PREDICTING SUPERMARKET SALES WITH BIG DATA ANALYTICS

A comparative study of machine learning techniques

Pranav Satish Garve

Department of Information Technology, Shri Guru Gobind Singhji Institute of Engineering and Technology (SGGSIET), Nanded

pranavgarve1@gmail.com

Abstract: This study explores the complex dynamics of supermarket shopping habits, with a focus on understanding customer preferences and decision-making processes. Through meticulous data analysis encompassing order frequency, day-of-week trends, and product preferences, we seek to uncover insights into consumer behavior patterns. Employing advanced techniques such as clustering and machine learning, we identify correlations between frequently purchased products, revealing underlying customer preferences and purchasing habits. Additionally, we delve into demographic factors to further refine our understanding of consumer behavior. Utilizing this analysis, we aim to improve the accuracy of future sales predictions. By extrapolating from past purchasing behaviors, including shopping frequency and timing, we can make informed projections about future consumer choices. This predictive capability is invaluable for supermarkets, enabling them to proactively plan and cater to customer demands, akin to possessing a foresight mechanism for anticipating consumer needs. Moreover, we explore the potential for personalized marketing strategies based on individual customer preferences. In conclusion, this research underscores the potential for supermarkets to harness data analytics to deepen their understanding of customers and drive strategic decision-making. Through the application of machine learning and comprehensive analysis of shopping behaviors, supermarkets can tailor their offerings to better align with customer preferences, fostering satisfaction and facilitating growth in today's competitive retail landscape. Furthermore, we discuss the implications of our findings for the wider retail industry and the opportunities for collaboration and innovation.

Keywords: Supermarket sales, big data analytics, forecasting, comparative study, clustering, etc

1. Introduction

Supermarkets are like treasure troves filled with all sorts of goodies, from groceries to household items. They've become essential in our daily lives, offering convenience and everything we need under one roof. But with so many supermarkets popping up, competition is fierce. That's where data analytics comes in. By crunching numbers and analyzing how people shop, supermarkets can stay ahead of the game and keep their shelves stocked with the stuff people love.

Data analytics plays a pivotal role in supermarkets by leveraging data to anticipate customers' future purchases. It enables markets to discern popular products and identify those in need of promotion. Through inventory management and insight into shopper behaviors, supermarkets can make informed choices regarding product selection and marketing strategies. Moreover, by gaining deeper insights into customer demographics and preferences, they can customize promotions and advertisements accordingly.

Big data analytics is the superhero of the supermarket world. With mountains of data pouring in from all directions – think checkout registers, loyalty programs, and even social media – big data analytics swoops in to save the day. It's like having a super-powered computer that can sift through all that info in record time, uncovering hidden patterns and trends. And with the help of machine learning, supermarkets can even predict future sales trends, giving them a leg up in the ever-changing retail landscape.

So, what's the point of all this number-crunching? Well, it's simple – to help supermarkets run smoother and make more money. By optimizing inventory management and understanding what makes shoppers tick, supermarkets can boost their bottom line and keep customers coming back for more. And at the end of the day, that's a win-win for everyone – supermarkets make more dough, and customers get the products and services they love.

2. Literature review

In the past few years, there has been a significant increase in interest in the application of big data analytics techniques to improve sales forecasting in the retail industry. Supermarkets, in particular, have been keen on tapping into this trend to refine their inventory management practices and bolster their financial performance. This section offers a glimpse into the various methods currently employed for sales forecasting in supermarkets, as well as an exploration of the big data analytics techniques that are increasingly being integrated into these strategies.

In the domain of retail, K-means clustering and PCA stand out as essential tools. K-means efficiently groups data points, revealing distinct customer segments based on order history and preferences. Meanwhile, PCA simplifies complex datasets, aiding interpretation while retaining crucial information. Research by Jain and Dubes (1988) underscores the efficacy of these techniques in diverse retail contexts.

Through K-means clustering and PCA, businesses gain actionable insights swiftly. These methods enable precise customer segmentation, facilitating targeted marketing strategies.

Studies by Jolliffe (2002) highlight their utility in deciphering consumer behaviour, empowering supermarkets to optimize sales and enhance customer satisfaction.

Alternatively, big data analytics techniques offer a comprehensive strategy for sales forecasting, examining extensive datasets across various origins. Employing machine learning algorithms, these methods discern consumer behaviour patterns and anticipate future sales, incorporating diverse factors like product preferences, promotions, and social media engagement. Moreover, big data analytics aids supermarkets in refining inventory management through demand prediction and waste reduction identification

Some of the commonly used machine learning algorithms for sales forecasting in the retail industry include unsupervised learning, principal component analysis (PCA). These algorithms are capable of handling large and complex datasets, and can provide more accurate forecasts than traditional statistical models. Moreover, large-scale data analysis methods can be employed to carry out market segmentation studies, aiding supermarkets in pinpointing various groups of customers with unique needs and tastes.

