

**PROJECT**

**Market Basket Analysis for Customer Purchase Insights**

*Project Report submitted on the fulfilment of the requirements of Post graduate Diploma in Big data Analytics*

**PG-DBDA September 2022**

**TEAM NO. - 6**

|  |  |
| --- | --- |
| **AUTHORS:** | **CO-ORDINATOR:** |
| PRADYUMN VIBHANDIK 230340325032 | MS. VINEETA SINGH (Course Coordinator) |
| PRADYUMNA PATHADE 230340325033 | MR. VISHAL KHETMALIS (Project guide) |
| PRANAL TAYRE (PL) 230340325034 |  |
| PRANAY PARIHAR (CL) 230340325035 |  |
| PRASAD AMLE 230340325036 |  |
| PRASAD SHELAR (TL) 230340325037 |  |

**INDEX**

|  |  |  |
| --- | --- | --- |
| SR. No. | TOPIC | PAGE No. |
| 1 | ABSTRACT | 2 |
| 2 | INTRODUCTOION | 3 |
| 1.1 | PROBLEM STATEMENT | 4 |
| 1.2 | OBJECTIVE | 4 |
| 2 | APPROACH | 5 |
| 2.1 | TECHNICAL APPROACH | 6 |
| 2.2 | DATA COLLECTION AND EXPLORATION | 6 |
| 2.3 | PROJECT SETUP: | 6 |
| 2.4 | TECHNOLOGY STACK | 6 |
| 3 | DATA DESCRIPTION | 9 |
| 4 | DATA PREPARATION | 12 |
| 4.1 | DATA CLEANING | 13 |
| 4.2 | DATA ANALYSIS | 13 |
| 4.3 | SUMMARY STATISTICS | 15 |
| 5 | STRATEGY FOR BASKET ANALYSIS | 19 |
| 5.1 | KEY TERMS AND CONCEPTS | 19 |
| 6 | CLOUD PLATFORM | 25 |
| 7 | PIPELINE CONNECTIVITY | 26 |
| 8 | CONCLUSION AND FUTURE WORK | 27 |

**ABSTRACT**

Market basket analysis is one of the most prevalent and effective aspects of information analysis for selling and marketing. Market basket analysis is a data processing technique used by merchants to understand their customer's shopping behaviour, such as which items customers are inclined to buy together from their stores, which can assist the distributor in making the right decisions. It operates by seeking a variety of products that appear in exchange now and again. Effective analysis can increase a retailer's profitability, service quality, and customer satisfaction. Enterprises collect and store massive amounts of data as information technology continues to develop. In changing markets, businesses must translate this data into valuable information and knowledge for decision-making. Market basket analysis provides value-added information that can be utilized to enhance decision-making

To facilitate reordering and maintaining adequate product stock, this paper will demonstrate how Instacart can use its customer transaction data. It will concentrate on descriptive analysis of customer purchase patterns, items that are frequently purchased together, and units that are frequently purchased from the store. Additionally, to locate client subgroups and clusters with similar purchasing habits. Furthermore, the data should be visualized in order to make useful recommendations for enhancing revenue and customer experience via segmentation and prediction models.

Our dataset includes variables relating to orders as well as order timing. As a result, order-related and time-based features were developed to predict whether or not a product will be reordered. This paper will enable Instacart to enhance the user experience by suggesting which products should be placed to be collected. During the ordering process, it will suggest products based on past orders

**1.INTRODUCTION**

In the modern digital age, Terabytes of commercial data are produced every second. The volume of data is significantly growing as a result of the enormous volumes of data that are produced in daily activities. Data management and mining communities are facing one of their biggest problems yet: extracting insight from the data that is being generated at an exponential rate. Furthermore, most well-known companies gather and keep a ton of information about customer transactions. However, just because an organization has a lot of data does not necessarily indicate it has valuable commercial information. This massive amount of data must produce valuable knowledge and information for the business sectors. This results in a market basket analysis.

By identifying relationships between various things that customers place in their shopping baskets, this technique determines the purchasing patterns of those customers. The goal of market basket analysis is to determine the products that buyers usually buy together. The analysis helps business owners in making a variety of crucial business decisions, including how to design a catalogue, recognize regular customers, and increase sales of products. The primary objective of market basket analysis is to identify relationships between the things people buy. By grouping like products together, businesses can better arrange their product placement on aisles.

People nowadays prefer to buy everything online, and supermarkets have recently entered that sector, where Instacart is gaining popularity. Because consumers are busy and have other responsibilities, they do not like to spend lots of time on an application or website. This is why they shop online for groceries.

The main objective of the project is to assist Instacart vendors in better comprehending the purchase patterns of their current consumers. Retailers must first analyse current consumer activity to predict potential customer buying behaviour. To retain consumers, increase revenue, and deepen customer relationships, customer transaction data may be utilized to analyse and visualize customer purchase behaviour, assortment planning, and inventory management.

* Examine the customer's purchasing patterns
* What are the products that customers are most likely to purchase together?
* Find a group of shoppers who make similar purchases.

Instacart market basket analysis will examine a dataset of Instacart and its various attributes. Eventually, it will help Instacart vendors to better understand their current customers' purchasing behaviours. The Instacart market basket analysis is significant because it creates a truly positive user experience by enabling users to discover frequently paired items. This reduces search time and facilitates quicker checkout. Additionally, it would boost vendor revenue in addition to the consumer experience.

**1.1. PROBLEM STATEMENT**

Nowadays people buy daily goods from super market nearby. There are many supermarkets that provide goods to their customer. The problem many retailers face is the placement of the items. They are unaware of the purchasing habits of the customer so they don’t know which items should be placed together in their store. With the help of this application shop managers can determine the strong relationships between the items which ultimately helps them to put products that co-occur together close to one another. Also, decisions like which item to stock more, cross selling, up selling, store shelf arrangement are determined

**1.2. OBJECTIVE**

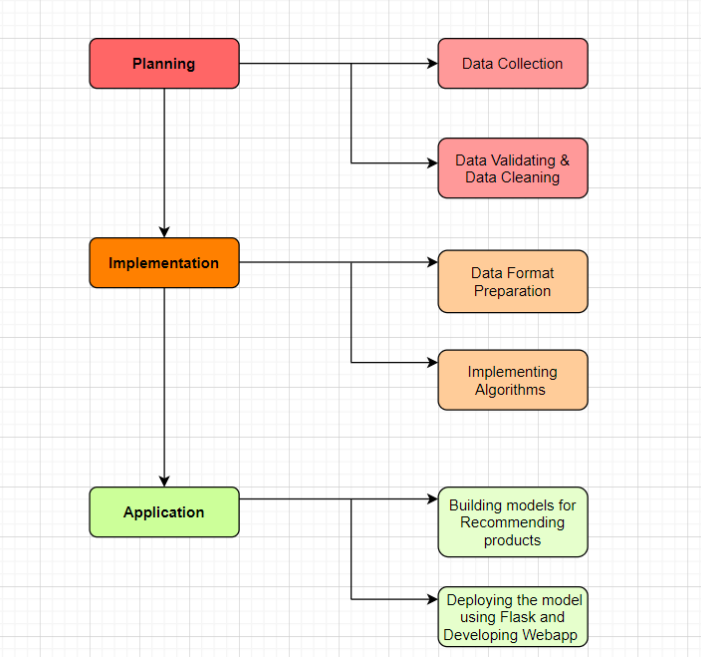
The main objective of the project is to make Instacart retailers to understand the current customer's behaviour and to predict future customers’ purchasing behaviour. Leveraging customer transaction data can help in understanding customers’ purchasing behaviour, offering right bundles and promotions, assortment planning and inventory management to retain customers, improve sales and extend their relationship with customers.

1. To identify the frequent items from the transaction on the basis of support and confidence
2. To generate the association rule from the frequent item sets.
3. To understand the purchasing pattern of products that comprise the customers’ basket.
4. To study the most likely products purchased by the customers along with a particular product category.
5. To recommend and suggest products to individual customers.

**2. APPROACH**

In today's customer-centred markets, businesses must develop adequate and low-cost advertising strategies that can react to changes in buyer impressions and product demands. To make crucial decisions on market strategy, you need reliable, concrete evidence data. Currently, database technology innovation has grown to the point where these information stacks are robust, hence this information enhances the value of an organization. The approach may also allow businesses to develop a completely new market approach that can be effectively addressed. Due to innovation, data mining has provided the best solution to this requirement. Data mining is the process of extracting relevant information from large databases using a wide range of statistical and computational techniques such as clustering, classification, and summarization.

