Institute of Emerging Careers

Portfolio Project # 01

COURSE: DATA ANALYTICS

SUBMITTED BY

GROUP E - COHORT 12

EYMAN AMEER

NAZISH ABDULLAH

MUHAMMAD JAWAD HASAN

SUMBITTED TO: SIR MISBAH UDDIN

SUBMISSION DATE: 01-06-2024

Contents

Introduction

In this report dataset for a Brazilian ecommerce store Olist is taken from kaggle website. The data is publicly provided by Olist for the period of Oct2016 to Sep2018.

The dataset has information on 100k orders from 2016 to 2018 made at multiple marketplaces in Brazil. This is real commercial data, it has been anonymized, and references to the companies and partners in the review text have been replaced with the names of Game of Thrones great houses.

Olist Store is one of the leading stores in the Brazilian e-commerce market which partners with businesses from all over Brazil to facilitate them to reach a large customer base in all states of Brazil.

The dataset covers all aspects of the revenue streams along with the geographical location of customers and sellers. The store has a significant number of product categories which have multiple products sold under them.

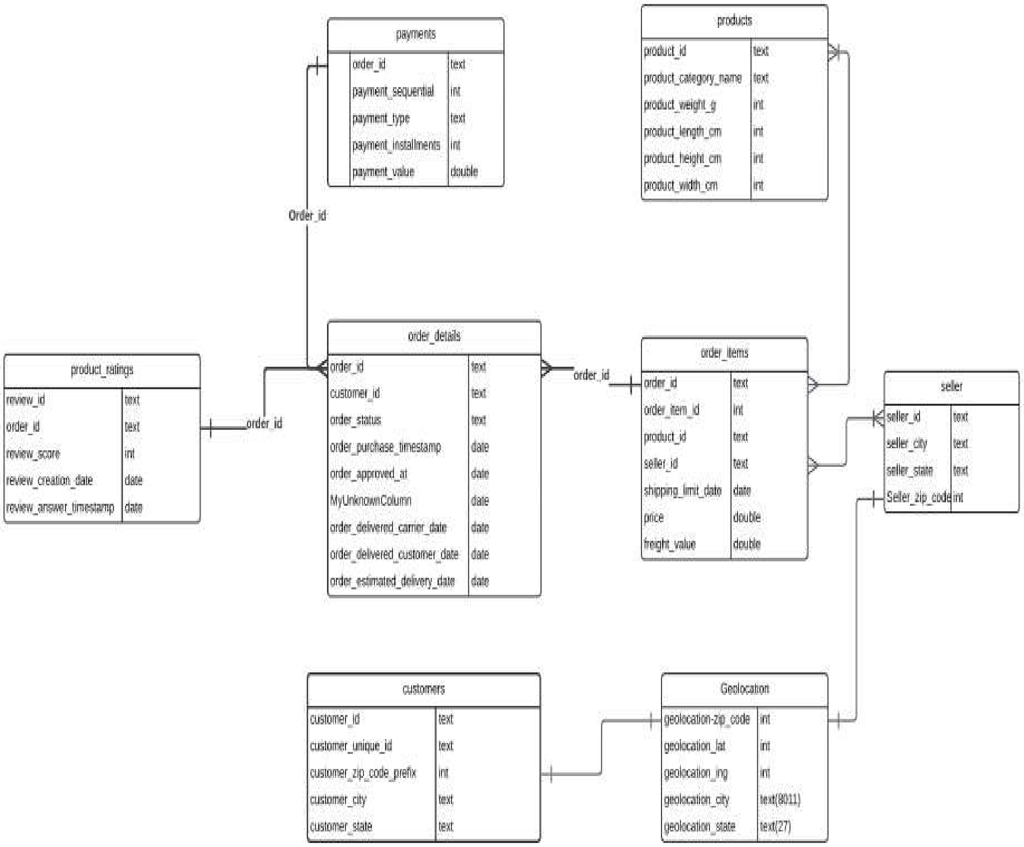
Following are the datasheets provided by the store:

1. Customers
2. Sellers
3. Products
4. Product category
5. Geolocation
6. Order details
7. Order items
8. Order payments
9. Order reviews

Data Dictionary

TABLE_SCHEMA	TABLE_NAME	COLUMN_NAME	ORD_POSITION	DATA_TYPE	CHAR_MAX_LENGTH
brazilianecommerce	customers	customer_city	4	text	99441
brazilianecommerce	customers	customer_id	1	text	99441
brazilianecommerce	customers	customer_state	5	text	99441
brazilianecommerce	customers	customer_unique_id	2	text	99441
brazilianecommerce	customers	customer_zip_code_prefix	3	int	NULL
brazilianecommerce	payments	order_id	1	text	99440
brazilianecommerce	payments	payment_installments	4	int	NULL
brazilianecommerce	payments	payment_sequential	2	int	NULL
brazilianecommerce	payments	payment_type	3	text	99440
brazilianecommerce	payments	payment_value	5	double	NULL
brazilianecommerce	product_cat_eng	product_category_name	1	text	74
brazilianecommerce	product_cat_eng	product_category_name_english	2	text	74
brazilianecommerce	product_ratings	order_id	2	text	99441
brazilianecommerce	product_ratings	review_answer_timestamp	5	text	99441
brazilianecommerce	product_ratings	review_creation_date	4	text	99441
brazilianecommerce	product_ratings	review_id	1	text	99441
brazilianecommerce	product_ratings	review_score	3	int	NULL
brazilianecommerce	products	product_category_name	2	text	99441
brazilianecommerce	products	product_height_cm	5	int	NULL
brazilianecommerce	products	product_id	1	text	99441
brazilianecommerce	products	product_length_cm	4	int	NULL
brazilianecommerce	products	product_weight_g	3	int	NULL
brazilianecommerce	products	product_width_cm	6	int	NULL
brazilianecommerce	seller	seller_city	2	text	3096
brazilianecommerce	seller	seller_id	1	text	3096
brazilianecommerce	seller	seller_zipcode	3	int	Null
brazilianecommerce	seller	seller_state	4	text	3096
brazilianecommerce	orders	customer_id	2	text	99441
brazilianecommerce	orders	order_approved_at	5	date	NULL
brazilianecommerce	orders	order_delivered_carrier_date	7	date	NULL
brazilianecommerce	orders	order_delivered_customer_date	8	date	NULL
brazilianecommerce	orders	order_estimated_delivery_date	9	date	NULL
brazilianecommerce	orders	order_id	1	text	65535
brazilianecommerce	orders	order_purchase_timestamp	4	date	NULL
brazilianecommerce	orders	order_status	3	text	99441
brazilianecommerce	order_items	freight_value	7	double	NULL
brazilianecommerce	order_items	order_id	1	text	99441
brazilianecommerce	order_items	order_item_id	2	int	NULL
brazilianecommerce	order_items	price	6	double	NULL
brazilianecommerce	order_items	product_id	3	text	112650
brazilianecommerce	order_items	seller_id	4	text	112650
brazilianecommerce	order_items	shipping_limit_date	5	date	NULL
brazilianecommerce	geolocation	geolocation-zip_code	1	int	NULL
brazilianecommerce	geolocation	geolocation_lat	2	int	NULL
brazilianecommerce	geolocation	geolocation_ing	3	text	NULL
brazilianecommerce	geolocation	geolocation_city	4	text	8011
brazilianecommerce	geolocation	geolocation_state	5	text	27

Entity Relationship Diagram



payments	
order_id	text
payment_sequential	int
payment_type	text
payment_installments	int
payment_value	double

products	
product_id	text
product_category_name	text
product_weight_g	int
product_length_cm	int
product_height_cm	int
product_width_cm	int

product_ratings	
review_id	text
order_id	text
review_score	int
review_creation_date	date
review_answer_timestamp	date

order_details	
order_id	text
customer_id	text
order_status	text
order_purchase_timestamp	date
order_approved_at	date
MyUnknownColumn	date
order_delivered_carrier_date	date
order_delivered_customer_date	date
order_estimated_delivery_date	date

order_items	
order_id	text
order_item_id	int
product_id	text
seller_id	text
shipping_limit_date	date
price	double
freight_value	double

seller	
seller_id	text
seller_city	text
seller_state	text
Seller_zip_code	int

customers	
customer_id	text
customer_unique_id	text
customer_zip_code_prefix	int
customer_city	text
customer_state	text

Geolocation	
geolocation-zip_code	int
geolocation_lat	int
geolocation_ing	int
geolocation_city	text(8011)
geolocation_state	text(27)

The above diagram defines one to one and one-to-many relationship between the fields.

