Machine Learning

E-commerce Consumer Behaviours

Mobin Nesari
Dr. Hadi Farahani

June 2, 2023

Introduction:

E-commerce consumer behaviour dataset is a collection of data that captures the actions, preferences, and attitudes of consumers in the context of online shopping. This type of dataset provides valuable insights into how consumers interact with e-commerce platforms, what factors influence their purchasing decisions, and how they respond to various marketing strategies. By analyzing this data, businesses can better understand their target audience and tailor their marketing efforts to meet their needs and preferences. The dataset typically includes information such as demographics, browsing behavior, purchase history, and customer feedback. Overall, e-commerce consumer behavior datasets offer a powerful tool for businesses looking to optimize their e-commerce operations and improve customer satisfaction.

Dataset Exploration :

This dataset contains 12 features with 2019501 number of records. Here is a general overview over all features and their data types.

#	Column	Data Type
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

#	Column	Data Type
10	department	object
11	product_name	object

Here is general description over numerical features

Crit \ Feature	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
count	2.019501E+06	2.019501E+06	2.019501E+06	2.019501E+06	2.019501E+06	1.895159E+06	2.019501E+06	2.019501E+06	2.019501E+06	2.019501E+06
mean	1.707013E+06	1.030673E+05	1.715138E+01	2.735367E+00	1.343948E+01	1.138603E+01	7.120590E+01	8.363173E+00	5.897427E-01	9.928349E+00
std	9.859832E+05	5.949117E+04	1.752576E+01	2.093882E+00	4.241008E+00	8.970980E+00	3.820727E+01	7.150059E+00	4.918804E-01	6.282933E+00
min	1.000000E+01	2.000000E+00	1.000000E+00	0.000000E+00	0.000000E+00	0.000000E+00	1.000000E+00	1.000000E+00	0.000000E+00	1.000000E+00
25%	8.526490E+05	5.158400E+04	5.000000E+00	1.000000E+00	1.000000E+01	5.000000E+00	3.100000E+01	3.000000E+00	0.000000E+00	4.000000E+00
50%	1.705004E+06	1.026900E+05	1.100000E+01	3.000000E+00	1.300000E+01	8.000000E+00	8.300000E+01	6.000000E+00	1.000000E+00	9.000000E+00
75%	2.559031E+06	1.546000E+05	2.400000E+01	5.000000E+00	1.600000E+01	1.500000E+01	1.070000E+02	1.100000E+01	1.000000E+00	1.600000E+01
max	3.421080E+06	2.062090E+05	1.000000E+02	6.000000E+00	2.300000E+01	3.000000E+01	1.340000E+02	1.370000E+02	1.000000E+00	2.100000E+01

