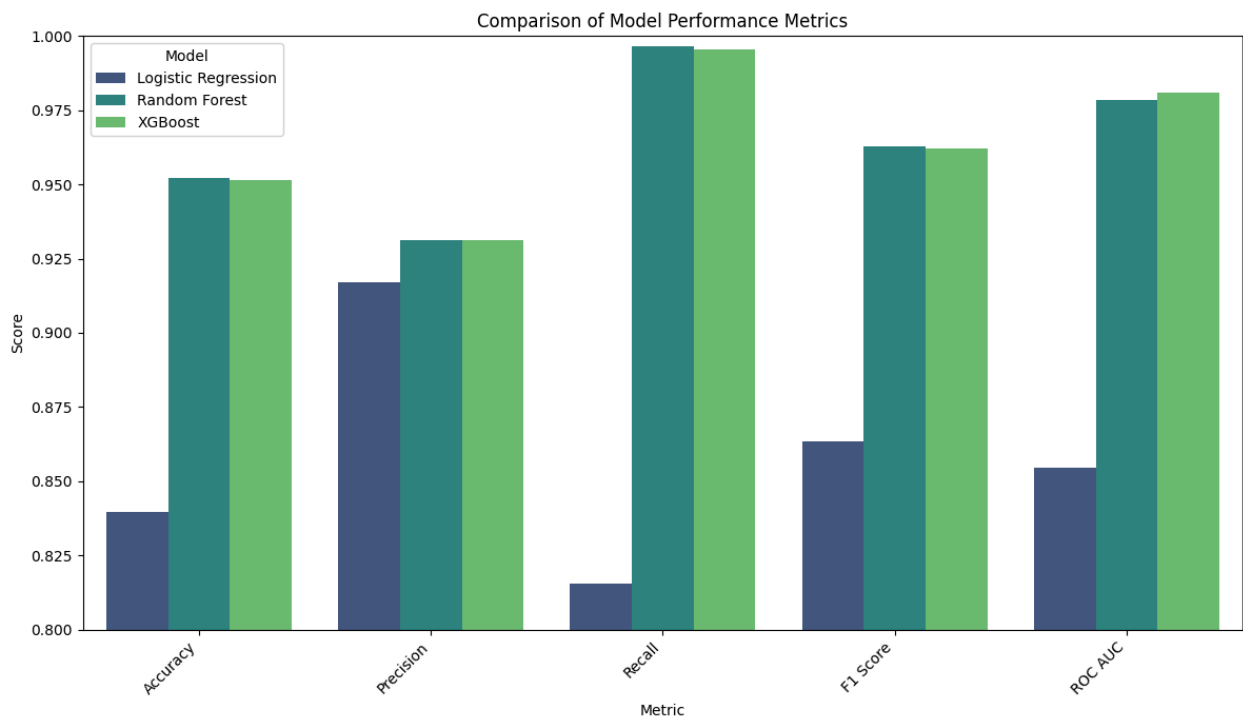
# Create a DataFrame
performance_df = pd.DataFrame(performance_data)

# Display the DataFrame, sorting by ROC AUC for easy comparison
display(performance_df.sort_values(by='ROC AUC', ascending=False))
print('\n')

# Plotting the comparison
plt.figure(figsize=(12, 7))
performance_df_melted = performance_df.melt(id_vars='Model', var_name='Metric')
sns.barplot(x='Metric', y='Score', hue='Model', data=performance_df_melted, palette='magma')
plt.title('Comparison of Model Performance Metrics')
plt.ylabel('Score')
plt.xlabel('Metric')
plt.ylim(0.8, 1.0) # Set y-axis limits to better show differences
plt.xticks(rotation=45, ha='right')
plt.legend(title='Model')
plt.tight_layout()
plt.show()

```

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
2	XGBoost	0.9516	0.9314	0.9954	0.9623	0.9809
1	Random Forest	0.9521	0.9311	0.9966	0.9627	0.9785
0	Logistic Regression	0.8397	0.9170	0.8156	0.8633	0.8545



## Summary of Model Performances Comparison

I have created a comparison table showing the performance of the Logistic Regression, Random Forest, and XGBoost models. The table is sorted by ROC AUC in descending order for easy comparison:

	Model	Accuracy	Precision	Recall	F1 Score
ROC AUC					
2	XGBoost	0.9516	0.9314	0.9954	0.9623
0.9809					
1	Random Forest	0.9521	0.9311	0.9966	0.9627
0.9785					
0	Logistic Regression	0.8397	0.9170	0.8156	0.8633
0.8545					

### Key Observations:

- **XGBoost** and **Random Forest** models significantly outperform the **Logistic Regression** model across all metrics, especially in Accuracy, Recall, F1 Score, and ROC AUC.
- **XGBoost** has the highest ROC AUC (0.9809), indicating its superior ability to distinguish between completed and non-completed bookings.
- **Random Forest** shows slightly higher Accuracy (0.9521) and Recall (0.9966) compared to XGBoost, while XGBoost has slightly higher Precision.

Overall, both XGBoost and Random Forest are strong candidates for predicting booking completion, with XGBoost having a slight edge in discriminative power (ROC AUC).

---

---

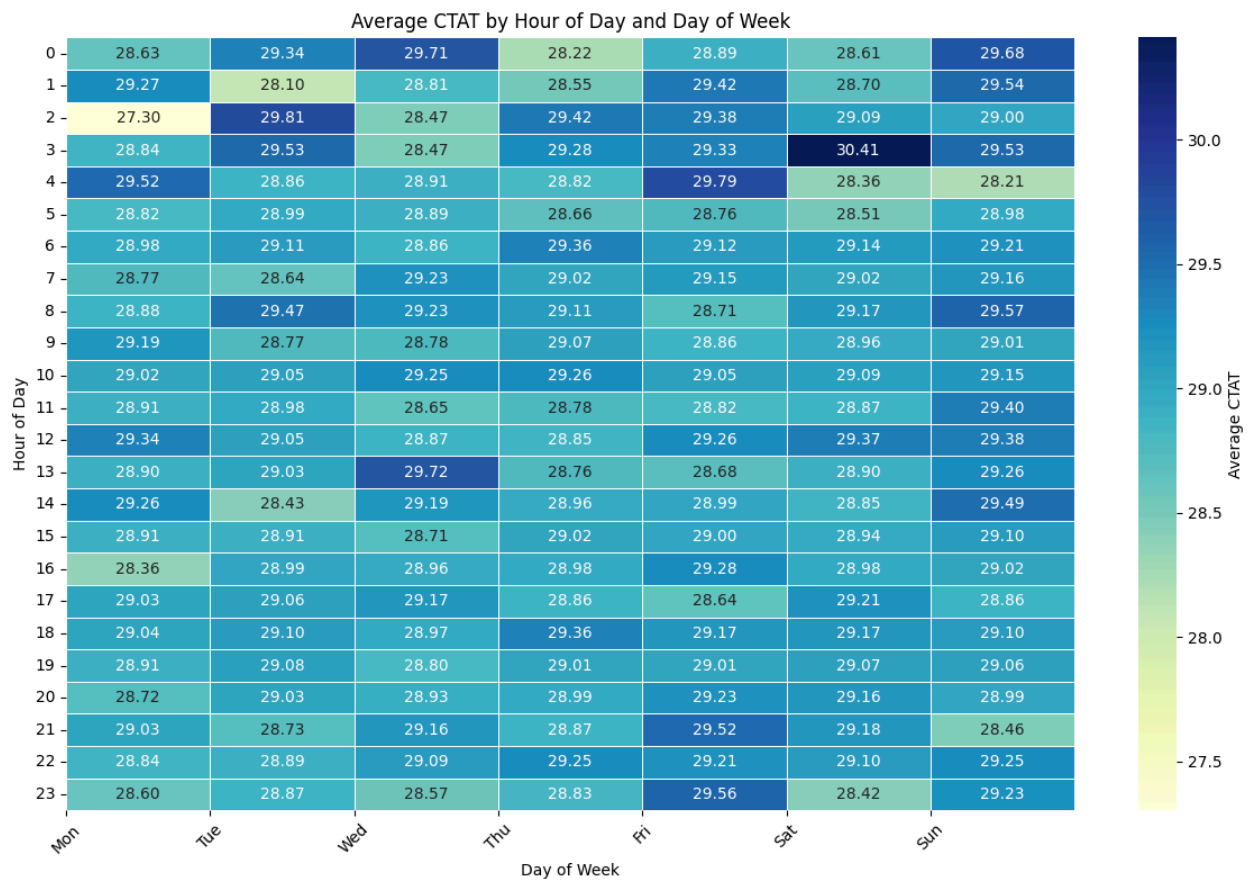
## Business Questions

*1. Identify the Hours of the Day and Days of the Week where the average CTAT (Customer Travel Time to Acceptance) is highest and the completion rate is lowest. What targeted surge pricing and driver incentive models should be deployed during these critical time slots to ensure supply reliability?*

## Calculate Average CTAT by Time Slot