Objective

In this report, we have used the given datasets to analyze Olist sales concerning different business variables.

Olist has over 70 product categories, which provides the customer with a single platform to fulfil their shopping needs. In our analysis, we aim to find the most selling products, high sales generating streams, and different variants associated with them.

Business Problem

Olist store is one of the leading e-commerce stores in Brazil. Its customer and supplier base is spread across the country with a diversified portfolio of products. To have a clear picture of how and which factors affect the sales of the company, we look into following four dimensions

- How many ***product categories*** and products does the store offer? Which are the most selling product categories and generate high sales revenues?
- What is the pattern in sales generated across ***different states*** in Brazil? How do customer demographics affect the sales?
- Which ***payment methods*** are used while purchasing from the Olist platform? And how does it impact sales?
- What are the ***seasonal trends*** in sales? In which season do customer are more inclined towards purchasing and what are their product preferences?

For our analysis, we imported the Olist datasheets in MYSQL Workbench and MS PowerBi.
In the following section, identified problems are explained using queries, run on MYSQL Workbench supported with visuals from the PowerBi dashboard.

Problems

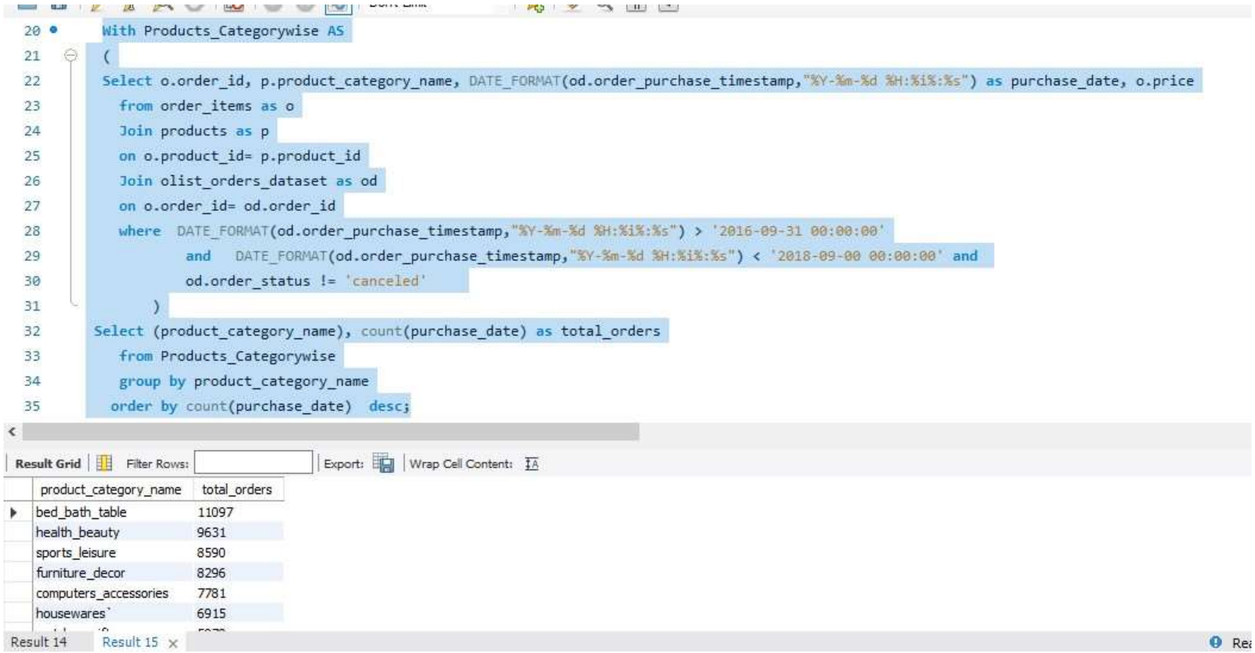
On searching through data, following statements were identified for analysis.

Query no.1

To calculate the number of orders in each product category placed from Sep 2016 to Aug 2018. This will result in finding out the high-in-demand product categories and the volume of sales they generate.

```
With Products_Categorywise AS
  (
  Select o.order_id, p.product_category_name,
  DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") as
  purchase_date, o.price
   from order_items as o
   Join products as p
   on o.product_id= p.product_id
   Join olist_orders_dataset as od
   on o.order_id= od.order_id
   where  DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") > '2016-
09-31 00:00:00'
            and   DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") <
'2018-09-00 00:00:00' and
            od.order_status != 'canceled'
                        )
 Select (product_category_name), count(purchase_date) as total_orders
    from Products_Categorywise
    group by product_category_name
   order by count(purchase_date)  desc;
```

Initially a CTE named Products_categorywise is designed to be used in few of the queries. This query converts the text type in date format in order items table and extract selected fields for the orders executed from Oct 2016 to Sep 2018.



```sql
 With Products_Categorywise AS
 (
 Select o.order_id, p.product_category_name, DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") as purchase_date, o.price
   from order_items as o
   Join products as p
   on o.product_id= p.product_id
   Join olist_orders_dataset as od
   on o.order_id= od.order_id
   where  DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") > '2016-09-31 00:00:00'
          and   DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") < '2018-09-00 00:00:00' and
          od.order_status != 'canceled'
       )
Select (product_category_name), count(purchase_date) as total_orders
   from Products_Categorywise
   group by product_category_name
  order by count(purchase_date)  desc;
```

product_category_name	total_orders
bed_bath_table	11097
health_beauty	9631
sports_leisure	8590
furniture_decor	8296
computers_accessories	7781
housewares`	6915

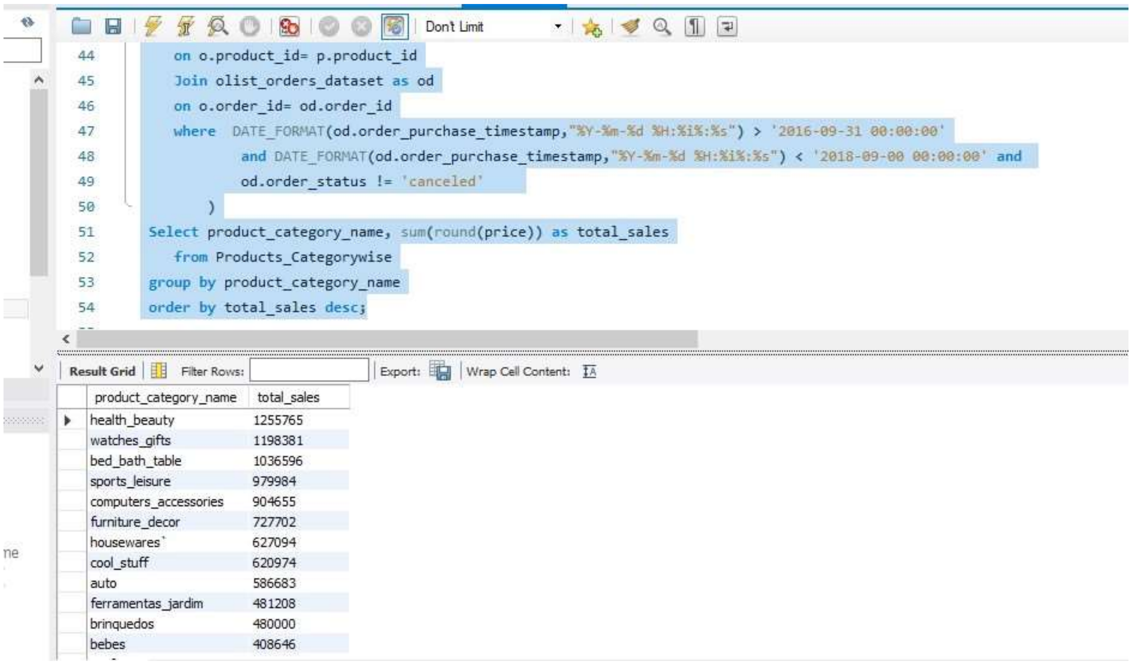
Insight

From the above result, it can be concluded that bed and bath had over 11000 orders during the period followed closely by health and beauty products.

