SQL DATA ANALYSIS

target

Project Introduction

This project focuses on analyzing sales, customer behavior, and operational data from a retail e-commerce dataset to derive actionable insights that can support decision-making and business optimization. The dataset includes information about customer orders, reviews, payment transactions, order logistics, and seller performance. Through data exploration, aggregation, and visualization, the project aims to uncover key trends, patterns, and areas for improvement within the business.

The primary objectives of this project are as follows:

1. **Sales Analysis**: Understand revenue patterns, order volumes, and the average value of orders over time to identify seasonal trends, peak sales periods, and potential areas for sales growth.
2. **Customer Behavior Analysis**: Investigate customer demographics, purchasing habits, and satisfaction levels, with a focus on analyzing Net Promoter Score (NPS) and customer reviews. This will provide insights into customer preferences and areas needing service improvement.
3. **Operational Efficiency**: Analyze logistics data to assess the efficiency of order fulfillment processes, focusing on delivery times, discrepancies between estimated and actual delivery dates, and freight costs across different regions. This will highlight regions with operational inefficiencies and high costs.
4. **Payment Trends**: Examine payment methods and installment preferences to understand customer purchasing power and offer recommendations for optimizing payment options to boost sales.
5. **Seller Performance**: Evaluate seller performance in terms of the number of orders handled per seller and customer satisfaction associated with different sellers. This will help in identifying high-performing sellers and those who may require support to improve their service levels.

Context:

Target is a globally renowned brand and a prominent retailer in the United States. Target makes itself a preferred shopping destination by offering outstanding value, inspiration, innovation and an exceptional guest experience that no other retailer can deliver.

This particular business case focuses on the operations of Target in Brazil and provides insightful information about 100,000 orders placed between 2016 and 2018. The dataset offers a comprehensive view of various dimensions including the order status, price, payment and freight performance, customer location, product attributes, and customer reviews.

By analyzing this extensive dataset, it becomes possible to gain valuable insights into Target's operations in Brazil. The information can shed light on various aspects of the business, such as order processing, pricing strategies, payment and shipping efficiency, customer demographics, product characteristics, and customer satisfaction levels.

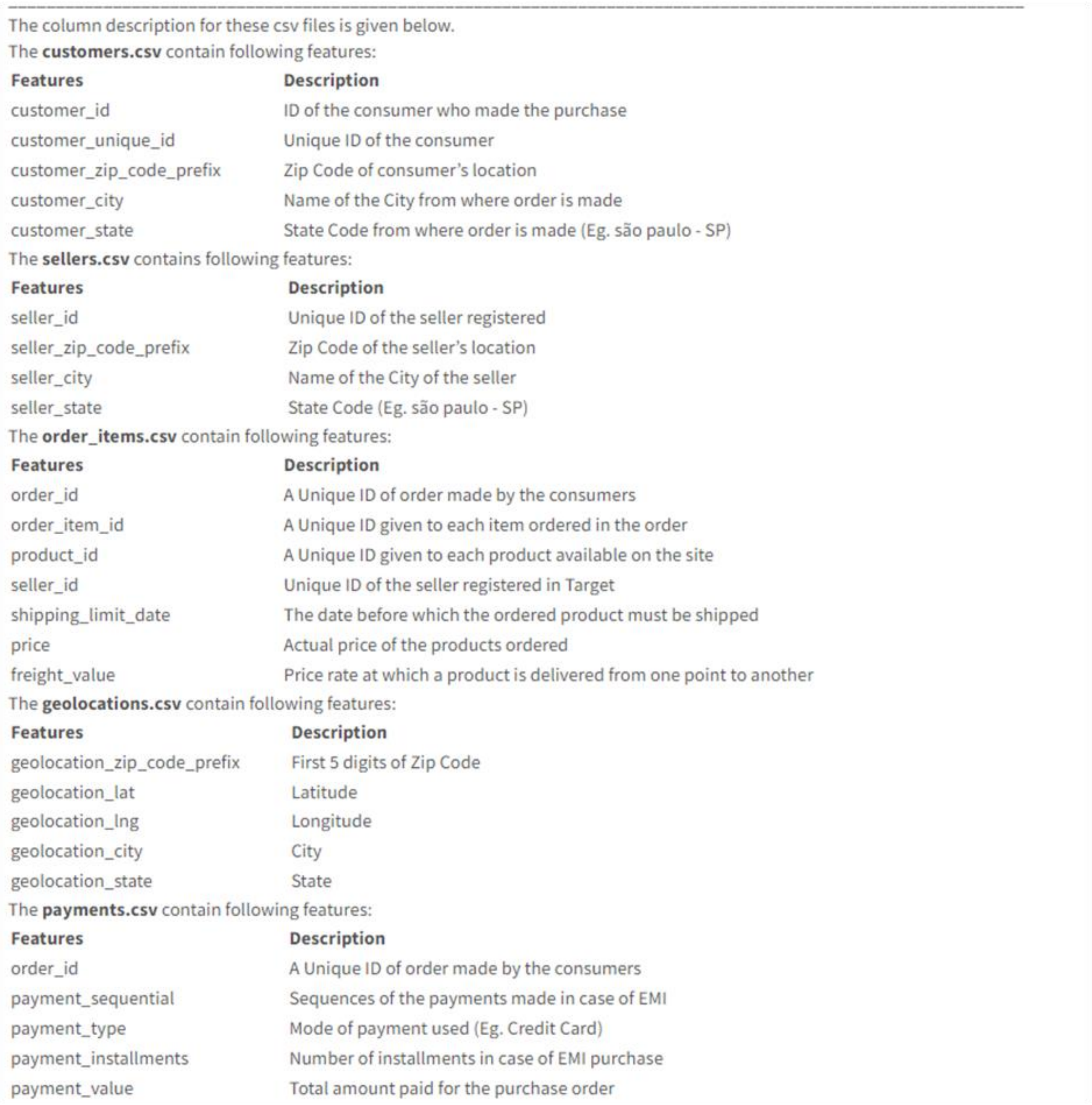
Dataset: https://drive.google.com/drive/folders/1TGEc66YKbD443nslRi1bWgVd238gJCnb

The data is available in 8 csv files:

1. customers.csv
2. sellers.csv
3. order_items.csv
4. geolocation.csv
5. payments.csv
6. reviews.csv
7. orders.csv
8. products.csv

The column description for these csv files is given below.

The customers.csv contain following features:



The column description for these csv files is given below.

The **customers.csv** contain following features:

Features	Description
customer_id	ID of the consumer who made the purchase
customer_unique_id	Unique ID of the consumer
customer_zip_code_prefix	Zip Code of consumer's location
customer_city	Name of the City from where order is made
customer_state	State Code from where order is made (Eg. são paulo - SP)

The **sellers.csv** contains following features:

Features	Description
seller_id	Unique ID of the seller registered
seller_zip_code_prefix	Zip Code of the seller's location
seller_city	Name of the City of the seller
seller_state	State Code (Eg. são paulo - SP)

The **order_items.csv** contain following features:

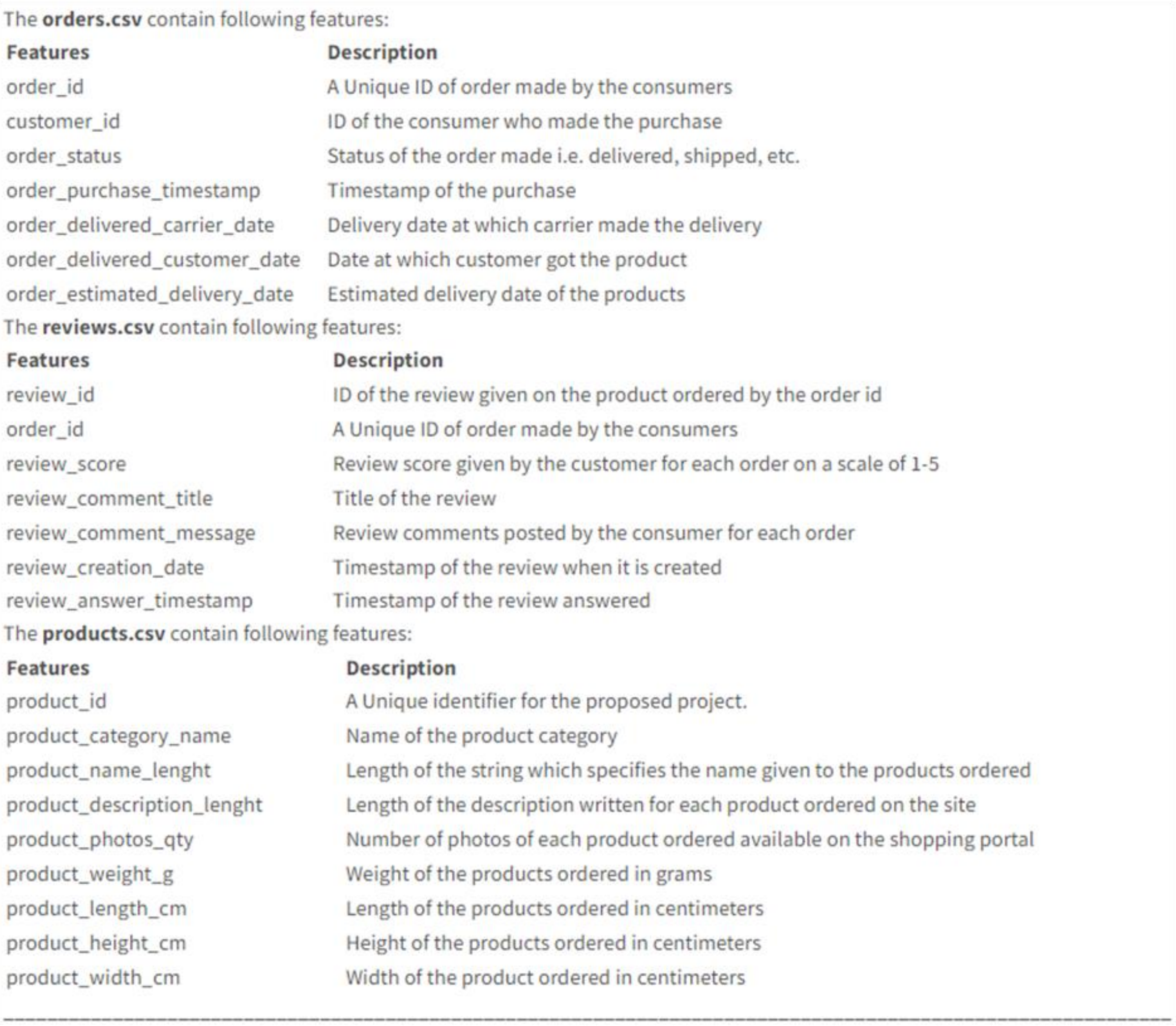
Features	Description
order_id	A Unique ID of order made by the consumers
order_item_id	A Unique ID given to each item ordered in the order
product_id	A Unique ID given to each product available on the site
seller_id	Unique ID of the seller registered in Target
shipping_limit_date	The date before which the ordered product must be shipped
price	Actual price of the products ordered
freight_value	Price rate at which a product is delivered from one point to another

The **geolocations.csv** contain following features:

Features	Description
geolocation_zip_code_prefix	First 5 digits of Zip Code
geolocation_lat	Latitude
geolocation_lng	Longitude
geolocation_city	City
geolocation_state	State

The **payments.csv** contain following features:

Features	Description
order_id	A Unique ID of order made by the consumers
payment_sequential	Sequences of the payments made in case of EMI
payment_type	Mode of payment used (Eg. Credit Card)
payment_installments	Number of installments in case of EMI purchase
payment_value	Total amount paid for the purchase order



The **orders.csv** contain following features:

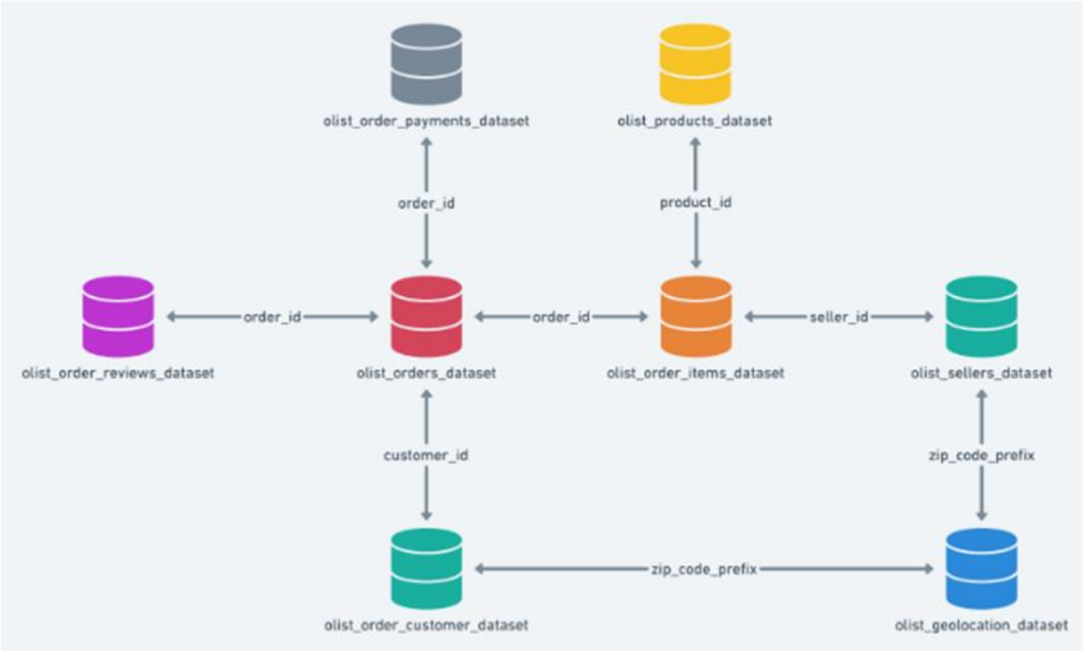
Features	Description
order_id	A Unique ID of order made by the consumers
customer_id	ID of the consumer who made the purchase
order_status	Status of the order made i.e. delivered, shipped, etc.
order_purchase_timestamp	Timestamp of the purchase
order_delivered_carrier_date	Delivery date at which carrier made the delivery
order_delivered_customer_date	Date at which customer got the product
order_estimated_delivery_date	Estimated delivery date of the products

The **reviews.csv** contain following features:

Features	Description
review_id	ID of the review given on the product ordered by the order id
order_id	A Unique ID of order made by the consumers
review_score	Review score given by the customer for each order on a scale of 1-5
review_comment_title	Title of the review
review_comment_message	Review comments posted by the consumer for each order
review_creation_date	Timestamp of the review when it is created
review_answer_timestamp	Timestamp of the review answered

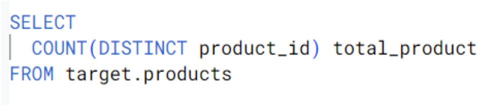
The **products.csv** contain following features:

Features	Description
product_id	A Unique identifier for the proposed project.
product_category_name	Name of the product category
product_name_lenght	Length of the string which specifies the name given to the products ordered
product_description_lenght	Length of the description written for each product ordered on the site
product_photos_qty	Number of photos of each product ordered available on the shopping portal
product_weight_g	Weight of the products ordered in grams
product_length_cm	Length of the products ordered in centimeters
product_height_cm	Height of the products ordered in centimeters
product_width_cm	Width of the product ordered in centimeters



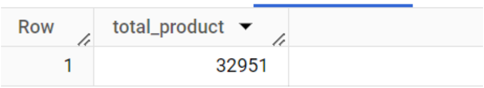
1.1 Total number of products

Query



```
SELECT
  COUNT(DISTINCT product_id) total_product
FROM target.products
```

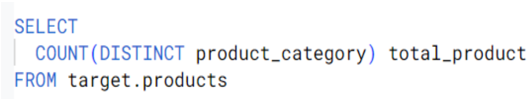
Output



Row	total_product
1	32951

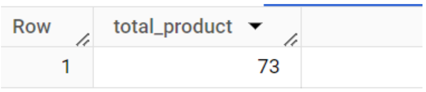
1.2 Total number of product category

Query



```
SELECT
  COUNT(DISTINCT product_category) total_product
FROM target.products
```

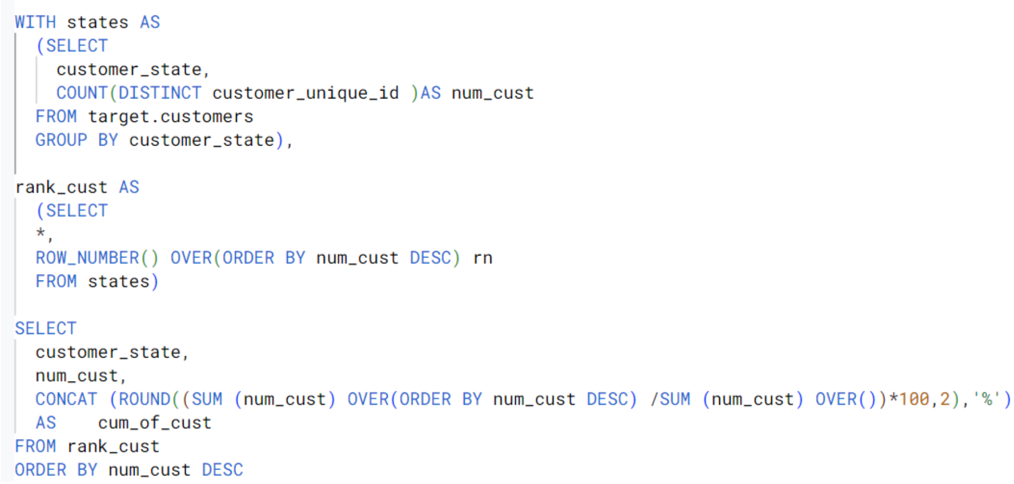
Output



Row	total_product
1	73

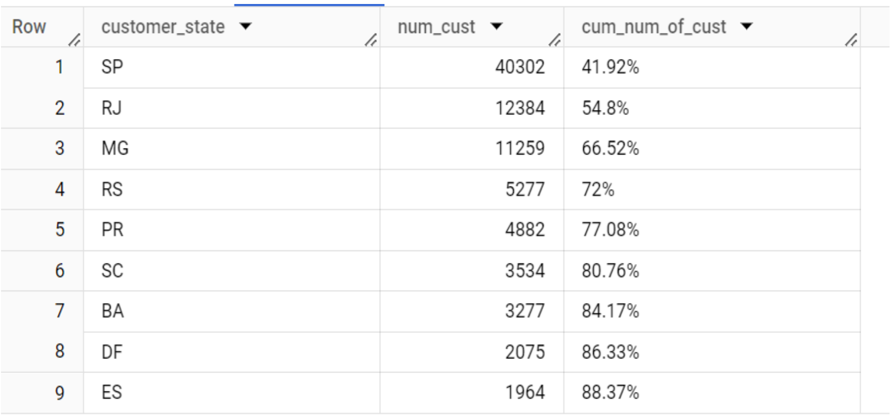
1.3 What is the distribution of unique customers across different states?

