MIS 587-INSTACART ANALYSIS PROJECT REPORT

Group 1

(Parth Dattani, Nikhil Kumar, Varun Kapuria, Gurleen Kohli)

Table of Contents

About the company

Instacart is an American company which operates as a same-day grocery delivery and pick-up service in the United States and Canada. The company's platform helps connect customers with personal shoppers who pick and pack items from local grocery stores, and then deliver them to the customer's doorstep or allow for pick-up. Instacart works on B2C model in which business sells goods or services to consumers. Customers use the Instacart online platform to browse products and grocery stores and place orders. It then facilitates the process of shopping and delivering the requested items to the consumers with the help of shoppers.

Problem Statement

We plan to focus on two problems, through business intelligence, which would enable Instacart to run the business more efficiently and profitably.

a) Increasing revenue through buyer behavior analysis and corresponding marketing strategies by analyzing past purchase history to identify patterns and purchases. This data would help in understanding consumer buying behavior, which can be used to tailor promotions, optimize product placement, and develop personalized recommendations. Additionally, understanding the price elasticity of products through price analysis helps in setting strategic pricing that can boost sales without eroding profit margins.
b) Maintenance of inventory by analyzing which products are more reordered. By analyzing purchase patterns, frequency of sales by product, and aisle traffic (which aisles products are picked from), a business can optimize stock levels, reduce inventory costs, and minimize stockouts or overstock situations. For instance, if certain products are consistently purchased together, they can be stocked closer to one another to speed up the picking process. Similarly, if some items sell more on specific days of the week or times of the day, inventory restocking schedules can be adjusted accordingly.

Data Description

The source of the dataset is Kaggle (Instacart, 2023) which describes the customer orders over time. There are 20,000 records with 15 attributes. The data contains departments, aisle, products, orders, and users' data.

Departments have the information of the categorical data such as produce, snacks, beverages, alcohol, breakfast, pantry, personal care, seafood and many more. Products are more detailed which tell you the items that are present in retail stores, like in produce there are going to be organic items such as bananas, strawberries, vegetables. Orders gives information about the orders placed and more attributes like the hour of the day at which the order was ordered. It also gives the information about the user who ordered using the user ID.

Data warehouse Design

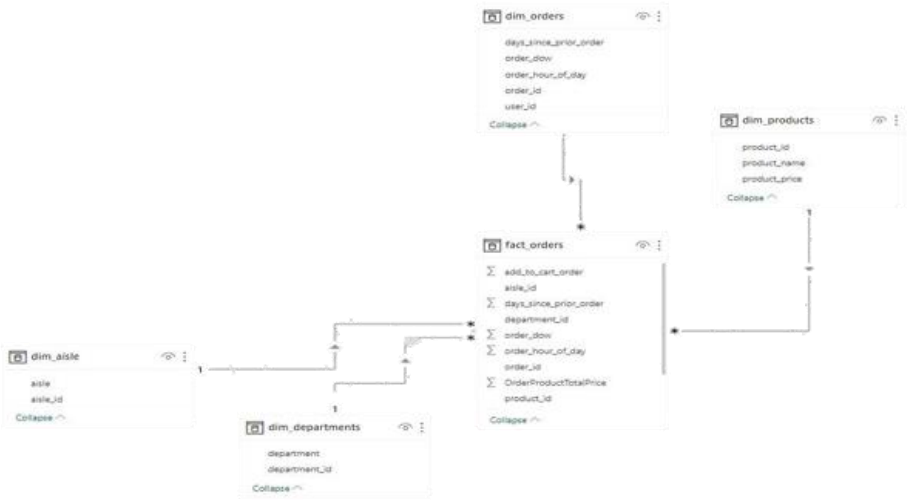
The data warehouse consists of 4 dimensions and 1 fact.

a) Dim_orders(order_id(PK),user_id,days_since_prior_order,order_dow,order_hour_of_day)
b) Dim_products (product_id(PK),product_name,product_price)
c) Dim_aisle(aisle_id(PK),aisle)
d) Dim_departments(department_id(PK),department)

e) Fact_orders(aisle_id(FK),order_id(FK),product_id(FK),department_id(FK),add_to_cart_order, days_since_prior_order,order_dow,order_hour_of_day,OrderProductTotalPrice)

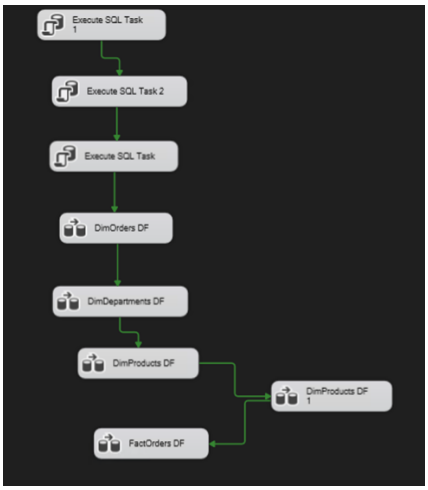
OrderProductTotalPrice is a derived column which is the product of product price and add to cart order.

