

Data Driven E-commerce Case Study for Olist

-Aditi Ganesh Joshi, Ashish Murlidhar Pagote, Surbhi Garg

Agenda

Olist Overview:

Company Insights

Key operations and datasets

Business Goals

Data Manipulation and Transformation using Excel Power Query

Feature Generation

Sorting

Aggregation

Merges

So on.....

Data Visualization using Tableau

Seller Analysis

Product Analysis

Customer Analysis

Conclusion and Recommendations

About Olist

- Connects SMBs with customers through its Online Marketplace
- Founded in 2015, Brazil by Tiago Dalvi
- 45000+ shopkeepers and retail stores as client
- Areas of operation: Brazil, Mexico
- Was valued at \$1.5 Billion in 2021

The Olist logo consists of a solid blue square. Inside the square, the word "olist" is written in a white, lowercase, sans-serif font. The letter 'o' is slightly larger and more prominent than the others.

Key operations and datasets



1. Order items
2. Orders
3. Product-Side
 - a. Products
 - b. Product Category Translation
4. Seller-Side
 - a. Sellers
5. Marketing
 - a. Closed Deals
 - b. Marketing Qualified Leads
6. Customer-Side
 - a. Customers
 - b. Order Payment Details
 - c. Order reviews

Business Goals

Data-driven analysis of an e-commerce firm in Brazil called “Olist”. Olist provides a marketplace for various SMBs to sell their products online. The aim of this project is divided into four parts as follows:

Sellers and Seller-Marketing Analysis

Understand the logistics and address inefficiencies by using seller performance data

Identify high impact sellers by analyzing sales metrics

Identify channels of marketing and quantify conversion metrics

Products Analysis

Perform exploratory data analysis on order data to identify the most frequently bought products, product categories, etc.

Conduct Market Basket Analysis on the e-commerce data to understand consumer behavior

Analyze trends and forecast future sales

Customer and Payment Analysis

To study the customer share across different regions of the country

Understand the modes of payments used the customer base and by how much

Data Manipulation & Transformation - Excel Power Query

Seller-Side

Data Source	Size (#rows)	Primary Keys
Order Items	112.6k	order_id, order_item_id, product_id
Orders	99.4k	order_id and customer_id
Sellers	3.1k	seller_id

Data Manipulation & Transformation - Excel Power Query

Seller Side

Feature Generation and Cleaning - Orders Dataset

- a. Time Difference between actual time of delivery and estimated time of delivery in days
- b. Order Breach Flag (1 or 0)
- c. Imputed Breach flag with 0 where delivery date was not available.

Merging, Feature selection and generation - Order Items Dataset

- b. Merged with Orders dataset
- c. Kept relevant columns (Shipping limit time, Breached Flag etc.)
- d. Created Time to Ship using order purchase time and shipping limit time
- e. Aggregated all the metrics at seller level
Metrics: Orders,Unique Orders,Breached Orders,Total Revenue,Total Freight Value,Average Shipping Time
- f. Calculate ***breach percentage*** orders by dividing breached orders by total orders

Power Query Snapshot- Query Editor and Seller level DB

Name
olist_orders_dataset

▼ **Applied steps**

- Source
- Promoted headers
- Changed column type
- Added custom
- Changed column type 1
- Added custom 1
- Replaced errors

▼ **Properties**

Name
olist_order_items_dataset

▼ **Applied steps**

- Source
- Promoted headers
- Changed column type
- Merged queries
- Expanded olist_orders_dataset
- Added custom
- Changed column type 1
- Grouped rows
- Added custom 1
- Changed column type 3

seller_id	1.2 Pri...	1.2 FreightVa...	1.2 ShipTi...	1.2 Ord...	1.2 UniqueOrd...	BreachedOrd...	% Breach_Percent...
48436dade18ac8b2bce089ec...	12271.71	2911.93	5.960264901	151	138	10	6.62%
dd7ddc04e1b6c2c614352b38...	9178.51	2893.49	8.468531469	143	122	18	12.59%
289cdb325fb7e7f891c38608...	13544.95	2094.34	3.626984127	126	110	2	1.59%
5b51032eddd24adc84c38ac...	3280	268.95	6.285714286	14	12	0	0.00%
4869f7a5dfa277a7dca6462dc...	229472.63	20168.07	6.011245675	1156	1132	130	11.25%
9d7a1d34a505240900642527...	1054.82	277.05	6.4375	16	13	1	6.25%
66922902710d126a0e7d26b0...	14362.3	3218.77	5.179487179	156	151	11	7.05%
df560393f3a51e74553ab940...	3661.18	606.15	10.20689655	29	28	2	6.90%
2c9e548be18521d1c43cde1c...	6109.44	2493.61	5.482758621	174	126	29	16.67%
6426d21aca402a131fc0a5d09...	1209.64	384.18	6.086956522	23	23	0	0.00%
8581055ce74af1daba164fdbd...	64925.3	9659	5.763218391	435	387	41	9.43%
7040e82f899a04d1b434b795...	9738.2	3217.78	5.785087719	228	212	20	8.77%
dc8798cbf453b7e0f98745e3...	2056.3	624.4	7.333333333	51	41	3	5.88%
5996cddab893a4652a15592f...	810	70.75	6	1	1	0	0.00%
16090f2ca825584b5a147ab2...	25716.44	5456.67	4.929268293	410	402	43	10.49%
a416b6a846a1172439302564...	25670.72	2999.8	5.900552486	181	162	22	12.15%
63b9ae557efed31d1f7687917...	139.5	102.81	4.142857143	7	5	0	0.00%

Data Manipulation & Transformation - Excel Power Query

Seller-Marketing Side

Data Source	Size (#rows)	Primary Keys
olist_marketing_qualified_leads	8k	mql_id
Orders	0.8k	mql_id

Data Manipulation & Transformation - Excel Power Query

Seller Marketing Side

Feature Generation and Cleaning - Marketing Dataset

- a. Converted or Not Flag
- b. Time to convert
- c. Aggregate at origin level to obtain converted leads, total leads and average time to convert
- d. Calculate conversion rate by converted leads and total leads
- e. Filtration of unknown origins

Power Query Snapshot- Query Editor and Marketing DB

Power Query Editor interface showing the query name and applied steps.