Query no.2

To calculate the total sales revenue for each product category purchased from Sep 2016 to Aug 2018. This will result in finding out the high sales revenue generating product categories.

```
  With Products_Categorywise AS
 (
 Select o.order_id, p.product_category_name,
DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") as purchase_date, o.price
   from order_items as o
   Join products as p
   on o.product_id= p.product_id
   Join olist_orders_dataset as od
   on o.order_id= od.order_id
   where  DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") > '2016-09-31 00:00:00'
          and DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") < '2018-09-00 00:00:00' and
          od.order_status != 'canceled'
                    )
 Select product_category_name, sum(round(price)) as total_sales
   from Products_Categorywise
 group by product_category_name
 order by total_sales desc;
```



```
    on o.product_id= p.product_id
    Join olist_orders_dataset as od
    on o.order_id= od.order_id
    where  DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") > '2016-09-31 00:00:00'
            and DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") < '2018-09-00 00:00:00' and
            od.order_status != 'canceled'
        )
Select product_category_name, sum(round(price)) as total_sales
   from Products_Categorywise
group by product_category_name
order by total_sales desc;
```

product_category_name	total_sales
health_beauty	1255765
watches_gifts	1198381
bed_bath_table	1036596
sports_leisure	979984
computers_accessories	904655
furniture_decor	727702
housewares`	627094
cool_stuff	620974
auto	586683
ferramentas_jardim	481208
brinquedos	480000
bebes	408646

Insight

Health & beauty products have brought in the highest revenue of 12.5 MN BRL, whereas watches and gifts have relatively fewer orders but have bought a high revenue of almost 1.2 MN BRL.

Query no.3

To find average review ratings for top-selling products.

Select p.product_category_name, (sum (pr. review_score)/count(pr.review_score)) as average_rating

from order_items as o

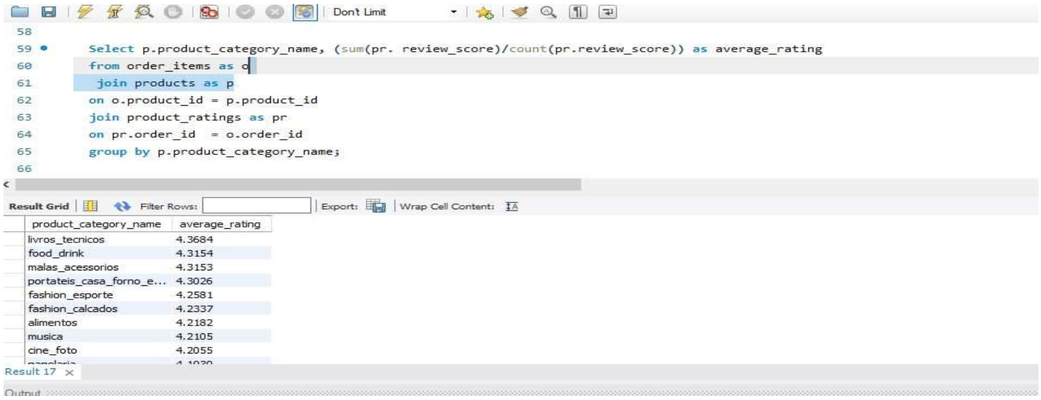
join products as p

on o.product_id = p.product_id

join product_ratings as pr

on pr.order_id = o.order_id

group by p.product_category_name;



```
Select p.product_category_name, (sum(pr. review_score)/count(pr.review_score)) as average_rating
from order_items as o
 join products as p
on o.product_id = p.product_id
join product_ratings as pr
on pr.order_id  = o.order_id
group by p.product_category_name;
```

product_category_name	average_rating
livros_tecnicos	4.3684
food_drink	4.3154
malas_acessorios	4.3153
portateis_casa_forno_e...	4.3026
fashion_esporte	4.2581
fashion_calcados	4.2337
alimentos	4.2182
musica	4.2105
cine_foto	4.2055

Insight

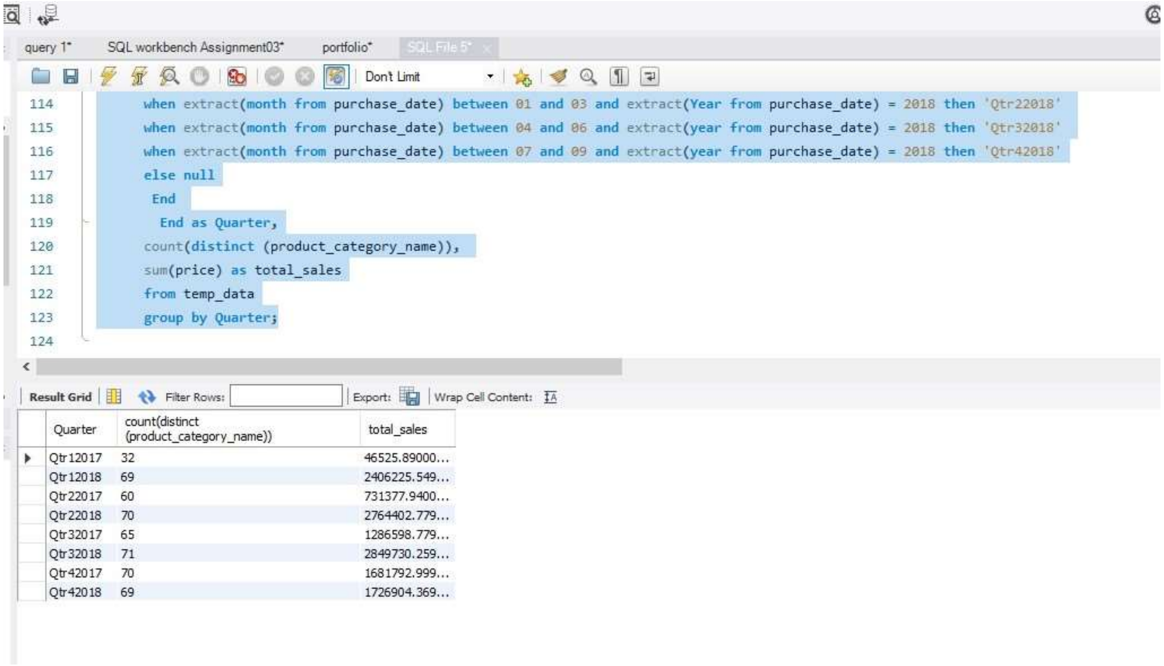
The maximum average review rate is 4.3, which means there are areas where customers' expectations are not met. The underlying reason may be delayed delivery time or high freight charges for low-priced products. The following table shows the average review rating for top-selling product categories

Product_Category_Name with no.orders >5000	Average_Rating
bed_bath_table	3.8957
health_beauty	4.1428
sports_leisure	4.108
furniture_decor	3.9035
computers_accessories	3.9308
housewares`	4.055
watches_gifts	4.0192

Query no.4

To find sales pattern over the quarters in two years.

```
Select
  Case
  when extract(month from purchase_date) between 10 and 12 and extract(year from
purchase_date) = 2016 then 'Qtr12017'
            when extract(month from purchase_date) between 01 and 03 and extract(Year from
purchase_date) = 2017 then 'Qtr22017'
      when extract(month from purchase_date) between 04 and 06 and extract(year from
purchase_date) = 2017 then 'Qtr32017'
            when extract(month from purchase_date) between 07 and 09 and extract(year from
purchase_date) = 2017 then 'Qtr42017'
      else
      case
            when extract(month from purchase_date) between 10 and 12 and extract(year from
purchase_date) = 2017 then 'Qtr12018'
            when extract(month from purchase_date) between 01 and 03 and extract(Year from
purchase_date) = 2018 then 'Qtr22018'
      when extract(month from purchase_date) between 04 and 06 and extract(year from
purchase_date) = 2018 then 'Qtr32018'
            when extract(month from purchase_date) between 07 and 09 and extract(year from
purchase_date) = 2018 then 'Qtr42018'
      else null
       End
        End as Quarter,
      count(distinct (product_category_name)),
      sum(price) as total_sales
            from temp_data
      group by Quarter;
```



```sql
        when extract(month from purchase_date) between 01 and 03 and extract(Year from purchase_date) = 2018 then 'Qtr22018'
        when extract(month from purchase_date) between 04 and 06 and extract(year from purchase_date) = 2018 then 'Qtr32018'
        when extract(month from purchase_date) between 07 and 09 and extract(year from purchase_date) = 2018 then 'Qtr42018'
        else null
         End
          End as Quarter,
        count(distinct (product_category_name)),
        sum(price) as total_sales
        from temp_data
        group by Quarter;
```

Quarter	count(distinct (product_category_name))	total_sales
Qtr12017	32	46525.89000...
Qtr12018	69	2406225.549...
Qtr22017	60	731377.9400...
Qtr22018	70	2764402.779...
Qtr32017	65	1286598.779...
Qtr32018	71	2849730.259...
Qtr42017	70	1681792.999...
Qtr42018	69	1726904.369...