Market basket analysis can be performed using a variety of algorithms. To develop a comprehensive understanding of the market basket, I decided to analyse product categories using Apriori algorithms and data mining techniques. I also decided to analyse association rule mining, which is utilized by market basket research to determine consumer purchasing behaviour and assist in revenue growth.



**FIG.1**

**2.1. TECHNICAL APPROACH:**

As part of this study, the first step is data collection followed by data preparation, data cleaning, and data manipulation in order to prepare it for EDA (Exploratory Data Analysis).

**2.2. DATA COLLECTION AND EXPLORATION:**

To undertake analysis, we gathered datasets from Kaggle, which was provided by the Instacart Company. Over 3 million purchases from over 200000 Instacart consumers are included in the databases. 50,000 unique products, multiple product aisles, departments, week, and time of purchase are included in both product and consumer data.

**2.3. PROJECT SETUP:**

After collecting data from the Kaggle platform, the next step is to set up a project to analyse the data for various operations. The primary language employed for the project is Python. It was primarily chosen because of Python's widespread use in data science projects and its simplicity of comprehend.

**2.4. TECHNOLOGY STACK:**

This section discusses the technology leveraged in building this project. The project uses various libraries such as Pandas, NumPy, Seaborn, Matplotlib etc. and the project also extensively developed on Jupyter Notebook and Databricks and used Azure Blob Storage as a storage backend.

1. **Jupyter Notebook:**

A popular tool for creating computation notebooks with visualization is Jupyter Notebook. In addition to ipynb, Jupyter Lab is an enhancement because it offers terminals, file viewers, and custom components.

1. **Databricks:**

Databricks is basically a cloud-based engineering tool that provides a simple collaborative environment to run interactive and scheduled data analysis workloads.

1. **AWS S3:**

Amazon S3 has a simple web services interface that you can use to store and retrieve any amount of data, at any time, from anywhere on the web. As part of the project, S3 was used to store CSV files of data.

1. **Azure Blob Storage:**

Azure Blob Storage is Microsoft's cloud object storage solution within the Microsoft Azure cloud platform. It offers features similar to Amazon S3 for storing and managing various types of data.

1. **PySpark:**

PySpark is a Python library that enables seamless integration with Apache Spark. Apache Spark is a powerful distributed computing framework designed for processing and analyzing large datasets across clusters of computers. PySpark combines the simplicity and expressiveness of Python with the performance and scalability of Spark.

1. **Spark SQL:**

Spark SQL is a module within Apache Spark that provides a programming interface for querying structured data using SQL (Structured Query Language). It seamlessly integrates with Spark's other components, such as Spark Core, Spark Streaming, and MLlib, enabling SQL queries to be performed on distributed data stored in various formats.

1. **Pandas:**

Pandas is an open source, Python licensed library that offers high-performance, easy-to-use data structures, and data analysis tools to the Python programming language. The Data Frame is the core data structure. Data frame allows tabular data to be stored and manipulated in observation rows and variable columns. A broad variety of stored data types is available, such as CSVs, TSV's (Tab separated values), JSONs (Hypertext Mark-Up Language), and more. Pandas can read different types. A Data Frame consists of both a row and a column index, a two-dimensional set of values. A series is a special collection of index values.

In our project, we have converted dataset CSV files to data frames:

* + - 1. aisles\_df
      2. departments\_df
      3. orders\_df
      4. order\_products\_prior\_df
      5. order\_products\_train\_df

1. **Numpy:**

NumPy is an array-processing application for general purposes. It stands for 'Numerical Python'. It is a library of multidimensional array objects, and a set of array processing routines. NumPy has functions built in for linear algebra and the generation of random numbers.

1. **Matplotlib:**

Is the art of displaying data through charts, icons, presentations and more. It is most common to translate complex data for a non-technical audience into comprehensible insights. Matplotlib is one of the most powerful Data Visualization Python packages used. This is a cross - platform framework designed to make Two dimensional graphs from records in arrays. This also provides an object-oriented API which helps, for example, to embed plots into implementations using Python GUI toolkits such as PyQt.

1. **Seaborn:**

Seaborn is an enhancement to matplotlib and not a substitution for it. The reason for this is that it is placed on top of matplotlib and you will often explicitly invoke matplotlib functions to draw simpler plots already available through the namespace pyplot. Matplotlib is completely scalable but it can be difficult to know what settings to change to achieve an appealing plot. Seaborn comes with a number of custom themes to track the matplotlib look and a high-level user interface. It is closely integrated with the PyData stack, including support of SciPy and stats models data structures for NumPy and Pandas, and statistical routines.

1. **MLxtend:**

MLxtend is a library which implements a range of core machine learning and data mining algorithms and utilities. The primary goal of MLxtend is to create widely used tools to focus solely consistency with existing machine learning libraries on user-friendly and intuitive APIs. While MLxtend enforces a wide range of functions, highlights include sequential selection feature algorithms, stacked generalization implementations for classification and regression, and frequent pattern mining algorithms. MLxtend offers a variety of utilities that draw on Python 's scientific computing stack and increasing its capabilities.

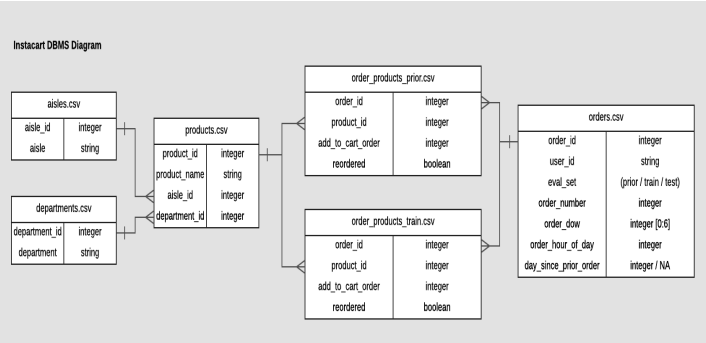
1. **Apriori:**

Apriori is an association rule mining algorithm used to find frequent itemsets in transactional datasets. It operates on the principle of the "apriori property," which states that if an itemset is frequent, then all of its subsets must also be frequent. The algorithm works through multiple iterations, gradually increasing the length of itemsets it searches for.

1. **FPGrowth:**

FP-Growth (Frequent Pattern Growth) is another association rule mining algorithm that aims to find frequent itemsets more efficiently than Apriori. It uses a tree-based data structure called the FP-tree to store the dataset and perform mining.

**3. DATA DESCRIPTION**

****

This section discusses details about the dataset provided by Instacart. A total of 6 datasets were provided which gave information on customer transactional and purchasing order details.

* **Section 1: Aisles**

Data provided here includes information about aisles, such as aisle names and aisle IDs, and how products were organized within them.

|  |  |
| --- | --- |
| Variable | Description |
| aisle\_id | aisle identifier |
| aisle | The name of the aisle in the retail store |

Table 1 - Details of Aisles data set

* **Section 2: Department**

This dataset provides information on the department such as department names and department ID

|  |  |
| --- | --- |
| Variable | Description |
| department\_id | Department identifier |
| department | Name of the department in retail store |

Table 2 - Details of Department data set

* **Section 3 : Order\_Products\_prior/Order\_Products\_train**

In this dataset, all order details for any prior orders are included, including information on orders, products, and reordered items.

Order\_Products\_train is the same as Order\_Products\_prior and it is trained dataset.

|  |  |
| --- | --- |
| Variable | Description |
| order\_id | Order identifier |
| product\_id | Product identifier |
| add\_to\_cart\_order | Order in which product added to cart |
| reordered | 1 if product was ordered by user in past else 0 |

Table 3 - Details of Order\_Products\_prior/Order\_Products\_train data set

* **Section 4 : Orders**

This dataset has information about customer orders like order ID, order number, weekday of the order, an hour of the order, user ID, and days since the prior order.

|  |  |
| --- | --- |
| Variable | Description |
| order\_id | Order identifier |
| user\_id | Customer identifier |
| eval\_set | Which evaluation set this order belongs in(prior/train/test) |
| order\_number | Sequence of the order placed by user |
| order\_dow | The day of week the order was placed |
| orders\_hours\_of\_day | The hour of the day the order was placed |
| day\_since\_prior\_order | Number of days since last order(NAs for order\_number = 1) |

Table 4 - Details of Orders data set

* **Section 5 : Products**

This dataset gives information on the products which were sold to customer such as product name, product ID, aisle and departments.

|  |  |
| --- | --- |
| Variable | Description |
| product\_id | Product identifier |
| product\_name | Name of the product purchased by customer |
| aisle\_id | aisle identifier |
| department\_id | department identifier |

Table 5 - Details of products data set

**4. DATA PREPARATION**

Data preparation is necessary for data analysis after data gathering. Six different datasets provided information about customer purchases and transaction details. Order\_products\_train and Order\_products\_prior were merged into a single data set based on order\_id and product\_id respectively. Later, the department, aisles, and orders datasets were merged with the Order\_products\_train and Order\_products\_prior combined dataset through department\_id, aisles\_id, and orders\_id. This created a master dataset to begin the analysis.