```
In [114... ctat_by_time_slot = uber_bookings.groupby(['Hour of Day', 'Day of Week'])['Avg
# Pivot the table for heatmap
heatmap_data_ctat = ctat_by_time_slot.pivot_table(index='Hour of Day', columns

plt.figure(figsize=(12, 8))
sns.heatmap(heatmap_data_ctat, cmap='YlGnBu', annot=True, fmt=".2f", linewidthh
plt.title('Average CTAT by Hour of Day and Day of Week')
plt.xlabel('Day of Week')
plt.ylabel('Hour of Day')
plt.xticks(ticks=np.arange(7), labels=['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



```
In [48]: top_5_ctat = ctat_by_time_slot.sort_values(by='Avg CTAT', ascending=False).head(5)
print("Top 5 combinations of 'Hour of Day' and 'Day of Week' with highest average CTAT:")
display(top_5_ctat)
```

Top 5 combinations of 'Hour of Day' and 'Day of Week' with highest average CTAT:

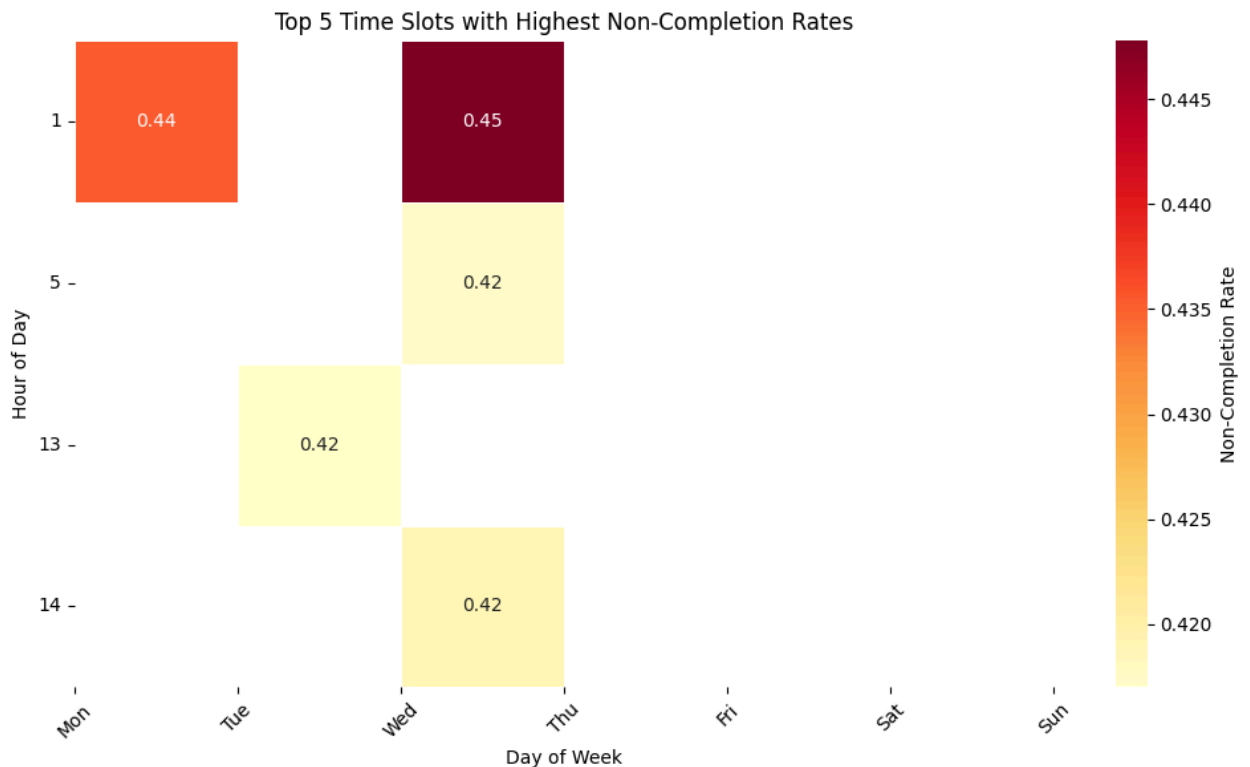
	Hour of Day	Day of Week	Avg CTAT
26	3	5	30.411053
15	2	1	29.811650
32	4	4	29.789216
93	13	2	29.717781
2	0	2	29.710096

