instacart

BIG DATA ANALYTICS

Semester Project Presentation

Aiza Islam - 343201

Shehryar Saqib - 347703

TABLE OF CONTENT

INTRODUCTION

- Analysis of Instacart's historical purchasing data
- To understand consumer behavior and shopping patterns.
- Exploratory Data Analysis (EDA)
- Frequent Pattern Growth Algorithm & Association Rule Mining
- Recommender System
- Clustering Segment customers
- Page Rank Algorithm

DATASET

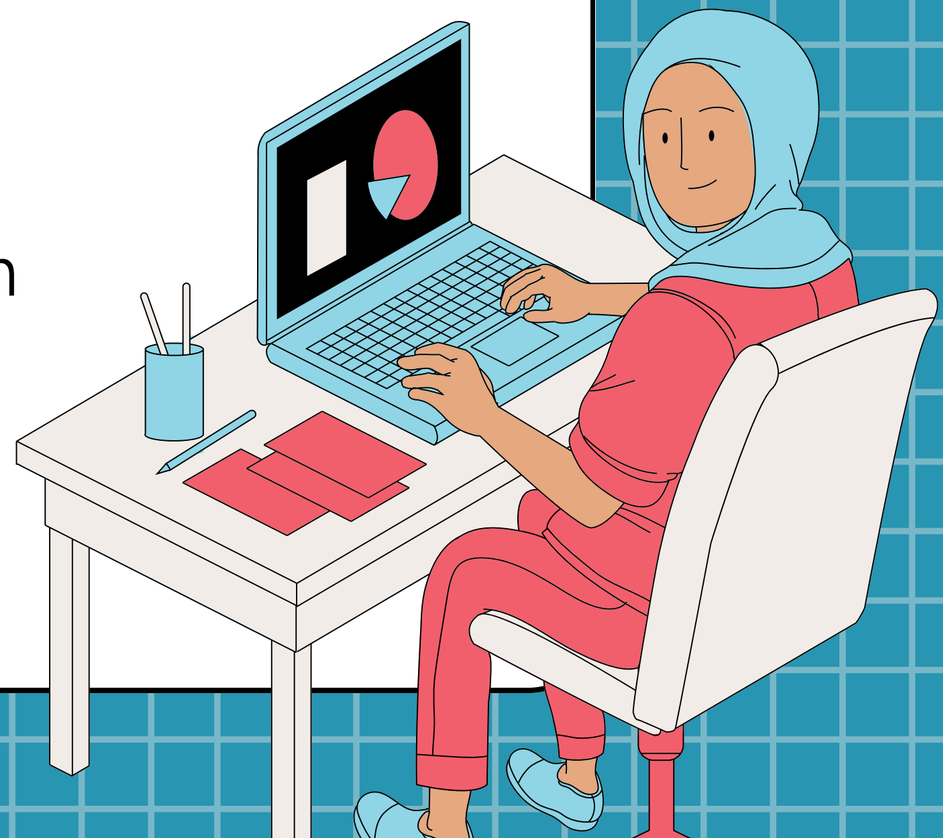
- Instacart Market Basket Analysis
- Dataset Overview:
  - Orders.
  - Products.
  - Order Products
  - Aisles and Departments.
- Data Characteristics:
  - Temporal Features
  - Categorical Data
  - High Dimensionality

USE CASE

- Improved Product Recommendation
- Business Benefits
- Efficient Inventory Management
- Targeted Marketing Campaigns
- Strategic Decisions for Customer Retention

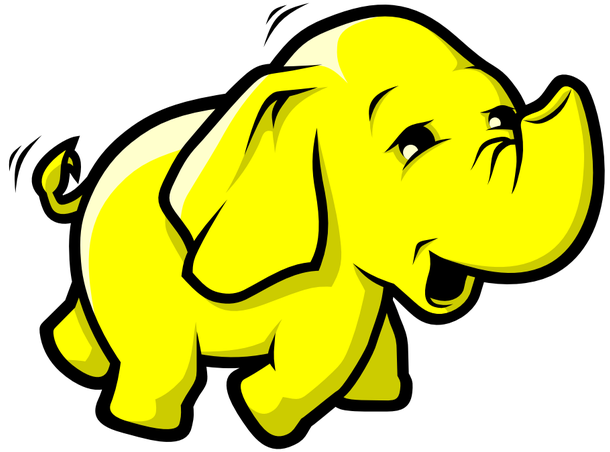
DATA PREPROCESSING

- Loading Data
- Merging Data
- Cleaning Data
  - Checked for missing values with isnull().sum().
  - Removed duplicates using drop_duplicates().