All the dimensions hold one -to-many relationship with the fact orders.

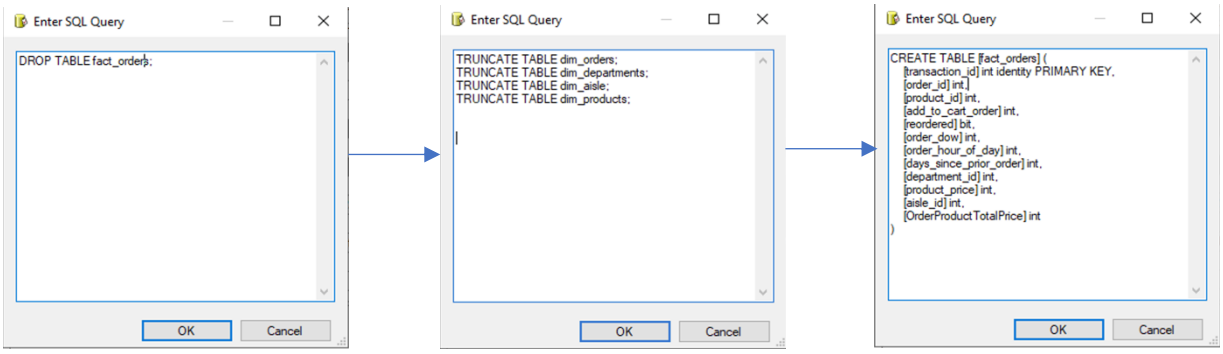


ETL Processing

Control Flow:



SQL Tasks:



SQL TASK 1 SQL TASK 2 SQL TASK 3

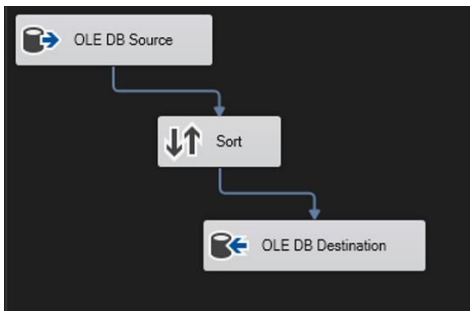
To establish a consistent and clean starting point, we truncate dimension tables before data loading, preventing conflicts and enhancing performance. However, due to foreign key references from the fact table, we address this by dropping the fact table in the first SQL task, truncating dimension tables in the second, and recreating the fact table in the third task.

Data Flow for Dimension Tables:

Individual Dimension Tables

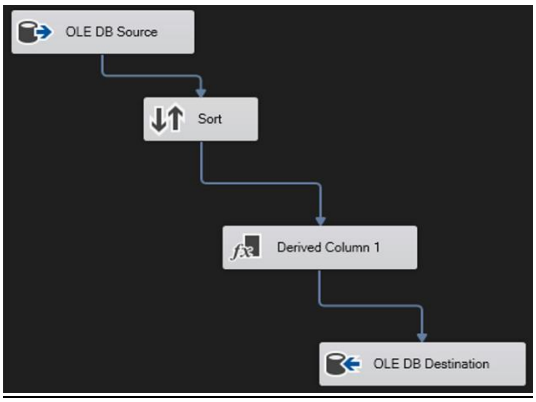
1. Orders Dimension: This table serves as the identity card for each order, recording its unique order ID, the user who placed it, the day of the week it was placed (order_dow), the hour of the day (order_hour_of_day), and even how many days have passed since the previous order (day_since_prior_order). Each order, like a unique brick, finds its place within this dimension.
2. Department Dimension: This one acts as a directory for our departments, mapping each department name to its corresponding ID. Think of it as the building's blueprint, clearly labeling each room (department) by its assigned number.
3. Products Dimension: Each product, from the smallest screw to the grandest chandelier, has its own story told in this table. It captures the unique product ID, the name that resonates with customers, and even the price that makes it shine. It's like a detailed catalog for our building's inventory.
4. Aisle Dimension: This table acts as a map, guiding customers (or analysts) to the exact location of each product. It pairs each aisle ID with its corresponding aisle name, like signposts leading you to the treasure you seek.