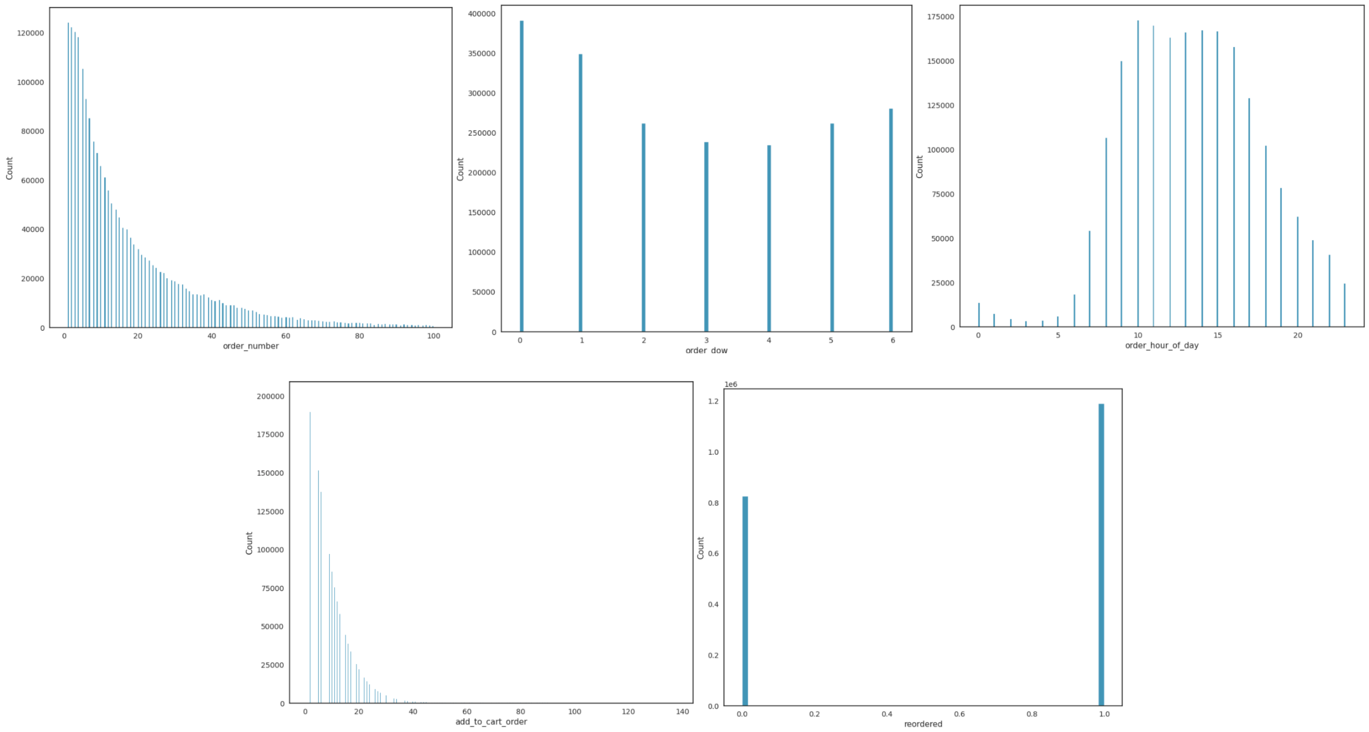
For further study on dataset, I’ve decided to seek missing data percentage. The results has been shown in following table

Feature	Percent Missing
order_id	0.000000
user_id	0.000000
order_number	0.000000
order_dow	0.000000
order_hour_of_day	0.000000
product_id	0.000000
add_to_cart_order	0.000000
reordered	0.000000
department_id	0.000000
department	0.000000
product_name	0.000000

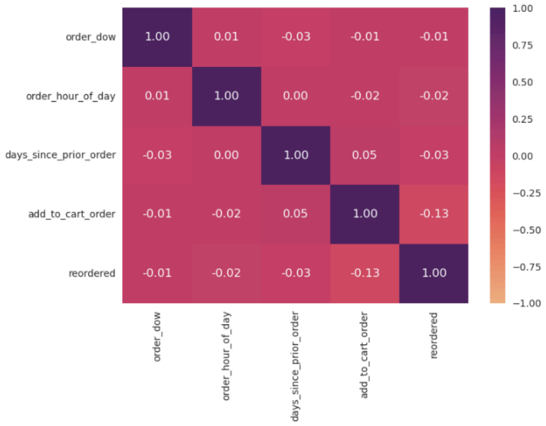
Feature	Percent Missing
days_since_prior_order	6.157066

That means we have a clean dataset. The days_since_prior_order feature indicates number of days since prior order which nan in it means zero. Due to this fact I substituted all nan values by zero.

In this dataset, there are 21 unique departments, 134 unique products and 105273 unique customers which gave us a vast distribution over consumer behaviour. In following, histogram plots of features have been shown:



At the end of this section, we shall review correlation matrix of this dataset to find highly correlated or weakly correlation between features:



As you can see, features of dataset are not strongly correlated and we can assume we are dealing with an sparse feature space.

K-Means and Elbow Approach:

K-means clustering is a popular unsupervised machine learning algorithm that is used for grouping similar data points together in a dataset. The goal of K-means clustering is to partition a given dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean value. The algorithm works by randomly selecting K centroids (cluster centers) from the dataset and iteratively assigning each data point to the closest centroid while updating the centroid based on the new assignments until convergence. K-means clustering can be applied to a wide range of fields such as image processing, natural language processing, and market segmentation. However, it is important to note that the quality of the clustering result heavily depends on the initial choice of centroids and the number of clusters chosen.

