Final Project Datamining

Member :

1. นาย ณภัทร พรรณเชษฐ์ 6610502013
2. นาย ชยกร ศรุตยาพร 6610505331

Subject : Uber Data Analytics

Dataset : Kaggle Uber Data Analytics

Exploratory Data Analysis (EDA) :

Data Overview :

ข้อมูลชุดนี้นำเข้ามาจากเว็บไซต์ Kaggle โดยใช้ชื่อชุดข้อมูลว่า Uber Ride Analytics Dashboard ซึ่งเป็นข้อมูลที่จำลองการให้บริการเรียกรถของ Uber ชุดข้อมูลนี้ประกอบด้วย 150,000 รายการ และมี 21 คอลลัมน์ โดยแต่ละรายการแทนการจองรถหนึ่งครั้ง ซึ่งบันทึกข้อมูลเกี่ยวกับ ลูกค้า คนขับ ยานพาหนะ สถานะการจอง มูลค่าการเดินทาง ระยะทาง คะแนนการให้บริการ และเวลาที่เกิดเหตุการณ์ เป็นต้น

ข้อมูลสามารถแบ่งออกได้เป็น 3 กลุ่มหลักคือ

1. Numerical Features เช่น Booking Value, Ride Distance, Driver Ratings, Customer Rating, Avg VTAT, Avg CTAT
2. Binary Features เช่น Cancelled Rides by Customer, Cancelled Rides by Driver, Incomplete Rides
3. Categorical Features เช่น Vehicle Type, Payment Method, Booking Status, Pickup Location, Drop Location

ข้อมูลนี้สะท้อนกระบวนการทำงานทั้งหมดของระบบเรียกรถ ตั้งแต่การจอง การให้บริการ ไปจนถึงการชำระเงินและการให้คะแนน ซึ่งสามารถนำมาใช้วิเคราะห์แนวโน้มการเดินทาง พฤติกรรมลูกค้า และประสิทธิภาพของคนขับได้อย่างครอบคลุม นอกจากนี้ ข้อมูลยังมีลักษณะของ time-series ทำให้สามารถวิเคราะห์แนวโน้มตามวันและช่วงเวลาได้อีกด้วย

Column Description :

Column Name	คำอธิบาย
Date	วันที่ของการจอง
Time	เวลาที่ทำการจอง
Booking ID	รหัสระบุเฉพาะของแต่ละการจอง
Booking Status	สถานะของการจอง เช่น Completed, Cancelled by Customer, Cancelled by Driver
Customer ID	รหัสประจำตัวของลูกค้า
Vehicle Type	ประเภทยานพาหนะ เช่น *Go Mini*, *Go Sedan*, *Auto*, *eBike/Bike*, *UberXL*, *Premier Sedan*
Pickup Location	จุดรับผู้โดยสาร
Drop Location	จุดหมายปลายทางของการเดินทาง
Avg VTAT	ระยะเวลาเฉลี่ยที่คนขับใช้ในการมาถึงจุดรับผู้โดยสาร (นาที)
Avg CTAT	ระยะเวลาเฉลี่ยของการเดินทางจากจุดรับถึงจุดหมาย (นาที)
Cancelled Rides by Customer	ตัวบ่งชี้ว่าลูกค้าเป็นผู้ยกเลิกการจอง
Reason for cancelling by Customer	เหตุผลที่ลูกค้ายกเลิกการจอง
Cancelled Rides by Driver	ตัวบ่งชี้ว่าคนขับเป็นผู้ยกเลิกการจอง
Driver Cancellation Reason	เหตุผลที่คนขับยกเลิกการจอง
Incomplete Rides	ตัวบ่งชี้ว่าการเดินทางไม่สมบูรณ์
Incomplete Rides Reason	เหตุผลที่การเดินทางไม่สมบูรณ์
Booking Value	ค่าโดยสารทั้งหมด
Ride Distance	ระยะทางที่เดินทางจริง (กิโลเมตร)
Driver Ratings	คะแนนที่ลูกค้าให้กับคนขับ (ระดับ 1–5)

Customer Rating	คะแนนที่คนขับให้กับลูกค้า (ระดับ 1–5)
Payment Method	วิธีการชำระเงิน เช่น *UPI*, *Cash*, *Credit Card*, *Uber Wallet*, *Debit Card*

เหตุผลที่ข้อมูลชุดนี้จัดเป็น **Complex Data :**

ข้อมูล Uber ชุดนี้มีคุณลักษณะตรงกับนิยามของ Complex Data ตามที่อาจารย์กำหนดไว้ ดังนี้

1. **Imbalanced Data**

   ค่าของ Booking Status มีความไม่สมดุลอย่างชัดเจน โดยกลุ่ม Completed มีจำนวนมากกว่ากลุ่ม Cancelled หรือ Incomplete หลายเท่า ส่งผลให้การวิเคราะห์เชิงจำแนก (classification) ต้องใช้เทคนิคจัดการข้อมูลไม่สมดุล

2. **High Dimensional Data**

   ข้อมูลประกอบด้วยตัวแปรถึง 21 คอลลัมน์ ที่มีทั้งตัวเลขและข้อความ บางคอลัมน์ เช่น Reason for cancelling, Location มีค่า string ที่ซับซ้อนและต้องแปลงก่อนนำไปวิเคราะห์ การทำความเข้าใจความสัมพันธ์ระหว่างหลายตัวแปรจำเป็นต้องใช้เทคนิคเชิงสถิติและ visualization

Data Summary :

```
<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
```

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

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

dtype: int64

จากภาพจะพบว่ามีบันทึกข้อมูลการจองทั้งหมด 150,000 รายการ และโดยมี 21 คอลลัมน์ และมี 13 คอลลัมน์ที่ยังมี missing value อยู่

```
df.describe()
```

	Avg VTAT	Avg CTAT	Cancelled Rides by Customer	Cancelled Rides by Driver	Incomplete Rides	Booking Value	Ride Distance	Driver Ratings	Customer Rating
count	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000	150000.000000
mean	7.864407	19.821753	0.070000	0.180000	0.060000	345.641220	16.753168	2.623215	2.730842
std	4.230640	15.452834	0.255148	0.384189	0.237488	403.423487	16.291118	2.082283	2.165548
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	4.700000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	7.800000	22.000000	0.000000	0.000000	0.000000	244.000000	13.060000	3.900000	4.100000
75%	11.000000	32.900000	0.000000	0.000000	0.000000	521.000000	30.650000	4.300000	4.600000
max	20.000000	45.000000	1.000000	1.000000	1.000000	4277.000000	50.000000	5.000000	5.000000

1. Booking Value มีค่าเฉลี่ยประมาณ 345 บาท และสูงสุดถึง 4,277 บาท แสดงถึงการกระจายของมูลค่าการเดินทางที่ค่อนข้างกว้าง
2. Ride Distance เฉลี่ยอยู่ที่ 16.75 กม. และสูงสุดถึง 50 กม.
3. Driver Ratings และ Customer Rating มีค่าเฉลี่ยประมาณ 2.6 – 2.7 คะแนน จาก 5 คะแนน สะท้อนถึงระดับความพึงพอใจปานกลาง
4. ค่าเฉลี่ยของเวลาเดินทาง Avg CTAT อยู่ที่ 19.82 นาที โดยมีค่าเบี่ยงเบนมาตรฐานสูงซึ่ง แสดงถึงความหลากหลายของระยะเวลาเดินทาง

Handle Missing Value :

