PRESENTATION ON

OLiST ECOMMERCE-ANALYTICS - DATA SET

PRESENTED BY :
ARABINDA SAHOO

INTRODUCTION

- E-commerce growth demands data-driven decisions to improve customer experience and operational efficiency.
- This project analyzes the List Brazilian E-commerce Dataset covering orders, customers, sellers, products, payments, reviews, delivery timelines, and geolocation.
- Despite strong sales, List faces challenges such as delivery delays, high logistics costs, uneven seller performance, and low customer ratings.
- The analysis aims to uncover key patterns in customer behavior, delivery performance, payment trends, and product categories.
- Insights from this study support logistics optimization, seller improvement, and better customer satisfaction.

DATASET OVERVIEW

- The dataset contains ~120,000 rows and ~45 columns, created by merging multiple List e-commerce datasets.

- It represents Brazilian online marketplace transactions, covering orders, customers, sellers, products, payments, reviews, delivery, and geolocation.

- Key features include order status, product category, price, freight value, payment details, and review scores.

- The dataset consists of categorical, numerical, and datetime columns, suitable for business and operational analysis.

KEY COLUMNS OVERVIEW

- **order_id** – Unique identifier for each order
- **order status** – Current status of the order (delivered, canceled, etc.)
- **order status** – Date and time when the order was placed
- **order status** – Date when the order was delivered to the customer
- **customer state** – State where the customer is located
- **customer state** – Product category in English
- **price** – Price of the product
- **freight value** – Shipping cost for the order item
- **seller state** – State where the seller is located
- **payment type** – Payment method used by the customer
- **review score** – Customer rating for the order (1 to 5)
- **delivery days** – Number of days taken to deliver the order

PROBLEM STATEMENT

- The List dataset represents Brazilian e-commerce transactions, including orders, customers, sellers, products, payments, reviews, delivery, and geolocation data.
- Despite high sales volume, List faces challenges such as delivery delays, high logistics costs, low customer ratings, cancellations, and regional inefficiencies.
- The objective of this analysis is to understand delivery performance, customer satisfaction, sales trends, payment behavior, and geographical patterns using the merged dataset.
- Insights from this analysis aim to identify operational bottlenecks, improve logistics efficiency, enhance seller performance, and increase customer satisfaction.

READING DATASET

1. Load All Raw CSV Files

```
import pandas as pd

customers = pd.read_csv("olist_customers_dataset.csv")
orders = pd.read_csv("olist_orders_dataset.csv")
order_items = pd.read_csv("olist_order_items_dataset.csv")
products = pd.read_csv("olist_products_dataset.csv")
sellers = pd.read_csv("olist_sellers_dataset.csv")
payments = pd.read_csv("olist_order_payments_dataset.csv")
reviews = pd.read_csv("olist_order_reviews_dataset.csv")
category = pd.read_csv("product_category_name_translation.csv")
geolocation = pd.read_csv("olist_geolocation_dataset.csv")
```

- This step loads all raw CSV files related to customers, orders, products, sellers, payments, reviews, categories, and geolocation using pandas.
- Loading these datasets separately is necessary before merging them into a single master dataset for analysis.

MERGE THE DATASETS

Merge the Datasets into One Master DataFrame

```
[4]: Olist = orders.merge(customers, on="customer_id", how="left")

Olist = Olist.merge(order_items, on="order_id", how="left")

Olist = Olist.merge(product, on="product_id", how="left")

Olist = Olist.merge(sellers, on="seller_id", how="left")

Olist = Olist.merge(payments, on="order_id", how="left", suffixes=("", "_payment"))

Olist = Olist.merge(reviews, on="order_id", how="left", suffixes=("", "_review"))
```

```
[5]: Olist = Olist.merge(category,on = 'product_category_name',how = 'left')
```

- This step merges all individual List datasets into a single master Data Frame using common keys such as order_id, customer, productid, and Selerix.
- Left joins are used to preserve all order records while enriching them with customer, product, seller, payment, review, and category details for complete analysis.

INNOMATICS RESEARCH LABS

DATA SET

```
Olist
```

	order_id	customer_id	order_status	order_purchase_timestamp	order_approved_at	order_delivered
0	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-1
1	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-1
2	e481f51cbdc54678b7cc49136f2d6af7	9ef432eb6251297304e76186b10a928d	delivered	2017-10-02 10:56:33	2017-10-02 11:07:15	2017-1
3	53cdb2fc8bc7dce0b6741e2150273451	b0830fb4747a6c6d20dea0b8c802d7ef	delivered	2018-07-24 20:41:37	2018-07-26 03:24:27	2018-0
4	47770eb9100c2d0c44946d9cf07ec65d	41ce2a54c0b03bf3443c3d931a367089	delivered	2018-08-08 08:38:49	2018-08-08 08:55:23	2018-0
...	...	...	...	...	...	
119146	63943bddc261676b46f01ca7ac2f7bd8	1fca14ff2861355f6e5f14306ff977a7	delivered	2018-02-06 12:58:58	2018-02-06 13:10:37	2018-0
119147	83c1379a015df1e13d02aae0204711ab	1aa71eb042121263aafbe80c1b562c9c	delivered	2017-08-27 14:46:43	2017-08-27 15:04:16	2017-0
119148	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	2018-01-08 21:28:27	2018-01-08 21:36:21	2018-0
119149	11c177c8e97725db2631073c19f07b62	b331b74b18dc79bcdf6532d51e1637c1	delivered	2018-01-08 21:28:27	2018-01-08 21:36:21	2018-0
119150	66dea50a8b16d9b4dee7af250b4be1a5	edb027a75a1449115f6b43211ae02a24	delivered	2018-03-08 20:57:30	2018-03-09 11:20:28	2018-0

119151 rows × 44 columns

INNOMATICS RESEARCH LABS

CHECK DATA STRUCTURE

Check Dataset Structure

```
Olist.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 119151 entries, 0 to 119150
Data columns (total 48 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   order_id                       119151 non-null  object
 1   customer_id                    119151 non-null  object
 2   order_status                   119151 non-null  object
 3   order_purchase_timestamp       119151 non-null  datetime64[ns]
 4   order_approved_at              118974 non-null  datetime64[ns]
 5   order_delivered_carrier_date   117065 non-null  datetime64[ns]
 6   order_delivered_customer_date  115730 non-null  datetime64[ns]
 7   order_estimated_delivery_date  119151 non-null  datetime64[ns]
 8   customer_unique_id             119151 non-null  object
 9   customer_zip_code_prefix       119151 non-null  int64
 10  customer_city                  119151 non-null  object
 11  customer_state                 119151 non-null  object
 12  order_item_id                  118318 non-null  float64
 13  product_id                     118318 non-null  object
 14  seller_id                      118318 non-null  object
 15  shipping_limit_date            118318 non-null  datetime64[ns]
 16  price                          118318 non-null  float64
 17  freight_value                  118318 non-null  float64
 18  product_category_name          116609 non-null  object
 19  product_name_lenght            116609 non-null  float64
 20  product_description_lenght     116609 non-null  float64
 21  product_photos_qty             116609 non-null  float64
```

```
 21  product_photos_qty             116609 non-null  float64
 22  product_weight_g               118298 non-null  float64
 23  product_length_cm              118298 non-null  float64
 24  product_height_cm              118298 non-null  float64
 25  product_width_cm               118298 non-null  float64
 26  seller_zip_code_prefix         118318 non-null  float64
 27  seller_city                    118318 non-null  object
 28  seller_state                   118318 non-null  object
 29  payment_sequential             119148 non-null  float64
 30  payment_type                   119148 non-null  object
 31  payment_installments           119148 non-null  float64
 32  payment_value                  119148 non-null  float64
 33  review_id                      119151 non-null  object
 34  review_score                   119151 non-null  int64
 35  review_comment_title           14189 non-null   object
 36  review_comment_message         51250 non-null   object
 37  review_creation_date           119151 non-null  datetime64[ns]
 38  review_answer_timestamp        119151 non-null  datetime64[ns]
 39  product_category_name_english  116584 non-null  object
 40  geolocation_state              119064 non-null  object
 41  geolocation_city               119064 non-null  object
 42  geolocation_lat                119064 non-null  float64
 43  geolocation_lng                119064 non-null  float64
 44  delivery_days                  115730 non-null  Int64
 45  delivered_flag                 119151 non-null  int64
 46  late_delivery_flag             119151 non-null  int64
 47  revenue                        118318 non-null  float64
dtypes: Int64(1), datetime64[ns](8), float64(17), int64(4), object(18)
memory usage: 43.7+ MB
```

