**Brazilian E-Commerce Public Dataset by Olist**

*100,000 Orders with product, customer and reviews info*

**This is a Brazilian eCommerce public dataset of orders made at “Olist Store”. The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. Its features allow viewing orders from various dimensions: from order status, price, payment, and freight performance to customer location, product attributes, and finally reviews written by customers. We also released a geolocation dataset that relates Brazilian zip codes to LAT/LNG coordinates.**

<https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce>

**BUSINESS PROBLEMS:**

1. **Identifying factors that lead to increased customer loyalty:** By analyzing customer location, past order history, and other relevant factors, we may be able to identify patterns and trends that can help us to develop targeted marketing campaigns that are more effective at increasing customer loyalty.
2. **Optimizing shipping and logistics:** By analyzing data on delivery times, customer locations, and other relevant factors, we may be able to identify opportunities to optimize our shipping and logistics processes and reduce delivery times, which could improve the overall customer experience.
3. **Improving order fulfillment and payment processing:** By analyzing data on common issues with order fulfillment and payment processing, we may be able to identify problems and implement solutions that will help us to increase the success rate of orders.
4. **Identifying high-satisfaction products:** By analyzing customer reviews and ratings, we may be able to identify which products or product categories have the highest customer satisfaction ratings. We can then use this information to focus our marketing efforts and drive sales.
5. **Identifying areas for improvement in product quality:** By analyzing customer reviews, we may be able to identify common issues or areas for improvement in the quality of our products. We can then use this information to make changes that will improve customer satisfaction and drive sales.
6. **Identifying customer segments with the highest potential for growth:** By analyzing customer

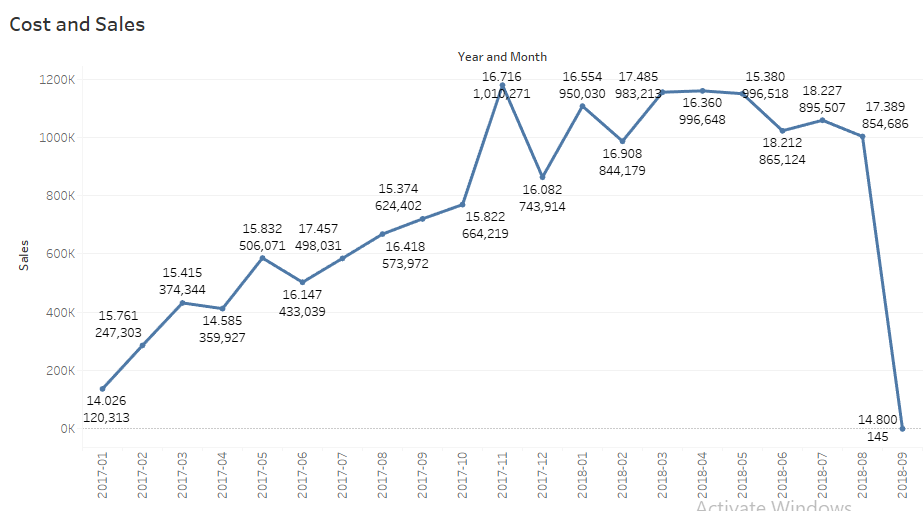
location, past order history, and other relevant factors, we may be able to identify specific customer segments that have the highest potential for growth. We can then use this information to develop targeted marketing campaigns and drive sales.

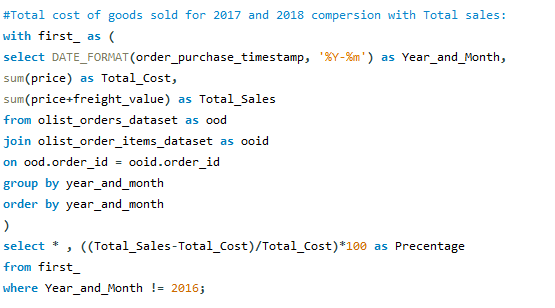
1. **Improving the efficiency of our marketing efforts:** By analyzing data on customer location, past order history, and other relevant factors, we may be able to identify patterns and trends that can help us to more effectively target our marketing efforts and improve the efficiency of our campaigns.
2. **Identifying opportunities to expand into new markets:** By analyzing data on customer location and product demand, we may be able to identify opportunities to expand into new markets and drive sales growth.

**BUSINESS QUESTIONS:**

1. **Total cost of goods sold for 2017 and 2018.**

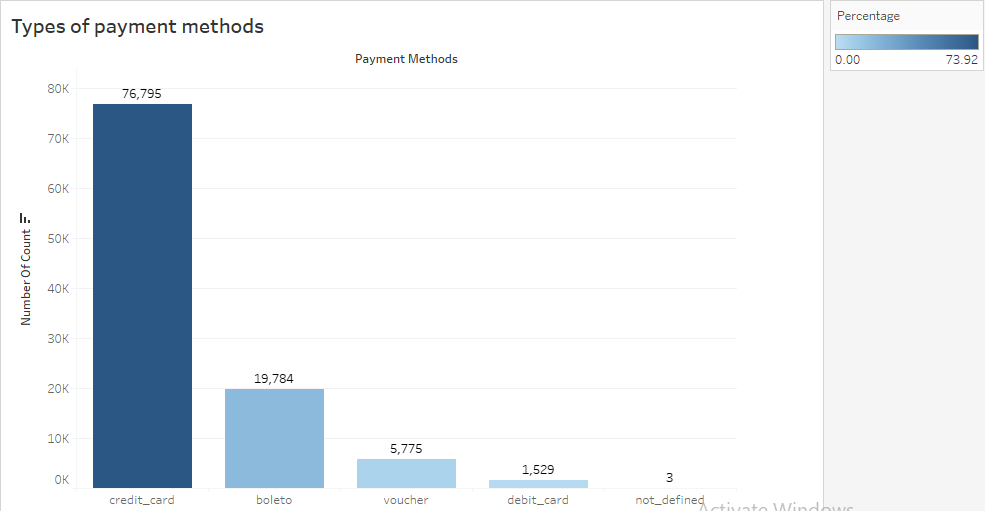
In June 2018 has 18.212% highest sales percentage where total sales is 1,022,677 and total cost is 865,124.