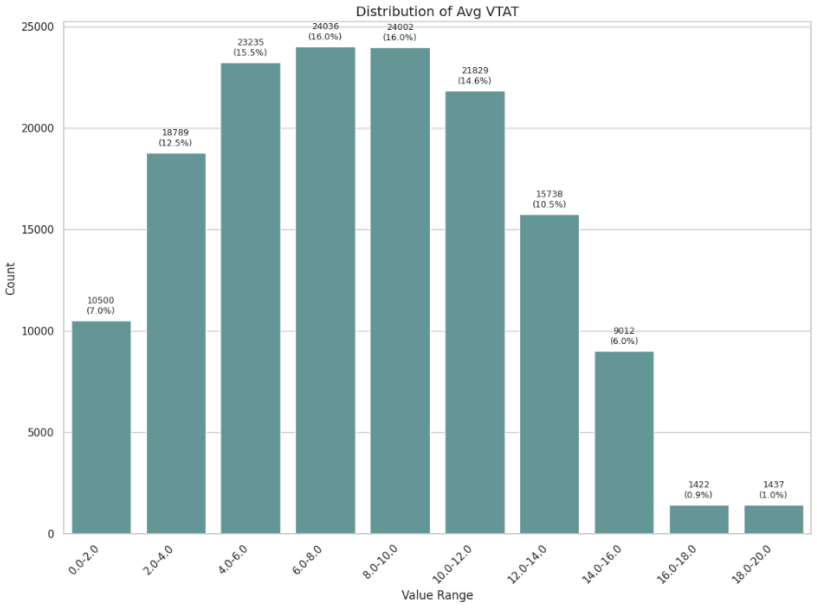
```
# แทนที่ค่าในช่องที่ว่างด้วย 0
df['Avg VTAT'] = df['Avg VTAT'].fillna(0)
df['Avg CTAT'] = df['Avg CTAT'].fillna(0)
df['Cancelled Rides by Customer'] = df['Cancelled Rides by Customer'].fillna(0)
df['Cancelled Rides by Driver'] = df['Cancelled Rides by Driver'].fillna(0)
df['Incomplete Rides'] = df['Incomplete Rides'].fillna(0)
df['Booking Value'] = df['Booking Value'].fillna(0)
df['Ride Distance'] = df['Ride Distance'].fillna(0)
df['Driver Ratings'] = df['Driver Ratings'].fillna(0)
df['Customer Rating'] = df['Customer Rating'].fillna(0)
# แทนที่ค่าในช่องที่ว่างด้วย "..."
df['Incomplete Rides Reason'] = df['Incomplete Rides Reason'].fillna("No Reason")
df['Reason for cancelling by Customer'] = df['Reason for cancelling by Customer'].fillna("No Reason")
df['Driver Cancellation Reason'] = df['Driver Cancellation Reason'].fillna("No Reason")
df['Payment Method'] = df['Payment Method'].fillna("Undefined")
```

จากการตรวจสอบด้วยคำสั่ง df.isnull().sum() พบว่าข้อมูลบางคอลัมน์มีค่าที่หายไป เป็นจำนวนมาก โดยเฉพาะในคอลัมน์ที่เกี่ยวข้องกับการยกเลิกการเดินทาง หรือเหตุผลของการเดินทางที่ไม่สมบูรณ์ ซึ่งค่าที่หายไปเหล่านี้เกิดขึ้นตามเงื่อนไขของธุรกิจจริง (เช่น ถ้าไม่มีการยกเลิก ระบบจะไม่บันทึกเหตุผลไว้) ดังนั้นจึงถือเป็น Missing by Design ไม่ใช่ข้อผิดพลาดของข้อมูล แต่เพื่อให้สามารถนำข้อมูลไปวิเคราะห์ต่อได้อย่างถูกต้อง ได้มีการแทนค่าที่หายไปด้วยค่าที่เหมาะสม โดยแบ่งแนวทางการจัดการออกเป็น 2 กลุ่มหลัก ดังนี้

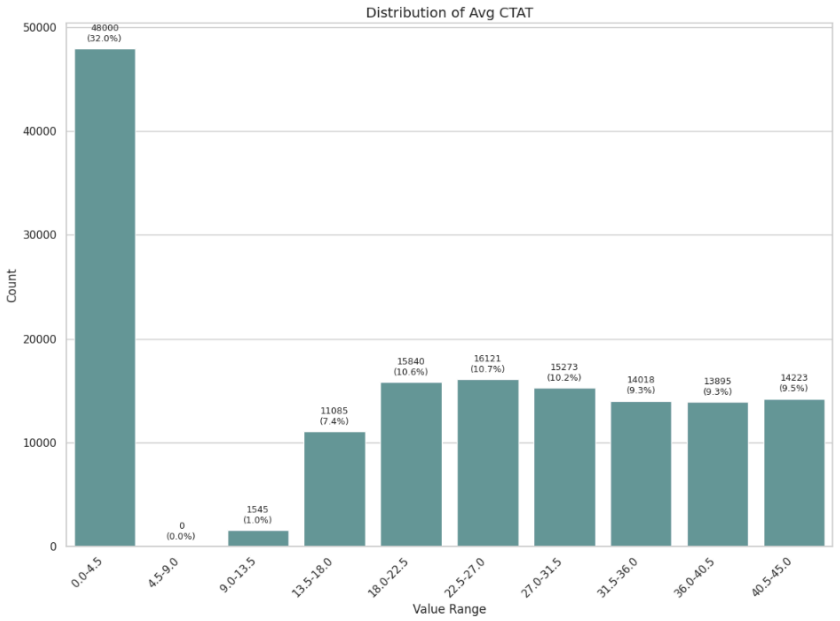
1. Numerical Columns แทนค่าที่หายไปด้วย 0
2. Categorical Columns แทนค่าที่หายไปด้วยข้อความที่อธิบายสถานการณ์
   เช่น No Reason หรือ Undefined

Numerical Features :

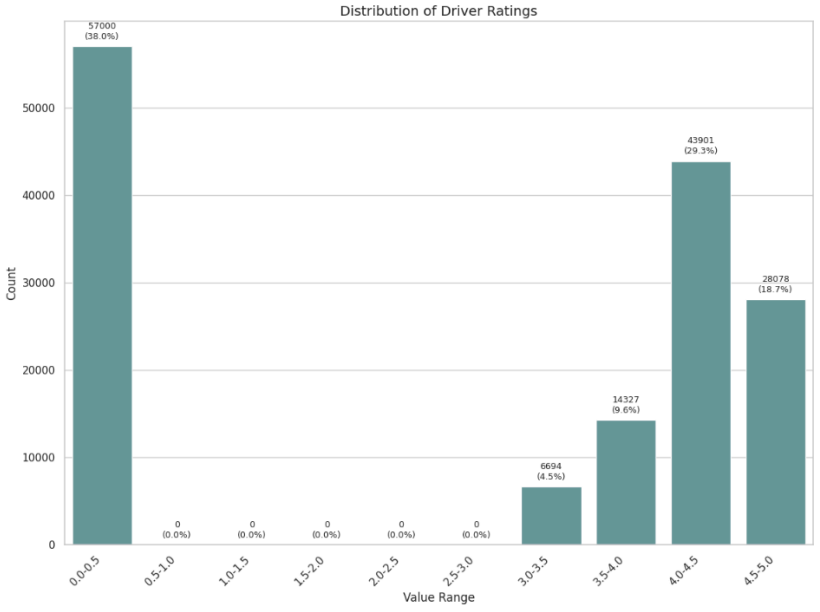
ในการสำรวจ Numerical Features ของชุดข้อมูลนี้ ได้ทำการวิเคราะห์ตัวแปรที่มีค่าต่อเนื่องและสามารถคำนวณเชิงสถิติได้ เช่น Booking Value, Ride Distance, Driver Ratings, Customer Rating, Avg VTAT, และ Avg CTAT การวิเคราะห์ในส่วนนี้มีวัตถุประสงค์เพื่อทำความเข้าใจ การกระจายของค่า (Distribution) ตรวจสอบแนวโน้ม ที่อาจส่งผลต่อขั้นตอนการสร้างโมเดลในภายหลัง โดยใช้กราฟประเภท Histogram

