AIT 580 BIG DATA ANALYTICS PROJECT

FINAL PROJECT REPORT ON

INSTACART

BY
THARUN KANAMNENI
G01235383

The dataset I have selected is **Instacart** which is about 800mb of data which includes data of aisles, departments, orders and products. Some of the data types are strings and integer values.

ABOUT: Instacart is US Software firm which delivers groceries which are ordered online and values at $8 billion and operates as same-day grocery delivery and pickup service in the U.S and Canada.[3] [4]

NEED: The data was collected to understand and analyze which products the consumers will purchase again based on past orders, what are some of the most reordered items and some more interesting findings about the orders and the products.[4]

PRIVACY : All the customer data is anonymized for privacy, and as it is posted by Instacart itself, it doesn't have any privacy issues and ethical issues.[4]

A structured collection of data representing customer orders over time is the dataset and it was posted by the Instacart as open source of 3 million grocery orders from more than 200,000 orders.[3][4] It also includes the data of the week and hour of the day order was placed.

TYPES OF DATATYPES :There are various attributes present in the dataset, they are order hour of the day, order day of the week, eval_set, days since prior order, product_id, product_type, reordered.
order_id, product_id, order_day_of_week, days_since_prior_order **:Numeric**
product_name, department_name : **Nominal**
eval_set :**Categorical**

The dataset of orders has about 3.4 million rows, and order_products has about 30 million rows.

Some of the questions i got to solve are :

1) what are the products that are sold the most?

2) what are the most reordered items?

3) How many orders are placed in the morning around 10am to 12pm?

Some of the software resources needed to study this data are :

- R analytics for data exploration and visualization
- Python analytics for data exploration
- SQL Work Bench

The Hardware I am using :

Processor: 2.6 GHz 6-Core Intel Core i7, 9th Gen

Memory : 16 GB 2400 MHz DDR4

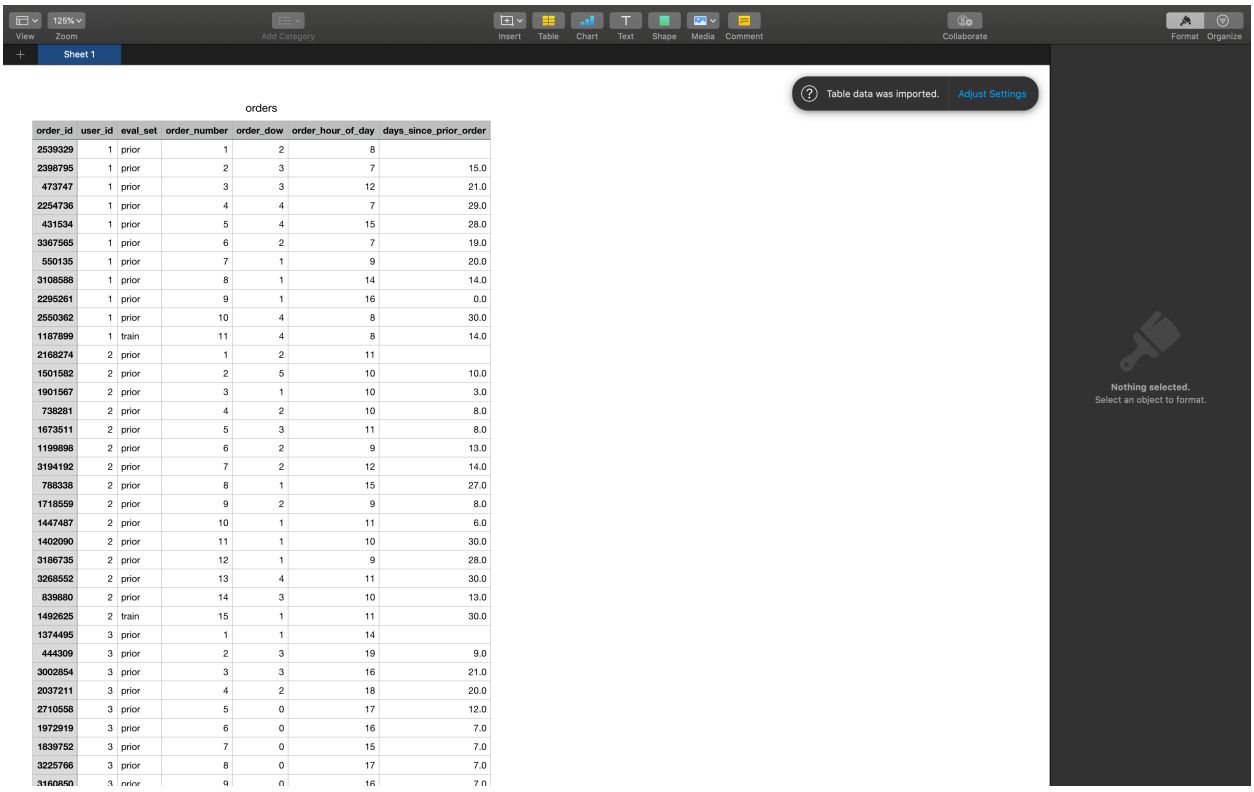
I have previously used ggplot2, SQL, R and Python in my assignments and did various visualizations, so in order to improve more and work with more large data, I take this as an opportunity to learn more and improve my skills.

Visualizations have been done using python, R studio, Tableau.

Visualizations are done for few major attributes where we can draw many insights and which help to answer our driving questions.

The dataset Orders is :

Similarly, I have used various datasets of Instacart like products, aisles, departments, order_products_prior, order_products_train to draw conclusions and analyze as much as possible



orders

order_id	user_id	eval_set	order_number	order_dow	order_hour_of_day	days_since_prior_order
2539329	1	prior	1	2	8	
2398795	1	prior	2	3	7	15.0
473747	1	prior	3	3	12	21.0
2254736	1	prior	4	4	7	29.0
431534	1	prior	5	4	15	28.0
3367565	1	prior	6	2	7	19.0
550135	1	prior	7	1	9	20.0
3108588	1	prior	8	1	14	14.0
2295261	1	prior	9	1	16	0.0
2550362	1	prior	10	4	8	30.0
1187899	1	train	11	4	8	14.0
2168274	2	prior	1	2	11	
1501582	2	prior	2	5	10	10.0
1901567	2	prior	3	1	10	3.0
738281	2	prior	4	2	10	8.0
1673511	2	prior	5	3	11	8.0
1199898	2	prior	6	2	9	13.0
3194192	2	prior	7	2	12	14.0
788338	2	prior	8	1	15	27.0
1718559	2	prior	9	2	9	8.0
1447487	2	prior	10	1	11	6.0
1402090	2	prior	11	1	10	30.0
3186735	2	prior	12	1	9	28.0
3268552	2	prior	13	4	11	30.0
839880	2	prior	14	3	10	13.0
1492625	2	train	15	1	11	30.0
1374495	3	prior	1	1	14	
444309	3	prior	2	3	19	9.0
3002854	3	prior	3	3	16	21.0
2037211	3	prior	4	2	18	20.0
2710558	3	prior	5	0	17	12.0
1972919	3	prior	6	0	16	7.0
1839752	3	prior	7	0	15	7.0
3225766	3	prior	8	0	17	7.0
3160850	3	prior	9	0	16	7.0