In their research, Wang and colleagues (2019) utilized machine learning techniques to forecast sales within a Chinese e-commerce platform. Their investigation showcased the effectiveness of their methodology in predicting sales across a wide array of product categories. Through their study, they provided significant insights aimed at improving inventory management practices and mitigating wastefulness. This work underscored the applicability of machine learning algorithms in enhancing sales forecasting accuracy and operational efficiency within the e-commerce domain.

Vafeiadis et al (2018) employed both time-series analysis and machine learning methodologies to predict sales within a Greek supermarket franchise. Their study demonstrated the efficacy of their approach in accurately forecasting sales across various store locations and product categories.

Ghosh et al. (2017) employed a blend of data mining and machine learning methodologies to forecast sales for a retailer based in the United States. Their study revealed the efficacy of their approach in pinpointing the variables affecting sales, including promotional activities, weather patterns, and economic factors.

Overall, these studies underscore the growing interest in leveraging machine learning algorithms to analyse and predict supermarket sales in big data analytic.

3. Methodology

A. Data Collection and Tool Used:

The gathering data to build the predictive model, we collected data of supermarket sales from (https://www.kaggle.com) a popular web portal.

Jupyter Notebook is an extremely popular tool among data scientists and machine learning practitioners for developing and presenting code, visualizations, and explanations. It's particularly well-suited for machine learning tasks due to its interactive nature and support for various programming languages, including Python. Its Graphical User Interface (GUI) communicates with all the choices required to analyze machine data and learning techniques.

Using python programming language, we performed Preprocessing, classification, regression, grouping, and visualization on this platform.

B. Data Understanding:

In this phase, we aim to delve deeper into the dataset to gain a comprehensive understanding. To accomplish this, we will employ essential libraries such as pandas, numpy, and matplotlib within the Jupyter notebook environment. Initially, we'll import these libraries and packages and proceed to read the dataset. Subsequently, we will store the dataset in a variable named 'data'.

	order_id	user_id	order_number	order_dow	order_hour_of_day	days_since_prior_order	product_id	add_to_cart_order	reordered	department_id	department	product_name
0	2425083	49125	1	2	18	NaN	17	1	0	13	pantry	baking ingredients
1	2425083	49125	1	2	18	NaN	91	2	0	16	dairy eggs	soy lactosefree
2	2425083	49125	1	2	18	NaN	36	3	0	16	dairy eggs	butter
3	2425083	49125	1	2	18	NaN	83	4	0	4	produce	fresh vegetables
4	2425083	49125	1	2	18	NaN	83	5	0	4	produce	fresh vegetables
...	...	...	...	...	...	...	...	...	...	...	...	...
2019496	3390742	199430	16	3	18	5.0	83	8	0	4	produce	fresh vegetables
2019497	458285	128787	42	2	19	3.0	115	1	1	7	beverages	water seltzer sparkling water
2019498	458285	128787	42	2	19	3.0	32	2	1	4	produce	packaged produce
2019499	458285	128787	42	2	19	3.0	32	3	1	4	produce	packaged produce
2019500	458285	128787	42	2	19	3.0	123	4	1	4	produce	packaged vegetables fruits

Read the dataset

Next, let's examine the dataset for missing values and its overall condition, including data types and the sum of rows.

```
df.info()
```

