MA705

Data science Final Project

Customer Value Analysis

—— Carol Yu

Content

- Dataset Introduction
- Project Goals
- Executive Summaries
- Math and Python behind

Dataset Introduction

✓

Data Source

From Kaggle, the final dataset is merged with two data frames, namely, ***customers*** and ***orders.***

✓

Data Overviews

After preprocessing, the suitable final dataset consists of 16 variables and 161,581 rows, contains KPIs reflecting customer ordering behaviors and demographic information.

✓

Data Justification

Real word marketing data sourced from Instacart, an online delivery service, is ideal for addressing and exploring real-world business questions.

Content

- Dataset Introduction
- Project Goals
- Executive Summaries
- Statistic and math behind

Project Goals

1
Data explorations and visualizations to identify patterns and trends.

- Understand the overview of customers.
- Know the busiest days of the week and hours of the day to perform marketing strategies timely.

2
Applying logistic regression to identify the factors that impact customer value.

- Utilizing a binary variable to categorize customer value according to their historical order behavior.
- identify the factors that impact customer value.

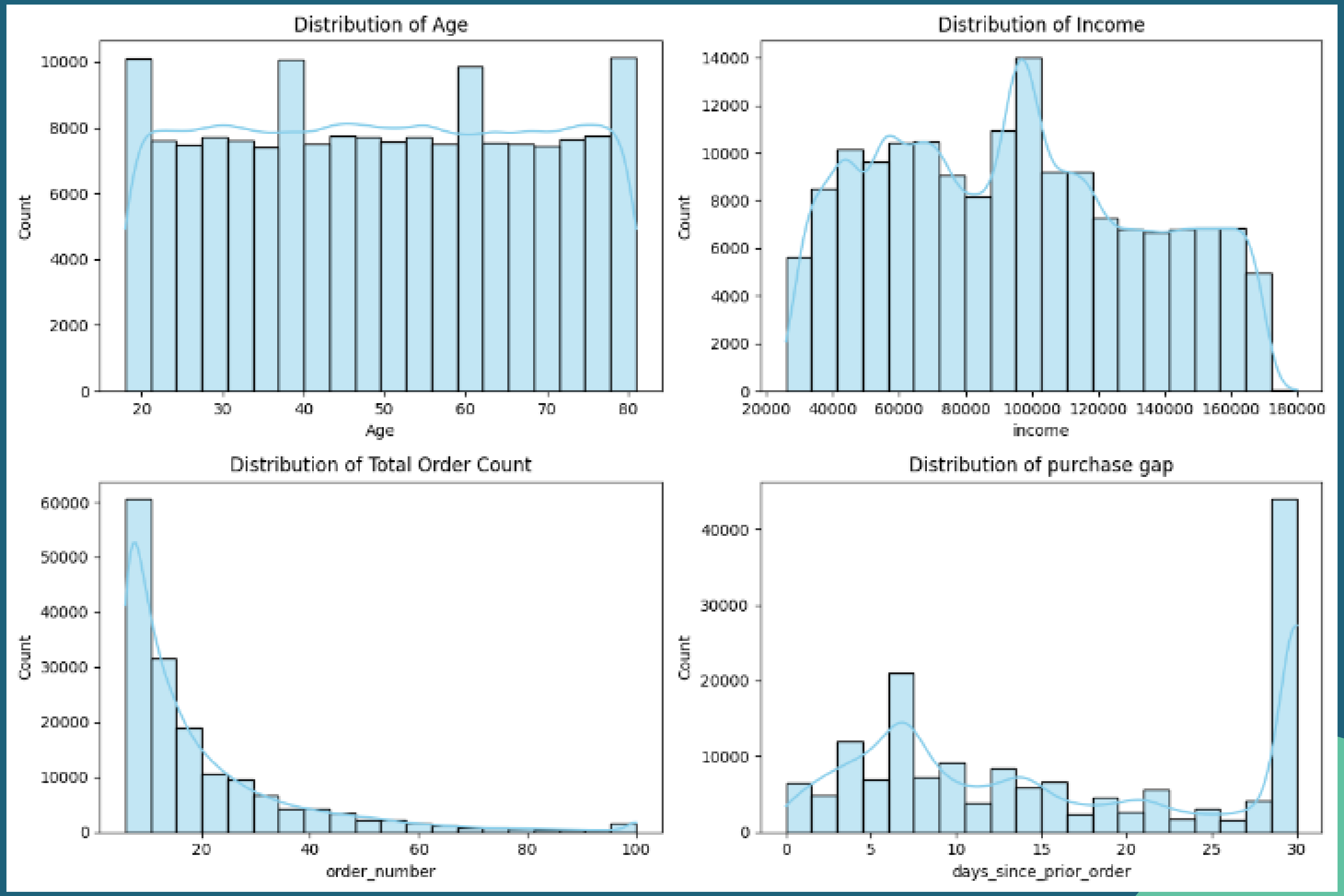
Content

- Dataset Introduction
- Project Goals
- Executive Summaries
- Statistic and math behind