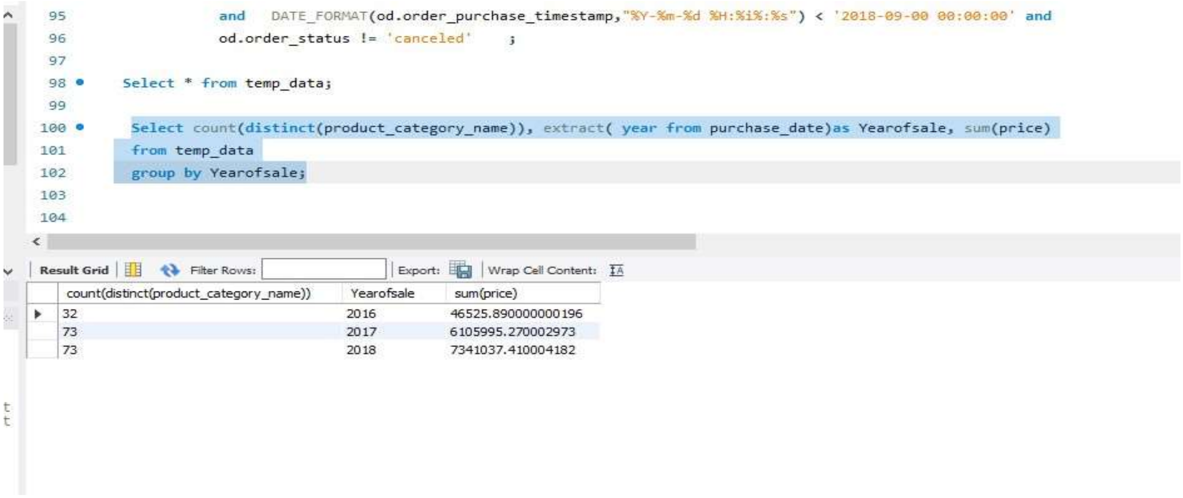
Insight

There was a sharp rise in sales from quarter 1 of 2017 to quarter 1 of 2018 because Olist introduced new product categories in Jan 2017. This is also verified by an increase in sales in 2nd quarter of 2017 only. The gap between the revenues in the following quarters shows a decline as we see more product categories subsequently introduced in 2017.

Query no.5

To find number of product categories sold during two years.
This query will support the result derived from our previous query.

Select count(distinct(product_category_name)), extract(year from purchase_date)as Yearofsale, sum(price)
from temp_data
group by Yearofsale;



```
                and   DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") < '2018-09-00 00:00:00' and
                od.order_status != 'canceled'    ;

  Select * from temp_data;

   Select count(distinct(product_category_name)), extract( year from purchase_date)as Yearofsale, sum(price)
   from temp_data
   group by Yearofsale;
```

```
  count(distinct(product_category_name))   Yearofsale   sum(price)
▶ 32                                       2016         46525.890000000196
  73                                       2017         6105995.270002973
  73                                       2018         7341037.410004182
```

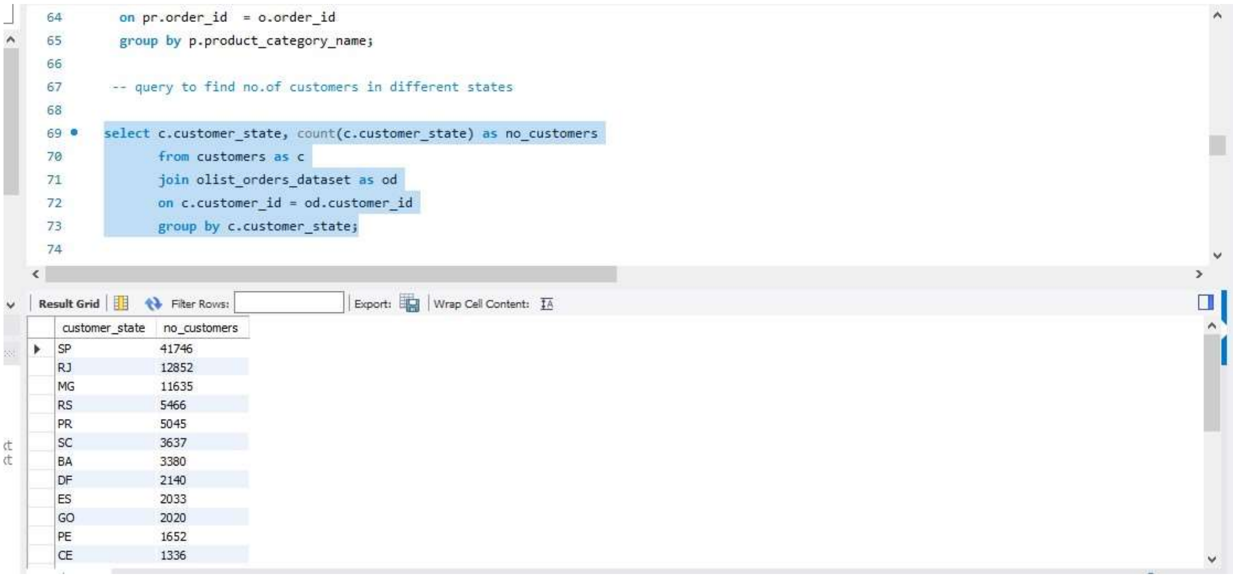
Insight

From the above table, it is evident that sale revenue has substantially increased in year 2017 due to the addition of multiple new product categories on the store’s platform.

Query no.6

To find number of customers residing in different states of Brazil.
This query will provide insight on customer demographics.

Select
c.customer_state, count(c.customer_state) as no_customers
from customers as c
join olist_orders_dataset as od
on c.customer_id = od.customer_id
Group by c.customer_state;



```
     on pr.order_id  = o.order_id
     group by p.product_category_name;

    -- query to find no.of customers in different states

   select c.customer_state, count(c.customer_state) as no_customers
          from customers as c
          join olist_orders_dataset as od
          on c.customer_id = od.customer_id
          group by c.customer_state;
```

customer_state	no_customers
SP	41746
RJ	12852
MG	11635
RS	5466
PR	5045
SC	3637
BA	3380
DF	2140
ES	2033
GO	2020
PE	1652
CE	1336