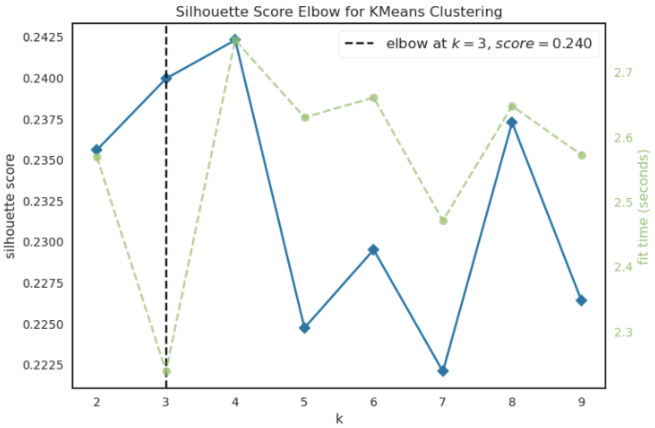
The elbow method is a popular approach used to determine the optimal number of clusters in K-means clustering. The idea behind this method is that the sum of squared distances between data points and their corresponding cluster centroids decreases as the number of clusters increases. However, adding too many clusters can lead to overfitting and reduced interpretability.

To use the elbow method, we plot the sum of squared distances against the number of clusters, usually denoted as K. We then look for the "elbow" point on the curve, which represents the point of diminishing returns in terms of reducing the sum of squared distances. The optimal number of clusters is usually chosen at the "elbow" point, where adding more clusters does not lead to a significant improvement in the quality of clustering.

Another way to evaluate the quality of clustering is to use silhouette analysis, which measures how well each data point fits into its assigned cluster compared to other nearby clusters. The silhouette score ranges from -1 to 1, where a score closer to 1 indicates good clustering while a score closer to -1 indicates poor clustering. By computing the average silhouette score for different values of K, we can determine the optimal number of clusters that maximizes the overall quality of clustering.

In practice, the elbow method and silhouette analysis are often used together to determine the optimal number of clusters in K-means clustering.

According to given facts, I've used 'KElbowVisualizer' method to search in interval $k \in (2,10)$ to find optimum number of clusters with respect to silhouette score. This fact must be noted that we fit k-means on standard scaled data.



As you can see in 'KElbowVisualizer', fit time is a criterion like silhouette score. Now we can take best silhouette score with high time trade off or take $k = 3$ as best silhouette-time tradeoff.

After fitting data on a k-means clustering with 4 classes, the number of datapoints in each cluster in a 10000 sample is

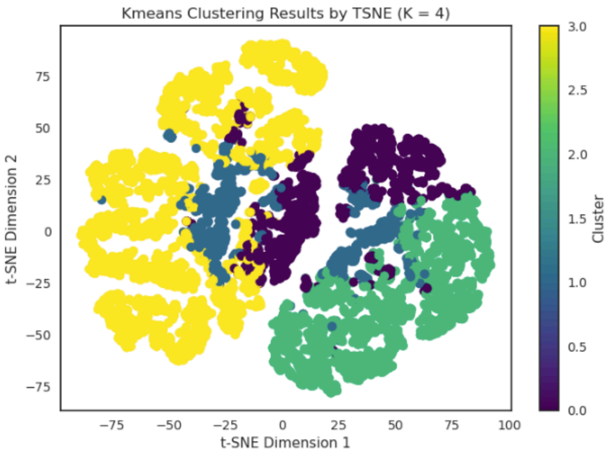
Cluster Number	Number of Samples
0	1773
1	1083