Query



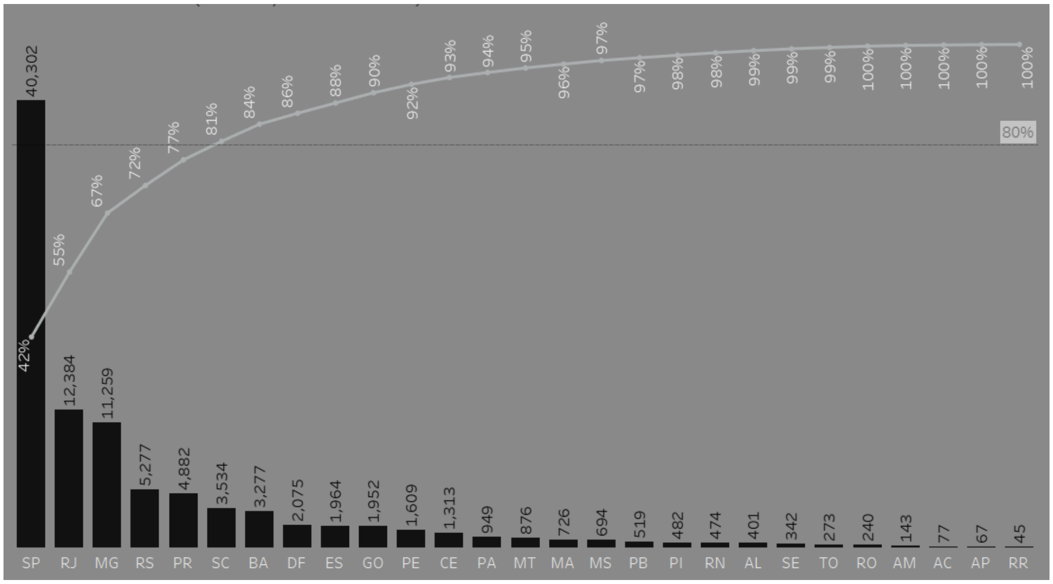
```
WITH states AS
  (SELECT
    customer_state,
    COUNT(DISTINCT customer_unique_id )AS num_cust
  FROM target.customers
  GROUP BY customer_state),

rank_cust AS
  (SELECT
  *,
  ROW_NUMBER() OVER(ORDER BY num_cust DESC) rn
  FROM states)

SELECT
  customer_state,
  num_cust,
  CONCAT (ROUND((SUM (num_cust) OVER(ORDER BY num_cust DESC) /SUM (num_cust) OVER())*100,2),'%')
  AS    cum_of_cust
FROM rank_cust
ORDER BY num_cust DESC
```

Output



Row	customer_state	num_cust	cum_num_of_cust
1	SP	40302	41.92%
2	RJ	12384	54.8%
3	MG	11259	66.52%
4	RS	5277	72%
5	PR	4882	77.08%
6	SC	3534	80.76%
7	BA	3277	84.17%
8	DF	2075	86.33%
9	ES	1964	88.37%

Visualization

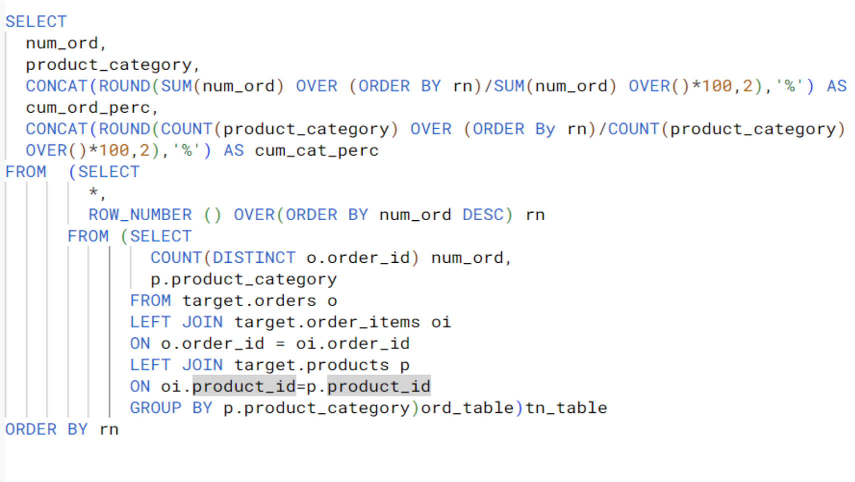


Insights

The cumulative percentage shows that the top three states (SP, RJ, and MG) account for nearly 67% of the total customer base. This insight can be vital for decision-making, particularly in determining where to allocate resources for customer service, marketing campaigns, or even logistics. The fact that a few states contribute to a large portion of the customer base can indicate opportunities for deeper market penetration in those areas.

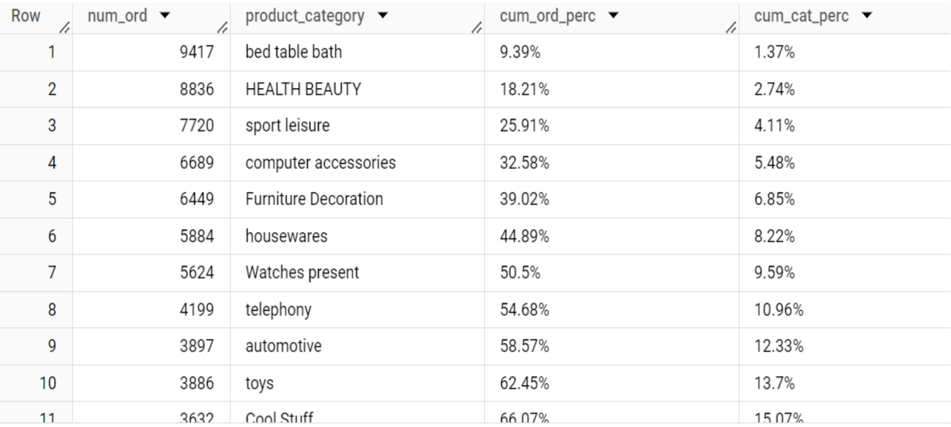
1.4 What are the top product categories based on the number of orders?

Query



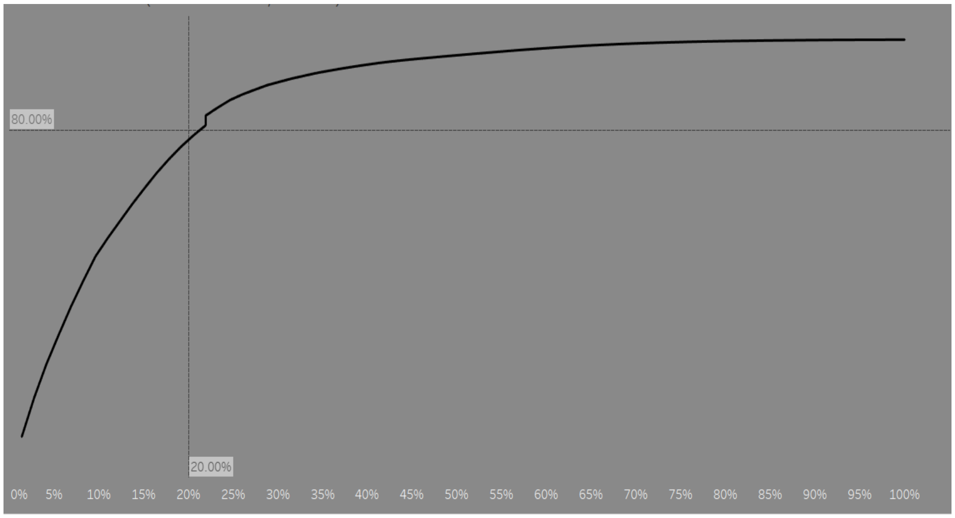
```
SELECT
  num_ord,
  product_category,
  CONCAT(ROUND(SUM(num_ord) OVER (ORDER BY rn)/SUM(num_ord) OVER()*100,2),'%') AS
  cum_ord_perc,
  CONCAT(ROUND(COUNT(product_category) OVER (ORDER By rn)/COUNT(product_category)
  OVER()*100,2),'%') AS cum_cat_perc
FROM  (SELECT
        *,
        ROW_NUMBER () OVER(ORDER BY num_ord DESC) rn
      FROM (SELECT
              COUNT(DISTINCT o.order_id) num_ord,
              p.product_category
            FROM target.orders o
            LEFT JOIN target.order_items oi
            ON o.order_id = oi.order_id
            LEFT JOIN target.products p
            ON oi.product_id=p.product_id
            GROUP BY p.product_category)ord_table)tn_table
ORDER BY rn
```

Output



Row	num_ord	product_category	cum_ord_perc	cum_cat_perc
1	9417	bed table bath	9.39%	1.37%
2	8836	HEALTH BEAUTY	18.21%	2.74%
3	7720	sport leisure	25.91%	4.11%
4	6689	computer accessories	32.58%	5.48%
5	6449	Furniture Decoration	39.02%	6.85%
6	5884	housewares	44.89%	8.22%
7	5624	Watches present	50.5%	9.59%
8	4199	telephony	54.68%	10.96%
9	3897	automotive	58.57%	12.33%
10	3886	toys	62.45%	13.7%
11	3632	Cool Stuff	66.07%	15.07%

Visualization

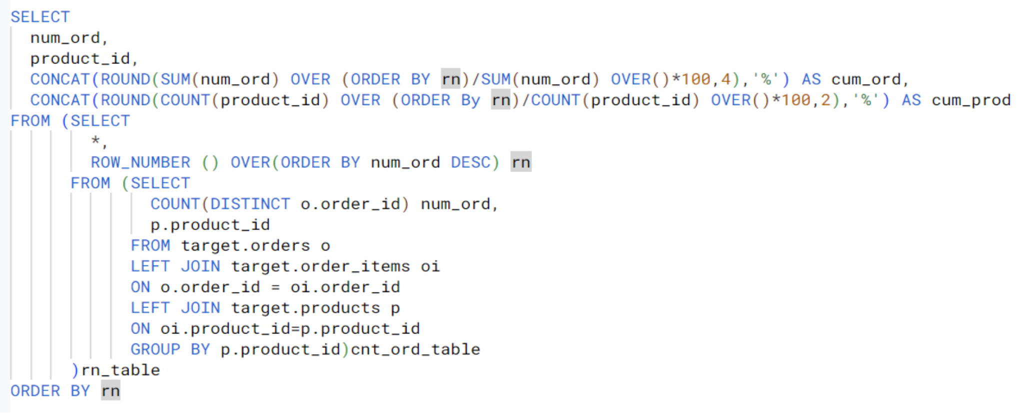


Insights

The results show that "bed table bath" is the top-selling category with 9,417 orders, accounting for 9.39% of total orders. Categories like "HEALTH BEAUTY" and "sport leisure" follow closely behind. These categories collectively represent a significant portion of the total sales, indicating which product segments are most popular among customers.

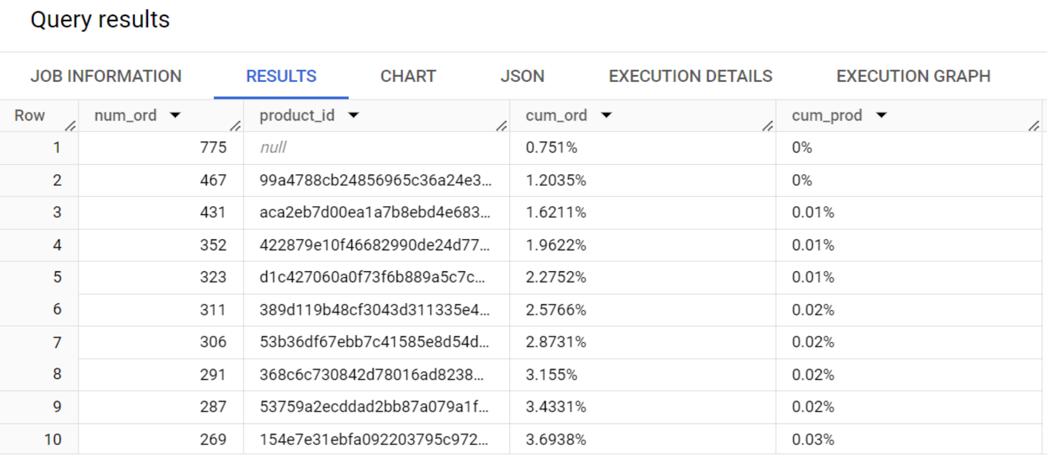
1.5 How can we analyze product performance based on the number of orders and cumulative percentages?