Name: olist_marketing_qualified_leads_dataset

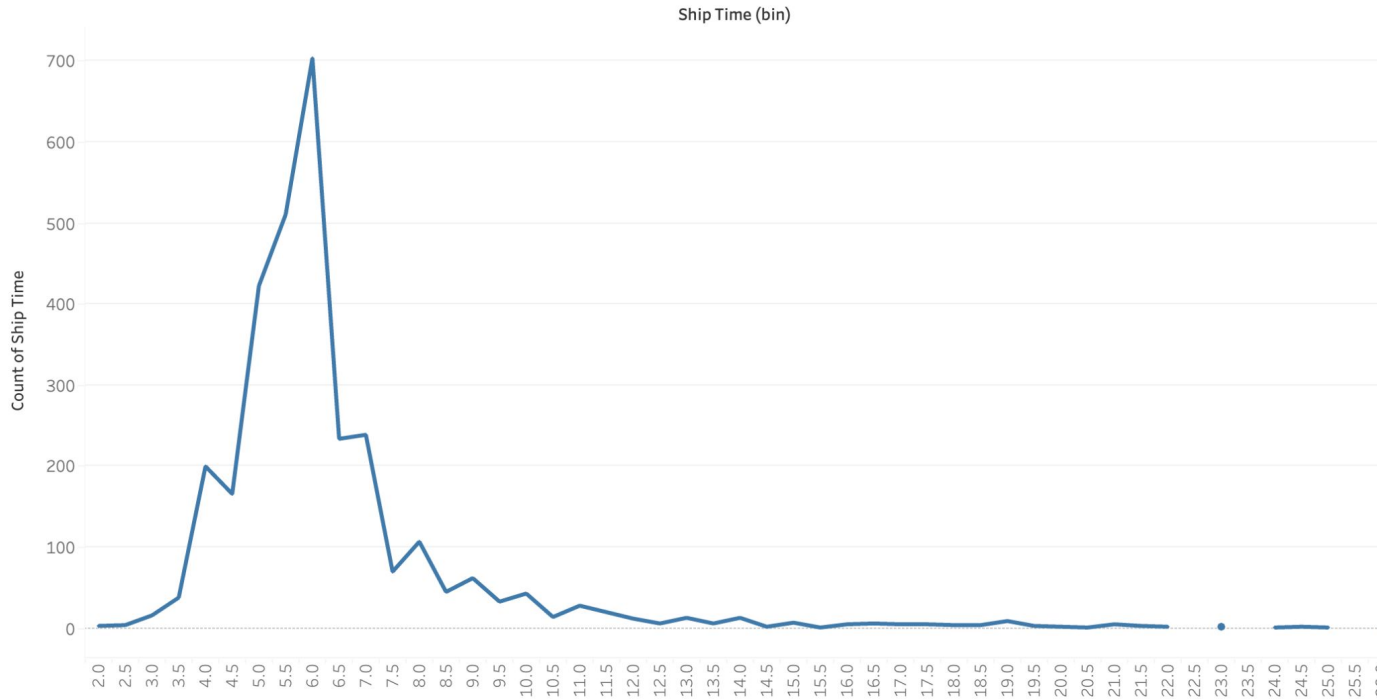
Applied steps:

- Source
- Promoted headers
- Changed column type
- Merged queries
- Expanded olist_closed_deals_dataset
- Added custom
- Changed column type 1
- Added custom 1
- Changed column type 2
- Grouped rows
- Added custom 2
- Changed column type 3
- Filtered rows

origin	TotalMarketingLe...	TotalConvertedLe...	TimeToConv...	% ConvertedPercent...
social	1350	75	30	5.56%
organic_search	2296	271	14	11.80%
paid_search	1586	195	15	12.30%
referral	284	24	18.5	8.45%
email	493	15	21	3.04%
direct_traffic	499	56	10	11.22%
display	118	6	8.5	5.08%
other_publicities	65	3	35	4.62%
other	150	4	9	2.67%

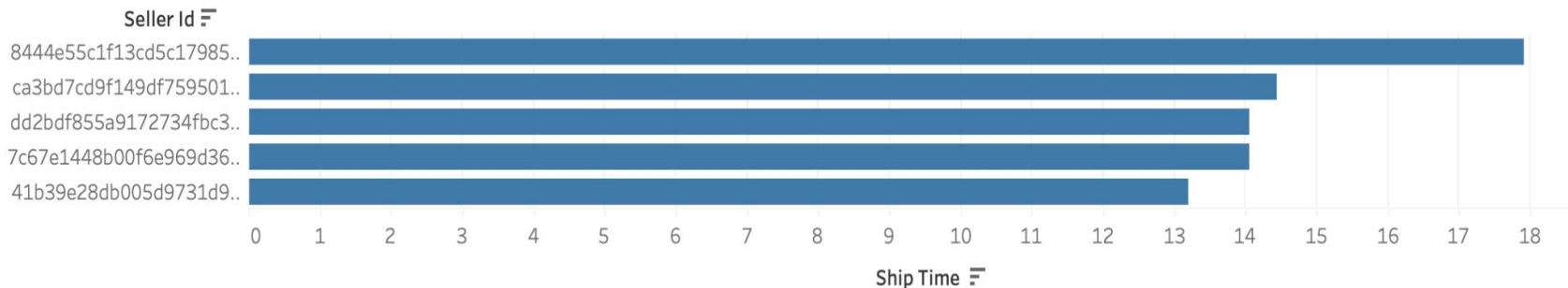
Data Visualisation using Tableau

Seller Analysis- Distribution of shipping time (Median around 6 days)

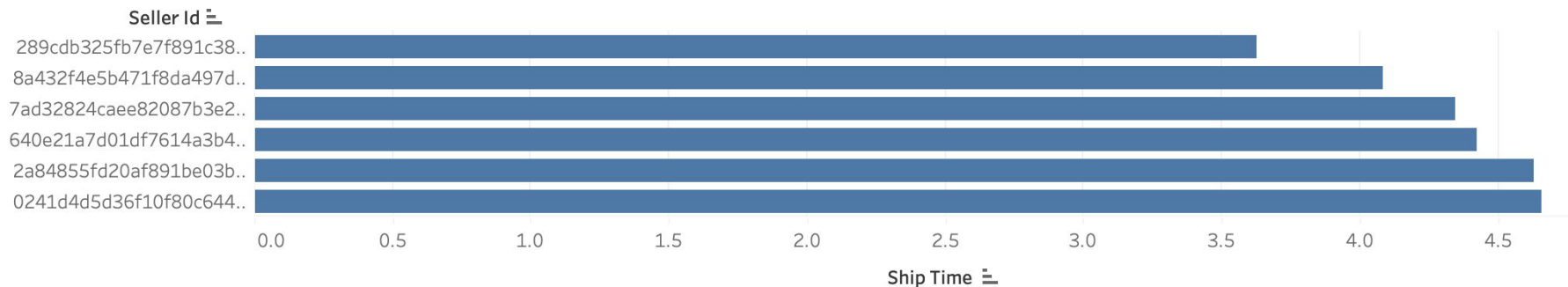


Data Visualisation using Tableau

Seller Performance- High Ship Time (Minimum 100 orders)

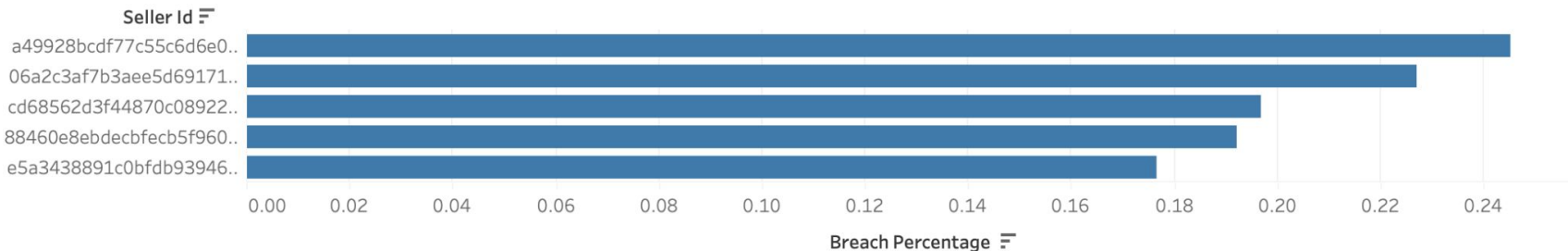


Seller Performance- Low Ship Time (Minimum 100 orders)