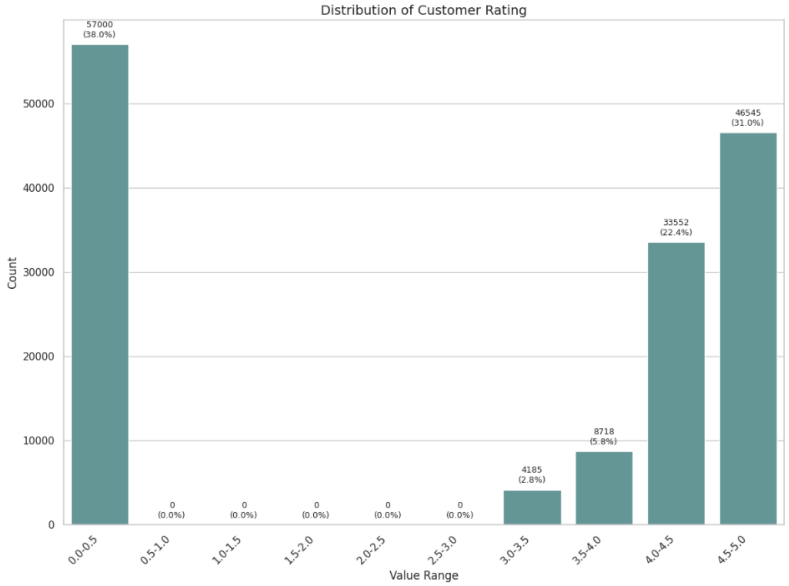


จากภาพข้อมูลมีการกระจายตัวแบบมาตรฐาน ซึ่งข้อมูลจะกระจุกตัวอยู่ในช่วง 4 - 10 นาที ซึ่งคิดเป็นประมาณ 48% ของข้อมูลทั้งหมด

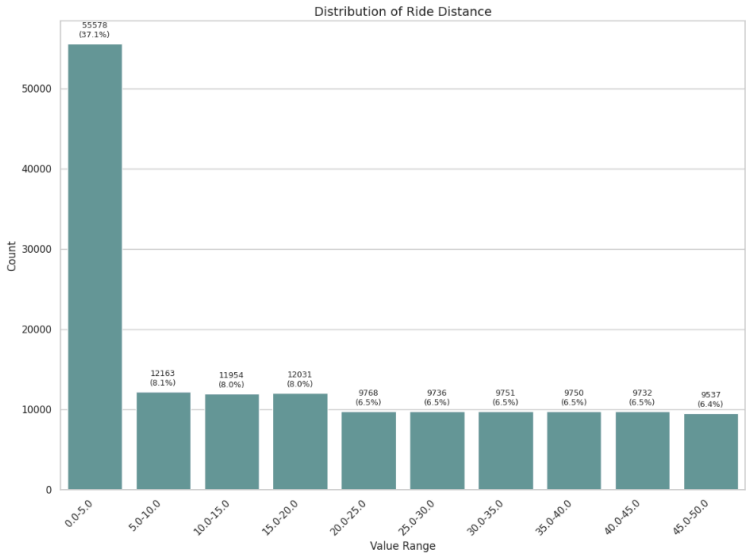


จากภาพจะเห็นว่าข้อมูลกระจุกตัวอยู่ในช่วง 0 - 4.5 นาที ประมาณ 32% ซึ่งเป็นข้อมูลที่ ไม่สมบูรณ์หรือยกเลิกก่อนเดินทาง ซึ่งถูกแทนด้วยค่า 0 หลังจากทำการเติมค่าที่หาย โดยจากข้อมูลเราจะเห็นว่าข้อมูลมีการกระจุกตัวกันในช่วง 18 - 45 นาที ซึ่งแต่ละช่วงมีสัดส่วนใกล้เคียงกัน

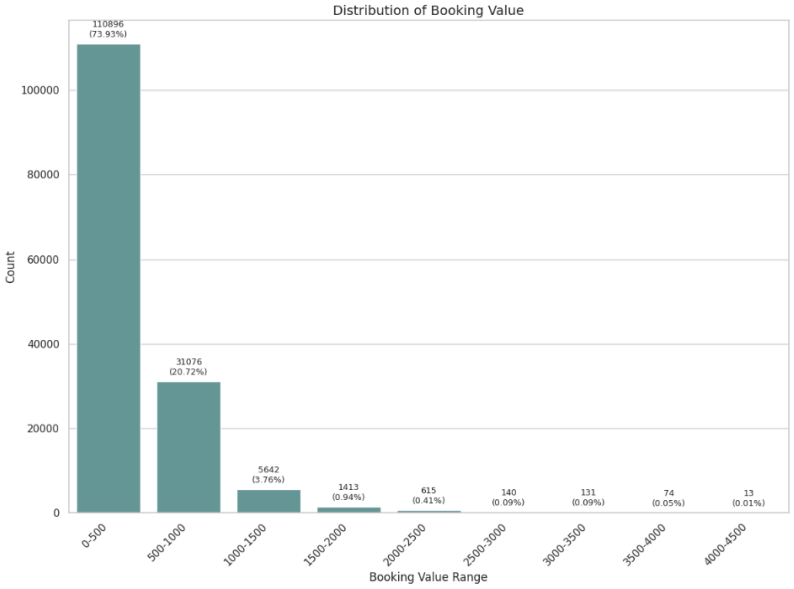




จากภาพจะเป็นการกระจายตัวของ Driver Rateing และ Customer Rating พบว่าประมาณ 38% ของข้อมูลอยู่ในช่วง 0.0–0.5 ซึ่งไม่ได้เกิดจากการให้คะแนนต่ำจริง แต่เกิดจากการจองที่ถูกยกเลิกหรือไม่สำเร็จ ทำให้ระบบไม่มีข้อมูลการให้คะแนนและถูกแทนด้วยค่า 0 หลังจากทำการเติมค่าที่หาย ในขณะที่การเดินทางที่สำเร็จส่วนใหญ่จะได้รับคะแนนเฉลี่ยในช่วง 4.0–5.0 แสดงถึงความพึงพอใจในระดับสูงของผู้ใช้ที่ใช้บริการจนจบการเดินทาง



จากกราฟเป็นการกระจายตัวของ Ride Distance ซึ่งมีข้อมูลกระจุกตัวอยู่ในช่วง 0 - 5 กิโลเมตรถึง 37.1% ซึ่งเป็นการเดินทางที่ ไม่สมบูรณ์หรือ ยกเลิกก่อนเดินทาง ซึ่งถูกแทนด้วยค่า 0 หลังจากทำการเติมค่าที่หายไป เกือบ 90% ซึ่งเมื่อลองหักออกไปแล้ว ทุกระยะจะมีการกระจายตัวที่ค่อนข้างคงที่อยู่ที่ประมาณ 6 - 8%

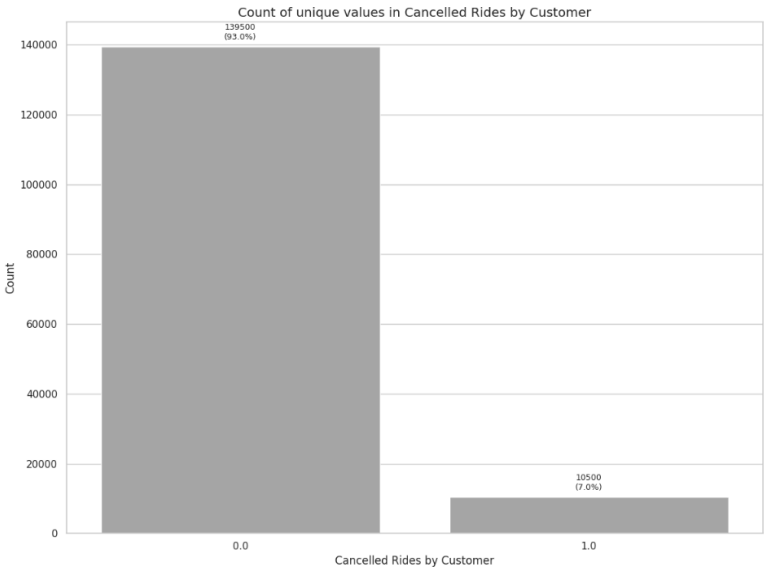


จากภาพเป็นการกระจายตัวของ Booking Value จะเห็นว่า ข้อมูลส่วนใหญ่ 73.9% จะอยู่ในช่วง 0 - 500 และอีก 20.72% จะอยู่ในช่วง 500-1000 ซึ่งอาจจะบอกได้ว่า รายการจองของ Uber ส่วนใหญ่เป็นการเดินทางที่มีราคาต่ำ ซึ่งอาจจะต้องดูปัจจัยอื่นๆนอกจาก Ride Distance ในการประเมิณ เช่น Vehicle Type