Query



```
SELECT
  num_ord,
  product_id,
  CONCAT(ROUND(SUM(num_ord) OVER (ORDER BY rn)/SUM(num_ord) OVER()*100,4),'%') AS cum_ord,
  CONCAT(ROUND(COUNT(product_id) OVER (ORDER By rn)/COUNT(product_id) OVER()*100,2),'%') AS cum_prod
FROM (SELECT
        *,
        ROW_NUMBER () OVER(ORDER BY num_ord DESC) rn
      FROM (SELECT
              COUNT(DISTINCT o.order_id) num_ord,
              p.product_id
            FROM target.orders o
            LEFT JOIN target.order_items oi
            ON o.order_id = oi.order_id
            LEFT JOIN target.products p
            ON oi.product_id=p.product_id
            GROUP BY p.product_id)cnt_ord_table
      )rn_table
ORDER BY rn
```

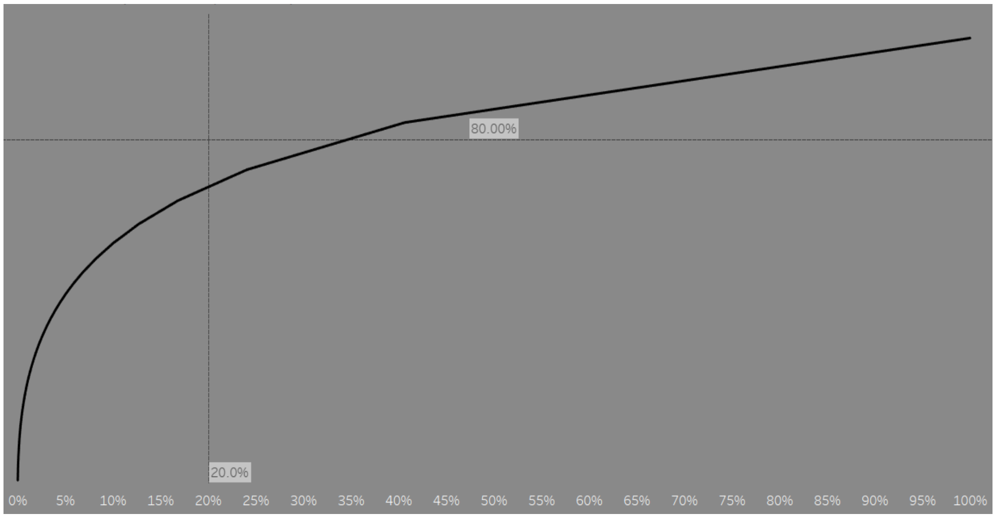
Output



Query results

Row	num_ord	product_id	cum_ord	cum_prod
1	775	null	0.751%	0%
2	467	99a4788cb24856965c36a24e3...	1.2035%	0%
3	431	aca2eb7d00ea1a7b8ebd4e683...	1.6211%	0.01%
4	352	422879e10f46682990de24d77...	1.9622%	0.01%
5	323	d1c427060a0f73f6b889a5c7c...	2.2752%	0.01%
6	311	389d119b48cf3043d311335e4...	2.5766%	0.02%
7	306	53b36df67ebb7c41585e8d54d...	2.8731%	0.02%
8	291	368c6c730842d78016ad8238...	3.155%	0.02%
9	287	53759a2ecddad2bb87a079a1f...	3.4331%	0.02%
10	269	154e7e31ebfa092203795c972...	3.6938%	0.03%

Visualization

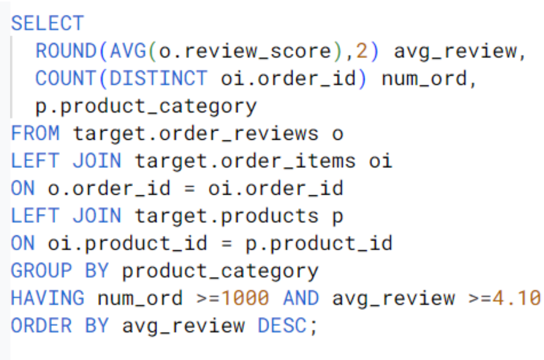


Insights

The top product with 775 orders contributes 0.751% to the overall number of orders. As the rank goes down, products like "99a4788cb24856965c36a24e339b6058" and "aca2eb7d00ea1a7b8ebd4e68314663af" contribute 1.2035% and 1.6211% of orders respectively. This kind of insight is valuable for identifying high-demand products that should be prioritized in inventory management and marketing campaigns.

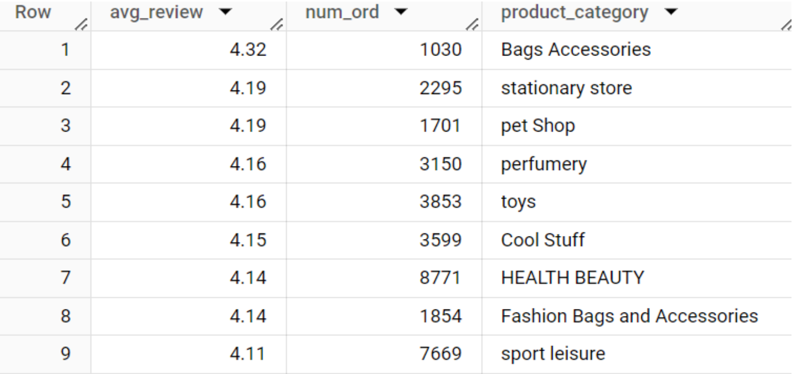
1.6 Which product categories have received the highest average customer reviews and a significant number of orders?

Query



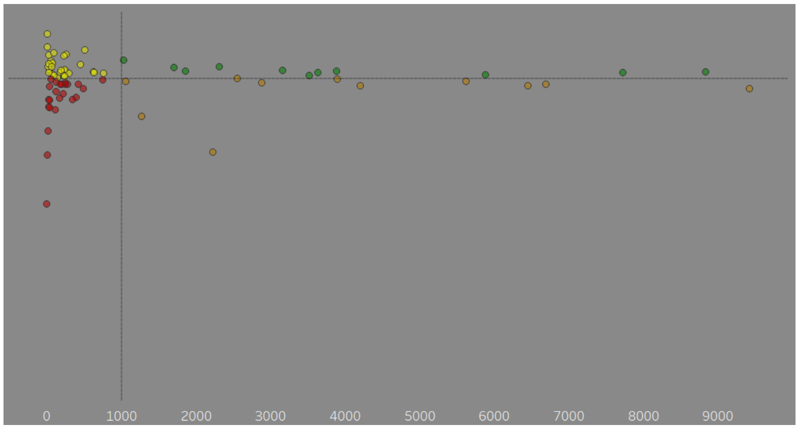
```
SELECT
  ROUND(AVG(o.review_score),2) avg_review,
  COUNT(DISTINCT oi.order_id) num_ord,
  p.product_category
FROM target.order_reviews o
LEFT JOIN target.order_items oi
ON o.order_id = oi.order_id
LEFT JOIN target.products p
ON oi.product_id = p.product_id
GROUP BY product_category
HAVING num_ord >=1000 AND avg_review >=4.10
ORDER BY avg_review DESC;
```

Output



Row	avg_review	num_ord	product_category
1	4.32	1030	Bags Accessories
2	4.19	2295	stationary store
3	4.19	1701	pet Shop
4	4.16	3150	perfumery
5	4.16	3853	toys
6	4.15	3599	Cool Stuff
7	4.14	8771	HEALTH BEAUTY
8	4.14	1854	Fashion Bags and Accessories
9	4.11	7669	sport leisure

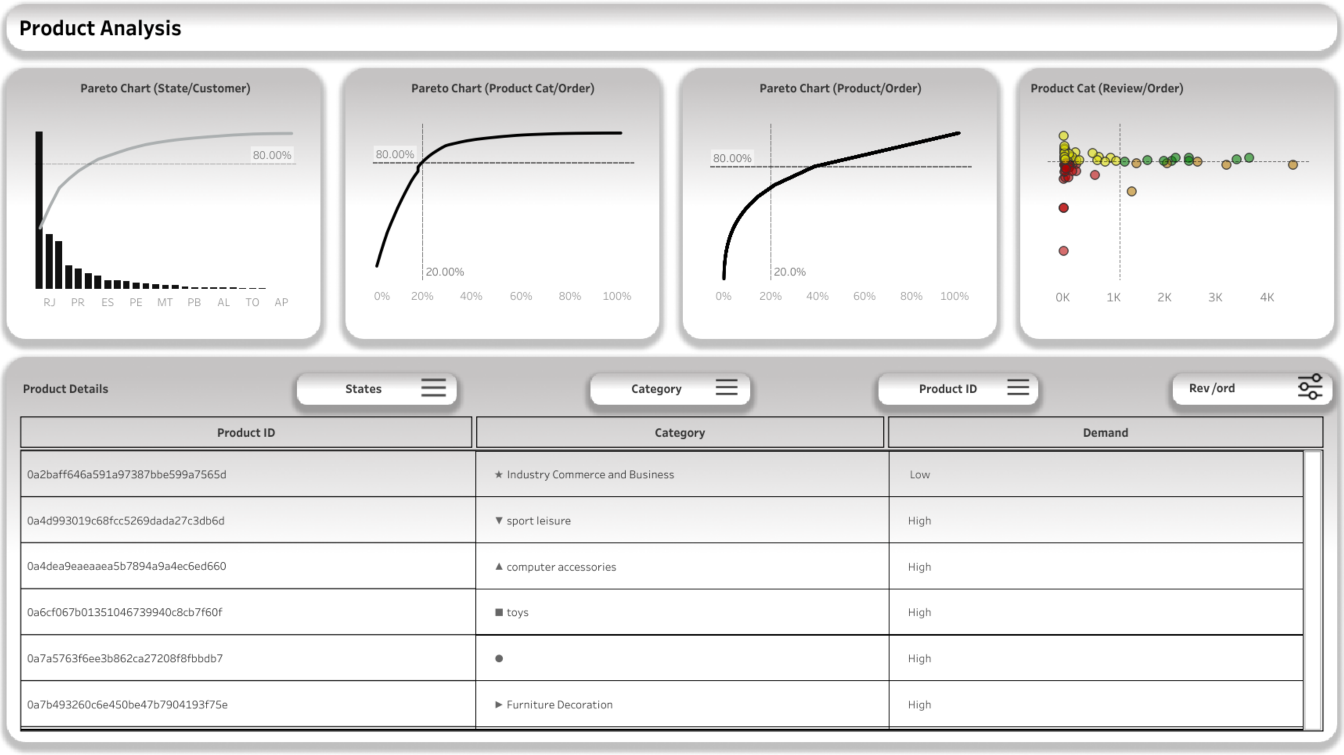
Visualization



Insights

Bags Accessories leads with an average review score of 4.32 across 1,030 orders, making it the highest-rated category. Customers are highly satisfied with products in this category, and despite a lower number of orders compared to other categories, it shows strong customer approval.

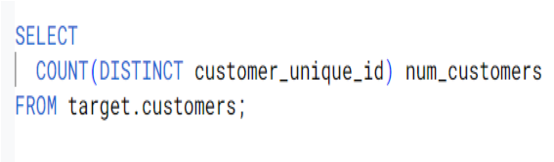
Dashboard



Product ID	Category	Demand
0a2baff646a591a97387bbe599a7565d	★ Industry Commerce and Business	Low
0a4d993019c68fcc5269dada27c3db6d	▼ sport leisure	High
0a4dea9eaeaaea5b7894a9a4ec6ed660	▲ computer accessories	High
0a6cf067b01351046739940c8cb7f60f	■ toys	High
0a7a5763f6ee3b862ca27208f8fbbdb7	●	High
0a7b493260c6e450be47b7904193f75e	▶ Furniture Decoration	High

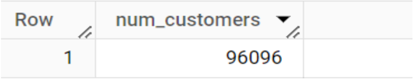
2.1 How many distinct customers are in the database?

Query



```
SELECT
  COUNT(DISTINCT customer_unique_id) num_customers
FROM target.customers;
```

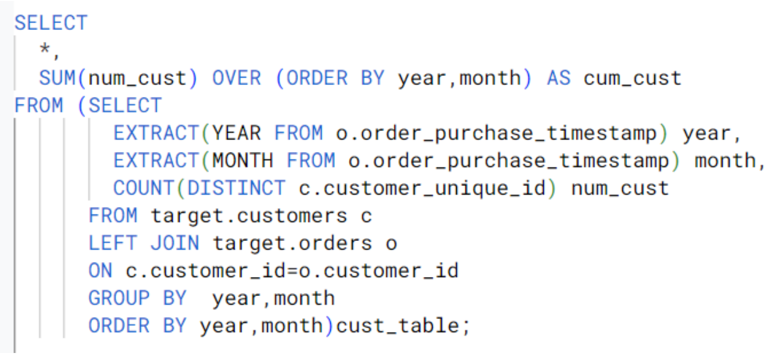
Output



Row	num_customers
1	96096

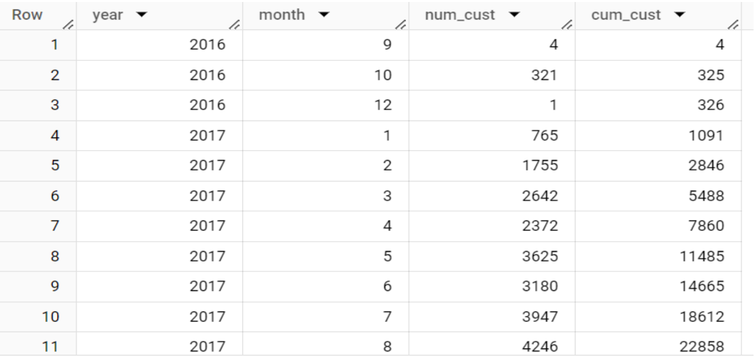
2.2 How has the number of unique customers grown over time, on a monthly basis?

Query



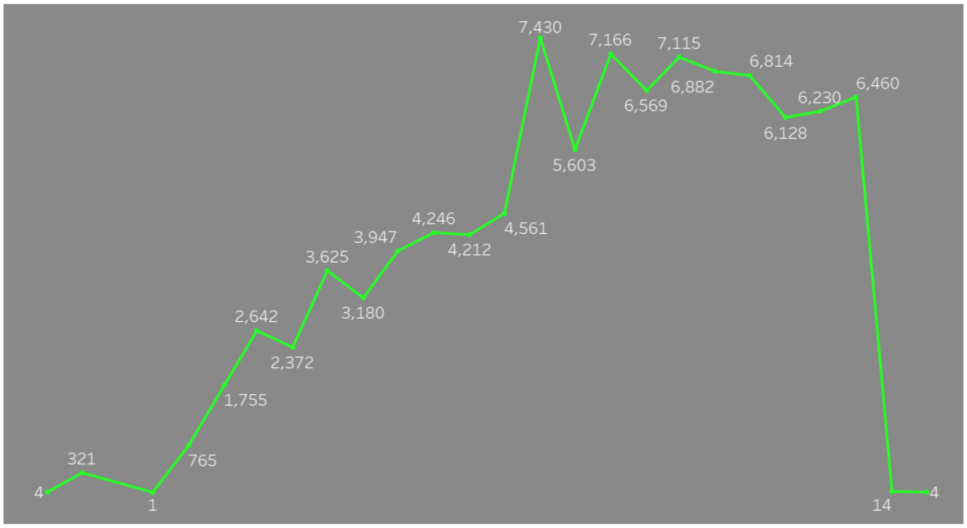
```
SELECT
  *,
  SUM(num_cust) OVER (ORDER BY year,month) AS cum_cust
FROM (SELECT
        EXTRACT(YEAR FROM o.order_purchase_timestamp) year,
        EXTRACT(MONTH FROM o.order_purchase_timestamp) month,
        COUNT(DISTINCT c.customer_unique_id) num_cust
      FROM target.customers c
      LEFT JOIN target.orders o
      ON c.customer_id=o.customer_id
      GROUP BY  year,month
      ORDER BY year,month)cust_table;
```

Output



Row	year	month	num_cust	cum_cust
1	2016	9	4	4
2	2016	10	321	325
3	2016	12	1	326
4	2017	1	765	1091
5	2017	2	1755	2846
6	2017	3	2642	5488
7	2017	4	2372	7860
8	2017	5	3625	11485
9	2017	6	3180	14665
10	2017	7	3947	18612
11	2017	8	4246	22858

Visualization



Insights

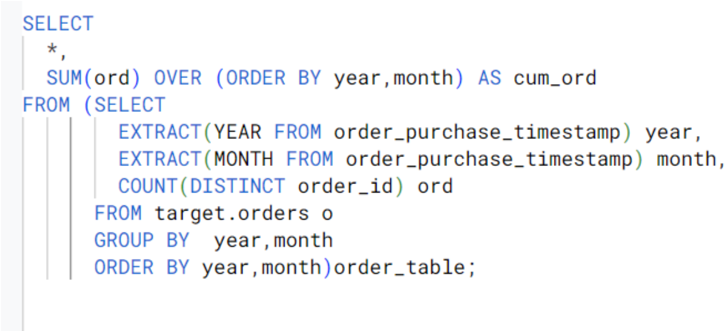
The company started with a few customers in **September 2016** and saw a substantial increase in the customer base as time progressed.

By **October 2017**, the cumulative number of unique customers reached **31,631**, indicating steady growth over the year.

The highest jump in customer acquisition occurred in **November 2017**, with **7,430 unique customers**, likely due to holiday season promotions or significant events.

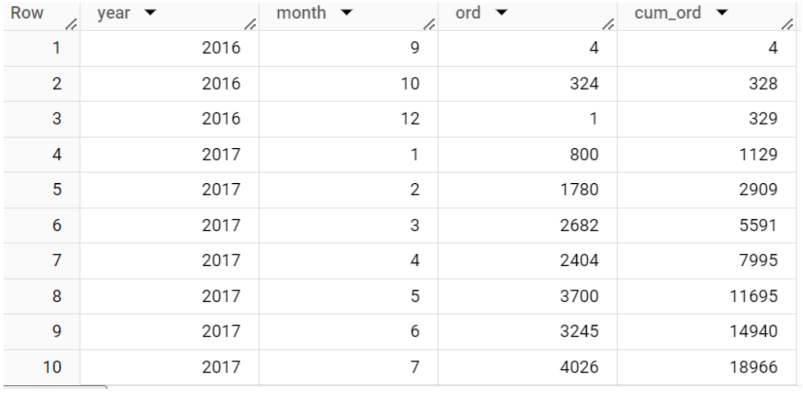
2.3 How has the number of orders grown over time on a monthly basis?