Figure 1 shows the final master dataset attributes.

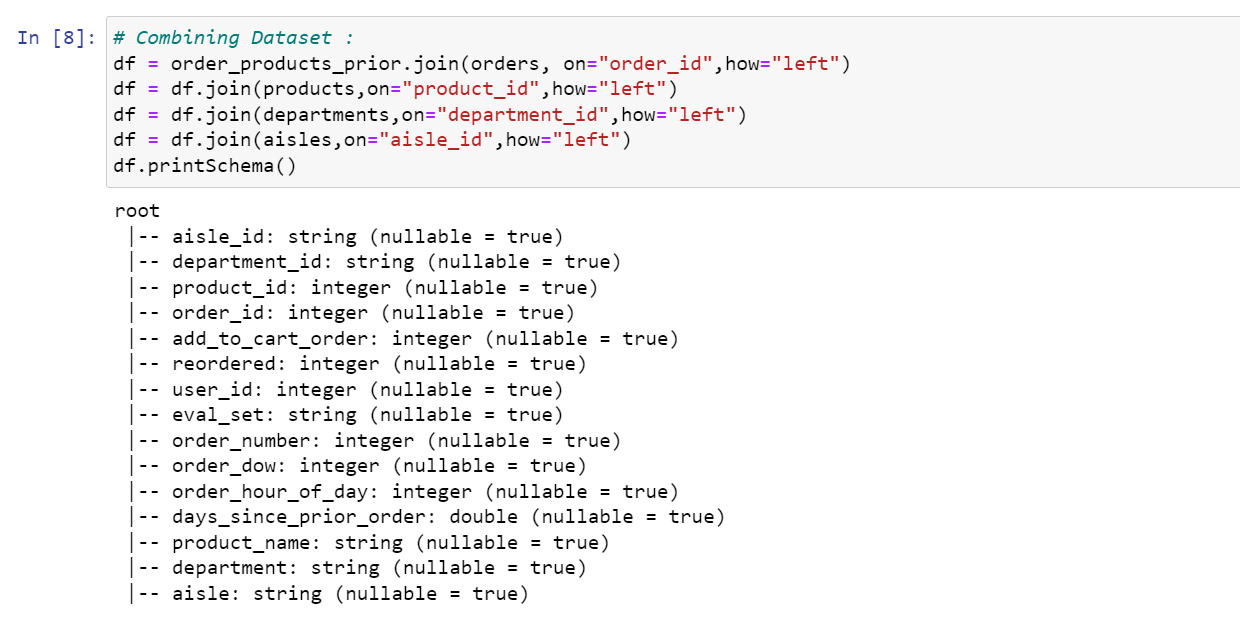


Figure 1 - Final dataset attributes

Along with exploring the length and attributes of a dataset, it was also checked for null data. When the dataset was computed using pandas there are only 2078068 missing values among the many 33819106 values. That computes around 5 percent of data as missing which is low enough of a threshold to be eliminated.

Figure 2 shows total attributes and null values in final dataset.

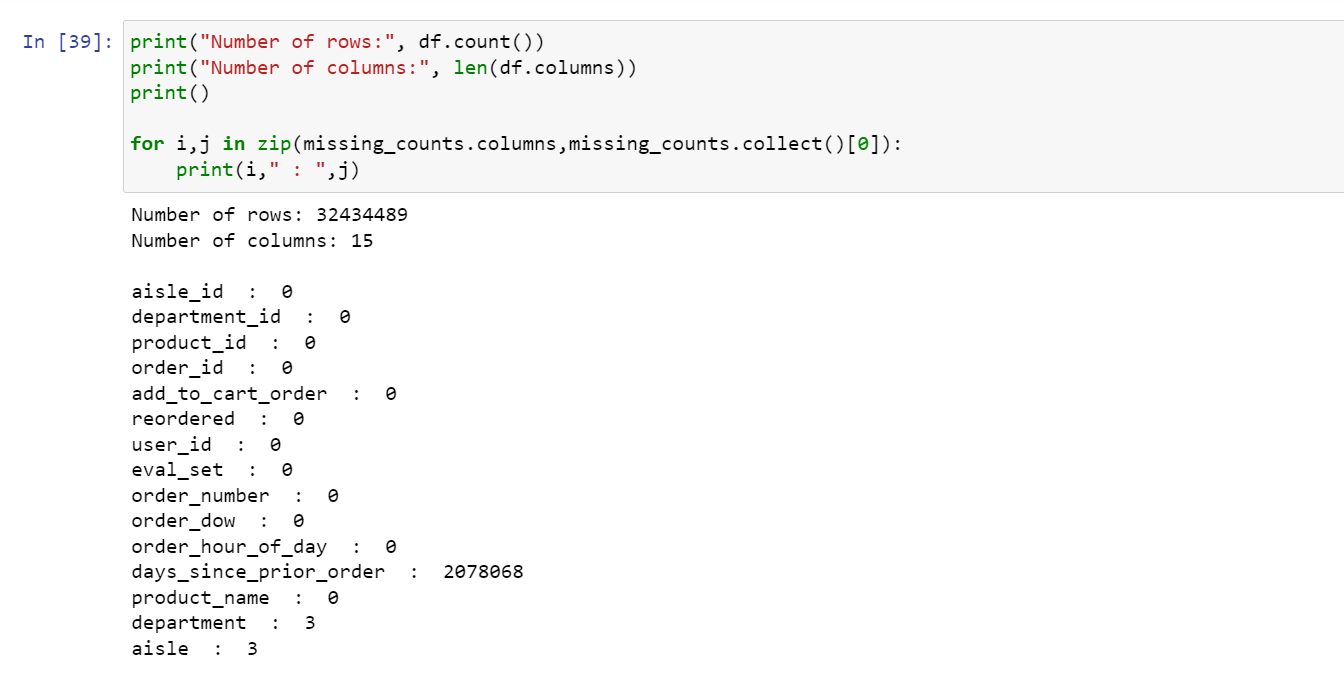


Figure 2 - Null values in dataset

Figure 2 shows that the data, which is acceptable for modelling because of its count and low percentage of null values, is of good quality.

**4.1. DATA CLEANING**

According to the previous section, the data has around 5 percent of null values with missing attributes. These values have been rejected. This is to ensure the data is used for modelling and to ensure that results are accurate. Since the percentage of null values is so low, the null values are removed, reducing the data size to 30356418 from 32434489.

Figure 3 show how the data is cleaned and data size after dropping null values.

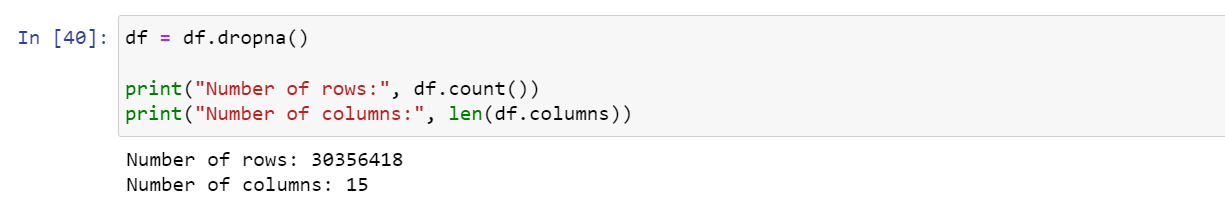


Figure 3 - Dropping null values

**4.2. DATA ANALYSIS**

Following data preparation and cleaning, descriptive analysis of the dataset is important to determine the number of distinct aisles, departments, orders, and customers.

Figure 4 shows that there are 134 distinct aisles in all, including five for prepared soups, salads, specialty cheeses, energy granola bars, rapid snacks, and meat marinades.