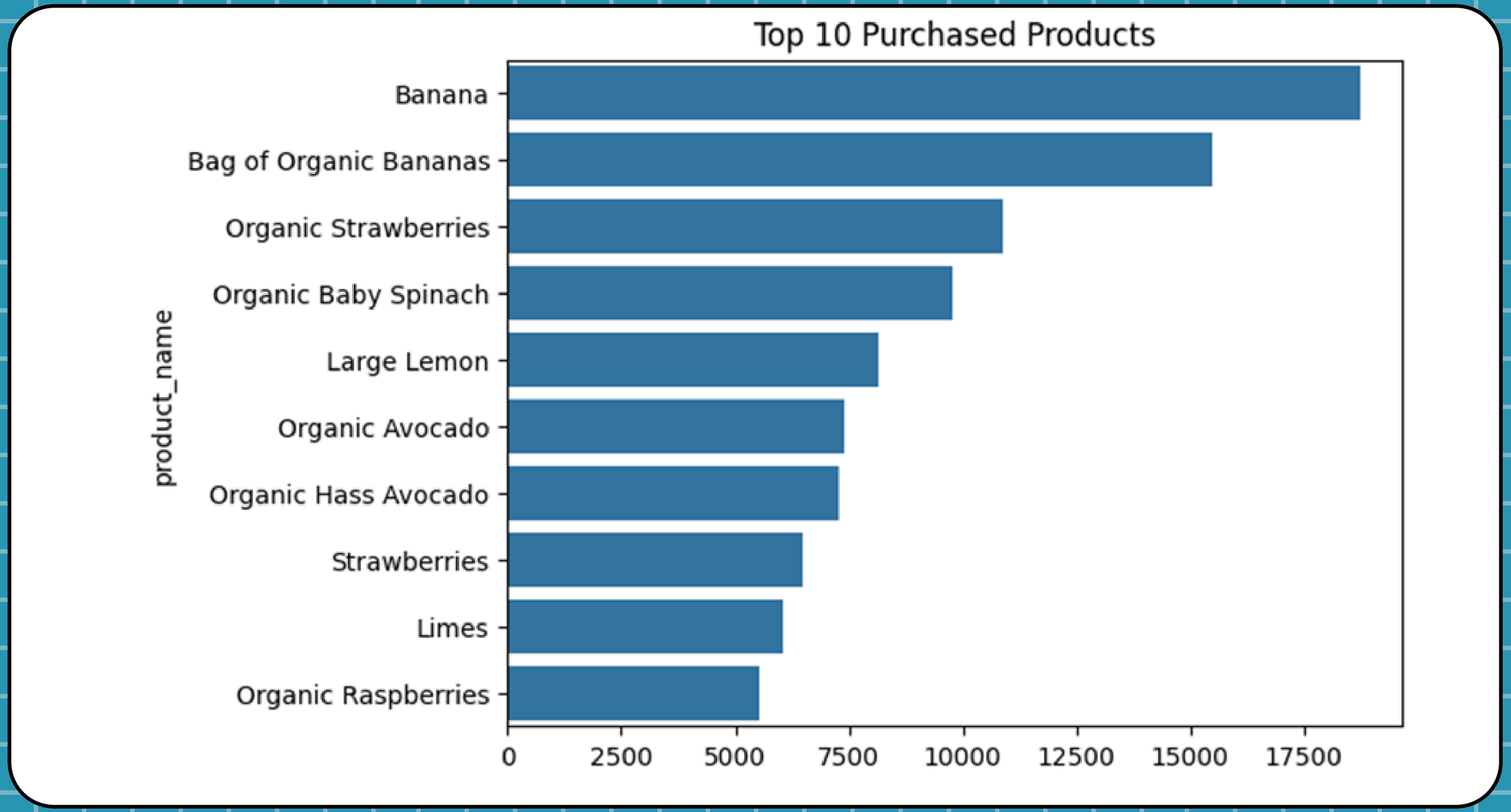


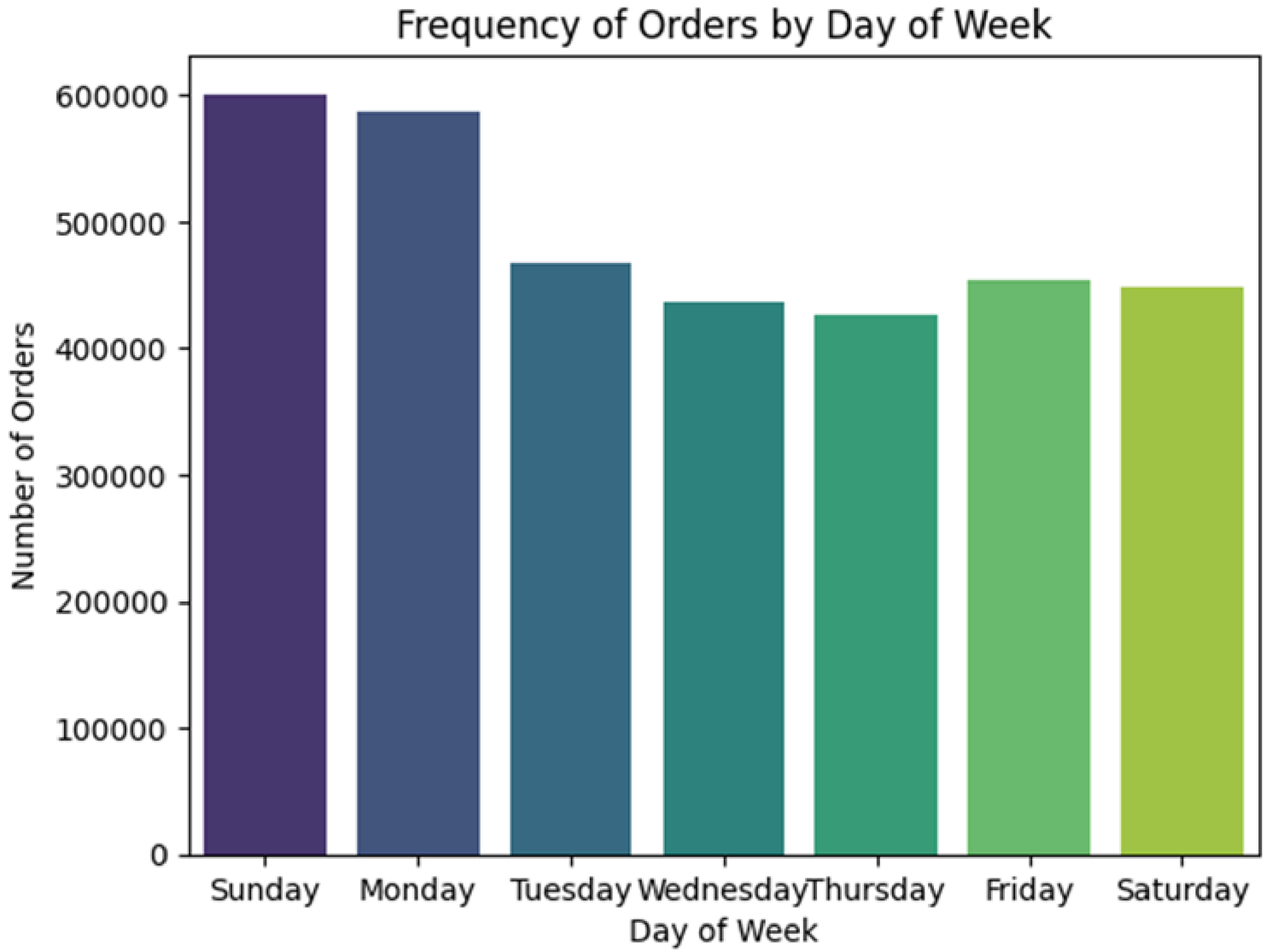
TECHNOLOGIES

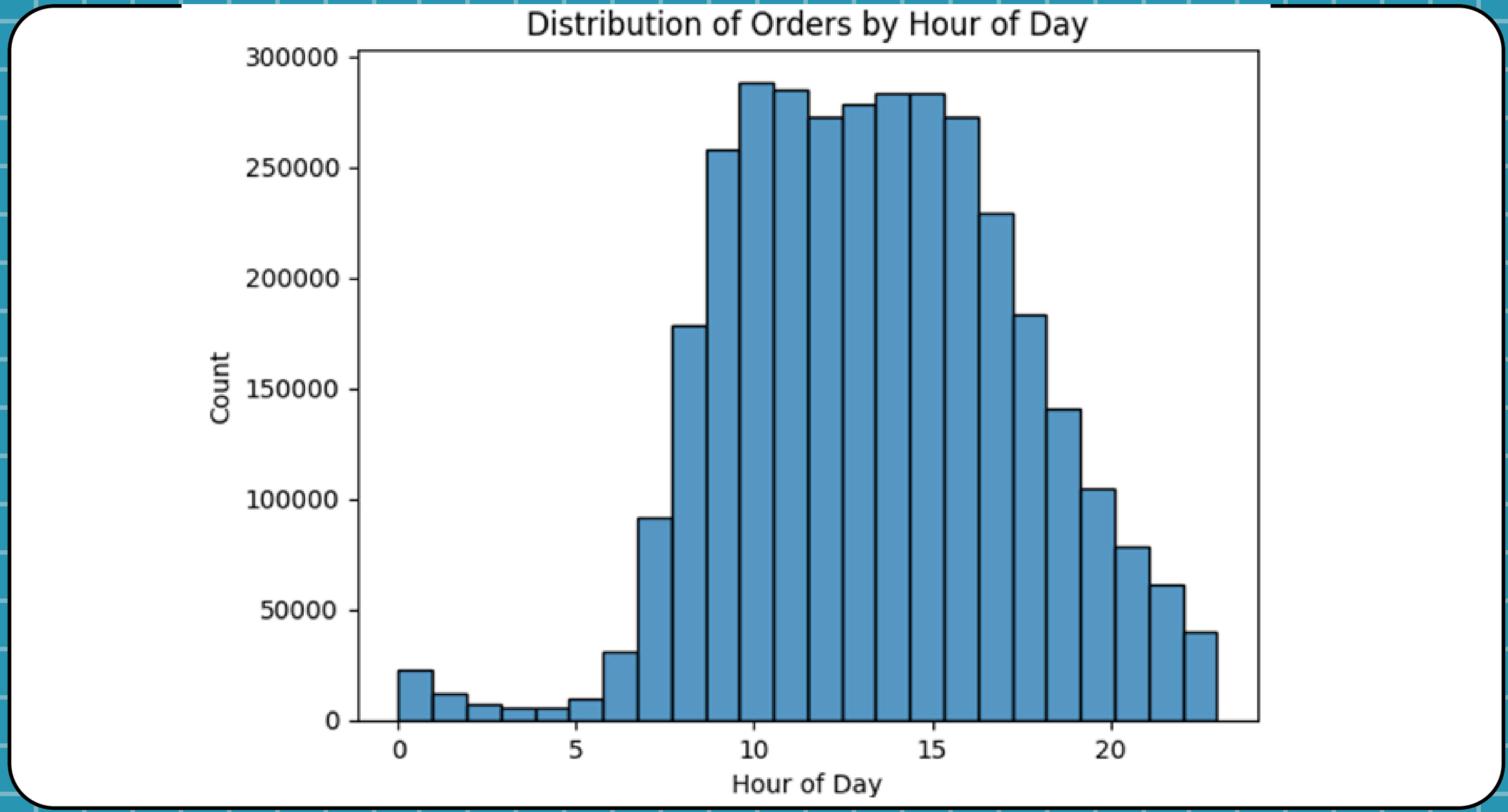
For EDA, BDA Techniques and Analytics:

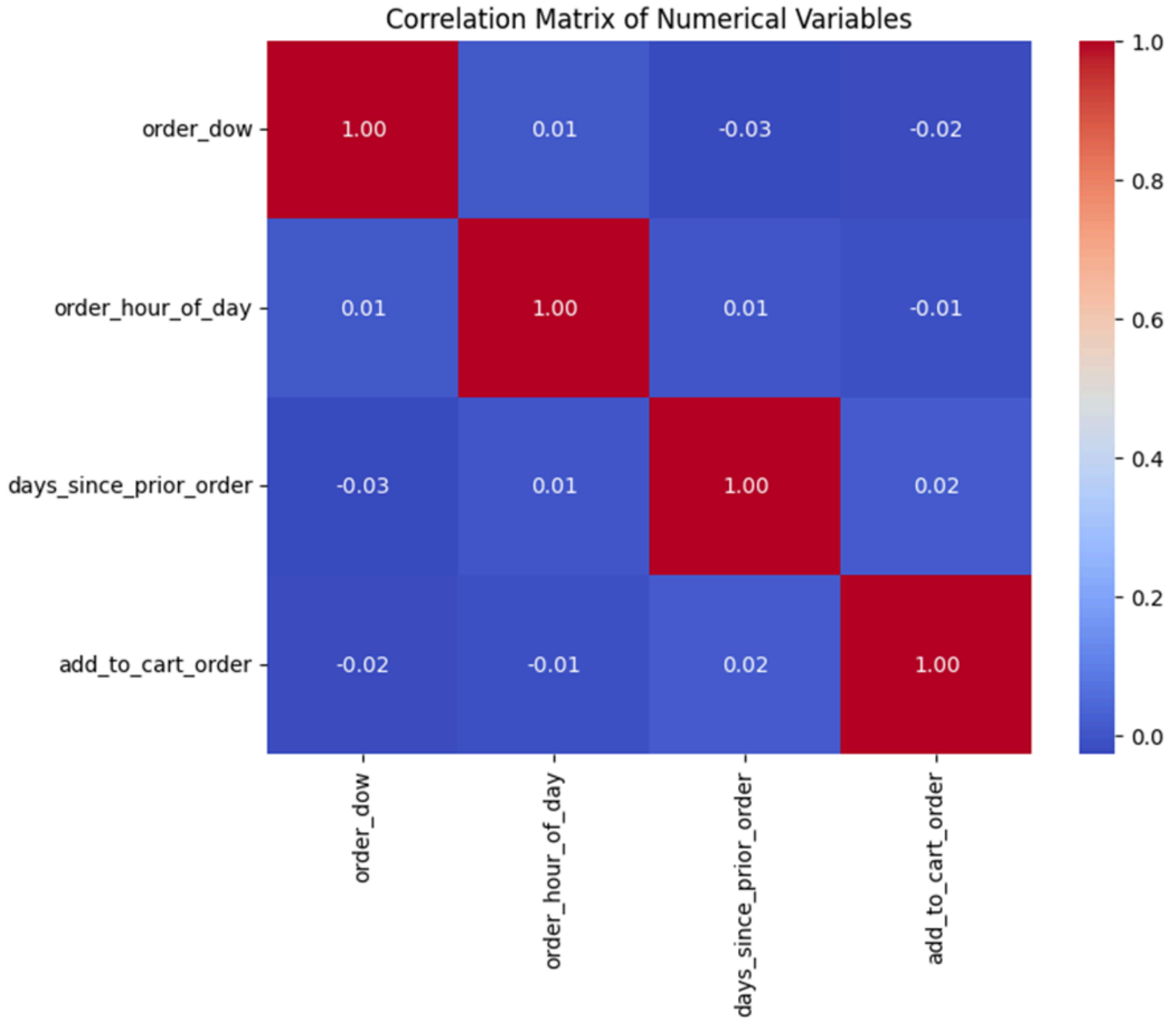


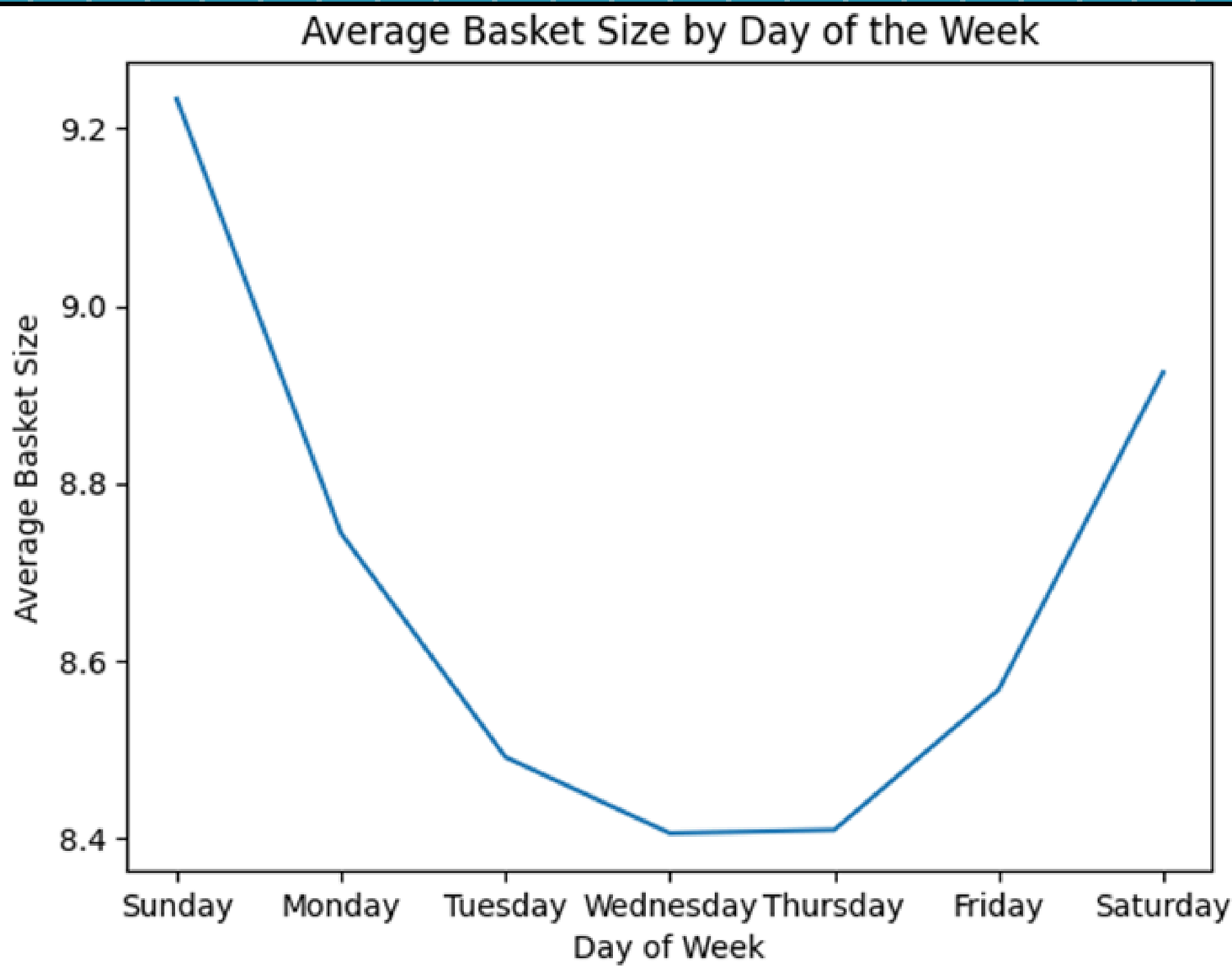
EXPLORATORY DATA AANALYSIS

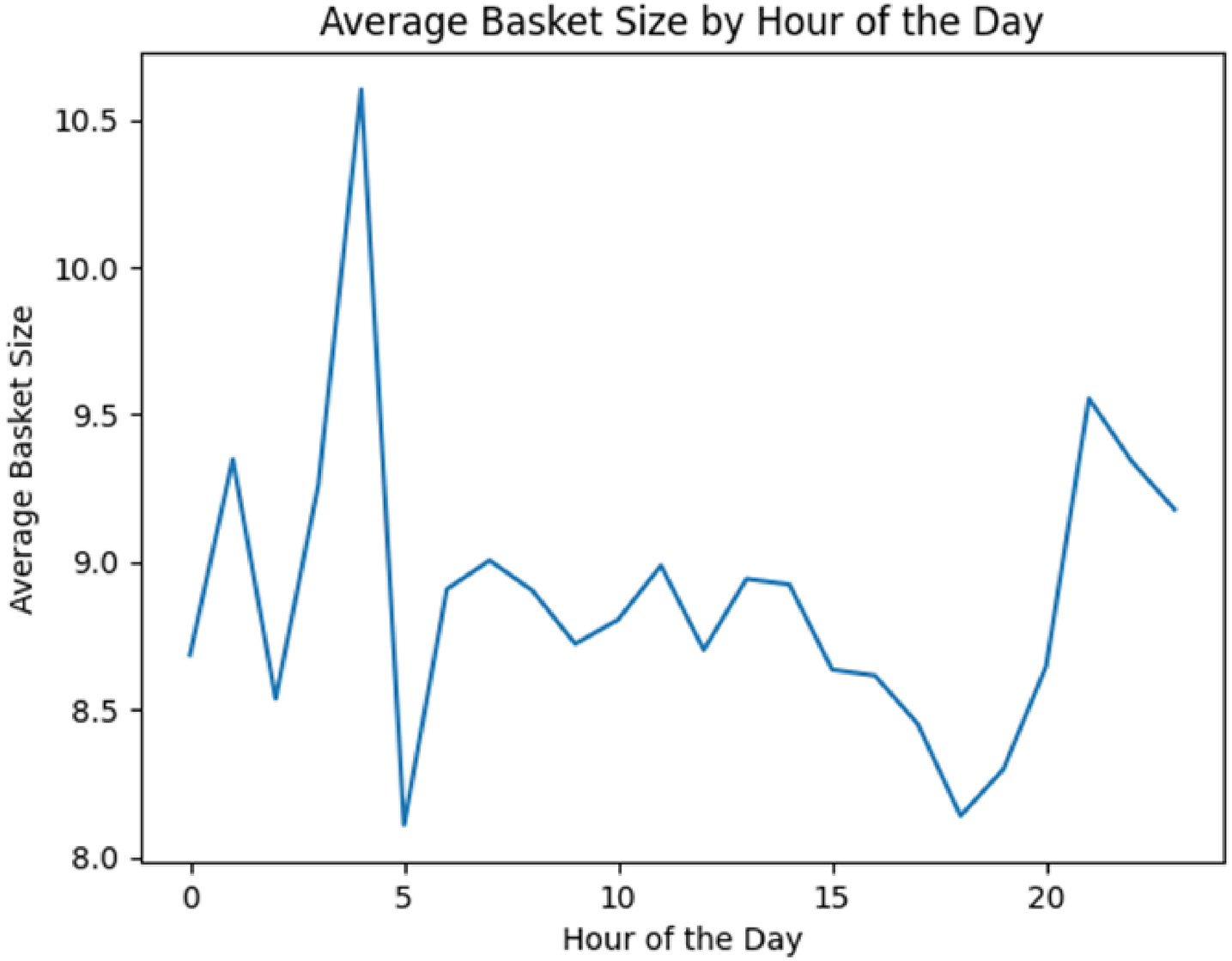

Top 10 Purchased Products
product_name
Banana
Bag of Organic Bananas
Organic Strawberries
Organic Baby Spinach
Large Lemon
Organic Avocado
Organic Hass Avocado
Strawberries
Limes
Organic Raspberries
0
2500
5000
7500
10000
12500
15000
17500


Frequency of Orders by Day of Week
Number of Orders
600000
500000
400000
300000
200000
100000
0
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week


Distribution of Orders by Hour of Day
300000
250000
200000
150000
100000
50000
0
Count
0
5
10
15
20
Hour of Day


Correlation Matrix of Numerical Variables
order_dow
order_hour_of_day
days_since_prior_order
add_to_cart_order
1.00
0.01
-0.03
-0.02
0.01
1.00
0.01
-0.01
-0.03
0.01
1.00
0.02
-0.02
-0.01
0.02
1.00
order_dow
order_hour_of_day
days_since_prior_order
add_to_cart_order
1.0
0.8
0.6
0.4
0.2
0.0


Average Basket Size by Day of the Week
Average Basket Size
9.2
9.0
8.8
8.6
8.4
Sunday
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Day of Week


Average Basket Size by Hour of the Day
Average Basket Size
10.5
10.0
9.5
9.0
8.5
8.0
0
5
10
15
20
Hour of the Day

