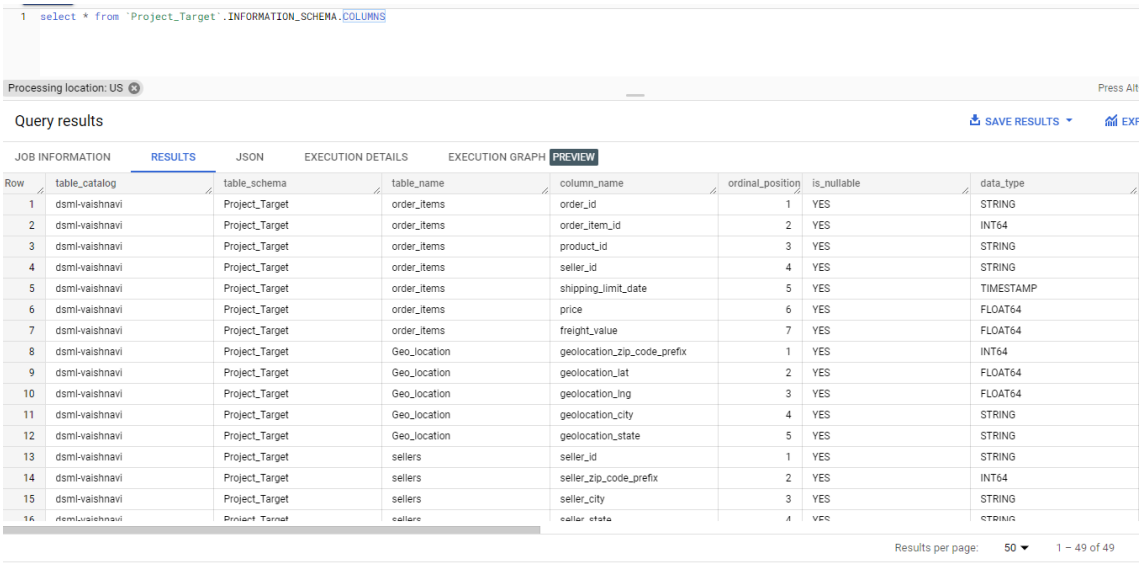
Import the dataset and do usual exploratory analysis steps like checking the structure & characteristics of the dataset

1. Data type of columns in a table?

```
select * from `Project_Target`.INFORMATION_SCHEMA.COLUMNS
```



Processing location: US

Query results

SAVE RESULTS

JOB INFORMATION | RESULTS | JSON | EXECUTION DETAILS | EXECUTION GRAPH | PREVIEW

Row	table_catalog	table_schema	table_name	column_name	ordinal_position	is_nullable	data_type
1	dsml-vaishnavi	Project_Target	order_items	order_id	1	YES	STRING
2	dsml-vaishnavi	Project_Target	order_items	order_item_id	2	YES	INT64
3	dsml-vaishnavi	Project_Target	order_items	product_id	3	YES	STRING
4	dsml-vaishnavi	Project_Target	order_items	seller_id	4	YES	STRING
5	dsml-vaishnavi	Project_Target	order_items	shipping_limit_date	5	YES	TIMESTAMP
6	dsml-vaishnavi	Project_Target	order_items	price	6	YES	FLOAT64
7	dsml-vaishnavi	Project_Target	order_items	freight_value	7	YES	FLOAT64
8	dsml-vaishnavi	Project_Target	Geo_location	geolocation_zip_code_prefix	1	YES	INT64
9	dsml-vaishnavi	Project_Target	Geo_location	geolocation_lat	2	YES	FLOAT64
10	dsml-vaishnavi	Project_Target	Geo_location	geolocation_lng	3	YES	FLOAT64
11	dsml-vaishnavi	Project_Target	Geo_location	geolocation_city	4	YES	STRING
12	dsml-vaishnavi	Project_Target	Geo_location	geolocation_state	5	YES	STRING
13	dsml-vaishnavi	Project_Target	sellers	seller_id	1	YES	STRING
14	dsml-vaishnavi	Project_Target	sellers	seller_zip_code_prefix	2	YES	INT64
15	dsml-vaishnavi	Project_Target	sellers	seller_city	3	YES	STRING

Results per page: 50 1 – 49 of 49

2. Time period for which the data is given

```
select
min(EXTRACT(date from order_purchase_timestamp)) as Strat_period,
max(EXTRACT(date from order_purchase_timestamp)) as End_period
from `Project_Target.orders`
```

Row	Strat_period	End_period
1	2016-09-04	2018-10-17

Insights: We have data from September 2016 to October 2018 only.