Executive Summaries for Goal 1

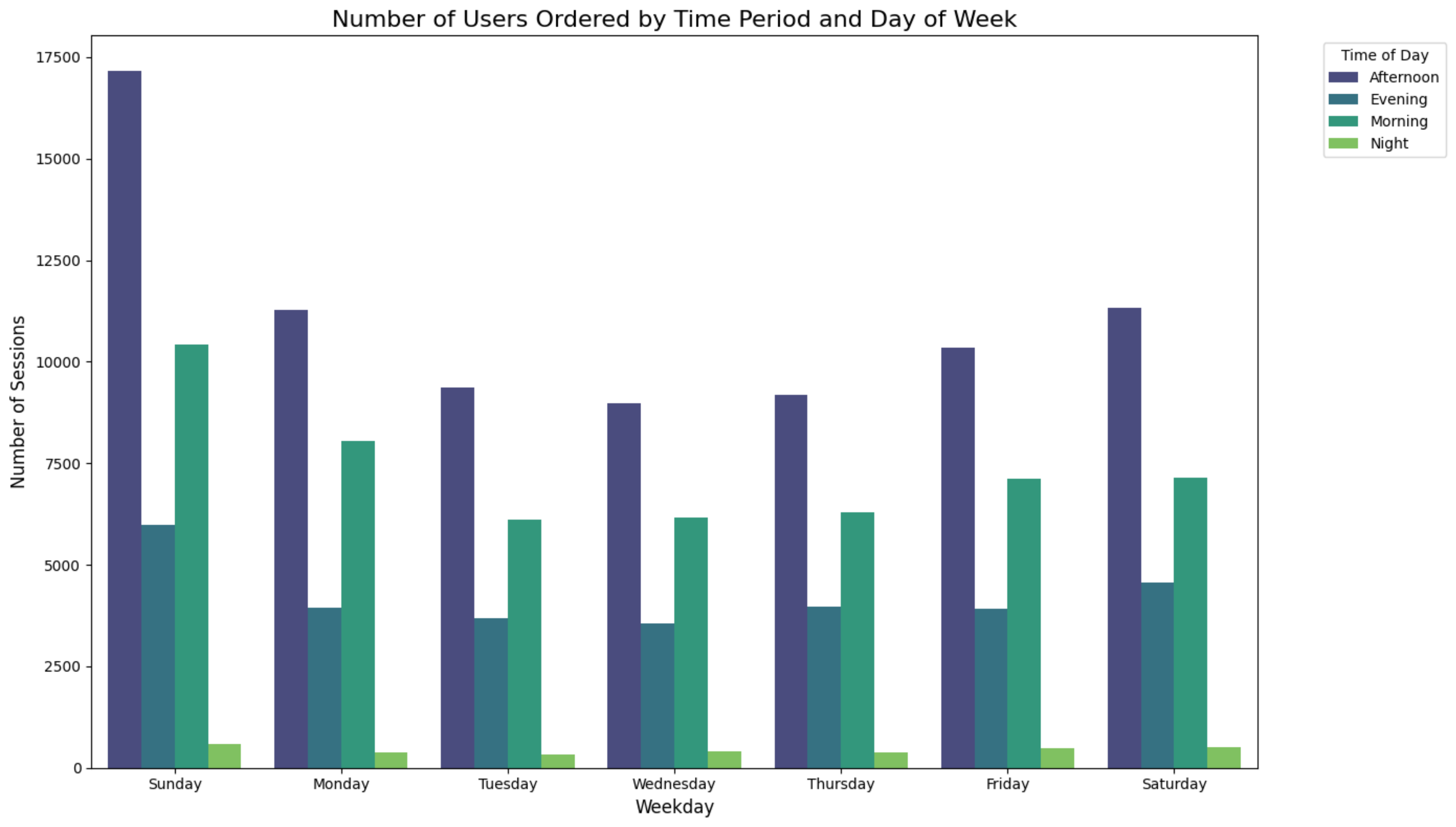
Understand the overview of customers

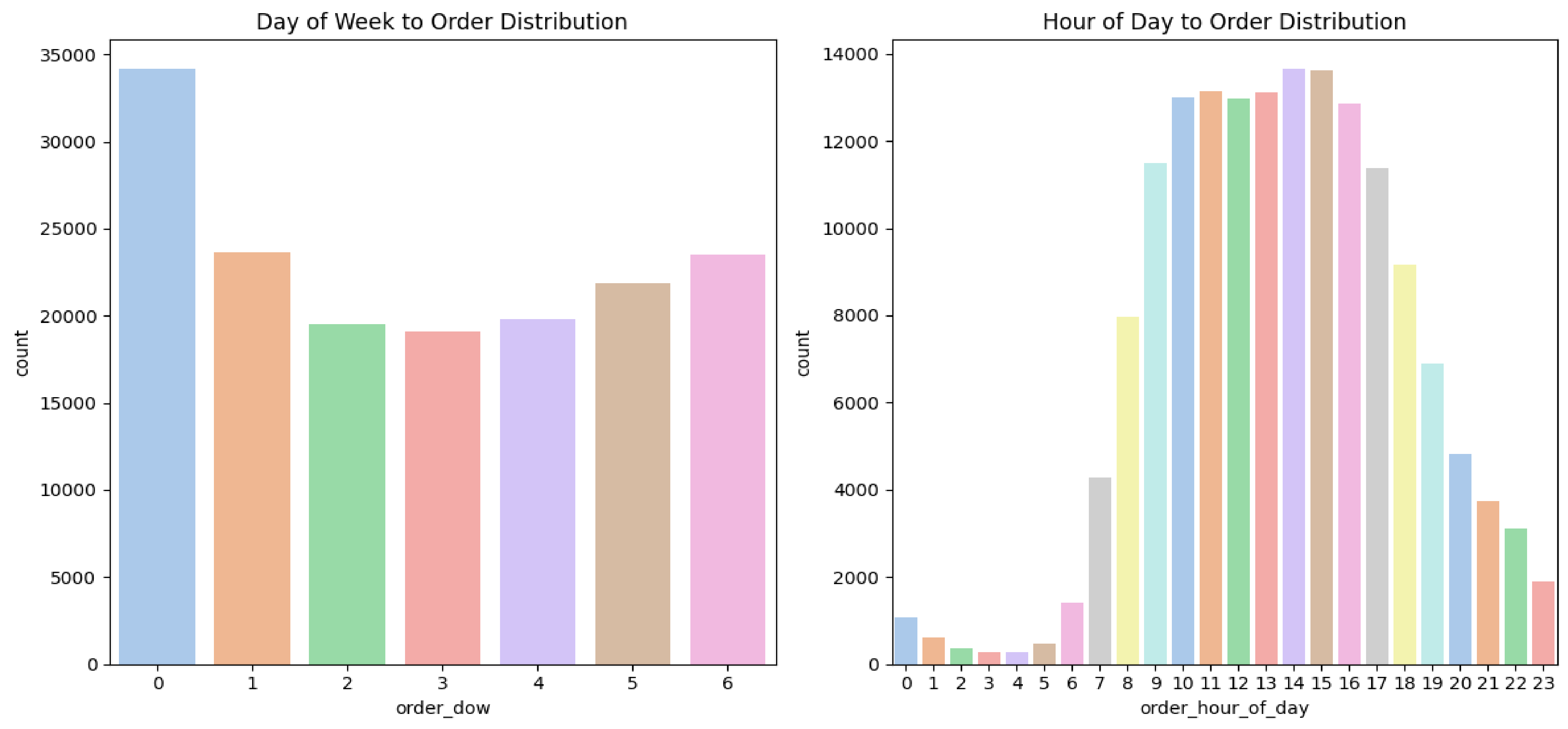
- ➢ The age range of Instacart customers spans from young to old.
- ➢ Customers' annual income levels exhibit a wide range, from $20,000 to $180,000.
- ➢ Instacart has a significant customer base with small order amounts (less than 12 orders in total), encompassing almost half of the customer population.
- ➢ A substantial portion of customers also opts for a longer waiting period, exceeding 28 days before placing the next order, constituting nearly half of the customer population.



Executive Summaries for Goal 1

Know the busiest days of the week and hours of the day to perform marketing strategies timely