Query



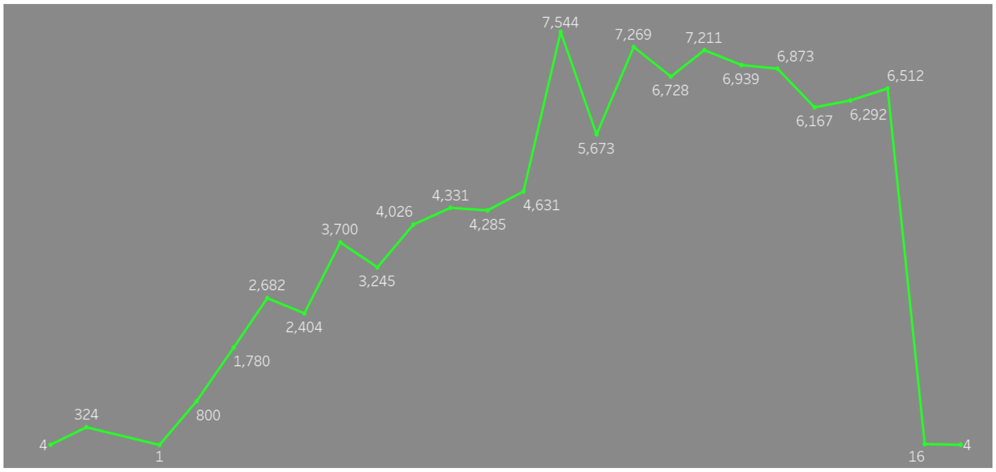
```
SELECT
  *,
  SUM(ord) OVER (ORDER BY year,month) AS cum_ord
FROM (SELECT
        EXTRACT(YEAR FROM order_purchase_timestamp) year,
        EXTRACT(MONTH FROM order_purchase_timestamp) month,
        COUNT(DISTINCT order_id) ord
      FROM target.orders o
      GROUP BY  year,month
      ORDER BY year,month)order_table;
```

Output



Row	year	month	ord	cum_ord
1	2016	9	4	4
2	2016	10	324	328
3	2016	12	1	329
4	2017	1	800	1129
5	2017	2	1780	2909
6	2017	3	2682	5591
7	2017	4	2404	7995
8	2017	5	3700	11695
9	2017	6	3245	14940
10	2017	7	4026	18966

Visualization



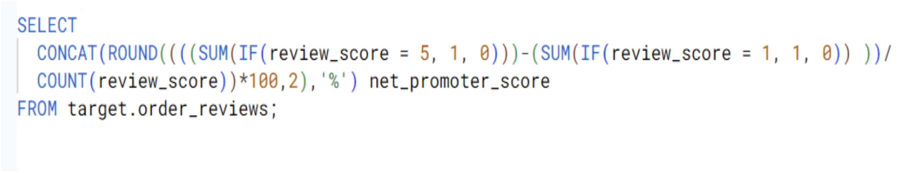
Insights

The number of orders placed saw a gradual rise from **September 2016**, where only **4 orders** were placed, to **April 2018**, where the cumulative number of orders reached **73,577**.

The largest single-month increase occurred in **November 2017** with **7,544 orders**, which could align with promotional events or holiday sales, making it a peak period.

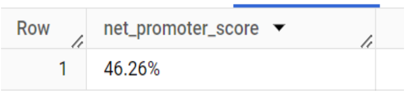
2.4 What is the Net Promoter Score (NPS) of the business based on customer reviews?

Query



```
SELECT
  CONCAT(ROUND((((SUM(IF(review_score = 5, 1, 0)))-(SUM(IF(review_score = 1, 1, 0)) ))/
  COUNT(review_score))*100,2),'%') net_promoter_score
FROM target.order_reviews;
```

Output



Row	net_promoter_score
1	46.26%

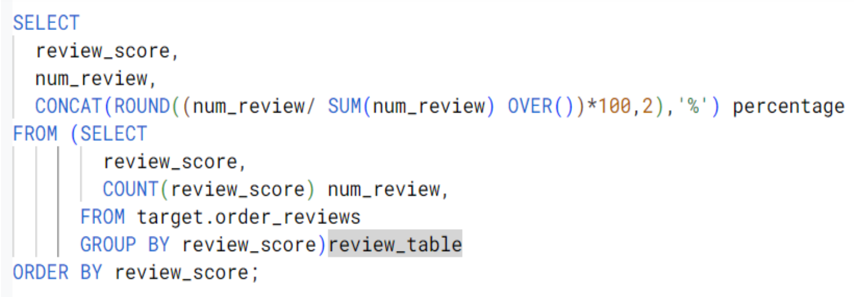
Insights

An NPS of **46.26%** is considered **moderately positive**. This indicates that a good proportion of customers are happy with their purchases and willing to recommend the company, though there is room for improvement.

Typically, an NPS above **50%** is considered excellent, and scores between **30-50%** reflect a satisfactory performance with potential for growth.

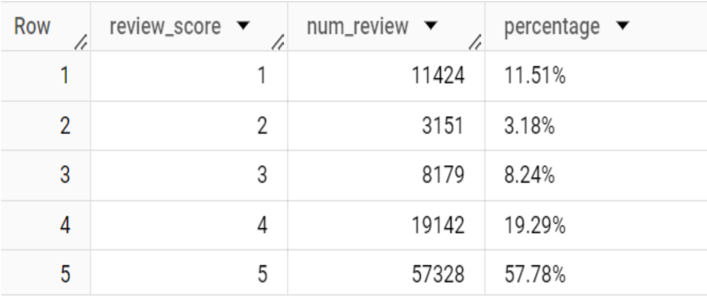
2.5 How are customer reviews distributed across different scores, and what is the percentage representation of each score?

Query



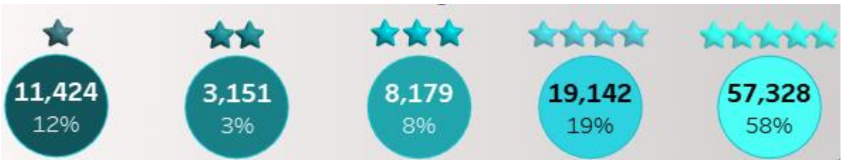
```
SELECT
  review_score,
  num_review,
  CONCAT(ROUND((num_review/ SUM(num_review) OVER())*100,2),'%') percentage
FROM (SELECT
        review_score,
        COUNT(review_score) num_review,
      FROM target.order_reviews
      GROUP BY review_score)review_table
ORDER BY review_score;
```

Output



Row	review_score	num_review	percentage
1	1	11424	11.51%
2	2	3151	3.18%
3	3	8179	8.24%
4	4	19142	19.29%
5	5	57328	57.78%

Visualization



Insights

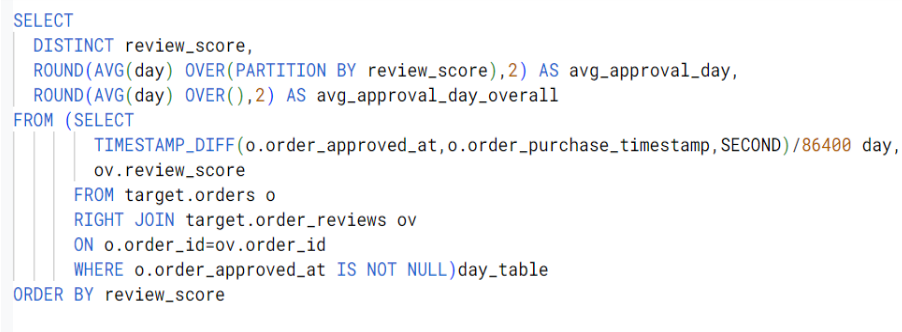
Majority Score: The **majority of reviews** (57.78%) are **5-star**, indicating a **strong level of customer satisfaction**. This is a positive sign, suggesting that most customers are happy with their purchases.

Detractors (Score 1 & 2): Combined, these scores represent only **14.69%** of total reviews, suggesting that the percentage of unsatisfied customers is relatively low.

Promoters (Score 4 & 5): Combined, they account for **77.07%**, indicating a high level of customer loyalty and satisfaction.

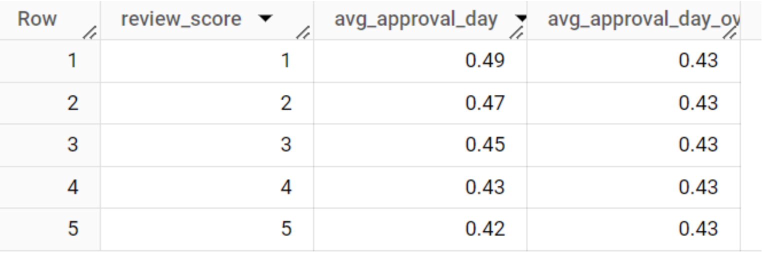
2.6 What is the average number of days taken for order approvals based on customer review scores?

Query



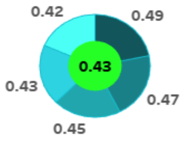
```
SELECT
  DISTINCT review_score,
  ROUND(AVG(day) OVER(PARTITION BY review_score),2) AS avg_approval_day,
  ROUND(AVG(day) OVER(),2) AS avg_approval_day_overall
FROM (SELECT
        TIMESTAMP_DIFF(o.order_approved_at,o.order_purchase_timestamp,SECOND)/86400 day,
        ov.review_score
      FROM target.orders o
      RIGHT JOIN target.order_reviews ov
      ON o.order_id=ov.order_id
      WHERE o.order_approved_at IS NOT NULL)day_table
ORDER BY review_score
```

Output



Row	review_score	avg_approval_day	avg_approval_day_ov
1	1	0.49	0.43
2	2	0.47	0.43
3	3	0.45	0.43
4	4	0.43	0.43
5	5	0.42	0.43

Visualization



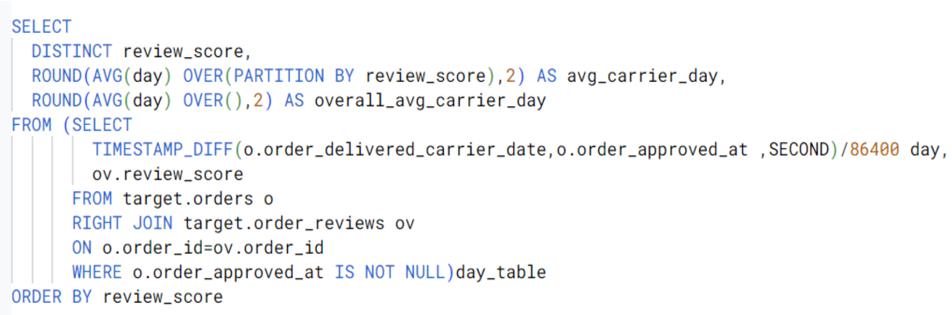
Insights

The average approval time decreases as the review score increases, suggesting that orders with higher satisfaction ratings tend to be approved slightly faster than those with lower scores.

Customers giving **1-star** ratings had the longest average approval time (0.49 days), indicating potential issues that may be reflected in their reviews.

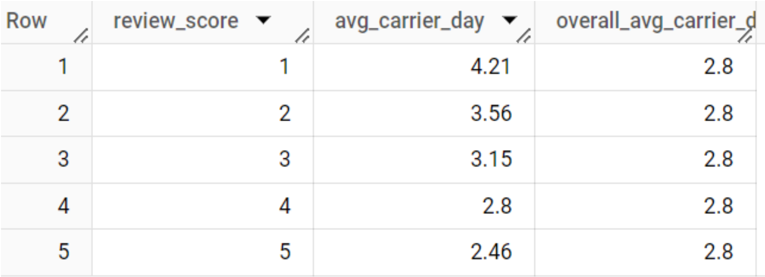
2.7 How do delivery carrier times vary by customer review score?

Query



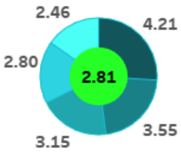
```
SELECT
  DISTINCT review_score,
  ROUND(AVG(day) OVER(PARTITION BY review_score),2) AS avg_carrier_day,
  ROUND(AVG(day) OVER(),2) AS overall_avg_carrier_day
FROM (SELECT
        TIMESTAMP_DIFF(o.order_delivered_carrier_date,o.order_approved_at ,SECOND)/86400 day,
        ov.review_score
      FROM target.orders o
      RIGHT JOIN target.order_reviews ov
      ON o.order_id=ov.order_id
      WHERE o.order_approved_at IS NOT NULL)day_table
ORDER BY review_score
```

Output



Row	review_score	avg_carrier_day	overall_avg_carrier_d
1	1	4.21	2.8
2	2	3.56	2.8
3	3	3.15	2.8
4	4	2.8	2.8
5	5	2.46	2.8

Visualization



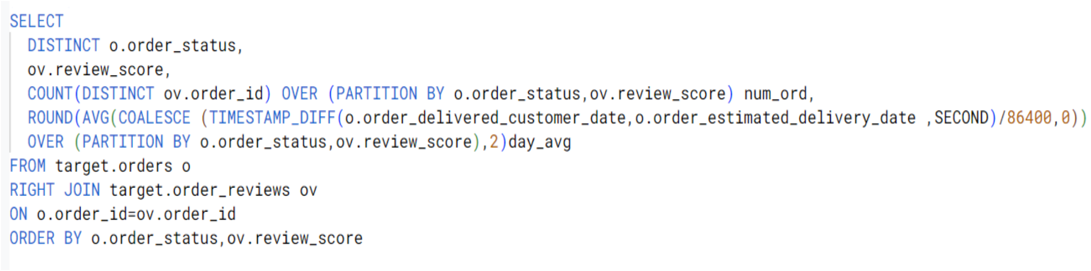
Insights

The data suggests that higher review scores are associated with shorter delivery times to the carrier. Customers who rated their experience **1-star** experienced the longest delay in carrier delivery (4.21 days on average), while **5-star** ratings are linked to quicker deliveries (2.46 days on average).

As the review score increases, the delivery to carrier time tends to decrease, indicating that faster delivery may contribute to higher customer satisfaction.

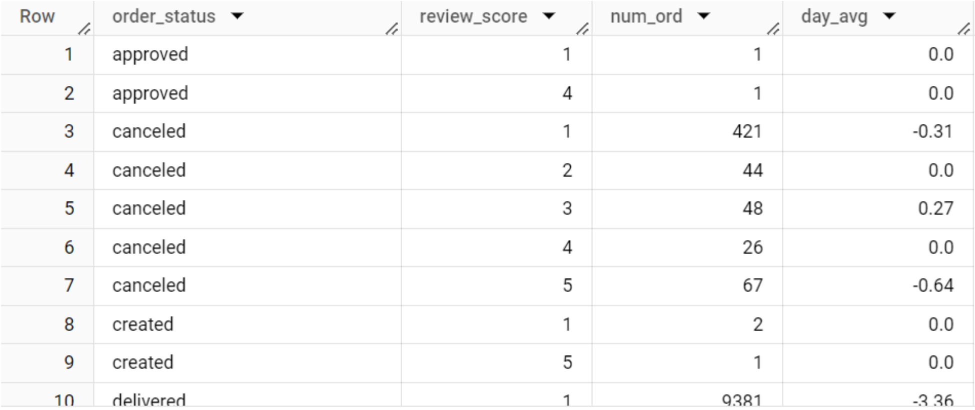
2.8 Analysis of Order Status and Review Score with Delivery Time

Query



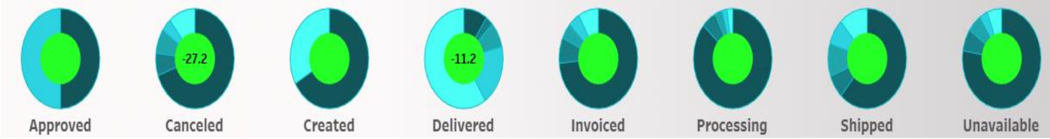
```
SELECT
  DISTINCT o.order_status,
  ov.review_score,
  COUNT(DISTINCT ov.order_id) OVER (PARTITION BY o.order_status,ov.review_score) num_ord,
  ROUND(AVG(COALESCE (TIMESTAMP_DIFF(o.order_delivered_customer_date,o.order_estimated_delivery_date ,SECOND)/86400,0))
  OVER (PARTITION BY o.order_status,ov.review_score),2)day_avg
FROM target.orders o
RIGHT JOIN target.order_reviews ov
ON o.order_id=ov.order_id
ORDER BY o.order_status,ov.review_score
```

Output



Row	order_status	review_score	num_ord	day_avg
1	approved	1	1	0.0
2	approved	4	1	0.0
3	canceled	1	421	-0.31
4	canceled	2	44	0.0
5	canceled	3	48	0.27
6	canceled	4	26	0.0
7	canceled	5	67	-0.64
8	created	1	2	0.0
9	created	5	1	0.0
10	delivered	1	9381	-3.36

Visualization



Insights

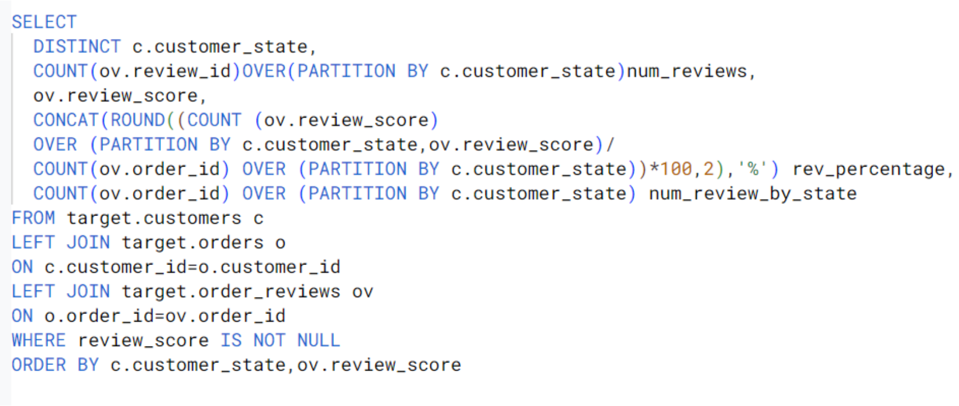
Early Delivery Impact: Orders being delivered significantly earlier than the estimated date (-11.22 days) could positively influence review scores, especially with a majority of 5-star ratings.