1. To view the Orders, Prior Dataset

```
>>> orders.head()
...
   order_id  user_id  ... order_hour_of_day  days_since_prior_order
0   2539329        1  ...                 8                     NaN
1   2398795        1  ...                 7                    15.0
2    473747        1  ...                12                    21.0
3   2254736        1  ...                 7                    29.0
4    431534        1  ...                15                    28.0

[5 rows x 7 columns]
```

```
>>> prior.head()
   order_id  product_id  add_to_cart_order  reordered
0         2       33120                  1          1
1         2       28985                  2          1
2         2        9327                  3          0
3         2       45918                  4          1
4         2       30035                  5          0
```

To Print number of Aisles.

```
>>> print(aisles.shape)
(134, 2)
```

```
>>> mt.head(20)
    order_id  product_id  ...  days_since_prior_order  aisle
0          2       33120  ...                     8.0   eggs
1         26       33120  ...                     7.0   eggs
2        120       33120  ...                    10.0   eggs
3        327       33120  ...                     8.0   eggs
4        390       33120  ...                     9.0   eggs
5        537       33120  ...                     3.0   eggs
6        582       33120  ...                    10.0   eggs
7        608       33120  ...                    12.0   eggs
8        623       33120  ...                     3.0   eggs
9        689       33120  ...                     3.0   eggs
10       689       35921  ...                     3.0   eggs
11       726       33120  ...                     7.0   eggs
12       726       32655  ...                     7.0   eggs
13       771       33120  ...                     7.0   eggs
14       800       33120  ...                     3.0   eggs
15       901       33120  ...                     8.0   eggs
16      1005       33120  ...                     6.0   eggs
17      1097       33120  ...                     8.0   eggs
18      1193       33120  ...                     9.0   eggs
19      1300       33120  ...                    20.0   eggs

[20 rows x 14 columns]
```

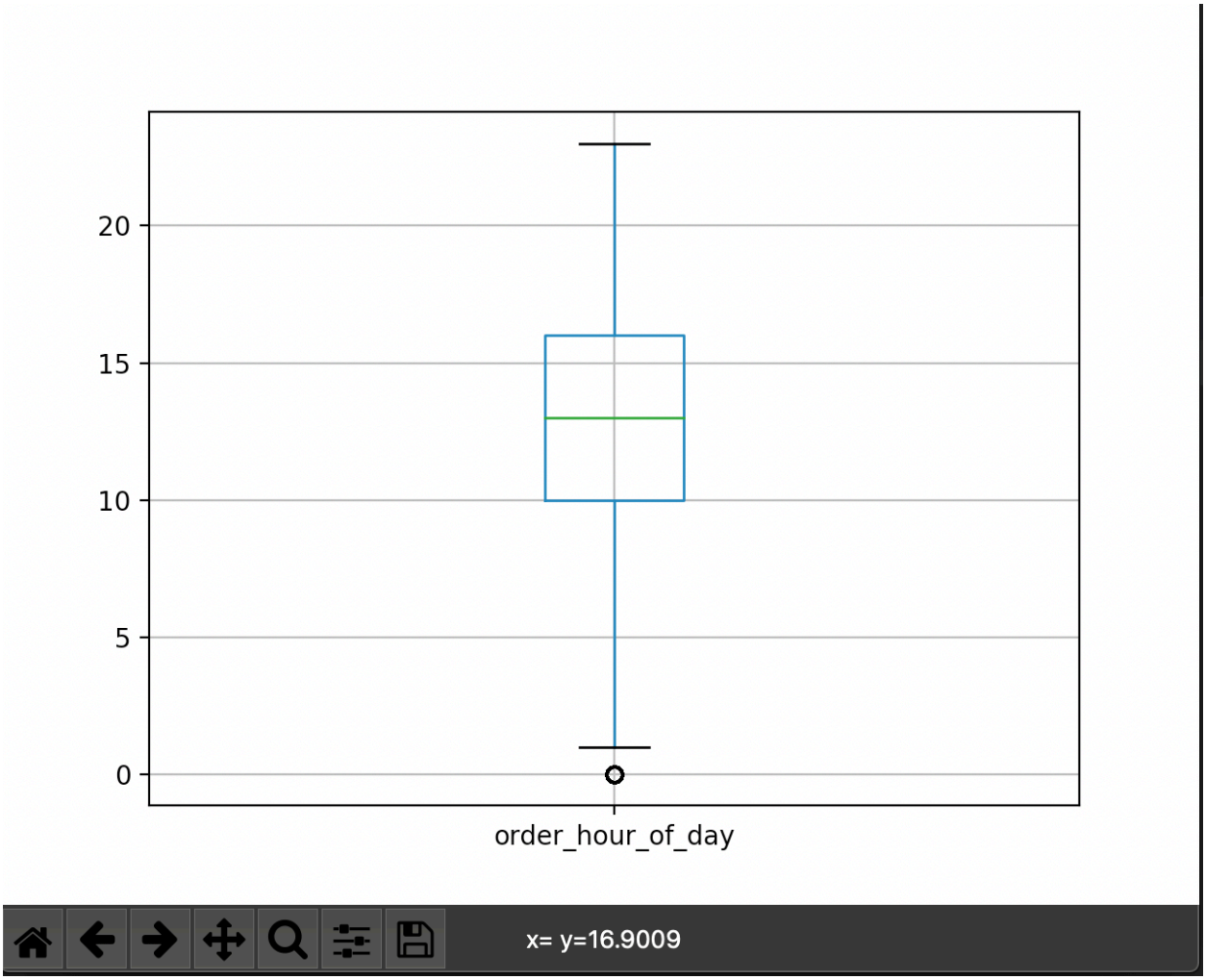
Count of Fresh fruits, fresh vegetables and the best selling goods.

```
>>> mt['aisle'].value_counts()[0:10]
fresh fruits                     3642188
fresh vegetables                 3418021
packaged vegetables fruits       1765313
yogurt                           1452343
packaged cheese                   979763
milk                              891015
water seltzer sparkling water     841533
chips pretzels                    722470
soy lactosefree                   638253
bread                             584834
Name: aisle, dtype: int64

>>>
```