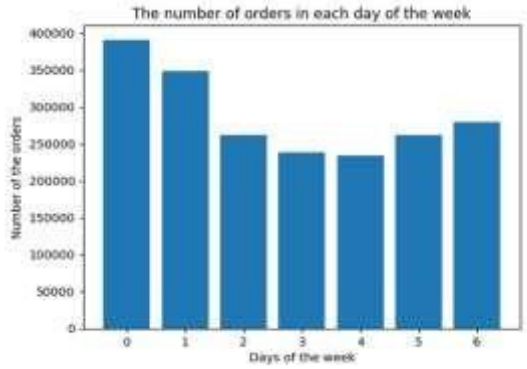
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2019501 entries, 0 to 2019500
Data columns (total 12 columns):
 #   Column                  Dtype
---  ------                  -----
 0   order_id                int64
 1   user_id                 int64
 2   order_number            int64
 3   order_dow               int64
 4   order_hour_of_day       int64
 5   days_since_prior_order  float64
 6   product_id              int64
 7   add_to_cart_order       int64
 8   reordered               int64
 9   department_id           int64
 10  department              object
 11  product_name            object
dtypes: float64(1), int64(9), object(2)
memory usage: 184.9+ MB
```

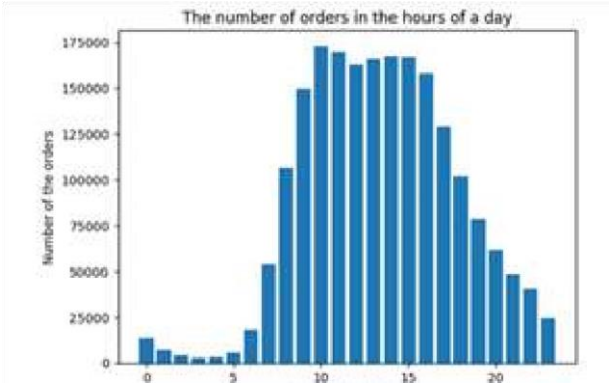
Data Information

C. Data Exploring:

Data exploration involves analyzing and understanding the structure and characteristics of a dataset, particularly focusing on aspects relevant to clustering analysis in the context of supermarket data in machine learning. It encompasses tasks such as identifying patterns, distributions, and relationships within the data, as well as detecting outliers and missing values. The goal is to gain insights that inform subsequent steps in the clustering process, aiding in the selection of appropriate features and algorithms for effectively grouping similar data points.

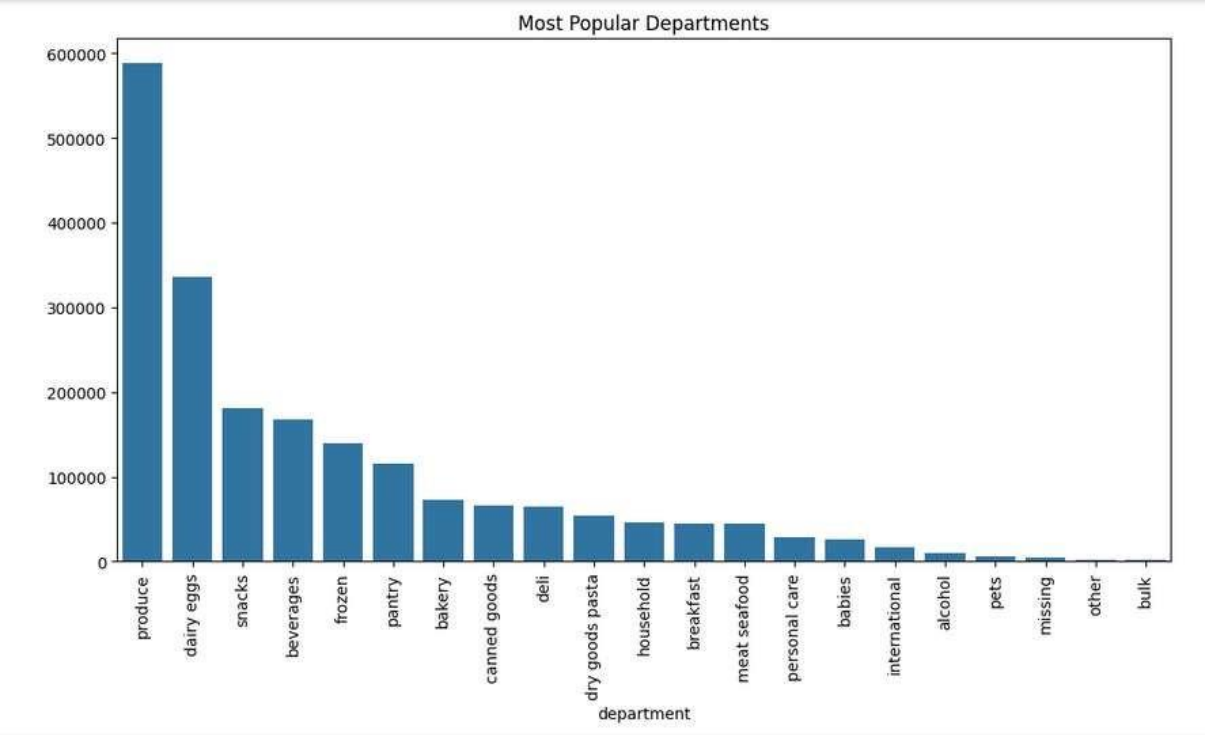
- You can see that the highest number of orders at Hunter's store are on Mondays and then Tuesdays. We see that it decreases during the week and increases slightly at the end of the week.



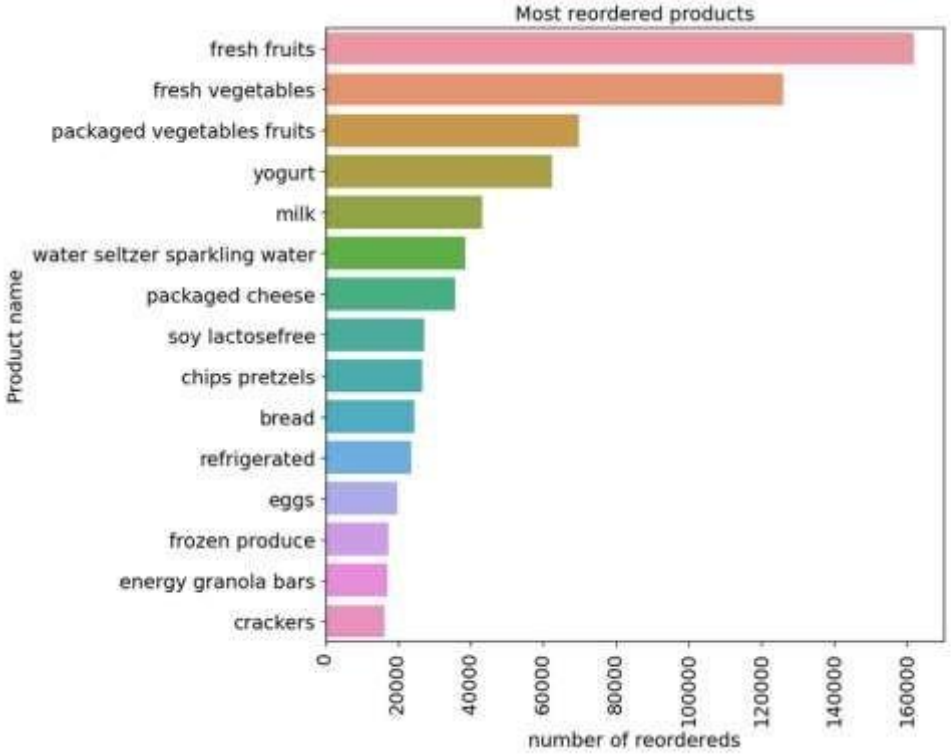


- As can be seen in the chart, the highest number of orders occurs around 10:00 to 16:00.