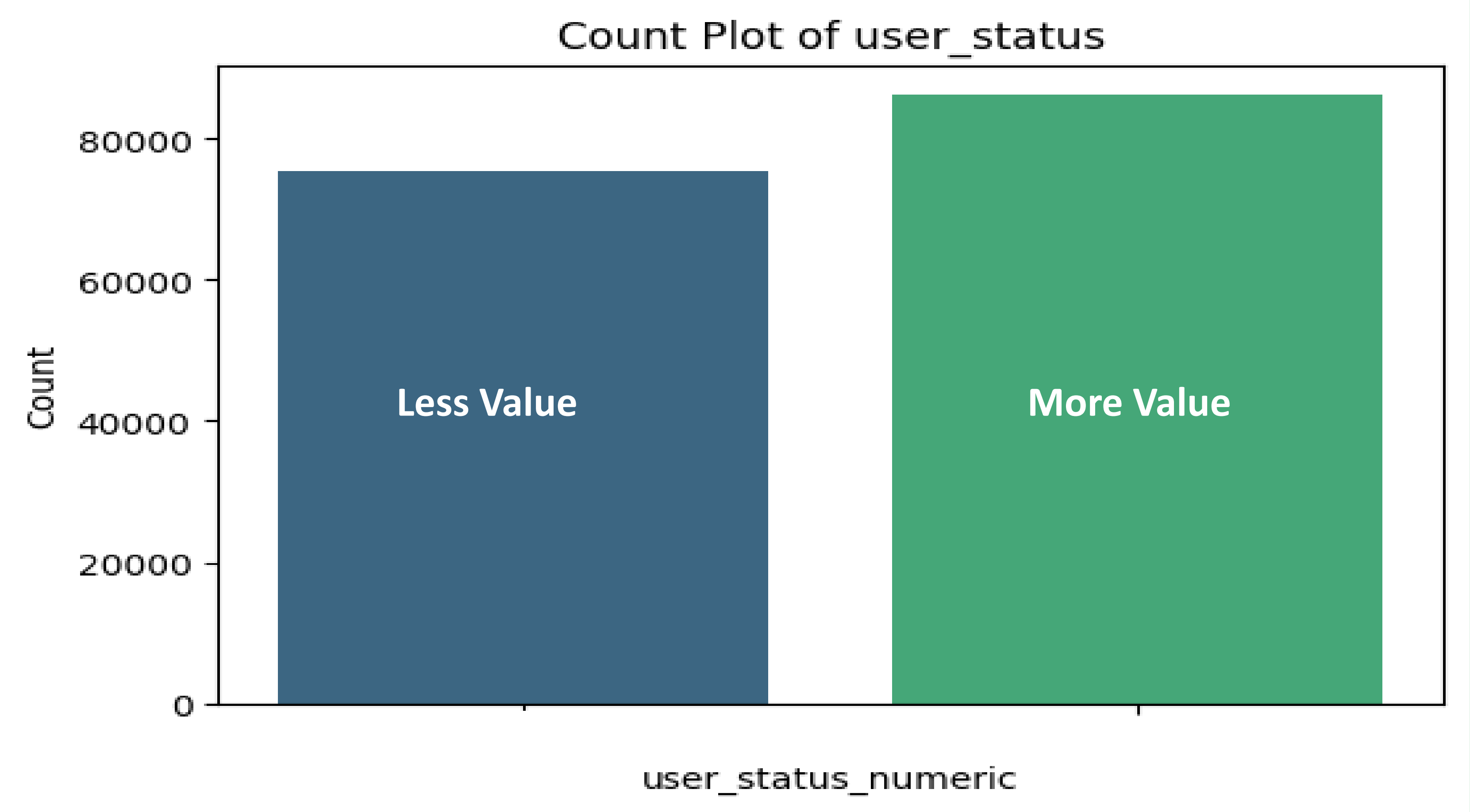


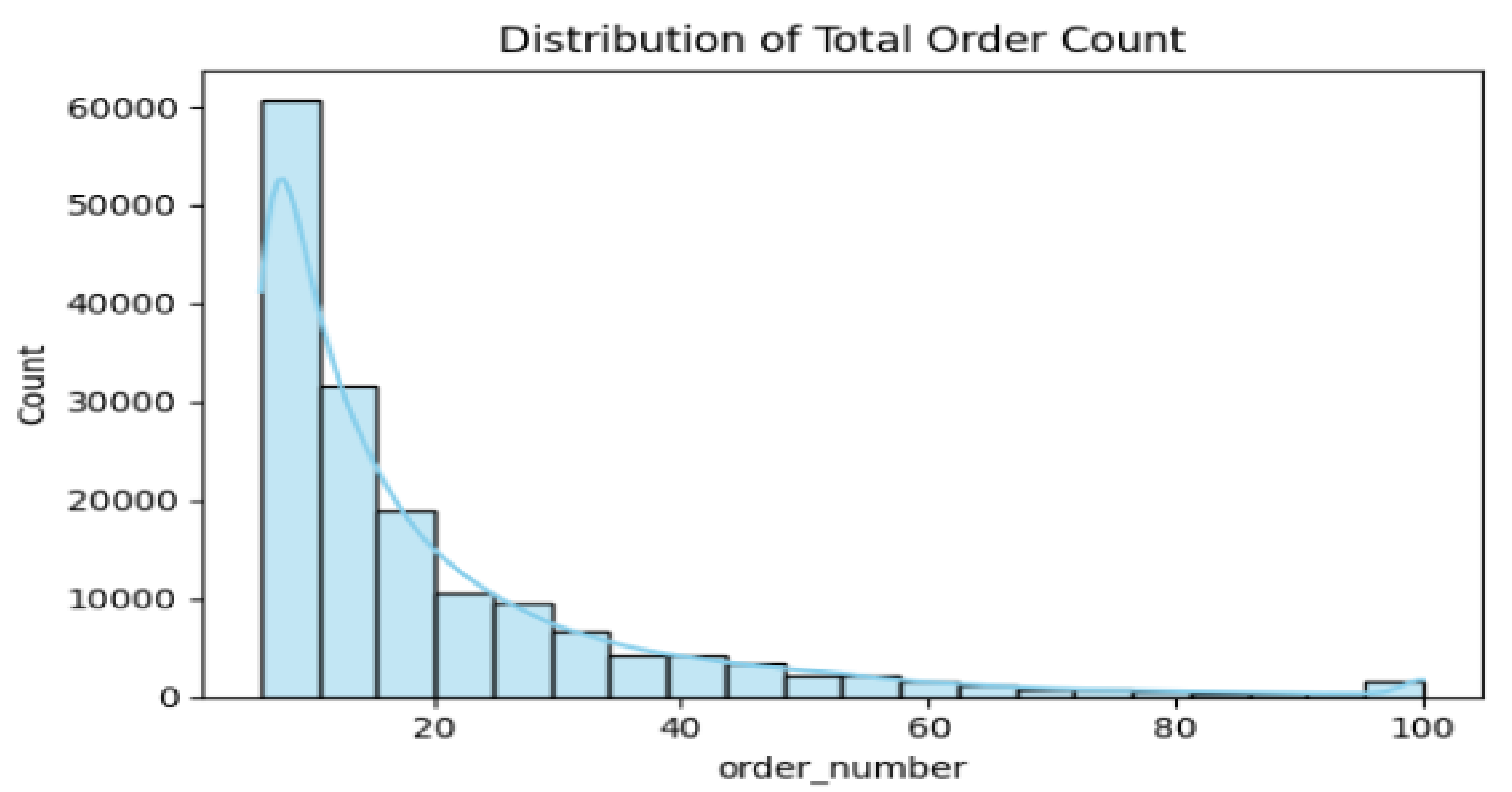


- **Sunday** experiences the highest level of order activity during the week, particularly in the afternoon between 12pm and 6pm. Followed with **Saturday and Monday**, during the same time slot .
- The Marketing team could schedule ads at times when there are fewer orders based on the findings

Executive Summaries for Goal 2

Utilizing a binary variable to categorize customer value according to their historical order behavior.





Categorizing customers based on their historical order count, those who have made more than 12 orders are labeled as more valuable, while those with fewer than 12 orders are categorized as less valuable for the logistic regression model fitting.