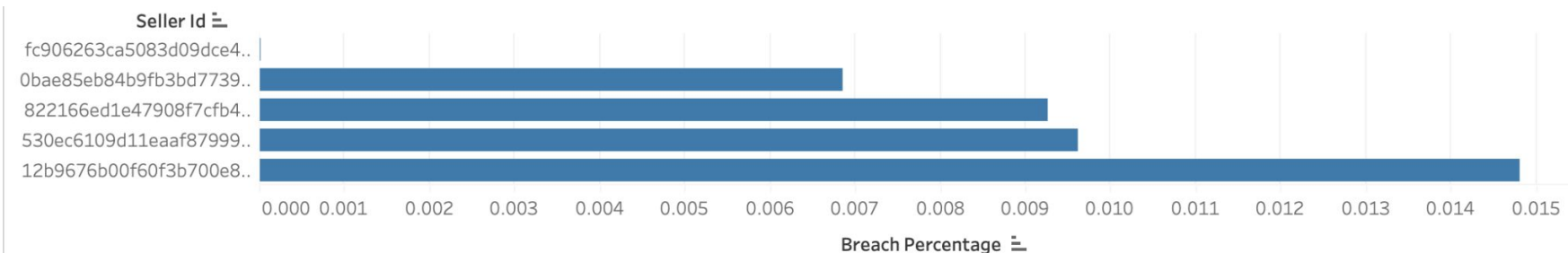


Data Visualisation using Tableau

Seller Performance- High Breach Percentage(Minimum 100 orders)



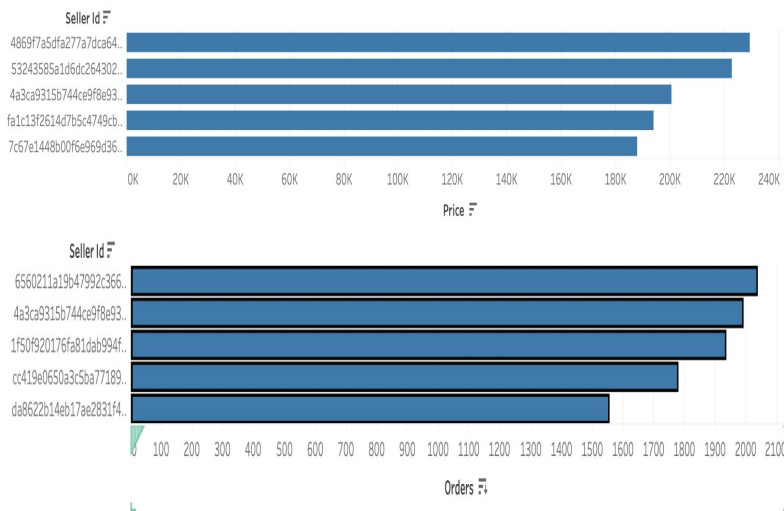
Seller Performance- Low Breach Percentage (Minimum 100 orders)