- we can see that Produce, Dairy Eggs, Snacks, Beverages and Frozen are the five most popular departments.

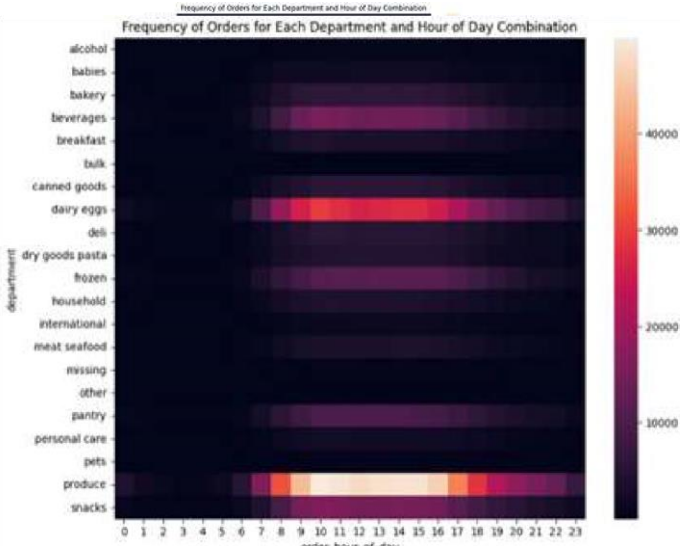


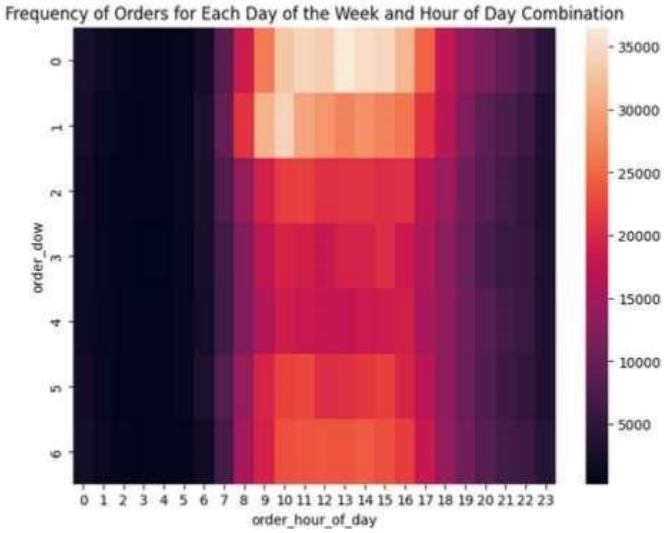
- As you can see, Fresh fruits, Fresh vegetables, Packaged vegetable fruits, Yogurt and Milk are the five most reordered products.



- we can see that a large number of orders per day are made from the "Produce" department of the store between 9:00 and 18:00. Departments “bulk” and “Pets” do not have significant sales during the day

- we can make the same observations as before. This second chart shows that the growth in the number of orders in each day starts at 7:00 and peaks at 13:00 or 14:00 for some days and 11:00 for others. It also shows that the highest number of orders occur between 10:00 and 15:00 on Mondays and 9:00 to 10:00 on Tuesdays.





4. Machine Learning Techniques

To forecast supermarket sales using big data analytics, the research team utilized various machine learning algorithms commonly applied in time-series prediction. These algorithms were chosen for their capacity to manage extensive datasets, identify intricate patterns, and produce precise forecasts.

Below is a simplified explanation of the machine learning algorithms employed in this study: **1.**

Clustering:

Clustering in unsupervised learning is a technique used to group data points into distinct clusters based on similarities in their features or characteristics, without the need for predefined labels or categories. The objective is to identify natural groupings within the data, where data points within the same cluster are more similar to each other compared to those in other clusters. Clustering algorithms aim to partition the dataset into clusters that maximize intra-cluster similarity and minimize inter-cluster similarity, facilitating data exploration, pattern recognition, and understanding of underlying structures within the data.

2. Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is a technique used in unsupervised machine learning to reduce the dimensionality of high-dimensional datasets while preserving the most important information. It identifies the principal components, which are new variables that are linear combinations of the original features, aiming to maximize the variance of the data. PCA is commonly employed for data visualization, noise reduction, and feature extraction, facilitating easier interpretation and analysis of complex datasets.

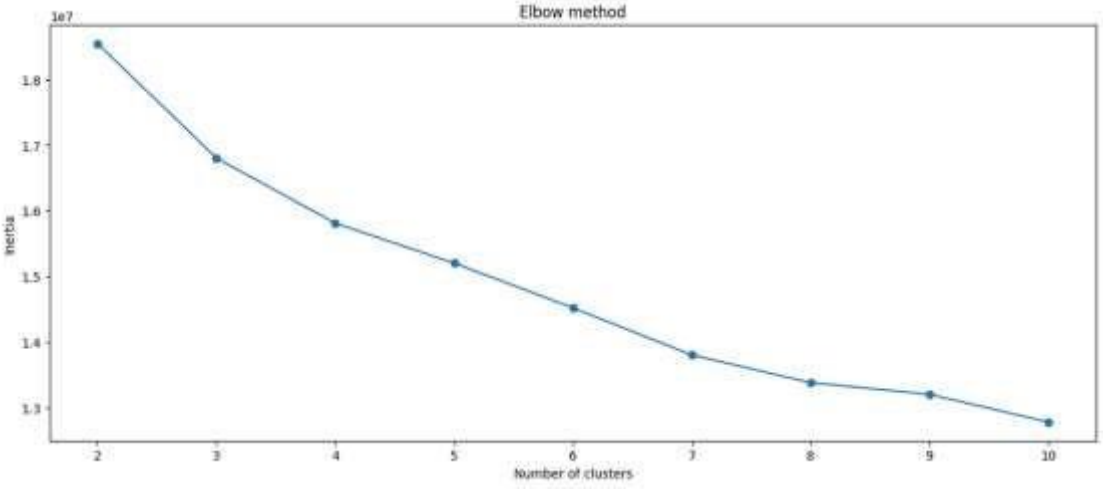
5. Comparative Analysis

In this study, the performance machine learning algorithms: K- means Clustering, PCA was evaluated and compared using the collected supermarket sales data. The purpose of this analysis was to determine which algorithm(s) performed the best in predicting sales for supermarket. The performance of each algorithm was evaluated using these metrics: K means Clustering, Elbow method, Silhouette method, and plotting graph using PCA.

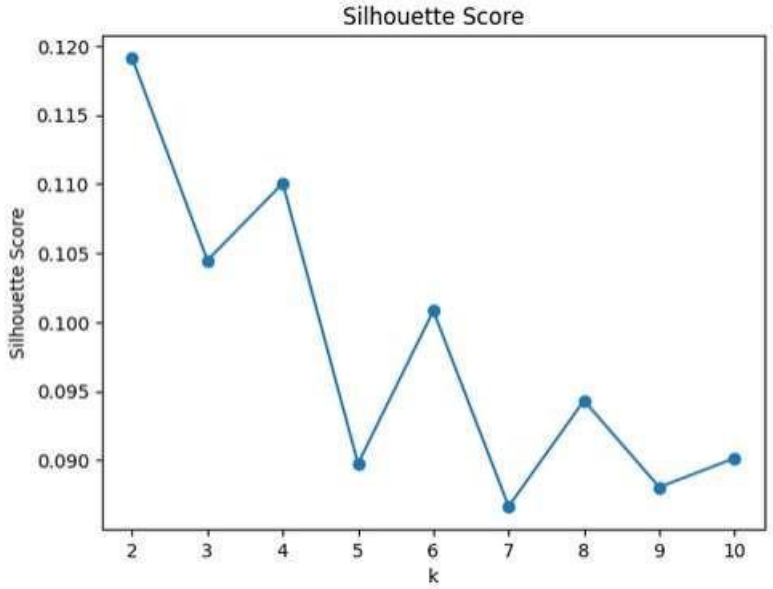
1. **K-means Clustering:**
   K means clustering, assigns data points to one of the K clusters depending on their distance from the center of the clusters. It starts by randomly assigning the clusters centroid in the space. Then each data point assign to one of the clusters based on its distance from centroid of the cluster. After assigning each point to one of the cluster, new cluster centroids are assigned. This process runs iteratively until it finds good cluster. In the analysis we assume that number of clusters is given in advanced and we have to put points in one of the groups.