INNOMATICS
RESEARCH LABS

STATISTICS SUMMARY

Get Summary Statistics

[24]:
```
Olist.describe()
```

[24]:

	order_purchase_timestamp	order_approved_at	order_delivered_carrier_date	order_delivered_customer_date	order_estimated_delivery_date	customer_zip
count	119151	118974	117065	115730	119151	1
mean	2017-12-29 18:31:42.703502080	2017-12-30 04:44:50.771109632	2018-01-03 08:19:59.011719936	2018-01-12 20:51:28.266957568	2018-01-22 15:17:47.119873024	
min	2016-09-04 21:15:19	2016-09-15 12:16:38	2016-10-08 10:34:01	2016-10-11 13:46:32	2016-09-30 00:00:00	
25%	2017-09-10 20:15:46	2017-09-11 15:50:48.500000	2017-09-14 19:52:12	2017-09-22 21:54:31.249999872	2017-10-02 00:00:00	
50%	2018-01-17 11:59:12	2018-01-17 16:49:49	2018-01-23 17:03:08	2018-02-01 03:17:55	2018-02-14 00:00:00	
75%	2018-05-03 13:11:15	2018-05-03 16:56:53	2018-05-07 14:54:00	2018-05-14 23:58:16	2018-05-25 00:00:00	
max	2018-10-17 17:30:18	2018-09-03 17:40:06	2018-09-11 19:48:28	2018-10-17 13:22:46	2018-11-12 00:00:00	
std	NaN	NaN	NaN	NaN	NaN	

8 rows × 30 columns

INNOMATICS RESEARCH LABS

```
[45]: Olist.describe(include = 'O')
```

[45]:

product_category_name	seller_city	review_id	review_comment_title	review_comment_message	geolocation_state	geolocation_city
116609	119151	119151	14189	51250	119064	119064
73	612	99173	4600	36921	27	4067
cama_mesa_banho	sao paulo	eef5dbca8d37dfce6db7d7b16dd0525e	Recomendo	Muito bom	SP	sao paulo
11990	29294	63	498	259	50259	18876

Numerical Summary (describe())

- Prices, freight values, and delivery days show high variation and right-skewness, indicating diverse order sizes and shipping costs.
- Delivery days contain long tails, confirming the presence of delayed orders.
- Product weight and dimensions have outliers, which are expected in an e-commerce marketplace.

Categorical Summary (describe('O'))

- Most orders are successfully delivered, making it the dominant order status.
- A small number of product categories, states, and sellers account for the majority of transactions.
- Credit card is the most commonly used payment method, and many reviews lack text feedback.