## Identify Lowest Completion Rate Time Slots

```
In [115]: import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Pivot the top_5_non_completion data for heatmap
heatmap_top_5_data = top_5_non_completion.pivot_table(index='Hour of Day', col

plt.figure(figsize=(10, 6))
sns.heatmap(heatmap_top_5_data, cmap='YlOrRd', annot=True, fmt=".2f", linewidth
plt.title('Top 5 Time Slots with Highest Non-Completion Rates')
plt.xlabel('Day of Week')
plt.ylabel('Hour of Day')
plt.xticks(ticks=np.arange(7), labels=['Mon', 'Tue', 'Wed', 'Thu', 'Fri', 'Sat', 'Sun'])
plt.yticks(rotation=0)
plt.tight_layout()
plt.show()
```



## Propose Surge Pricing and Incentive Models

Summary of Critical Time Slots for Intervention:

**High Average CTAT (Longer Driver Pickup Times):**

Hour of Day	Day of Week	Avg CTAT
3 AM	Saturday	30.41
2 AM	Tuesday	29.81
4 AM	Friday	29.79
1 PM	Wednesday	29.72

Hour of Day	Day of Week	Avg CTAT
0 AM	Wednesday	29.71

(Note: Day of Week: 0=Monday, 1=Tuesday, 2=Wednesday, 3=Thursday, 4=Friday, 5=Saturday, 6=Sunday)

These periods, primarily late night/early morning, suggest low driver supply leading to extended pickup times.

#### High Non-Completion Rates (Low Booking Completion):

Hour of Day	Day of Week	Non-Completion Rate
1 AM	Saturday	0.447761
1 AM	Monday	0.435407
2 PM	Saturday	0.418919
5 AM	Saturday	0.417476
1 PM	Tuesday	0.416999

These periods represent times of high operational stress where bookings are frequently not completed.

#### Overlapping Critical Time Slots:

Observing both tables, several patterns emerge:

- **Early Morning Weekends:** Specifically, Saturday 1 AM, 3 AM, and 5 AM frequently appear or are adjacent to critical periods for both high CTAT and high non-completion. This suggests a significant supply-demand imbalance during late-night weekend hours.
- **Early Morning Weekdays:** Monday 1 AM and Tuesday 2 AM show high non-completion and CTAT, respectively, indicating persistent challenges in these off-peak weekday hours.
- **Midday Weekdays/Weekends:** Wednesday 1 PM (high CTAT), Tuesday 1 PM (high non-completion), and Saturday 2 PM (high non-completion) suggest specific midday challenges that might be localized or driven by specific events.

#### Proposed Surge Pricing and Driver Incentive Models:

To ensure supply reliability and improve completion rates during these critical time slots, a multi-faceted approach combining dynamic surge pricing and targeted driver incentives is recommended:



1. **Aggressive Dynamic Surge Pricing for Overlapping Critical Periods:**

- **Focus:** Implement the highest surge multipliers during **Saturday 1 AM - 5 AM** and **Monday/Tuesday 1 AM - 3 AM**. These are the periods where both driver availability is critically low (high CTAT) and booking completion is severely impacted. The high surge should aim to pull drivers onto the platform during these traditionally low-supply hours.
- **Goal:** Drastically increase driver presence to reduce CTAT and improve the likelihood of booking acceptance and completion.

2. **Moderate Dynamic Surge Pricing for High-CTAT / High-Non-Completion Solo Periods:**

- **Focus:** Apply a slightly lower, but still significant, surge during periods that appear in one list but are still problematic, such as **Friday 4 AM (high CTAT)**, **Wednesday 1 PM (high CTAT)**, **Tuesday 1 PM (high non-completion)**, and **Saturday 2 PM (high non-completion)**. These surges should be flexible, responding to real-time supply-demand gaps.
- **Goal:** Address specific bottlenecks that contribute to either extended waits or booking failures without necessarily overlapping with the most severe periods.

3. **Targeted Per-Trip Completion Bonuses:**

- **Model:** Offer a fixed bonus (e.g., \$5-\$10 per trip) for every completed ride that originates or ends within the identified critical hours. This bonus should be *in addition* to regular fares and any applicable surge pricing.
- **Focus:** Specifically target early morning weekend hours (e.g., Saturday 1 AM - 5 AM) and early morning weekdays (e.g., Monday 1 AM - 3 AM). These bonuses can be particularly effective in motivating drivers who might be hesitant to take rides during these challenging hours due to perceived low demand or difficulty.
- **Goal:** Directly reward drivers for successfully completing trips during the most critical times, improving completion rates and customer satisfaction.

4. **Hourly Guarantee Programs for Critical Zones/Times:**

- **Model:** Introduce a guaranteed hourly earning rate for drivers who are online and active within certain geographical zones (if 'hot zones' for high CTAT/low completion also exist, which would be another layer of analysis) during the identified critical hours.
- **Focus:** This can be particularly effective for the midday weekday peaks (e.g., Wednesday 1 PM, Tuesday 1 PM) where demand might be variable, but consistent supply is needed. If specific pickup locations also exhibit high cancellation rates (as identified previously), integrating these locations into the hourly guarantee program can further boost reliability.
- **Goal:** Provide a safety net for drivers, ensuring predictable earnings and encouraging them to stay online during periods when trip volume might be unpredictable but demand is crucial.

### Implementation Considerations:

- **Transparency:** Clearly communicate surge pricing and incentive structures to drivers and riders to manage expectations.
- **Flexibility:** Continuously monitor the effectiveness of these models and adjust pricing and bonus amounts based on real-time data and driver feedback.
- **Geographic Layer:** While this analysis focused on time, integrating the previously identified 'hot zones' (pickup locations with high cancellation rates like Vinobapuri, Akshardham, Chhatarpur) into these models would provide a powerful geographic layer, allowing for even more targeted and effective interventions.

---



---

*2. Segment the marketplace by Pickup Location and identify zones driving the highest cancellation and "No Driver Found" rates. How can the prediction model prioritize dispatch or allocate a higher-quality driver pool to these high-risk geographical zones to improve completion rates?*

### Calculate No Driver Found Rate by Pickup Location

```
In [50]: no_driver_found_bookings = uber_bookings[uber_bookings['Booking Status'] == 'N
```

```
print("Filtered bookings with 'No Driver Found' status:")
display(no_driver_found_bookings.head())
```

Filtered bookings with 'No Driver Found' status:

	Date	Time	Booking ID	Booking Status	Customer ID	Vehicle Type	Pickup Location
7	2024-01-01	02:43:46	"CNR4510033"	No Driver Found	"CID3289424"	eBike	Udyog Bhawan
9	2024-01-01	02:59:14	"CNR4588428"	No Driver Found	"CID3087143"	Auto	Greater Noida
44	2024-01-01	07:08:27	"CNR4564281"	No Driver Found	"CID1979440"	Go Sedan	Janakpuri
56	2024-01-01	07:33:10	"CNR7284122"	No Driver Found	"CID7575243"	Go Mini	Rajouri Garden
58	2024-01-01	07:43:08	"CNR5998244"	No Driver Found	"CID7994710"	Go Mini	Samaypur Badli