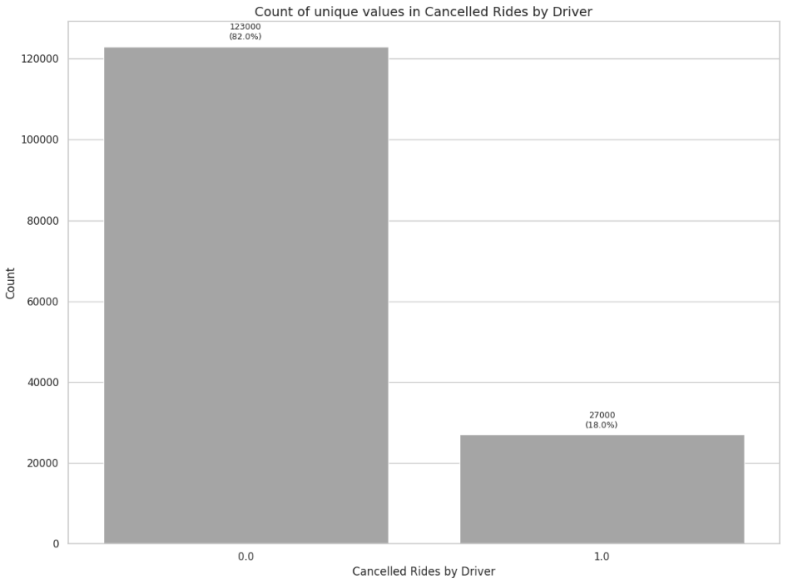
Binary Features :

Binary Features คือชุดข้อมูลที่มีค่าได้เพียงสองสถานะ เช่น 0/1 หรือ Yes/No ในชุดข้อมูลนี้ ตัวแปรกลุ่มนี้ถูกใช้เพื่อระบุเหตุการณ์ที่เกิดขึ้นหรือไม่เกิดขึ้น เช่น Cancelled Rides by Customer, Cancelled Rides by Driver, และ Incomplete Rides การวิเคราะห์ในส่วนนี้มุ่งเน้นเพื่อทำความเข้าใจ สัดส่วนของเหตุการณ์ ว่ามีการยกเลิกหรือการเดินทางที่ไม่สมบูรณ์เกิดขึ้นบ่อยเพียงใด รวมถึงตรวจสอบ ความไม่สมดุลของข้อมูล (Imbalanced Data) ซึ่งอาจส่งผลให้โมเดลเรียนรู้ลำเอียงไปทางกลุ่มที่มีจำนวนมากกว่า

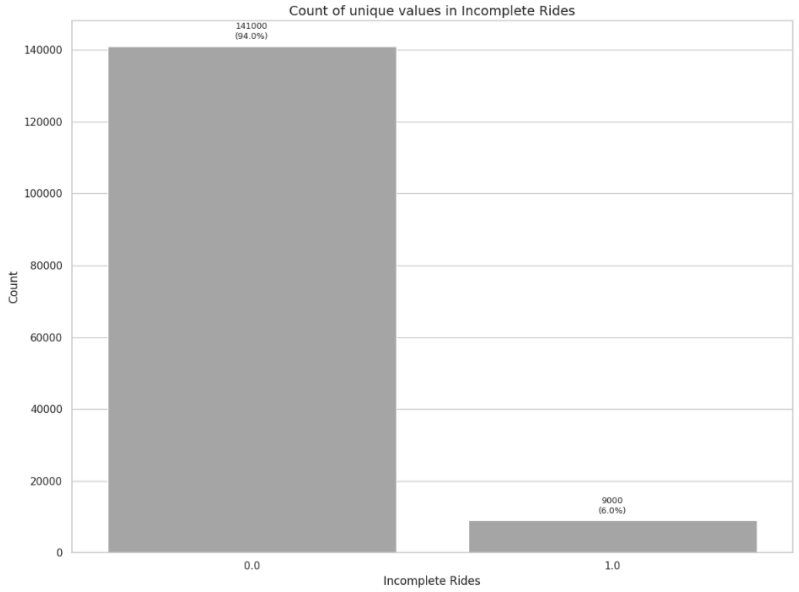
โดยใช้กราฟแท่ง เพื่อเปรียบเทียบจำนวนของแต่ละสถานะในแต่ละตัวแปร



จากภาพจะเห็นสัดส่วนการยกเลิกโดย Customer ซึ่ง 7% ของข้อมูลทั้งหมด อยู่ที่ค่า 1 หมายถึงมีแค่ส่วนน้อยของลูกค้าเท่านั้นที่ยกเลิกการเดินทาง



จากภาพจะเห็นสัดส่วนการยกเลิกโดย Driver ซึ่ง 18% ของข้อมูลทั้งหมด อยู่ที่ค่า 1 หมายถึงจากการจองทั้งหมด มี 18% ที่เป็นการยกเลิกโดย Driver



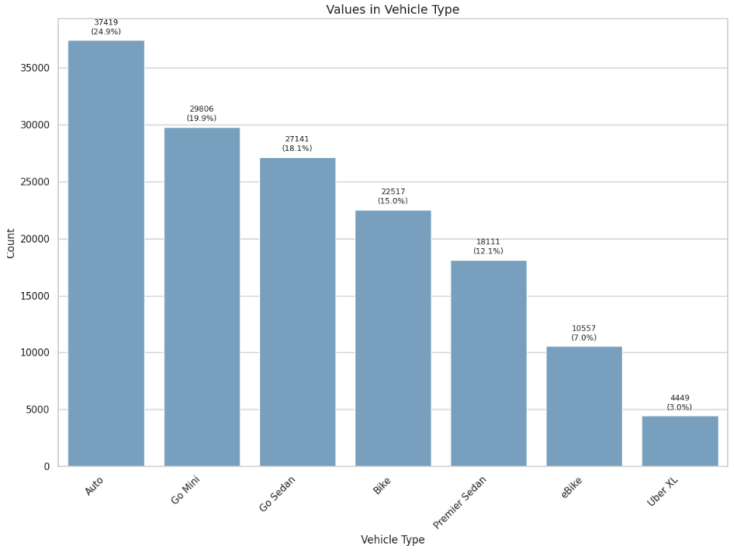
จากภาพจะเห็นสัดส่วนการเดินทางที่ไม่สำเร็จ 6% ของข้อมูลทั้งหมด อยู่ที่ค่า 1 หมายถึง จากข้อมูลการจองทั้งหมด มี 6% ที่เดินทางไม่สำเร็จ

Catagorical Features :

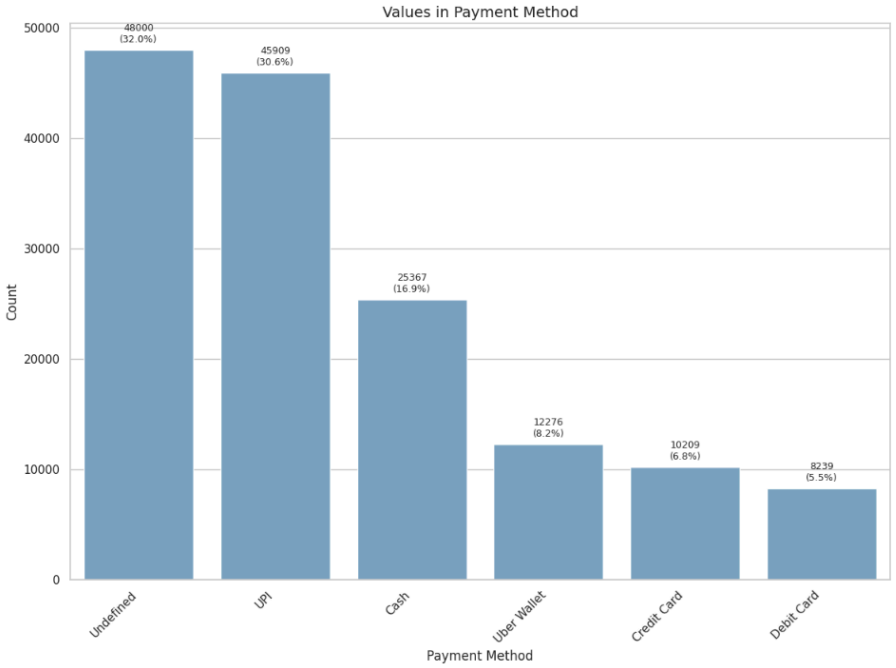
Categorical Features คือข้อมูลที่แสดงถึง หมวดหมู่ ของข้อมูล โดยไม่สามารถนำมาคำนวณทางคณิตศาสตร์ได้โดยตรง ในชุดข้อมูลนี้ ตัวแปรประเภทนี้ได้แก่