Most number of reorders:

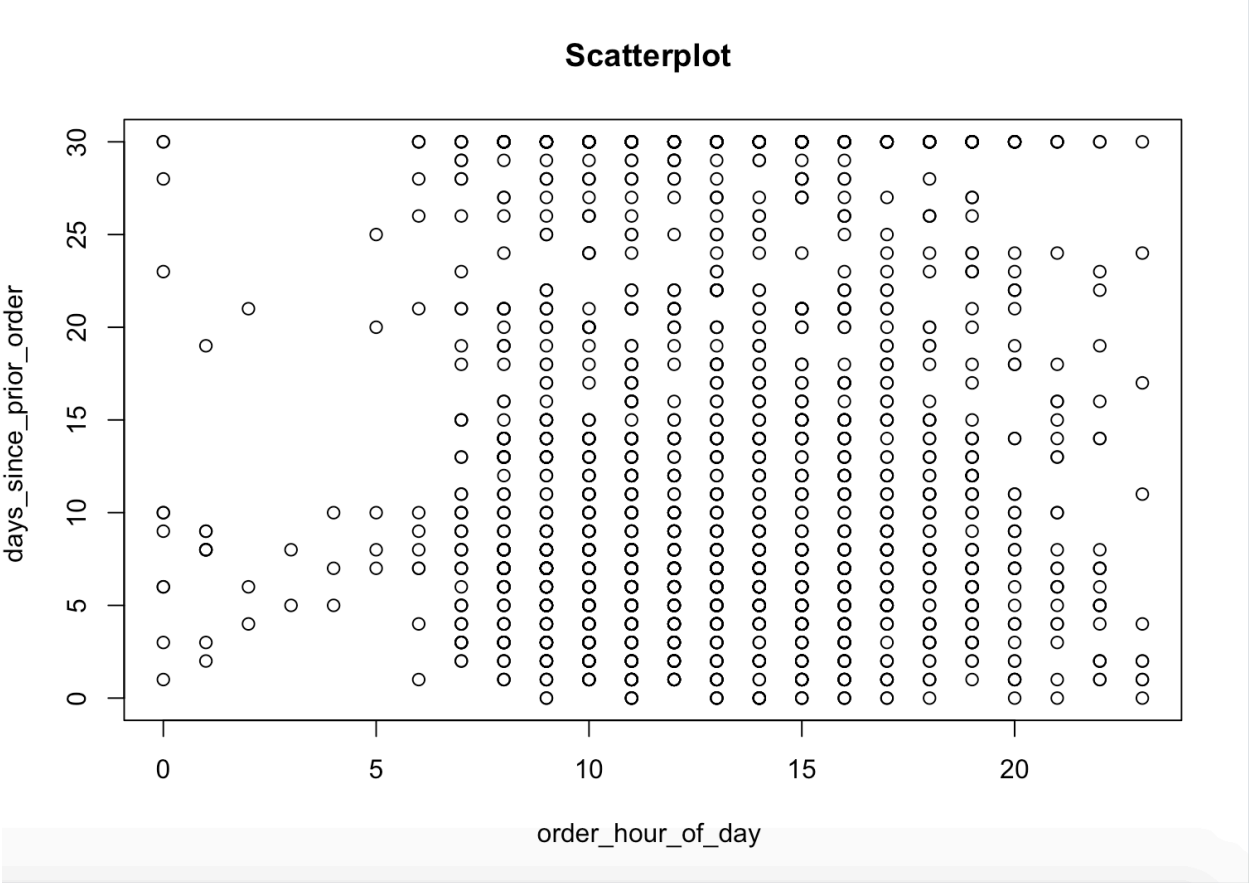
	product_id	Total_reorders	product_name
24848	24852	472565	Banana
13172	13176	379450	Bag of Organic Bananas
21133	21137	264683	Organic Strawberries
21899	21903	241921	Organic Baby Spinach
47198	47209	213584	Organic Hass Avocado
47755	47766	176815	Organic Avocado
47615	47626	152657	Large Lemon
16793	16797	142951	Strawberries
26204	26209	140627	Limes
27839	27845	137905	Organic Whole Milk

1. Total Number of Orders in 24 Hour scale



The above visualization is done in Pycharm using Python to understand that the maximum orders that were done are between 10:00 and 16:00 .