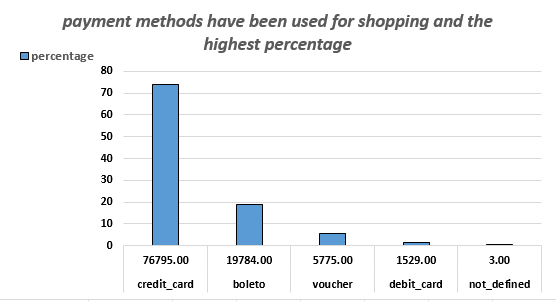
****

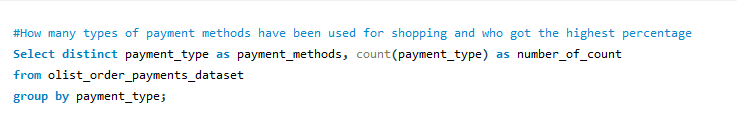


1. **How many types of payment methods have been used for shopping and who has the highest percentage?**

Four different payment methods can be found: **credit card,** **boleto**, **voucher** and **debit card**. **3** payments had no payment defined against them. Credit cards have the highest **73.92%** used for payments.



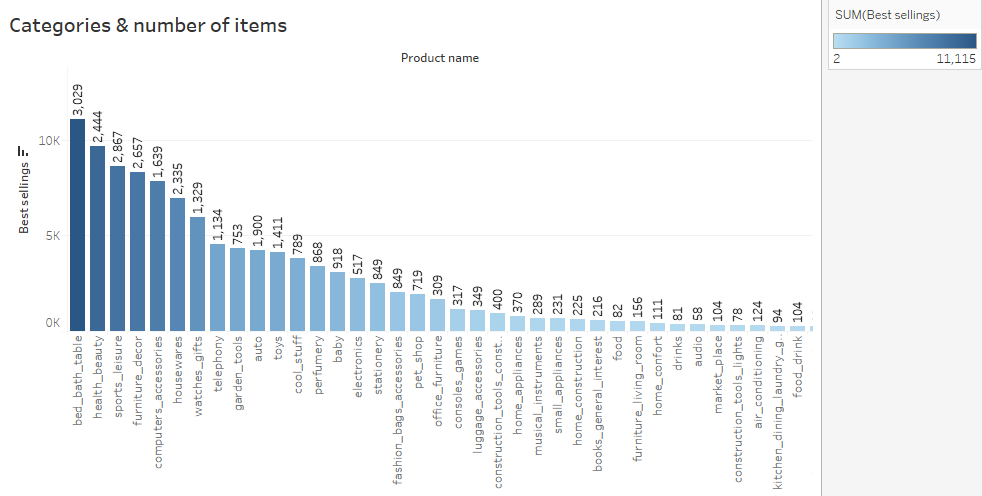
****

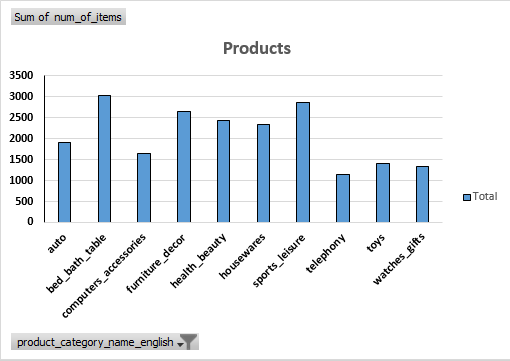
****

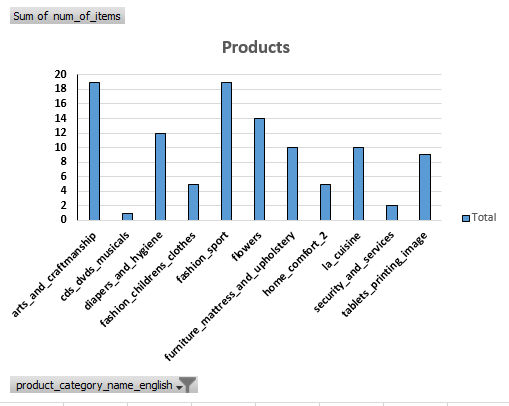
1. **Show the number of categories present in the OLIST and how many items are present in that category.**

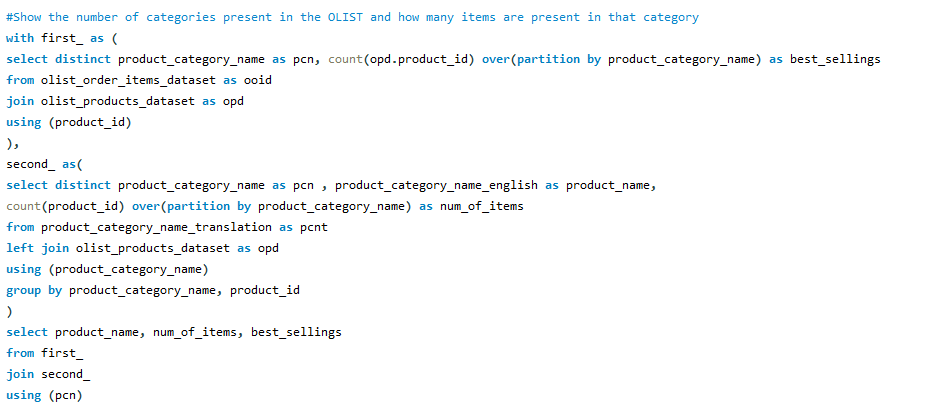
There are a total **71** categories and multiple items present in each category, also we make a correlation between number of items and number of items selling.

Bed\_bath\_table have **max** **3029** items and cds\_dvds\_musicals have only **min** **1** item.