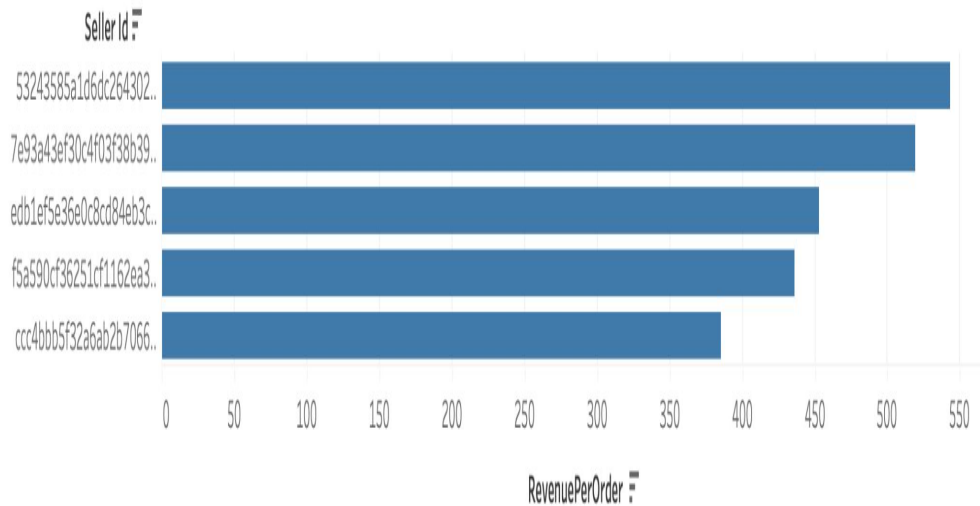
Data Visualisation using Tableau

Seller Performance- Highest Revenue, Orders and Revenue per Orders

High Revenue Sellers

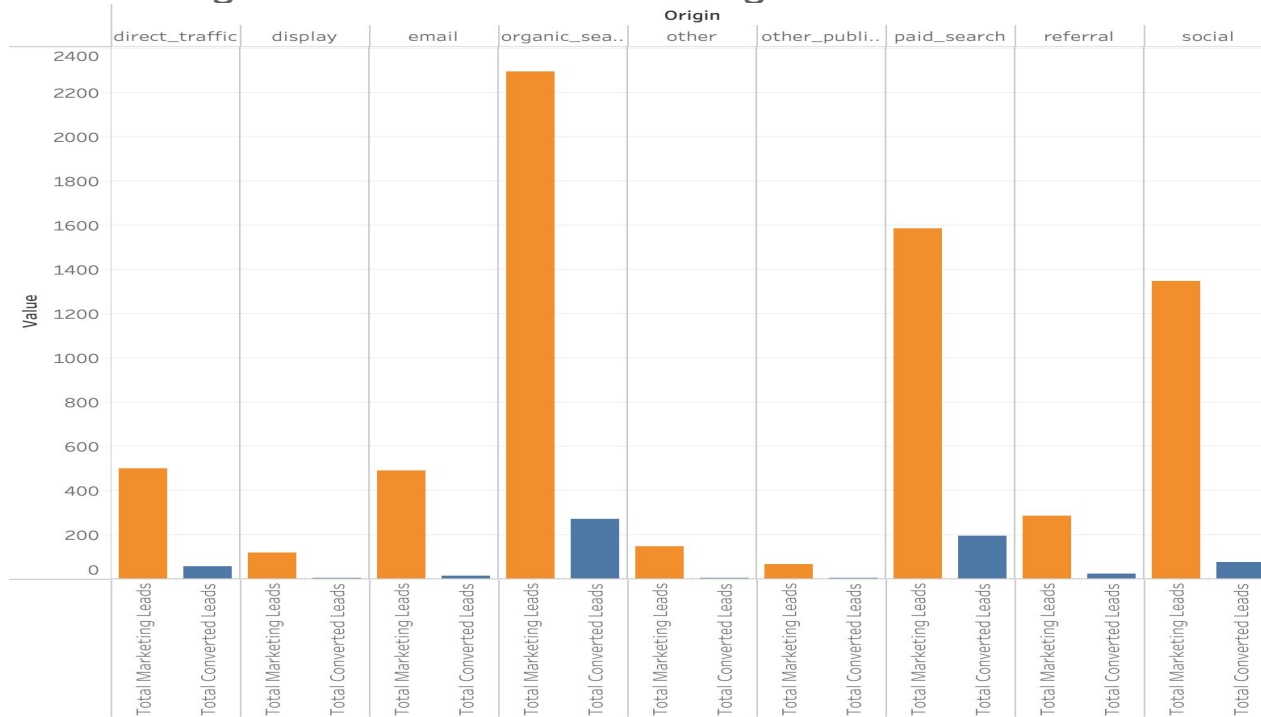


Sellers with Highest Revenue Per Order



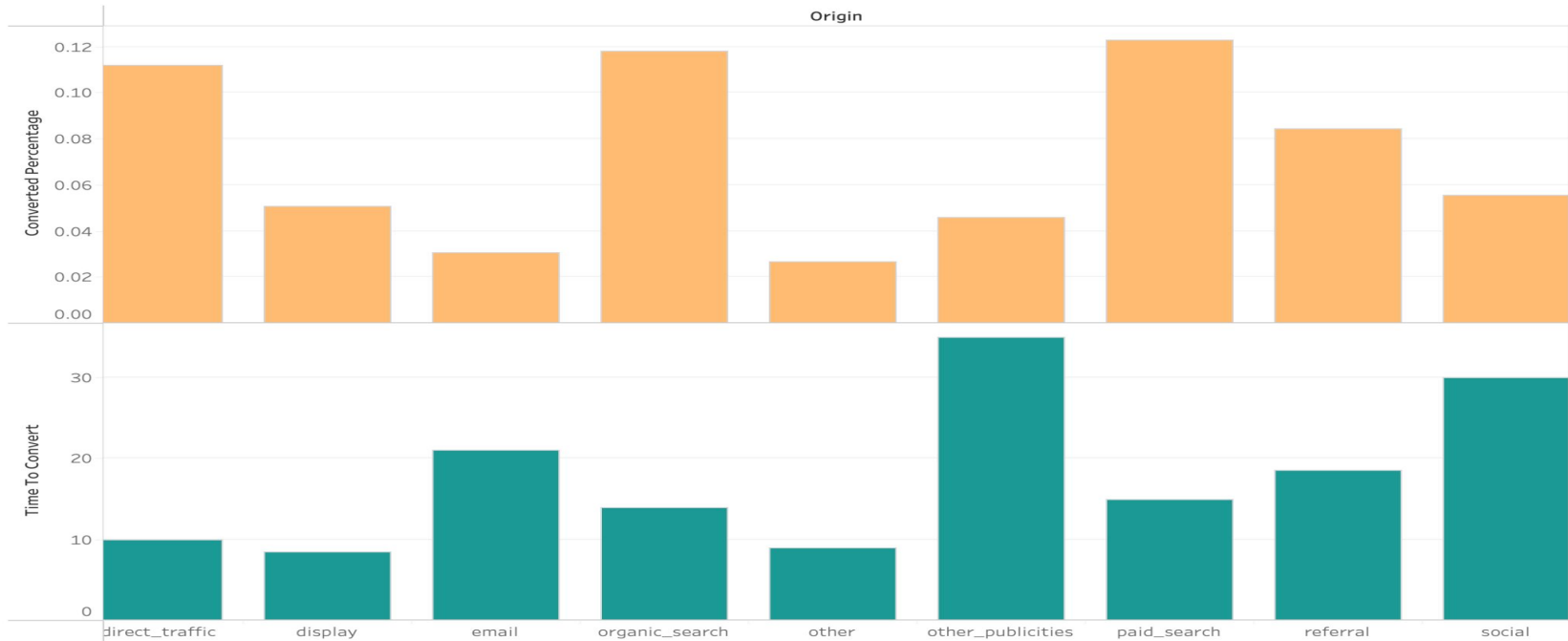
Data Visualisation using Tableau

Marketing Performance - Marketing Leads and Converted Leads



Data Visualisation using Tableau

Marketing Performance - Conversion Rate and Time to convert



Data Manipulation & Transformation - Excel Power Query

Product-Side

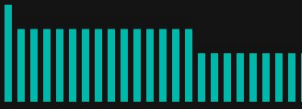



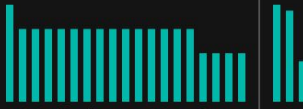
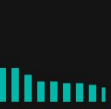

Data Source	Size (#rows)	Primary Keys
Order Items	112.6k	order_id, order_item_id, product_id
Orders	99.4k	order_id and customer_id
Products	33k	product_id
Product Category Translation	72	product_category_name

Data Manipulation & Transformation - Excel Power Query

Product Side

Data Quality Checks

- No duplicates found with respect to primary keys
- Excel Power Query Insights on Column Quality - ensured non-nulls for relevant columns

	order_id	order_item...	product_id	seller_id	shipping_limit...	price	freight
	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>891 distinct, 798 unique</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>4 distinct, 0 unique</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>121 distinct, 57 unique</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>115 distinct, 48 unique</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>890 distinct, 796 unique</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>163 distinct, 6...</div></div>	<div><div><div>Valid 100%</div><div>Error 0%</div><div>Empty 0%</div></div><div></div><div>388 distinct, 388 unique</div></div>
1	00010242fe8c5a6d1ba2dd79...	1	4244733e06e7ecb4970a6e26...	48436dade18ac8b2bce089ec...	9/19/2017, 9:45:35 AM	58.9	
2	130898c0987d1801452a8ed9...	1	4244733e06e7ecb4970a6e26...	48436dade18ac8b2bce089ec...	7/5/2017, 2:44:11 AM	55.9	
3	532ed5e14e24ae1f0d735b915...	1	4244733e06e7ecb4970a6e26...	48436dade18ac8b2bce089ec...	5/23/2018, 10:56:25 AM	64.9	
4	00018f77f2f0320c557190d7a...	1	e5f2d52b802189ee658865ca...	dd7ddc04e1b6c2c614352b38...	5/3/2017, 11:05:13 AM	239.9	
5	048cc42e03ca8d43c729adf6...	1	6a2fb4dd53d2cdb88e0432f1...	7040e82f899a04d1b434b795...	11/23/2017, 9:31:31 PM	16.9	

Data Manipulation & Transformation - Excel Power Query

Product Side

Feature Generation - Orders Dataset

- a. Time For Approval - Difference between purchase time and time when the order was approved (in min)
- b. Delivery Time - Time required for delivery starting from purchase time
- c. Delay - Difference between actual time of delivery and estimated time of delivery

Data Manipulation & Transformation - Excel Power Query

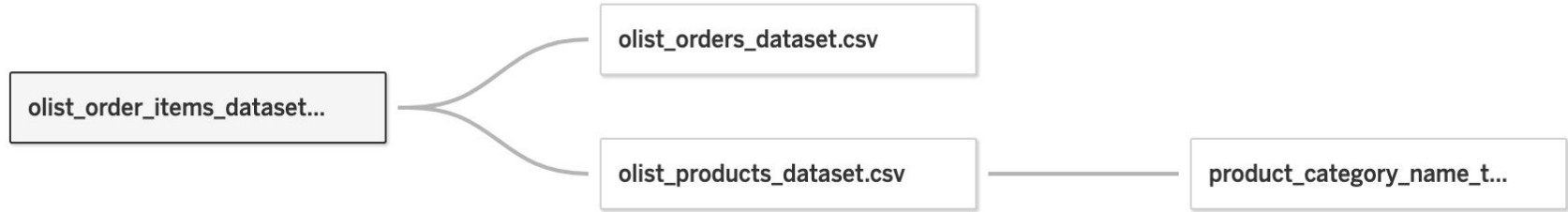
Product Side

Data Integration

Dataset 1	Dataset 2	Key	Join Type	Details
Order Items	Orders	order_id	Left	All entries matched
Products	Product Category Translation	product_category_name	Left	32.33k out of 32.95k matched
Order Items	Products	product_id	Left	All entries matched