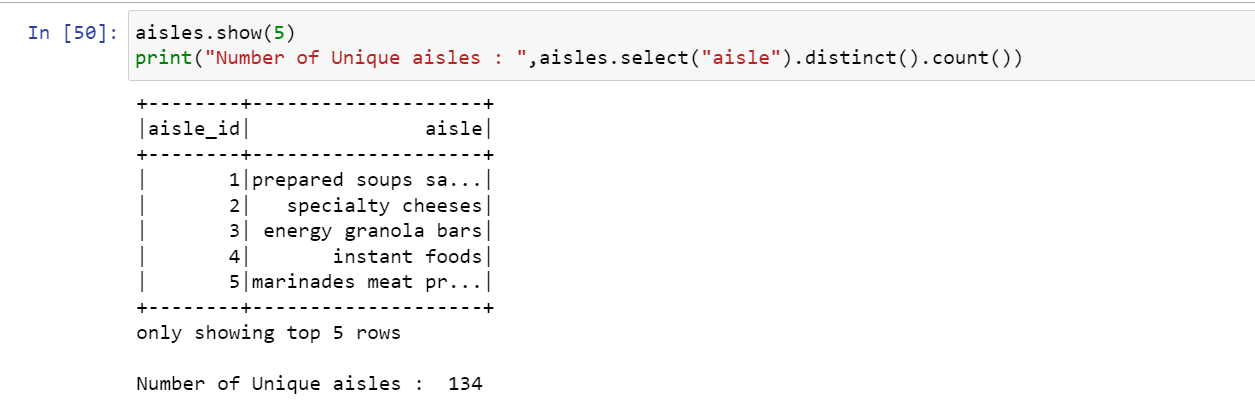
****

Figure 4 - Aisles attributes

Figure 5 shows that there are 21 distinct departments in all, including five for frozen, bakery, produce, alcohol and others.

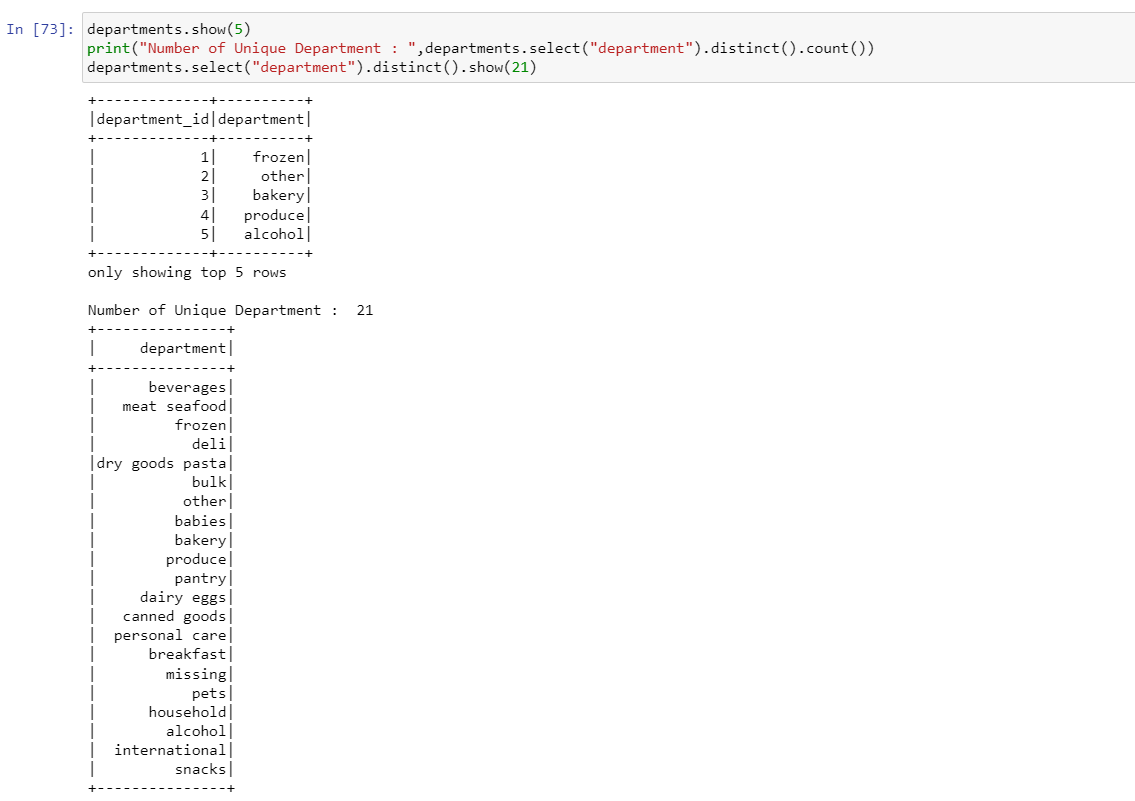
****

Figure 5 - Department attributes

Figure 6 shows that there are 49688 distinct products, including Chocolate Sandwich Cookies, All-Seasons Salt, Robust Golden Unsweetened Oolong Tea, Artisan Baguettes, etc

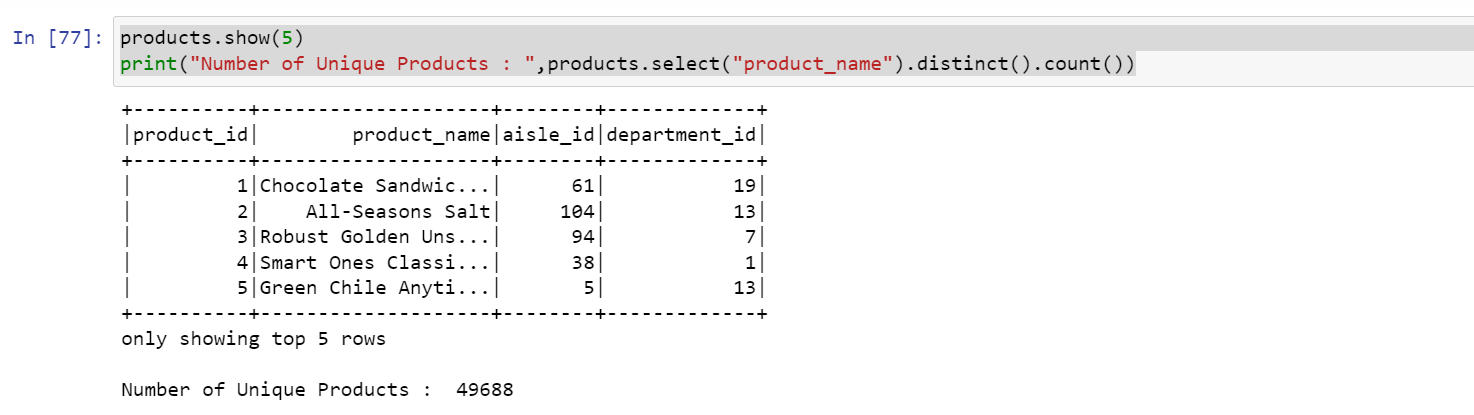
****

Figure 6 - Product attributes

Figure 7 shows that there is total 3421083 orders made by total 206209 users.

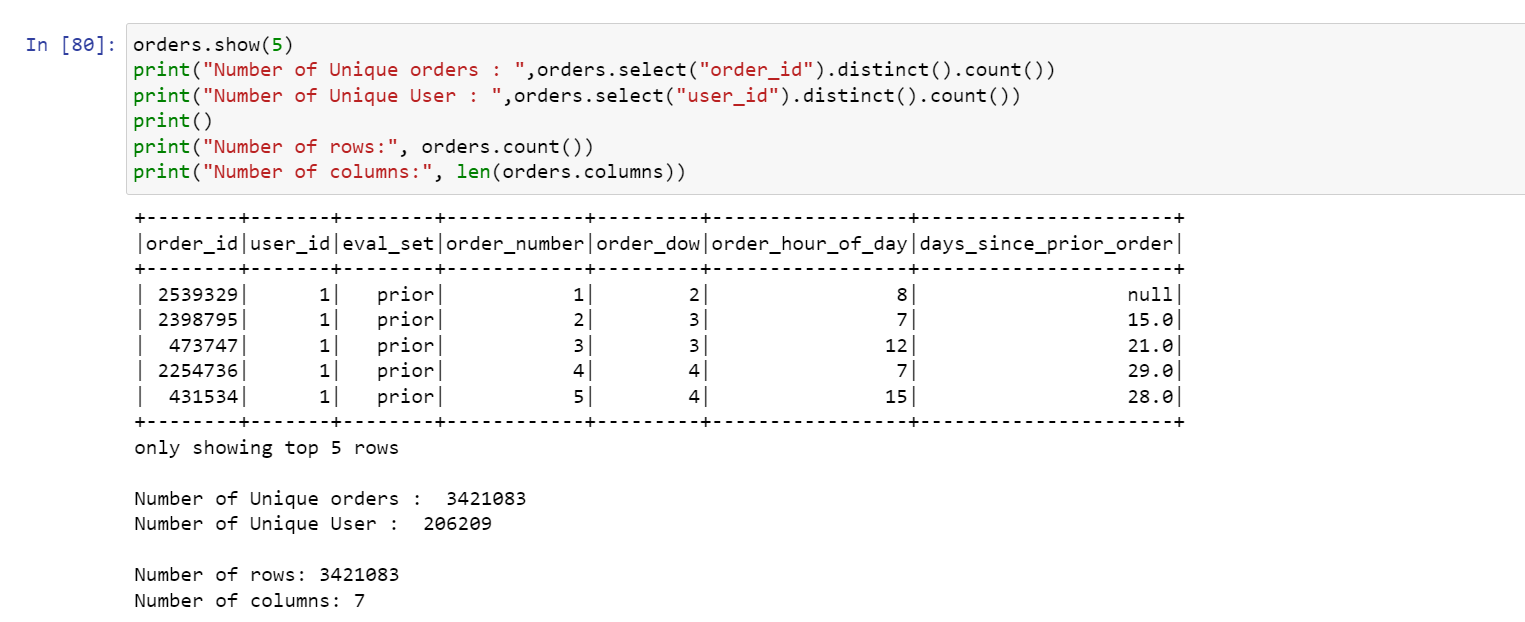
****

Figure 7 - Orders attributes

Figure 8 displays a dataset with descriptive statistics and an order summary. There are three sets of orders, such as prior, train, and test, as shown in the figure. While the distribution of orders in the train and test sets is comparable, it is different for the prior set, and the total number of orders per customer ranges from 1 to 100.