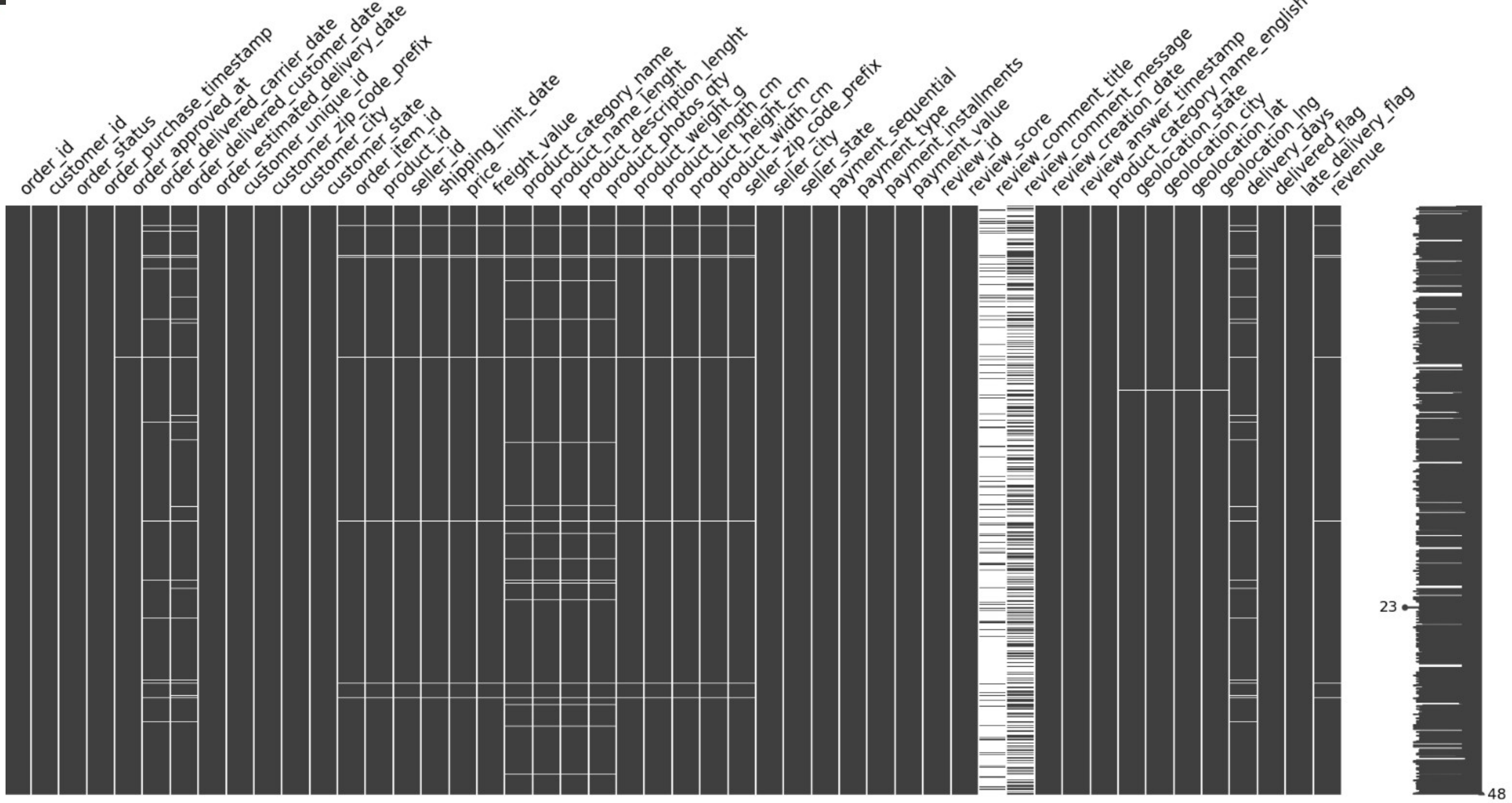
IDENTIFY MISSING VALUES

Identify Missing Values

```
Olist.isnull().sum()
```

```
order_id                            0
customer_id                         0
order_status                        0
order_purchase_timestamp            0
order_approved_at                 177
order_delivered_carrier_date     2086
order_delivered_customer_date    3421
order_estimated_delivery_date       0
customer_unique_id                  0
customer_zip_code_prefix            0
customer_city                       0
customer_state                      0
order_item_id                     833
product_id                        833
seller_id                         833
shipping_limit_date               833
price                             833
freight_value                     833
product_category_name            2542
product_name_lenght              2542
product_description_lenght       2542
product_photos_qty               2542
product_weight_g                  853
product_length_cm                 853
product_height_cm                 853
product_width_cm                  853
seller_zip_code_prefix            833
seller_city                       833
seller_state                      833
payment_sequential                  3
payment_type                        3
payment_value                       3
review_id                           0
review_score                        0
review_comment_title           104962
review_comment_message          67901
review_creation_date                0
review_answer_timestamp             0
product_category_name_english       0
geolocation_state                  87
geolocation_city                   87
geolocation_lat                    87
geolocation_lng                    87
delivery_days                    3421
delivered_flag                      0
late_delivery_flag                  0
revenue                           833
dtype: int64
```

INNOMATICS RESEARCH LABS



- Missing values are mainly concentrated in delivery-related and review comment columns, representing undelivered orders and customers who did not leave written feedback.
- These missing values are expected in real-world e-commerce data and reflect actual business situations rather than data quality issues.

HANDLING MISSING VALUES

```
[ ]: product_cols = [
         'product_description_lenght','product_photos_qty','product_weight_g',
         'product_length_cm','product_height_cm','product_width_cm'
     ]

     for col in product_cols:
         Olist[col] = Olist[col].fillna(Olist[col].median())
```

```
[ ]: ## Filling the NaN values with 'No Comments'
     Olist['review_comment_title'] = Olist['review_comment_title'].fillna("No Comment")

     Olist['review_comment_message'] = Olist['review_comment_message'].fillna("No Comment")

     Olist['product_category_name_english'] = Olist['product_category_name_english'].fillna("Unknown")
```

Filling the Geolocation_state and Geolocation_city with 'Unknown'.

```
[ ]: Olist['geolocation_state'] = Olist['geolocation_state'].fillna("Unknown")
     Olist['geolocation_city'] = Olist['geolocation_city'].fillna("Unknown")
```

Filling the Product_name_length with Median

```
[ ]: Olist['product_name_lenght'] = Olist['product_name_lenght'].fillna(
         Olist['product_name_lenght'].median()
     )
```

- Missing numerical product attributes were filled using the median to reduce the impact of outliers and preserve realistic values.
- Categorical and text fields were filled with meaningful labels such as “Unknown” and “No Comment”, ensuring data completeness without distorting business meaning.

INNOMATICS RESEARCH LABS

Filling the Olist[['payment_sequential','payment_type','payment_installments','payment_value']] with Mode

```
]: Olist[['payment_sequential','payment_type','payment_installments','payment_value']].isna().sum()
```

```
]: fill_cols = ['payment_sequential', 'payment_type',
             'payment_installments', 'payment_value']

for col in fill_cols:
    Olist[col] = Olist[col].fillna(Olist[col].mode()[0])
```

Filling the Olist['geolocation_lat','geolocation_lng'] with "0" because They lat,lng are nan values.

```
]: Olist['geolocation_lat'] = Olist['geolocation_lat'].fillna(0)
Olist['geolocation_lng'] = Olist['geolocation_lng'].fillna(0)
```

- Payment-related columns were filled using the **mode**, preserving the most common payment behavior and avoiding distortion of transaction patterns.
- Missing geolocation latitude and longitude values were set to **0** to maintain dataset completeness while indicating unavailable location information.

CHECKING NULL VALUES AFTER PREPROCESSING

```
order_id                          0
customer_id                       0
order_status                      0
order_purchase_timestamp          0
order_approved_at               177
order_delivered_carrier_date   2086
order_delivered_customer_date  3421
order_estimated_delivery_date     0
customer_unique_id                0
customer_zip_code_prefix          0
customer_city                     0
customer_state                    0
order_item_id                   833
product_id                      833
seller_id                       833
shipping_limit_date             833
price                           833
freight_value                   833
product_category_name          2542
product_name_lenght            2542
product_description_lenght     2542
product_photos_qty             2542
product_weight_g                853
product_length_cm               853
product_height_cm               853
product_width_cm                853
seller_zip_code_prefix          833
seller_city                       0
seller_state                      0
payment_sequential                3
payment_type                      0
payment_installments              3
payment_value                     3
```

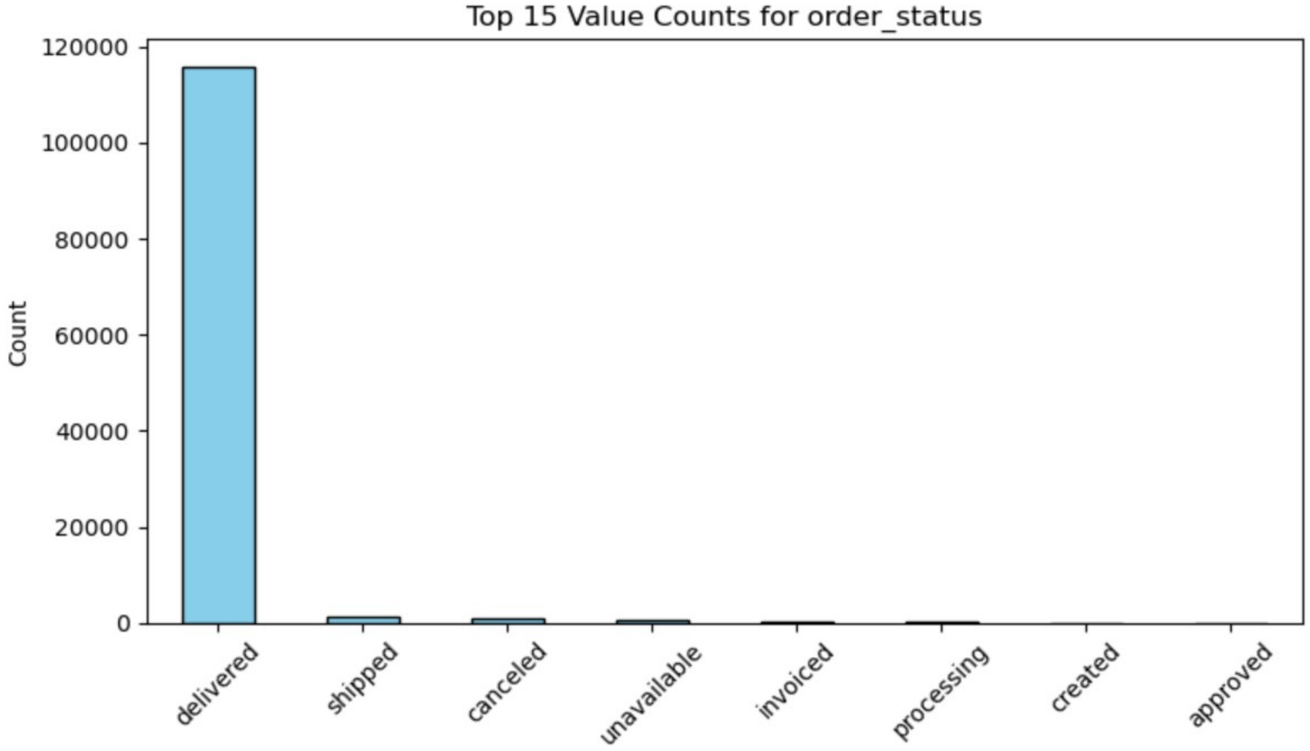
```
review_id                             0
review_score                          0
review_comment_title             104962
review_comment_message            67901
review_creation_date                  0
review_answer_timestamp               0
product_category_name_english         0
geolocation_state                    87
geolocation_city                     87
geolocation_lat                      87
geolocation_lng                      87
delivery_days                      3421
delivered_flag                        0
late_delivery_flag                    0
revenue                             833
dtype: int64
```