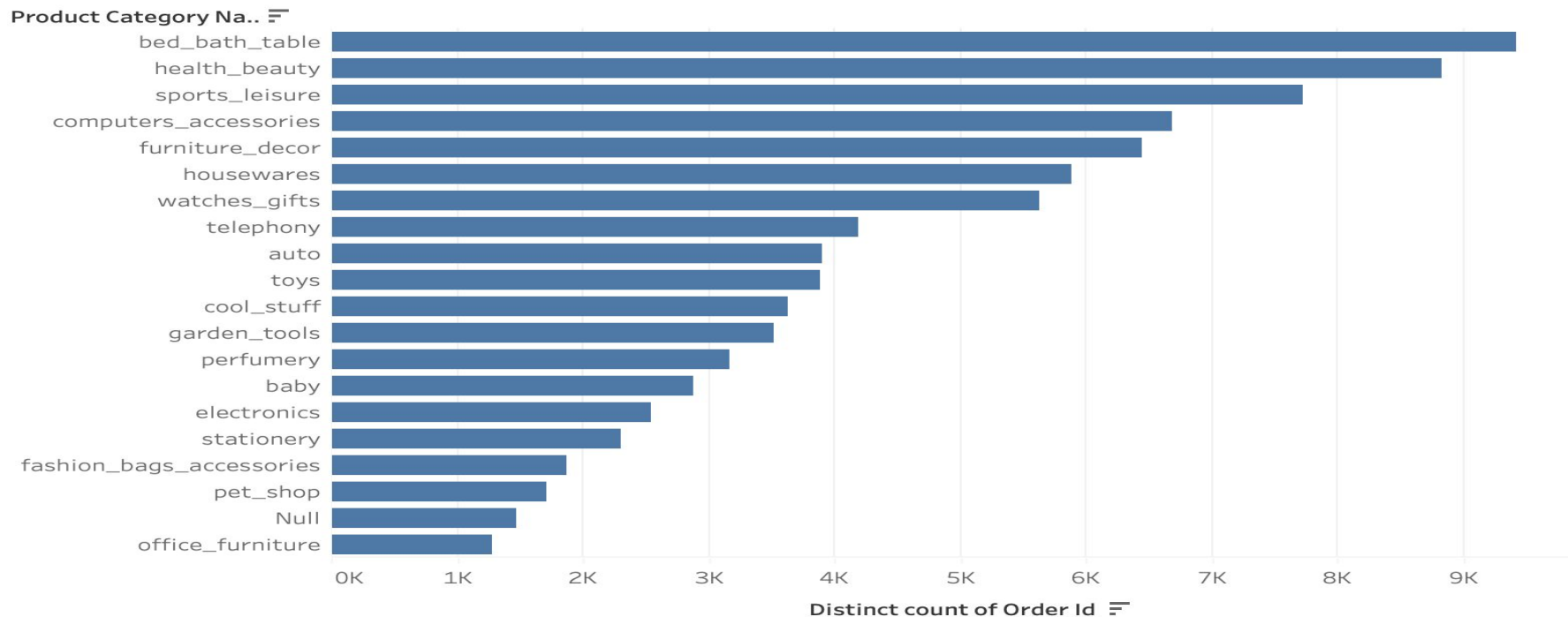
Data Manipulation & Transformation - Excel Power Query