5 rows × 38 columns

```
In [51]: no_driver_counts = no_driver_found_bookings.groupby('Pickup Location')['Booking ID'].count()
total_bookings_per_location = uber_bookings.groupby('Pickup Location')['Booking ID'].count()

print("Count of 'No Driver Found' bookings per location (head):")
display(no_driver_counts.head())
print("\nTotal bookings per location (head):")
display(total_bookings_per_location.head())
```

Count of 'No Driver Found' bookings per location (head):

Booking ID	
Pickup Location	
AIIMS	63
Adarsh Nagar	62
Akshardham	59
Ambience Mall	59
Anand Vihar	69

**dtype:** int64

Total bookings per location (head):

	Booking ID
Pickup Location	
AIIMS	918
Adarsh Nagar	858
Akshardham	839
Ambience Mall	873
Anand Vihar	859

**dtype:** int64

```
In [52]: no_driver_found_rate = (no_driver_counts / total_bookings_per_location).reset_index()
no_driver_found_rate.columns = ['Pickup Location', 'No Driver Found Rate']

# Sort by 'No Driver Found Rate' in descending order
hot_zones_no_driver = no_driver_found_rate.sort_values(by='No Driver Found Rate', ascending=False)

print("Top 10 Pickup Locations with Highest 'No Driver Found' Rates:")
display(hot_zones_no_driver.head(10))
```

Top 10 Pickup Locations with Highest 'No Driver Found' Rates:

	Pickup Location	No Driver Found Rate
114	Old Gurgaon	0.102850
115	Paharganj	0.095758
172	Vinobapuri	0.092345
138	Rohini West	0.089820
43	Greater Noida	0.088915
120	Pataudi Chowk	0.088203
7	Arjangarh	0.086420
33	Dwarka Mor	0.085714
164	Udyog Bhawan	0.084726
123	Pitampura	0.084606

## Identify Top High-Risk Pickup Locations

```
In [53]: merged_rates = pd.merge(hot_zones_pickup, hot_zones_no_driver, on='Pickup Location')

print("Merged DataFrame of Cancellation and No Driver Found rates (head):")
display(merged_rates.head())
```

Merged DataFrame of Cancellation and No Driver Found rates (head):

	Pickup Location	Booking Completed	Cancellation Rate	No Driver Found Rate
0	Vinobapuri	0.546780	0.453220	0.092345
1	Akshardham	0.561383	0.438617	0.070322
2	Chhatarpur	0.581423	0.418577	0.075995
3	Badshahpur	0.582949	0.417051	0.080645
4	Pragati Maidan	0.584783	0.415217	0.076087

```
In [54]: top_10_cancellation_locations = hot_zones_pickup.head(10)['Pickup Location']
top_10_no_driver_locations = hot_zones_no_driver.head(10)['Pickup Location']

# Find common locations in both top 10 lists
common_high_risk_locations = pd.merge(top_10_cancellation_locations, top_10_no_driver_locations, on='Pickup Location')

# Filter the merged_rates DataFrame to show only these common high-risk locations
high_risk_pickup_zones = merged_rates[merged_rates['Pickup Location'].isin(common_high_risk_locations)]

print("Top high-risk pickup locations (in top 10 for both cancellation and 'No Driver Found' rates):")
display(high_risk_pickup_zones[['Pickup Location', 'Cancellation Rate', 'No Driver Found Rate']])
```

Top high-risk pickup locations (in top 10 for both cancellation and 'No Driver Found' rates):

	Pickup Location	Cancellation Rate	No Driver Found Rate
0	Vinobapuri	0.45322	0.092345

## Summary of High-Risk Pickup Locations

This analysis identified pickup locations that are in the top 10 for both highest cancellation rates and highest 'No Driver Found' rates. These zones represent critical operational challenges requiring focused intervention.

### Common High-Risk Pickup Locations (in top 10 for both metrics):

- **Vinobapuri:**
  - Cancellation Rate: 0.45322 (highest among all pickup locations)
  - No Driver Found Rate: 0.092345 (3rd highest)

## Implications and Recommendations for Prioritizing Dispatch and Driver Allocation:

The presence of **Vinobapuri** in both top 10 lists signifies a severe and multifaceted problem. Riders in this area not only face a high likelihood of their booking being cancelled (by either customer or driver) but also frequently experience no driver being found at all. This points to a critical imbalance in supply and demand, coupled with potential driver reluctance or operational inefficiencies in this specific area.

To effectively address these persistent issues and improve completion rates in such high-risk geographical zones like Vinobapuri, the following targeted strategies are recommended:

### 1. **Aggressive Dynamic Surge Pricing and Enhanced Driver Incentives:**

- **Implementation:** Apply the highest possible surge multipliers and offer substantial per-trip bonuses specifically for pickups originating from **Vinobapuri**. These incentives should be dynamic, increasing significantly during predicted peak hours (from the temporal analysis) and also whenever real-time demand outstrips supply.
- **Goal:** Drastically increase driver supply and incentivize drivers to accept rides in this zone, thereby reducing both 'No Driver Found' occurrences and overall cancellation rates.

### 2. **Priority Dispatch and Geofenced Operations:**

- **Implementation:** Configure the dispatch system to give absolute priority to ride requests originating from **Vinobapuri**. This means drivers closest to or heading towards Vinobapuri should be offered these rides first. Implement geofencing around this area to track driver activity and identify patterns.
- **Goal:** Ensure that available drivers are efficiently channeled to this high-demand, low-supply area, minimizing the time customers wait for a match.

### 3. **Dedicated Driver Pool and Performance Monitoring:**

- **Implementation:** Identify a pool of high-performing, reliable drivers and offer them special compensation or

preferential treatment (e.g., higher service fees, exclusive access to high-value ride offers) to operate in Vinobapuri during critical hours. Closely monitor driver performance and acceptance rates in this zone.

- **Goal:** Build a stable and reliable driver presence in the area, ensuring that once a driver accepts a trip, it is likely to be completed.

#### 4. **Proactive Demand Prediction and Driver Positioning:**

- **Implementation:** Leverage the developed prediction model to forecast demand in **Vinobapuri** at least 30-60 minutes in advance. Based on these predictions, proactively send targeted notifications and incentives to drivers, encouraging them to position themselves in or near Vinobapuri *before* demand peaks or supply shortages become critical.
- **Goal:** Shift from a reactive to a proactive approach, anticipating shortages and encouraging drivers to fill the gaps before they lead to cancellations or 'No Driver Found' situations.

