GROUP 36

Cleaning Our Dataset

- We had already cleaned our dataset by inserting value 0 whenever we encountered null values in the attribute for example in the case of days_since_prior_order attribute of the orders table. We also converted the values for our attributes to integer and text types as required.

```
def main():
    s_t = time.time()
    conn = psycopg2.connect(database="project", host="localhost", user="maitreyakocharekar", password="project123", port="5432")
    print(conn)
    cursor = conn.cursor()
    """
    pd1 = pd.read_csv(r'/Users/maitreyakocharekar/Documents/Sem2(Spring2022-23)/Big Data/project/instacart-market-basket-analysis/orders.csv')
    print(pd1.head())
    pd1['days_since_prior_order'] = pd1['days_since_prior_order'].fillna(0)
    pd1['days_since_prior_order'] = pd1['days_since_prior_order'].astype('int')
    pd1 = pd1.loc[pd1['eval_set'] == 'prior']
    orders = pd1[["order_id","user_id","order_number", "order_dow", "order_hour_of_day","days_since_prior_order"]]
    orders.to_csv('/Users/maitreyakocharekar/Documents/Sem2(Spring2022-23)/Big Data/project/instacart-market-basket-analysis/ordersnew.csv', encoding
    """



    sql_query = """
    CREATE TABLE temp(
```

Integration of Dataset

- For data integration we created different kinds of views, one of which, dept_aisle(view2), sorts the aisles according to departments and displays an array of aisles for each department. For this we integrated the product department and aisle table and created a view out of it to find which different aisles belong to what departments.

```
project=# select * from view2;
 department_name |                                                                                                                   aisles

-----------------+--------------------------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------
 alcohol         | {"beers coolers","red wines","specialty wines champagnes",spirits,"white wines"}
 babies          | {"baby accessories","baby bath body care","baby food formula","diapers wipes"}
 bakery          | {"bakery desserts",bread,"breakfast bakery","buns rolls","tortillas flat bread"}
 beverages       | {"cocoa drink mixes",coffee,"energy sports drinks","juice nectars",refrigerated,"soft drinks",tea,"water seltzer sparkling water"}
 breakfast       | {"breakfast bars pastries",cereal,granola,"hot cereal pancake mixes"}
 bulk            | {"bulk dried fruits vegetables","bulk grains rice dried goods"}
 canned goods    | {"canned fruit applesauce","canned jarred vegetables","canned meals beans","canned meat seafood","soup broth bouillon"}
 dairy eggs      | {butter,cream,eggs,milk,"other creams cheeses","packaged cheese","refrigerated pudding desserts","soy lactosefree","specialty cheeses",yogurt}
 deli            | {"fresh dips tapenades","lunch meat","prepared meals","prepared soups salads","tofu meat alternatives"}
 dry goods pasta | {"dry pasta","fresh pasta","grains rice dried goods","instant foods","pasta sauce"}
 frozen          | {"frozen appetizers sides","frozen breads doughs","frozen breakfast","frozen dessert","frozen juice","frozen meals","frozen meat seafood","frozen pizza","froz
,"ice cream ice"}
 household       | {"air fresheners candles","cleaning products","dish detergents","food storage","kitchen supplies",laundry,"more household","paper goods","plates bowls cups fl
 international   | {"asian foods","indian foods","kosher foods","latino foods"}
 meat seafood    | {"hot dogs bacon sausage","meat counter","packaged meat","packaged poultry","packaged seafood","poultry counter","seafood counter"}
 missing         | {missing}
 other           | {other}
 pantry          | {"baking ingredients","baking supplies decor",condiments,"doughs gelatins bake mixes","honeys syrups nectars","marinades meat preparation","oils vinegars","pi
preads","salad dressing toppings","spices seasonings",spreads}
 personal care   | {beauty,"body lotions soap","cold flu allergy",deodorants,digestion,"eye ear care","facial care","feminine care","first aid","hair care","muscles joints pain
replacements","shave needs","skin care",soap,"vitamins supplements"}
 pets            | {"cat food care","dog food care"}
 produce         | {"fresh fruits","fresh herbs","fresh vegetables","packaged produce","packaged vegetables fruits"}
 snacks          | {"candy chocolate","chips pretzels","cookies cakes",crackers,"energy granola bars","fruit vegetable snacks","ice cream toppings","mint gum","nuts seeds dried
ck mix"}
(21 rows)

project=# select count(*) from department;
 count
-------
    21
(1 row)

project=# 
```