Product Level - Data Integration



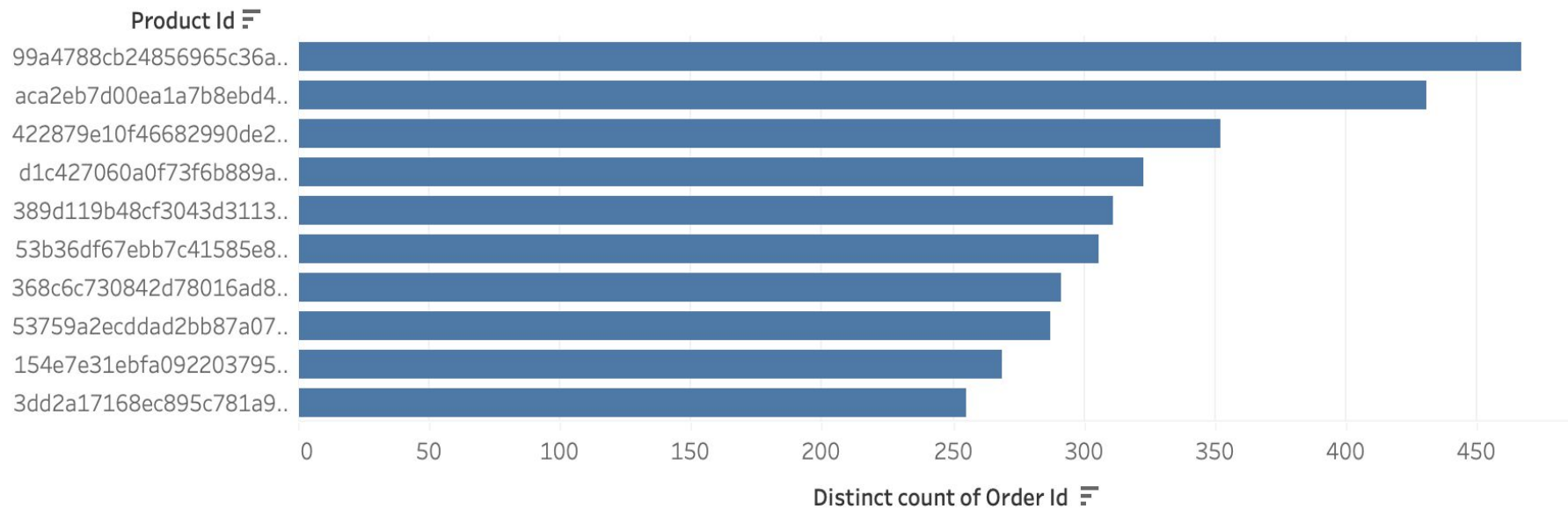
Data Visualisation using Tableau

Product Analysis - Top 10 Product Categories



Data Visualisation using Tableau

Product Analysis - Top 10 Products



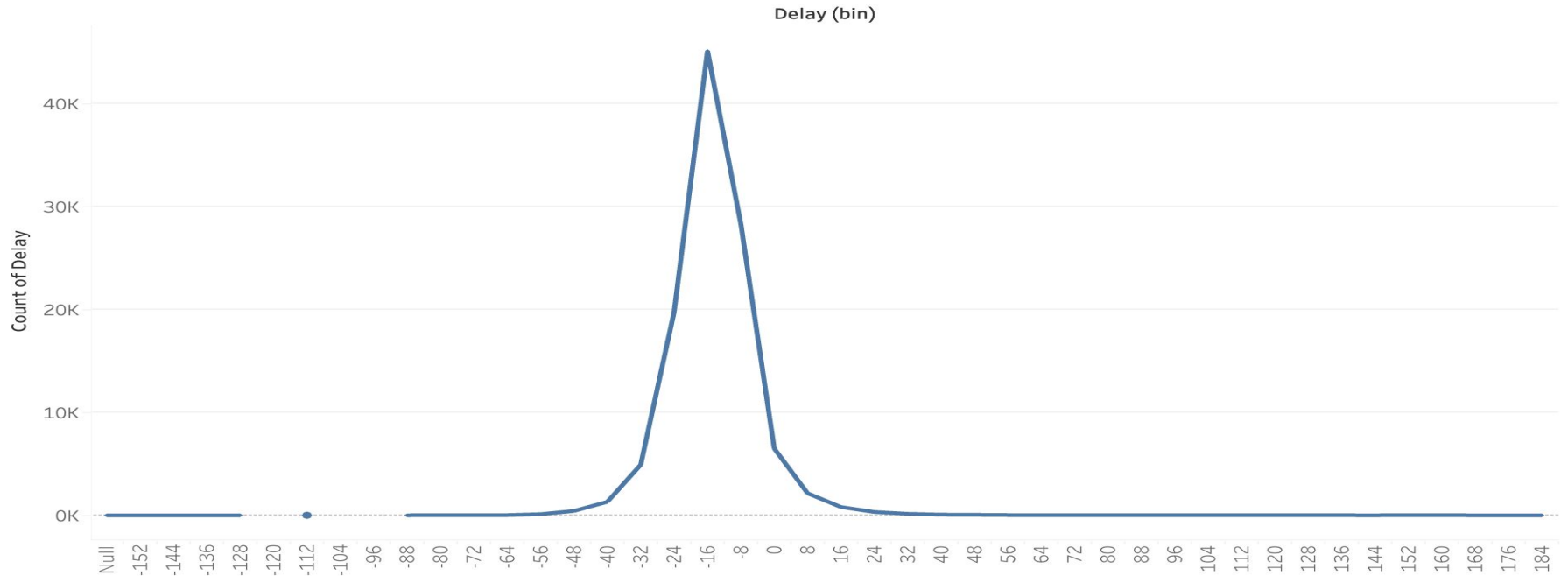
Data Visualisation using Tableau

Product Analysis - Market Basket Analysis

Product Category Name English										
Product Category Na..	auto	bed_bat..	compute..	furnitur..	health_b..	housewa..	sports_l..	telephony	toys	watches..
auto	3,897									
bed_bath_table	2	9,417								
computers_accessories	5	1	6,689							
furniture_decor	2	70	1	6,449						
health_beauty	1	11	1	3	8,836					
housewares	2	20	4	24	1	5,884				
sports_leisure	3	1	3	3	14	11	7,720			
telephony	3		6		2			4,199		
toys		2			1		4	1	3,886	
watches_gifts	1		1	7	3	2	5	2	1	5,624