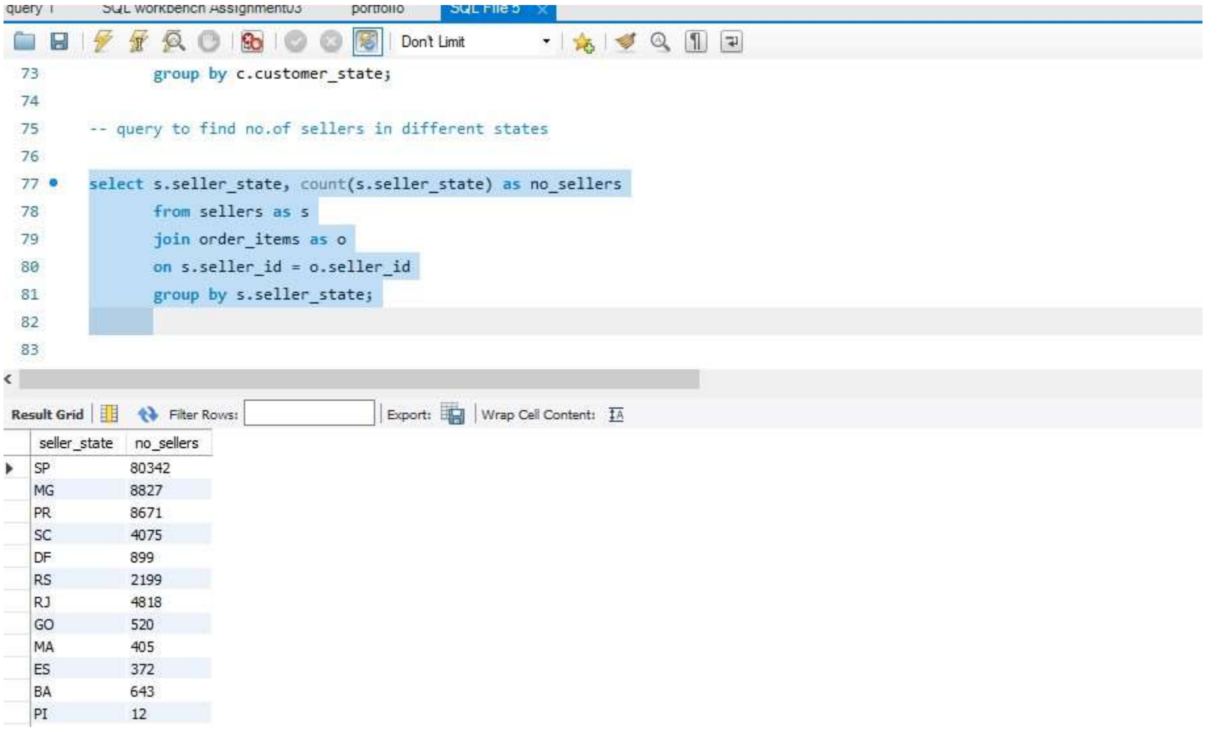
Insight

It is observed that the highest number of customers are from SP, and that's comprehensible as São Paulo is the largest and most populous state in Brazil, located in the Southeast Region. It has more than 600 municipalities, among which few are known for being extremely advanced in technology.
Due to high awareness and accessibility to internet and networks Olist makes high profit from urban areas.

Query no.7

To find where are the most sellers located in different states of Brazil.

```
Select
      s.seller_state, count(s.seller_state) as no_sellers
      from sellers as s
      join order_items as o
      on s.seller_id = o.seller_id
      group by s.seller_state;
```



```
        group by c.customer_state;

-- query to find no.of sellers in different states

select s.seller_state, count(s.seller_state) as no_sellers
        from sellers as s
        join order_items as o
        on s.seller_id = o.seller_id
        group by s.seller_state;
```

seller_state	no_sellers
SP	80342
MG	8827
PR	8671
SC	4075
DF	899
RS	2199
RJ	4818
GO	520
MA	405
ES	372
BA	643
PI	12

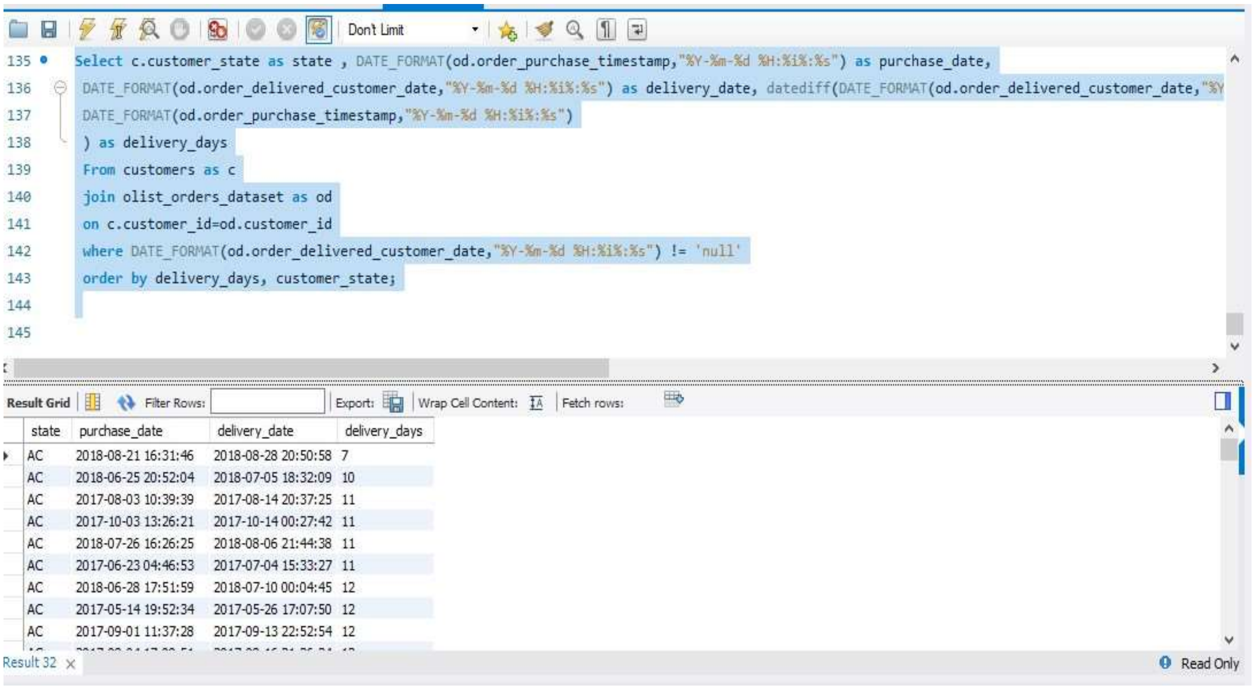
Insights

The majority of the suppliers are from SP. As SP is the most populated and developed state in Brazil, local businesses in this area are well aware of the usage of e-commerce platforms and modern modes of executing business transactions.

Query no.8

To find delivery days in different states

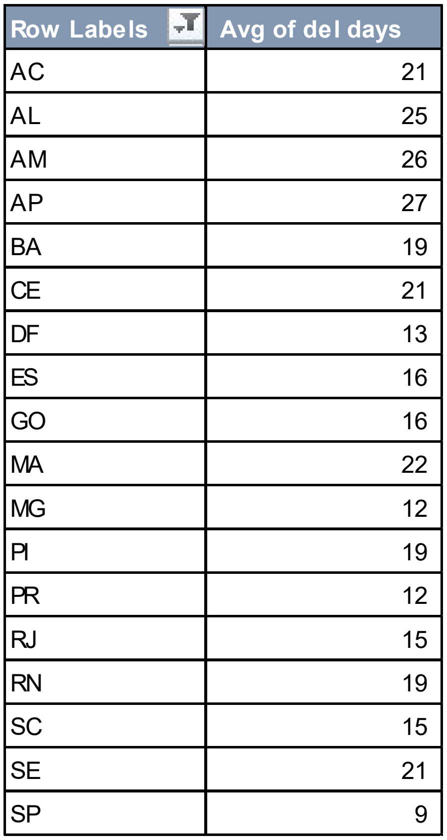
```
Select c.customer_state as state , DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-
%d %H:%i%:%s") as purchase_date,
 DATE_FORMAT(od.order_delivered_customer_date,"%Y-%m-%d %H:%i%:%s") as
delivery_date, datediff(DATE_FORMAT(od.order_delivered_customer_date,"%Y-%m-%d
%H:%i%:%s"),
 DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s")
 ) as delivery_days
 From customers as c
 join olist_orders_dataset as od
 on c.customer_id=od.customer_id
 where DATE_FORMAT(od.order_delivered_customer_date,"%Y-%m-%d %H:%i%:%s") != 'null'
 order by delivery_days, customer_state;
```



```
Select c.customer_state as state , DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s") as purchase_date,
 DATE_FORMAT(od.order_delivered_customer_date,"%Y-%m-%d %H:%i%:%s") as delivery_date, datediff(DATE_FORMAT(od.order_delivered_customer_date,"%Y
 DATE_FORMAT(od.order_purchase_timestamp,"%Y-%m-%d %H:%i%:%s")
 ) as delivery_days
 From customers as c
 join olist_orders_dataset as od
 on c.customer_id=od.customer_id
 where DATE_FORMAT(od.order_delivered_customer_date,"%Y-%m-%d %H:%i%:%s") != 'null'
 order by delivery_days, customer_state;
```

state	purchase_date	delivery_date	delivery_days
AC	2018-08-21 16:31:46	2018-08-28 20:50:58	7
AC	2018-06-25 20:52:04	2018-07-05 18:32:09	10
AC	2017-08-03 10:39:39	2017-08-14 20:37:25	11
AC	2017-10-03 13:26:21	2017-10-14 00:27:42	11
AC	2018-07-26 16:26:25	2018-08-06 21:44:38	11
AC	2017-06-23 04:46:53	2017-07-04 15:33:27	11
AC	2018-06-28 17:51:59	2018-07-10 00:04:45	12
AC	2017-05-14 19:52:34	2017-05-26 17:07:50	12
AC	2017-09-01 11:37:28	2017-09-13 22:52:54	12