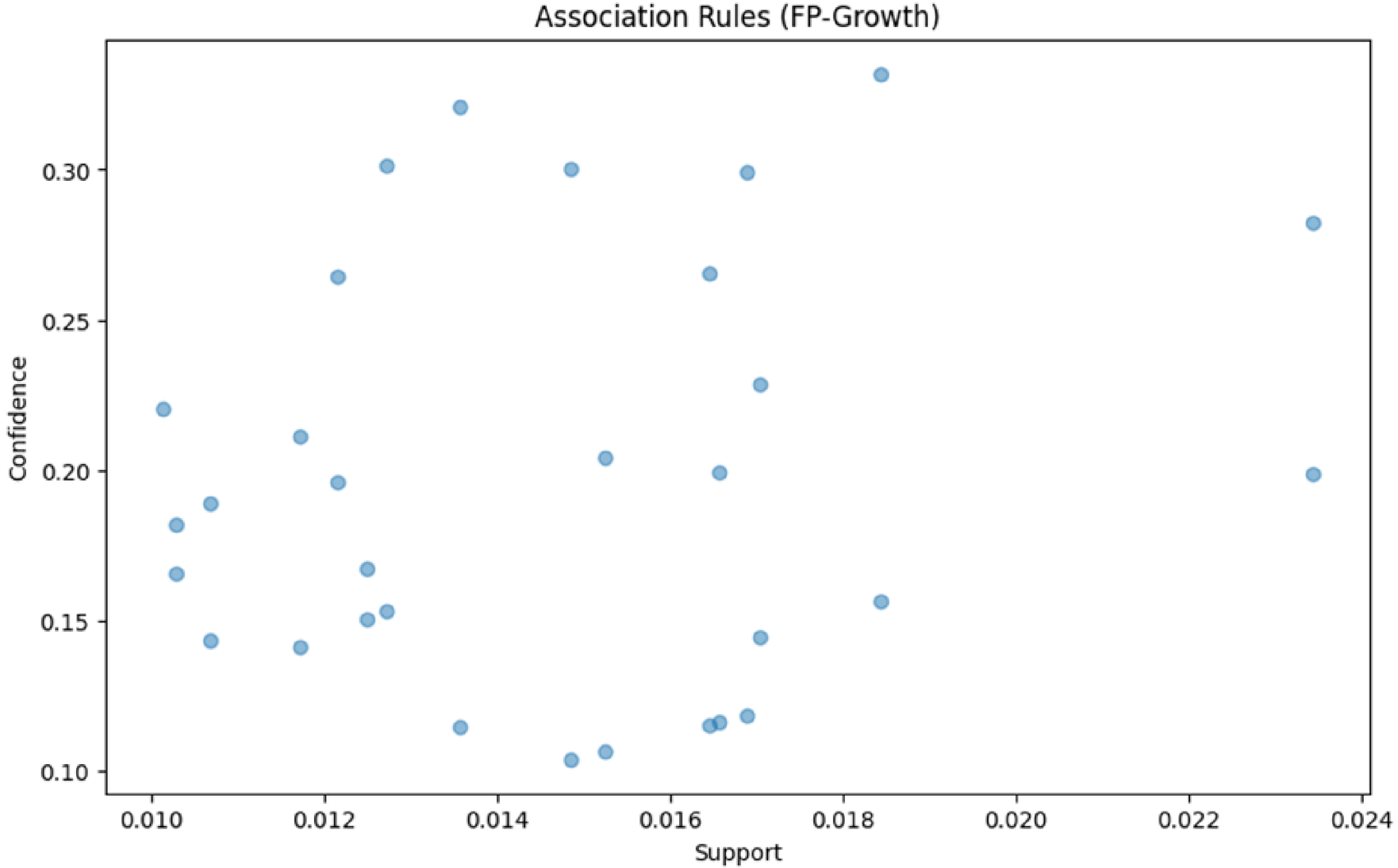
1.APRIORI ALGORITHM

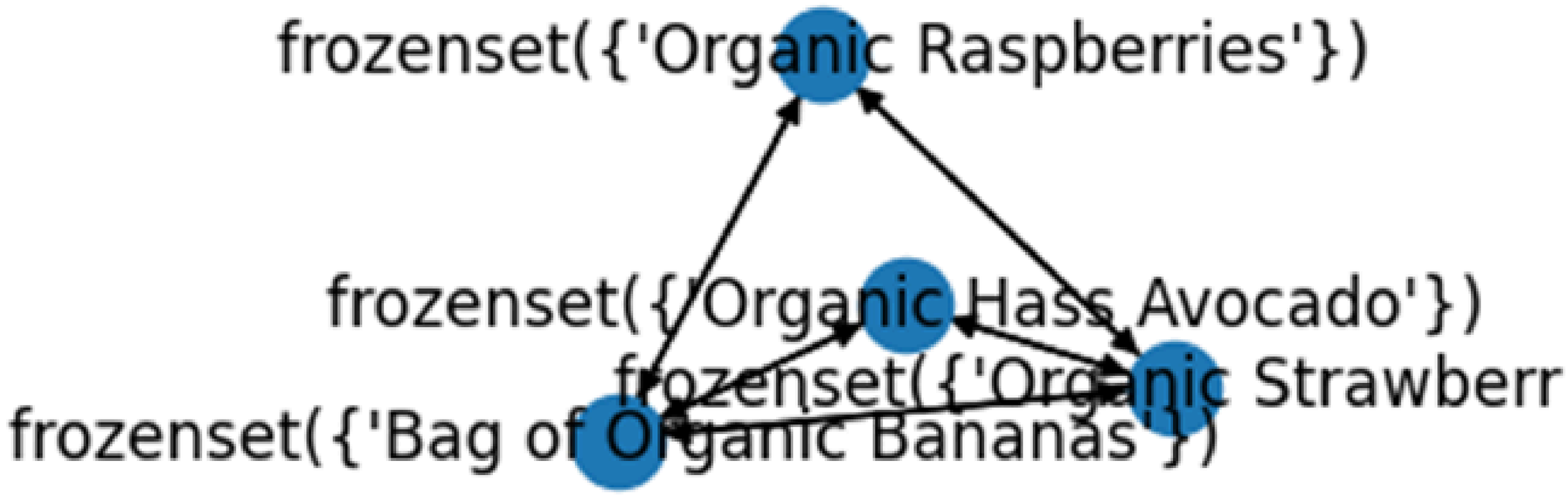
- Created a list of transactions grouped by order_id, transformed into a one-hot encoded DataFrame using a TransactionEncoder.
- Utilized the apriori function with a minimum support threshold to identify frequently appearing itemsets.
- Generated rules from itemsets using the association_rules function, focusing on 'confidence'.
- Used histograms to display the distribution of 'confidence' and 'lift', assessing the strength of item associations.