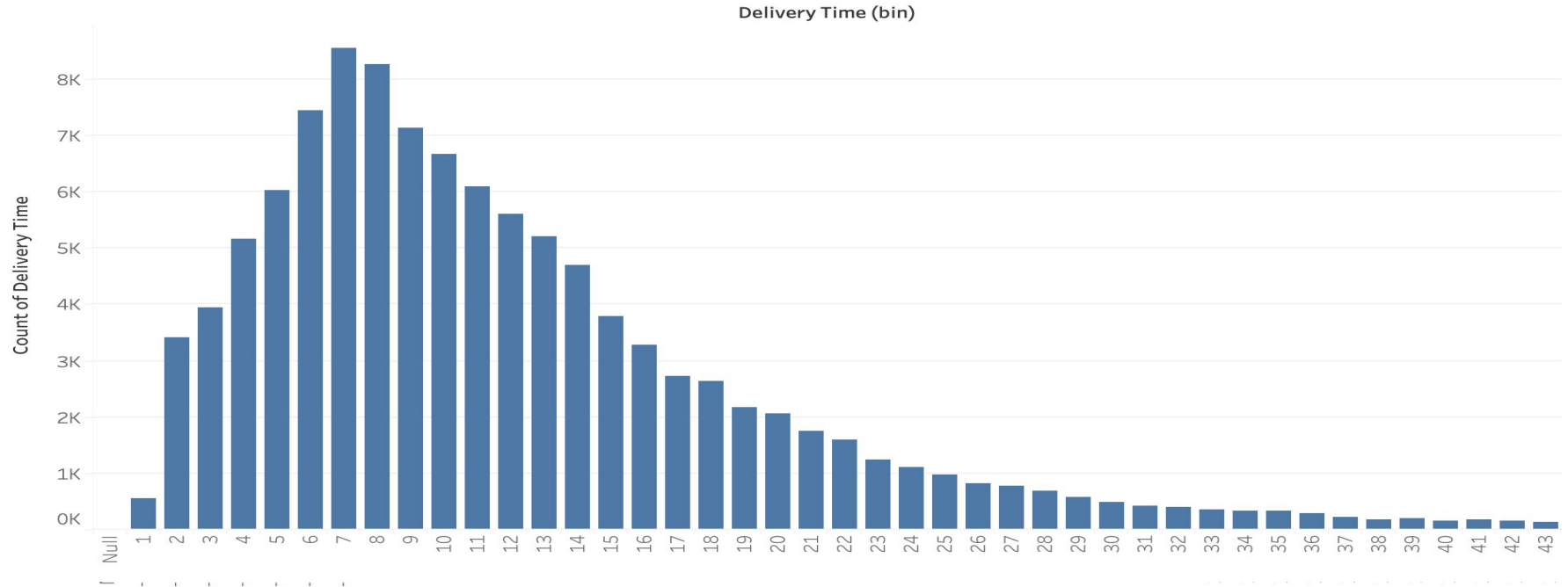
Data Visualisation using Tableau

Delay Distribution



Data Visualisation using Tableau

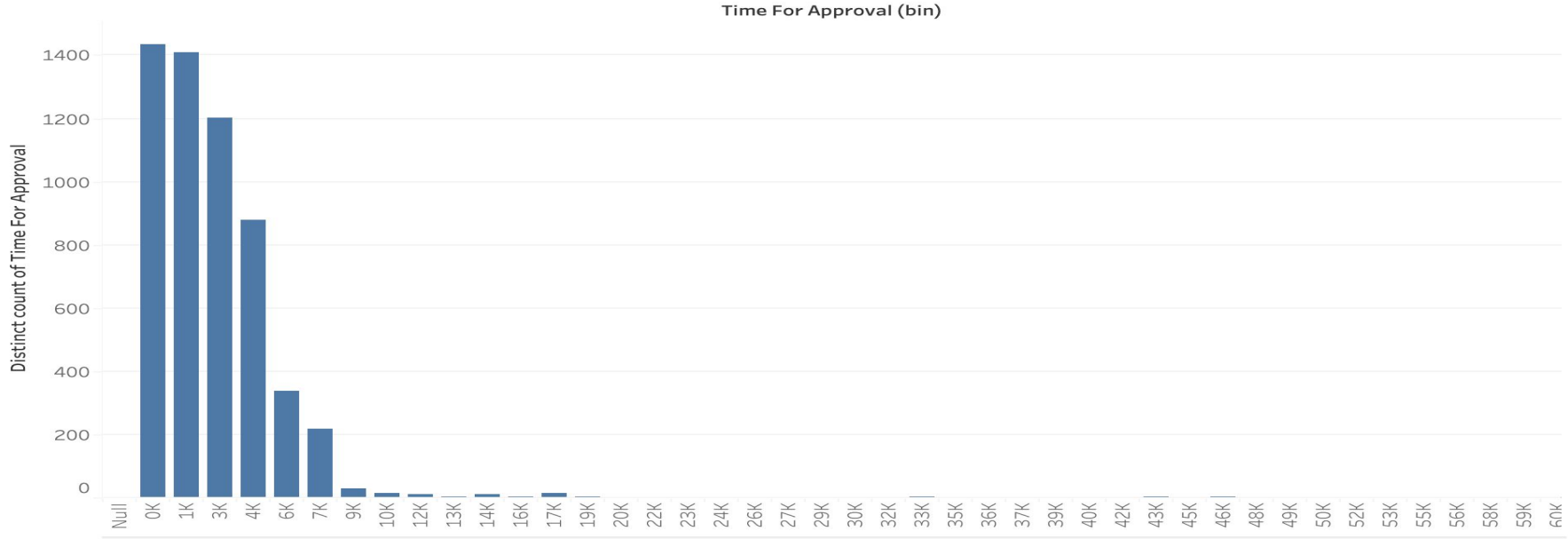
Delivery Time Distribution



Data Visualisation using Tableau

Approval Time Distribution

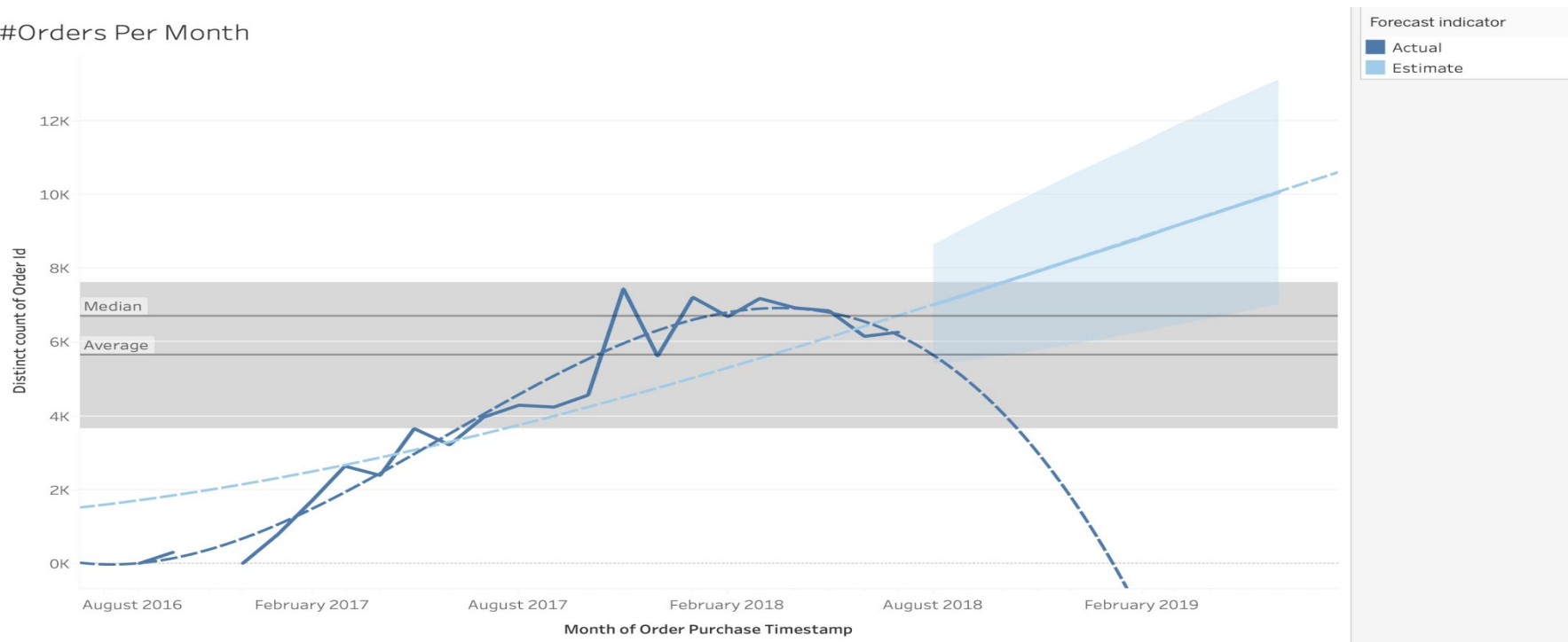
Approval Time Distribution



Data Visualisation using Tableau

Product Analysis - Trends in #Orders over time

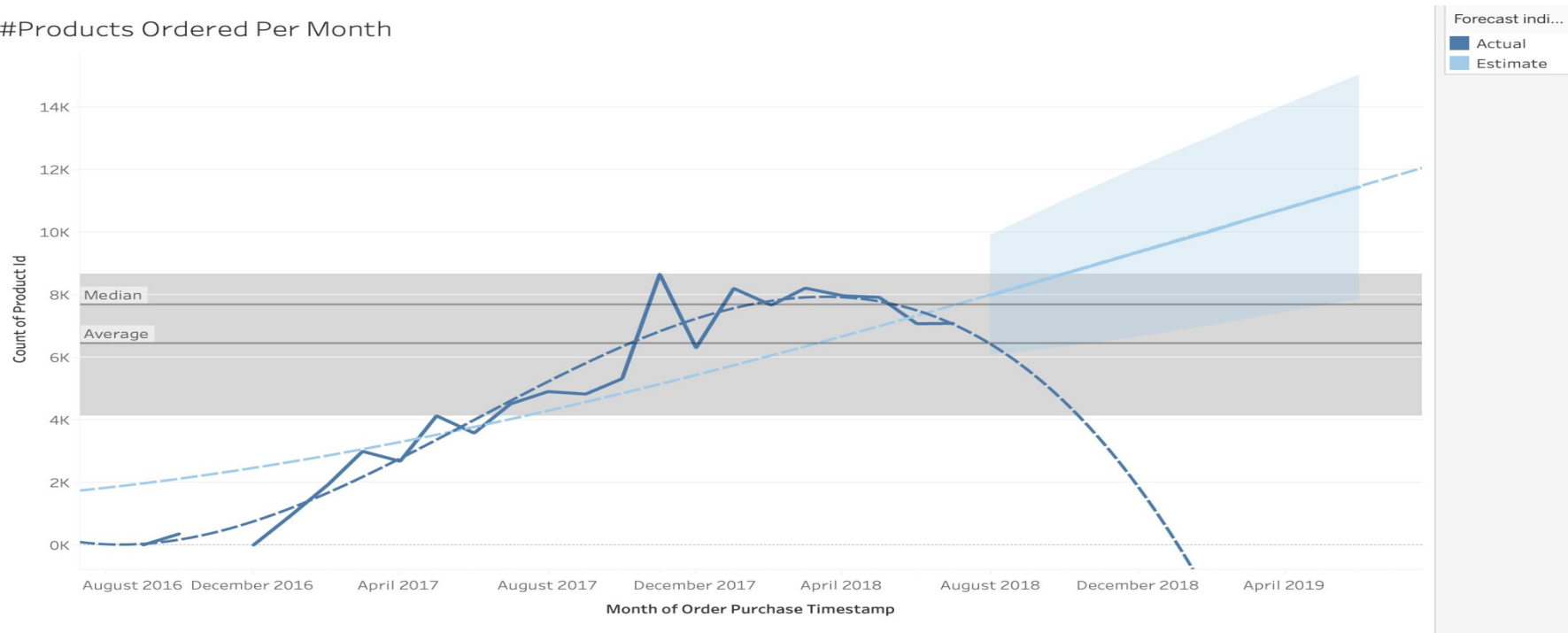
#Orders Per Month



Data Visualisation using Tableau

Product Analysis - Trends in #Products Ordered over time

#Products Ordered Per Month



Data Manipulation & Transformation - Excel Power Query

Customer - Side

Data Source	Size (#rows)	Primary Keys
customers	99.44k	customer_id
Payments	116.34k	order_id, payment_type, payment_sequence, payment_installments, payment_value
Reviews	116.34k	Order_id, review_id

Data Manipulation & Transformation - Excel Power Query

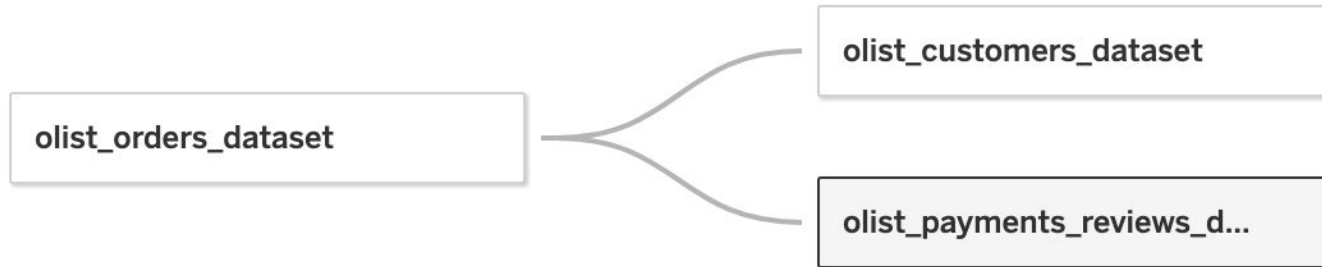
Customer Side

Merging, Feature Generation and Selection - customers

- a. Merged payments dataset with reviews dataset on order_id
- b. Took average of reviews to get unique set of rows based on the columns: order_id, payment_type, payment_sequence, payment_installments, payment_value
- c. Removed irrelevant columns review comments, order_dates etc