Insights



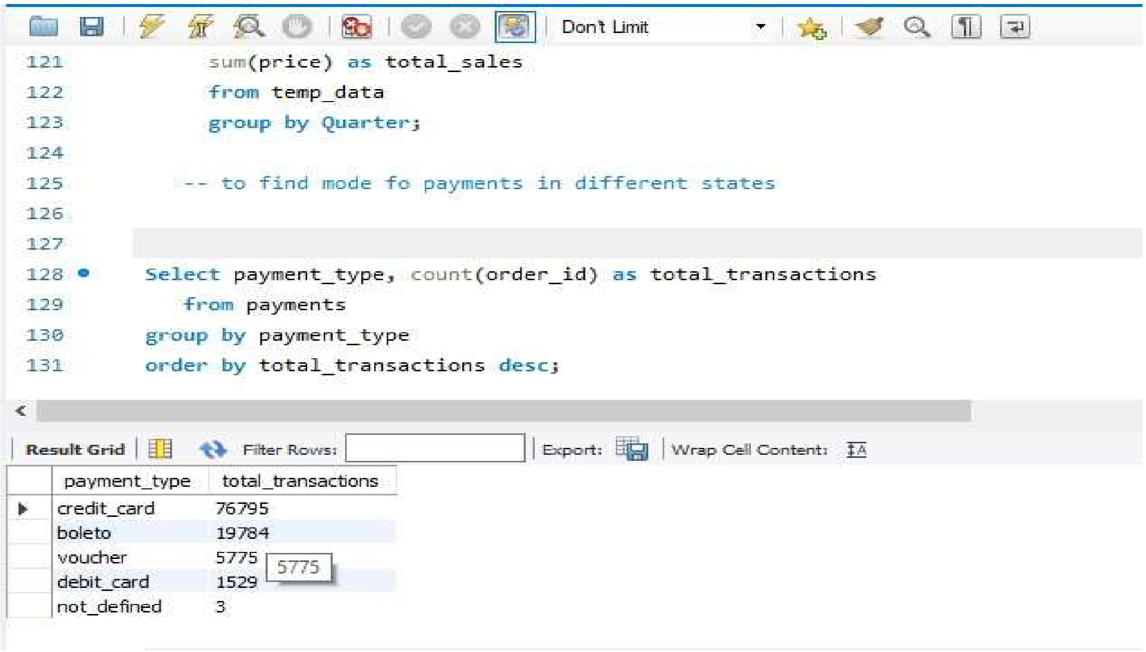
Row Labels	Avg of del days
AC	21
AL	25
AM	26
AP	27
BA	19
CE	21
DF	13
ES	16
GO	16
MA	22
MG	12
PI	19
PR	12
RJ	15
RN	19
SC	15
SE	21
SP	9

The above table clearly shows that average delivery period offered to customers in states having low sales is above 20 days.

Query no.9

To find customer preferences for using various modes of payment in different states.

```
Select payment_type, count(order_id) as total_transactions
   from payments
 group by payment_type
 order by total_transactions desc;
```



```
        sum(price) as total_sales
        from temp_data
        group by Quarter;

     -- to find mode fo payments in different states


Select payment_type, count(order_id) as total_transactions
   from payments
group by payment_type
order by total_transactions desc;
```

payment_type	total_transactions
credit_card	76795
boleto	19784
voucher	5775
debit_card	1529
not_defined	3

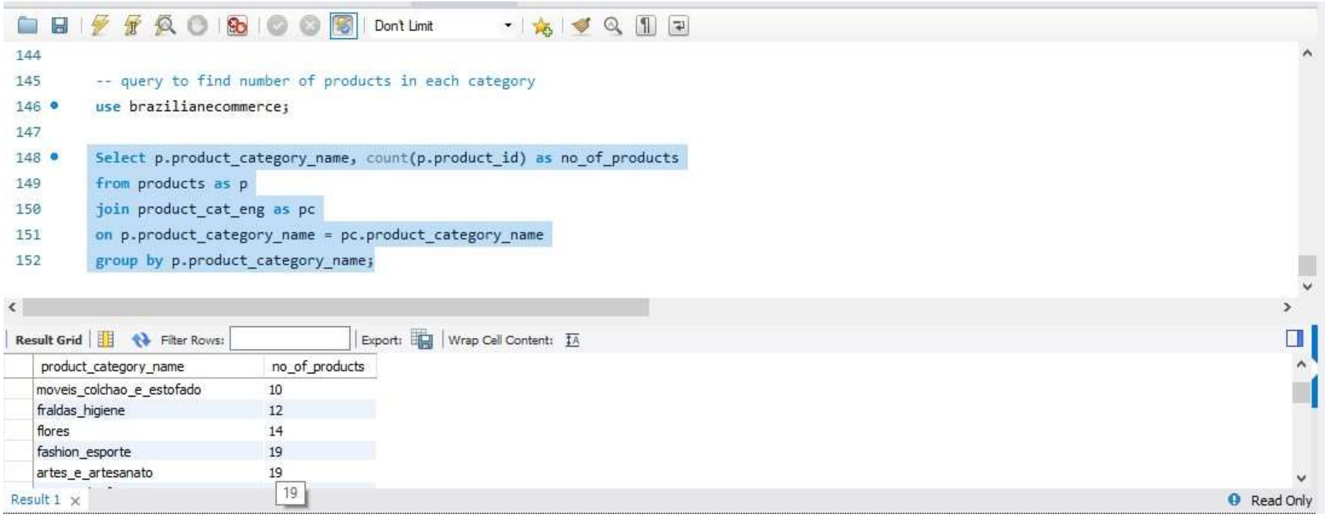
Insight

The above query result reflects on extensive usage of credit cards for payments by customers. The reason behind this trend may be promotional campaigns introduced by Olist in partnership with credit card companies. Further, it is also observed that payments are received in multiple installments when purchases are made via credit cards.

Query no. 10

To find number of products in each category

```
Select p.product_category_name, count(p.product_id) as no_of_products
 from products as p
 join product_cat_eng as pc
 on p.product_category_name = pc.product_category_name
 group by p.product_category_name;
```