****

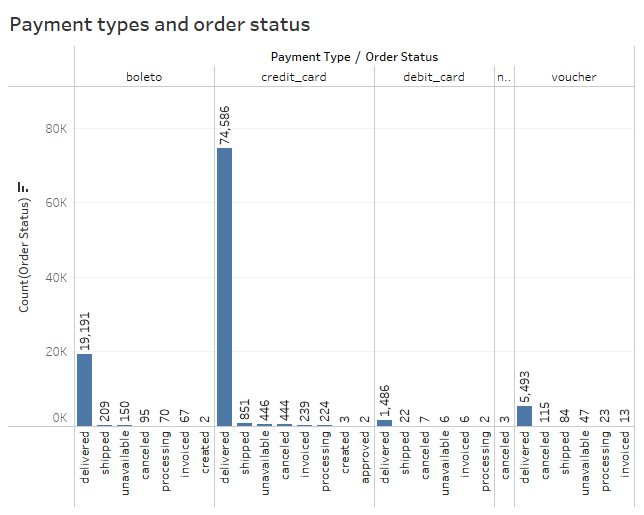


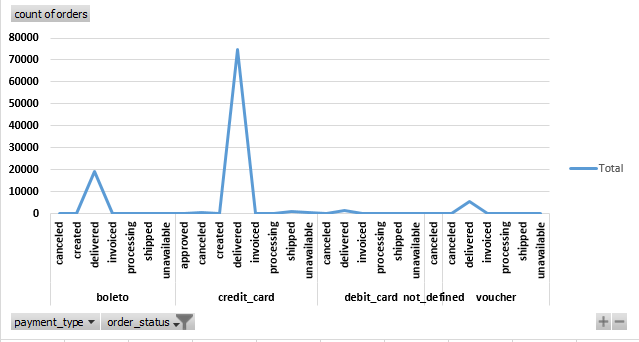


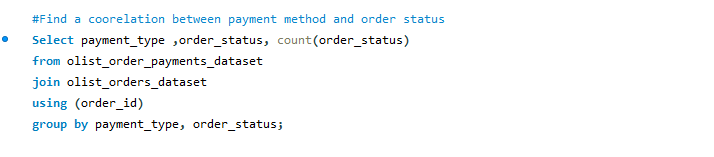
****

1. **Find a correlation between payment method and order status.**

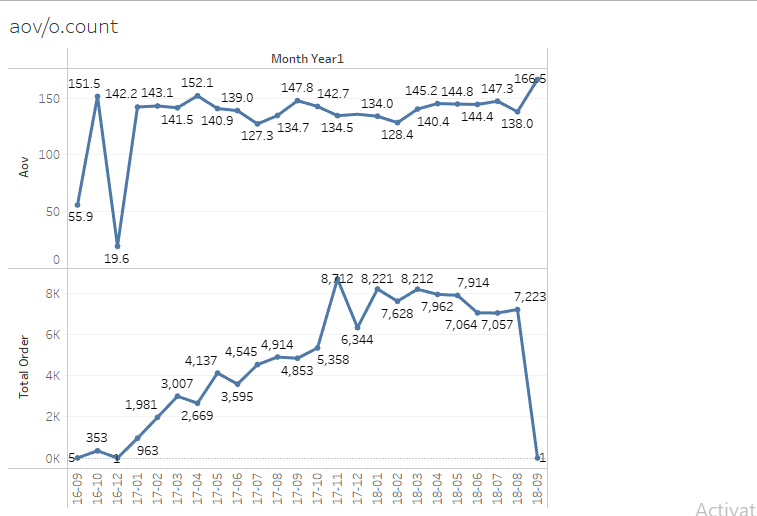
We can find four payment methods: **credit card,** **boleto**, **voucher** and **debit card** and a total eight order statuses for each payment method. Here we analyze the correlation between payment method and order status: highest number of orders is **74586**, placed by credit card payment method and the status is delivered.

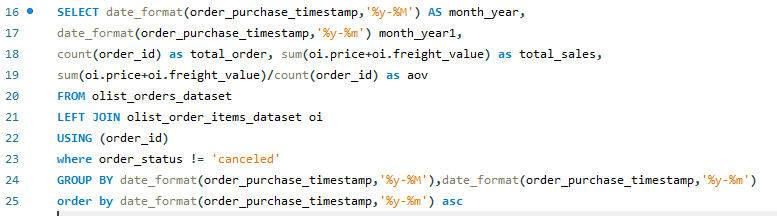




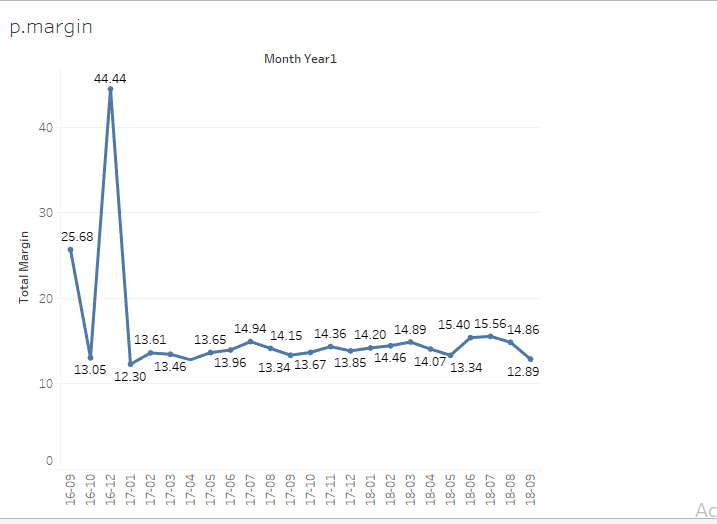
****

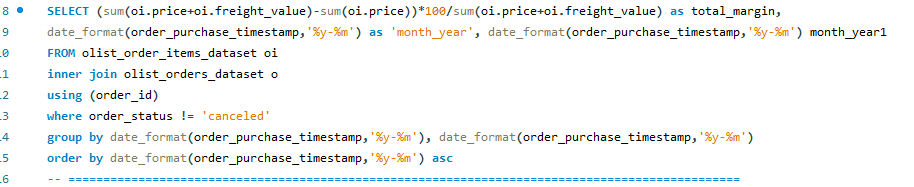
1. **Find average order value and order count, monthly and yearly:**