- ➨ **Similar task was done for Aisle_Product, wherein we sorted different unique products into their aisle by joining aisle and product table. This way any user can simply find the location of their desired product.(This one may look complex as there are a lot of products for an aisle)**

```
 asian foods           | {"1 Step-1 Minute Noodles Toasted Sesame","100% Pure Sesame Seed Oil","100% Whole Soy Organic Gluten Free Tamari Soy Sauce","100% Whole Wheat Panko","37% Less Sodium Soy Sauce","Agar Agar
Sea Vegetable Flakes","Aji-Mirin Sweet Cooking Rice Seasoning","Albacore Tuna in Thai Coconut Lemongrass","Albacore Tuna in Yellow Coconut Curry","All Natural Hoisin Sauce","All Natural Hokkien Stir-Fry Noodles - 2
 CT","All Natural Oyster Sauce","All Natural Soba Stir-Fry Noodles - 2 CT","All Natural Udon Stir-Fry Noodles","All Natural Won Ton Wraps","All Purpose Soy Sauce","Asian BBQ Sauce","Asian Noodles Teriyaki","Asian S
auces, Gourmet, Thai Peanut","Asian Vegetable Ramen","Atlantic Dulse","Baked Chipotle Tofu","Bamboo Rolling Mat and Paddle","Bangkok Curry Instant Rice Noodle Soup","Bangkok Peanut Dipping Sauce","Banh Pho Rice Noo
dles","Barley Miso, Premium, Organic, Country","Basil Chili Rice","Bean Sprouts in Water","Bean Threads","Bean Threads, 3 Pack Tray","Beef Flavor Cup Noodles","Beef Flavor Top Ramen","Bifun Rice Pasta","Big Cup Noo
dles Homestyle Shrimp Flavor","Big Kick Wasabi","Biscuits With Milk Cream","Black & Tan Gomasio","Black Bean Garlic Sauce","Black Roasted Sesame Seed","Bluegrass Soy Sauce","Bonito Flakes, Aged and Dried","Broccoli
 Beef Stir Fry Seasoning Mix","Broth, Vietnamese Pho","Brown Rice Sticky Noodles","Brown Sugar & Sea Salt Seaweed Snacks","Butter Chicken Simmer Sauce","California Roll","Cantonese Oyster Flavored Sauce","Chana Dal
 Split Desi Chickpeas","Chicken Chow Mein","Chicken Flavor Cup Noodles","Chicken Flavor Japanese Style Udon Noodles With Soup Base","Chicken Flavor Top Ramen Noodles","Chicken Pad Thai","Chicken-Style Seitan","Chic
kpea Rice Miso","Chili Garlic Sauce","Chili Oil","Chili Paste, Ground Fresh","Chili Pepper, Assorted","Chinese Barbecue Char Siu Seasoning Mix","Chinese Barbecue Sauce","Chinese Duck Sauce","Chinese Five Spices Pow
der","Chinese Fried Rice Seasoning Mix","Chinese Noodles","Chinese Style Chicken Soup Bowl","Chinese-Style Extra Hot Mustard","Chocolate Brownie Mochi","Chow Mein","Chow Mein Noodle","Chow Mein Noodles","Chow Mein
Seasoning Mix","Chow Mein Stir-Fry Noodles","Chow Mein Wide Noodles","Classic Stir-fry Sauce","Coconut Aminos Garlic Soy Sauce","Coconut Aminos Teriyaki Sauce","Coconut Milk","Cooking Sauce, Thai, Red Curry, Medium
","Cracked Pepper & Herbs Roasted Seaweed Snacks","Cup Noodles Ramen Noodle Soup with Shrimp","Curry Paste, Panang","Curry Powder, Oriental","Cut Baby Corn","Dark Japanese Style Soy Sauce","Diakon Radish Kimchi","D
ulse, Wild Atlantic Sea Vegetable","Edamame Fettucine With Thai Coconut Sauce","Egg Drop Soup Mix","Egg Roll Wrappers","Egg Roll Wraps","Fancy Bamboo Shoots","Fancy Mixed Chinese Vegetables","Fancy Sliced Bamboo Sh