- After preprocessing, all critical columns have been successfully handled, and the dataset contains no missing values that impact analysis.
- This confirms that the dataset is clean, consistent, and ready for exploratory analysis and business insights

DATA ANALYSIS THROUGH VISUALIZATIONS

UNIVARIATE ANALYSIS

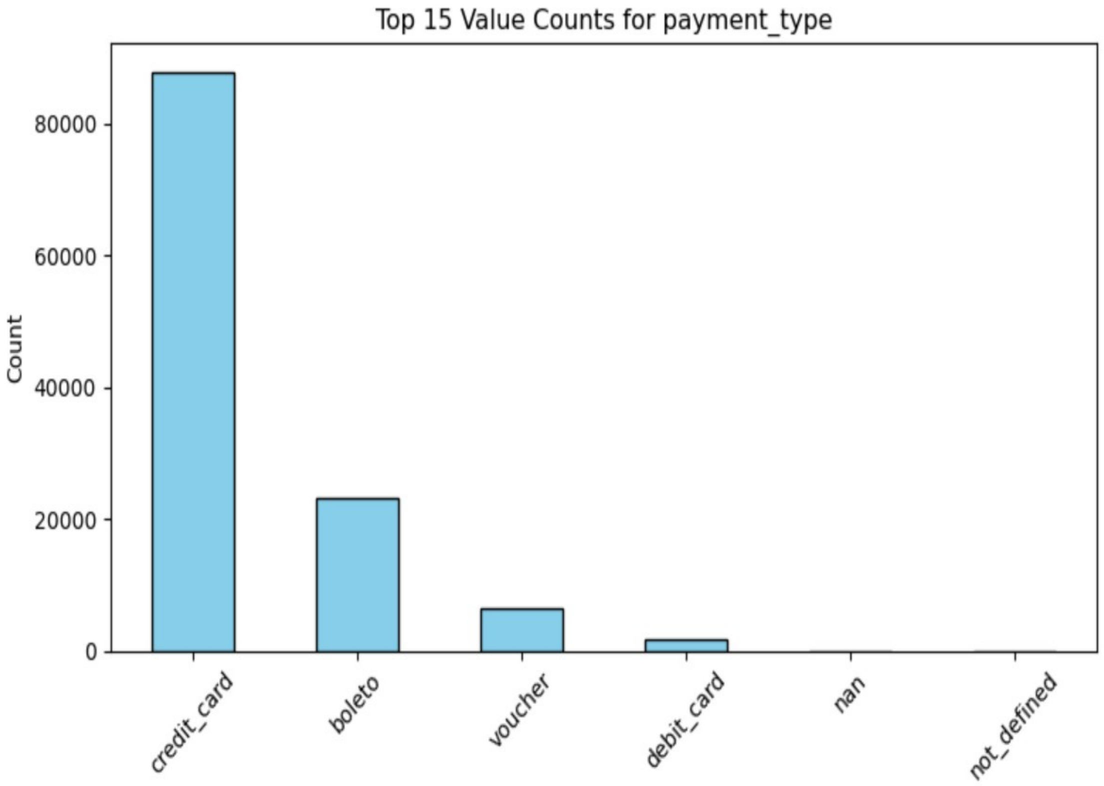
```
Value Counts for order_status:
order_status
delivered      115731
shipped          1256
canceled          750
unavailable       652
invoiced          378
processing        376
created             5
approved            3
Name: count, dtype: int64
```



Order Status:

The vast majority of orders are **delivered**, indicating strong order fulfillment performance, while cancellations and unavailable orders form only a small fraction of total transactions.

```
Value Counts for payment_type:
payment_type
credit_card    87784
boleto         23190
voucher         6465
debit_card      1706
nan                3
not_defined        3
Name: count, dtype: int64
```



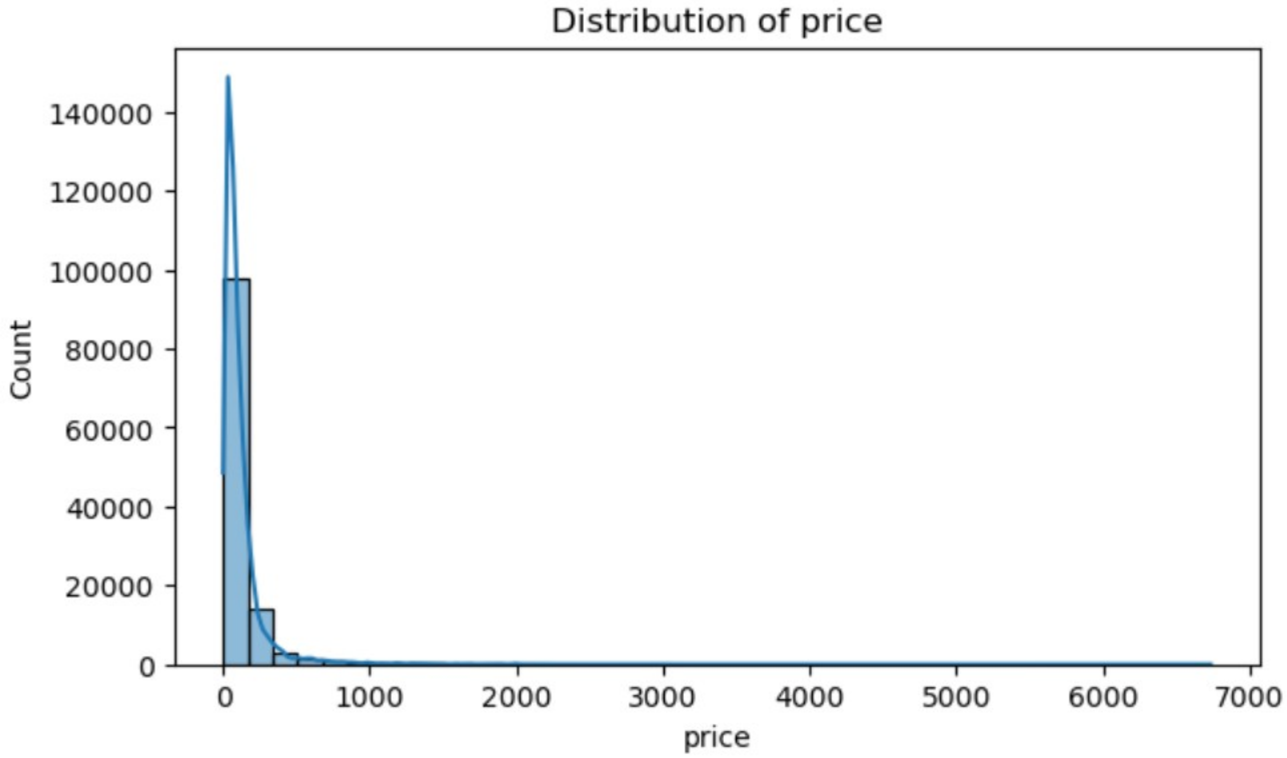
Insights from Categorical Distributions:

Payment Type:

Credit cards dominate as the primary payment method, followed by bolete, showing a strong preference for card-based payments among customers.