Vehicle Type, Booking Status, Payment Method, Reason for cancelling by Customer, Driver Cancellation Reason และ Incomplete Rides Reason

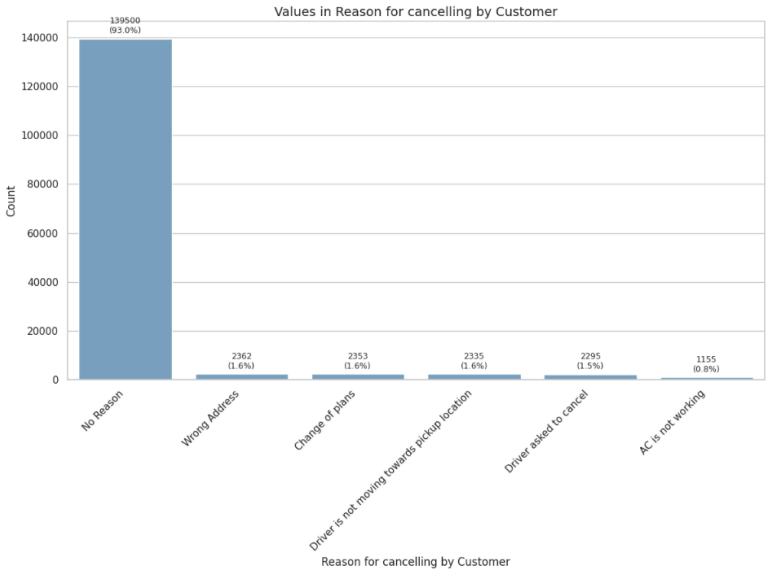
การวิเคราะห์ข้อมูลในส่วนนี้มุ่งเน้นเพื่อดู ความถี่ของแต่ละหมวดหมู่ และเพื่อระบุหมวดหมู่ที่ได้รับความนิยมสูงสุด กราฟที่ใช้ในการนำเสนอ ได้แก่ Bar Chart และ Count Plot เพื่อช่วยให้เห็นการกระจายของแต่ละประเภทได้อย่างชัดเจน

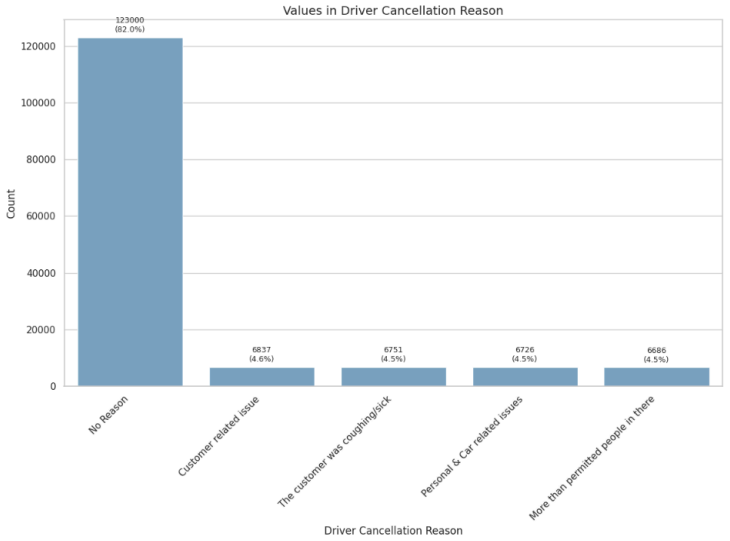


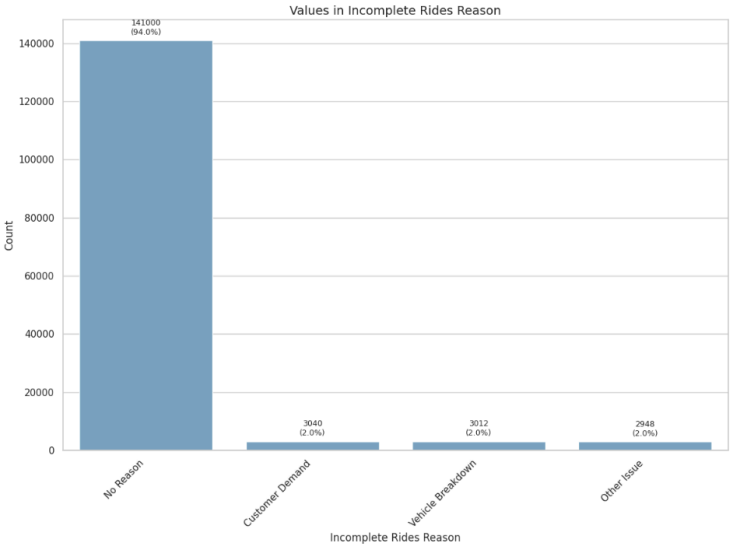
จากกราฟจะพบว่า Auto เป็นประเภทที่ได้รับความนิยมสูงสุด รองลงมาคือ Go Mini และ Go Sedan ส่วนประเภทพรีเมียมและขนาดใหญ่ เช่น Premier Sedan, Uber XL มีสัดส่วนน้อยกว่า



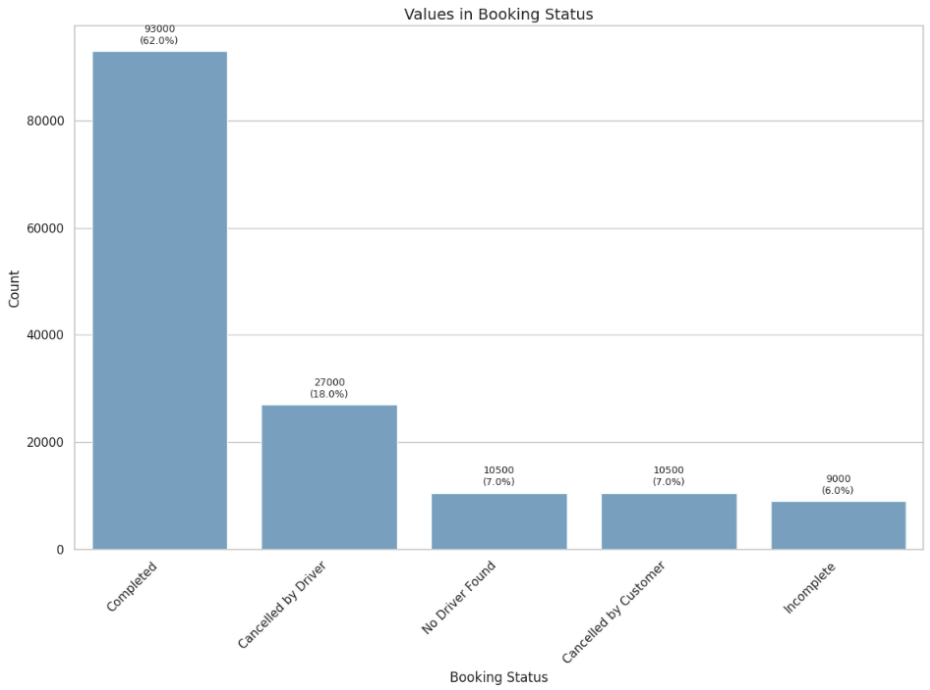
จากกราฟจะเห็นว่าข้อมูลส่วนใหญ่ อยู่ที่ Undefined และ UPI ซึ่งค่าของ Undefined จะเป็นกรณีที่การจองไม่สำเร็จทั้งหมด ซึ่งเราก็สามารถบอกได้ว่า ระบบชำระเงินแบบดิจิทัล UPI ได้รับความนิยมสูงสุด