Executive Summaries for Goal 2

identify the factors that predict customer value in logistic regression model.

All factors have meaningful impacts on weather a customer can be identified as more valuable or less valuable for the business.

	coef
Age	0.0047
income	3.395e-06
n_dependants	0.1691
days_since_prior_order	-0.0540
fam_status_single	0.3963
Gender_Male	0.0806

- **When customer are a bit older**
- **When customer has higher income level**
- **When someone has more dependents (like family members)**

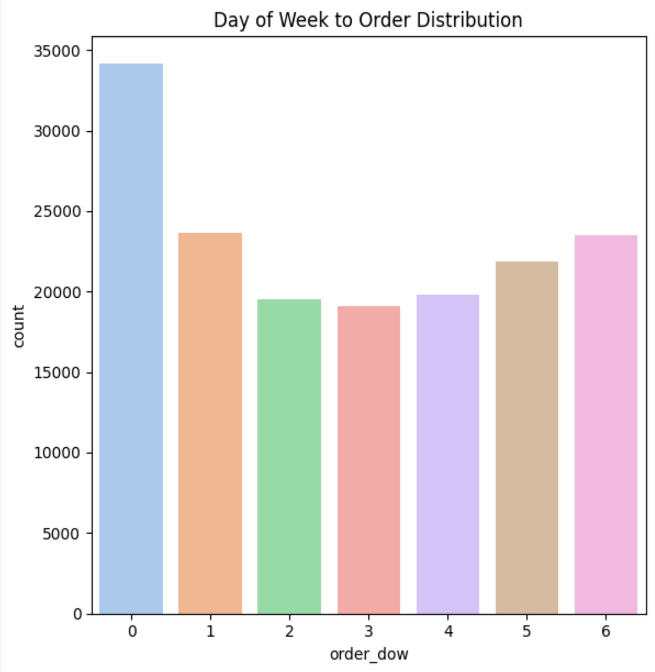
- **When customer take shorter to make the second order**
- **When customer is single**
- **When customer is a male**

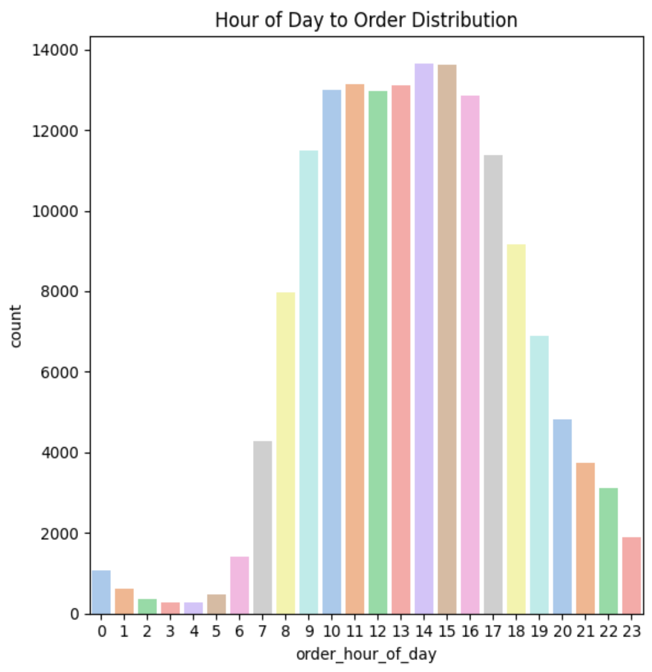
Code Notebook

what I performed to find out the busiest days of the week and hours of the day

The initial dataset includes two columns: "order_dow," indicating the day of the week when a customer placed an order, and "order_hour_of_day," indicating the specific time of day when the order occurred.

user_id int64	order_hour_of_day i..	days_since_prior_...
26711	16	22
33890	18	28
65803	9	16
125935	10	23