Distribution Plots (Histogram + KDE)

```
for col in num_cols:
    plt.figure(figsize=(7,4))
    sns.histplot(Olist[col].dropna(), kde=True, bins=40)
    plt.title(f"Distribution of {col}")
    plt.show()
```



Distribution Plot Insights

- Product prices and other monetary variables are highly right-skewed, with most values at the lower range and a few high-priced outliers typical of e-commerce data.

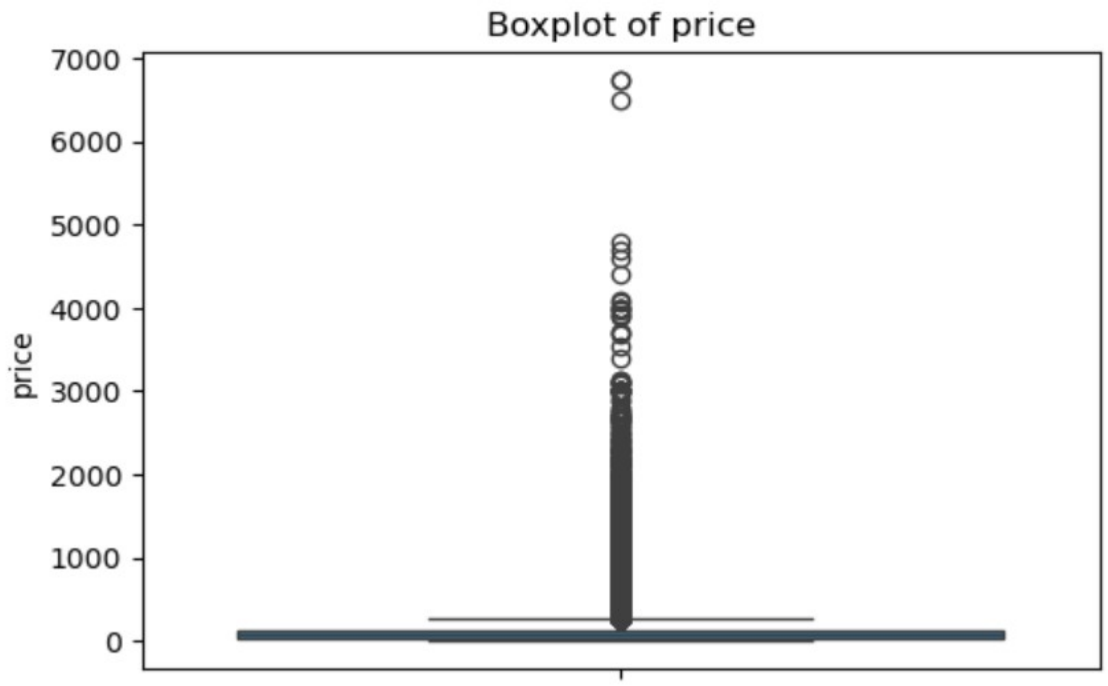
Delivery Days Distribution – Insights

- Delivery time is right-skewed, with most orders delivered quickly but a long tail of delayed deliveries highlighting logistics inefficiencies that negatively impact customer satisfaction.

Boxplot (Outlier) Analysis

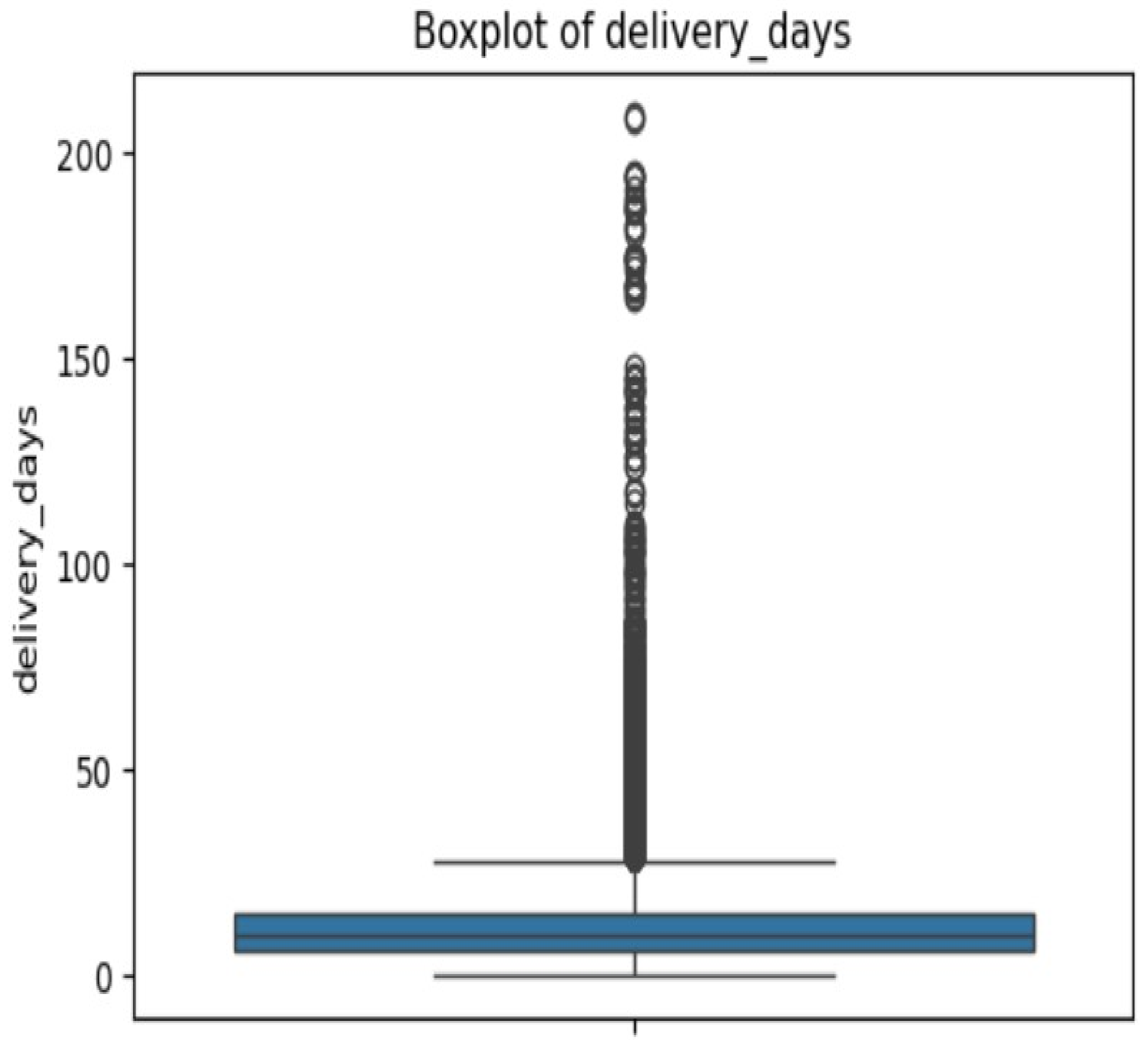
```
9]: product_cols = [
        'price','freight_value','payment_value','payment_installments',
        'delivery_days','product_weight_g','product_length_cm',
        'product_height_cm','product_width_cm'
    ]

    for col in product_cols:
        plt.figure(figsize=(6,4))
        sns.boxplot(y=Olist[col])
        plt.title(f"Boxplot of {col}")
        plt.show()
```



- Boxplots show significant upper outliers in price, freight, delivery days, and product dimensions, reflecting real e-commerce diversity such as premium products, bulky items, and remote deliveries.