To construct each dimension table, we follow a meticulous three-step process:



1. Data Extraction: We leverage the power of OLE DB source tasks to extract relevant data from our master database table. This targeted selection ensures each dimension table contains only the necessary columns, optimizing storage and processing efficiency.
2. Data Sorting: Following extraction, each data stream undergoes a rigorous sorting process. This ensures consistent ordering within each dimension, facilitating efficient joins and aggregations later in the data analysis pipeline.
3. Data Destination: Finally, we utilize OLE DB destination tasks to precisely place each completed dimension table within the database. This organized placement allows for optimal data retrieval and utilization throughout the analysis process.

Data Flow for Fact Table:



In the fact table creation process, we adhere to four steps. Three of these steps (OLE DB Source, Sort, and OLE DB Destination) mirror the procedures employed in Dimension Table creation. The additional step introduced here involves the incorporation of a Derived Column.

Derived Column Name	Derived Column	Expression	Data Type	Le
OrderProductTotalPrice	<add as new column>	product_price * add_to_cart_order	four-byte signed integ...	

Derived Column: This step is essential for deriving a new column from the dataset. Specifically, in our dataset, we have information on the total quantity purchased and the price per unit. The Derived Column process is utilized to calculate the total price of a product for an order, achieved through the multiplication of the quantity purchased and the price per unit.

Exploratory Data Analysis

We did some exploratory data analysis using SQL to see the behavior of customers and identify the scope of improvements in our existing business model. In the first analysis, we checked the number of orders placed by each department to check which department has the maximum number of orders and which department has the minimum number of orders.

```
select count(order_id) as "Count of Orders",d.department_id,d.department from dbo.dim_departments d
JOIN fact_orders fo ON fo.department_id=d.department_id
group by d.department_id,d.department
order by count(order_id) DESC;
```

In this query we took the count of order IDs and grouped it on the department. To keep the highest performing department on top, we sorted the count in a descending order. Below is the result of the query

	Count of Orders	department_id	department
1	29940	4	produce
2	15649	16	dairy eggs
3	8559	19	snacks
4	8169	7	beverages
5	7368	1	frozen

From the above query, we see that the produce has the maximum number of orders followed by dairy eggs. The third highest number of orders are from the department snacks which is followed by beverages. The least number of orders are from the department frozen.
The next analysis shows the users which have not ordered for 30 days. We did this with RANK clause but we could also use the where clause and see how many users are there in the 30 days bracket.

```
WITH CTE AS(select RANK() OVER(order by dim_orders.days_since_prior_order DESC) rank_users,
user_id,
dim_orders.days_since_prior_order
from fact_orders JOIN dim_orders
ON fact_orders.order_id=dim_orders.order_id
group by dim_orders.days_since_prior_order,user_id,product_id)

select * from cte where rank_users=1;
```

The results of this query show that out of 90,000 users there are around 10,000 users which have not ordered for 30 days. Below is the result of the query

	rank_users	user_id	days_since_prior_order
276	1	6101	30
277	1	6228	30
278	1	6613	30
279	1	6654	30
280	1	6654	30
281	1	6688	30
282	1	6709	30
283	1	6832	30
284	1	6832	30
285	1	6877	30
286	1	6877	30

The third analysis is to analyze the most reordered products where we have taken the count of reordered and grouped it on product

```
SELECT
    COUNT(f.reordered) AS reorder_count,
    f.product_id,
    p.product_name
FROM
    fact_orders AS f
JOIN
    dim_products AS p ON f.product_id = p.product_id
GROUP BY
    f.product_id, p.product_name
ORDER BY
    reorder_count DESC;
```

The result shows the most reordered product is banana followed by the bag of organic bananas. The result set is given below.

	reorder_count	product_id	product_name
1	1337	24852	Banana
2	1129	13176	Bag of Organic Bananas
3	776	21137	Organic Strawberries
4	705	21903	Organic Baby Spinach
5	594	47626	Large Lemon

The fourth analysis we did was using the hour of the day in which we analyze how the sales differ depending on what hour of the day it is. We took the count of product Ids from fact orders and grouped it on the hour of the day.

```
select count(dim_products.product_id) as count_product_id,order_hour_of_day from fact_orders
JOIN dim_products on fact_orders.product_id=dim_products.product_id
group by order_hour_of_day
order by count_product_id DESC ;
```

The result shows that around 11 AM to 2 PM, the highest selling of products takes place with the maximum at 2 PM. Below is the result.