Distribution of Confidence
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
0.10
0.15
0.20
0.25
0.30
confidence

2.FP GROWTH

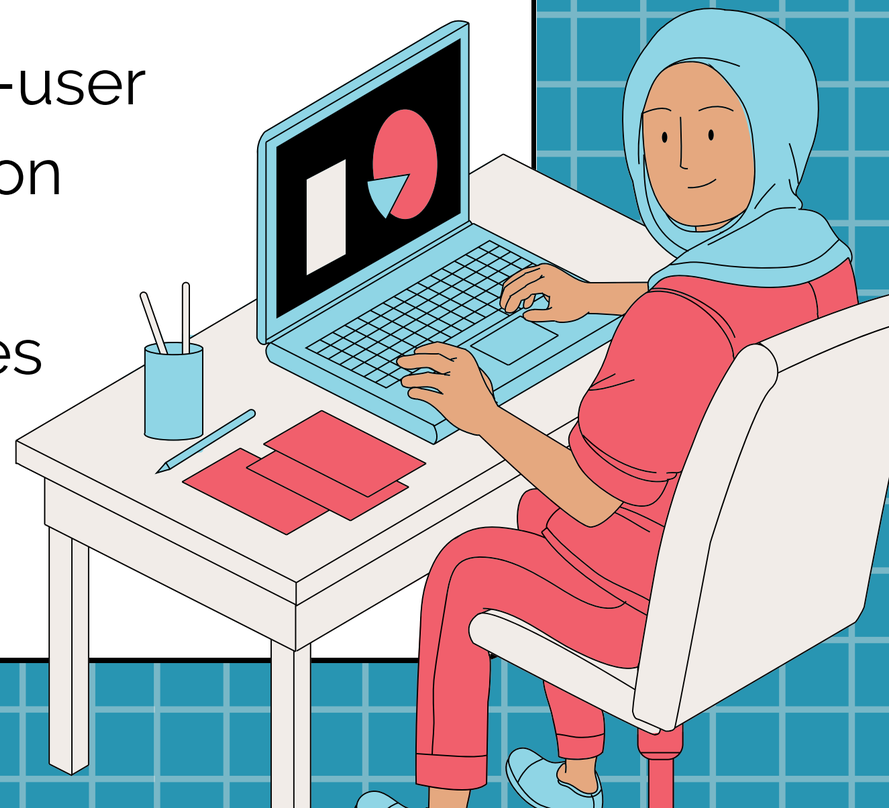
- FP-Growth Algorithm for Efficient Itemset Mining
- Used same transaction list and encoding as in Apriori.
- Applying the Algorithm:
- Employed the fpgrowth function to directly find frequent itemsets from the encoded DataFrame using a minimum support threshold.
- Generated rules from itemsets using the association_rules function, focusing on 'confidence' to identify strong associations efficiently.

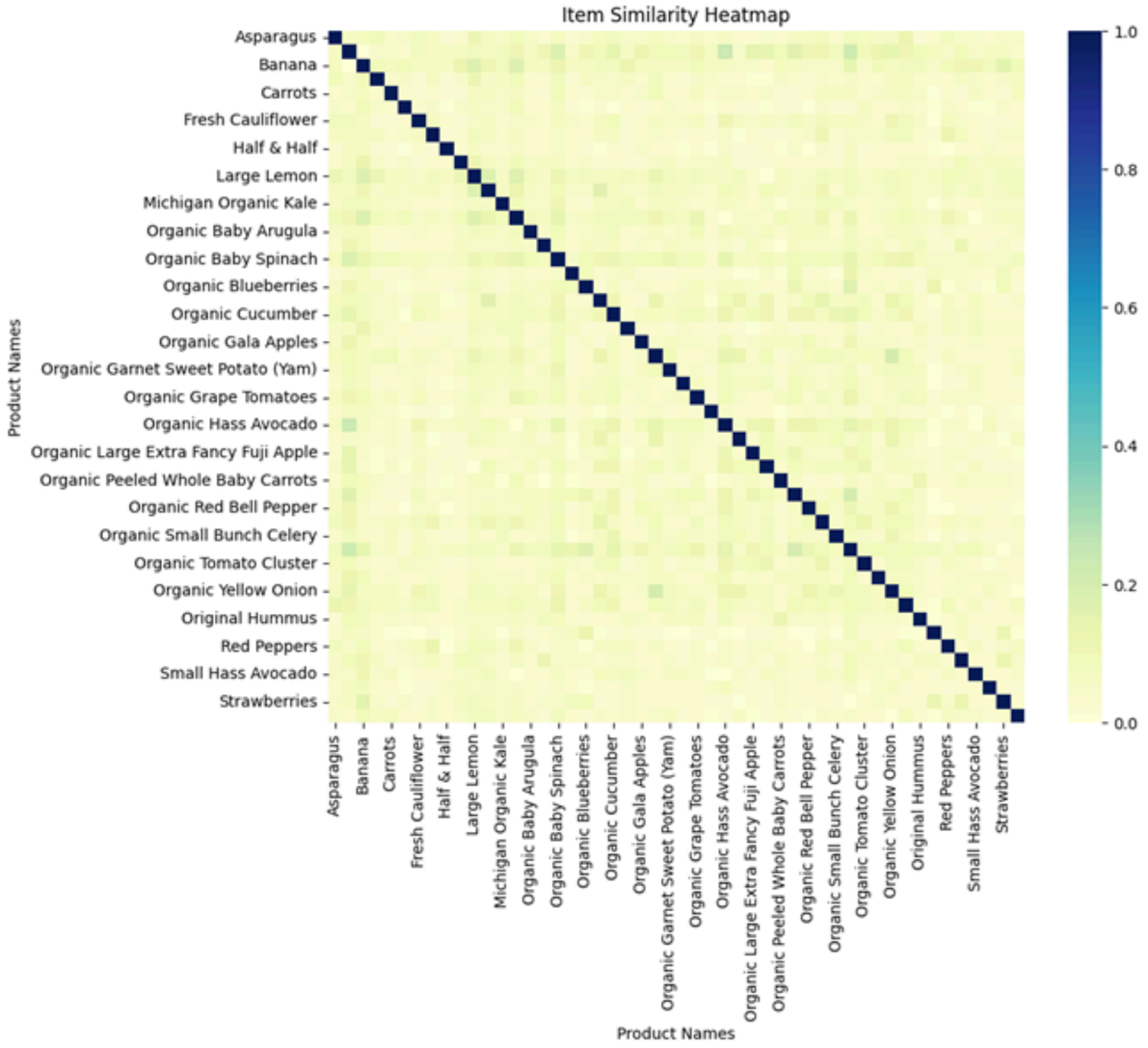

Association Rules (FP-Growth)
Confidence
0.30
0.25
0.20
0.15
0.10
0.010
0.012
0.014
0.016
0.018
0.020
0.022
0.024
Support


frozenset({'Organic Raspberries'})
frozenset({'Organic Hass Avocado'})
frozenset({'Organic Strawberr
frozenset({'Bag of Organic Bananas'})

3.RECOMMENDER SYSTEM

- Identified top 50 most frequently purchased products
- Created a pivot table with users as rows and products as columns
- Converted the item-user matrix to a sparse format using CSR
- Computed cosine similarity for the sparse item-user matrix to determine product similarities based on purchasing patterns.
- Used a heatmap to display item similarity scores among the top 50 items