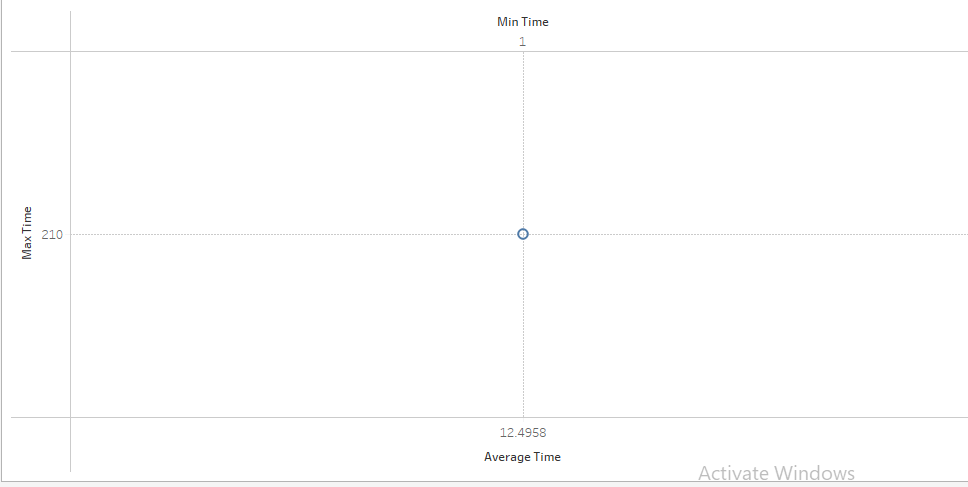


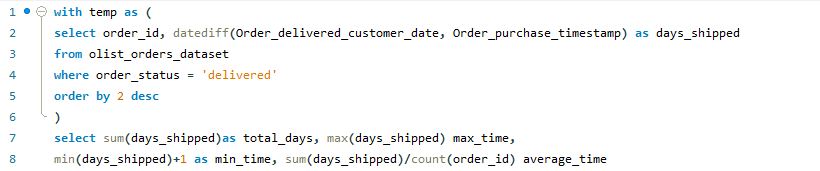
1. **Find net profit margins:**



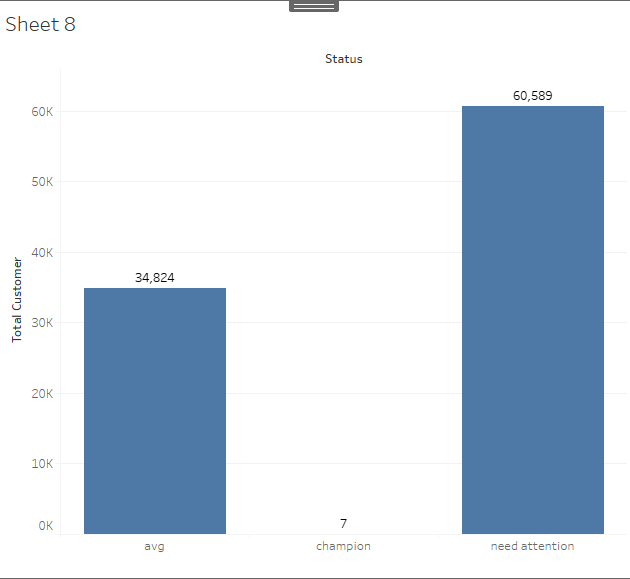
****

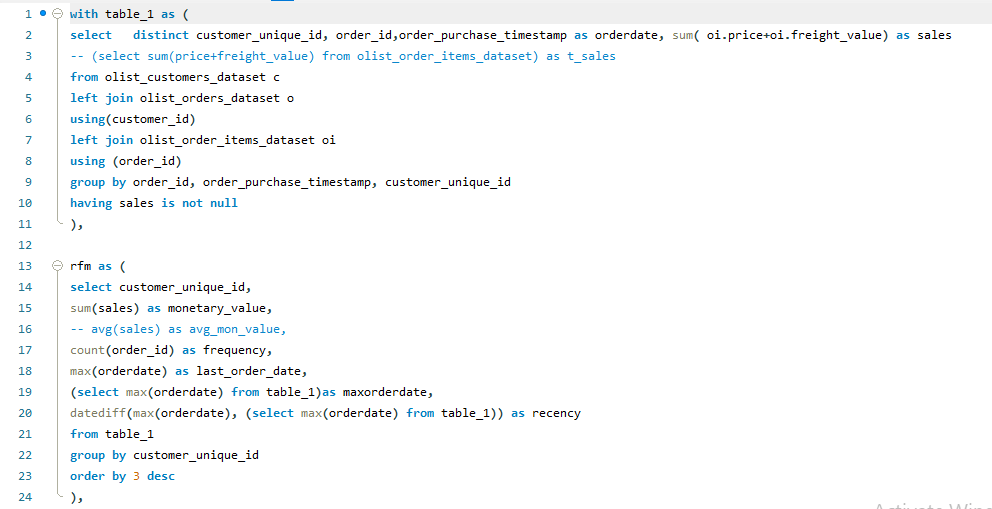
1. **Find average delivery time:**

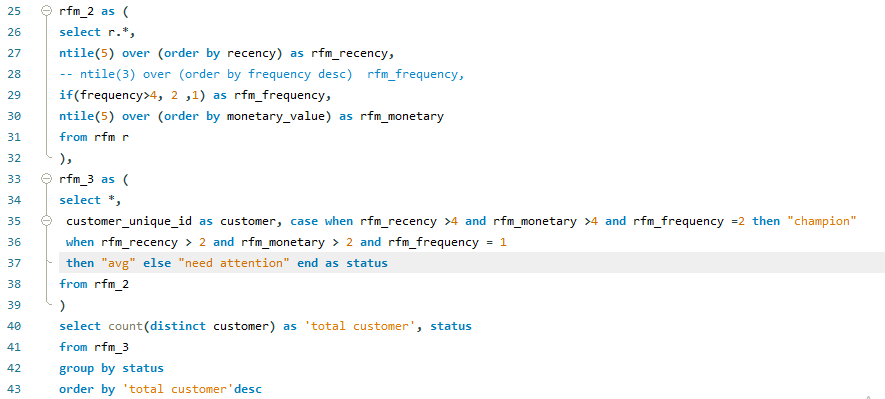


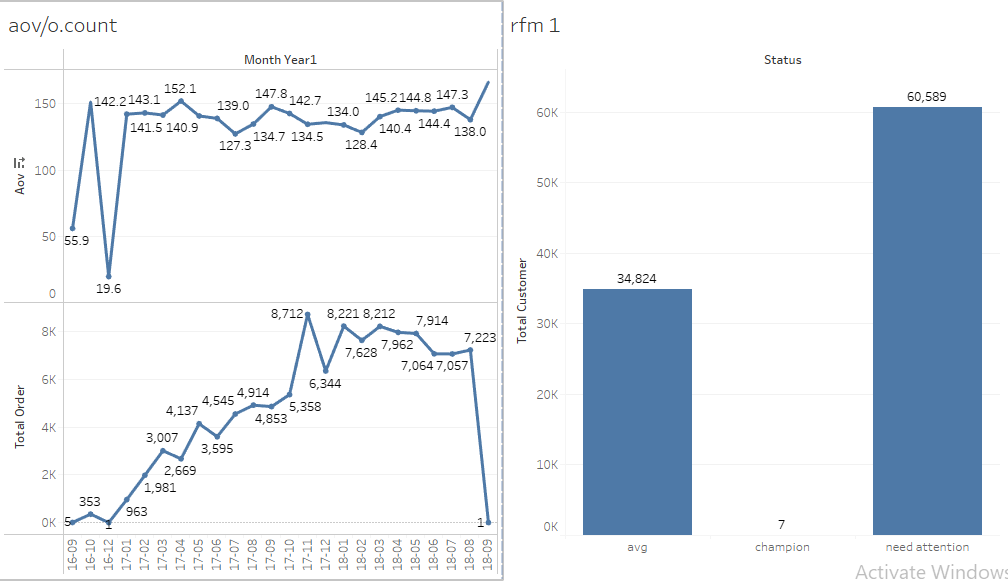
****

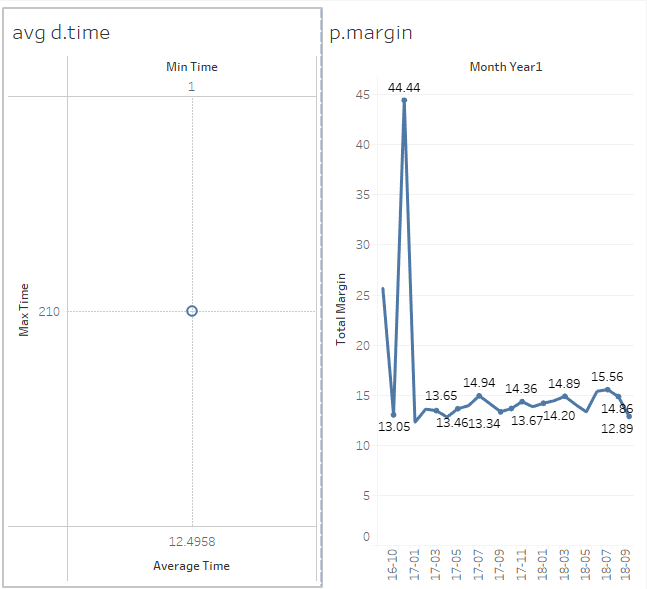
1. **Conduct RFM analysis for this dataset:**



****

****





1. **What is the best-selling category?**

Best-selling category: cama\_mesa\_banho