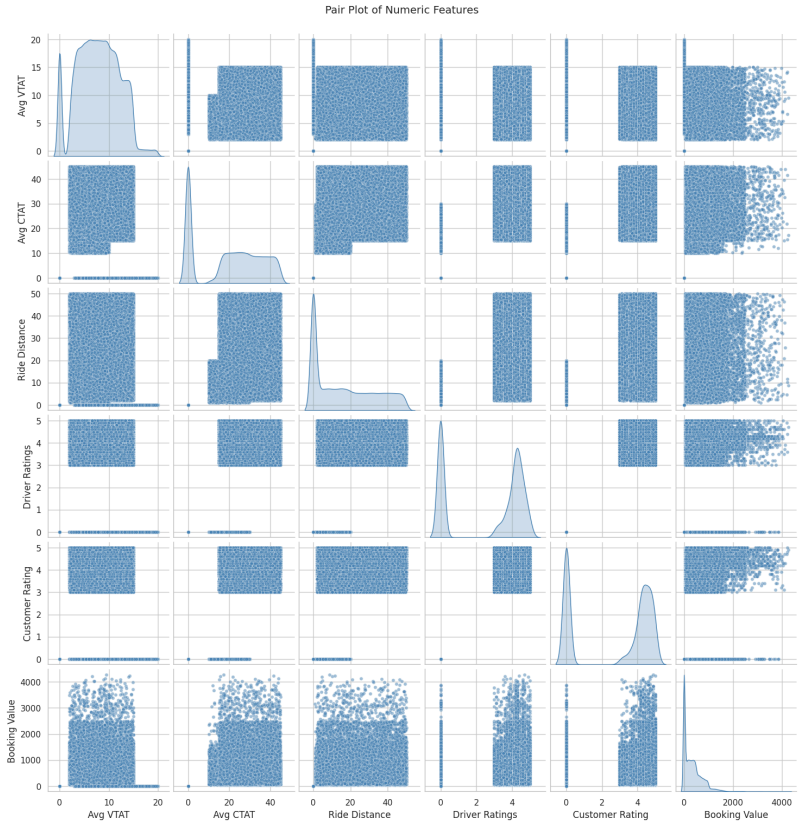
จากกราฟทั้ง 3 จะบอกได้ว่าข้อมูลส่วนใหญ่ 80% ขึ้นไป อยู่ในส่วนของ No Reason ซึ่งเป็ค่าที่ถูกแทนที่ หลังจากทำการเติมค่าที่หายไป ซึ่งจะบอกว่าการเดินทางส่วนใหญ่สำเร็จ



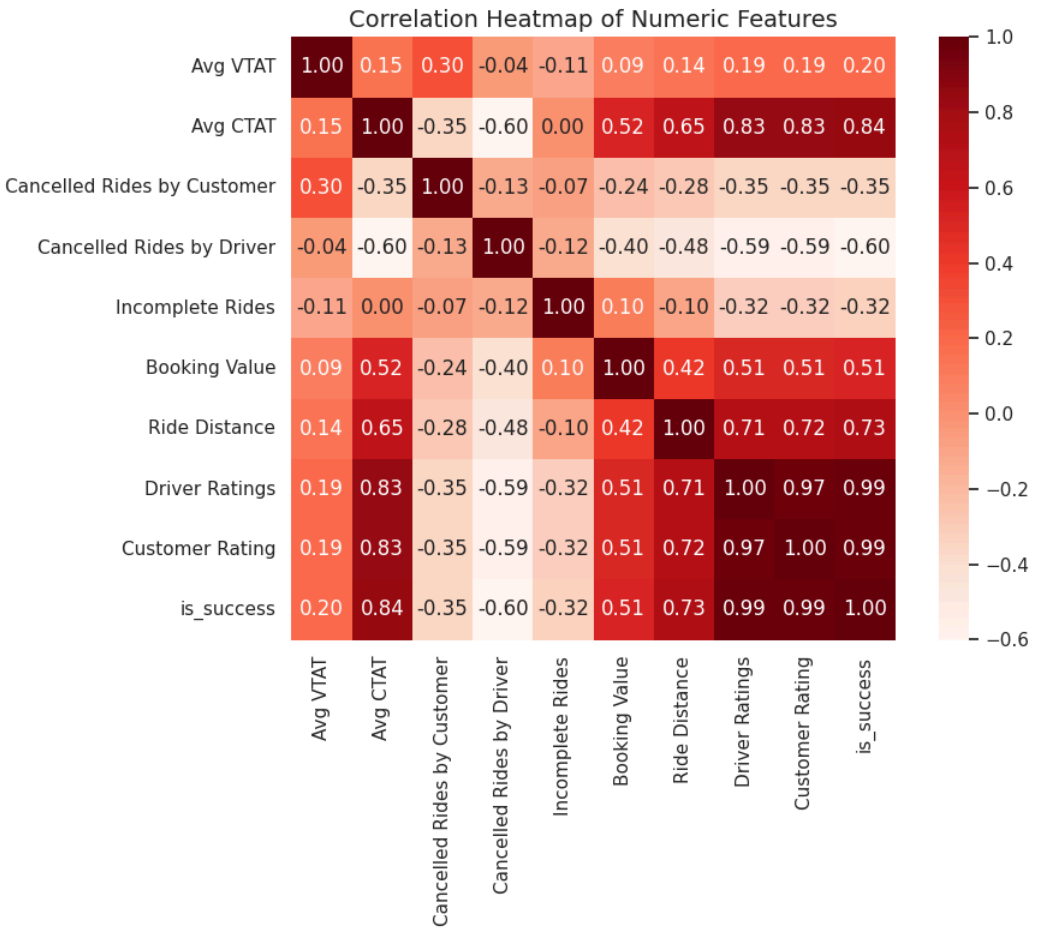
จากกราฟจะพบว่าข้อมูลส่วนใหญ่กว่า 62% เป็น Completed หมายถึงการเดินทางสำเร็จ แต่สัดส่วนการยกเลิกโดยคนขับยังถือว่าสูงพอสมควร ซึ่งอาจบ่งบอกถึงปัญหาในฝั่งการจัดการคนขับ เช่น การจับคู่หรือระยะทางรับผู้โดยสาร

Relationship between Features :

จากการวิเคราะห์ความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลขโดยใช้ Pair Plot และ Correlation Heatmap พบว่าข้อมูลในชุดนี้ ไม่มีความสัมพันธ์เชิงเส้นที่ชัดเจน ระหว่างตัวแปรส่วนใหญ่ ซึ่งสะท้อนถึงลักษณะของข้อมูลที่ซับซ้อนและมีความแปรปรวนสูงในแต่ละตัวแปร



จาก Pair Plot of Numeric Features ซึ่งแสดงการกระจายและความสัมพันธ์ของตัวแปรเชิงตัวเลขทุกคู่ พบว่า จุดข้อมูลส่วนใหญ่ ไม่แสดงแนวโน้มเชิงเส้นระหว่างกันอย่างชัดเจน เช่น ความสัมพันธ์ระหว่าง Booking Value กับ Ride Distance, Avg CTAT กับ Driver Ratings, หรือ Avg VTAT กับ Customer Rating ล้วนแสดงการกระจายแบบกระจายกว้าง ไม่มีเส้นแนวโน้มชัดเจน การกระจายของค่าหลายคอลัมน์ (เช่น Avg VTAT, Avg CTAT, Ride Distance) มีลักษณะ กระจุกตัวในช่วงค่าต่ำ (right-skewed) แสดงว่าการเดินทางส่วนใหญ่มีระยะเวลาและระยะทางไม่ยาวมาก และถ้าดู จาก Correlation Heatmap ซึ่งใช้วัดระดับความสัมพันธ์ระหว่างตัวแปรเชิงตัวเลข