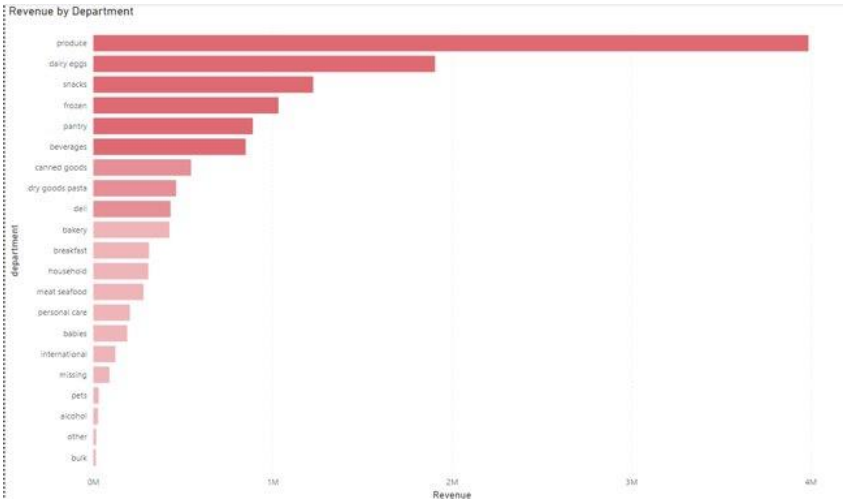
	count_product_id	order_hour_of_day
1	8340	14
2	8260	11
3	8238	10
4	8100	13
5	8074	16
6	8070	12
7	8055	15
8	7620	9

Data Visualizations

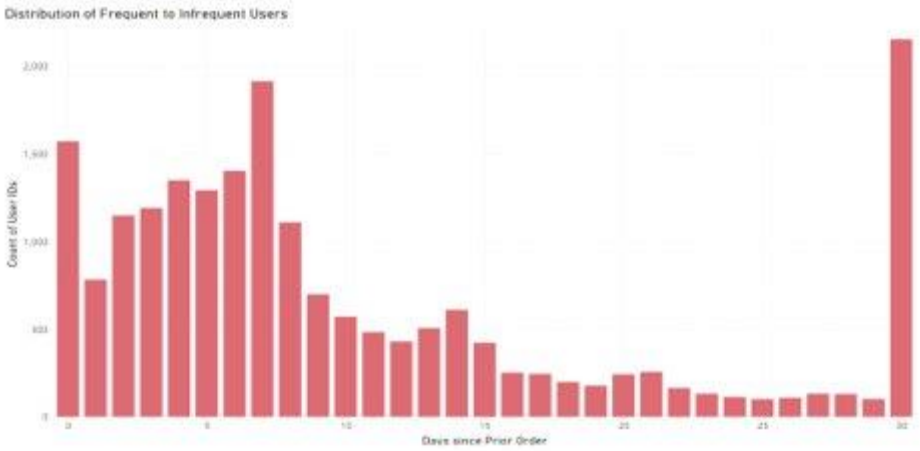
The market basket analysis is done based on the visualizations that we have created. It gives us a lot of insight into how we can improve our strategies and how the data can be utilized by different stakeholders to design some processes or work on the inventory more effectively. These visualizations will also help us to answer a lot of motivating questions which could resolve the demand/supply problem, decrease churn rate, and increase the sales of products that are low selling or not very popular among consumers.

The first visualization (Microsoft, 2023) revolves around the first SQL of exploratory data analysis. It gives the department wise revenue. The analysis is a clustered bar chart. Produce has the highest revenue; dairy eggs have the second and snacks the third. Bulk and alcohol have the lowest revenue generation which is less than 1 million. This could help inventory and supply chain managers with forecasting demand and planning for future inventory needs. The high-selling departments should be well stocked, and stock units can be adjusted accordingly. Financial analysis can do the budgeting as per the supply and demand. They can analyze the reasons behind variations in revenue generation among departments. This data also becomes the basis for return on investment for various departments. Since this is historical data, analysts can develop accurate forecasts and projections.