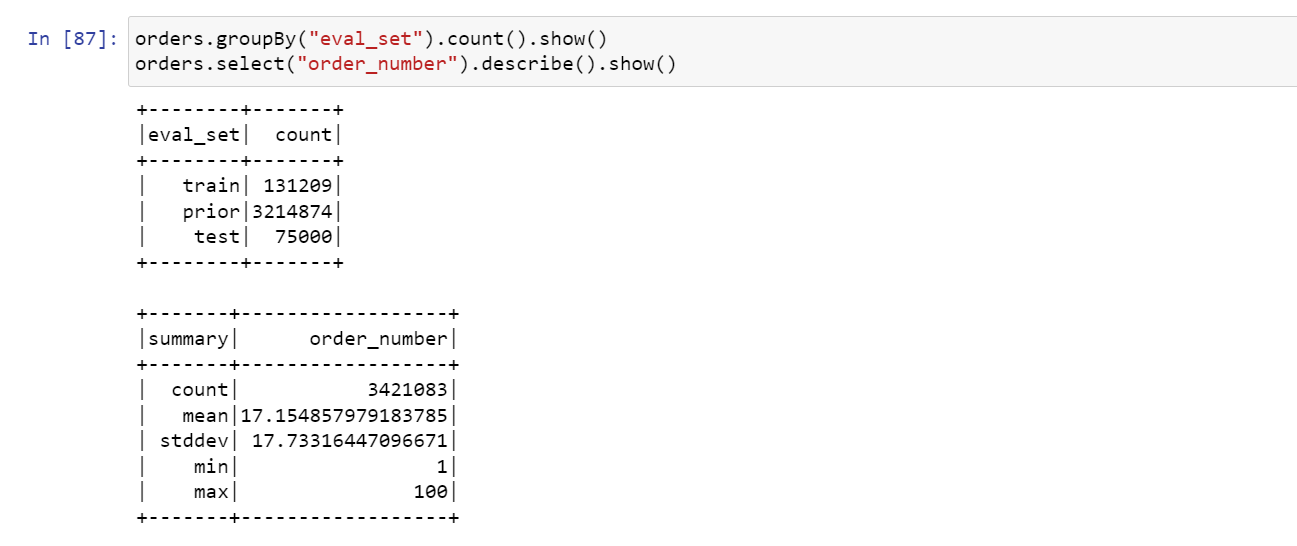
****

Figure 8 - statistics and summary of orders dataset

**4.3. SUMMARY STATISTICS**

The charts that will offer a succinct explanation of the final dataset's summary statistics are included below.

Figure 9 displays the most popular department based on sales. According to the graph, the most profitable sales department is snacks and products, while the lowest seven categories are bulk, alcohol, and personal care.

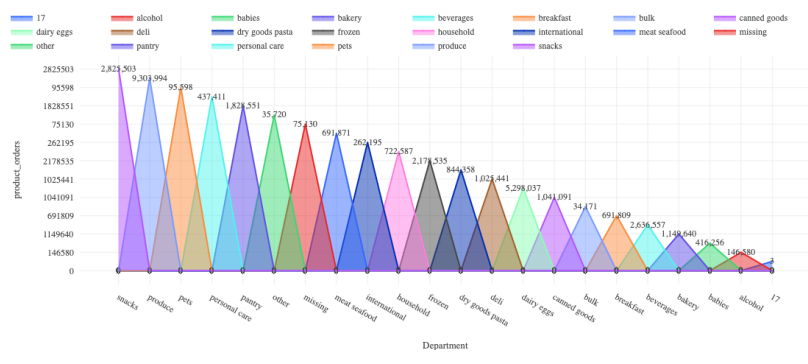
****

Figure 9 - Most popular department according to the sale

The busiest shopping days, as depicted in Figure 10, are shown. According to the data, Sunday has the biggest sales volume, followed by Monday, with a little decline in sales throughout the week.

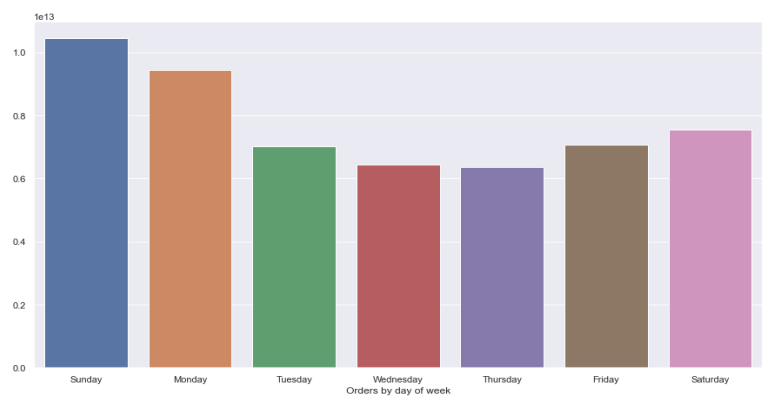


Figure 10 - Most popular days of week to order

The most common time that customers like to place orders is shown in Figure 11. The figure displays the hours in military time for analysis. The research shows that during regular business hours, from 9:00 am to 5:00 pm, are the busiest times for shopping. Before 9:00 am and after 5:00 pm, there aren't many shopping hours where we may presume people who need certain things are buying them.

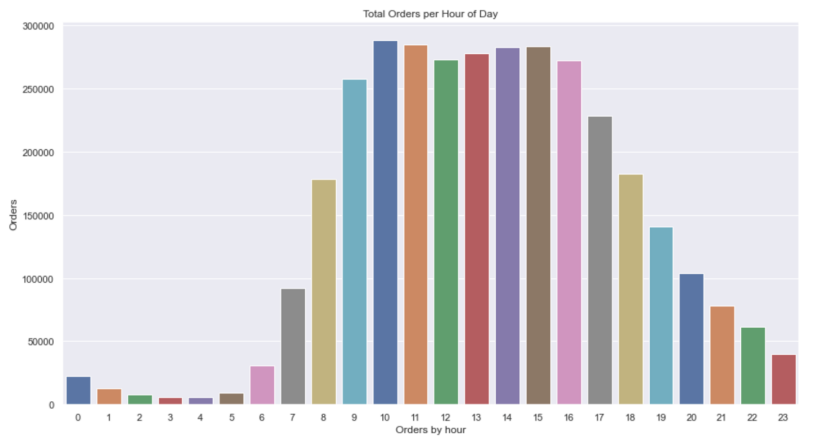


Figure 11 - Most popular time of the day to order

Figure 12 shows the top 20 most popular products ordered by the majority of customers. Fruits and vegetables are among the top 20 selling products, according to this list. Additionally, we can see that the most popular among them are organic goods.

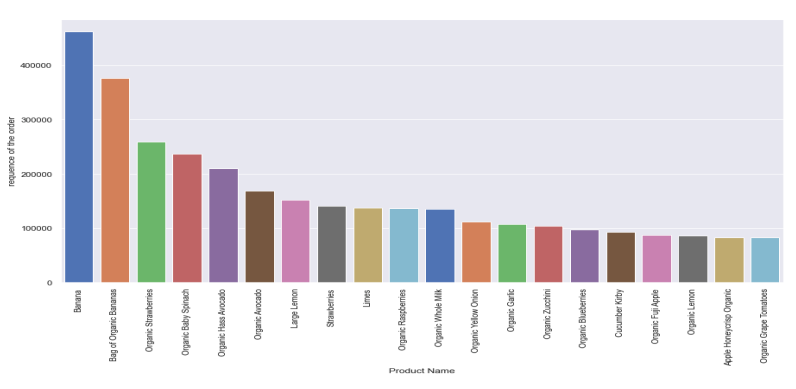


Figure 12 - Top 20 most popular products

Figure 13 shows from which department customers reordered the products. According to the figure, the biggest rate of reordering is for dairy eggs. The second-place category, which likewise shows the same reorder ratio, is beverages and produce. The department with the lowest reorder ratio is personal care.