oots","Fancy Sliced Water Chestnuts","Fenugreek Seeds","Fortune Cookies","Fried Rice Seasoning Mix","Fuero Wakami Dried Seaweed","Futonaga Udon Noodles","Garlic & Green Onion Teriyaki Sauce","Garlic & Vegetable Ins
tant Rice Noodle Soup","Garlic Pepper Ramen","Garlic Sesame Rice Noodle Soup Bowl","General Tsao Stir-Fry Sauce","General Tso's Chicken Season","General Tso's Sauce","General Tso's Sauce & Glaze","Genmai Miso Aged
and Fermented Soy and Brown Rice","Genuine Brewed Rice Vinegar","Ginger Sushi","Ginger Teriyaki Stir-Fry  Sauce","Ginger, Pickled","Ginger, Pickled Sushi","Gluten Free Miso Soup Tofu","Gluten Free Soy Sauce Tamari
Lite","Gluten Free Teriyaki Sauce","Gluten-Free Garlic Goodness Vietnamese Brown Rice Noodle Soup","Gluten-Free Soy Sauce","Glutin Free Low Fat Rice Crackers Black Sesame and Soy Sauce Bag","Go Chu Jang Sweet and S
picy Sauce","Gochujang Fermented Chile Paste Concentrate","Gochujang Fermented Garlic Chile Paste","Gochujang Fermented Sesame Chile Paste","Gourmet Teriyaki Stir Fry Sauce","Green Curry Paste","Ground Sesame Tahin
a","Hacho Miso Aged & Fermented Soybeans","Happy Pho Zesty Ginger Brown Rice Noodle Soup","Hello Panda Biscuits with Strawberry Cream","Hello Panda Choco Cream Biscuits","Hi-Chew Fruit Chews Green Apple","Hi-Chew M
ango Fruit Chews","Hi-Chew Strawberry Fruit","Hi-Chew Strawberry Fruit Chews","Hoisen Sauce","Hoisin Sauce","Homestyle Beef Flavor Ramen Noodle Soup","Hot & Sour  Chinese Style Egg Flower Soup Mix","Hot & Sour Soup
 Bowl","Hot & Spicy Noodle Bowl Soup","Hot Chili Oil","Hot Chili Sesame Oil","Hot Garlic Baked Shrimp Flavored Chips","Hot Golden Curry Sauce Mix","Hot Kim Chee","Hot Thai Red Curry Paste","Hot Wasabi Coated Green
Peas","Hot Wasabi Peas","House Napa Cabbage","Individual Jumbo Udon Serving Packets with Soup Base","Instant Aka Miso Soup","Instant Noodle Soup, Miso","Instant Rice Vermicelli Noodles","Instant Sea Vegetable Wakam
e","Instant Soup, Soybean Paste with Tofu, Tofu Miso","Iso Maki, Seaweed Wrapped Rice Cracker","Jade Pearl Rice Ramen","Japanese Buckwheat Noodles Soba","Japanese Soba Buckwheat Noodle","Japanese Style Extra Crispy
 Tempura Batter Mix","Japanese Style Noodles & Chicken Flavored Sauce","Japanese Style Rice Crackers","Japanese Thick Udon Noodles","Japanese Udon Noodles","Jasmine Brown Rice","Jasmine Green Iced Tea","Kaedama Ram
en Noodles","Katsu Sauce","Kelp Atlantic Kombu","Kelp Noodles","Kelp Noodles with Green Tea","Kelp, Wild Atlantic Kombu",Kimchi,"Kimchi Flavor Noodle Soup",Kombu,"Kombu Sea Vegetable","Korean BBQ Sauce","Korean Goc
hujang Sauce","Korean Kalbi Bulgogi BBQ Sauce","Korean Kimchi","Korean Stir Fry Simmer Sauce","Korean Sweet Chili Noodle Bowl","Korean Teriyaki Stir-Fry Sauce","Kung Pao Noodle Bowl","Kung Pao Noodles","Kung Pao Sa
uce","Kuzu Root Starch","Laksa Noodles","Lemon Sliced Grass","Lemongrass & Chili Instant Rice Noodle Soup","Less Salt Soy Sauce","Less Sodium Soy Sauce","Less Sodium Soy Sauce Dispenser Bottle","Less Sodium Teriyak
i Marinade & Sauce","Light Coconut Milk","Lightly Salted Roasted Edamame","Lime Ponzu Citrus Seasoned Dressing & Sauce","Lite Coconut Milk","Lite Rice Vinegar","Lite Soy Sauce","Lo Mein Egg Noodles","Lotus Forbidde
:
```