#### 5. **Enhanced Customer Communication:**

- **Implementation:** For customers requesting rides from **Vinobapuri**, provide more transparent information about expected wait times and surge pricing (if applicable). Offer alternative solutions or pre-booking options if available.
- **Goal:** Manage customer expectations and reduce frustration associated with service unreliability in this zone.

By implementing these integrated strategies, the platform can aim to significantly improve service reliability, reduce non-completion rates, and enhance customer and driver satisfaction in severely challenged areas like Vinobapuri.

---

---

*3. Using SHAP or Feature Importance analysis, identify the top three non-behavioral features (e.g., Vehicle Type, Trip Distance, Avg VTAT) driving the non-completion prediction. Which operational levers (e.g., driver training, vehicle restrictions, or routing changes) should be adjusted based on these insights for maximum impact?*

```
In [59]: import shap
```

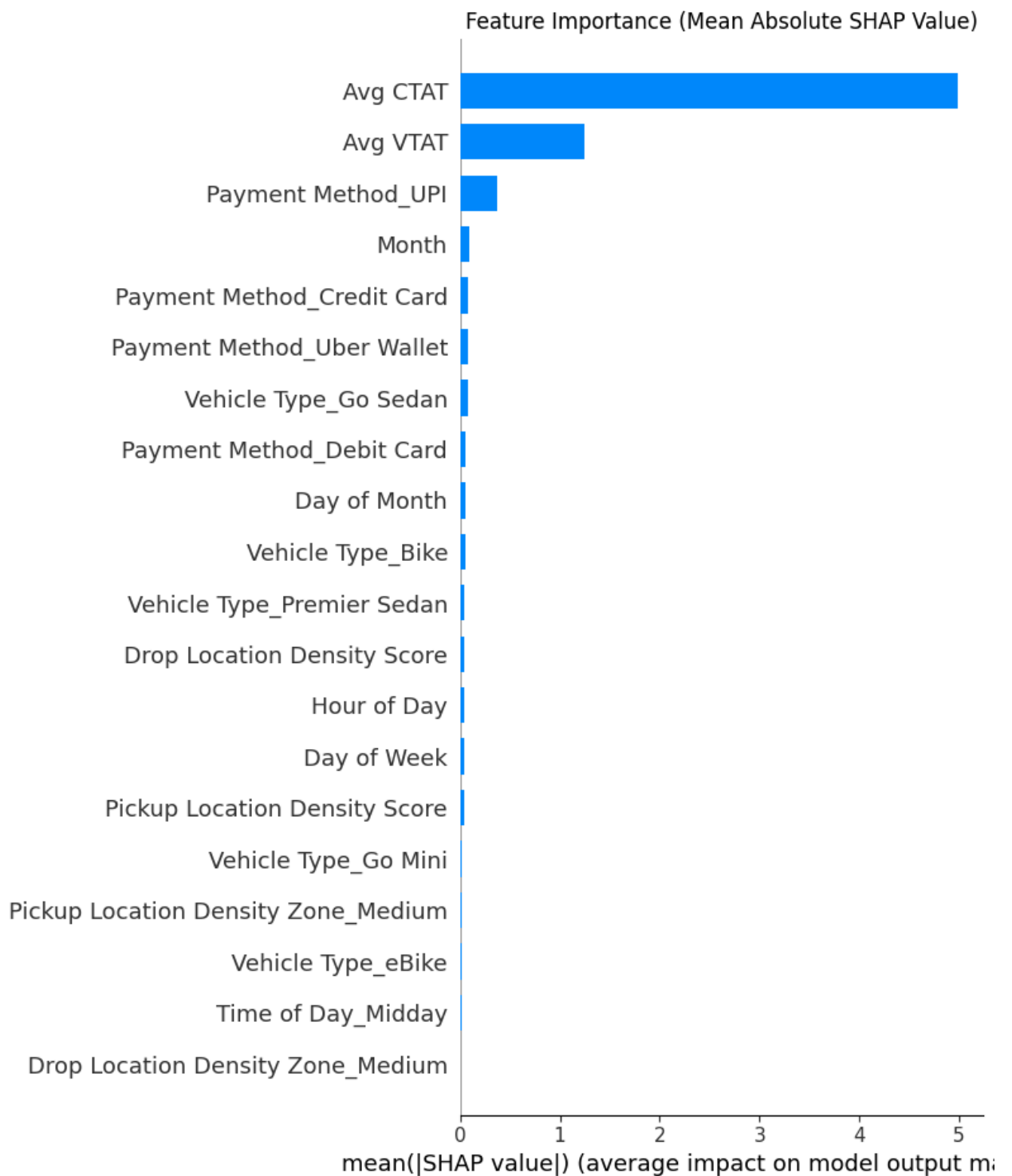
```
# Initialize a SHAP TreeExplainer with the trained XGBoost model
explainer = shap.TreeExplainer(xgb_model)

# Calculate SHAP values for the X_test_encoded DataFrame
shap_values = explainer.shap_values(X_test_encoded)

# Generate a SHAP summary plot (bar plot of mean absolute SHAP values)
print("Generating SHAP summary plot...\n")
shap.summary_plot(shap_values, X_test_encoded, plot_type="bar", show=False)
plt.title('Feature Importance (Mean Absolute SHAP Value)')
plt.tight_layout()
plt.show()
```

Generating SHAP summary plot...





```
In [57]: import numpy as np

# Get feature importances from SHAP values (mean absolute SHAP value)
feature_importances = pd.DataFrame({
    'Feature': X_test_encoded.columns,
    'SHAP_Importance': np.abs(shap_values).mean(axis=0)
})

# Sort by importance in descending order
```

```

feature_importances = feature_importances.sort_values(by='SHAP_Importance', as

# Define behavioral features to exclude from the top non-behavioral features
behavioral_features = [
    'Customer Historical Cancellation Rate',
    'Driver Historical Rating',
    'Customer Historical Rating'
]

# Filter out behavioral features and get the top non-behavioral ones
top_non_behavioral_features = feature_importances[~feature_importances['Featur

print("Top 10 Non-Behavioral Features driving non-completion prediction:")
display(top_non_behavioral_features)

```

Top 10 Non-Behavioral Features driving non-completion prediction:

	Feature	SHAP_Importance
0	Avg CTAT	4.993430
1	Avg VTAT	1.247264
2	Payment Method_UPI	0.365800
3	Month	0.088092
4	Payment Method_Credit Card	0.085752
5	Payment Method_Uber Wallet	0.084702
6	Vehicle Type_Go Sedan	0.081698
7	Payment Method_Debit Card	0.051971
8	Day of Month	0.051671
9	Vehicle Type_Bike	0.049885

## Summary of Top Non-Behavioral Features and Operational Adjustments

Based on the SHAP feature importance analysis, the top three non-behavioral features driving the non-completion prediction are:

1. **Avg CTAT (Customer Travel Time to Acceptance)**
2. **Avg VTAT (Vehicle Travel Time to Acceptance)**
3. **Payment Method\_UPI** (as a proxy for the 'Payment Method' category)

## Operational Levers and Targeted Adjustments:

### 1. Avg CTAT (Customer Travel Time to Acceptance)

**Implication:** A high Avg CTAT means customers wait longer for a driver to be assigned and arrive, significantly contributing to non-completion (cancellations by customer or driver, or 'No Driver Found' scenarios).