3. Cities and States of customers ordered during the given period

```
select
distinct C.customer_city,
C.customer_state,
from `Project_Target.Customer` as C join `Project_Target.orders` as O on C.cus-
tomer_id=O.customer_id
join `Project_Target.Geo_location` as G on G.geolocation_zip_code_prefix = C.cus-
tomer_zip_code_prefix
where O.order_purchase_timestamp is not null
group by C.customer_city,C.customer_state
```

Row	customer_city	customer_state
1	acu	RN
2	ico	CE
3	ipe	RS
4	ipu	CE
5	ita	SC
6	itu	SP
7	jau	SP
8	luz	MG
9	poa	SP
10	uba	MG
11	una	BA
12	anta	RJ

In-depth Exploration:

Brazil

You can select the following states for Brazil:

State or Province	2-digit Code
Acre	AC
Alagoas	AL
Amapa	AP
Amazonas	AM
Bahia	BA
Ceara	CE
Distrito Federal	DF
Espirito Santo	ES
Goias	GO
Maranhao	MA
Mato Grosso	MT
Mato Grosso do Sul	MS
Minas Gerais	MG
Para	PA
Paraiba	PB
Parana	PR
Pernambuco	PE
Piaui	PI
Rio Grande do Norte	RN
Rio Grande do Sul	RS
Rio de Janeiro	RJ
Rondonia	RO
Roraima	RR
Santa Catarina	SC
Sao Paulo	SP
Sergipe	SE
Tocantins	TO

1. Is there a growing trend on e-commerce in Brazil? How can we describe a complete scenario? Can we see some seasonality with peaks at specific months?

```
select
extract(MONTH from O.order_purchase_timestamp) as month,
extract(YEAR from O.order_purchase_timestamp) as year,
count(O.order_id) as No_of_orders
from `Project_Target.orders` as O join `Project_Target.payments` as P
on O.order_id = P.order_id
group by year,month
order by Purchase_value desc
```

Row	month	year	Purchase_value
1	11	2017	1194882.80...
2	4	2018	1160785.47...
3	3	2018	1159652.11...
4	5	2018	1153982.14...
5	1	2018	1115004.18...
6	7	2018	1066540.75...
7	6	2018	1023880.49...
8	8	2018	1022425.32...
9	2	2018	992463.340...

Row	month	year	Purchase_value
16	6	2017	511276.380...
17	3	2017	449863.600...
18	4	2017	417788.030...
19	2	2017	291908.009...
20	1	2017	138488.039...
21	10	2016	59090.4800...
22	9	2018	4439.54000...
23	10	2018	589.670000...
24	9	2016	252.24
25	12	2016	19.62

Insight: Compared to 2016 and 2017, first part of the year 2018 has good sales. However, there is a huge drop in sales from September 2018. Since the data is limited, seasonality cannot be determined accurately. However, within the year of 2017, November has seen highest sales.

1. What time do Brazilian customers tend to buy (Dawn, Morning, Afternoon or Night)?

```
with test as
(SELECT
O.order_purchase_timestamp,
extract(hour from O.order_purchase_timestamp)as Ext_time,
P.payment_value
From `Project_Target.orders`as O join `Project_Target.payments`as P on O.or-
der_id = P.order_id)
```

```sql
select
case
when test.Ext_time between 0 and 6 then "Dawn"
when test.Ext_time between 6 and 12 then "Morning"
when test.Ext_time between 13 and 18 then "Afternoon"
else "Night" end as Bin,
count(test.order_purchase_timestamp) as No_of_purchases
from test
group by bin
order by No_of_purchases desc
```

Row	Bin	No_of_purchase
1	Afternoon	39691
2	Night	29739
3	Morning	28950
4	Dawn	5506

Insight: Brazil customers tend to purchase more in the afternoon (13hrs to 18 hrs). Significant orders also are during night and morning respectively.

