**Data Warehousing**

**Final project**

**Group 24**

**Sneha Sharma**

**Sai Kumar Kanna**

**Akhila Reddy Kondi**

**Chaitanya Kopparthi**

**Srujana Turaga**

**Table of Contents**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | Title | Page no | | Executive Summary | **3** | | Problem Statement | **4** | | Literature Review | **5** | | Data Collection and Preparation | **5-6** | | Exploratory Data Analysis (EDA) | **6-11** | | Database Design | **12-13** | | Reporting, Modeling and Storytelling | **13** | | Query Visualizations | **14-17** | | Conclusion  References | **18**  **19** | |

**Executive Summary**

A data warehouse is a system that is designed to store large amounts of data that can be used for data analysis purposes. Data warehouses are specialized systems that can store huge amounts of data while simultaneously processing requests to process, retrieve or change information from the database server. Data warehouses have become an indispensable part of companies due to their dependency on analytics to get meaningful insights from data which drives decision-making.

The following report provides an outline of how a data warehouse works using Oracle SQL developer. Oracle SQL Developer is a consolidated development environment that is used for data modelling without requiring a database administration.

Diagram, schematic

Description automatically generated

The dataset we used for our final project was the ‘**Instacart Online Grocery Shopping Dataset**’ released by Instacart in 2017.

The report is divided in several sections, generally increasing in level of detail and specificity.

1. The first section describes the problem statement and the systematic approach we have chosen.
2. The second section describes collection and segregation of the data, design of the database, preparing and exploring it through analytical methods.
3. The final section of the report contains a graphical representation of our data which provides insights into what we found after understanding and analyzing the trends in the Instacart dataset.

**Problem Statement**

**Logo, company name

Description automatically generated**

We used “The Instacart Online Grocery Shopping Dataset 2017” which is an open-sourced dataset provided by Instacart. Instacart is a grocery delivery service that allows its users to shop from prime grocery stores online and delivers it on the same day. Users can conveniently shop online from stores within the same locality. The Instacart dataset contains 3 million grocery orders from more than 200,000 users with at least 4-100 orders for each user in the database.

The grocery industry has been rapidly expanding for the last few years. A research study from 2019 states that the average US household spends about 8000$/year on groceries. Companies can no longer rely on traditional practices to determine sale strategies in the industry and hence, there is pressure to stay competitive and keep the ever-demanding customer base happy by amplifying the overall experience while increasing the profitability of the business.

We used a systematized approach to get a good understanding of the data we chose and started to create the database after learning all the information we could gather. We assessed the following scenarios using the Instacart dataset:

1. Customers who have reordered a product.
2. The number of days since a customer last purchased a product.
3. Top products that are purchased from stores during the weekend.
4. Top aisles in grocery stores with the highest count of orders.
5. The average amount of products ordered by a customer.

The different tables in our database are Aisles, Department, OrderProducts, Orders and Products. OrderProducts is a fact table while the rest are dimension tables in our database.

Using the tables, we created analytical SQL queries to gather insight for analysis and then created a visual representation of the analysis we came up with.

The objective of our analysis is to collect data and establish a relationship between products and user purchase history, identify recurring patterns and track inventory changes and management.

**Literature Review**

Research shows that Instacart’s online grocery store solution has not only offered convenient deliveries all over the country, but also created more than 70,000 brick-and-mortar retail jobs and singlehandedly increased the country’s grocers revenue by 3.5 billion since the COVID pandemic. A study describes Instacart as a catalyst for the grocery industry. Statistics show a constructive relationship between the company’s growth and the growth of the grocery industry which benefited customers, retailers, workers and even the country’s economy. This is known as ‘The Instacart Effect’.

Naturally, the data released by Instacart was very interesting to explore since the grocery industry plays a very important part in our lives.

In this project, we have focused on the analytical SQL queries, using SQL developer, to create tables and aggregates for analysis. We also developed corresponding visuals using Tableau to helps us better understand and interpret the results.