Delivery Days – Boxplot Insights

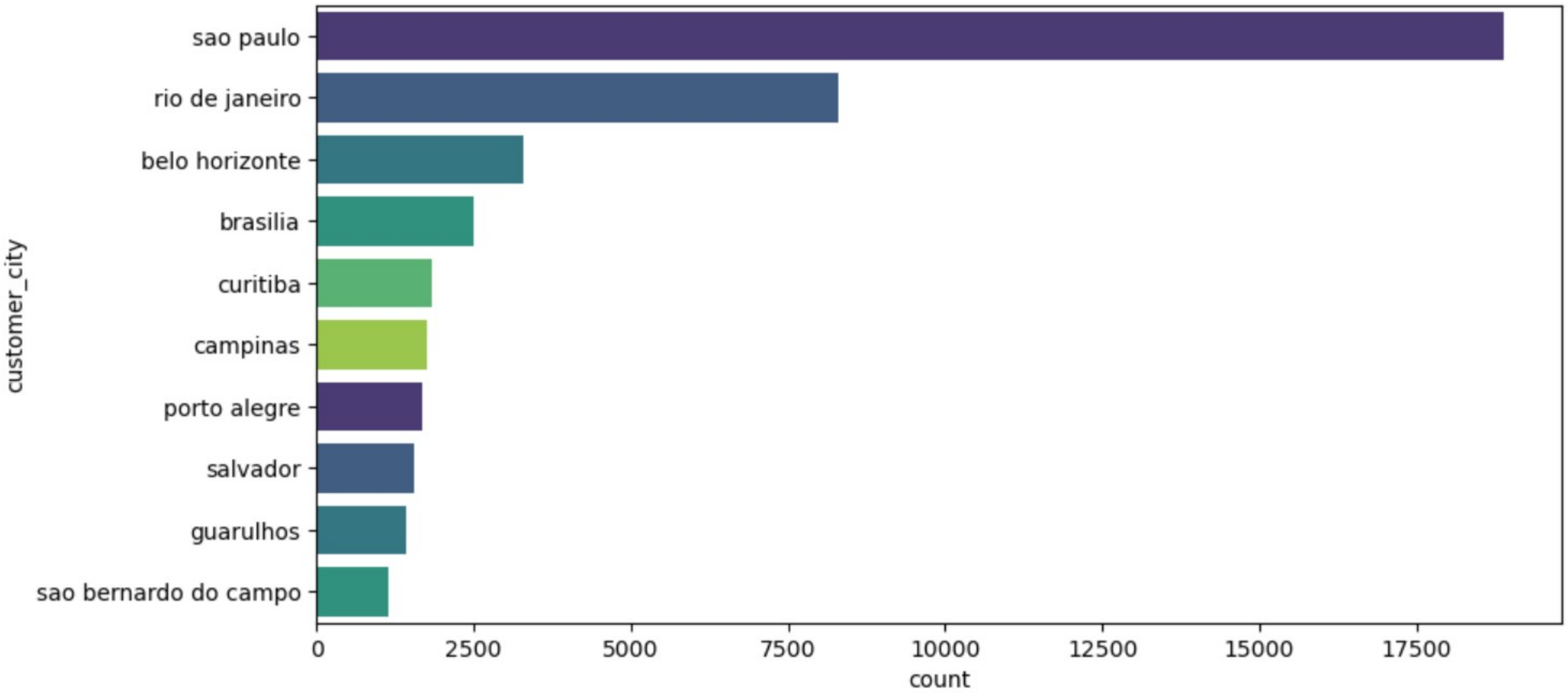


- Most orders are delivered within a consistent time range, but extreme delivery delays reveal regional or seller-level logistics inefficiencies that negatively impact customer satisfaction.

INNOMATICS RESEARCH LABS

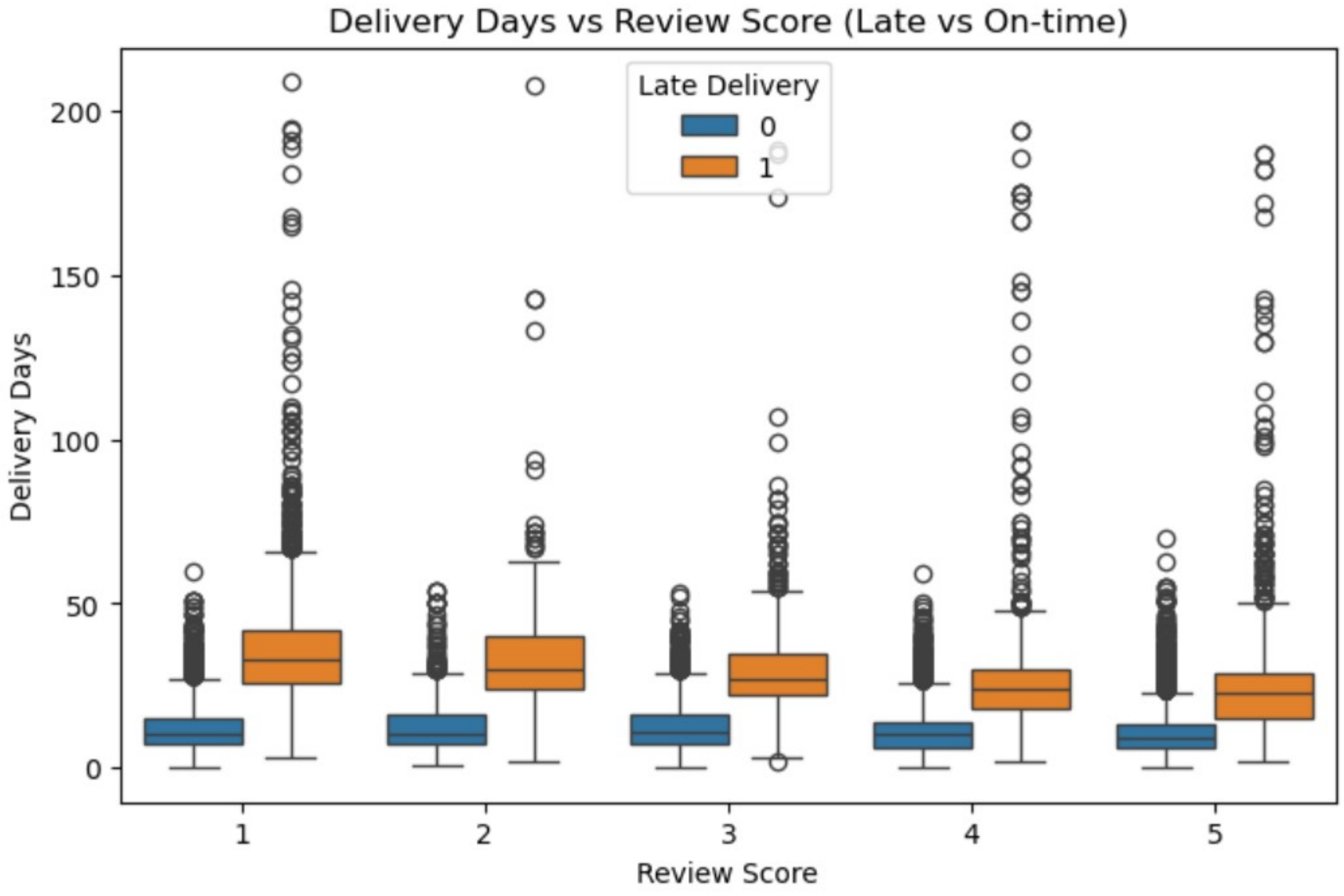
Customer City Distribution – Insights

```
top10_cities = Olist['customer_city'].value_counts().head(10).index
df_top10 = Olist[Olist['customer_city'].isin(top10_cities)]
plt.figure(figsize=(10,5))
sns.countplot(data=df_top10, y='customer_city', order=top10_cities,palette=sns.color_palette("viridis"))
plt.show()
```