Evolution of E-commerce orders in the Brazil region:

1. Get month on month orders by states

```sql
---3(1)month on month orders-----DONE

select
C.customer_state,
extract(month from O.order_Purchase_timestamp) as month,
Sum(P.payment_value) as sales
from `Project_Target.Customer` as C join `Project_Target.orders` as O on C.customer_id
 = O.customer_id join `Project_Target.payments` as P
on O.order_id = P.order_id
group by month,C.customer_state
order by month
```

Row	customer_state	month	sales
1	RJ	1	159492.909...
2	SP	1	477677.920...
3	DF	1	22861.6499...
4	RS	1	71027.7999...
5	CE	1	18178.5799...
6	PE	1	20688.5500...
7	PR	1	59385.1000...
8	BA	1	49422.1500...
9	MG	1	153717.829...
10	RN	1	8531.69999...

2. Distribution of customers across the states in Brazil

```
----3(2)distribution of customer across states------DONE

Select
Count (customer_id) as Customer_count,
customer_state
From `Project_Target.Customer`
group by customer_state
order by Customer_count desc
```

Row	Customer_count	customer_state
1	41746	SP
2	12852	RJ
3	11635	MG
4	5466	RS
5	5045	PR
6	3637	SC
7	3380	BA
8	2140	DF
9	2033	ES

Insight: The month over month sales in the states SP , RJ, MJ are higher where as in the state RR , it is low (when ordered by sales). Also, the customer count is highest in the state SP. The states RR, AP and AC has customer count less than 100.

Impact on Economy: Analyze the money movement by e-commerce by looking at order prices, freight and others.

1. Get % increase in cost of orders from 2017 to 2018 (include months between Jan to Aug only) - You can use "payment_value" column in payments table

```
With COO as (SELECT
extract (year from O.order_purchase_timestamp) as Year,
Sum(P.payment_value) as Cost_of_order
from `Project_Target.payments` as P join `Project_Target.orders` as O on P.or-
der_id = O.order_id
where extract (year from O.order_purchase_timestamp) between 2017 and 2018 and ex-
tract (month from O.order_purchase_timestamp) between 1 and 8
group by Year)

Select
COO.year,
COO.Cost_of_order,
(COO.Cost_of_order-lag(COO.Cost_of_order)over (or-
der by COO.year asc))/lag(COO.Cost_of_order)over(order by COO.year asc)*100 as Per-
centage_Increase
from COO
```

Row	year	Cost_of_order	Percentage_Incr
1	2017	3669022.12...	*null*
2	2018	8694733.83...	136.976871...

2. Mean & Sum of price and freight value by customer state

```
Select
C.customer_state,
avg(OI.price) as Mean_of_Price,
Sum(OI.price) as Sum_of_Price,
Avg(OI.freight_value) as Mean_of_freight,
Sum(OI.freight_value) as Sum_of_freight,
from `Project_Target.order_items` as OI
join `Project_Target.orders`as O on OI.order_id = O.order_id join `Project_Target.Cus-
tomer` as C on O.customer_id = C.customer_id
Group by C.customer_state
```

Row	customer_state	Mean_of_Price	Sum_of_Price	Mean_of_freight	Sum_of_freight
1	SP	109.653629...	5202955.05...	15.1472753...	718723.069...
2	RJ	125.117818...	1824092.66...	20.9609239...	305589.310...
3	PR	119.004139...	683083.760...	20.5316515...	117851.680...
4	SC	124.653577...	520553.340...	21.4703687...	89660.2600...
5	DF	125.770548...	302603.939...	21.0413549...	50625.4999...
6	MG	120.748574...	1585308.02...	20.6301668...	270853.460...
7	PA	165.692416...	178947.809...	35.8326851...	38699.3000...
8	BA	134.601208...	511349.990...	26.3639589...	100156.679...
9	GO	126.271731...	294591.949...	22.7668152...	53114.9799...
10	RS	120.337453...	750304.020...	21.7358043...	135522.740...