**select product\_id, order\_item\_id, order\_id, product\_category\_name**

**from (**

**SELECT**

**product\_id, order\_item\_id, order\_id, product\_category\_name,**

**SUM(order\_item\_id) AS order\_sales,**

**RANK() OVER (PARTITION BY product\_category\_name ORDER BY SUM(order\_item\_id) DESC) as rn**

**FROM olist\_order\_items**

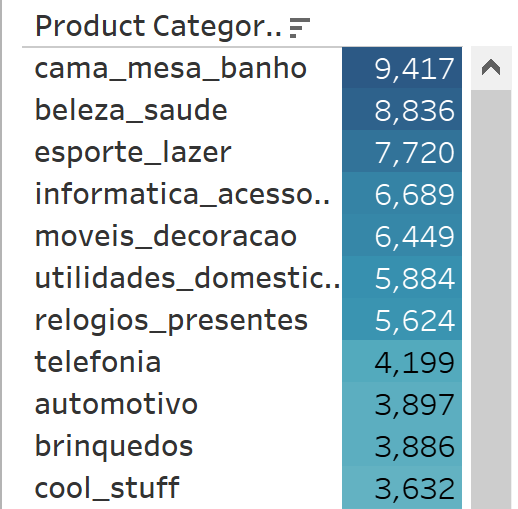
**INNER JOIN olist\_products\_dataset**

**USING (product\_id)**

**GROUP BY product\_id**

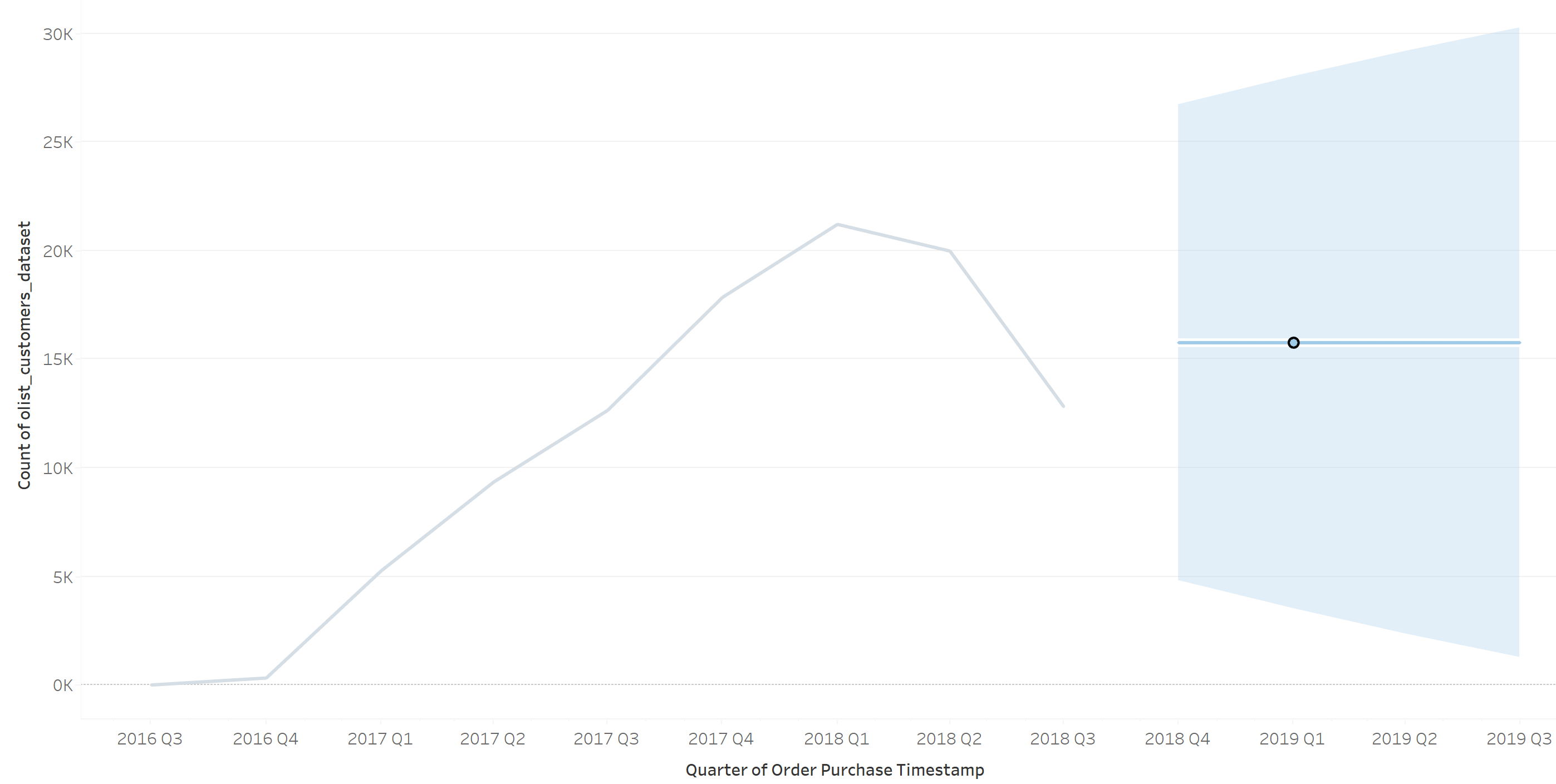
**ORDER BY product\_category\_name ASC ) as max\_productSales\_category**