BIVARIATE ANALYSIS

```
plt.figure(figsize=(8,5))
sns.boxplot(data=Olist,x='review_score',y='delivery_days',hue='late_delivery_flag')
plt.title("Delivery Days vs Review Score (Late vs On-time)")
plt.xlabel("Review Score")
plt.ylabel("Delivery Days")
plt.legend(title="Late Delivery")
plt.show()
```



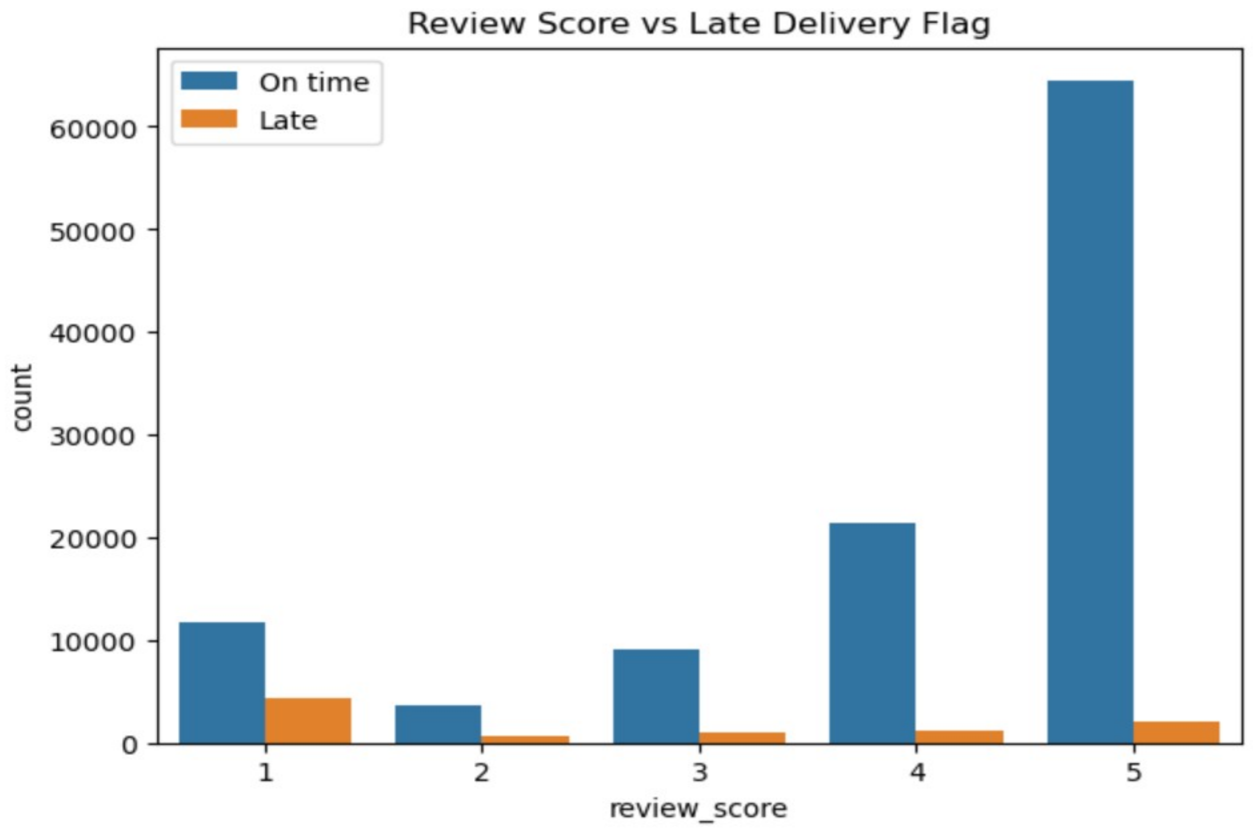
Delivery Days vs Review Score – Insights

- Late deliveries are associated with significantly longer delivery times and lower review scores, while on-time deliveries show shorter durations and higher customer satisfaction.

Late Delivery vs Review Score – Insights

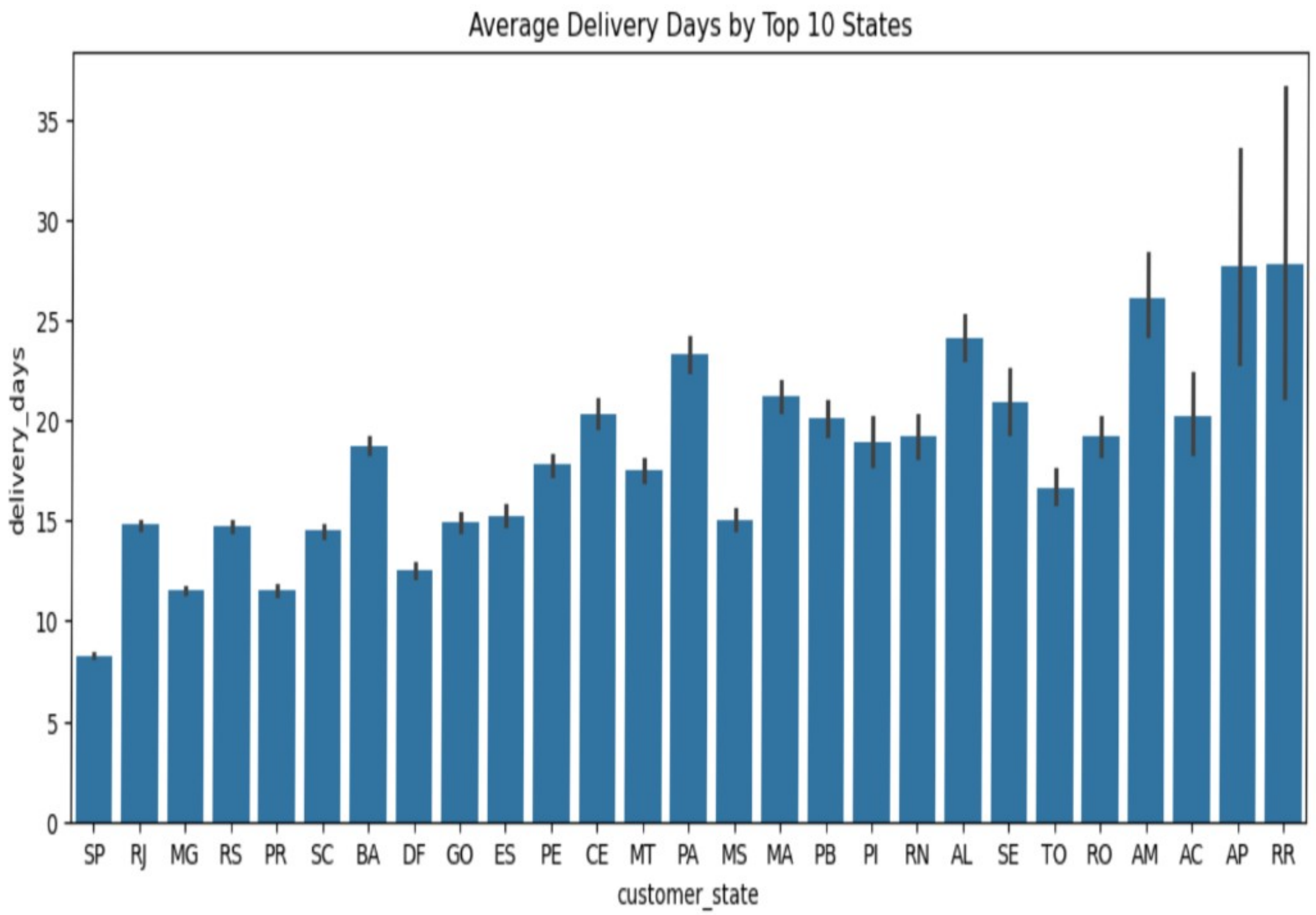
Late Delivery Flag vs Review Score

```
plt.figure(figsize=(7,5))
sns.countplot(data=Olist, x='review_score', hue='late_delivery_flag')
plt.title("Review Score vs Late Delivery Flag")
plt.legend(["On time", "Late"])
plt.show()
```



- On-time deliveries receive higher ratings while late deliveries are strongly associated with low reviews, confirming delivery timeliness as a key driver of customer satisfaction.