Cluster Number	Number of Samples
2	2730
3	4414

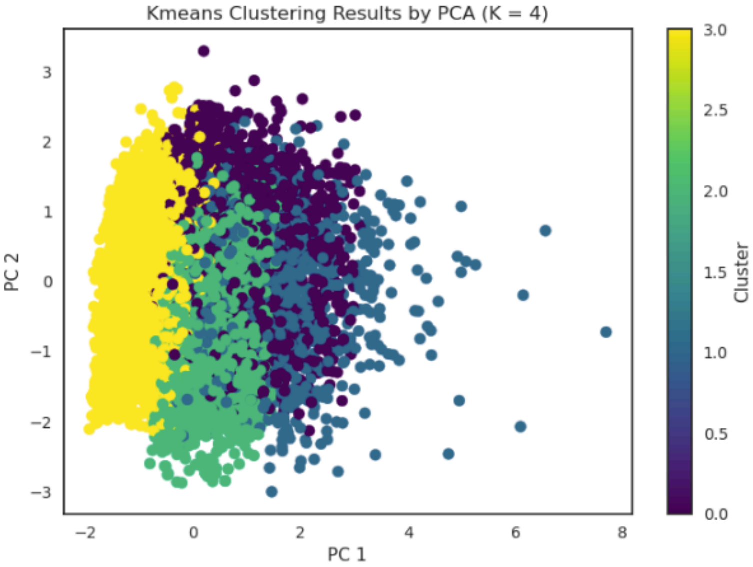
As you can see, we have an imbalanced data clustering and its biased on last cluster.

Clustering Visualization:

In this section, visualization over clustered sample dataset will be shown. Its worth to mention that this dataset is not available to plot in 2D or 3D plot due to high dimension. In first plot I've used TSNE to reduce dimension to two features and plot it with their clusters as label. After it I've used Principal Component Analysis to reduce dimension of dataset into two features and plot their clusters as label.



As you can see in TSNE dimension reduction, we can't specify a characteristic for each class because they have overlap in many data points but we can give a general point of view over each classes' characteristics. The third and second clusters are in most left and most right side of this plot which can give us a bound for dataset. The zeroth and first clusters a in middle of this plot which indicates they contains main core of dataset. Note that one of the reasons behind cluster overlapping is the fact that we used k-means in higher dimension and then visualize it on lower dimension.



On the other hand, in PCA dimension reduction algorithm, we found an interesting results. As you can see in plot, cluster is independent from second principal component and we can say it doesn't play a role in indicating cluster of a datapoint. In opposite, first principal component plays a crucial rule in cluster indication. As you can see in the lowest PC1, we have a dense region of the third cluster, then we have second cluster in PC1=0 and after that second and first clusters came by.

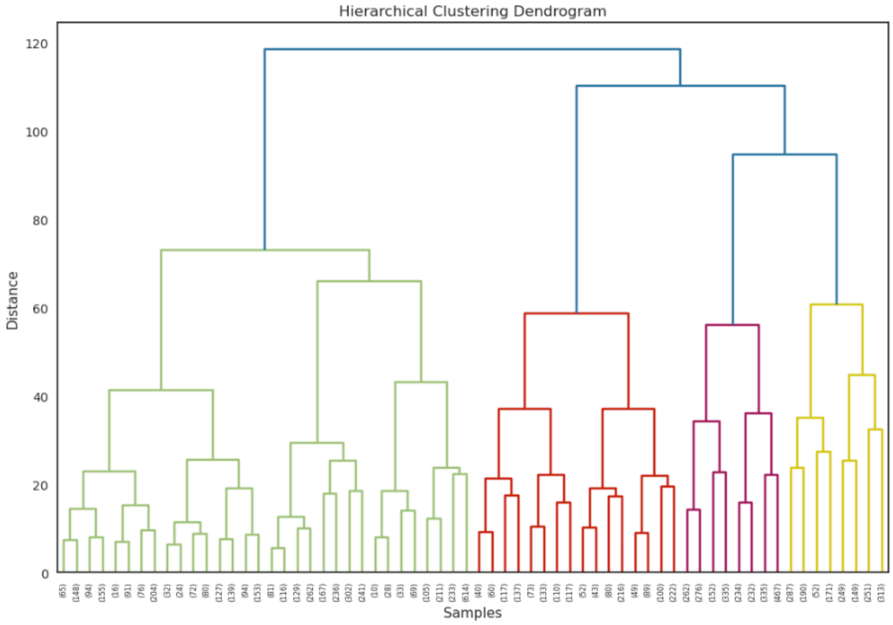
Hierarchical Clustering:

Agglomerative clustering algorithm is a bottom-up hierarchical clustering method that starts with each data point as its own cluster and iteratively merges the closest pair of clusters until all data points belong to a single cluster. The algorithm builds the hierarchy of clusters in a dendrogram, which represents the history of cluster merging. Agglomerative clustering can be performed using different distance metrics and linkage criteria, such as single linkage, complete linkage, or average linkage. The choice of distance metric and linkage criterion affects the cluster structure and can have a significant impact on the results. Agglomerative clustering is widely used in various fields, including biology, social science, and computer science, for exploring patterns in high-dimensional data and identifying natural groupings.