The consistent early delivery across all review scores may highlight that factors beyond just delivery timeliness contribute to lower review ratings (1-star or 2-star).

Review Variance: Even though the order was canceled, customers still left reviews. Interestingly, some canceled orders received 5-star reviews, which may suggest that customer service and handling of the cancellation process were satisfactory in certain cases.

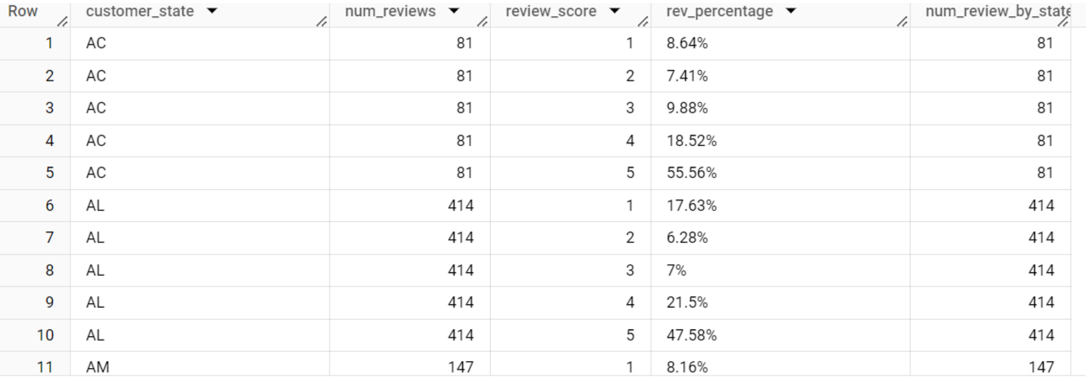
2.9 Provides the count of reviews, their associated scores, and the percentage breakdown of those scores by state.

Query



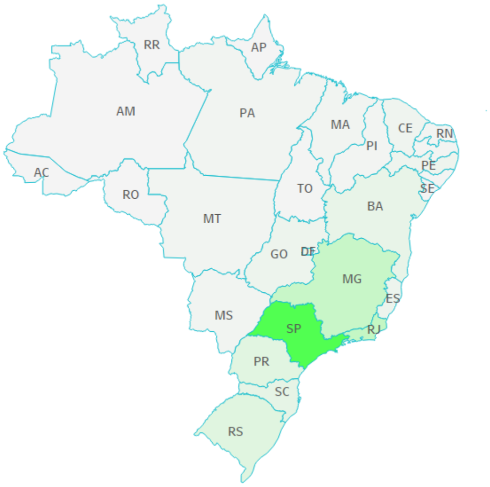
```sql
SELECT
  DISTINCT c.customer_state,
  COUNT(ov.review_id)OVER(PARTITION BY c.customer_state)num_reviews,
  ov.review_score,
  CONCAT(ROUND((COUNT (ov.review_score)
  OVER (PARTITION BY c.customer_state,ov.review_score)/
  COUNT(ov.order_id) OVER (PARTITION BY c.customer_state))*100,2),'%') rev_percentage,
  COUNT(ov.order_id) OVER (PARTITION BY c.customer_state) num_review_by_state
FROM target.customers c
LEFT JOIN target.orders o
ON c.customer_id=o.customer_id
LEFT JOIN target.order_reviews ov
ON o.order_id=ov.order_id
WHERE review_score IS NOT NULL
ORDER BY c.customer_state,ov.review_score
```

Output



Row	customer_state	num_reviews	review_score	rev_percentage	num_review_by_state
1	AC	81	1	8.64%	81
2	AC	81	2	7.41%	81
3	AC	81	3	9.88%	81
4	AC	81	4	18.52%	81
5	AC	81	5	55.56%	81
6	AL	414	1	17.63%	414
7	AL	414	2	6.28%	414
8	AL	414	3	7%	414
9	AL	414	4	21.5%	414
10	AL	414	5	47.58%	414
11	AM	147	1	8.16%	147

Visualization

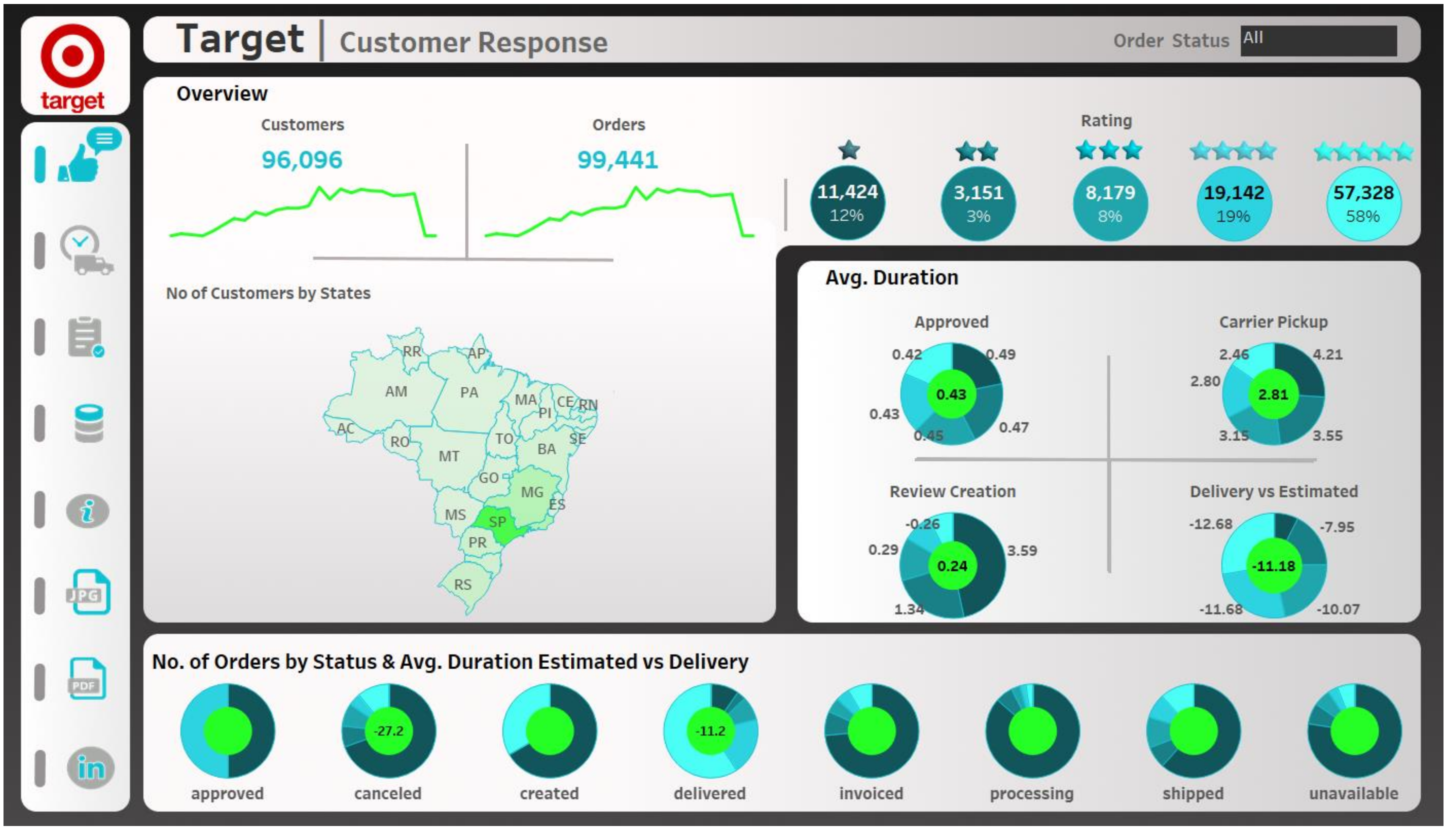


Insights

For the state **SP**, 60% of the reviews are **5 stars**, while 25% are **4 stars**. This indicates that most customers in SP are satisfied, with high ratings.

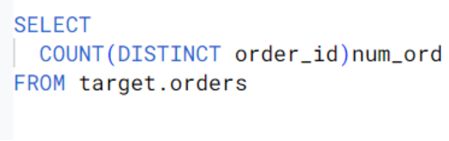
The lower review percentages for **1-star** and **2-star** reviews suggest that negative feedback is minimal in this state.

Dashboard



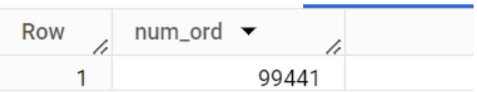
3.1 number of distinct orders

Query



```
SELECT
  COUNT(DISTINCT order_id)num_ord
FROM target.orders
```

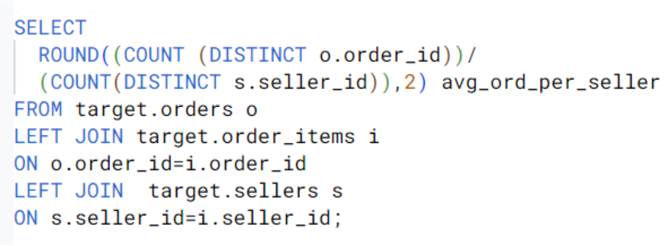
Output



Row	num_ord
1	99441

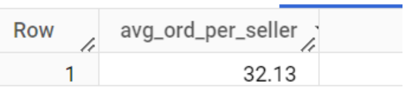
3.2 What is the average number of orders per seller

Query



```
SELECT
  ROUND((COUNT (DISTINCT o.order_id))/
  (COUNT(DISTINCT s.seller_id)),2) avg_ord_per_seller
FROM target.orders o
LEFT JOIN target.order_items i
ON o.order_id=i.order_id
LEFT JOIN  target.sellers s
ON s.seller_id=i.seller_id;
```

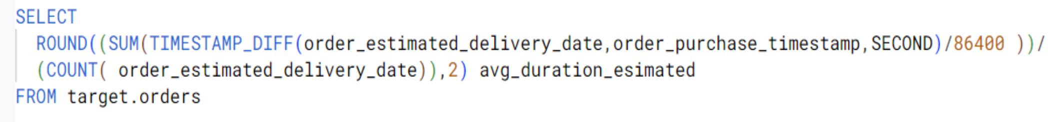
Output



Row	avg_ord_per_seller
1	32.13

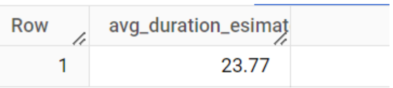
3.3 What is the average duration between order purchase and estimated delivery date across all orders

Query



```
SELECT
  ROUND((SUM(TIMESTAMP_DIFF(order_estimated_delivery_date,order_purchase_timestamp,SECOND)/86400 ))/
  (COUNT( order_estimated_delivery_date)),2) avg_duration_esimated
FROM target.orders
```

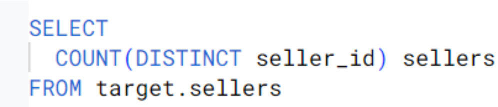
Output



Row	avg_duration_esimat
1	23.77

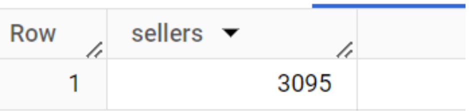
3.4 How many distinct sellers are there in the database

Query



```
SELECT
  COUNT(DISTINCT seller_id) sellers
FROM target.sellers
```

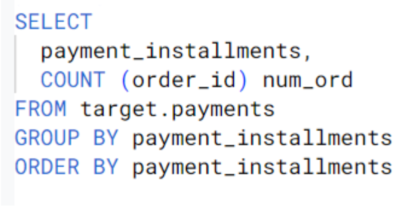
Output



Row	sellers ▾
1	3095

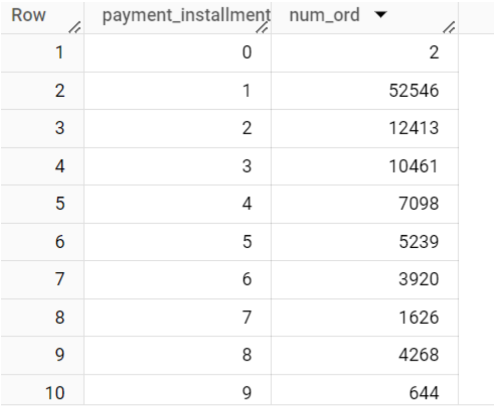
3.5 How do the number of order payments vary across different payment installment plans?

Query



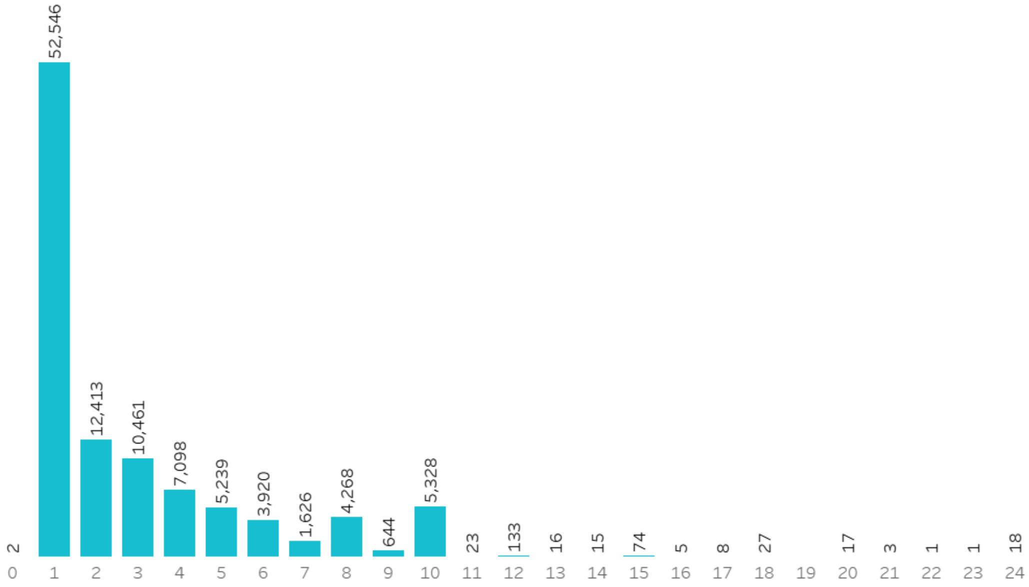
```
SELECT
  payment_installments,
  COUNT (order_id) num_ord
FROM target.payments
GROUP BY payment_installments
ORDER BY payment_installments
```

Output



Row	payment_installment	num_ord ▼
1	0	2
2	1	52546
3	2	12413
4	3	10461
5	4	7098
6	5	5239
7	6	3920
8	7	1626
9	8	4268
10	9	644

Visualization



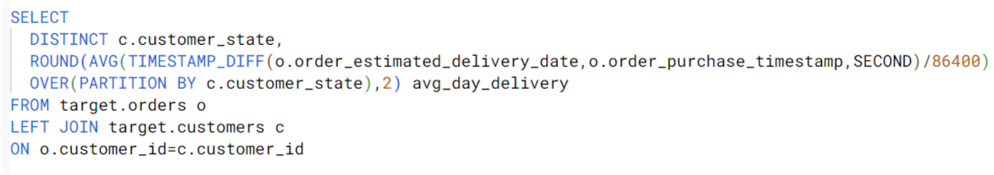
Insights

The highest number of orders (52,546) is associated with a single payment installment, indicating that most customers prefer immediate payment without installment plans. This may suggest a strong preference for cash flow management among customers.