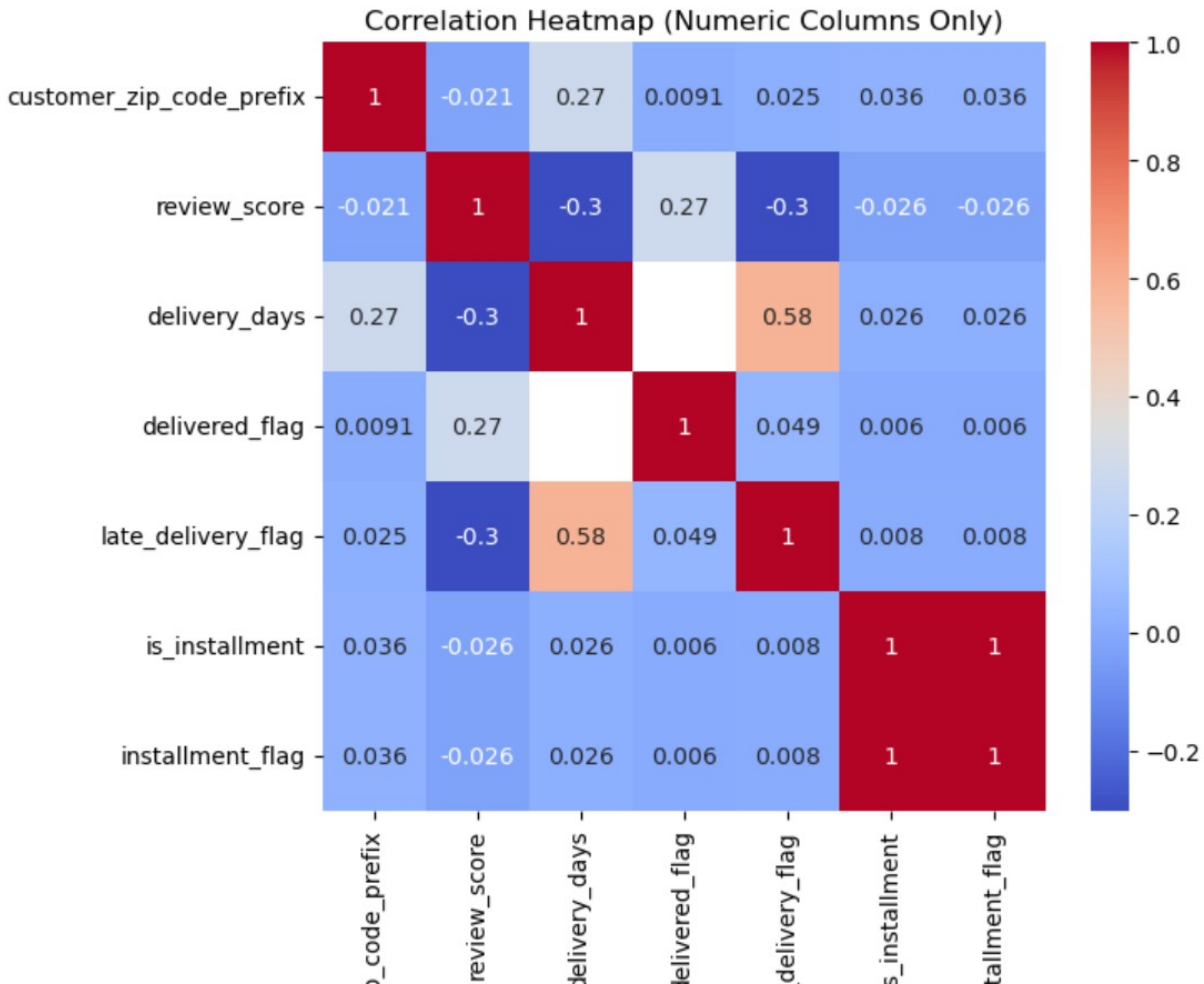
```
top_states = Olist['customer_state'].value_counts().tail(30).index
plt.figure(figsize=(12,5))
sns.barplot(data=Olist[Olist['customer_state'].isin(top_states)],
            x='customer_state', y='delivery_days',order = top_states)
plt.title("Average Delivery Days by Top 10 States")
plt.show()
```



Average Delivery Days by State

- States near major logistics hubs have shorter delivery times, while remote northern regions face longer deliveries due to infrastructure and distance challenges.

```python
numeric_df = Olist.select_dtypes(include=['int'])
corr_matrix = numeric_df.corr()
plt.figure(figsize=(7, 6))
sns.heatmap(corr_matrix, cmap='coolwarm',annot = True)
plt.title("Correlation Heatmap (Numeric Columns Only)")
plt.show()
```



Correlation Analysis – Key Insights

- Longer delivery times strongly lead to late deliveries and lower review scores, while payment behavior shows little influence on customer satisfaction.

Summary — Olist E-commerce Analysis

Category	Key Insight
Order Status	Majority of orders are successfully delivered, indicating strong fulfillment performance.
Delivery Performance	Delivery time varies significantly, certaln regions es-prience frequent late deliveries.
Customer Satisfaction	Late deliveries have a strong negative impact on customer review scores
Product Categories	A small number of product categories contribute to most orders and revenue
Sales Distribution	Order value and freight cost show right-skewed distributions with noticeable outliers
Seller Performance	Few sellers dominate sales, while some sellers consistently ship orders late.
Payment Behavior	Credit card is the most preferred payment method. followed by boleto and vouchers
Geography	Orders are highly concentrated in urban and southeastern states such as SP, RJ, and MG.
Logistics Cost	Freight cost increases with product weight and delivery distance.

CONCLUSION – LIST E-COMMERCE ANALYSIS

- The analysis provides a comprehensive view of customer behavior, seller performance, logistics efficiency, payment patterns, and product trends by integrating multiple List datasets.

- Delivery performance is the most critical factor affecting customer satisfaction; late deliveries strongly lead to low review scores, especially in remote regions.

- Sales follow an 80/20 pattern, where a small number of product categories and sellers generate most of the revenue.

- Credit cards dominate payment methods, and higher-value orders are commonly paid in installments, with no major impact on delivery timelines.

- Orders are highly concentrated in southeastern states (SP, RJ, MG), while distant regions face higher freight costs and longer delivery times.

- Operational outliers in price, weight, and freight reflect real business diversity, but consistent late shipping by some sellers highlights areas for logistics optimization.

Thank You

INNOMATICS
RESEARCH LABS