For this task, we used same 10000 samples from previous section and fitted agglomerative clustering for $k \in (3,6)$. The silhouette score for each clustering is

Number of Clusters	Silhouette Score
3	0.189455014817243
4	0.196862410263609
5	0.149254030166196
6	0.141957735661436

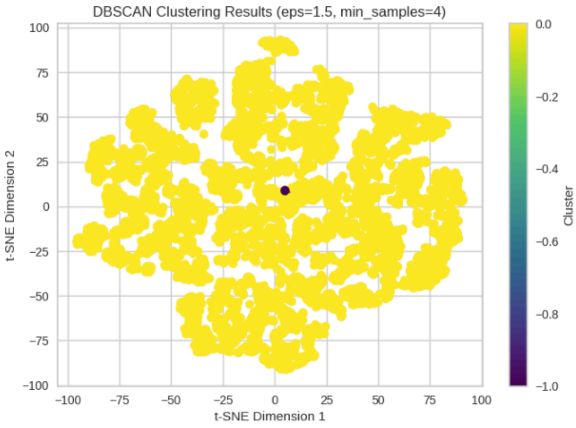
According to this table, the best number of clusters in 4 with 0.1969 as silhouette score. Here is dendrogram plot for four clusters

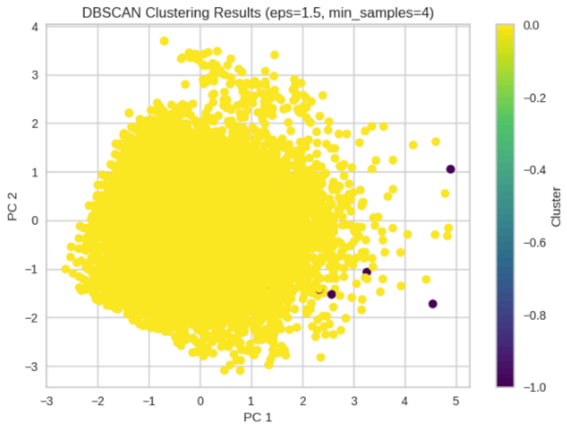


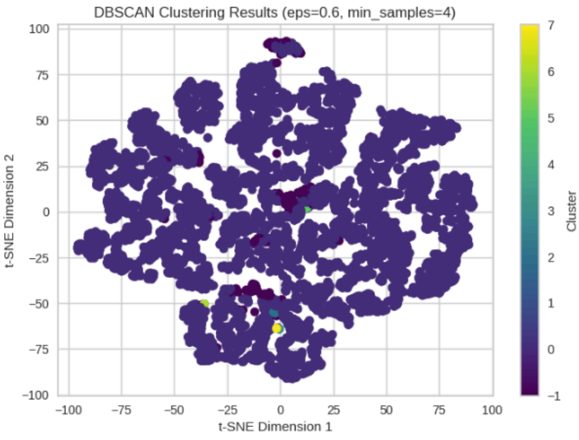
Density Based Clustering:

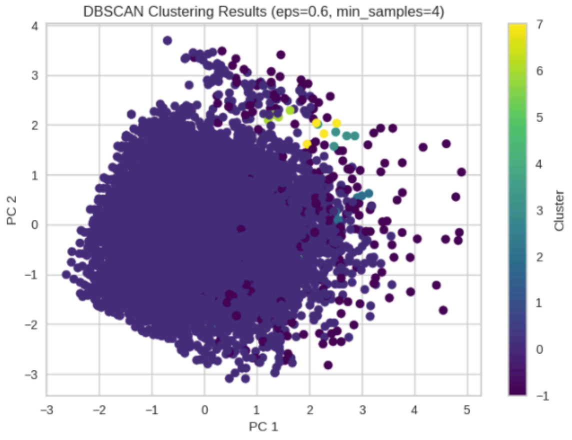
Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a popular clustering algorithm used in machine learning and data mining. The algorithm is designed to cluster together points that are close together in space, while identifying outliers as noise. It works by defining a neighborhood around each point and grouping together all the points within that neighborhood that meet certain density criteria. DBSCAN has several advantages over other clustering algorithms, such as being able to handle clusters of varying shapes and sizes, not requiring the number of clusters to be specified beforehand, and being robust to noise and outliers in the data. Despite its strengths, DBSCAN can also be sensitive to the choice of parameters and may struggle with high-dimensional data.