```
checke with 2 clusters | Inertia : 18547911.47710687
checke with 3 clusters | Inertia : 16803264.887409553
checke with 4 clusters | Inertia : 15809126.918312903
checke with 5 clusters | Inertia : 15201104.61931771
checke with 6 clusters | Inertia : 14519396.474782713
checke with 7 clusters | Inertia : 13800267.961906062
checke with 8 clusters | Inertia : 13378748.456463097
checke with 9 clusters | Inertia : 13200161.50608976
checke with 10 clusters | Inertia : 12778006.322304506
```

2. **Elbow Method:**
   The Elbow Method in K-Means clustering helps to determine the optimal number of clusters by plotting the within-cluster sum of squares against the number of clusters and identifying the "elbow" point where the rate of decrease sharply changes.



3. **Silhouette Method**:
   The silhouette algorithm is one of the many algorithms to determine the optimal number of clusters for an unsupervised learning technique. In the Silhouette algorithm, we assume that the data has already been clustered into k clusters by a clustering technique.

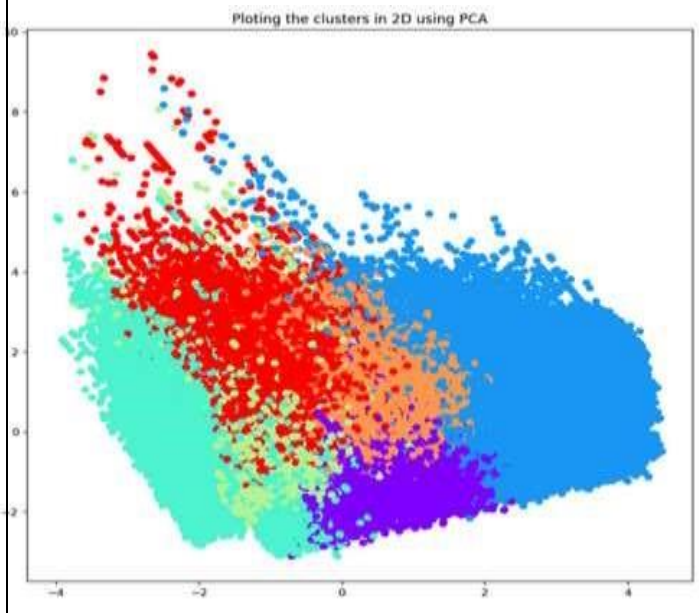


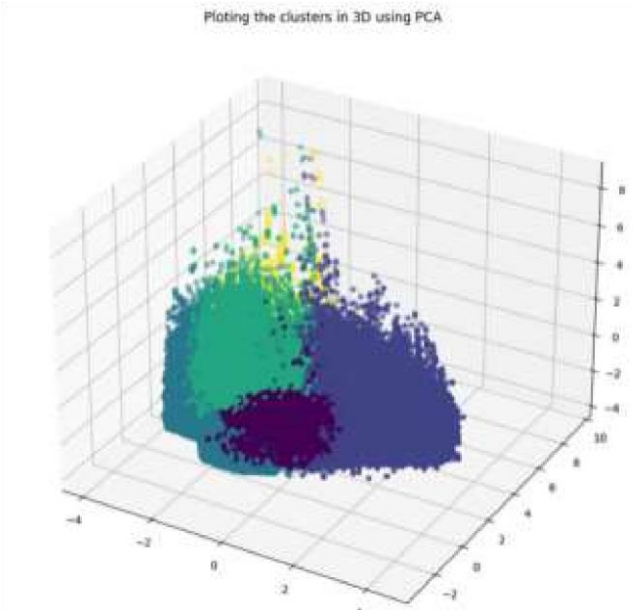
The plot below also shows that the silhouette score of 6 clusters is higher than 5 clusters, which gives more confidence to our choice.

4. **PCA:**

Principal Component Analysis is an unsupervised learning algorithm that is used for the dimensionality reduction in machine learning. It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.

Plotting the cluster in 2D and 3D graph





6.Results and Discussion:

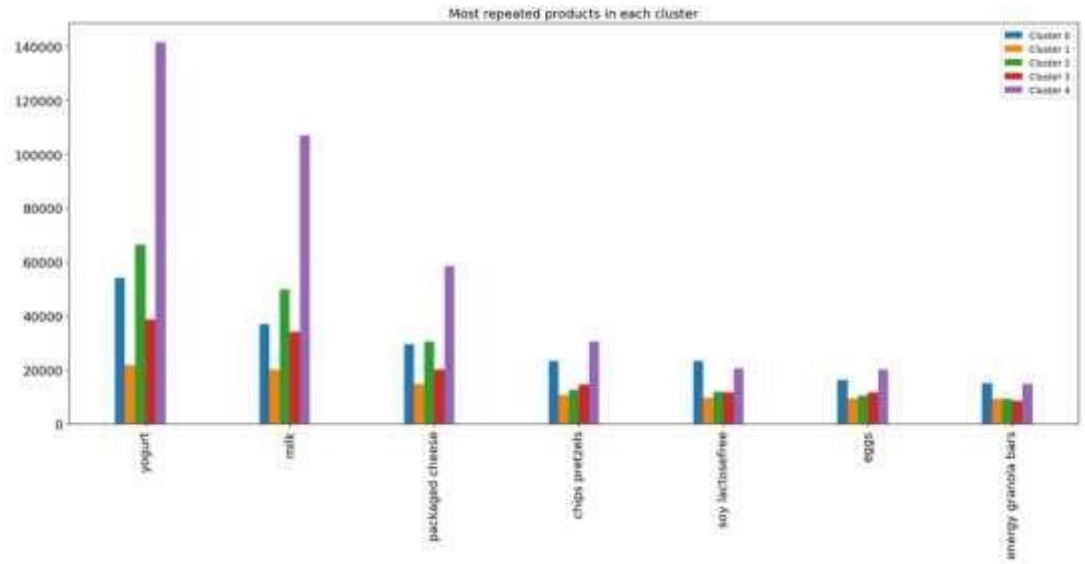
The comparative analysis results showed that K means clustering and PCA outperformed with the algorithms in terms of forecasting supermarket sales. This indicates that K means clustering and PCA are more accurate and better at capturing the patterns in the sales data compared to the other algorithms.

These findings are consistent with previous studies that have used machine learning techniques for supermarket sales forecasting. For example, Wang et al. (2019) found that LSTM performed better than other algorithms for sales forecasting in a Chinese supermarket.

After analysis and performing unsupervised machine learning algorithm to predict sales of supermarket we have a resultant cluster .

product_name	Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4
yogurt	54101.0	21899.0	NaN	14579.0	NaN
milk	36940.0	9437.0	NaN	8638.0	NaN
packaged cheese	29441.0	20233.0	NaN	11713.0	NaN
chips pretzels	23384.0	14723.0	NaN	NaN	NaN
soy lactosefree	23254.0	NaN	NaN	NaN	NaN
eggs	16356.0	NaN	NaN	NaN	NaN
energy granola bars	14993.0	9373.0	NaN	NaN	NaN
baking ingredients	NaN	10641.0	NaN	NaN	NaN
crackers	NaN	9792.0	NaN	NaN	NaN
fresh vegetables	NaN	NaN	66359.0	38754.0	107097.0
fresh fruits	NaN	NaN	49720.0	34218.0	141602.0