Analysis on sales, freight and delivery time

1. Calculate days between purchasing, delivering and estimated delivery

```
   ----
   5(1)Calculate days between purchasing, delivering and estimated delivery____DONE
Select
order_purchase_timestamp,
order_status,
DATE_DIFF(order_purchase_timestamp,order_delivered_customer_date , day) as Pur-
chase_to_deliver,
Date_diff(order_purchase_timestamp,order_estimated_delivery_date, day) as Pur-
chase_to_estimateDelivery,
Date_diff(order_delivered_customer_date,order_estimated_delivery_date, day) as deliv-
ery_to_estimateDelivery  ----positive value is late delivery
from `Project_Target.orders`
where order_status = "delivered"
order by order_status
```

Row	order_purchase_timestamp	order_status	Purchase_to_deliver	Purchase_to_estimateDelivery	delivery_to_estimateDelivery
1	2017-04-14 22:06:32 UTC	delivered	-23	-33	-9
2	2017-05-10 14:03:27 UTC	delivered	-12	-7	5
3	2017-04-22 15:50:30 UTC	delivered	-12	-25	-12
4	2017-05-09 17:42:45 UTC	delivered	-7	-8	-1
5	2017-04-26 01:01:39 UTC	delivered	-12	-21	-9
6	2017-05-10 20:47:02 UTC	delivered	-1	-7	-5
7	2017-05-10 15:34:59 UTC	delivered	-6	-7	0
8	2017-04-18 21:20:40 UTC	delivered	-21	-29	-7
9	2017-05-10 22:02:40 UTC	delivered	-7	-7	0
10	2017-04-15 15:37:38 UTC	delivered	-30	-32	-1
11	2017-04-22 13:55:16 UTC	delivered	-20	-25	-5

2. Find time_to_delivery & diff_estimated_delivery. Formula for the same given below:
    1. time_to_delivery = order_purchase_timestamp-order_delivered_customer_date
    2. diff_estimated_delivery = order_estimated_delivery_date-order_delivered_customer_date

```
5(2)Find time_to_delivery & diff_estimated_delivery. Formula for the same given below-
--DONE
Select
order_status,
date_diff(order_purchase_timestamp , order_delivered_customer_date,day) as time_to_del
ivery,
date_diff(order_estimated_delivery_date , order_delivered_customer_date, day) as diff_
estimated_delivery
from `Project_Target.orders`
```

Row	order_status	time_to_delivery	diff_estimated_c
1	canceled	-30	-12
2	canceled	-30	28
3	canceled	-35	16
4	delivered	-30	1
5	delivered	-32	0
6	delivered	-29	1
7	delivered	-43	-4
8	delivered	-40	-4
9	delivered	-37	-1
10	delivered	-33	-5
11	delivered	-38	-6

3. Group data by state, take mean of freight_value, time_to_delivery, diff_estimated_delivery

```
Select
C.customer_state,
avg(OI.freight_value) as Mean_FreightValue,
date_diff(O.order_purchase_timestamp,O.order_delivered_cus-
tomer_date,day) as time_to_delivery,
date_diff(O.order_estimated_delivery_date,O.order_delivered_cus-
tomer_date,day) as diff_estimated_delivery
from `Project_Target.Customer` as C Join `Project_Target.orders` as O on O.cus-
tomer_id = C.customer_id join `Project_Target.order_items`as OI on OI.order_id = O.or-
der_id
Group by C.customer_state, O.order_purchase_timestamp,O.order_delivered_cus-
tomer_date,O.order_estimated_delivery_date
```

Row	customer_state	Mean_FreightVa	time_to_delivery	diff_estimated_c
1	MG	14.1	-30	-12
2	SC	18.51	-30	28
3	RJ	14.11	-35	16
4	RS	19.43	-30	1
5	MT	44.73	-32	0
6	SE	20.8	-29	1
7	CE	30.94	-43	-4
8	SC	19.07	-40	-4
9	PE	35.24	-37	-1
10	RJ	15.56	-33	-5

5. Top 5 states with highest/lowest average freight value - sort in desc/asc limit 5

Lowest_Freight_value

```
select
C.customer_state,
Avg(OI.freight_value) as L_Freight_value
from `Project_Target.order_items` as OI join `Project_Target.orders` as O on OI.or-
der_id = O.order_id join `Project_Target.Customer` as C on O.customer_id=C.cus-
tomer_id
group by C.customer_state
order by L_Freight_value asc
limit 5
```

customer_state	L_Freight_value
SP	15.1472753...
PR	20.5316515...
MG	20.6301668...
RJ	20.9609239...
DF	21.0413549...

Highest_Freight_Value

```
select
C.customer_state,
```

```
Avg(OI.freight_value) as H_Freight_value
from `Project_Target.order_items` as OI join `Project_Target.orders` as O on OI.or-
der_id = O.order_id join `Project_Target.Customer` as C on O.customer_id=C.customer_id
group by C.customer_state
order by H_Freight_value desc
limit 5
```

Row	customer_state	H_Freight_value
1	RR	42.9844230...
2	PB	42.7238039...
3	RO	41.0697122...
4	AC	40.0733695...
5	PI	39.1479704...