2. This Visualization is a scatter plot made using R studio.

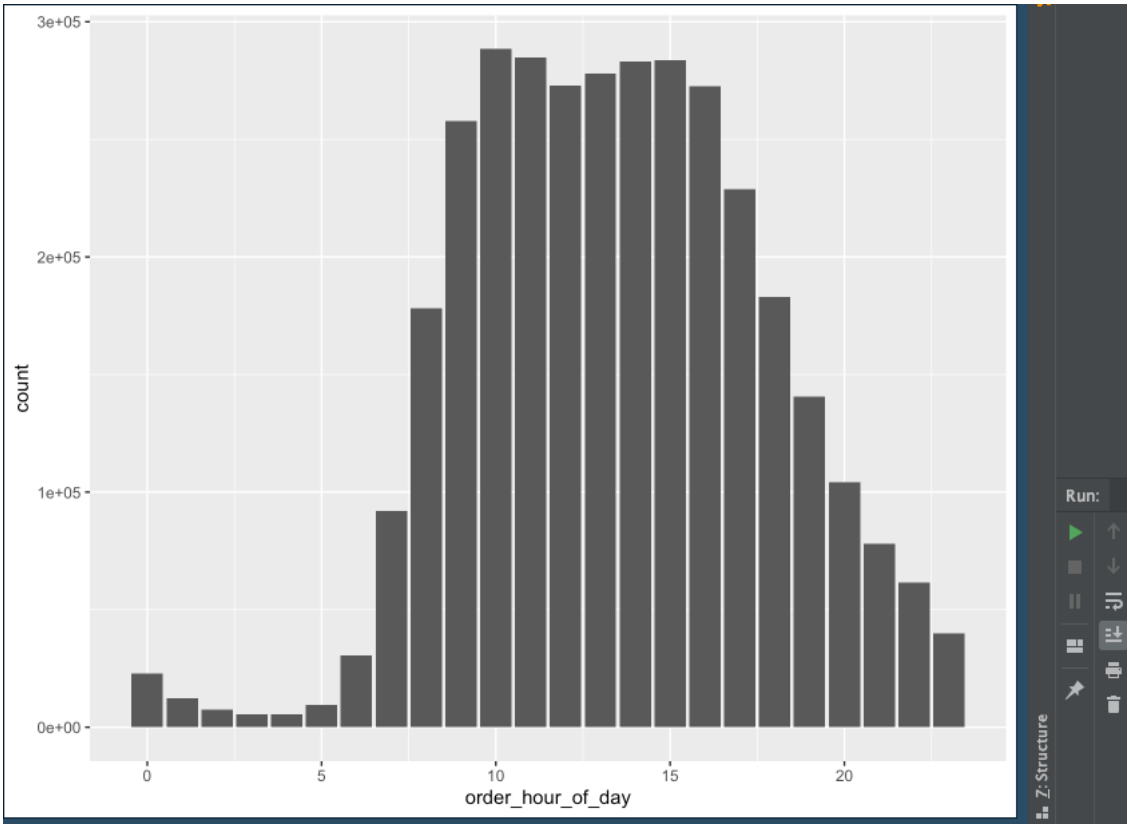


I used R to visualize this,

As I don’t have many attributes which relates to a good scatter plot, I don’t think this helps me to get information for me as there are no attributes where I can draw a scatter plot to get insights.

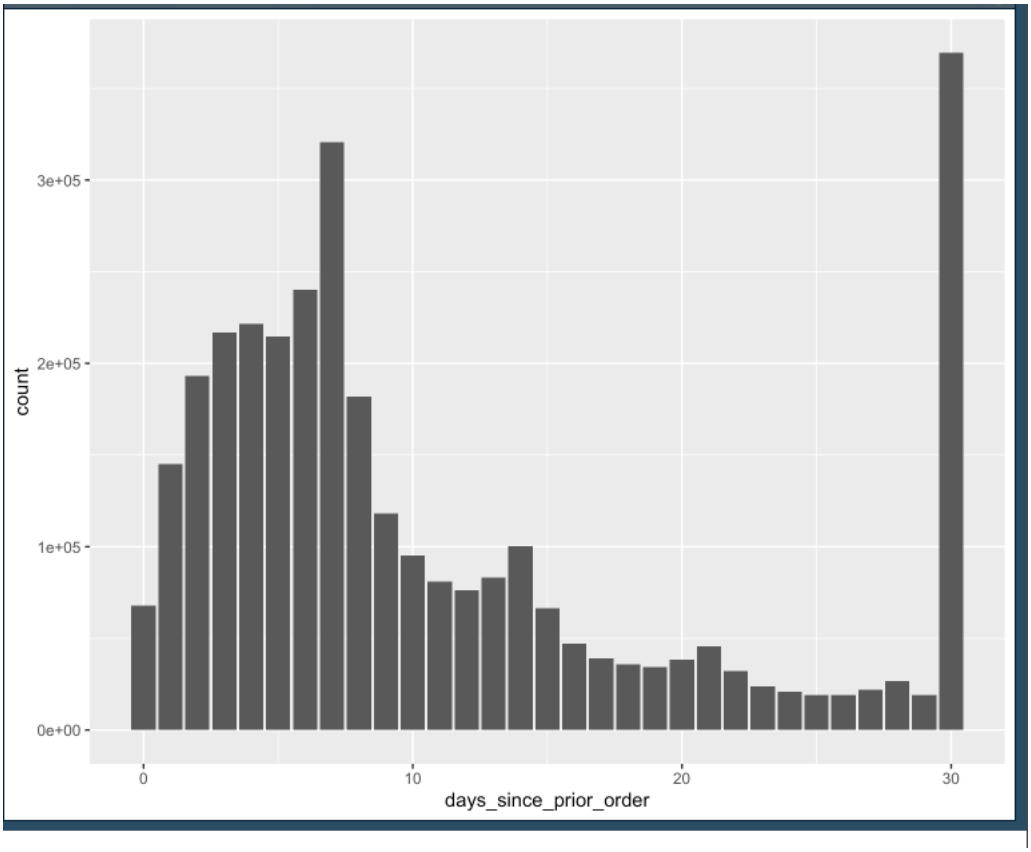
3. Total number of orders in according with the order hour of the day.

The below visualization helps me in finding out the number of items ordered in 24 hour scale where the highest number of items that were sold were between 9:00 to 15:00.



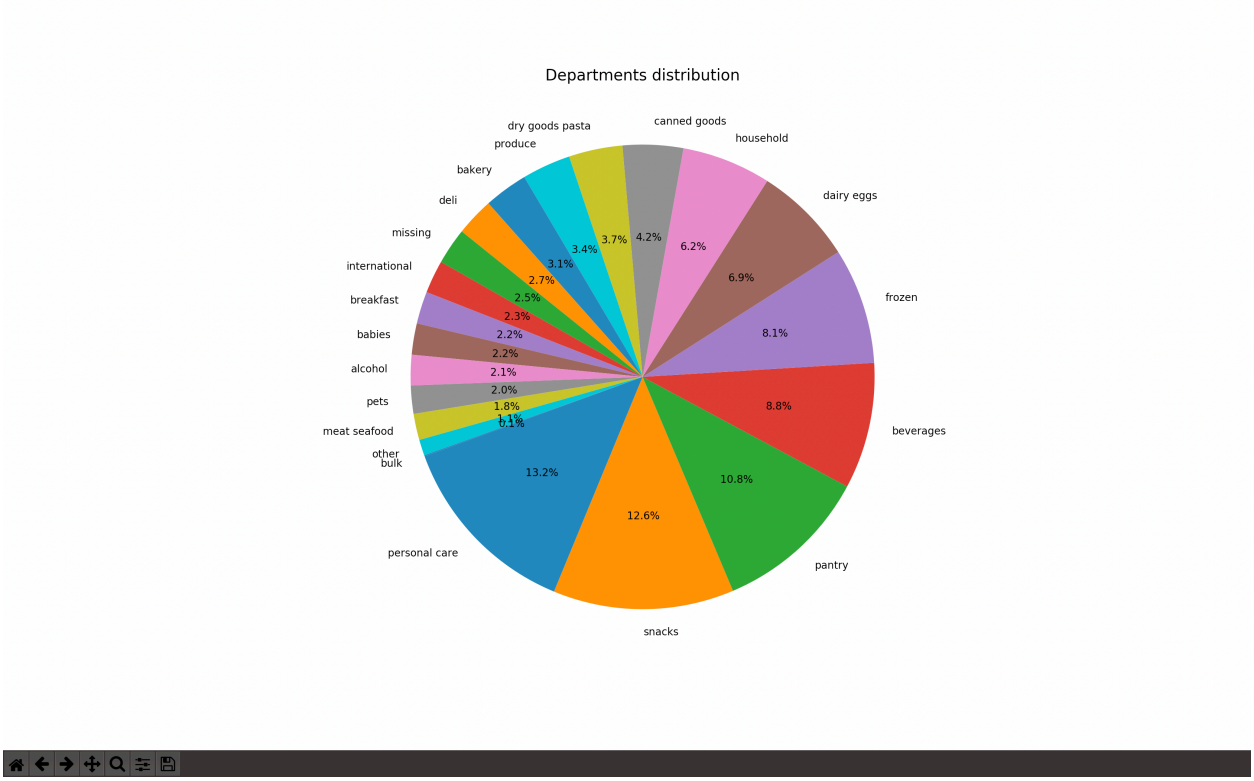
4. Total Number of orders done with respect to days since prior order

The below visualization describes the customers who order again since their prior order. The highest number of customers who order after 30 days are more.