**where rn = 1;**



1. **Visualize payment method and order status frequency.**
2. **Find any hidden patterns that are counter-intuitive for a layman.**
3. **Can we predict the number of orders, item categories, or the number of customers/amounts in advance?**

Forecast indicator predicts the *number of customers* to reach 15,762 by 1st Quarter of 2019.



*Item Categories* can be predicted in that, it can be inferred what categories will continue to contribute most to the net profit. However, an exact number is difficult to predict because of algorithmic limitations due to too many missing values.

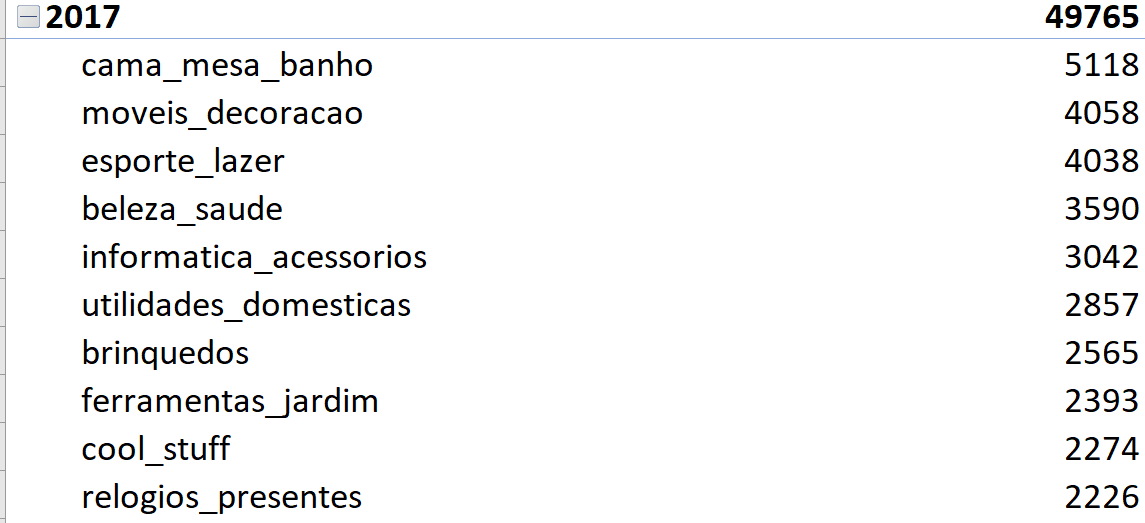
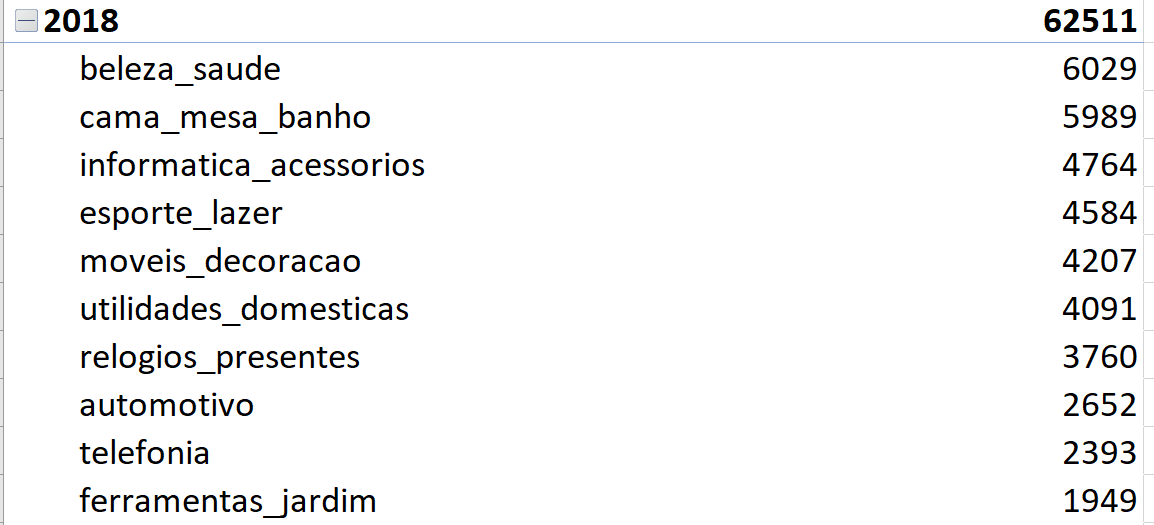
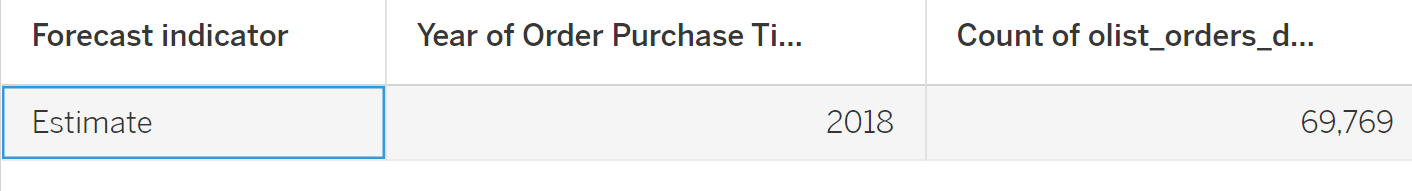
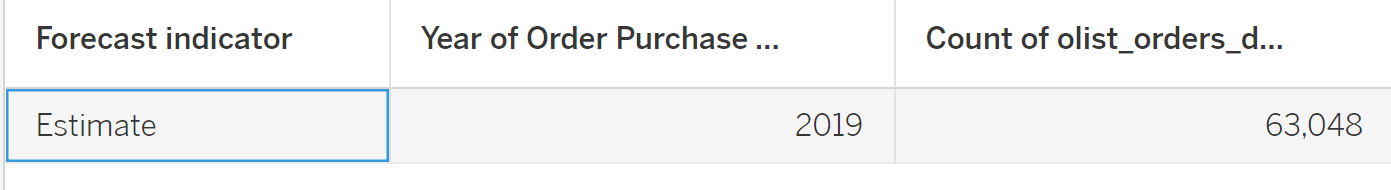
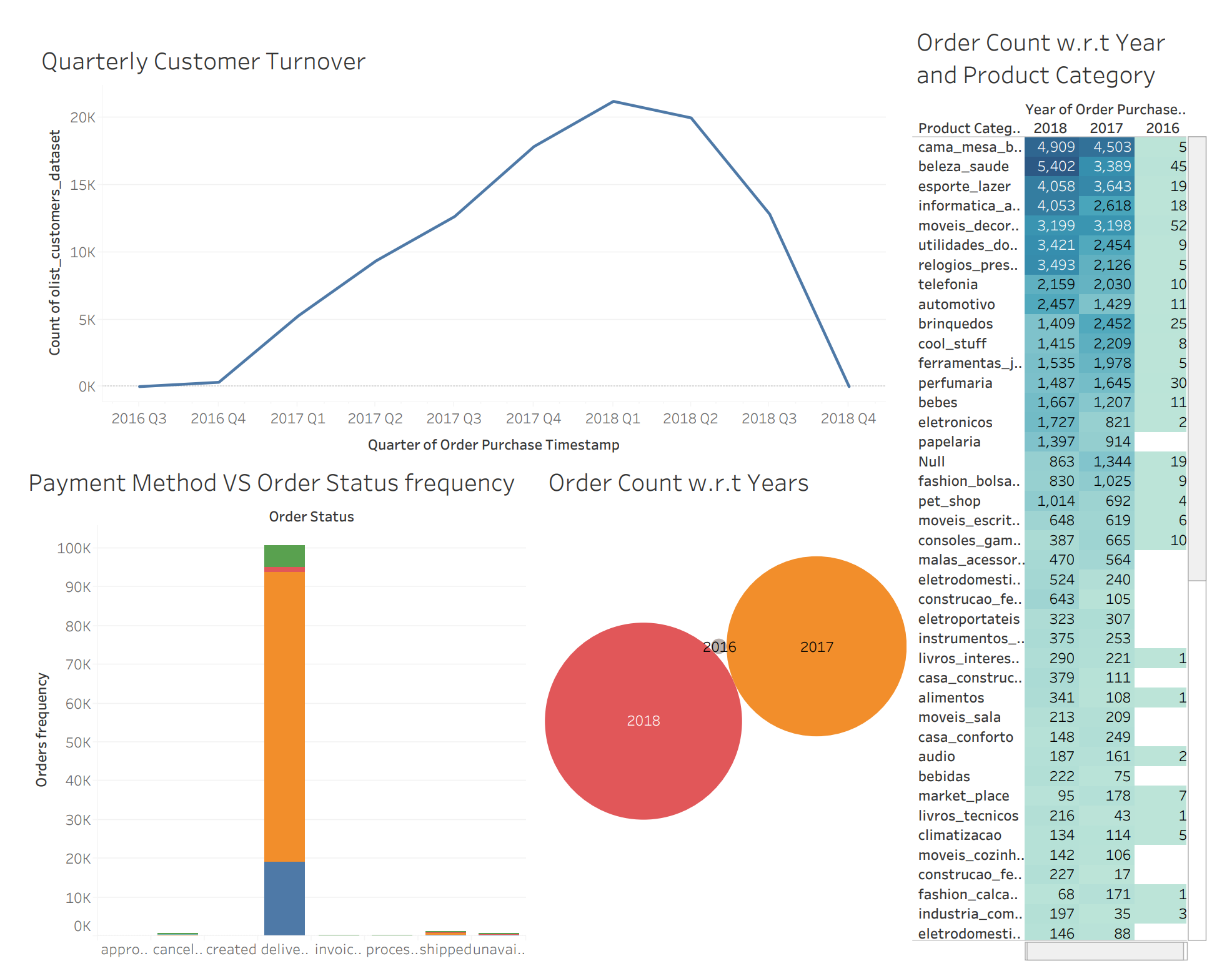
 