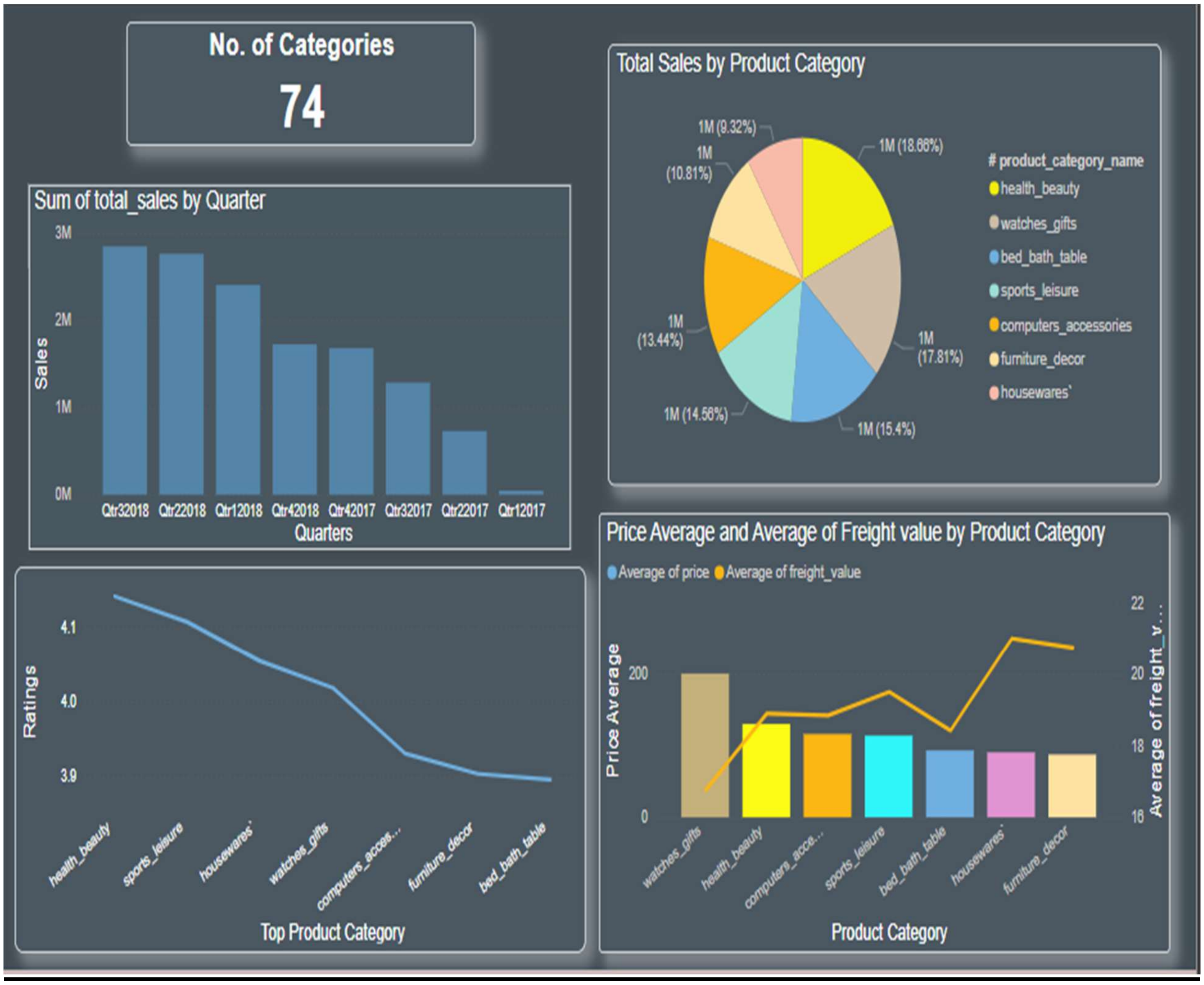
```
-- query to find number of products in each category
use brazilianecommerce;

Select p.product_category_name, count(p.product_id) as no_of_products
from products as p
join product_cat_eng as pc
on p.product_category_name = pc.product_category_name
group by p.product_category_name;
```

product_category_name	no_of_products
moveis_colchao_e_estofado	10
fraldas_higiene	12
flores	14
fashion_esporte	19
artes_e_artesanato	19

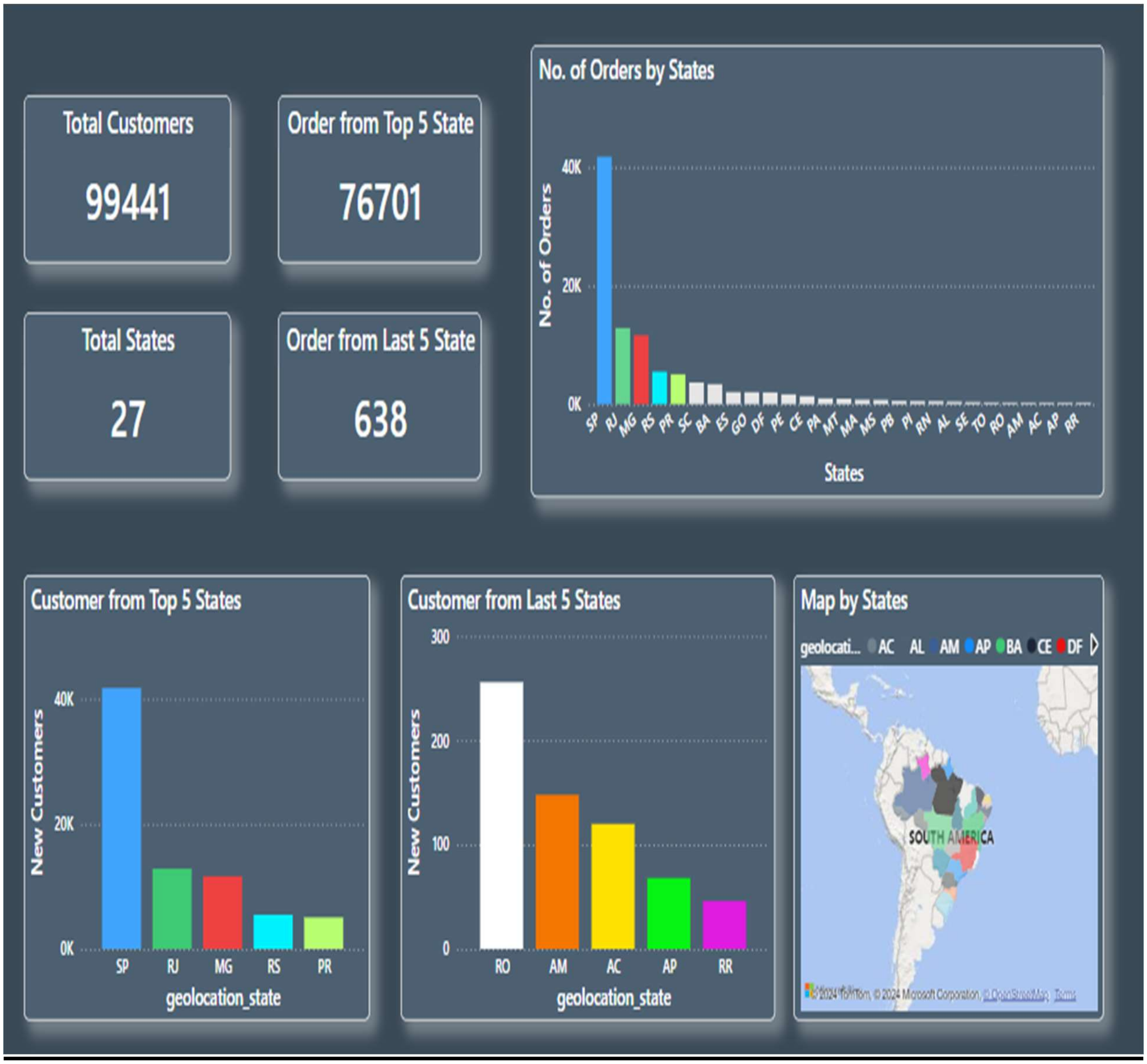
Dashboards

The dataset has been studied in detail using MS PowerBi and dashboards have been designed to give a holistic synopsis of our report.