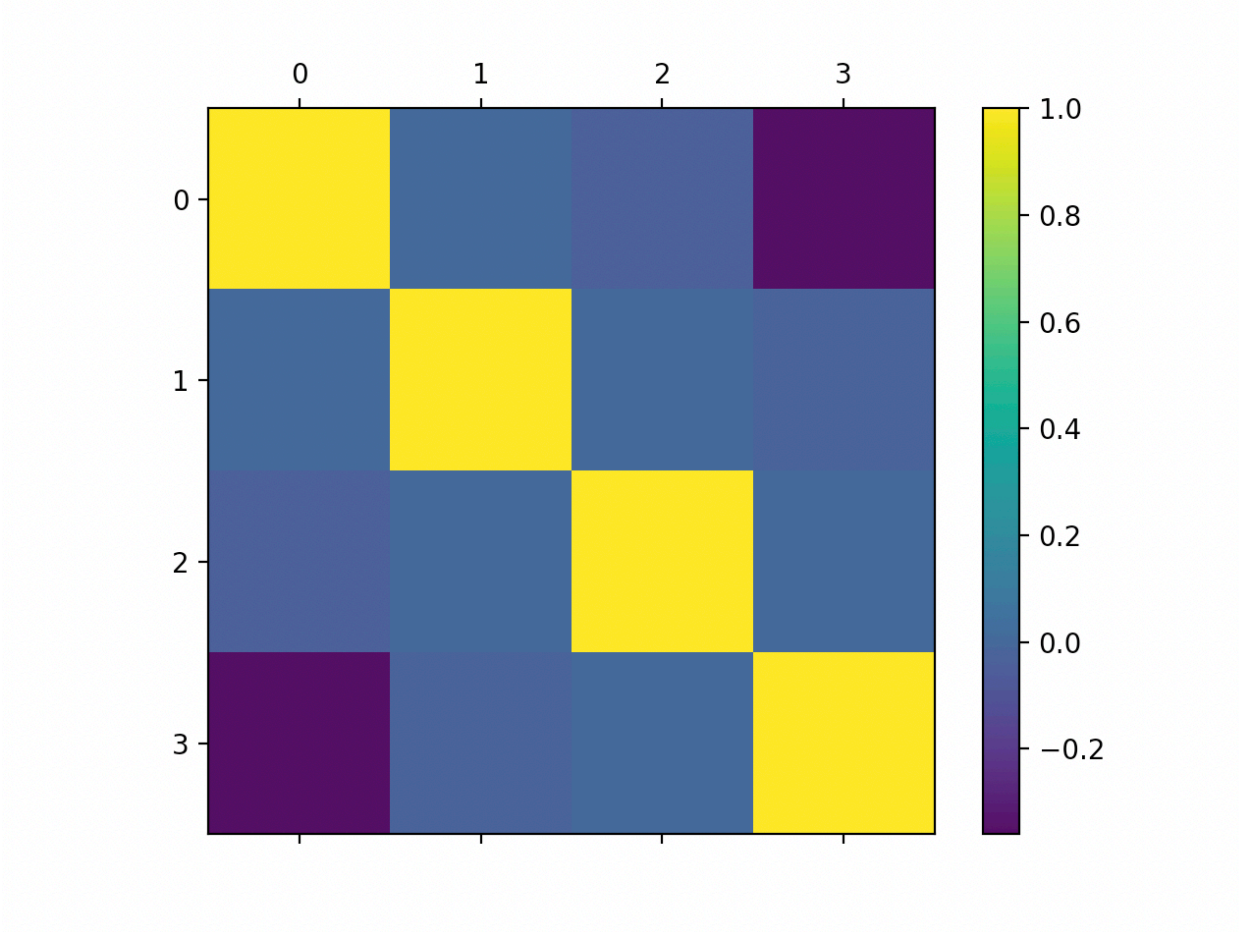


5. Percentage of items sold in each department

The below visualization helps us to understand which goods from the department are sold the most and how much percentage of goods are sold in each department.