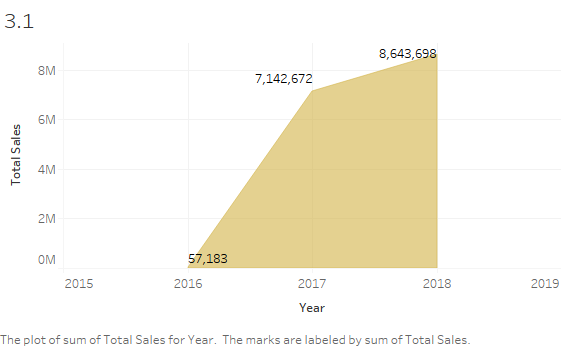
Tableau forecast indicator predicts total *order count* to reach 69,769 by 2018, and dip down to 63,048 in 2019.







1. **Total sales for 2016, 2017 and 2018:**



SELECT EXTRACT(YEAR FROM olist\_orders\_dataset.order\_purchase\_timestamp) AS year,

SUM(olist\_order\_items\_dataset.price + olist\_order\_items\_dataset.freight\_value) AS total\_sales

FROM olist\_orders\_dataset

JOIN olist\_order\_items\_dataset

ON olist\_orders\_dataset.order\_id = olist\_order\_items\_dataset.order\_id

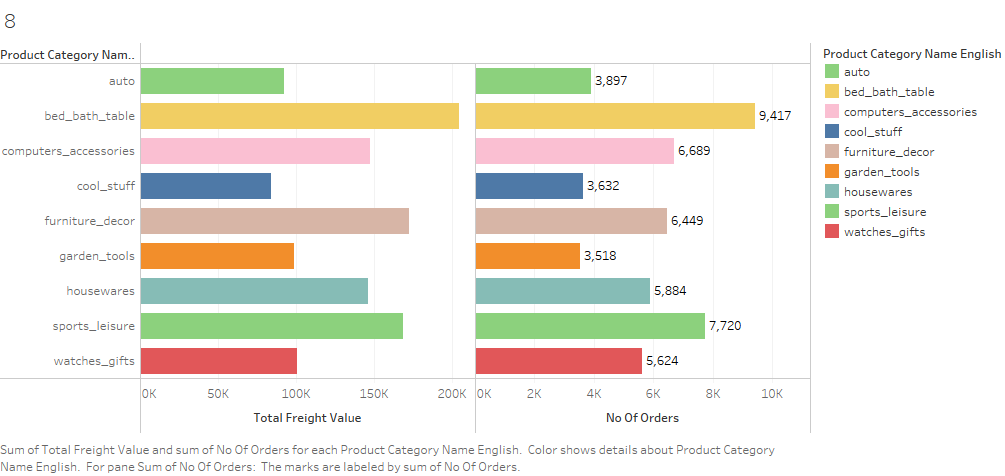
GROUP BY year

ORDER BY year ASC

1. **Which category cost the most transportation cost and why?**

Category Includes: bed\_bath\_table,health\_beauty, furniture\_decor, sports\_leisure, computers\_accessories, Housewares and watches\_gifts.