It opens doors for cross selling opportunities and to implement product mix strategy to grow revenue and focus on areas requiring attention. Marketing executives can plan promotions and campaigns to increase sales. Below is the visualization which depicts department wise revenue



The second visualization (Microsoft, 2023) is a bar chart which provides the count of user IDs against the days since prior order. It helps to identify users which order frequently and users which do not order frequently. It gives information about churn rate which is an opportunity for marketing team to design engagement opportunities to attract customers. Also, there is a need to find ways of how stakeholders can increase customer lifetime value.



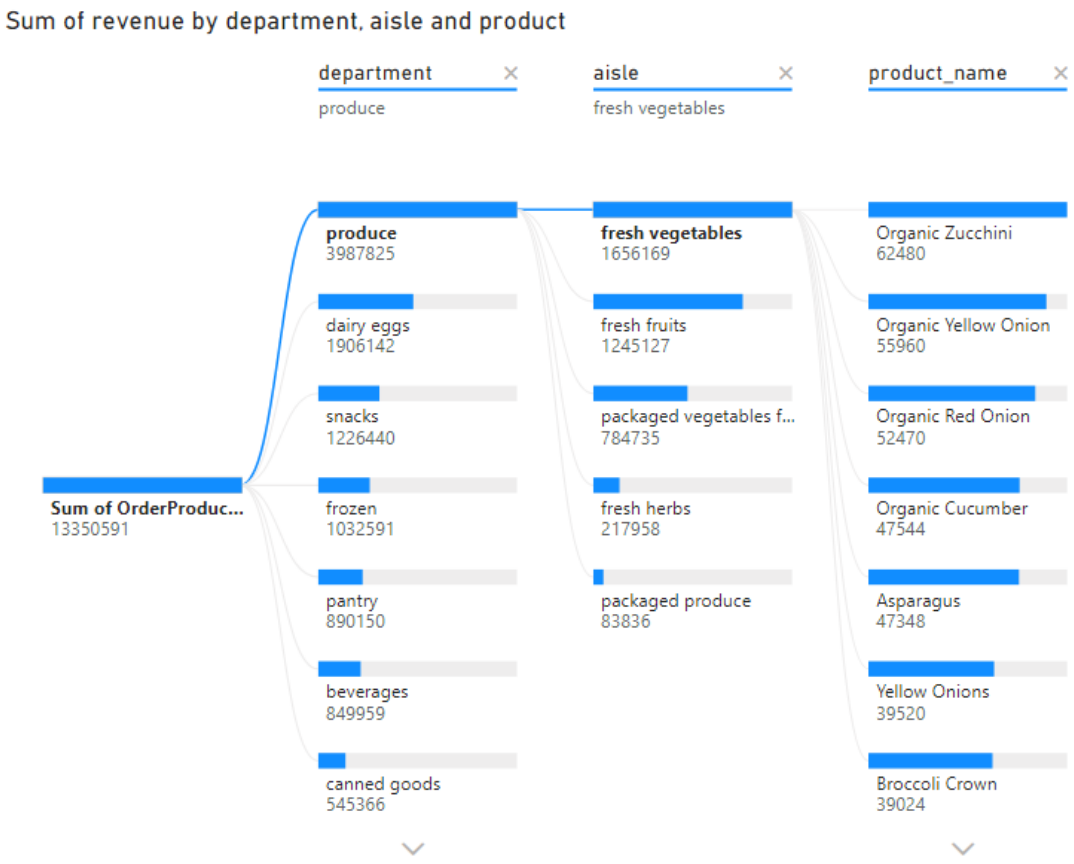
The third visualization (Microsoft, 2023) is a heatmap matrix which depicts the timeslots which have the most traction. The best insight that this visualization gives is shopper allocation. The busiest hours need more shoppers to be allocated and the light hours need less shoppers. This will help operations managers leverage this information to optimize staffing and operational activities based on the busy days of the week. The busiest hours are between 11 AM to 3 PM and the busiest day is Monday. The data suggests that the best day to stock the products is Sunday since the sales are expected to rise on Monday.