6. Correlation between order_number, order_dow, order_hour_of_day, days_since_prior_order



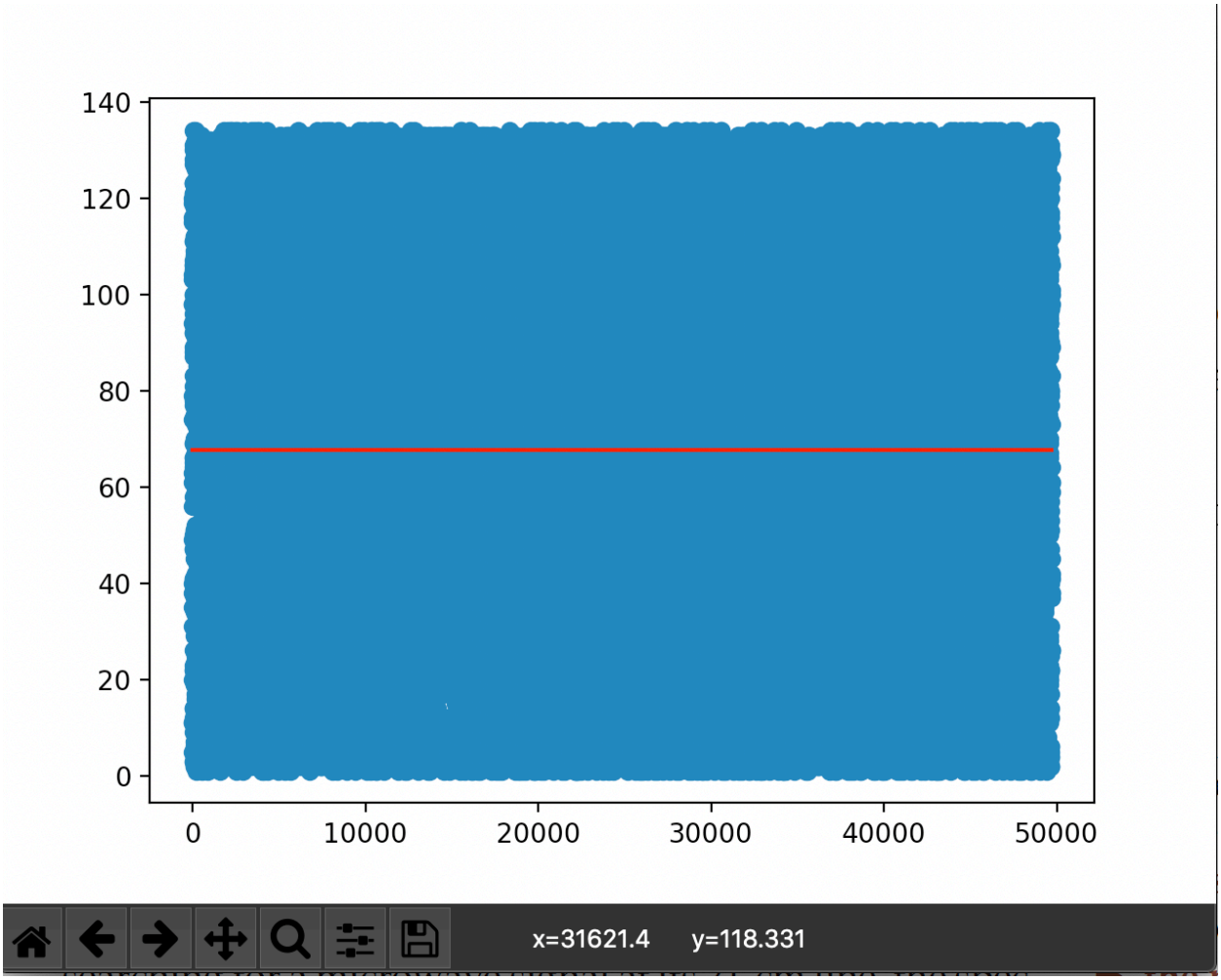
Where order_number, order_dow, order_hour_of_day, days_since_prior_order is 0,1,2,3 respectively.

By observing the Graph, we can observe that there is no **correlation.**

Null Hypothesis

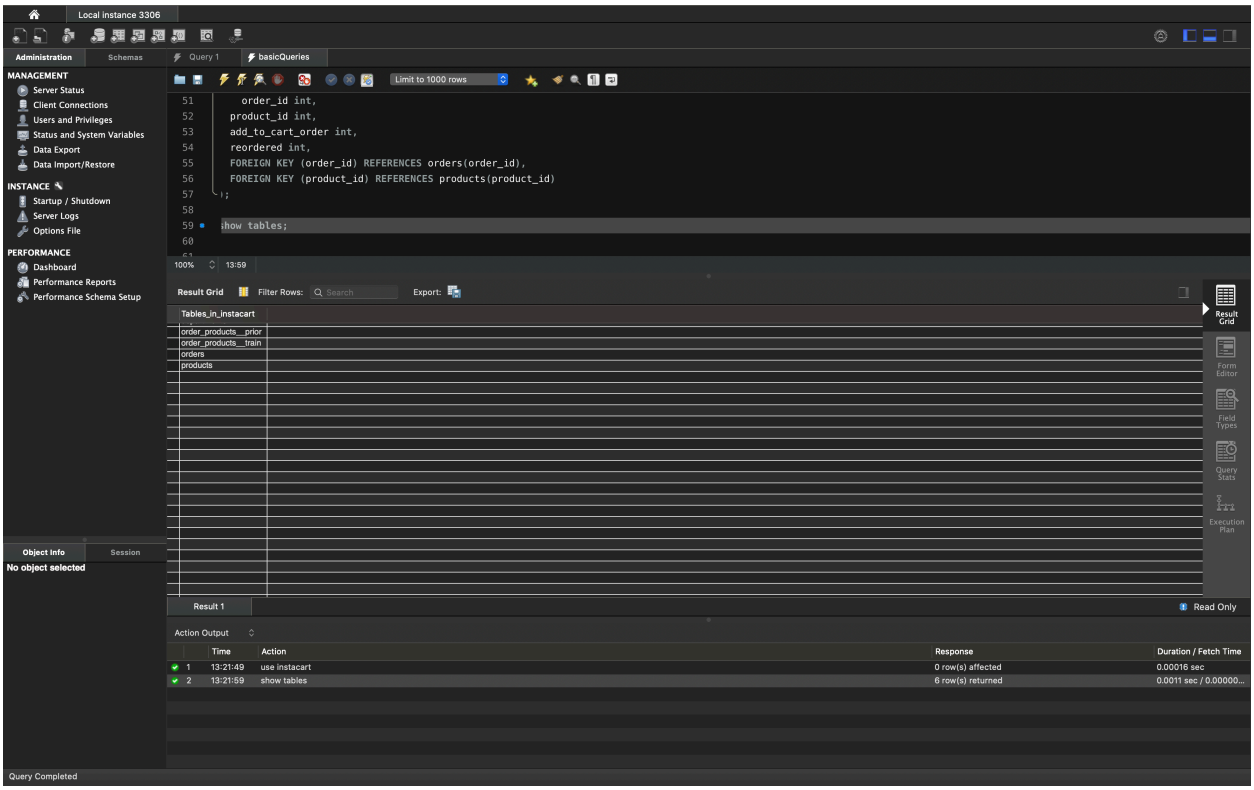
```
/Users/tharun/PycharmProjects/project01/venv/bin/python /Users/tharun/PycharmProjects/project01/tharun/script_01.py
13.45201534134074
p - values 0.0
 we are rejecting null hypothesis

Process finished with exit code 0
```