The ultimate goal of our research is to create a prediction algorithm that could predict a customer’s future order based on previous purchase history. We also want to build a recommender system that can recommend products together based on correlation between them in the previous purchase history of that customer. This Data Warehouse offers a great starting point for creating methods that can help us combine all the data into a single Fact Table.

**Data Collection and Preparation**

We divided the dataset into Fact Tables and Dimension Tables.

**Fact tables:**

1. **Aisles**: This table contains the aisle number of the products and aisle product description.
2. **Department**: This table contains the id of the department along with the name of the department.
3. **Order Products**: This table gives information on which products were purchased in each order.It also contains previous order contents for all customers.
4. **Products**: This table contains details on the product id, name along with foreign keys aisle\_id to the aisle table and department\_id to the department table.

**Dimension Tables:**

1. **Orders**: This table tells us to which type (prior,train, test) an order belongs. We are predicting reordered items only for the test set orders. The column ‘order\_dow’ denotes the day of week.

**Exploratory Data analysis (EDA)**

As part of the exploratory data analysis, we created a few analytical SQL queries intending to get deeper insight into the dataset. There were many relationships that we were able to investigate. Furthermore, the queries we came up with were used as a foundation for the visualizations in our report.

**Query 1:**

**Estimating the number of products in each aisle.**

Graphical user interface, text, application

Description automatically generated

**Query 2:**

**Customers who reordered products.**

Graphical user interface

Description automatically generated with medium confidence

**Query 3:**

**Estimating the number of days since a customer last purchased a product.**

Graphical user interface, table

Description automatically generated

**Query 4:**

**Top 10 products that are ordered on weekends (Saturdays and Sundays)**

Graphical user interface, text

Description automatically generated with medium confidence

**Query 5:**

**Top 10 Aisles with highest orders count.**

**Graphical user interface, text

Description automatically generated**

**Query 6.**

**Average number of items a customer ordered. Calculated by ( Total ordered items by a customer / Total number of orders by a customer )**

**Graphical user interface, application

Description automatically generated**

**Database Design (EDA)**

The database contains information about Instacart customer purchase patterns for products in grocery stores. The different tables in the database are:

* Aisles
* Department
* Orderproducts
* Orders
* Products

Graphical user interface, application, Teams

Description automatically generated

The database connections made sense. Aisles contain products within different departments which are ordered by customers.

**Reporting, Modeling and Storytelling**

By using Tableau software, we were able to visualize our findings from the Instacart dataset to understand the purchasing patterns for different products in grocery stores.  We used a sequence of visualizations to convey the results of our analysis

Diagram

Description automatically generated

**Visualizations for the Queries**

**Visualization 1. Finding the total number of products in each aisle category**

A screenshot of a computer

Description automatically generated with medium confidence

**Visualization 2: First 10 Users with their number of reordered products**

**Chart, pie chart

Description automatically generated**

**Visualization 3: Top 10 products that are ordered on weekends (Saturdays and Sundays)**

**Chart, bar chart

Description automatically generated**

**A screenshot of a computer

Description automatically generated with medium confidence**

**Visualization 4: Top 10 Aisles with highest orders count.**

**Chart, treemap chart

Description automatically generated**

**Visualization 5: Busiest day of Week**

**Chart, bar chart

Description automatically generated**

**Visualization 6: Busiest Hour of the Day**

**Chart, line chart

Description automatically generatedVisualization 7: Shelf space by Department**

**Chart, bubble chart

Description automatically generated**

**Discussion**

From the analysis, we found:

* The number of purchases increases during the weekends.
* During the weekends (Saturdays and Sundays) we have seen that Organic bananas and normal bananas and Organic strawberries are the top three most ordered products.
* We have also tried to find out what aisles have highest order count and found that aisles 24,83,123 have stocks that are most ordered. This shows that the inventory management of these aisles must be properly taken care to avoid ‘Out Of Stock’ problems.
* Moreover, we have also seen the average number of items a customer ordered based on the total items ordered combined and the number of orders they did previously.  This data helps to identify frequent customers and customers who order in bulk, to provide them with more offers or discounts and increase sales. This also helps us to understand and predict a customer’s next order based on the items and frequency of items they bought previously.
* The busiest day of a week or the weekday with an aggregate count of total orders is Sunday and Monday. This indicates that there is an increase in order count typically during weekends and can be used to provide offers and discounts.
* We have also found out the aggregate sum of orders with respect to the hour time of the day and found out that midnight to 6 am is the time with relatively low orders. We can see in the visualization that after 7am there is a significant increase in the number of orders and goes to peak between 10am and 4pm. Post 4pm, we see a gradual decline in the total number of orders.
* We also visualized the shelf space by department i.e the types of different products within each department on Instacart and found out that personal care, snacks, and pantry are the top 3 departments with a wide range of products.
* This dataset is taken from relational data files which describe customer orders over time. We have performed multiple queries and analysis, used visualizations to further simplify the understanding of the results and in future, we would use these datasets to build a prediction model that can successfully predict what a customer’s next cart will consist of and also provide recommendations to customers based on their purchase history.
* We would also like to predict the correlations between products under similar order id by a customer to provide a ‘Frequently added together’ recommendations.
* We have used Oracle SQL developer relational database management system for structuring the data, cleaning it, transforming it and successfully deploying it on server before we used it for analytical purposes.

Our data warehousing model can be used by stakeholders as a fundamental backbone to test different approaches and make decisions for inventory management, promotional strategies, distribution of the groceries network and creating a loyal customer base. It can also be used by grocers on the Instacart distribution network to make decisions based on what’s best for the store.

**References**

[*“The Instacart Online Grocery Shopping Dataset 2017”, Accessed from https://www.instacart.com/datasets/grocery-shopping-2017 on <07/10/2022>*](https://www.instacart.com/datasets/grocery-shopping-2017)

[*https://www.kaggle.com/c/instacart-market-basket-analysis/data*](https://www.kaggle.com/c/instacart-market-basket-analysis/data)

[*https://www.instacart.com/company/updates/instacarts-impact-on-americas-grocery-industry-by-the-numbers/*](https://www.instacart.com/company/updates/instacarts-impact-on-americas-grocery-industry-by-the-numbers/)

[*https://www.instacart.com/company/updates/the-instacart-effect-giving-grocers-and-the-economy-a-boost/*](https://www.instacart.com/company/updates/the-instacart-effect-giving-grocers-and-the-economy-a-boost/)

[*https://www.businessofapps.com/data/instacart-statistics/#:~:text=While%20Instacart%20doesn't%20publish,%2435%20billion%20in%20grocery%20sales.*](https://www.businessofapps.com/data/instacart-statistics/#:~:text=While%20Instacart%20doesn't%20publish,%2435%20billion%20in%20grocery%20sales.)

[*https://www.thekitchn.com/i-had-my-groceries-delivered-by-instacart-and-heres-how-it-went-214795*](https://www.thekitchn.com/i-had-my-groceries-delivered-by-instacart-and-heres-how-it-went-214795)

[*https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2*](https://tech.instacart.com/3-million-instacart-orders-open-sourced-d40d29ead6f2)

[*https://www.netsuite.com/portal/resource/articles/data-warehouse/data-warehouse.shtml*](https://www.netsuite.com/portal/resource/articles/data-warehouse/data-warehouse.shtml)

[*https://www.foodlogistics.com/warehousing/grocery-retail/article/22105965/local-express-from-data-to-insights-the-value-of-data-analysis-for-grocers*](https://www.foodlogistics.com/warehousing/grocery-retail/article/22105965/local-express-from-data-to-insights-the-value-of-data-analysis-for-grocers)

[*https://www.marketingtutor.net/swot-analysis-of-grocery-store/*](https://www.marketingtutor.net/swot-analysis-of-grocery-store/)

[*https://www.forbes.com/sites/chloesorvino/2021/01/27/instacart-survived-covid-chaos---but-can-it-keep-delivering-after-the-pandemic/?sh=5c3b8159bfa1*](https://www.forbes.com/sites/chloesorvino/2021/01/27/instacart-survived-covid-chaos---but-can-it-keep-delivering-after-the-pandemic/?sh=5c3b8159bfa1)