**Phase 1 Report**

**Group 36**

TOPIC - Instacart Market Basket Analysis

For this project, we have selected the "Instacart Market Basket Analysis" dataset available on Kaggle. The dataset is anonymized and contains a sample of over 3 million grocery orders from more than 200,000 Instacart users. For each user, it provides between 4 and 100 of their orders, with the sequence of products purchased in each order. It also provides the week and hour of day the order was placed, and a relative measure of time between orders. The dataset can be downloaded from the following link:

<https://www.kaggle.com/c/instacart-market-basket-analysis/data>

The dataset consists of the following files:

orders.csv: This file contains information about each order, such as the order ID, user ID, order number, order day of the week, order hour of day, and days since the user's previous order.

products.csv: This file contains information about each product, such as the product ID, product name, aisle ID, and department ID.

order\_products\_\_prior.csv: This file contains information about each product that was ordered in the past, such as the order ID, product ID, and the order in which the product was added to the cart.

order\_products\_\_train.csv: This file contains information about each product that was ordered in the train set, such as the order ID, product ID, and whether the product was reordered.

aisles.csv: This file contains information about each aisle, such as the aisle ID and aisle name.

departments.csv: This file contains information about each department, such as the department ID and department name.

To create a meaningful relational model for this dataset, we propose the following schema:

User(user\_id)

Order(order\_id, user\_id, order\_number, order\_dow, order\_hour\_of\_day, days\_since\_prior\_order)

Product(product\_id, product\_name, aisle\_id, department\_id)

Aisle(aisle\_id, aisle\_name)

Department(department\_id, department\_name)

Order\_Product(order\_id, product\_id, add\_to\_cart\_order, reordered)

The relational model is designed to represent the entities in the dataset and their relationships. The User table contains only the userID. The Order table contains information about each order, such as the order ID, user ID, order number, day of the week, hour of the day, and days since the user's previous order. The Product table contains information about each product, such as the product ID, product name, aisle ID, and department ID. The Aisle and Department tables contain information about each aisle and department, respectively. The Order\_Product table contains information about each product that was ordered, such as the order ID, product ID, the order in which the product was added to the cart, and whether the product was reordered.

Text

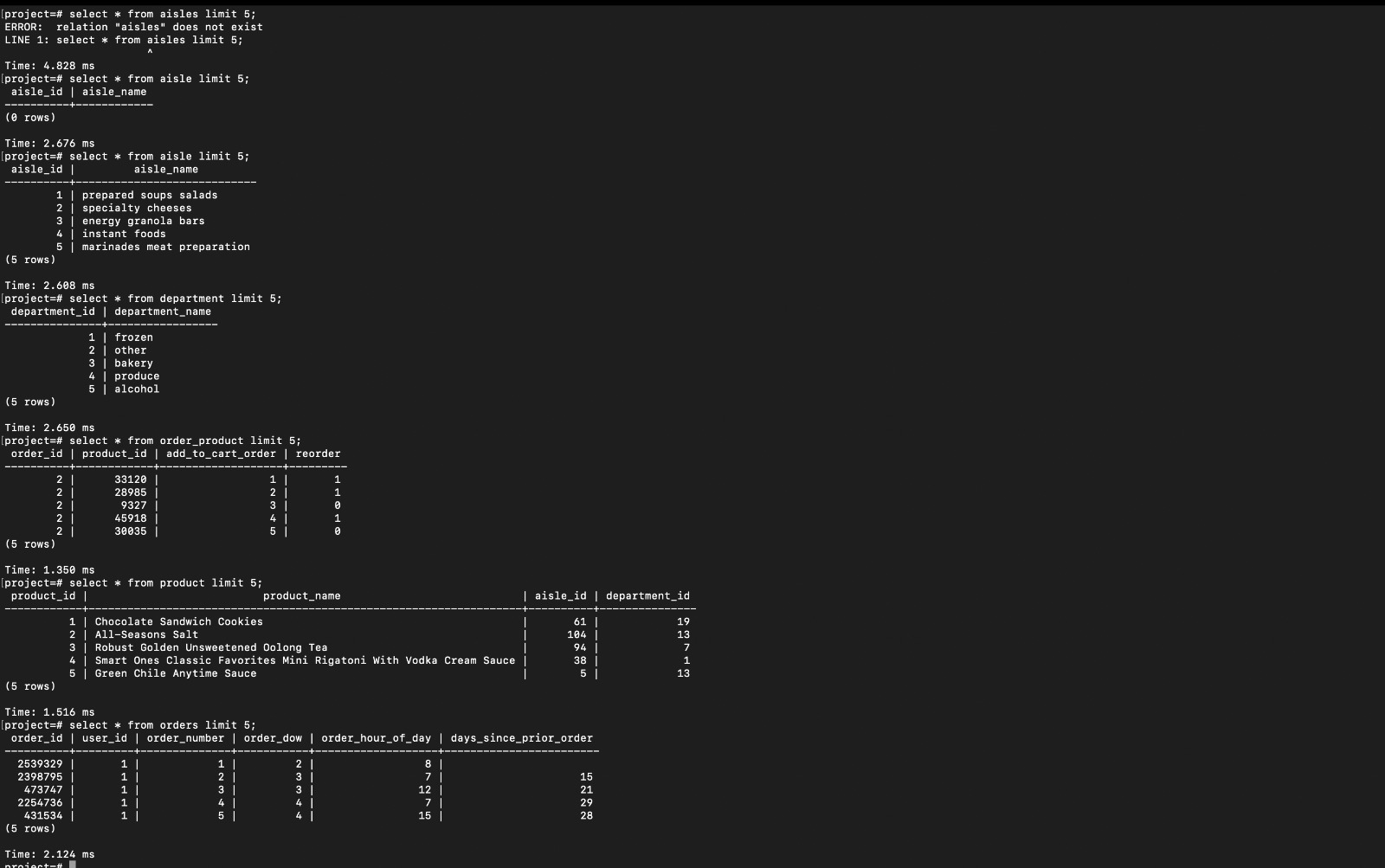
Description automatically generated

Text

Description automatically generated

Text

Description automatically generated



1. Find the top 10 products that were most frequently reordered by customers:

Shape

Description automatically generated with low confidence

1. Find the average days since a customer's last order for each department:

Text

Description automatically generated

1. Find the top 10 aisles with the highest percentage of products that were reordered:

Text

Description automatically generated with medium confidence

1. Find the number of products in each department that were purchased during each hour of the day:

Text

Description automatically generated

Time taken by our 5 queries:

Graphical user interface, text

Description automatically generated

Time taken by our 5 queries after adding indexing:

Graphical user interface, text, application

Description automatically generated

As we can see the time got decreased by a considerable amount for query1, and 3. For the last query it still takes a lot of time, as it deals with 4 tables.