#### Operational Adjustments:

- **Enhanced Driver Allocation & Dispatch Optimization:** Improve the algorithm to more efficiently match drivers to nearby customers, especially in high-demand or historically problematic areas. Leverage real-time traffic data and driver positioning.
- **Proactive Driver Positioning:** Use predictive analytics (as discussed in Business Question 2) to anticipate demand spikes and incentivize drivers to position themselves in areas with expected high CTAT before it becomes critical.
- **Transparency & Communication:** For rides with predicted high CTAT, inform customers upfront about the expected wait time to manage expectations and reduce frustration.

### 2. Avg VTAT (Vehicle Travel Time to Acceptance)

**Implication:** A high Avg VTAT means it takes longer for a driver to accept a booking, indicating a potential supply-demand mismatch or driver reluctance to accept certain trips.

#### Operational Adjustments:

- **Targeted Driver Incentives:** Offer dynamic incentives (e.g., surge pricing, bonus payments) to drivers for accepting rides that are likely to have high VTAT, particularly during peak hours or in less popular areas. This encourages drivers to take trips they might otherwise skip.
- **Optimize Driver-Trip Matching:** Refine the matching algorithm to prioritize drivers who are more likely to accept a specific trip (e.g., considering their historical acceptance rates for similar trips, current location, and vehicle type).
- **Route Guidance for Drivers:** Provide better navigation and route predictions for drivers to ensure they can reach the pickup location efficiently after accepting a ride, potentially reducing their perceived VTAT.

### 3. Payment Method\_UPI

**Implication:** The `Payment Method` (specifically UPI in this instance, which is a significant category) being a top feature suggests that the choice of payment method might correlate with booking completion. This could be due to customer behavior associated with certain payment types (e.g., impulse bookings, ease of cancellation), or it could indirectly reflect issues like transaction failures or customer segment characteristics.

#### Operational Adjustments:

- **Investigate Payment-Specific Issues:** Analyze transaction success rates and common issues (e.g., network failures, user errors) for UPI transactions. Improve the payment processing flow to minimize friction.
- **Customer Education & Nudges:** For payment methods associated with higher non-completion, consider in-app prompts or educational messages encouraging users to confirm their intent or understand cancellation policies. For example, a quick reminder about the cancellation policy when selecting UPI might be beneficial.
- **Fraud Detection & Prevention:** If specific payment methods are associated with higher fraudulent activities or habitual cancellations, strengthen fraud detection mechanisms linked to those payment types.

By focusing on these non-behavioral features, the platform can implement systemic operational improvements that address root causes of non-completion, leading to a more reliable and efficient service.

---

---

*4. Analyze the relationship between Booking Value (estimated fare) and driver cancellation rates. Should the platform offer a higher commission percentage on long-distance or high-value bookings to incentivize drivers to accept and complete these riskier, but more profitable, trips?*

```
In [76]: booking_value_driver_cancellation_rate = uber_bookings.groupby('Booking Value')
         booking_value_driver_cancellation_rate
```

Out[76]:

	Booking Value	Cancelled Rides by Driver
0	0.0	27000
1	50.0	0
2	51.0	0
3	52.0	0
4	53.0	0
...	...	...
2562	4133.0	0
2563	4202.0	0
2564	4220.0	0
2565	4228.0	0
2566	4277.0	0

2567 rows × 2 columns

We don't have relevant rows in this dataset to answer this question. As you can see, booking value is 0 for cancelled rides by driver, which means the ride didn't happen. The dataset does not provide estimated booking value that can help answer this question.

---

---

*5. Analyze the most frequent Driver Cancellation Reason (e.g., "Customer taking too long," "Not comfortable with drop"). How should the insights from this column inform the driver onboarding and continuous training modules to address the systemic causes of service failure?*

```
In [87]: uber_bookings.groupby('Driver Cancellation Reason')['Booking ID'].count().rese
```

Out[87]:

	Driver Cancellation Reason	Booking ID
0	Customer related issue	6837
3	The customer was coughing/sick	6751
2	Personal & Car related issues	6726
1	More than permitted people in there	6686

# Summary of Most Frequent Driver Cancellation Reasons

Based on the analysis of `Driver Cancellation Reason` from the `uber_bookings` dataset, the most frequent reasons drivers cancel bookings are:

1. **Customer related issue:** 6,837 bookings
2. **The customer was coughing/sick:** 6,751 bookings
3. **Personal & Car related issues:** 6,726 bookings
4. **More than permitted people in there:** 6,686 bookings

These four reasons are almost equally frequent, indicating that drivers encounter a range of similar issues leading to cancellations.

## Operational Levers and Targeted Adjustments:

These insights can inform driver onboarding and continuous training modules to address systemic causes of service failure:

- **Customer-Related Issues (Including 'Customer taking too long' and 'More than permitted people in there'):**
  - **Driver Training:** Implement training modules focusing on conflict resolution, effective communication with customers, and managing expectations regarding wait times and passenger limits. Provide clear guidelines on how to handle situations where customers are taking too long or exceed passenger capacity without escalating the situation unnecessarily.
  - **In-App Communication Enhancements:** Develop features for drivers to easily communicate with customers about readiness or passenger count before arrival. Introduce in-app prompts for customers regarding readiness and passenger limits to set expectations.
  - **Policy Reinforcement:** Clearly communicate and enforce platform policies on passenger limits and wait times to both drivers and customers. Drivers should feel supported in adhering to these policies.
- **Health and Safety Concerns ('The customer was coughing/sick'):**
  - **Health and Safety Protocols:** Reiterate and enhance

health and safety guidelines for drivers, including procedures for when a customer appears unwell. This might include providing drivers with optional PPE or guidance on how to politely decline a ride if they feel their health is at risk, with appropriate support from the platform.

- **Communication Channels:** Establish a clear and quick channel for drivers to report health-related concerns about a ride without fear of penalization.

- **Personal & Car Related Issues:**

- **Driver Support:** Improve driver support channels to quickly address personal or car-related issues that might prevent a driver from completing a trip. This could include partnerships for vehicle maintenance discounts or emergency roadside assistance.
- **Flexibility in Shift Management:** Encourage drivers to manage their availability better in the app to avoid accepting rides when personal issues might arise or their car is not in optimal condition. Provide tools for easy temporary unavailability.