As the number of payment installments increases, the number of orders decreases. This trend could indicate that while some customers may value flexibility in payment options, the majority may not opt for installment payments, possibly due to preferences for budgeting or simplicity in transactions

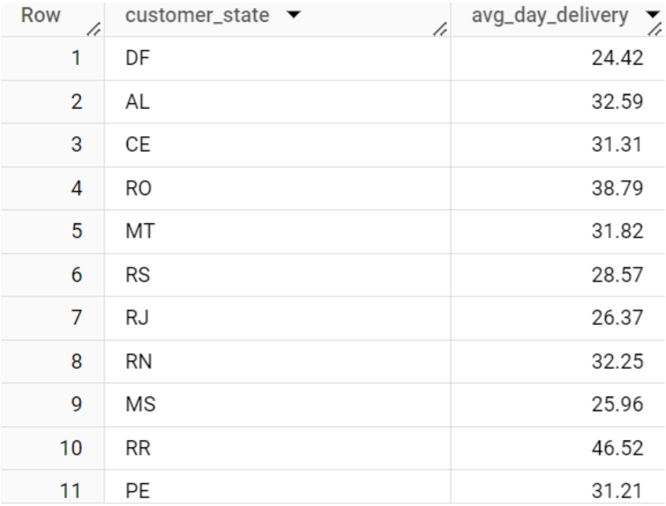
3.6 What is the average delivery time for orders across different customer states?

Query



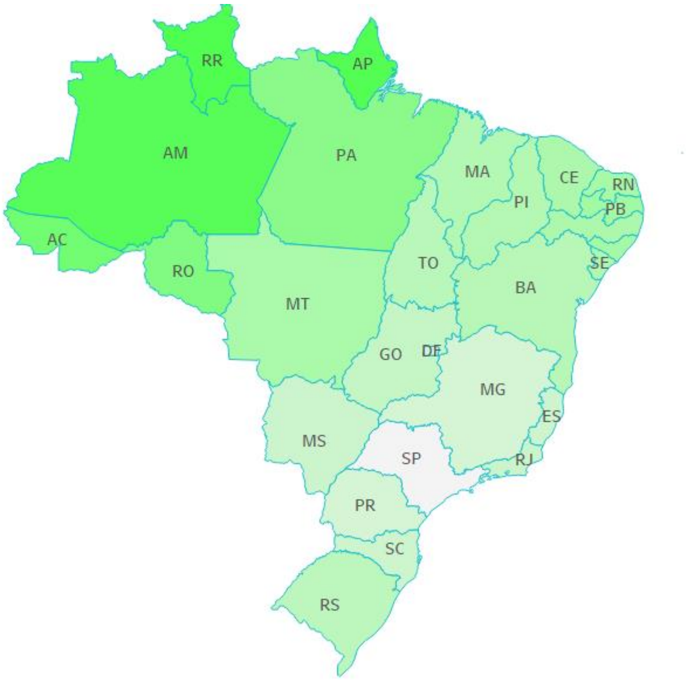
```
SELECT
  DISTINCT c.customer_state,
  ROUND(AVG(TIMESTAMP_DIFF(o.order_estimated_delivery_date,o.order_purchase_timestamp,SECOND)/86400)
  OVER(PARTITION BY c.customer_state),2) avg_day_delivery
FROM target.orders o
LEFT JOIN target.customers c
ON o.customer_id=c.customer_id
```

Output



Row	customer_state	avg_day_delivery
1	DF	24.42
2	AL	32.59
3	CE	31.31
4	RO	38.79
5	MT	31.82
6	RS	28.57
7	RJ	26.37
8	RN	32.25
9	MS	25.96
10	RR	46.52
11	PE	31.21

Visualization



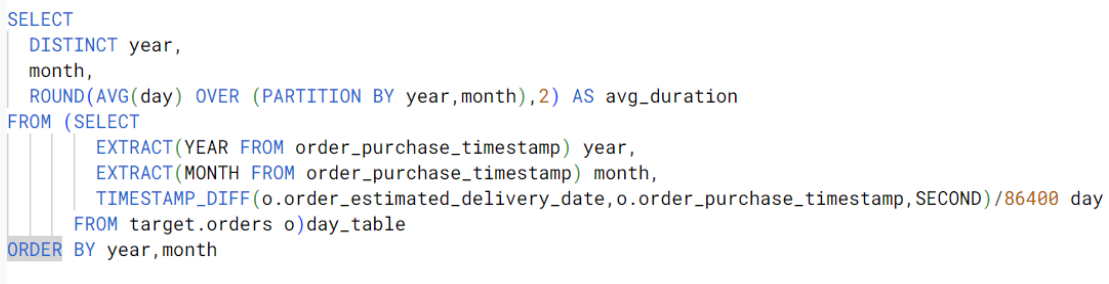
Insights

The query provides insights into how delivery times vary by customer state, highlighting geographical differences in logistics and service efficiency.

By calculating the average delivery time (in days) for each state, businesses can identify which states may require improvements in delivery processes. States with significantly higher average delivery times might indicate logistical challenges or inefficiencies.

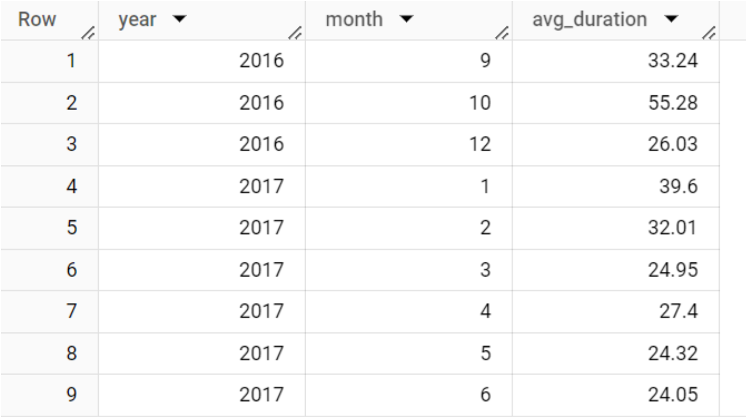
3.7 What is the average duration from order purchase to estimated delivery, broken down by year and month?

Query



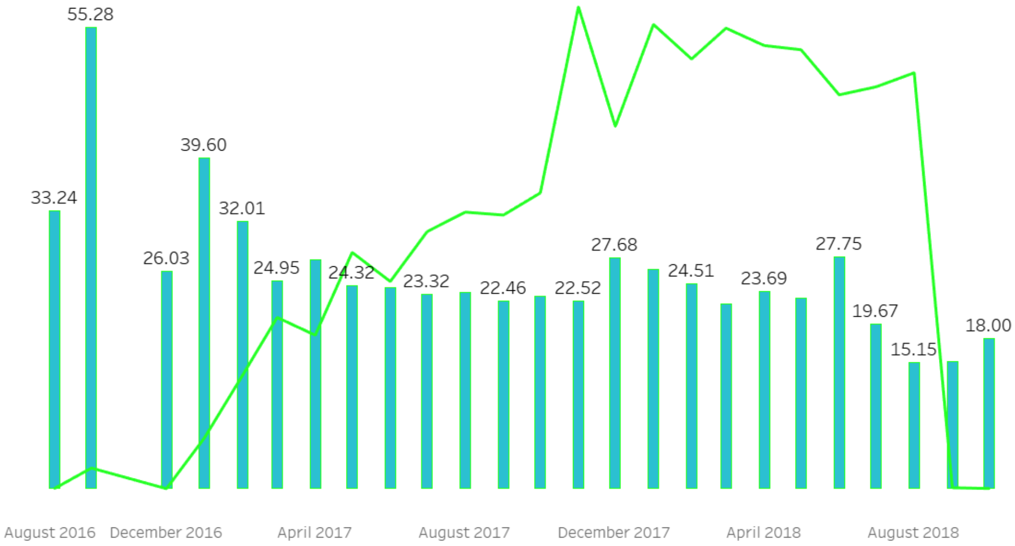
```
SELECT
  DISTINCT year,
  month,
  ROUND(AVG(day) OVER (PARTITION BY year,month),2) AS avg_duration
FROM (SELECT
        EXTRACT(YEAR FROM order_purchase_timestamp) year,
        EXTRACT(MONTH FROM order_purchase_timestamp) month,
        TIMESTAMP_DIFF(o.order_estimated_delivery_date,o.order_purchase_timestamp,SECOND)/86400 day
      FROM target.orders o)day_table
ORDER BY year,month
```

Output



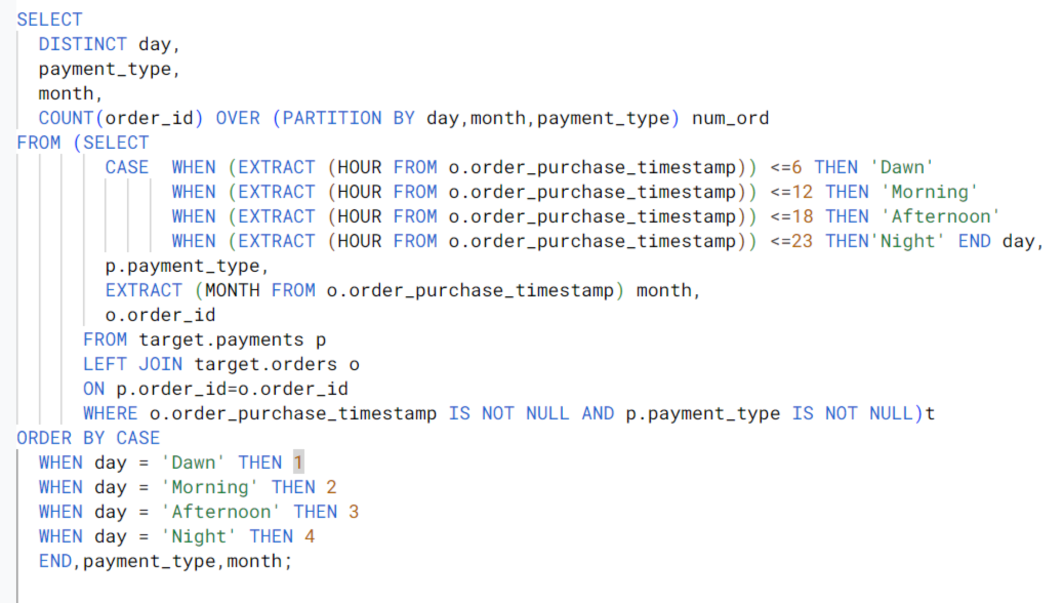
Row	year	month	avg_duration
1	2016	9	33.24
2	2016	10	55.28
3	2016	12	26.03
4	2017	1	39.6
5	2017	2	32.01
6	2017	3	24.95
7	2017	4	27.4
8	2017	5	24.32
9	2017	6	24.05

Visualization



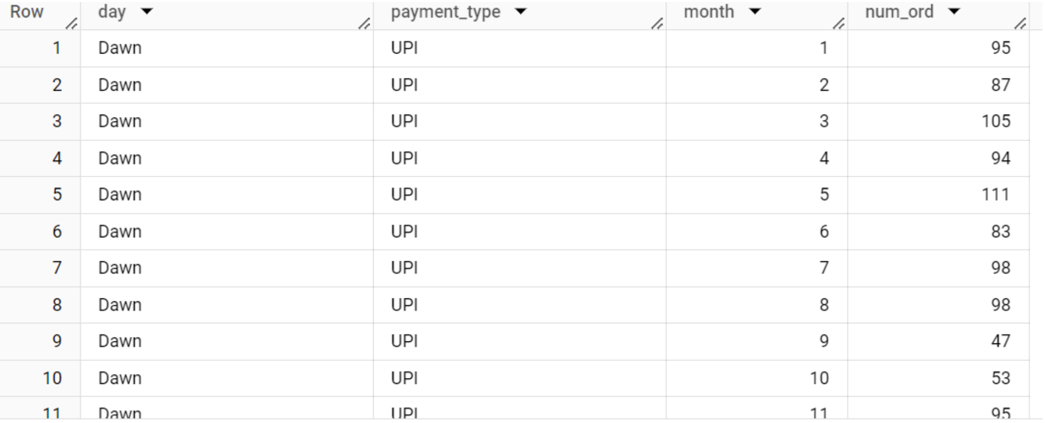
3.8 How does the number of orders vary by the time of day and payment type for each month?

Query



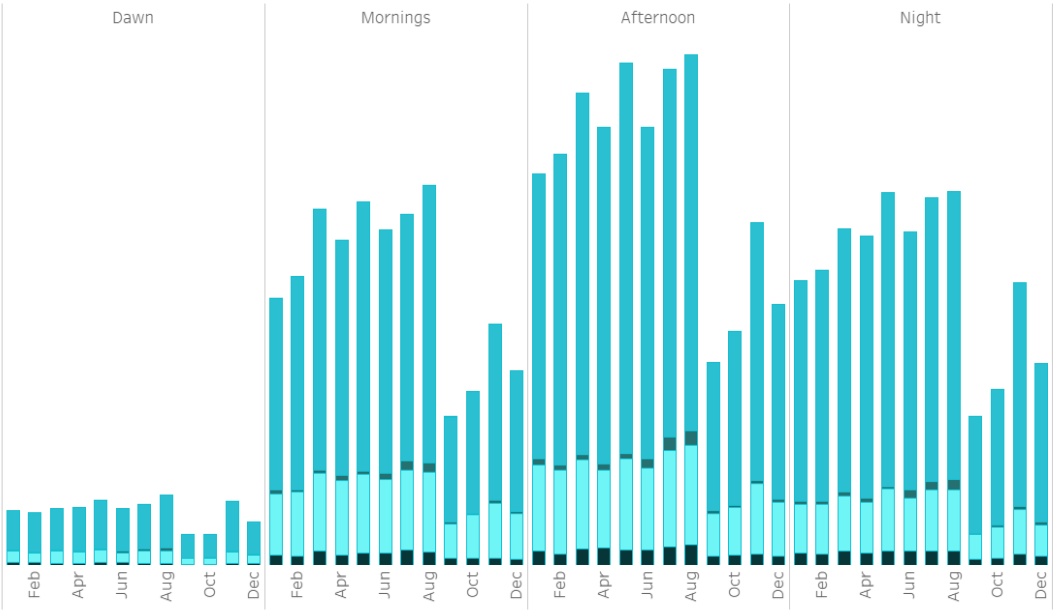
```
SELECT
  DISTINCT day,
  payment_type,
  month,
  COUNT(order_id) OVER (PARTITION BY day,month,payment_type) num_ord
FROM (SELECT
        CASE  WHEN (EXTRACT (HOUR FROM o.order_purchase_timestamp)) <=6 THEN 'Dawn'
              WHEN (EXTRACT (HOUR FROM o.order_purchase_timestamp)) <=12 THEN 'Morning'
              WHEN (EXTRACT (HOUR FROM o.order_purchase_timestamp)) <=18 THEN 'Afternoon'
              WHEN (EXTRACT (HOUR FROM o.order_purchase_timestamp)) <=23 THEN'Night' END day,
        p.payment_type,
        EXTRACT (MONTH FROM o.order_purchase_timestamp) month,
        o.order_id
      FROM target.payments p
      LEFT JOIN target.orders o
      ON p.order_id=o.order_id
      WHERE o.order_purchase_timestamp IS NOT NULL AND p.payment_type IS NOT NULL)t
ORDER BY CASE
  WHEN day = 'Dawn' THEN 1
  WHEN day = 'Morning' THEN 2
  WHEN day = 'Afternoon' THEN 3
  WHEN day = 'Night' THEN 4
  END,payment_type,month;
```

Output



Row	day	payment_type	month	num_ord
1	Dawn	UPI	1	95
2	Dawn	UPI	2	87
3	Dawn	UPI	3	105
4	Dawn	UPI	4	94
5	Dawn	UPI	5	111
6	Dawn	UPI	6	83
7	Dawn	UPI	7	98
8	Dawn	UPI	8	98
9	Dawn	UPI	9	47
10	Dawn	UPI	10	53
11	Dawn	UPI	11	95

Visualization



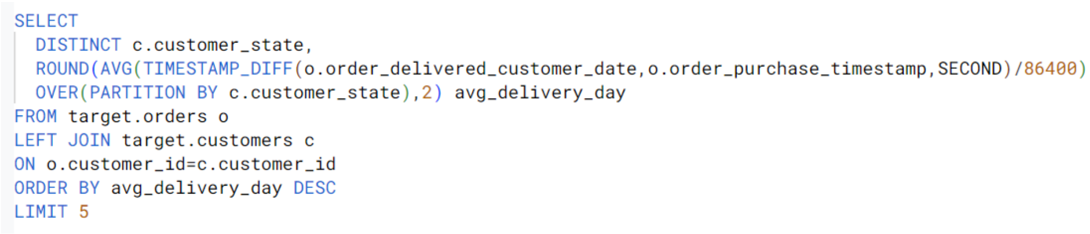
Insights

The query segments order placements into four-time categories: Dawn, Morning, Afternoon, and Night. This categorization allows for a detailed analysis of customer behavior based on the time they prefer to make purchases.