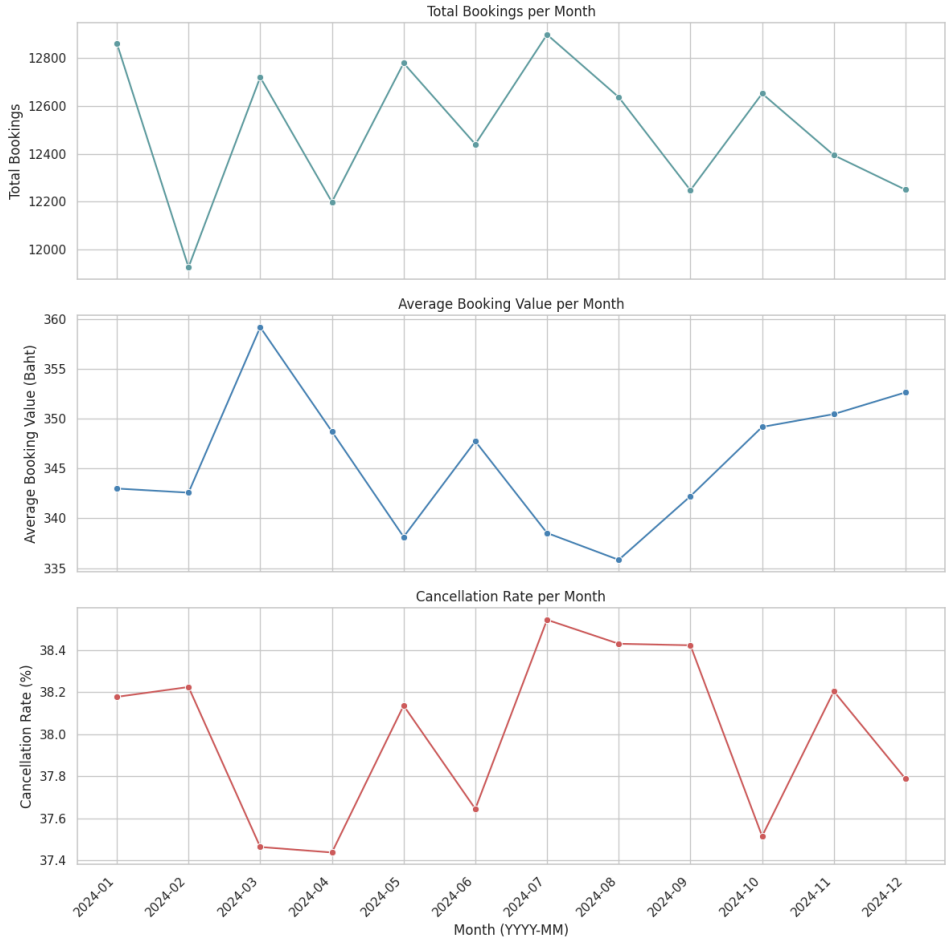


ค่าสัมประสิทธิ์ความสัมพันธ์ส่วนใหญ่มีค่าน้อยกว่า 0.5 ทั้งในทางบวกและลบ แสดงให้เห็นว่า คู่ตัวแปรส่วนใหญ่ไม่มีความสัมพันธ์เชิงเส้นที่แข็งแรง แต่ก็มีคู่ตัวแปรที่มีความสัมพันธ์เชิงเส้นกันบ้าง เช่น

Driver Ratings กับ Customer Rating คะแนนทั้งสองมีแนวโน้มไปในทิศทางเดียวกัน ซึ่งเป็นสิ่งที่สมเหตุสมผลในเชิงพฤติกรรมการให้บริการ

Driver Ratings, Customer Rating กับ AVG CTAT ซึ่งจะตีความได้ว่า ยิ่งระยะเวลาในการเดินทางมาก คะแนนทั้งสองมีแนวโน้มที่จะเพิ่มขึ้น

Driver Ratings, Customer Rating กับ Rider Distance ซึ่งจะตีความได้ว่า ยิ่งระยะทางในการเดินทางไกล คะแนนทั้งสองมีแนวโน้มที่จะเพิ่มขึ้น



	Month	total_bookings	avg_booking_value	avg_distance	cancel_rate
0	2024-01	12861	342.980250	16.740923	38.177436
1	2024-02	11927	342.566446	16.499969	38.224197
2	2024-03	12719	359.162513	16.926141	37.463637
3	2024-04	12199	348.699811	16.803902	37.437495
4	2024-05	12778	338.134215	16.555120	38.135859
5	2024-06	12440	347.721865	16.925898	37.644695
6	2024-07	12897	338.522370	16.670147	38.543847
7	2024-08	12636	335.826923	16.513798	38.429883
8	2024-09	12248	342.210402	16.661079	38.422600
9	2024-10	12651	349.155798	17.023503	37.514821
10	2024-11	12394	350.443360	16.775997	38.203970
11	2024-12	12250	352.621878	16.937714	37.787755

จากภาพจะเป็น Time-Based Analysis แบบ Monthly Trends โดยแสดงแนวโน้มรายเดือนของ

1. Total Bookings
2. Average Booking Value
3. Cancellation Rate

ซึ่งจะได้ว่า ตลอดทั้งปีแนวโน้มการจองค่อนข้างคงที่ โดยถ้าช่วงไหนที่มี การจองมากและ อัตราการยกเลิกต่ำ มูลค่าการจองเฉลี่ยก็จะเพิ่มขึ้นตามไปด้วย และ อัตราการยกเลิกค่อนข้างคงที่ จะอยู่ในช่วง 37-39% ตลอดทั้งปี ซึ่งบ่งชี้ว่าปัญหาการยกเลิกอาจไม่ได้ขึ้นอยู่กับฤดูกาล แต่เกิดจากปัจจัยภายในระบบ

Training model :

กลุ่มของเราเลือกใช้เป็น **XGBoost (Extreme Gradient Boosting)** แนวคิดหลักคือ “สร้างต้นไม้หลาย ๆ ต้นแบบลำดับต่อกัน โดยให้แต่ละต้นใหม่ช่วยแก้ข้อผิดพลาดของต้นก่อนหน้า” แนวคิดคล้ายๆการทำ backpropagation ใน neural network โดย XGBoost เป็นหนึ่งในโมเดลที่ “แม่นยำ” และ “เร็ว” ในสาย tree-base algorithms

Training Step :

1. **Data preprocessing**
    a. แทนค่าลงในข้อมูลที่เป็น null แบ่งเป็น 3 ประเภท
        i. แทนค่าด้วย 0 สำหรับข้อมูลที่เป็น binary

```
# impute with zero
X['Avg VTAT'].fillna(0, inplace=True)
X['Avg CTAT'].fillna(0, inplace=True)
X['Booking Value'].fillna(0, inplace=True)
X['Ride Distance'].fillna(0, inplace=True)
X['Driver Ratings'].fillna(0, inplace=True)
X['Customer Rating'].fillna(0, inplace=True)
X['Cancelled Rides by Customer'].fillna(0, inplace=True)
X['Cancelled Rides by Driver'].fillna(0, inplace=True)
X['Incomplete Rides'].fillna(0, inplace=True)
```

        ii. แทนค่า “Unknow” สำหรับข้อมูล Payment method ที่หายไปเพราะคิดว่า Payment method อาจส่งผลต่อการทำนาย

```
X['Payment Method'].fillna("Unknown", inplace=True)
```

        iii. ลบทิ้ง 5 ข้อมูลที่คาดว่าไม่มีผลต่อการทำนาย
            1. เหตุผลที่ลูกค้ายกเลิกการจอง
            2. เหตุผลที่ผู้ขับยกเลิกการจอง
            3. เหตุผลที่รายการนั้นไม่สำเร็จ
            4. customerID + bookingID

b. เปลี่ยน format เวลา

```
X['Date'] = pd.to_datetime(X['Date'], errors='coerce')
X['Time'] = pd.to_datetime(X['Time'], errors='coerce')

X['Hour'] = X['Time'].dt.hour
X['Weekday'] = X['Date'].dt.weekday
X['IsWeekend'] = X['Weekday'].isin([5,6]).astype(int)

X.drop(columns=['Date', 'Time'], inplace=True)
```

c. Scaling numerical data + One hot encoding categorical data

```
preprocessor = ColumnTransformer(
    transformers=[
        ('num', StandardScaler(), num_cols),
        ('cat', OneHotEncoder(handle_unknown='ignore', sparse_output=False), cat_cols)
    ]
)
```

d. ทำ oversampling ด้วย smote `('smote', SMOTE(random_state=42,)),`

2. Train model

a. แบ่งข้อมูลเป็น 2 ส่วน

i. Training data = 80%

ii. Testing date = 20%

```
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42, stratify=y
)
```

b. นำข้อมูลหลังจากการทำ preprocessing มา train model XGBoost
โดยปรับค่าพารามิเตอร์ดังนี้

i. n estimators คือจำนวนต้นไม้ทั้งหมดที่จะถูกสร้างขึ้นมา = 50
ii. learning rate คือค่าคูณเพื่อปรับค่า parameter กำหนดความเร็วในการเรียนรู้ของโมเดล = 0.1
iii. max depth คือความลึกมากสุดของต้นไม้ = 5
iv. subsample คืออัตราส่วน sample ที่หยิบมาพิจารณาใน tree = 0.8
v. col sample by tree คืออัตราส่วน features ที่หยิบมาพิจารณาใน tree = 0.8

```
('model', XGBClassifier(
    n_estimators=50,
    learning_rate=0.1,
    max_depth=5,
    subsample=0.8,
    colsample_bytree=0.8,
    random_state=42
))
```

vi. random_state = 42 seed ของการสุ่ม

c. train model + test on testing set

```
pipeline.fit(X_train, y_train)
y_pred = pipeline.predict(X_test)
```