**High Logistic cost due to larger size, fragility and high no. of orders.**



**SQL Query:**

SELECT pcnt.product\_category\_name\_english, SUM(oi.freight\_value) as total\_freight\_value, COUNT(DISTINCT oi.order\_id) as no\_of\_orders, SUM(op.product\_length\_cm \* op.product\_width\_cm \* op.product\_height\_cm) as total\_size

FROM olist\_order\_items\_dataset oi

JOIN olist\_products\_dataset op

ON oi.product\_id = op.product\_id

JOIN product\_category\_name\_translation pcnt

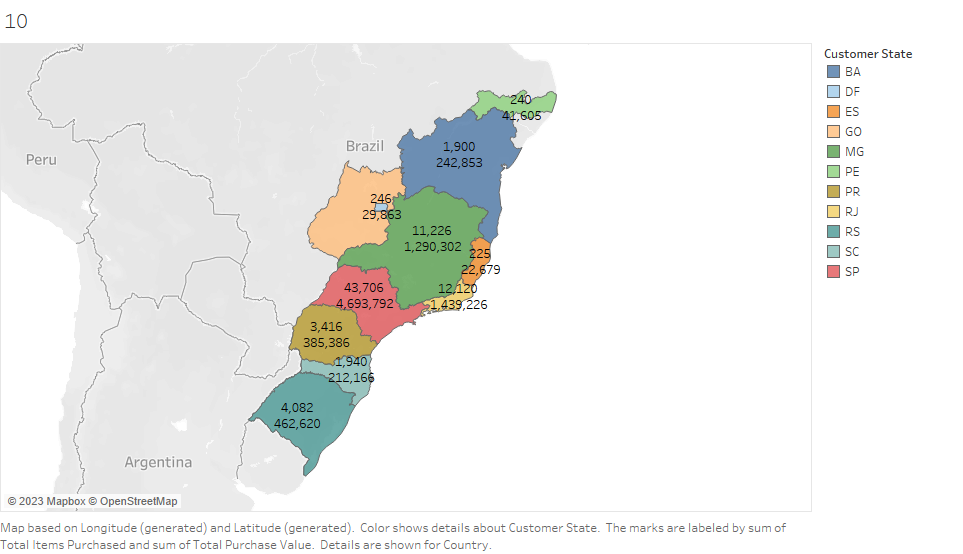
ON op.product\_category\_name = pcnt.product\_category\_name

GROUP BY pcnt.product\_category\_name\_english

ORDER BY total\_freight\_value DESC

LIMIT 9;

1. **Show the shopping patterns of OLIST customers:**



**SQL Query:**

SELECT 'Brazil' as country, ocd.customer\_state as state, SUM(oi.freight\_value) as total\_freight\_value,

COUNT(oo.order\_id) as total\_orders,

SUM(oi.price) as total\_purchase\_value

FROM olist\_customers\_dataset ocd

JOIN olist\_orders\_dataset oo ON ocd.customer\_id = oo.customer\_id

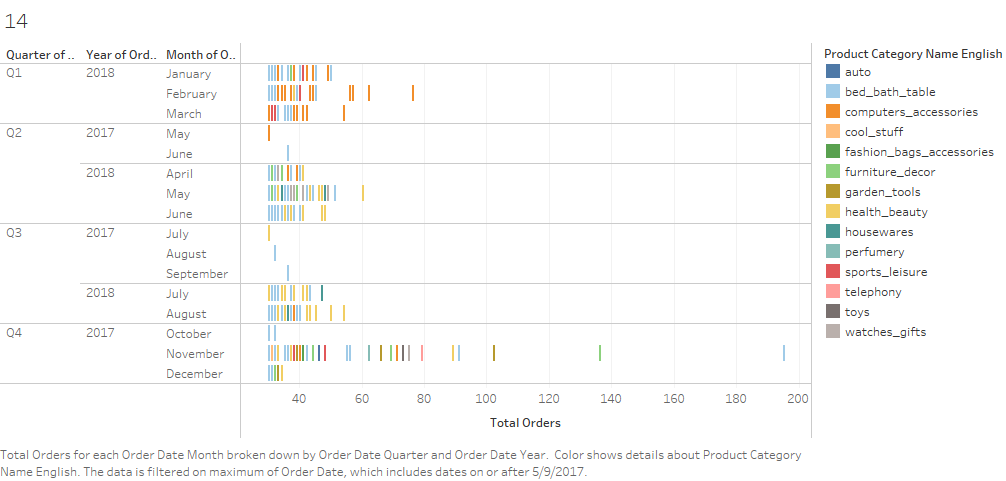
JOIN olist\_order\_items\_dataset oi ON oo.order\_id = oi.order\_id

GROUP BY ocd.customer\_state

ORDER BY total\_purchase\_value DESC

limit 11;

1. **Find a correlation between the order date and item category:**



**SQL Query:**

WITH item\_categories AS (

SELECT oi.order\_id, pcnt.product\_category\_name\_english

FROM olist\_order\_items\_dataset oi

JOIN olist\_products\_dataset op

ON oi.product\_id = op.product\_id

JOIN product\_category\_name\_translation pcnt

ON op.product\_category\_name = pcnt.product\_category\_name

)

SELECT DATE(oo.order\_purchase\_timestamp) as order\_date,

ic.product\_category\_name\_english,

COUNT(ic.order\_id) as total\_orders

FROM olist\_orders\_dataset oo

JOIN item\_categories ic

ON oo.order\_id = ic.order\_id

GROUP BY order\_date, ic.product\_category\_name\_english

ORDER BY total\_orders DESC