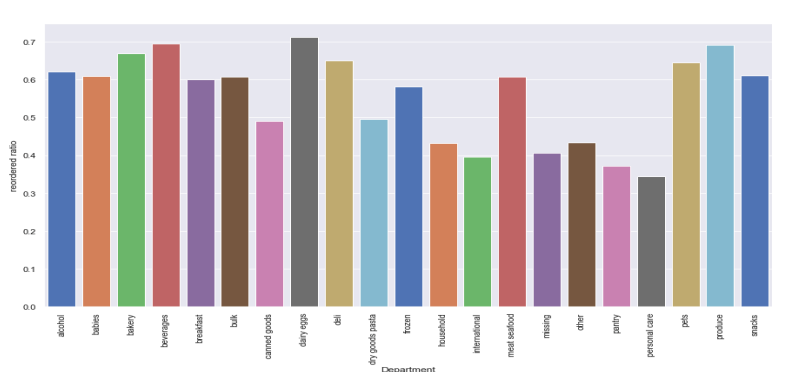


Figure 13 - Department reorder ratio

**5. Strategy for Basket Analysis**

In the section on modelling technique, we first review the fundamental concepts of association analysis. We then discuss over the algorithms that are used to extract the common patterns from larger datasets. Data should be in a transactional format before employing any rule mining algorithms so that the end result shows all the products purchased in a single row.

**5.1. KEY TERMS AND CONCEPTS**

**Association rule**

The what-with-what statement is tied to the association rule. A rule of association is of the form X=> Y, where X is the antecedent and Y is the consequent, and it states that customers who buy X are more likely to buy. Three metrics, namely support, confidence, and lift, can be used to determine an associative rule's magnitude.

**Support:**

Support is defined as a percentage of transactions in a data set containing the number of product items divided by the total number of transactions. Support indicates how many times an itemset has appeared in total transactions.

As an example, suppose 4000 customers visited a retailer shop to buy a product. It has been observed that 400 of them purchased product A, 600 purchased product B, and 200 purchased both product A and product B.

Equation 1: To find Support of an item

According to the equation 1, the support of product A is 10%, and the support of products A and B is 5%. The value of support aids in considering the rules that are worthwhile for further analysis on product correlation with other existing products in the store.

**Confidence**

Confidence is a measure of the like hood that product B is also bought is product A is bought. It is represented in a form of (item set A) => (item set B) where A is a precedent and B is a consequence. Confidence gives a probability of consequence occurring in cart provided with pre-existing antecedent.

Equation 2: To find confidence between two items

Of 400 customer who purchased product A and 200 of them purchased product A and product B which means that 50% customer who purchased Product A and Product B.

**Lift**

Lift refers to the increase in the ratio of the product B when you sell Product A. In general, lifting is greater than one implies that the rule is useful. A lift greater than one indicates that the presence of A in this transaction has increased the likelihood of generating B.

Equation 3: To find Lift between two items

If support of product A is 10 % and confidence of product A and B is 50% than lift is 5 % which means that customers are buying product A and product B together is 5 times more than that purchasing product A alone.

They represent the usefulness and certainty of the revealed rules. The association rules must meet the minimal support and confidence requirements. The two most fundamental algorithms for identifying frequent item sets and identifying relationships between products are Apriori and FP growth.

**Apriori Algorithm**

R. Agrawal and R. Srikant originally presented Apriori in 1994 as the most fundamental approach for locating frequent item sets for Boolean association rules. According to the Apriori principle, if an itemset is frequent, then its items in the subset will also be frequent. The item set is common if support for it exceeds the support threshold. Apriori uses a strategy called level-wise search. Here, scanning the database for often occurring 1-item sets that meet a minimal support is done first. Once more, frequent 1 -item sets are used to find frequent 2 -item sets, and so on until frequent k -item sets are discovered. Actually, Apriori performs a breadth-first search to count the candidate items frequently and follows to the antimonotone principle that every subset of a frequent itemset must likewise be frequent.

This process consists of two basic steps: Prune and Join

1. Generate all frequent item sets - A frequent item set is an item set that has minimum transaction support above minimum support.

2. Generate all confident association rules from frequent item sets – A confident association rule is a rule with confidence above minimum confidence.

The Apriori function from the mlxtend.frequent patterns library is used to apply the Apriori algorithm to the Instacart dataset. The apriori algorithm requires data to be in one hot encoded format before it can determine which items in a particular dataset are most popular. Figure 15 depicts the data structure for a one hot encoder as a result.

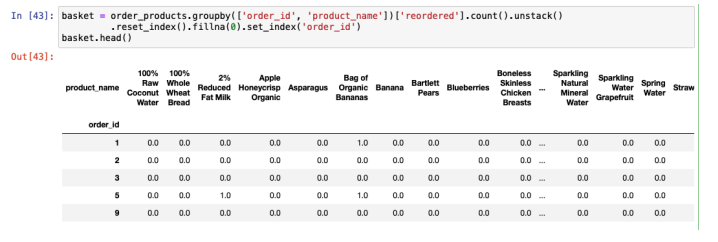


Figure 14 - Data transformation to the one hot encoded

After converting a data to one hot encoder apriori function which is imported from the mlxtend. frequent pattern library used to find items which has minimum support of 2 percentage.

Various input parameters for apriori algorithm are as follows.

* Dataframe(basket): one hot encoded data frame that has 0 and 1 or True and False. This project utilized 0 and 1 format.
* min\_support: the minimal support necessary for an itemset to be chosen, expressed as a floating-point value between 0 and 1.
* use\_colnames: This makes it easier to read by preserving column names for the itemset.
* low\_memory: If True, the algorithm searches for combinations over min support using an iterator.

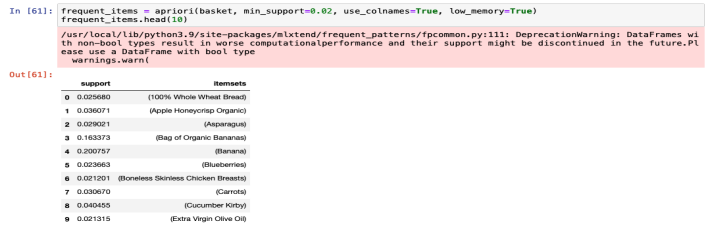
****

Figure 15 - Apriori algorithm on dataset

This project used the association rules function from the mlxtend.frequent patterns library to determine the correlation between the various products using a frequent itemset generated via the apriori technique. The rules generation based on the lift metrics with a minimum threshold value of 1 is shown in Figure.

The table produced by the association rule mining contain various output parameter as follows

* antecedents: This phrase refers to the first product that is supposed to be sold first.
* consequents: The next product that is anticipated to be sold after the first product is referenced in this term.
* antecedent support: The probability of observing the first product referenced by this term.
* consequent support: The probability of observing the second product referenced by this term.
* Support: The probability of observing the first and second product together.
* Confidence: The probability of observing the next product when first product sold.
* Lift: This term refers that when the first product sold the probability of selling the next product increase by a factor of lift.

**FP-Growth Algorithm**

The Apriori algorithm has various drawbacks, such as the cost of computing or the length of time required to complete a run. It takes place because in order to build the itemset, the algorithm must repeatedly iterate over the entire dataset.

The FP growth algorithm provides an alternate method for quantifying frequent item collections by compressing transaction records using the FP-tree visual data structure. To put it another way, FP growth is the process of converting datasets into a graph representation. While the FP growth algorithm builds the FP tree in the form of a lexicographic tree structure, the Apriori approach used candidate creation and check methods. Following the creation of the FP-tree, each frequent item is separated into a series of conditional FP-trees. Additionally, a group of conditional FP-Trees can be mined and evaluated independently. The issues with repeatedly looking for minor suffixes and integrating them into broad specific model discovery are transformed by FP-growth methods. Strong efficacy is provided by the use of slightly repeating objects as a suffix. The cost of the search is greatly reduced by this method.