3. Model evaluation

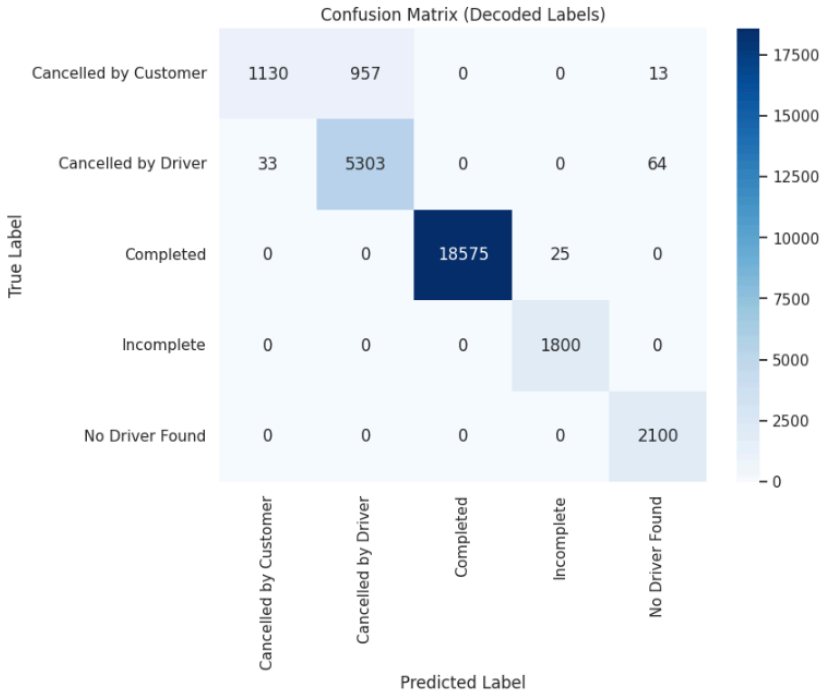
a. Accuracy = 96.36%

b. Classification report

Classification Report:

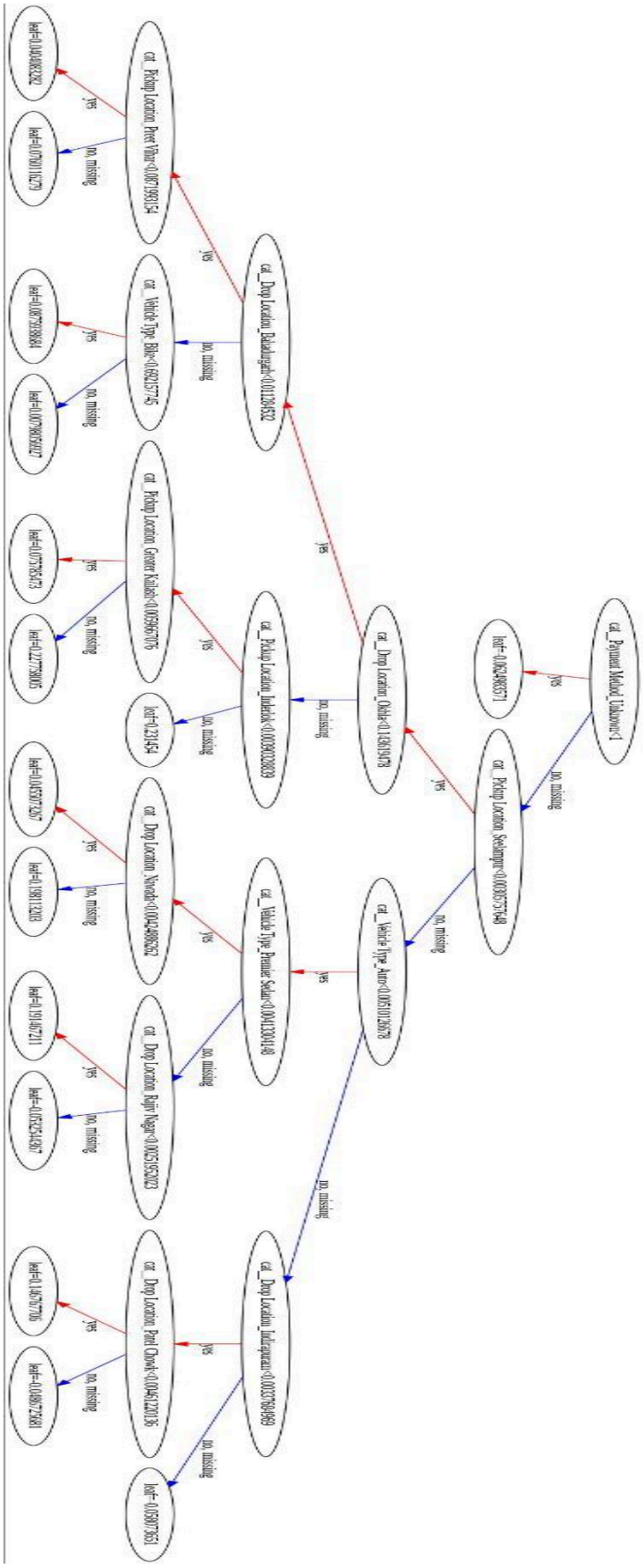
	precision	recall	f1-score	support
Cancelled by Customer	0.97	0.54	0.69	2100
Cancelled by Driver	0.85	0.98	0.91	5400
Completed	1.00	1.00	1.00	18600
Incomplete	0.99	1.00	0.99	1800
No Driver Found	0.96	1.00	0.98	2100
accuracy			0.96	30000
macro avg	0.95	0.90	0.92	30000
weighted avg	0.97	0.96	0.96	30000

c. Confusion matrix



จะเห็นว่าจะมีแค่บาง class เท่านั้นที่มีค่าน้อยกว่า class อื่นเล็กน้อยcancel by customer มีค่า recall = 54% จาก confusion matrix จะพบว่ามีการทำนายผิดไป 970 จาก 2100 ตัวอย่าง หรือ cancel by driver มีค่า precision = 85% เพราะมีค่าที่ทายผิด 97 จาก 5400 ตัวอย่าง

d. Result tree



Possible application : จากโมเดลที่ได้ทำออกมาสามารถนำไปใช้ได้ดังนี้

1. Predict ride outcome ตรงตัวที่สุดใช้ประโยชน์ได้ในระบบ
    a. แจ้งเตือนระบบล่วงหน้าว่าการจองนี้ “มีแนวโน้มจะถูกยกเลิก
    b. ช่วยให้ platform เตรียมหาคนขับสำรองทันที
    c. ช่วยลด cancellation rate และเพิ่ม ride completion rate
2. Dynamic pricing & demand forecasting (ราคาขึ้นลงตามโอกาสยกเลิก) เช่น ถ้าช่วงเวลา 18:00–20:00 น. ใน “Downtown” มีแนวโน้มยกเลิกสูง
   ระบบอาจปรับราคาเพื่อคงสมดุล

Future Work:

1. Classification ⟶ Probability forecasting ทำนายว่าการจองนั้นๆมีโอกาสที่จะถูกยกเลิกกี่เปอร์เซ็นต์
2. Time Series / Sequential model ตอนนี้โมเดล XGBoost ยังไม่เข้าใจถึงลำดับเวลาโดยตรงสามารถใช้ LSTM, Prophet

References:

XGBoost library document :
https://xgboost.readthedocs.io/en/stable/
Sklearn library document : https://scikit-learn.org/0.21/documentation.html
Kaggle Uber dataset :
https://www.kaggle.com/datasets/yashdevladdha/uber-ride-analytics-dashboard/data?select=ncr_ride_bookings.csv