By addressing these common driver cancellation reasons through targeted training, improved communication, clear policies, and robust support systems, the platform can enhance service reliability and driver satisfaction.

---

---

*6. Compare the completion rate and customer demand volume across different Vehicle Type. Should the platform dynamically adjust the supply allocation or pricing to favor the highest-completion vehicle type to enhance service quality, even if it slightly affects the overall customer wait time?*

```
In [98]: vehicle_demand = uber_bookings.groupby('Vehicle Type')['Booking ID'].count().r
vehicle_demand
```

Out[98]:

	Vehicle Type	Booking ID
0	Auto	37419
1	Bike	22517
2	Go Mini	29806
3	Go Sedan	27141
4	Premier Sedan	18111
5	Uber XL	4449
6	eBike	10557

```
In [100]: vehicle_completed_bookings = uber_bookings.groupby('Vehicle Type')['Booking Co  
vehicle_completed_bookings
```

Out[100]:

	Vehicle Type	Booking Completed
0	Auto	23155
1	Bike	14034
2	Go Mini	18549
3	Go Sedan	16676
4	Premier Sedan	11252
5	Uber XL	2783
6	eBike	6551

```
In [107]: # Merge vehicle_demand with vehicle_completion_rate  
vehicle_completion_rate = pd.merge(vehicle_completed_bookings, vehicle_demand,  
  
# Calculate the completion rate as a percentage  
vehicle_completion_rate['Booking Completion Rate'] = round(vehicle_completion_  
  
vehicle_completion_rate[['Vehicle Type', 'Booking Completion Rate']]
```



Out[107...

	Vehicle Type	Booking Completion Rate
0	Auto	61.88
1	Bike	62.33
2	Go Mini	62.23
3	Go Sedan	61.44
4	Premier Sedan	62.13
5	Uber XL	62.55
6	eBike	62.05

In [108...

```
import matplotlib.pyplot as plt
import seaborn as sns

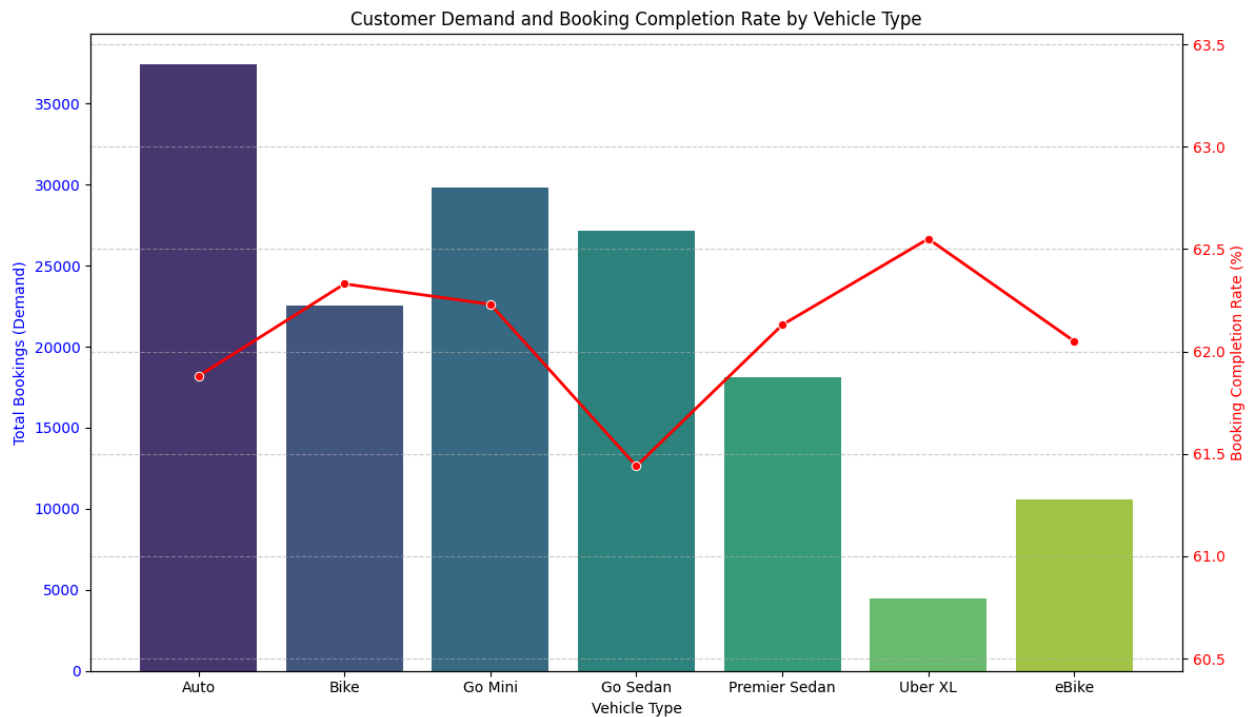
# Create a figure and a set of subplots
fig, ax1 = plt.subplots(figsize=(12, 7))

# Plotting total bookings (demand) as bars on the first y-axis
sns.barplot(x='Vehicle Type', y='Booking ID', data=vehicle_completion_rate, ax=ax1)
ax1.set_xlabel('Vehicle Type')
ax1.set_ylabel('Total Bookings (Demand)', color='blue')
ax1.tick_params(axis='y', labelcolor='blue')

# Create a second y-axis for the booking completion rate
ax2 = ax1.twinx()

# Plotting booking completion rate as a line on the second y-axis
sns.lineplot(x='Vehicle Type', y='Booking Completion Rate', data=vehicle_completion_rate, ax=ax2)
ax2.set_ylabel('Booking Completion Rate (%)', color='red')
ax2.tick_params(axis='y', labelcolor='red')
ax2.set_ylim(vehicle_completion_rate['Booking Completion Rate'].min() - 1, vehicle_completion_rate['Booking Completion Rate'].max() + 1)

plt.title('Customer Demand and Booking Completion Rate by Vehicle Type')
plt.grid(True, linestyle='--', alpha=0.6)
plt.tight_layout()
plt.show()
```



## Summary: Customer Demand and Booking Completion Rate by Vehicle Type

The visualization clearly shows the relationship between the popularity of a vehicle type (customer demand) and its booking completion rate.