packaged vegetables fruits	NaN	NaN	30619.0	20157.0	58597.0
ice cream ice	NaN	NaN	12579.0	NaN	NaN
frozen produce	NaN	NaN	11813.0	NaN	14811.0
water seltzer sparkling water	NaN	NaN	10391.0	11563.0	30474.0
bread	NaN	NaN	9250.0	NaN	20732.0
refrigerated	NaN	NaN	NaN	NaN	20073.0

And create a bar graph for resultant cluster:



7. Limitations and Ethical Considerations

There are several factors to consider regarding the scope and ethical implications of this study:

The study was conducted within a limited supermarket chain, potentially limiting the applicability of findings to broader contexts. Future research could involve multiple chains to enhance the generalizability of results.

The study relied on historical sales data without considering external influences like shifts in consumer behavior or market dynamics is a constraint. Incorporating additional data sources such as weather patterns, social media trends, and economic indicators could offer a more comprehensive understanding of factors impacting supermarket sales.

8. Benefits and Challenges

The two-seasons data analytics study for supermarket sales forecasting offers several advantages and presents some hurdles.

Benefits:

1. Precise sales prediction: The study demonstrates the effectiveness of machine learning algorithms in anticipating supermarket sales. By analyzing historical sales data and other relevant factors, these models can reliably forecast future sales.

2. Enhanced stock management: Accurate sales forecasts can assist supermarket chains in optimizing their inventory levels, minimizing waste, and boosting operational efficiency.

3. Enhanced shopper satisfaction: By accurately predicting demand, supermarket chains can ensure adequate availability of popular items, reducing instances of stockouts and enhancing the overall customer experience.

Challenges:

1. Limited scope: The study utilized data from just limited supermarket chain in India and examined sales data from only two seasons in a single year. This restricts how widely the findings can be applied to different areas and situations.

2. Data reliability and accessibility: The effectiveness of the models hinges on the reliability and accessibility of the data. If the data is incorrect or incomplete, it can result in inaccurate sales predictions.

9. Conclusion:

The study on supermarket sales forecasting using two seasons of data has revealed how machine learning algorithms can enhance sales predictions and inventory control. By analyzing past sales and other factors, the study identified effective models like K-means clustering and PCA for accurate forecasts. It emphasizes the significance of reliable data for precise sales predictions.

Future research should explore:

1. Testing various machine learning algorithms and data sources for sales forecasting.