By including payment type, the query helps to understand which payment methods are favored during different times of the day. For example, customers may prefer specific payment options in the morning versus the evening, indicating potential trends in consumer preferences.

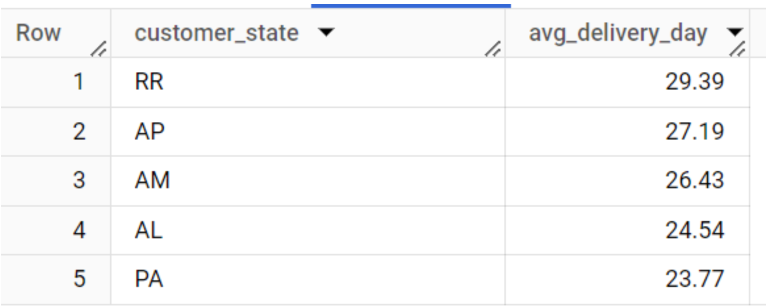
3.9 Which customer states experience the longest average delivery times?

Query



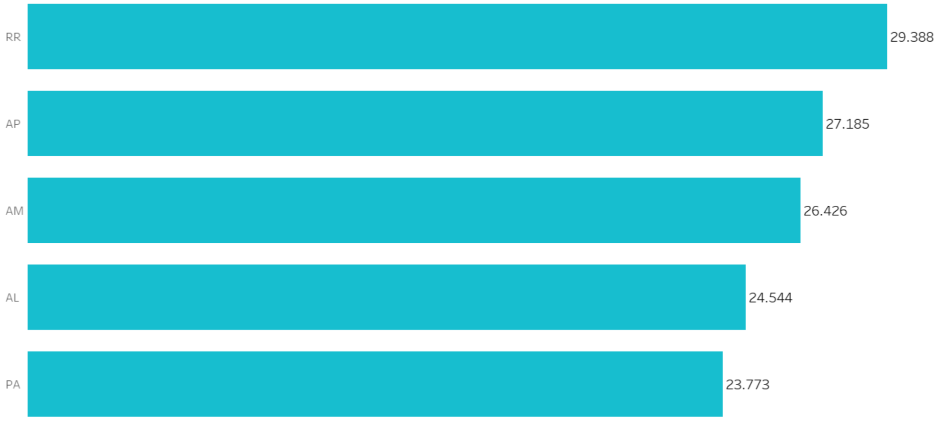
```
SELECT
  DISTINCT c.customer_state,
  ROUND(AVG(TIMESTAMP_DIFF(o.order_delivered_customer_date,o.order_purchase_timestamp,SECOND)/86400)
  OVER(PARTITION BY c.customer_state),2) avg_delivery_day
FROM target.orders o
LEFT JOIN target.customers c
ON o.customer_id=c.customer_id
ORDER BY avg_delivery_day DESC
LIMIT 5
```

Output



Row	customer_state	avg_delivery_day
1	RR	29.39
2	AP	27.19
3	AM	26.43
4	AL	24.54
5	PA	23.77

Visualization



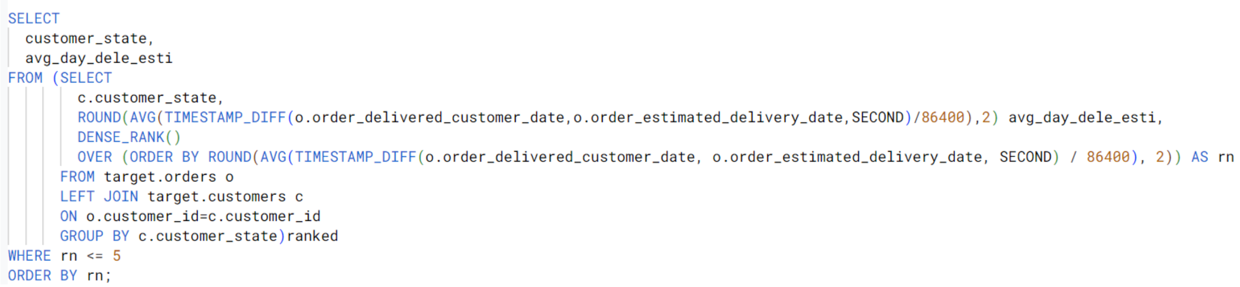
Insights

States like **RR, AP, and AM** have significantly longer delivery times. This could indicate logistical challenges such as geographic remoteness, poor infrastructure, or limited access to faster delivery methods.

Businesses may want to investigate the reasons behind these long delivery times in these specific states. Possible solutions could include optimizing delivery routes, collaborating with more efficient local carriers, or offering expedited shipping options.

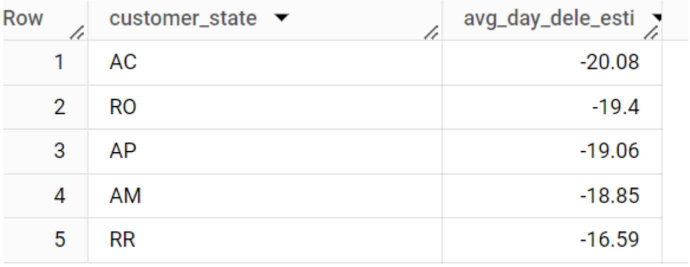
3.10 Which customer states have the largest negative deviation from their estimated delivery dates?

Query



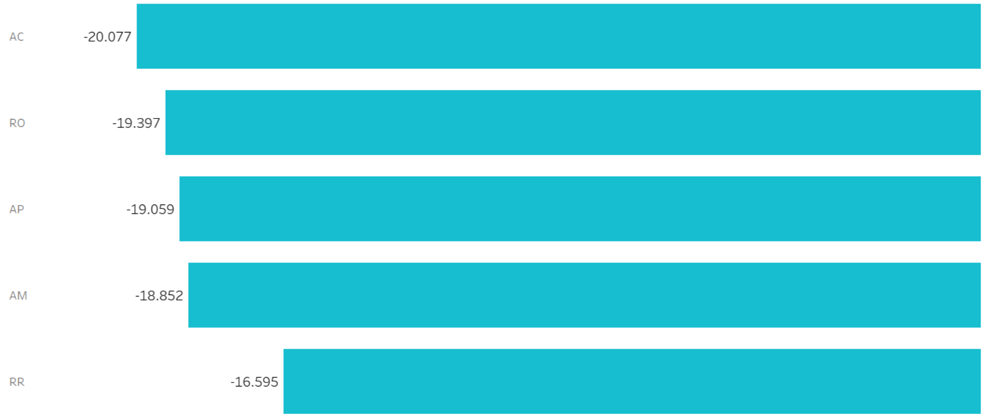
```
SELECT
  customer_state,
  avg_day_dele_esti
FROM (SELECT
        c.customer_state,
        ROUND(AVG(TIMESTAMP_DIFF(o.order_delivered_customer_date,o.order_estimated_delivery_date,SECOND)/86400),2) avg_day_dele_esti,
        DENSE_RANK()
        OVER (ORDER BY ROUND(AVG(TIMESTAMP_DIFF(o.order_delivered_customer_date, o.order_estimated_delivery_date, SECOND) / 86400), 2)) AS rn
      FROM target.orders o
      LEFT JOIN target.customers c
      ON o.customer_id=c.customer_id
      GROUP BY c.customer_state)ranked
WHERE rn <= 5
ORDER BY rn;
```

Output



Row	customer_state	avg_day_dele_esti
1	AC	-20.08
2	RO	-19.4
3	AP	-19.06
4	AM	-18.85
5	RR	-16.59

Visualization



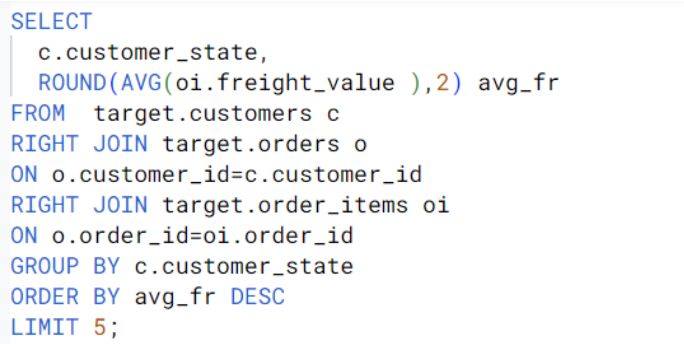
Insights

The large negative deviations suggest that delivery times in these regions are consistently overestimated. Businesses may need to reassess their estimated delivery algorithms or methods to provide more accurate predictions.

Early deliveries can lead to enhanced customer satisfaction as orders arrive well before the expected time, creating a positive impression. However, consistently large deviations could also lead to confusion if customers are not prepared for the early arrival of goods.

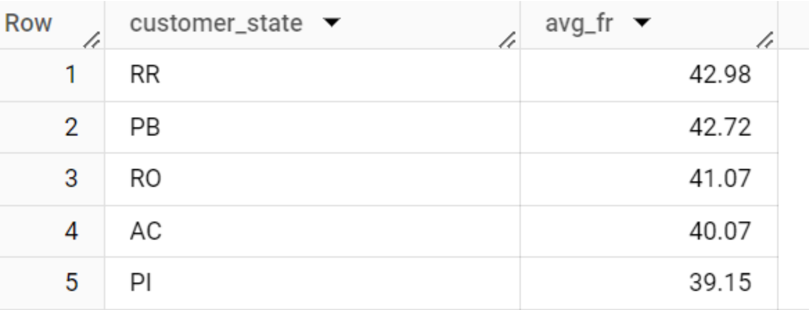
3.11 Which customer states incur the highest average freight costs?

Query



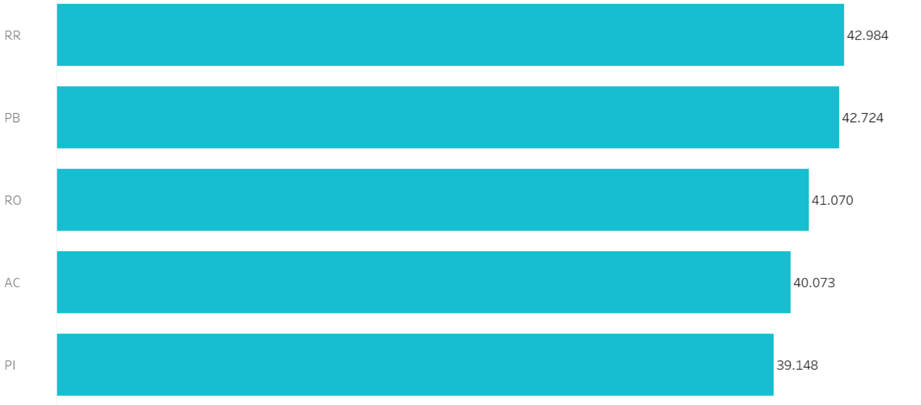
```
SELECT
  c.customer_state,
  ROUND(AVG(oi.freight_value ),2) avg_fr
FROM  target.customers c
RIGHT JOIN target.orders o
ON o.customer_id=c.customer_id
RIGHT JOIN target.order_items oi
ON o.order_id=oi.order_id
GROUP BY c.customer_state
ORDER BY avg_fr DESC
LIMIT 5;
```

Output



Row	customer_state	avg_fr
1	RR	42.98
2	PB	42.72
3	RO	41.07
4	AC	40.07
5	PI	39.15

Visualization



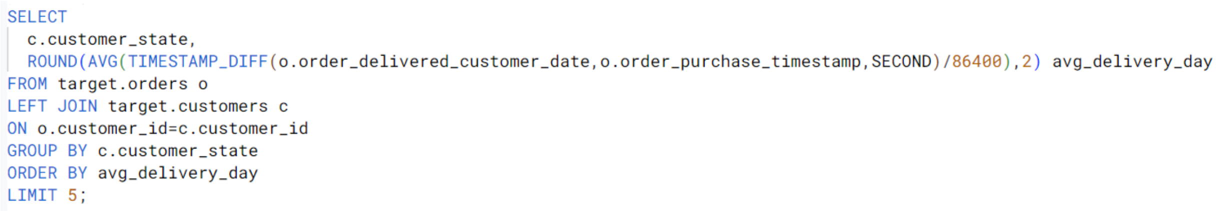
Insights

These states likely face higher freight charges due to their **geographical location**. They may be farther from major distribution hubs or have limited infrastructure, increasing transportation costs.

Customers in these regions could experience higher total purchase costs, which might affect purchasing behavior. Businesses might want to consider strategies to mitigate these costs, such as offering **discounts on freight charges** to maintain competitiveness in these areas.

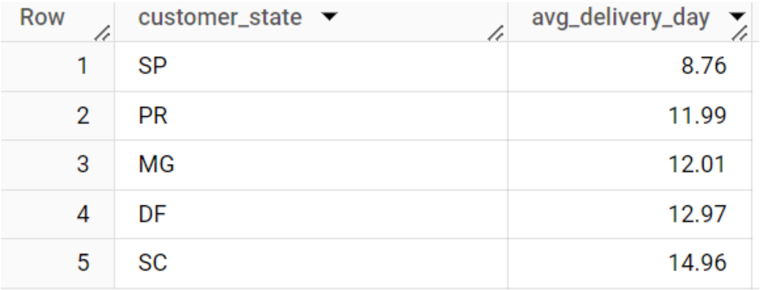
3.12 Which customer states have the shortest average delivery times?