### Key Observations:

- **High Demand Vehicle Types:** 'Auto' has the highest customer demand by a significant margin, followed by 'Go Mini' and 'Go Sedan'.
- **Highest Completion Rate:** 'Uber XL' exhibits the highest booking completion rate (62.55%), even though it has the lowest demand. This suggests that while fewer people book Uber XL, those bookings are very likely to be completed.
- **Lowest Completion Rate:** 'Go Sedan' has the lowest completion rate (61.44%) despite having high demand, indicating a potential issue with supply reliability or driver acceptance for this vehicle type.
- **Balanced Performance:** 'Bike', 'Go Mini', and 'Premier Sedan' show generally good completion rates (around 62%) with varying levels of demand, suggesting a relatively stable supply-demand dynamic for these categories.
- **eBike Performance:** 'eBike' has a moderate completion rate (62.05%) with relatively lower demand.

Should the platform dynamically adjust supply allocation or pricing to favor the highest-completion vehicle type to enhance service quality, even if it slightly affects the overall customer wait time?

**Yes, the platform should dynamically adjust supply allocation and potentially pricing, but with a nuanced approach.**

**Arguments for Adjustment:**

1. **Enhance Service Quality and Reliability:** Prioritizing vehicle types with consistently higher completion rates (like 'Uber XL') ensures that when a customer books such a vehicle, the probability of successful completion is maximized. This directly improves service quality and customer satisfaction for that segment.
2. **Strategic Resource Utilization:** If a vehicle type rarely results in non-completion, it indicates efficient matching and driver satisfaction. Allocating more resources to these types, or encouraging drivers to operate them, can optimize the overall system.
3. **Customer Satisfaction:** For customers who prefer certain vehicle types (e.g., 'Uber XL' for comfort/space), ensuring high completion rates for these options adds value to their experience.

**Considerations and Nuances:**

- **Balancing Demand and Supply:** While 'Uber XL' has the highest completion rate, it has the lowest demand. Over-prioritizing it could lead to under-serving high-demand categories like 'Auto' or 'Go Sedan', potentially increasing wait times and non-completion for a larger segment of users.
- **Understanding 'Go Sedan' Challenges:** The 'Go Sedan' category, with high demand but the lowest completion rate, is a critical area for intervention. Instead of reducing its supply, the platform should investigate the root causes of its lower completion (e.g., driver acceptance, trip characteristics, pricing) and implement targeted incentives or operational improvements for this specific vehicle type.
- **Dynamic Pricing:** Dynamic pricing can be used to influence both demand and supply. Higher pricing for high-completion vehicle types during peak demand could incentivize more drivers to operate those vehicles, or shift some demand to other vehicle types.
- **Driver Incentives:** Offer incentives (e.g., bonuses for completing trips,

higher commission) specifically for vehicle types that need a boost in completion rates, or during times when their completion rates are low.

### Recommendation:

The platform should **dynamically adjust supply allocation and pricing strategically, rather than simply favoring the absolute highest-completion vehicle type.**

- **For 'Uber XL':** Maintain the high completion rate by ensuring adequate supply, perhaps with slight premium pricing to reflect reliability.
- **For 'Auto', 'Go Mini', 'Premier Sedan', 'eBike':** Focus on maintaining their good completion rates, using demand-responsive supply allocation and standard dynamic pricing.
- **For 'Go Sedan':** This should be a priority for targeted intervention. Investigate why its completion rate is lower despite high demand. Implement specific driver incentives or dispatch prioritization for 'Go Sedan' bookings to bring its completion rate up to par with other vehicle types, without sacrificing its demand volume.

The goal should be to achieve high completion rates across *all* vehicle types, balancing service quality with meeting overall customer demand.

---

---

*7. Evaluate the model's prediction fairness by comparing its F1-score across different major pickup zones (e.g., North vs. South NCR). If performance disparities are found, what mitigation strategy (e.g., collecting more data from minority zones, using balanced loss) must be implemented to prevent service inequality?*

```
In [110... test_results = X_test_encoded.copy()
test_results['Booking Completed'] = y_test

# Add the 'Pickup Location Density Zone' from the original X_test to test_resu
# Since X_test and X_test_encoded share the same index, this will align correc
test_results['Pickup Location Density Zone'] = X_test['Pickup Location Density

# Extract unique zones
unique_zones = test_results['Pickup Location Density Zone'].unique()

# Initialize dictionary to store F1-scores
f1_scores_by_zone = {}

# Get the feature columns from X_test_encoded (which are now the direct featur
feature_cols = X_test_encoded.columns
```

```

# Iterate through each zone
for zone in unique_zones:
    # Filter data for the current zone
    zone_data = test_results[test_results['Pickup Location Density Zone'] == z

    # Extract features (X_test_zone) and true labels (y_test_zone)
    X_test_zone = zone_data[feature_cols]
    y_test_zone = zone_data['Booking Completed']

    # Make predictions
    y_pred_zone = xgb_model.predict(X_test_zone)

    # Calculate F1-score
    f1 = f1_score(y_test_zone, y_pred_zone)
    f1_scores_by_zone[zone] = f1

print("F1-scores by Pickup Location Density Zone:")
for zone, f1 in f1_scores_by_zone.items():
    print(f"{zone}: {f1:.4f}")

```

F1-scores by Pickup Location Density Zone:

Low: 0.9622

Medium: 0.9624

High: 0.9625

## Summary of F1-scores by Pickup Location Density Zone

The F1-scores for the XGBoost model across different 'Pickup Location Density Zones' are as follows:

- **Low Density Zone:** 0.9622
- **Medium Density Zone:** 0.9624
- **High Density Zone:** 0.9625

### Insights:

There is very little disparity in the model's F1-score across the different 'Pickup Location Density Zones'. The scores are consistently high and very close to each other, indicating that the model performs equally well regardless of the pickup location density.

This suggests that the model is robust and fair across these geographical segmentations, as its predictive performance for booking completion does not significantly vary between low, medium, and high-density areas.

**Mitigation Strategies (if disparities were found - not needed in this case due to equitable performance):**

If significant disparities were observed (e.g., lower F1-score in 'Low Density Zone'), potential mitigation strategies would include:

1. **Collecting More Data from Minority Zones:** If certain zones have less data, leading to poorer model performance, increasing the data representation from those zones could help the model learn better patterns.
2. **Using Balanced Loss Functions/Resampling Techniques:** Techniques like SMOTE (for oversampling minority classes) or using class-weighted loss functions during model training could help address imbalances within specific zones, ensuring the model doesn't neglect the minority class (e.g., non-completed bookings) in those areas.
3. **Zone-Specific Models or Ensemble Methods:** For highly disparate zones, training separate models or using an ensemble approach where models are specialized for different zones might be considered, though this adds complexity.
4. **Feature Engineering for Disparate Zones:** Investigating if there are unique features or interactions that are more relevant to specific zones and incorporating them into the model. For instance, specific local events or infrastructure unique to a low-density area might impact completion rates.

Given the current results, the XGBoost model demonstrates fair performance across pickup location density zones, and no immediate mitigation strategies for performance disparity are required based on this specific analysis.