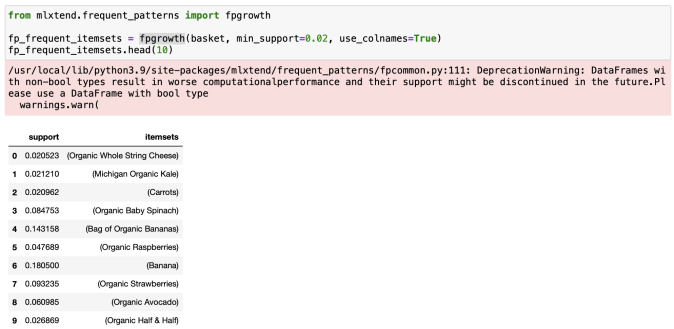


Figure 16 - FP growth on dataset

**Analysis for Apriori and FP Growth**

In this project compare the performance of Apriori and FP Growth algorithm I utilized a python time library to calculate the execution time of each algorithm. The time required for Apriori, and FP growth algorithm given in figure 17.

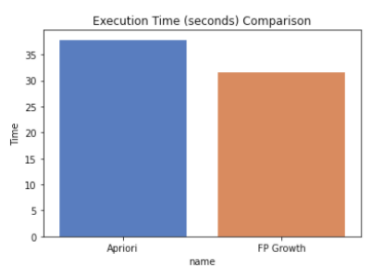


Figure 17- Required time for Apriori and FP Growth Algorithms

The result in figure 17 indicates that FP Growth algorithm takes a shorter time than Apriori algorithm.

This study's primary objective is to compare the performance of the FP Growth and Apriori algorithms. In order to detect links between frequent item sets that satisfy a given minimum lift, we first identify all frequent item sets that satisfy a predefined minimum support for both techniques. Apriori generates a significant amount of candidate sets when the database is large, and the FP Growth method cannot be used to construct a primary memory-based FP tree in this case. Therefore, if we can find a way to simplify our computation it will be more productive. My suggested approach is to decrease dataset items that are best-selling goods. So, using the goods that customers purchased the most of, I reshaped the datasets.

**Comparison between Apriori & FP-Growth Algorithm**

|  |  |  |
| --- | --- | --- |
| Evaluation Criteria | Apriori | FP Growth |
| Techniques | Breadth first search | Divide and Conquer |
| Generation of Patterns | Apriori makes pattern by combining the items into individual tons, sets, and triplets. | FP growth creates structure by building an FP tree. |
| Generation of Candidates | Apriori uses the list of candidates. | There is no generation of candidates. |
| Time | Takes more time to execute | Takes less time compared to Apriori |
| Memory Usage | The variations of candidates are kept in memory | A Compact Database Copy is saved |

Table 6 – Comparison between Apriori & FP-Growth Algorithm

Table 6 illustrates the general comparison between Apriori and FP-Growth Algorithm.

1. **CLOUD PLATFORM**

**Azure Cloud Services:**

Azure Cloud Services offer a comprehensive suite of cloud-based solutions designed to empower businesses with scalable, flexible, and cost-effective computing resources. As part of Microsoft's Azure cloud platform, these services provide organizations with the ability to build, deploy, and manage applications, infrastructure, and data in a secure and highly available environment.

**Key Features and Benefits:**

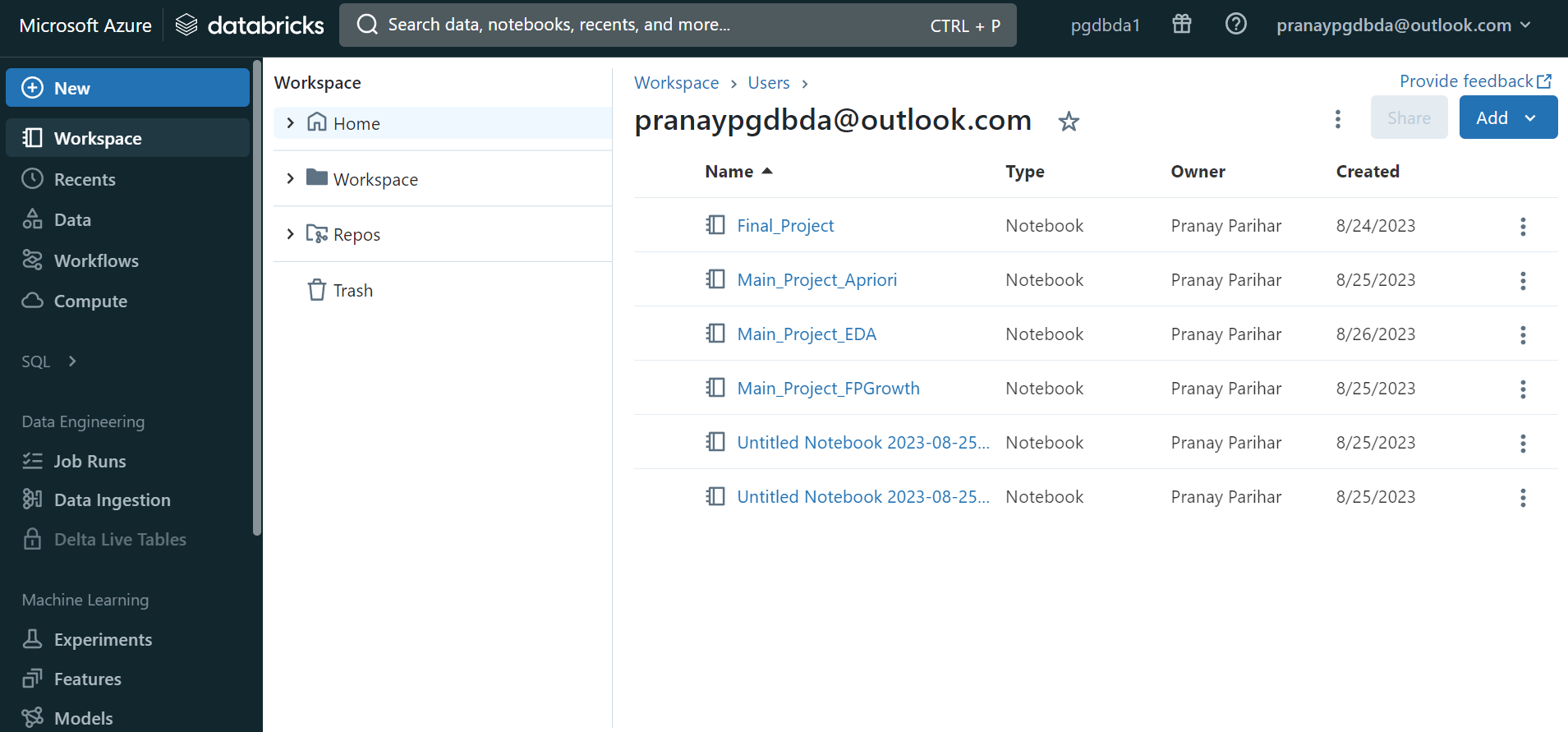
Scalability: Azure Cloud Services provide the ability to scale resources up or down based on demand, ensuring optimal performance and cost efficiency.

Flexibility: With a wide range of services including virtual machines, databases, analytics, AI, and more, Azure allows businesses to tailor solutions to their specific needs.

Global Reach: Azure's vast network of data centers spans the globe, enabling organizations to deploy applications and services in regions of their choice for improved performance and compliance.

Security and Compliance: Azure offers robust security measures, including encryption, identity and access management, and compliance certifications to ensure data protection and regulatory adherence.

Pay-as-You-Go Model: Azure's pay-as-you-go pricing allows organizations to pay only for the resources they consume, optimizing cost management.



**Azure Blob Storage:**

Azure Blob Storage is a scalable object storage service provided by Microsoft Azure, designed to store and manage large amounts of unstructured data. It is a core component of the Azure Data Services ecosystem and offers versatile data storage options for various use cases.

**Key Features and Benefits:**

Scalability and Durability: Azure Blob Storage scales seamlessly to accommodate growing data needs, and its built-in redundancy ensures high data durability.