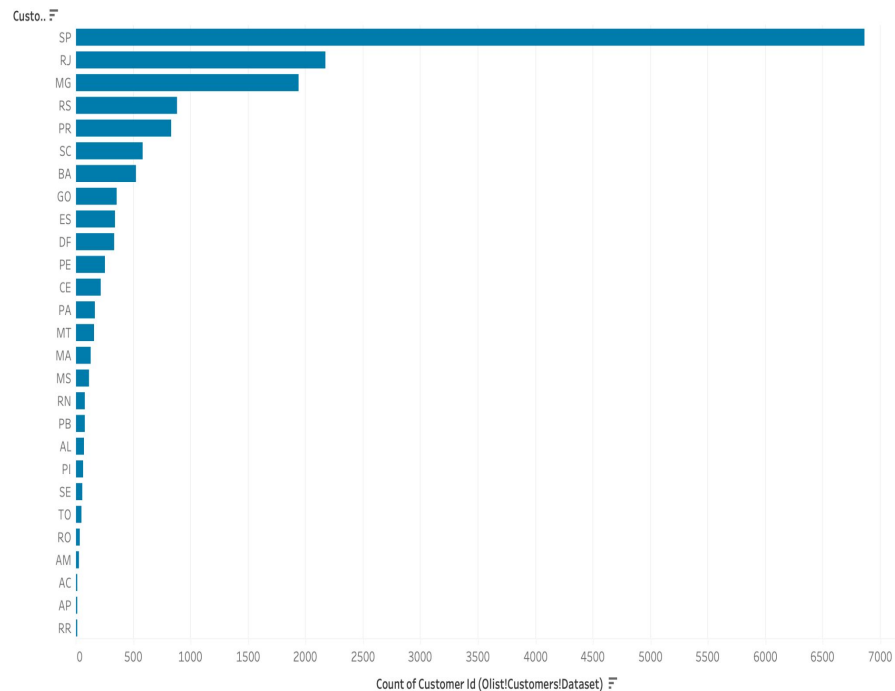
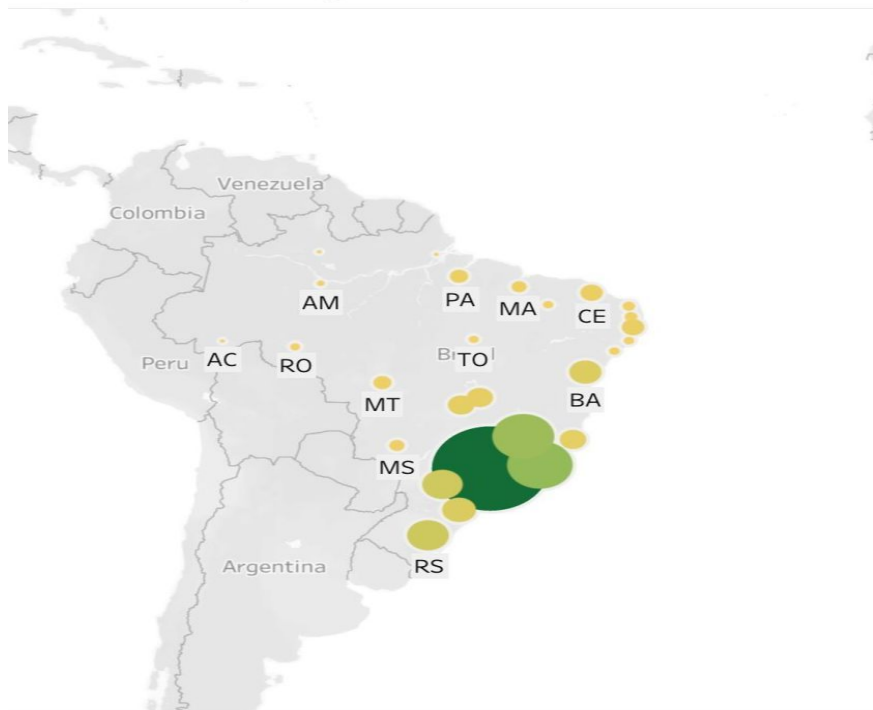
Customer Level Analysis

Data Integration



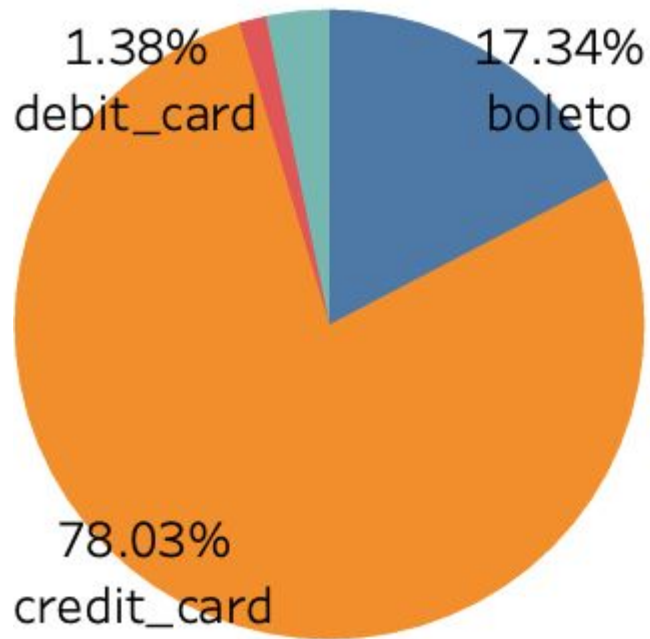
Data Visualisation using Tableau

Customer Analysis



Data Visualisation using Tableau

Payment Analysis



Conclusion and Recommendations

- Identified **high selling product and product categories** on the Olist platform.
- Performed cross category analysis, and observed **few cross selling patterns**.
- Identified the estimated time of delivery are generally higher than actual delivers and hence there are lot of early deliveries. Olist can **promise a faster delivery**, which might **increase their conversion rates**.
- Analyzed average order **approval time and delivery time** which can be used with other features to improve the prediction of estimated time of delivery in future.
- Analyzed sales pattern and created **linear trends to forecast** sales that can be used for better planning (traffic and inventory)
- Identified sellers with low and high performance in terms of **shipping time and delayed deliveries**. Olist should develop strategies and incentives to identify the root cause of these problems and mitigate delayed deliveries to restore customer satisfaction
- Identified **high impact sellers** in terms of revenue and orders. Olist should strengthen its partnership with these brands and sellers as they highly contribute to its business.
- **Organic search and direct traffic** have high conversion rates. On inorganic side, **Paid Search** turns out to be a best channel for marketing.
- Customers have **high affinity for credit card** as evident by its lion share in the pie chart. Olist can partner with the Credit Card Companies to promote various deals on the platform.
- Sao Paulo is the biggest market for Olist. **Huge gap between Sao Paulo and next best state**. Olist can focus on expanding its market by acquiring more customers and sellers in other states

Thank You