Regression Analysis

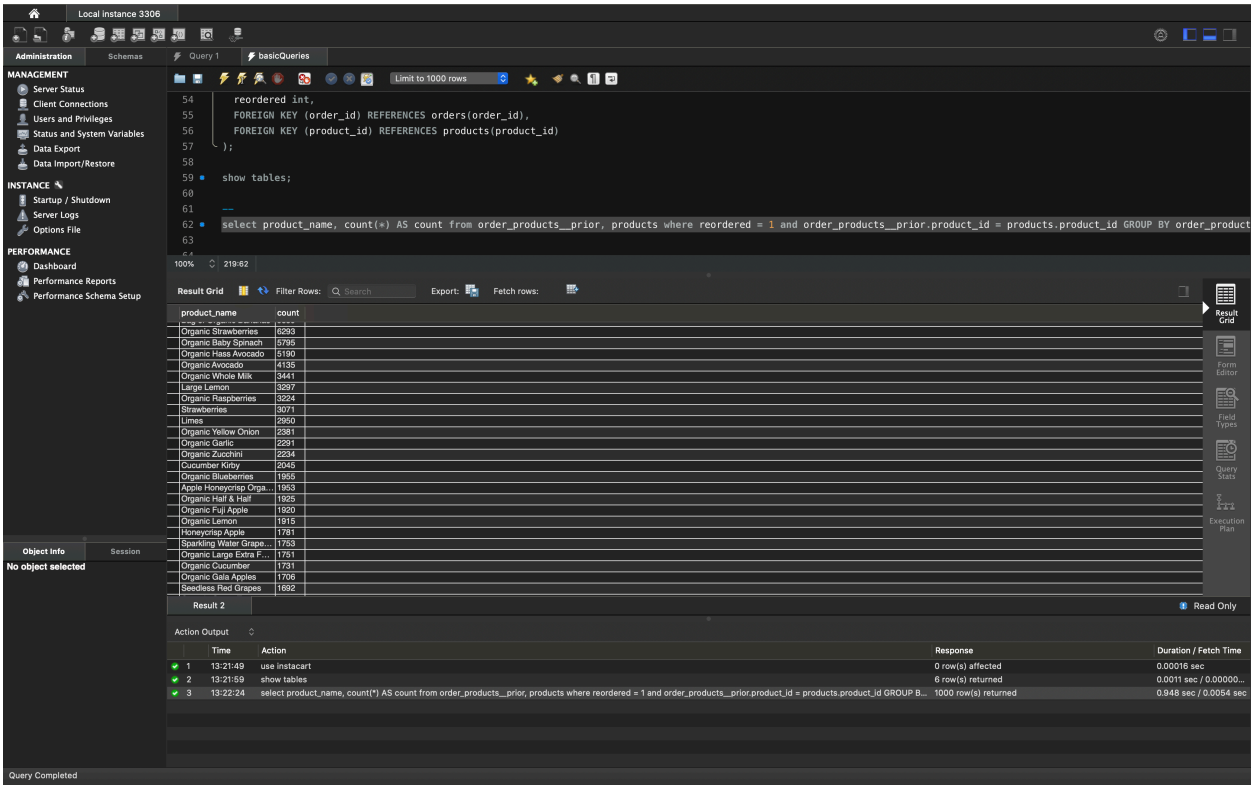


SQL Schema:

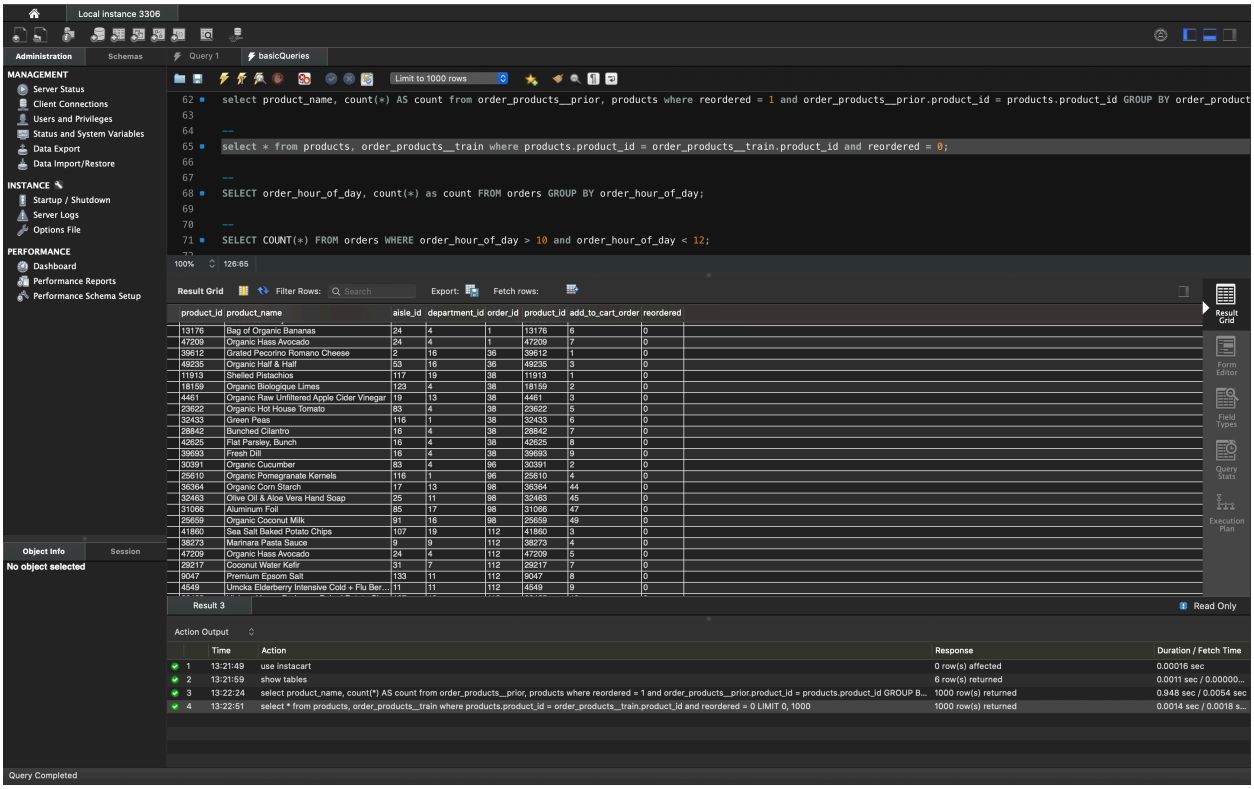
Show Tables:



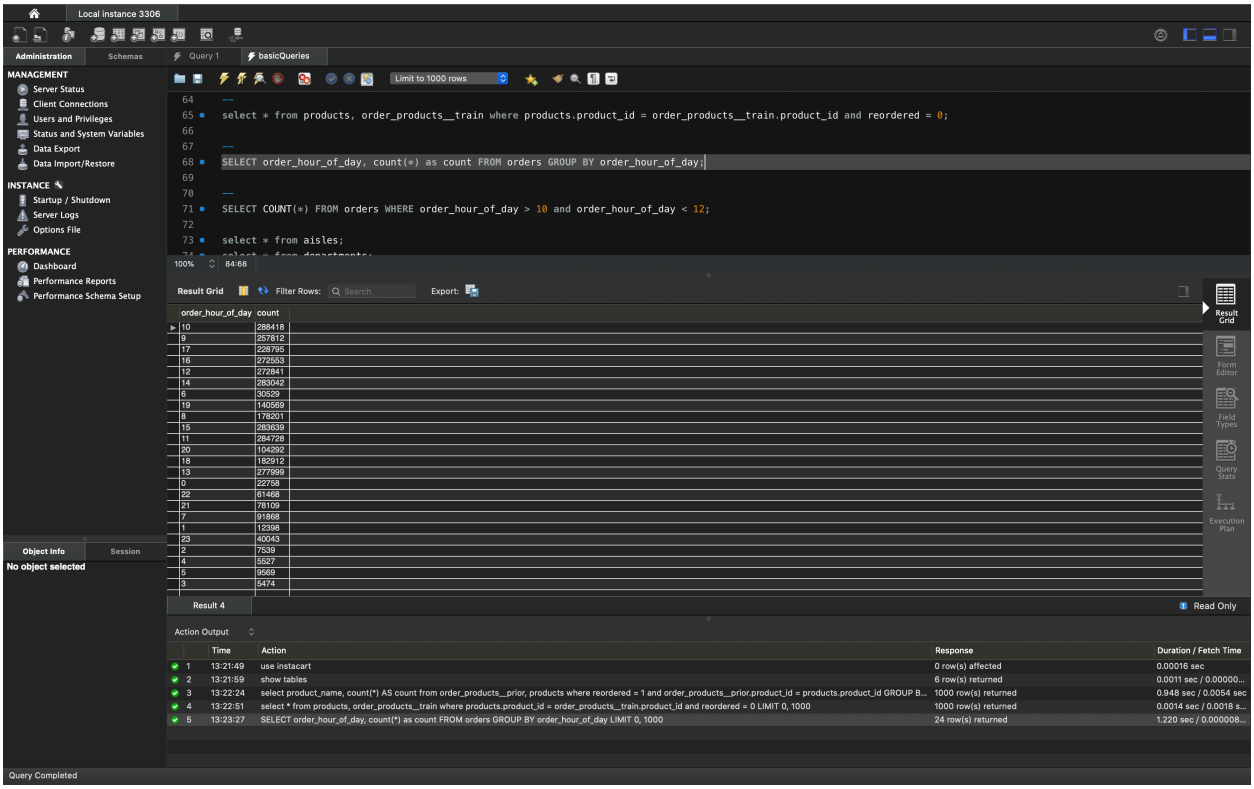
Products Names and Count Where Reordered=1



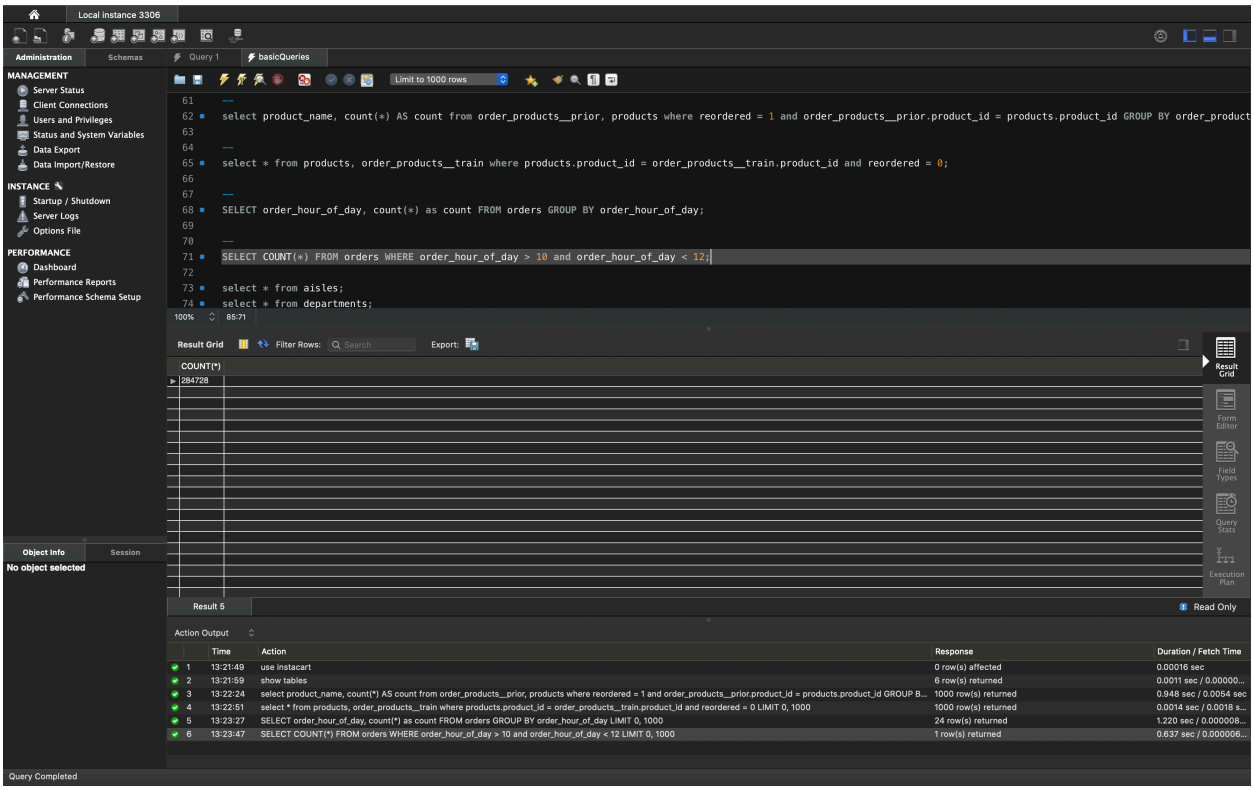
Product Names and product name with aisle and department ID where reordered =0



Orders are grouped by order hour of the day



Order Count done between 10:00 in the morning and 12:00



CONCLUSION:

From above visualizations, we can observe that the most products sold, and interesting insights, which are drawn from graphs. It is good to understand which are the products the customer need more, and what times do the customers order more so that more employees can be employed in that stipulated time. It also shows how many customers reordered the items at least once and how many customer didn't reorder.

REFERENCES:

References:

[1] “The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 [Accessed: 21-Nov-2019]

[2] “Instacart Market Basket Analysis.” [Online]. Available: https://kaggle.com/c/instacart-market-basket-analysis. [Accessed: 21-Nov-2019]

[3] “What Is Instacart? A Guide on How to Use the Grocery Delivery App,” Digital Trends, 19-Sep-2018. [Online]. Available: https://www.digitaltrends.com/home/what-is-instacart/. [Accessed: 21-Nov-2019]

[4] J. Stanley, “3 Million Instacart Orders, Open Sourced,” *Medium*, 04-May-2017. [Online]. Available: https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2. [Accessed: 21-Nov-2019]