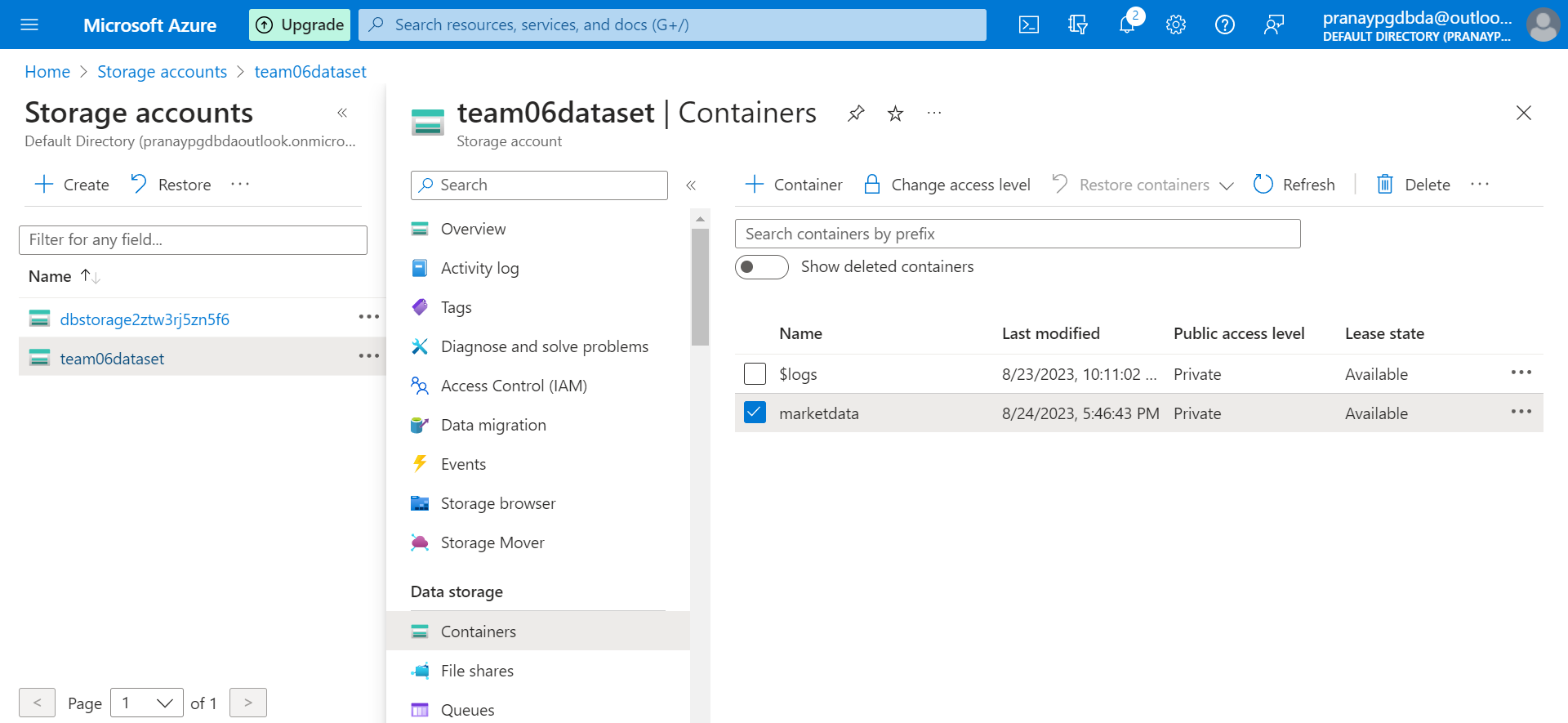
Data Accessibility: It enables easy access to stored data from anywhere in the world through HTTP/HTTPS protocols.

Multiple Data Types: Azure Blob Storage supports multiple data types, including text, binary, images, and large-scale analytics data.

Tiered Storage: Azure Blob Storage offers multiple storage tiers, such as hot, cool, and archive, allowing organizations to optimize costs by aligning data access patterns with storage costs.

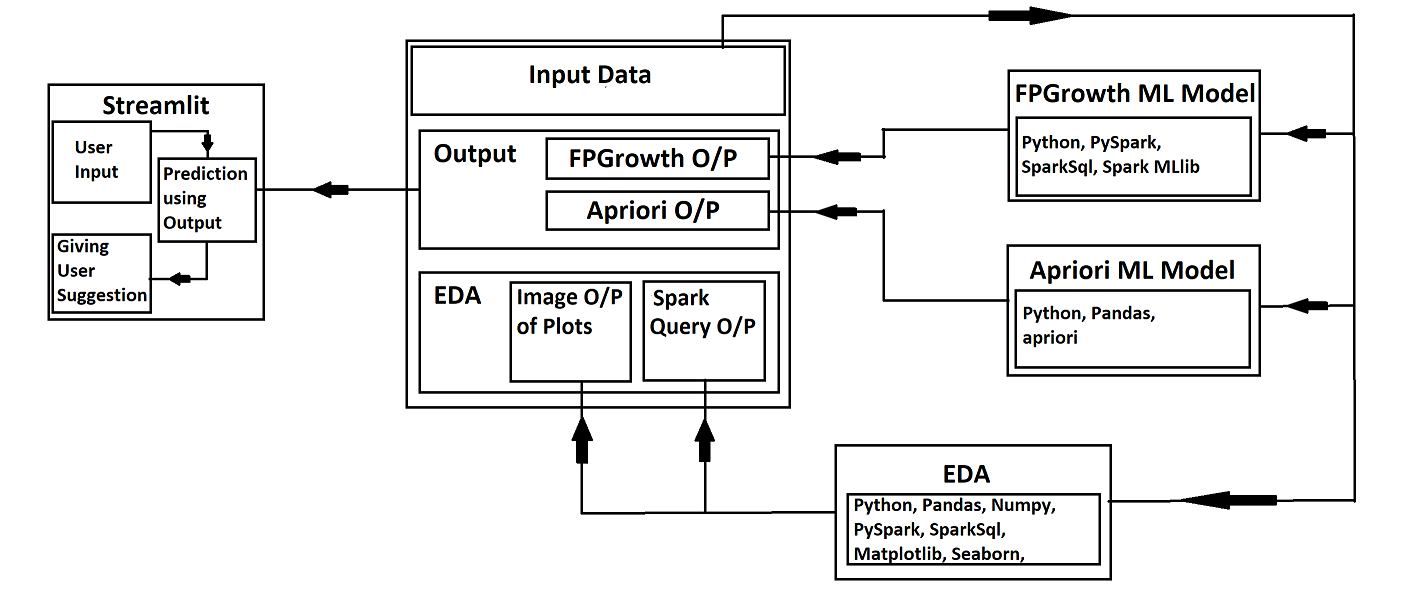
Data Security: It provides encryption at rest and in transit, along with integration with Azure Active Directory for access control and identity management.

Data Movement: Azure Blob Storage integrates seamlessly with various Azure services, facilitating data movement, analytics, and machine learning workflows.



**7.** **PIPELINE CONNECTIVITY**

Pipeline Connectivity between Azure Blob Storage and Databricks:



The pipeline connectivity between Azure Blob Storage and Databricks establishes a seamless link for transferring, processing, and analyzing data. Azure Blob Storage serves as a scalable and reliable storage solution, while Databricks offers a powerful analytics platform built on Apache Spark.

This integration enables data engineers, analysts, and data scientists to effortlessly move data from Azure Blob Storage to Databricks clusters, where it can be processed, transformed, and analyzed using advanced analytics, machine learning, and collaborative tools.

The pipeline ensures secure data movement by leveraging Azure Active Directory for authentication and authorization. It simplifies the process of accessing data stored in various formats within Azure Blob Storage and accelerates data-driven decision-making within the Databricks environment.

By connecting Azure Blob Storage and Databricks, organizations can leverage the strengths of both platforms, benefiting from the scalability of cloud storage and the computational capabilities of Databricks for efficient data exploration, analysis, and insights generation.

**8. CONCLUSION AND FUTURE WORK**

1. This study showcases a comprehensive data analytics project conducted on the Instacart market basket dataset.
2. The research involved crucial stages like data understanding, collection, and preparation for modeling, which form a significant aspect of data analysis projects.
3. The dataset underwent thorough cleaning, examination, and feature engineering to enhance its quality, resulting in noteworthy attributes for statistical analysis and integration into the clustering methodology.
4. Visual aids such as plots and graphs were employed to delve into the data, enabling a deeper comprehension and visualization of customer preferences across departments, aisles, order timings, and days of the week.
5. The graphical representation provided insights into the most popular product categories, frequently reordered departments, and peak hours and days for order placement.
6. Future enhancements for the project could involve the incorporation of advanced mining algorithms, including Apriori and FP-Growth, to bolster performance and achieve quicker results, especially for sparse datasets.
7. The current approach primarily utilizes association rules to tap into collective information, recommending similar associated items to customers based on product associations.
8. Future work could explore leveraging association rules for content-based information as well, suggesting products based on similarities.
9. A potential recommendation system could combine item-based and customer-based approaches into a hybrid framework, capitalizing on the strengths of both methodologies.
10. The application's scope extends beyond market basket analysis and could be extended to domains like sales tracking, product monitoring, discount calculation, and pricing strategies.
11. As a valuable method for large databases with limited memory space, future refinements could focus on enhancing efficiency and performance.
12. Overall, this study represents a significant step towards understanding customer behavior and preferences, paving the way for further exploration and refinement in future projects.