Query



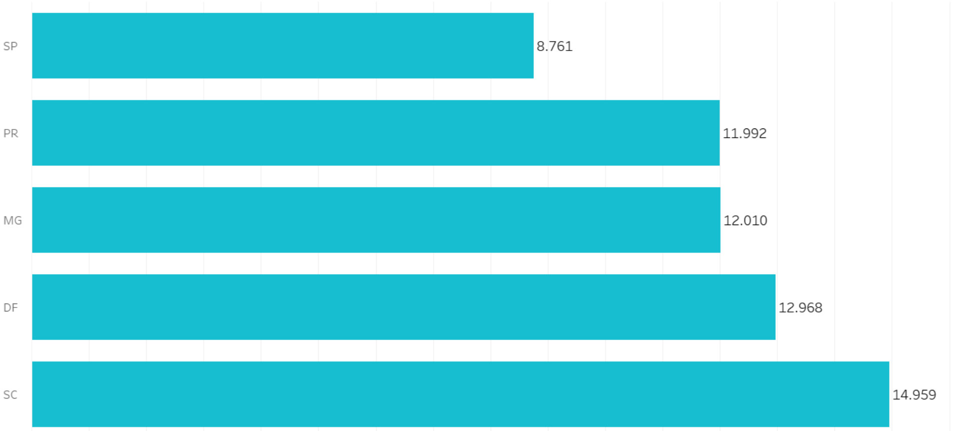
```
SELECT
  c.customer_state,
  ROUND(AVG(TIMESTAMP_DIFF(o.order_delivered_customer_date,o.order_purchase_timestamp,SECOND)/86400),2) avg_delivery_day
FROM target.orders o
LEFT JOIN target.customers c
ON o.customer_id=c.customer_id
GROUP BY c.customer_state
ORDER BY avg_delivery_day
LIMIT 5;
```

Output



Row	customer_state	avg_delivery_day
1	SP	8.76
2	PR	11.99
3	MG	12.01
4	DF	12.97
5	SC	14.96

Visualization



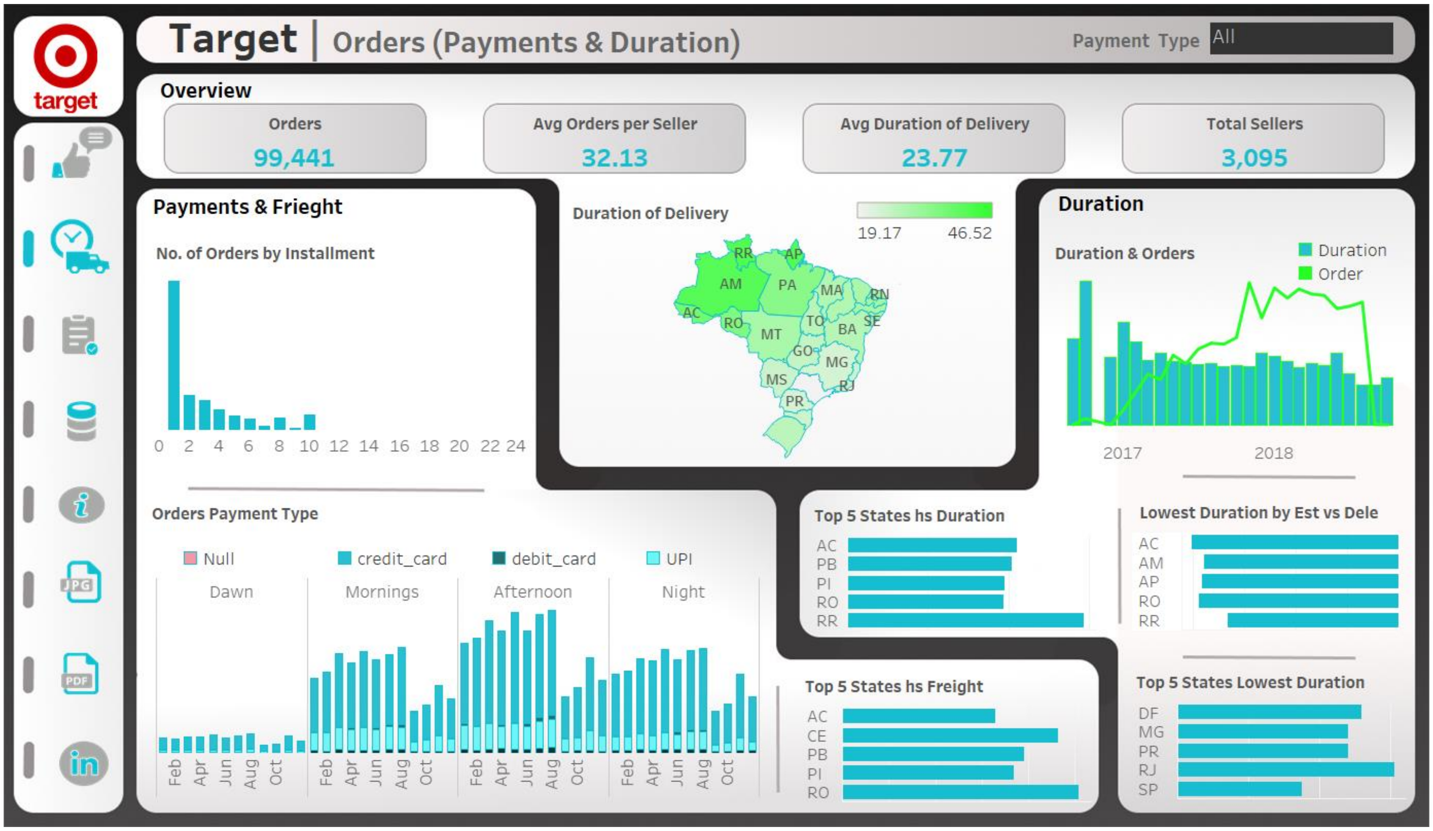
Insights

Proximity to logistics hubs: São Paulo, for example, is a major economic and logistics center, which contributes to faster shipping times.

Developed transportation infrastructure: States like PR, MG, and DF may have better road networks and efficient logistics operations.

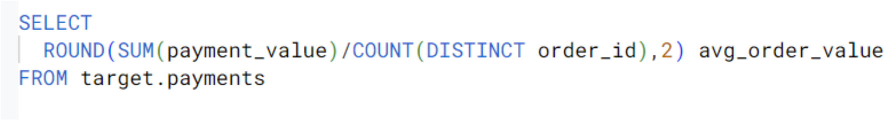
Higher concentration of distribution centers: These states may have more warehouses and fulfillment centers nearby, leading to quicker order processing and delivery.

Dashboard



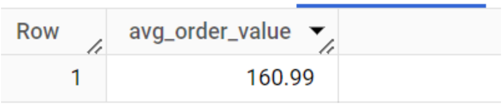
4.1 What is the average order value across all transaction

Query



```
SELECT
  ROUND(SUM(payment_value)/COUNT(DISTINCT order_id),2) avg_order_value
FROM target.payments
```

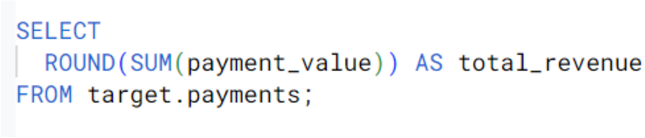
Output



Row	avg_order_value
1	160.99

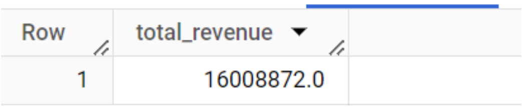
4.2 What is the total revenue generated from all transactions?

Query



```
SELECT
  ROUND(SUM(payment_value)) AS total_revenue
FROM target.payments;
```

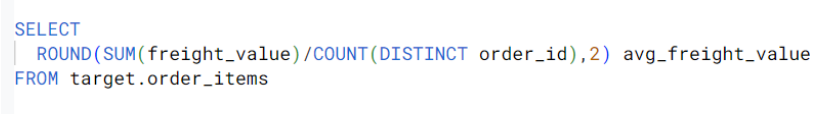
Output



Row	total_revenue
1	16008872.0

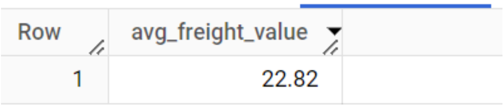
4.3 What is the average freight (shipping) cost per order?

Query



```
SELECT
  ROUND(SUM(freight_value)/COUNT(DISTINCT order_id),2) avg_freight_value
FROM target.order_items
```

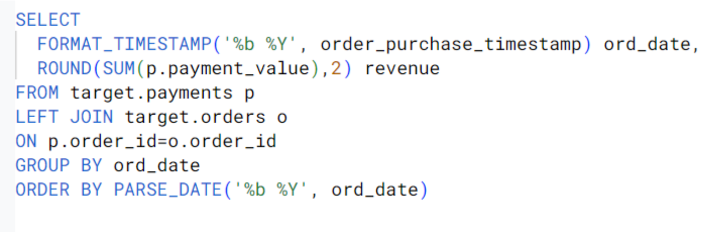
Output



Row	avg_freight_value
1	22.82

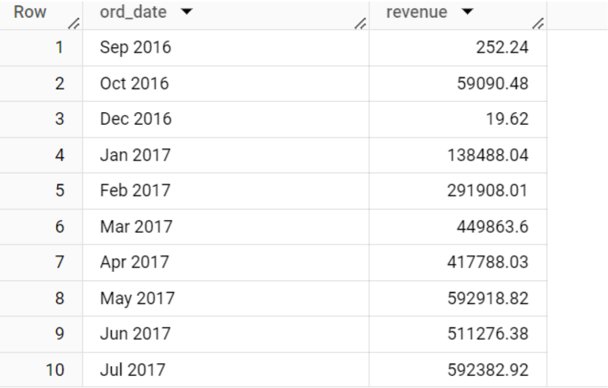
4.4 What is the monthly revenue trend?

Query



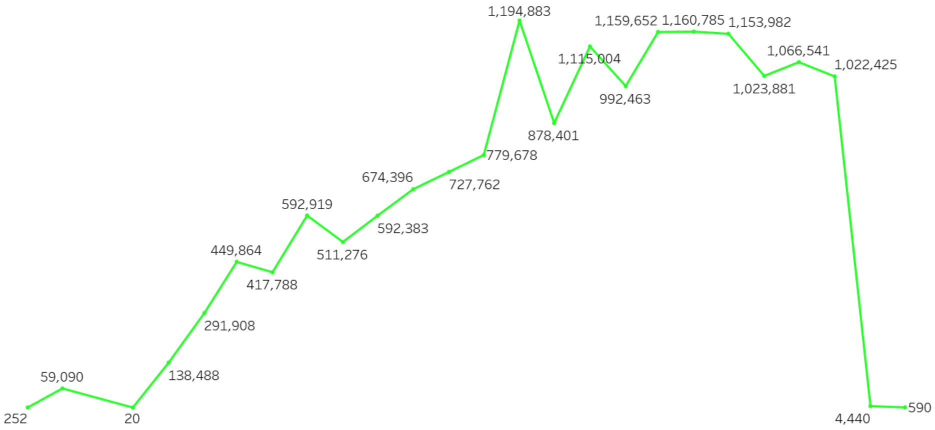
```
SELECT
  FORMAT_TIMESTAMP('%b %Y', order_purchase_timestamp) ord_date,
  ROUND(SUM(p.payment_value),2) revenue
FROM target.payments p
LEFT JOIN target.orders o
ON p.order_id=o.order_id
GROUP BY ord_date
ORDER BY PARSE_DATE('%b %Y', ord_date)
```

Output

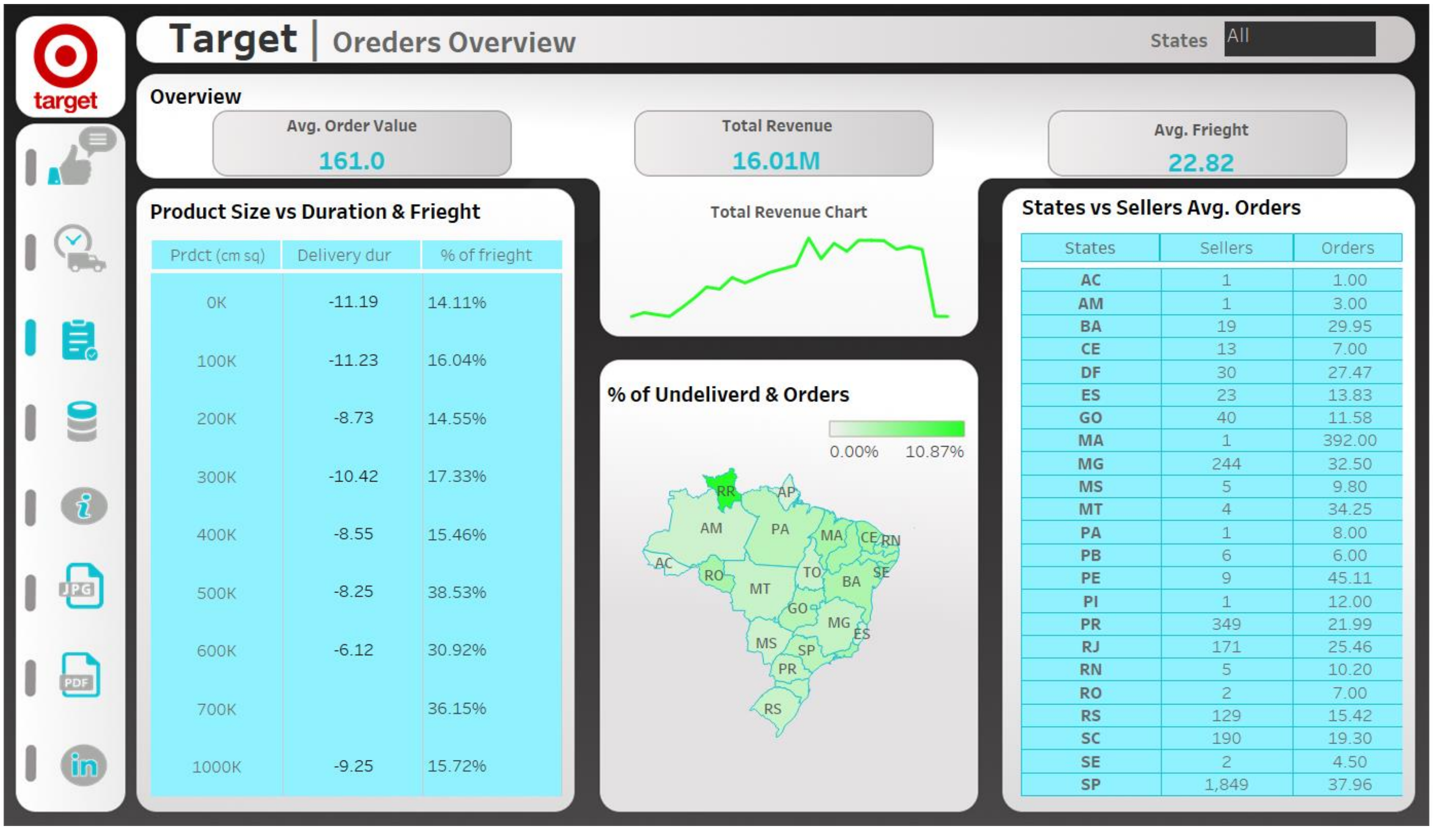


Row	ord_date	revenue
1	Sep 2016	252.24
2	Oct 2016	59090.48
3	Dec 2016	19.62
4	Jan 2017	138488.04
5	Feb 2017	291908.01
6	Mar 2017	449863.6
7	Apr 2017	417788.03
8	May 2017	592918.82
9	Jun 2017	511276.38
10	Jul 2017	592382.92

Visualization



Dashboard



Target | Oreders Overview

States: All

Overview

Avg. Order Value: 161.0

Total Revenue: 16.01M

Avg. Frieght: 22.82

Product Size vs Duration & Frieght

Prdct (cm sq)	Delivery dur	% of frieght
0K	-11.19	14.11%
100K	-11.23	16.04%
200K	-8.73	14.55%
300K	-10.42	17.33%
400K	-8.55	15.46%
500K	-8.25	38.53%
600K	-6.12	30.92%
700K		36.15%
1000K	-9.25	15.72%

% of Undeliverd & Orders

0.00% 10.87%

States vs Sellers Avg. Orders

States	Sellers	Orders
AC	1	1.00
AM	1	3.00
BA	19	29.95
CE	13	7.00
DF	30	27.47
ES	23	13.83
GO	40	11.58
MA	1	392.00
MG	244	32.50
MS	5	9.80
MT	4	34.25
PA	1	8.00
PB	6	6.00
PE	9	45.11
PI	1	12.00
PR	349	21.99
RJ	171	25.46
RN	5	10.20
RO	2	7.00
RS	129	15.42
SC	190	19.30
SE	2	4.50
SP	1,849	37.96

Project Conclusion

The project involved an in-depth analysis of sales, customer behavior, and operational performance data from the provided datasets. The key focus was on deriving actionable insights that could inform business decisions, improve customer satisfaction, and optimize logistics and sales strategies. Here's a summary of the conclusions drawn from the analysis:

1. **Sales Performance**:
    a. Monthly revenue trends were identified, showing fluctuations that could correlate with specific marketing campaigns, holidays, or external factors.
    b. The average order value was calculated, providing insight into customer purchasing behavior, which can be used to optimize pricing and upselling strategies.
2. **Customer Behavior**:
    a. Analysis of customer distribution by state revealed geographical patterns in delivery times and freight charges, highlighting regions with longer delivery durations and higher freight costs, which could be targeted for operational improvements.
    b. Net Promoter Score (NPS) provided a direct measurement of customer satisfaction, revealing areas where service improvements are needed.
3. **Operational Efficiency**:
    a. The average time taken for deliveries was analyzed across states, showing discrepancies between estimated and actual delivery times in certain regions. This insight can be used to adjust delivery logistics and set more realistic expectations for customers.

    b. Freight value analysis across states identified areas where shipping costs were higher, which may indicate inefficiencies or potential for renegotiation of shipping contracts.

4. **Payment and Revenue**:
    a. Payment installment trends revealed that most customers preferred single or short-term installment plans, providing insights for offering customized payment options.
    b. Total revenue and monthly revenue patterns were calculated, offering a comprehensive view of the business's financial performance, aiding in long-term financial planning.

Recommendations:

- **Target High Revenue Regions**: Focus on optimizing operations in states with high delivery times or freight charges to enhance customer satisfaction and reduce costs.
- **Improve Customer Service**: Use NPS and review data to address issues in lower-scoring segments, improving overall customer retention and satisfaction.
- **Optimize Payment Options**: Introduce more flexible payment plans or promotions in regions with lower installment usage to encourage higher-value purchases.
- **Forecasting and Planning**: Use the insights from monthly revenue patterns to forecast demand more accurately, plan inventory, and schedule marketing efforts during peak months.

In conclusion, the project successfully uncovered multiple opportunities for business improvements, and the insights generated will be crucial for enhancing customer experience, optimizing operations, and driving revenue growth.