2. Analyzing how different marketing strategies affect sales.

3. Investigating ethical concerns related to gathering and analyzing customer data.

10. References:

1. 1. Amirhossein Shamsaei's "Supermarket Dataset | EDA | Clustering" on Kaggle offers data about a supermarket's business, which can be useful for exploratory data analysis and clustering tasks.

2. In his book "Machine Learning for Predictive Analytics," Steven Finlay discusses the application of machine learning techniques for predictive analysis in various domains.

3. Abhyuday, Mishra, and Singh (2020) examine the application of machine learning in sales forecasting within the retail sector. Their work is published in the International Journal of Scientific Research in Computer Science, Engineering and Information Technology.

4. Siddique, Hussain, Khan, and Mahmood (2019) present a case study on using machine learning algorithms for sales forecasting in the retail industry at the International Conference on Computing, Mathematics and Engineering Technologies (iCoMET).

5. Garg and Rani (2020) focus on utilizing machine learning algorithms for forecasting supermarket sales, as outlined in the International Journal of Computer Science and Information Security.

6. Kumar and Gupta (2020) discuss using machine learning algorithms to predict sales in the retail sector at the 4th International Conference on Inventive Computation Technologies (ICICT
2019).

7. Anastasia Griva and colleagues (2018) delve into customer visit segmentation using market basket data in the retail industry, with their findings published in Expert Systems with Applications, Volume 100.

8. Philip Doganis and colleagues present research in the Journal of Food Engineering (2006) on forecasting sales of short shelf-life food products using neural networks and evolutionary algorithms.

9. Yonatan Rabinovich's Kaggle dataset includes sales data across 10 stores, featuring 1,559 products from various cities.

10. Beheshti-Kashi, Karimi, Thoben, Lutjen, and Teucke (2015) conduct research on fashion market retail sales forecasting, published in Systems Science & Control Engineering, Vol. 3(1).

11. Das and Chaudhury (2018) analyze the use of feed-forward and recurrent neural networks to predict sales in the retail footwear industry.

12. Cheriyan, Ibrahim, Mohanan, and Treesa (2018) investigate machine learning techniques for sales prediction in their research presented at the International Conference on Computing, Electronics, and Communications Engineering.

13. Pan and Zhou explore convolutional neural networks' application in data mining and ecommerce sales forecasting.

14. Yuan, Xu, Li, and colleagues examine topic sentiment analysis for predicting sales trends and how it can affect forecasting strategies.