Item Similarity Heatmap
Product Names
Asparagus
Banana
Carrots
Fresh Cauliflower
Half & Half
Large Lemon
Michigan Organic Kale
Organic Baby Arugula
Organic Baby Spinach
Organic Blueberries
Organic Cucumber
Organic Gala Apples
Organic Garnet Sweet Potato (Yam)
Organic Grape Tomatoes
Organic Hass Avocado
Organic Large Extra Fancy Fuji Apple
Organic Peeled Whole Baby Carrots
Organic Red Bell Pepper
Organic Small Bunch Celery
Organic Tomato Cluster
Organic Yellow Onion
Original Hummus
Red Peppers
Small Hass Avocado
Strawberries
Product Names
1.0
0.8
0.6
0.4
0.2
0.0

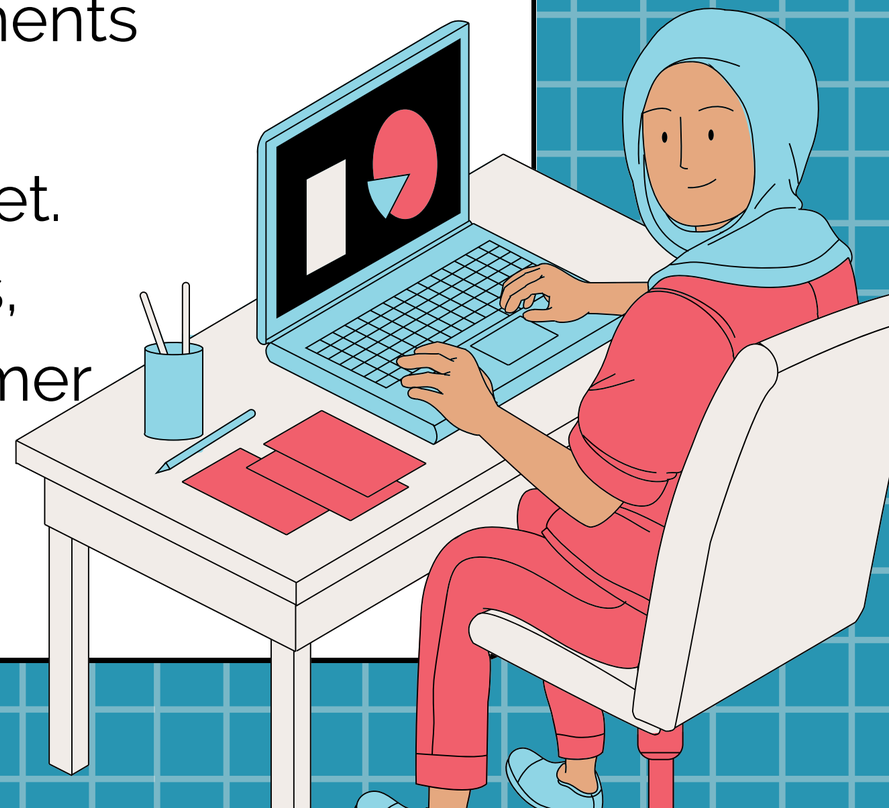
- A function calculates recommendations based on item similarity scores and user's purchasing history.
- For each product similar products from the item similarity is retrieved
- The top 'n' products with the highest aggregated similarity scores are recommended.
- The recommender system outputs a list of products that a user might be interested in,

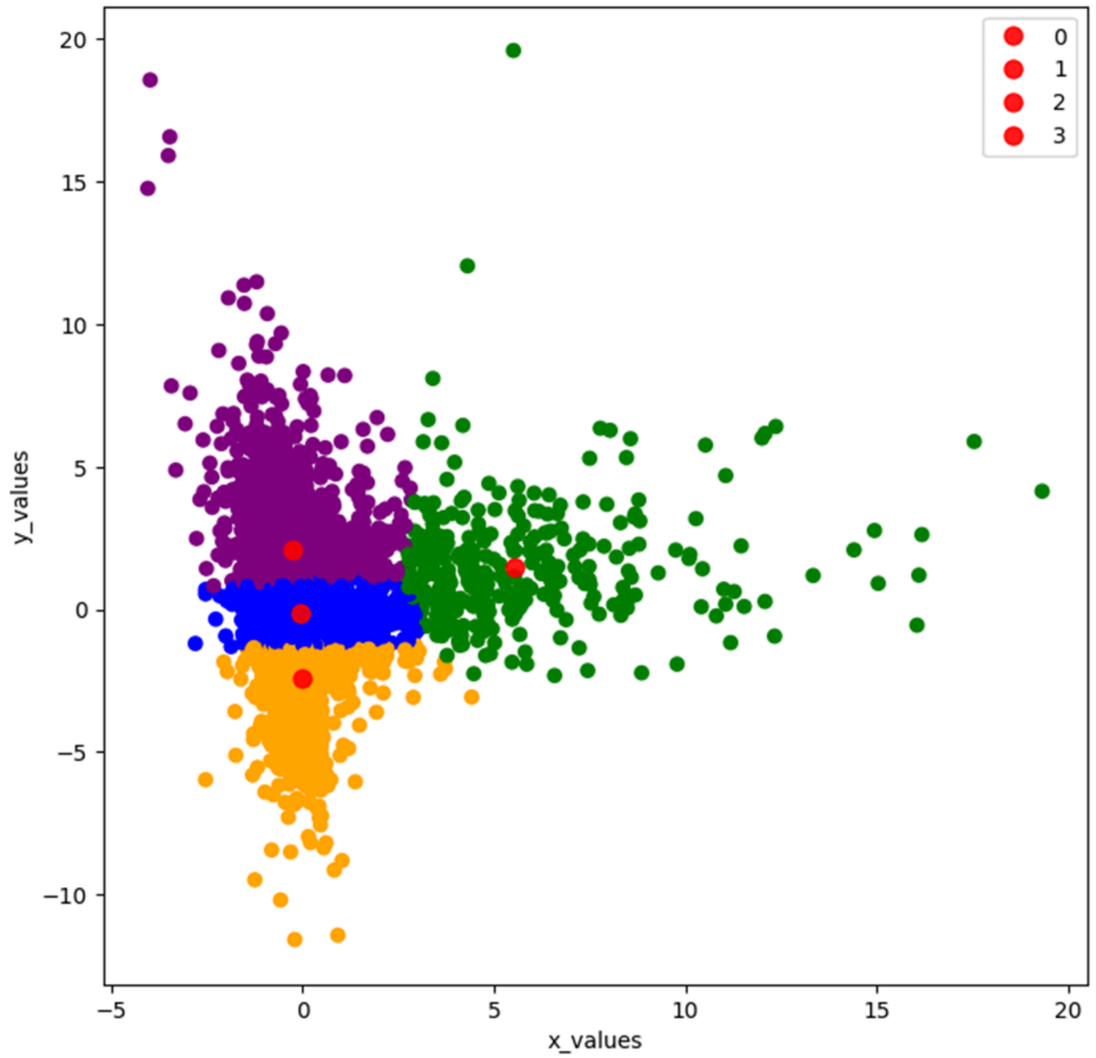
```python
# Example: Recommend items for user_id 264
print(recommend_items(264, item_similarity_df, reduced_data, top_n=5))
# Example: Recommend items for user_id 406
print(recommend_items(406, item_similarity_df, reduced_data, top_n=5))
```

```
['Large Lemon', 'Organic Baby Spinach', 'Organic Strawberries', 'Honeycrisp Apple', 'Limes']
['Organic Yellow Onion', 'Organic Red Onion', 'Organic Cilantro', 'Large Lemon', 'Red Peppers']
```