6. Top 5 states with highest/lowest average time to delivery

Lowest Avg time

```
with CTE as
(select
C.customer_state,
date_diff(O.order_delivered_customer_date,O.order_pur-
chase_timestamp,day) as time_to_delivery
from `Project_Target.orders` as O join `Project_Target.Customer` as C on O.cus-
tomer_id=C.customer_id
where date_diff(O.order_delivered_customer_date,O.order_pur-
chase_timestamp,day) is not null)

select
CTE.customer_state,
avg(CTE.time_to_delivery) over (order by CTE.time_to_delivery asc) as low-
est_AVG_Time,
from CTE
Group by CTE.time_to_delivery,CTE.customer_state
limit 5
```

Row	customer_state	lowest_AVG_Tim
1	SP	0.0
2	RJ	0.0
3	BA	0.0
4	SP	0.76923076...
5	RJ	0.76923076...

Highest avg time

```sql
with CTE as
(select
C.customer_state,
date_diff(O.order_delivered_customer_date,O.order_pur-
chase_timestamp,day) as time_to_delivery
from `Project_Target.orders` as O join `Project_Target.Customer` as C on O.cus-
tomer_id=C.customer_id
where date_diff(O.order_delivered_customer_date,O.order_pur-
chase_timestamp,day) is not null)  ---removing not delivered orders

select
CTE.customer_state,
avg(CTE.time_to_delivery) over (order by CTE.time_to_delivery desc) as high-
est_AVG_Time,
from CTE
Group by CTE.time_to_delivery,CTE.customer_state
limit 5
```

Row	customer_state	highest_AVG_Tir
1	ES	209.0
2	RJ	208.5
3	PA	204.0
4	PI	200.0
5	SE	200.0

7. Top 5 states where delivery is really fast/ not so fast compared to estimated date

Fast delivery

```sql
SELECT
C.customer_state,
Min(date_diff(order_delivered_customer_date, order_estimated_deliv-
ery_date,day)) as Fast_delivery,
from `Project_Target.orders` as O join `Project_Target.Customer` as C on O.cus-
tomer_id = C.customer_id
group by C.customer_state
Order by Fast_delivery
limit 5
```

Row	customer_state	Fast_delivery
1	SP	-146
2	MA	-139
3	RS	-134
4	RJ	-108
5	MG	-77

Late delivery (tried with CTE)

```
With CTE2 as
(SELECT
C.customer_state,
date_diff(order_delivered_customer_date, order_estimated_delivery_date,day) as Deliv-
ery_delay
from `Project_Target.orders` as O join `Project_Target.Customer` as C on O.cus-
tomer_id = C.customer_id
)

Select
CTE2.customer_state,
max(CTE2.Delivery_delay) as late_delivery
from CTE2
group by CTE2.customer_state
order by late_delivery desc
Limit 5
```

Row	customer_state	late_delivery
1	RJ	188
2	ES	181
3	SP	175
4	SE	166
5	PA	165

Payment type analysis:

1. Month over Month count of orders for different payment types

```
---6(1)Month over Month count of orders for different payment types---DONE
select
Extract(month from O.order_purchase_timestamp) as month,
P.payment_type,
count(O.order_id) as No_of_orders
from `Project_Target.payments` as P join `Project_Target.orders` as O on P.order_id =
O.order_id
group by payment_type,month
order by month
```

Row	month	payment_type	No_of_orders
1	1	credit_card	6103
2	1	UPI	1715
3	1	voucher	477
4	1	debit_card	118
5	2	UPI	1723
6	2	credit_card	6609
7	2	voucher	424
8	2	debit_card	82
9	3	credit_card	7707
10	3	UPI	1942
11	3	debit_card	109

Insight: Highest payment method used is credit card. 3 orders are purchased with not-defined payment type in the month of Aug and Sept.