Itemset Mining

- ➨ First, we formed a table named temp2. This table contains all the important information that we need. The table is formed by joining order and order_product table on order_id, call this newly formed table temp1, and then joining product table on the temp1 table on product_id thus resulting in our temp2 table.
- ➨ We then applied itemset mining on this table for product_id, as our goal here was to find the products that were bought together frequently in multiple different orders.

- The basic process is a python program that keeps on writing sql queries to form k-items lattice, till the rows in the lattices are zero or no more k number of elements are being brought together at least n number of times where n is our threshold set by us.
- We have set the threshold to 10,000. This means only those elements are included who were being brought in minimum 10000 number of orders.
- We then take each row of the final lattice and take each unique element from them and print at the last the most frequently brought items.
- In Total we got three lattice and we have printed the third lattice for you. The fourth lattice contains 0 rows.

```
project=#
project=# select * from l3;
 product_id1 | product_id2 | product_id3 | count
-------------+-------------+-------------+-------
       13176 |       21137 |       47209 | 15066
       13176 |       21903 |       47209 | 12196
       13176 |       21137 |       27966 | 11584
       13176 |       27966 |       47209 | 11409
       13176 |       21137 |       21903 | 10967
       21903 |       24852 |       47766 | 10770
(6 rows)

project=# select * from l3p;
  id   |          name
-------+------------------------
 47209 | Organic Hass Avocado
 21903 | Organic Baby Spinach
 21137 | Organic Strawberries
 24852 | Banana
 47766 | Organic Avocado
 13176 | Bag of Organic Bananas
 27966 | Organic Raspberries
(7 rows)

project=#
```

```
project=# select * from triplets;
        product1        |       product2       |       product3       | count
------------------------+----------------------+----------------------+-------
 Bag of Organic Bananas | Organic Strawberries | Organic Hass Avocado | 15066
 Bag of Organic Bananas | Organic Baby Spinach | Organic Hass Avocado | 12196
 Bag of Organic Bananas | Organic Strawberries | Organic Raspberries  | 11584
 Bag of Organic Bananas | Organic Raspberries  | Organic Hass Avocado | 11409
 Bag of Organic Bananas | Organic Strawberries | Organic Baby Spinach | 10967
 Organic Baby Spinach   | Banana               | Organic Avocado      | 10770
(6 rows)

project=# select * from l3;
 product_id1 | product_id2 | product_id3 | count
-------------+-------------+-------------+-------
       13176 |       21137 |       47209 | 15066
       13176 |       21903 |       47209 | 12196
       13176 |       21137 |       27966 | 11584
       13176 |       27966 |       47209 | 11409
       13176 |       21137 |       21903 | 10967
       21903 |       24852 |       47766 | 10770
(6 rows)
```

Which model is a best fit for our dataset?

- In the case of our project, the Instacart Market Basket Analysis dataset, a relational model is a better fit for the task of itemset mining to discover interesting association rules.
- The dataset consists of structured data with well-defined entities such as orders, products, aisles, and departments. Each entity has its own set of attributes that can be easily mapped to columns in a relational database.
- This makes it easy to create a schema that represents the entities and their relationships, as we described in our proposed schema.
- Relational databases are designed for structured data and excel at handling large volumes of structured data with complex relationships. They also provide powerful query capabilities for joining and aggregating data from multiple tables, which is important for tasks such as itemset mining.
- On the other hand, document-oriented databases are better suited for unstructured or semi-structured data that can vary in schema and format. They provide more flexibility in terms of schema design and can handle large volumes of data with high write and read throughput.
- However, in the case of the Instacart dataset, the structured nature of the data and the well-defined relationships between entities make a relational model a more appropriate choice for itemset mining.