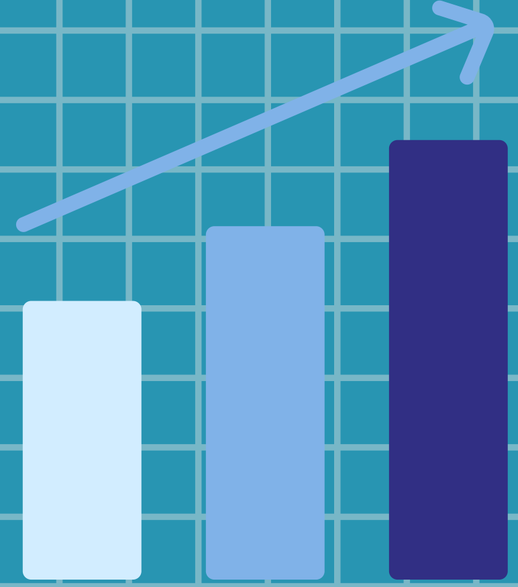
3.PCA & CLUSTERING

- Merged multiple data sources: user transactions, product details, and order information.
- Applying PCA:
  - Reduced the dataset's dimensionality
  - Transformed variables into principal components
- Clustering with K-Means:
- Applied K-Means to the PCA-transformed dataset.
- Clusters formed based on proximity to centroids,
- Examined each cluster to identify distinct customer segments.

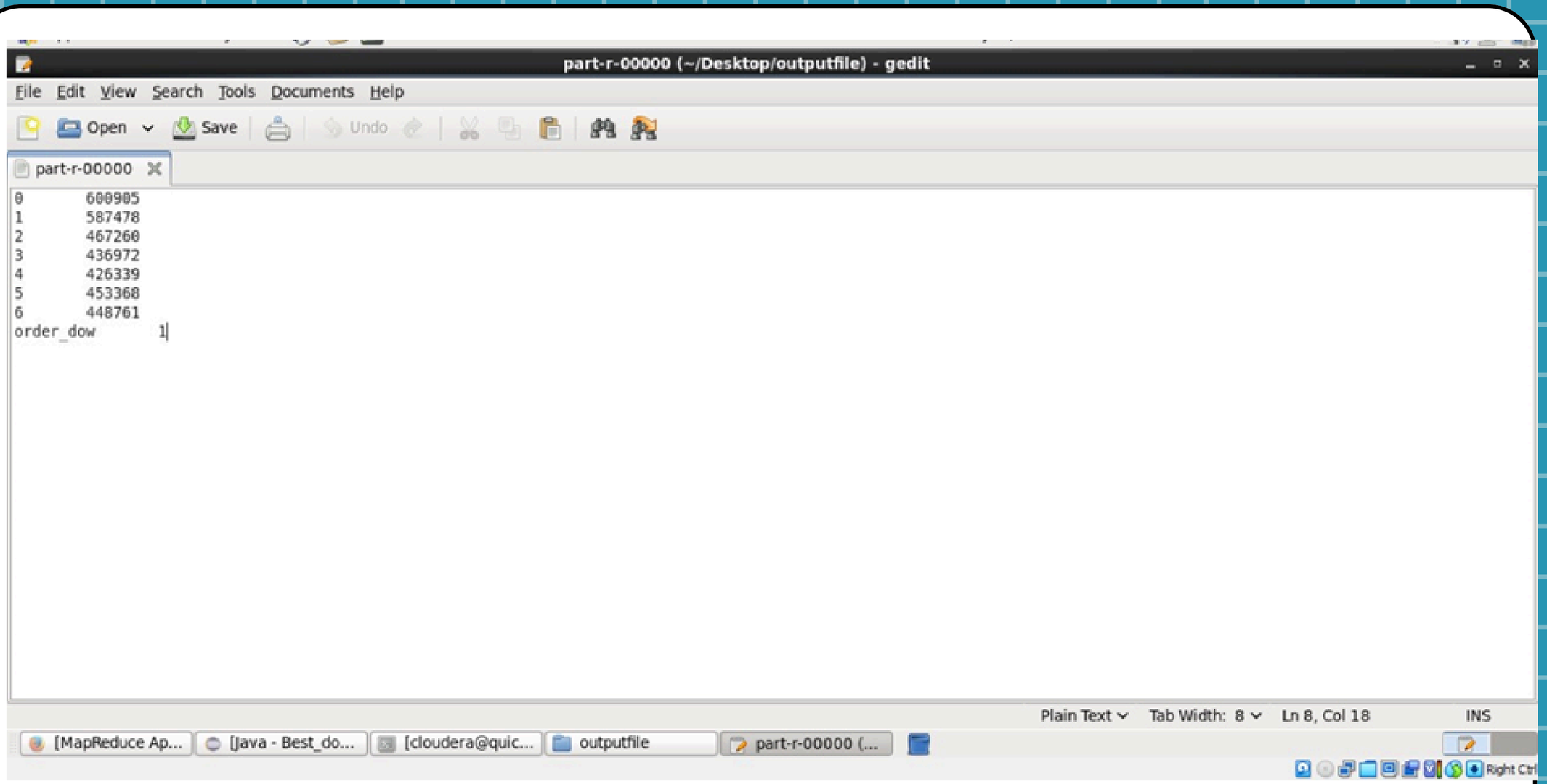



0
1
2
3
20
15
10
5
0
−5
−10
y_values
−5
0
5
10
15
20
x_values

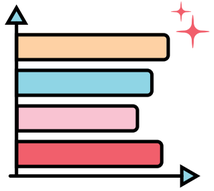
5. PAGE RANK (HADOOP)



- Mapper: Extracts day of the week from order data and emits it with a count of one.
- Reducer: Aggregates counts for each day to compute total orders.
- Hadoop Framework
- Configuration and job setuo
- Cloudera.
- Results Analysis:
  - Sunday (Day 0) has the highest number of orders.
  - Thursday (Day 4) has the lowest, indicating it's the least busy.

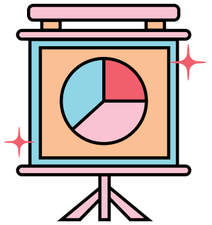

part-r-00000 (~/Desktop/outputfile) - gedit
File Edit View Search Tools Documents Help
Open
Save
Undo
part-r-00000
0 600905
1 587478
2 467260
3 436972
4 426339
5 453368
6 448761
order_dow 1
Plain Text
Tab Width: 8
Ln 8, Col 18
INS
[MapReduce Ap...
[Java - Best_do...
[cloudera@quic...
outputfile
part-r-00000 (...
Right Ctrl

PROBLEMS FACED



The effectiveness of PCA and clustering depends heavily on the quality and completeness of the data.

As the volume of data increases, the computational load also increases.



The recursive nature of the FP-Growth algorithm can lead to performance issues

FUTURE WORK

1.Incorporating hybrid approaches in your recommender systems

2.Enhancing your models by integrating more diverse datasets

3.Continuously refining and optimizing the existing algorithms

4.Developing interactive tools and dashboards that allow end-users to explore the data,

CONCLUSION

- Integration of Advanced Analytics
- Insights Gained
- Continual adaptation to the latest technological .
- Business Impact:
  - Provides actionable insights.
  - Supports data-driven decision-making processes.

THANK YOU

ARE THERE ANY QUESTIONS?