2. Count of orders based on the no. of payment installments

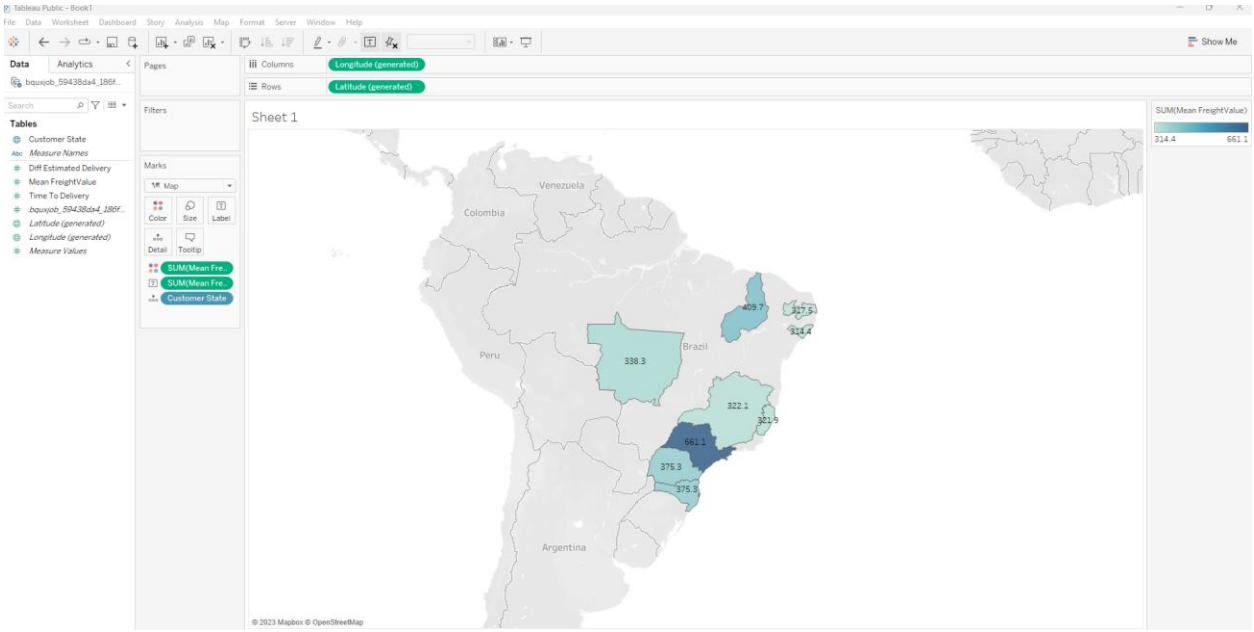
```
6(2)Month over Month count of orders for different payment types----DONE
Select
count(order_id) No_of_orders,
payment_installments
from `Project_Target.payments`
group by payment_installments
```

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

Actionable insights and recommendations:

1) The estimated delivery date has to met. According to the data, there are significant number of late deliveries which might be one of the reason for the drop in sales.
2) The mean freight value is higher for the states (see screenshot below). Hence, we can reduce the freight cost by having a local fulfilment center in the state of SP, in order to increase the willingness of the customer to purchase and cost to customer will reduce. This will help in long run. (visualized in Tableau)

Row	customer_state	Mean_FreightVa	time_to_delivery	diff_estimated_c
1	PI	409.68	-11	19
2	SC	375.28	-10	11
3	PR	375.28	-8	5
4	SP	339.59	-8	15
5	MT	338.3	-31	12
6	MG	322.1	-18	14
7	ES	321.88	-27	1
8	SP	321.46	-7	6
9	PB	317.47	-18	5
10	AL	314.4	-27	1



3) There are 2 unidentified payment methods in the month of Aug and Sept which might require further investigation. If a newer method of payment is to be accepted, it will also increase the willingness to purchase.
4) Since the credit card purchases are the most , offers on credit card purchase will be appealing to the customers.
5) Least number of orders are using full payment where as maximum orders are placed using installment methods of one month.
6) The month over month sales in the states SP , RJ, MJ are higher where as in the state RR , it is low (when ordered by sales). Also, the customer count is highest in the state SP. The states RR, AP and AC has customer count less than 100. Most of the revenue is from SP, RJ and MJ. We need to investigate more on why the count of customers is low in RR, AP and AC.
7) Brazil customers tend to purchase more in the afternoon (13hrs to 18 hrs). Significant orders also are during night and morning respectively.
8) More data is required to understand the seasonal peaks within an year.
9) Compared to 2016 and 2017, first part of the year 2018 has good sales. However, there is a huge drop in sales from September 2018. We need to investigate more on why there is a drop in the sales.