Number of Orders by Day of Week and Hour of Day

order_dow	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
0	115	105	10	45	10	20	85	380	910	1210	1310	1465	1500	1460	1560	1495	1330	1165	820	775	460	420	390	200
1	85	45	25	20	35	30	130	480	990	1460	1638	1525	1435	1435	1340	1455	1435	1095	890	635	610	360	230	190
2	155	40	35	40	25	50	195	350	769	1170	1220	1105	1005	1130	1100	1080	1120	1000	715	515	440	325	270	195
3	90	40	40	10	25	30	100	325	730	940	1115	1045	1110	1010	995	995	1085	775	735	535	500	295	235	155
4	119	60	35	5	20	70	135	335	610	930	915	945	860	945	1105	965	1059	800	565	605	360	265	260	100
5	105	35	20	20	10	25	165	415	675	985	1090	1095	1165	1005	1100	1025	960	930	685	495	455	265	160	170
6	110	40	30	20	15	40	60	375	635	925	950	1080	995	1115	1140	1040	1085	1005	735	495	385	295	335	190

The fourth visualization (Microsoft, 2023) is a scatter chart of frequently ordered by aisle. The best insight that this visualization gives is most reordered by aisle. A category manager and an inventory manager would benefit from this by stocking frequently reordered items accordingly. They can also use this to get meaningful insights such as fresh fruits might be ordered more than fresh vegetables due to higher shelf life and the fact that fruits can be eaten as snacks and require less preparation to be eaten. Patterns in reordering can assist with forecasting demand and planning for future inventory needs. If certain aisles show different reorder patterns, it may be an opportunity to review pricing strategies to either capitalize on high-demand items or improve the sales of less frequently purchased products.

The fifth visualization (Microsoft, 2023) is a decomposition tree which gives a high-level overview of departments which are high performing. The department is further segregated into aisle data which in turn is broken into products. This can help the retail store managers considerably since they can track which aisles need to be well-stocked. Department heads can be well informed on how they should design strategies for low performing departments. Data analysts can use this to create a story which can be presented to stakeholders in combination with other analysis so that they can bring changes in the areas of improvement based on actual insights.



Business Implications

We created a data warehouse using facts and dimensions. This has enabled us to EDA and develop visualizations which has helped us to take a deeper dig into the data. This would help business stakeholders to make informed decisions and have some business implications. We can see implications in the below categories

a. **Resource Allocation:** Visualizations showing revenue by department can guide where to allocate more resources, staff, or budget. Departments with high revenue might be prioritized for investment, while those with lower revenue may require strategy reevaluation.
b. **Inventory Management:** Insights from product and aisle sales performance can help in optimizing inventory levels, reducing waste due to overstocking, and minimizing lost sales due to stockouts.
c. **Marketing and Promotions**: Understanding purchasing patterns and customer behavior can lead to more targeted marketing campaigns. Promotions can be timed based on peak buying times or focused on high-margin products.
d. **Operational Efficiency:** Analysis of sales trends over time can help in streamlining operations, such as adjusting staff schedules to meet demand or optimizing store layouts to improve customer flow.
e. **Risk Management:** Understanding customer order frequency and departmental revenue can help in forecasting and managing risks associated with inventory and supply chain.

Conclusion

The data warehouse design on instacart data consists of dimensions and facts. The dimensions are orders, departments, aisles and products. The fact is fact_orders. We have populated the fact orders using ETL and created foreign key relationships. The ETL tool used for this is SSIS and the database used is SQL server. For the dimension tables, each datastream undergoes a sorting process.This ensures that consistent ordering within each dimension. Post the population of data in dimensions and facts, we do some exploratory data analysis using SQL. We analyzed the most reordered products, which hours of the day are noticed to have more sales than the others and identify the users which haven't ordered from Instacart for 30 days (about 4 and a half weeks). We used PowerBI to create data visualizations such as bar chart, clustered bar chart, scatter chart, heat matrix and decomposition tree. The clustered bar chart gives the department wise revenue.The second bar chart gives the count of userIDs against the days since prior order. The third visualization depicts the heat matrix which gives insight to which days of the week and hours of the day have the highest orders. The fourth visualization is a scatter chart of frequently ordered by aisle. The fifth visualization is a decomposition tree which gives a high-level overview of departments which are high performing. The department is further segregated into aisle data which in turn is broken into products. All these visualizations can help different stakeholders to plan strategies which can help to increase revenue and maintain the inventory in an efficient manner. Lastly, we discussed the business implications and suggestions that can be implemented.

References

Instacart. (2023). *Instacart Market Basket Analysis*. Retrieved from Kaggle: https://www.kaggle.com/competitions/instacart-market-basket-analysis/data

Microsoft. (2023). *Visualization Samples*. Retrieved from PowerBI: https://learn.microsoft.com/en-us/power-bi/developer/visuals/samples