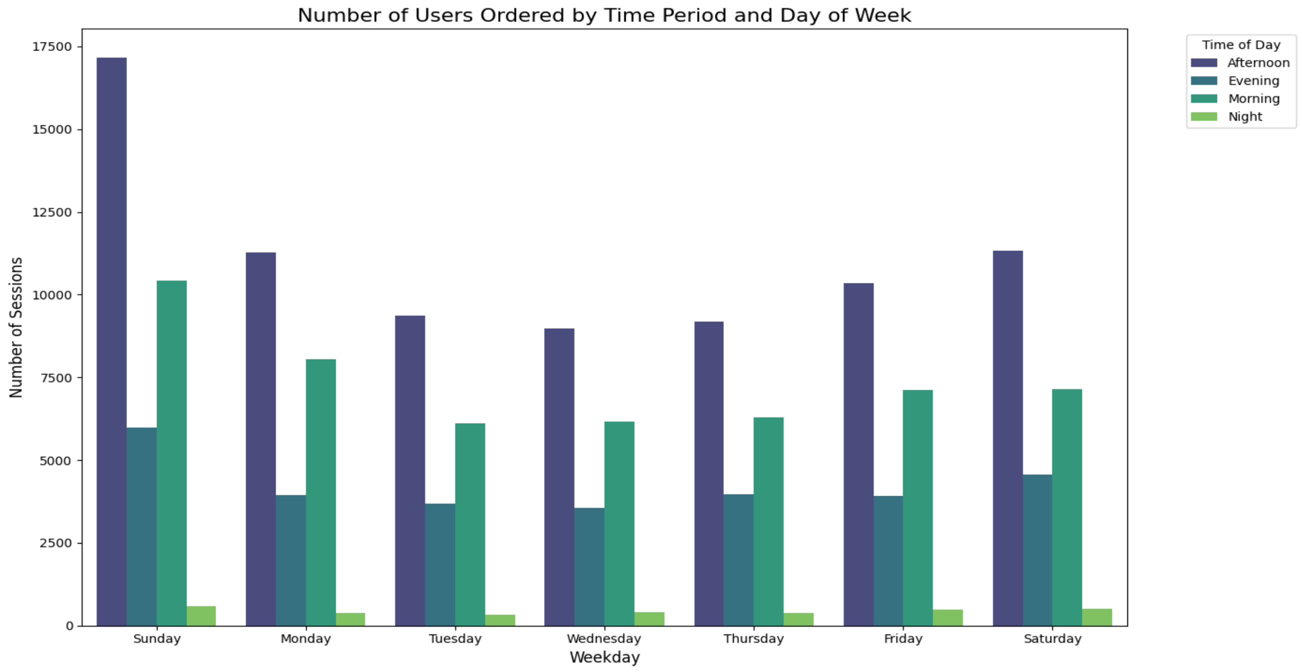




BEFORE

Through feature Engineering, I was able to convert order_hour_of_day to categorical variable time_period, and order_dow to specific day_of_week, both aggregated by the usercount on each time slot.

day_of_week object	time_period object	usercount int64
Friday	Afternoon	10333
Friday	Evening	3922
Friday	Morning	7115
Friday	Night	482



AFTER

Content

- Dataset Introduction
- Project Goals
- Executive Summaries
- Statistic and math behind

Data Merge

df_customers

	user_id int64	First Name object	Surnam object	Gender object	STATE object	Age int64
	Sorted asc...					
134..	1	Linda	Nguyen	Female	Alabama	31
318..	2	Norma	Chapman	Female	Alaska	68
36...	3	Janice	Fry	Female	Arizona	33
45...	4	Bobby	Reed	Male	Arkansas	31
112..	5	Janet	Lester	Female	California	75
100..	6	Alice	Blevins	Female	Colorado	48
156..	7	Peter	Villegas	Male	Connecticut	39
64...	8	Anna	Allison	Female	Delaware	32
192..	9	Nicole	Conrad	Male	District of Colum...	79
106..	10	Stephen	Oconnell	Male	Florida	34

df_orders

```
df_order = pd.read_csv( '/work/orders.csv' )
df_order
```

	order_id int64	user_id int64	eval_set object	order_number int64	order_dow int64	order_hour_of_day i..
		Sorted asc...				
0	2539329	1	prior	1	2	8
10	1187899	1	train	11	4	8
9	2550362	1	prior	10	4	8
7	3108588	1	prior	8	1	14
6	550135	1	prior	7	1	9
8	2295261	1	prior	9	1	16
4	431534	1	prior	5	4	15
3	2254736	1	prior	4	4	7
2	473747	1	prior	3	3	12
1	2398795	1	prior	2	3	7
5	3367565	1	prior	6	2	7

How and Why merge?

- Final dataset is merged on the shared column user_id of two data frames.
- Blending customer shopping behaviors with their demographic details to uncover additional insights for identifying their values.

EDA Check List

Data Collection:
Collected from Kaggle, merged on the unique key from both data frames

Data Cleaning:
Identify and handle missing values, outliers, and any inconsistencies in the dataset.
Perform data removing for irrelevant data points

Descriptive Statistics:
Compute statistical summaries to gain an overall understanding of the data's central tendencies and variability.

Univariate Analysis:
Examine and visualized individual variables one at a time.

Feature Engineering:
Create binary variable for logistic regression
Create aggregated variables for plot

EDA Check List

Data Cleaning:

	user_id int64 (Sorted asc...)	First Name object
134..	1	Linda
318..	2	Norma
36...	3	Janice
45...	4	Bobby
112..	5	Janet

206209 rows, showing 5 per page

	user_id int64 (Sorted asc...)	Gender object
134..	1	Female
318..	2	Female
36...	3	Female
45...	4	Male
156..	7	Male

161581 rows, showing 5 per page

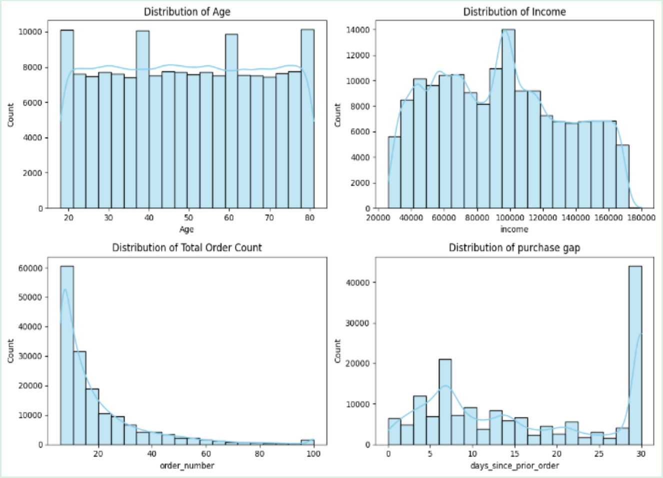
Descriptive Statistics:

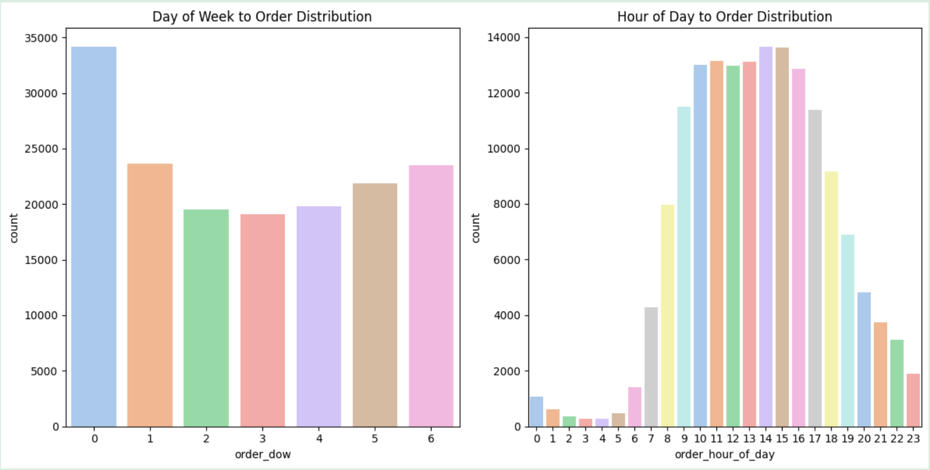
Summary Statistics:

	user_id	Age	n_dependants	income
count	206209.000000	206209.000000	206209.000000	206209.000000
mean	103105.000000	49.501646	1.499823	94632.852548
std	59527.555167	18.480962	1.118433	42473.786988
min	1.000000	18.000000	0.000000	25903.000000
25%	51553.000000	33.000000	0.000000	59874.000000
50%	103105.000000	49.000000	1.000000	93547.000000
75%	154657.000000	66.000000	3.000000	124244.000000
max	206209.000000	81.000000	3.000000	593901.000000

	order_number	order_dow	order_hour_of_day	days_since_prior_order
count	206209.000000	206209.000000	206209.000000	206209.000000
mean	16.590367	2.773957	13.585304	17.061782
std	16.654774	2.123616	4.221405	10.672178
min	4.000000	0.000000	0.000000	0.000000
25%	6.000000	1.000000	10.000000	7.000000
50%	10.000000	3.000000	14.000000	15.000000
75%	20.000000	5.000000	17.000000	30.000000
max	100.000000	6.000000	23.000000	30.000000

Univariate Analysis:





Feature Engineering:

user_status object	user_status_nume...
Lless_value	0
more_value	1
more_value	1
Lless_value	0
more_value	1

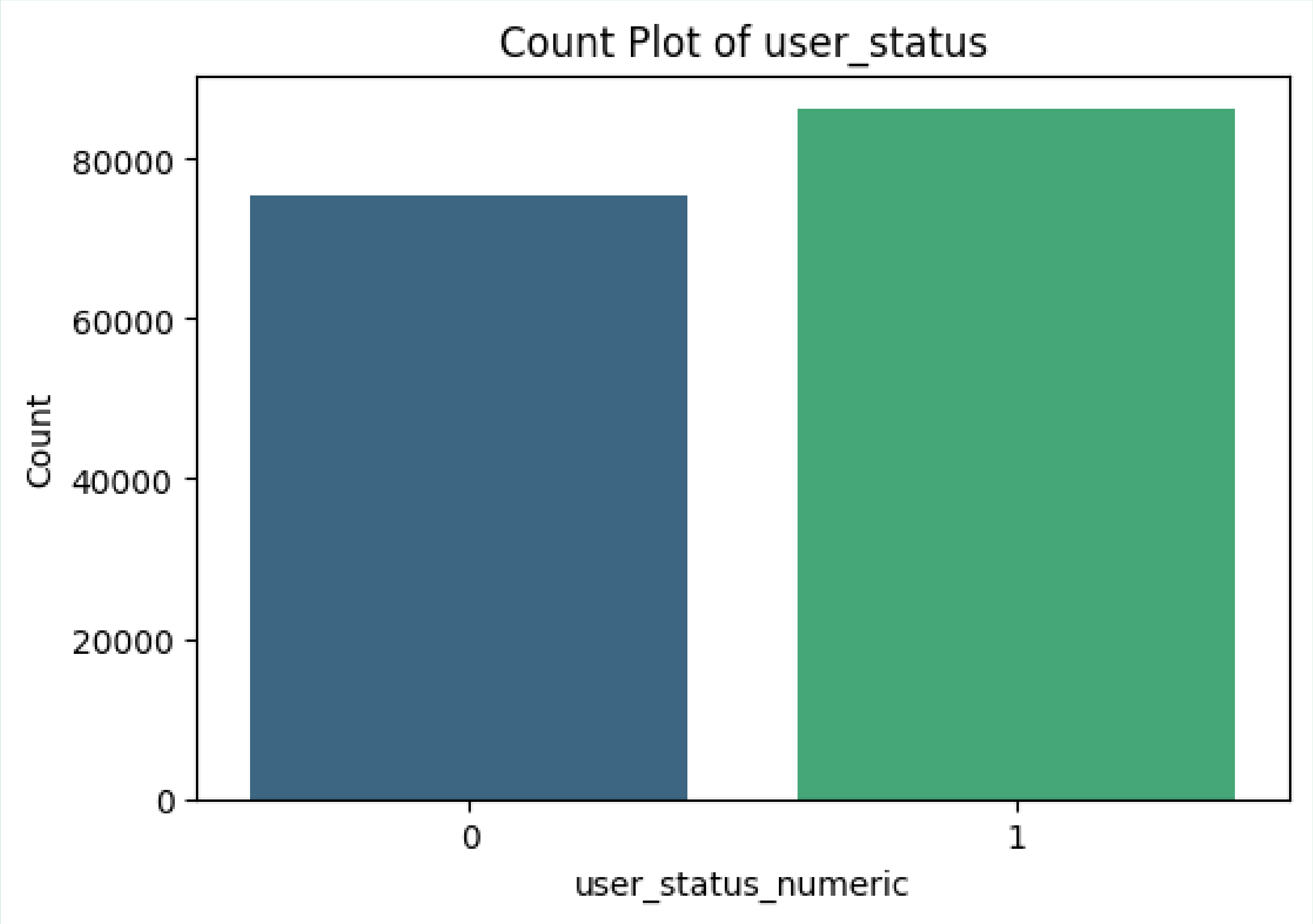
	day_of_week obj...	time_period object	usercount int64
0	Friday	Afternoon	10333
1	Friday	Evening	3922
2	Friday	Morning	7115
3	Friday	Night	482
4	Monday	Afternoon	11272