Analysis by Product Category



Regional Disparity Analysis



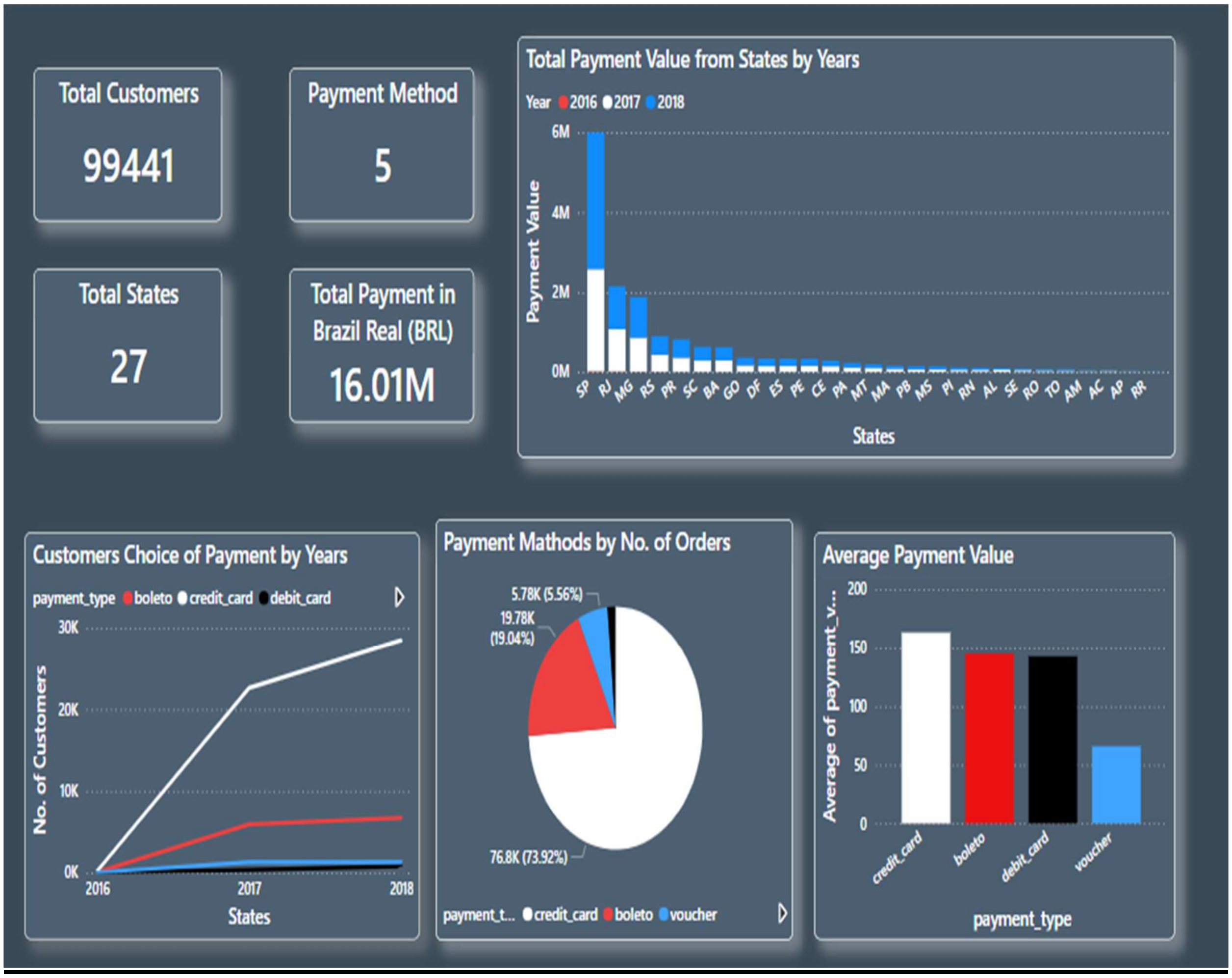
Total Customers 99441

Order from Top 5 State 76701

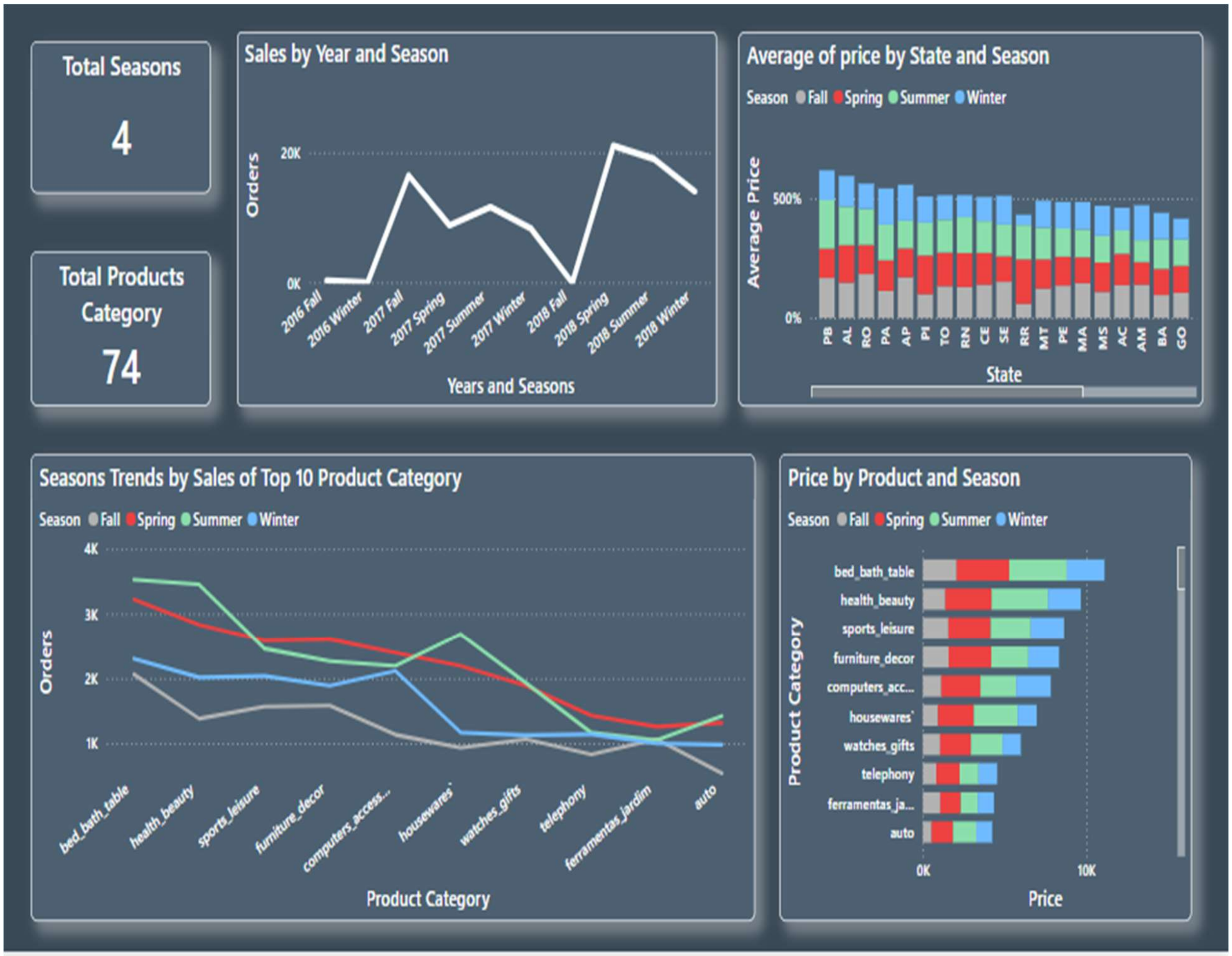
Total States 27

Order from Last 5 State 638

Payment Method Analysis


Total Customers
99441
Payment Method
5
Total States
27
Total Payment in Brazil Real (BRL)
16.01M
Total Payment Value from States by Years
Year ● 2016 ● 2017 ● 2018
6M
4M
2M
0M
Payment Value
SP RJ MG RS PR SC BA GO DF ES PE CE PA MT MA PB MS PI RN AL SE RO TO AM AC AP RR
States
Customers Choice of Payment by Years
payment_type ● boleto ● credit_card ● debit_card
30K
20K
10K
0K
No. of Customers
2016
2017
2018
States
Payment Mathods by No. of Orders
5.78K (5.56%)
19.78K (19.04%)
76.8K (73.92%)
payment_t... ● credit_card ● boleto ● voucher
Average Payment Value
200
150
100
50
0
Average of payment_v...
credit_card
boleto
debit_card
voucher
payment_type

Seasonal Analysis



Recommendations

- Increasing trend in sales indicates opportunities for further growth, hence Olist should strengthen its network across the country and adopt strategies to reach more number of customers.
- Olist should focus on partnering with sellers who provide competitive products in the categories of bed & bath, health & beauty, sports leisure, and furniture decoration.
- The Brazilian potential customer base is enormous and with appropriate market analysis, Olist can penetrate untapped regions.
- At the same time by adopting technology advancements and understanding customer demands, the increasing trend in sales cannot only be retained but also taken to the next level.
- In SP and other developed regions, further partnerships should be made to capitalize on the increasing trend of online shopping.
- For low-sales regions, implement targeted marketing campaigns to raise awareness and stimulate demand.
- Ensure the availability of popular payment methods in each region.
- Launch marketing campaigns and promotions around a high-demand season of summer to maximize sales.