5 rows, showing 10 per page

Statistical Analysis

The primary statistical analysis conducted in this study revolves around fitting a logistic regression model.

Explanation: The reason behind this approach is rooted in the column "total_order_number" in the dataset. This column can be split into binary values (1 for more valuable customers and 0 for less valuable customers). Subsequently, a logistic regression prediction is performed based on additional columns capturing customers' demographic information and shopping behaviors.



Statistical Analysis

Model Summary:

Due to significant p-value, factors have meaningful impacts on weather a customer can be identified as more valuable or less valuable for the business.

Model:

```
X = merged_df[['Gender','Age','fam_status','income','n_dependants','days_since_prior_order']]
y = merged_df['user_status_numeric’]
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
model = LogisticRegression()
model.fit(X_train, y_train)
```

	coef	std err	z	P>|z|	[0.025	0.975]
Age	0.0047	0.000	14.998	0.000	0.004	0.005
income	3.395e-06	1.58e-07	21.434	0.000	3.08e-06	3.71e-06
n_dependants	0.1691	0.006	29.196	0.000	0.158	0.180
days_since_prior_order	-0.0540	0.001	-99.963	0.000	-0.055	-0.053
fam_status_single	0.3963	0.015	27.103	0.000	0.368	0.425
Gender_Male	0.0806	0.011	7.087	0.000	0.058	0.103

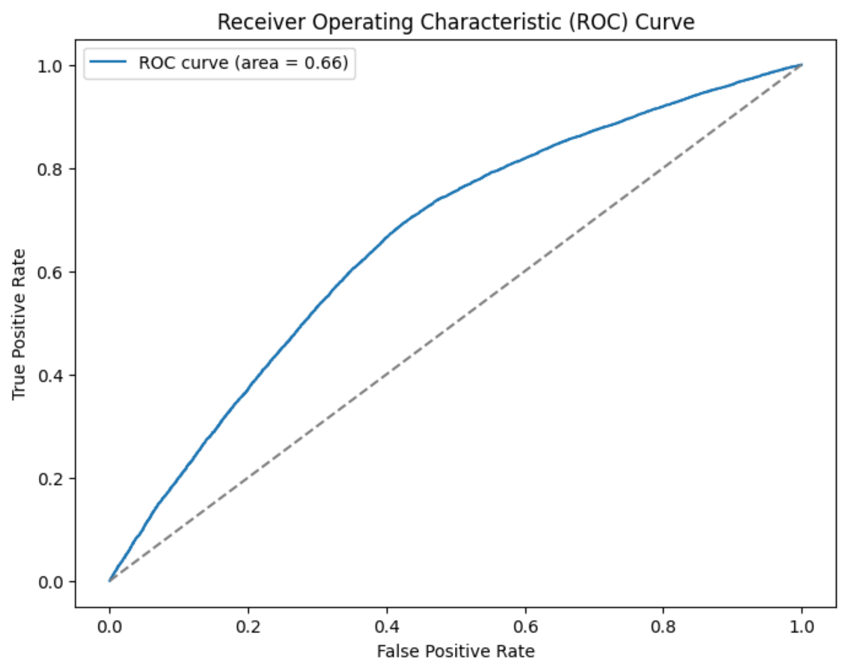
Statistical Analysis

Model validation:
the model have moderate performance in prediction customer value

Variable	VIF
const	23.615028
Age	1.287260
income	1.230264
n_dependants	1.953196
days_since_prior_order	1.000841
fam_status_single	2.012192
Gender_Male	1.000021

Small VIF indicates there is no multicollinearity among the predictor variables. the predictor variables are providing unique and independent information to the regression model.

AUC =0.66 indicates the model is performing better than random guessing to distinct between the positive and negative classes in binary variable.



Confusion Matrix: [[8846 6220] [5516 11735]
The model **correctly** identified 11,735 cases as positive.
It **correctly** identified 8,846 cases as negative.
It **incorrectly** predicted 6,220 cases as positive when they were actually negative.
It **incorrectly** predicted 5,516 cases as negative when they were actually positive.
Accuracy = 63%

THANKS

Presented by Carol Yu