For this task, I've consider two epsilons with same min samples for DBSCAN. For $\epsilon = 0.6$ and min_samples = 4, silhouette score comes around 0.06 and for $\epsilon = 1.5$ and min_samples = 4 silhouette score comes to 0.66.





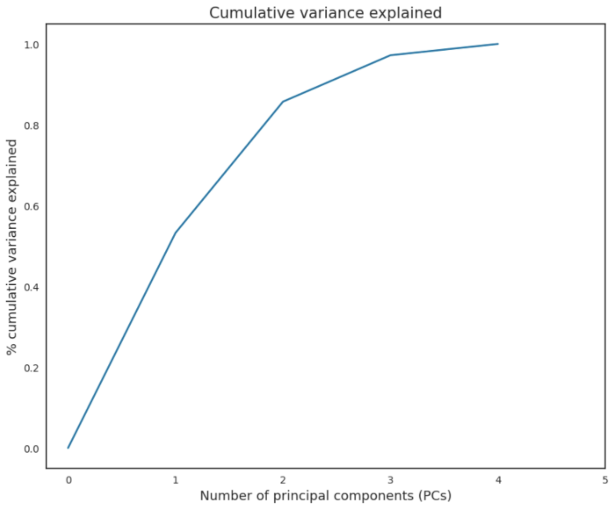




Clustering after Dimension Reduction:

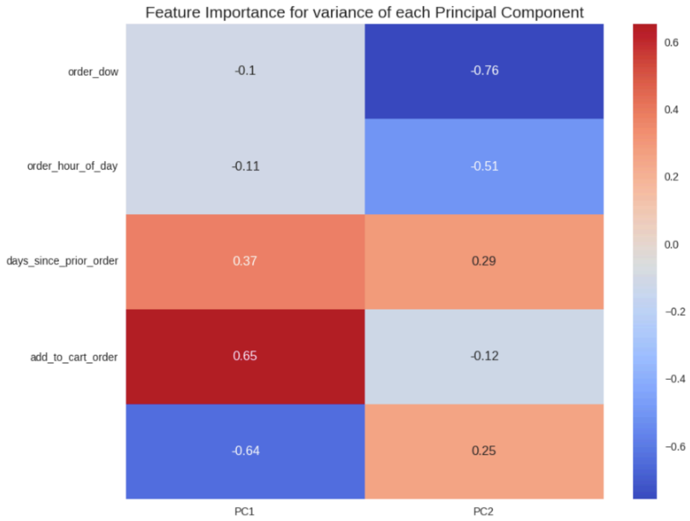
To explore clustering after dimension reduction, we shall know cumulative variance explained on dataset with different principal components and their affect on clustering.

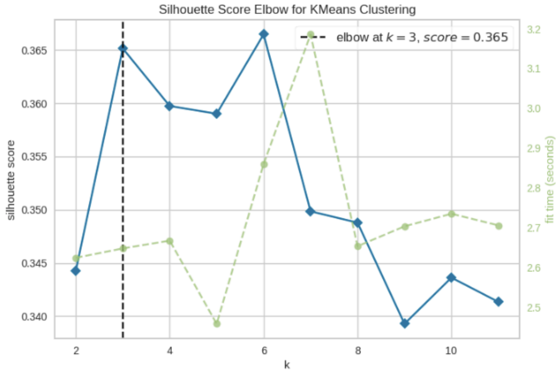
PCA, or principal component analysis, is a popular technique used in data science and machine learning for dimensionality reduction. One important aspect of PCA is the concept of cumulative variance explained, which measures how much of the variability in the original dataset can be captured by each successive principal component. The cumulative variance explained tells us how much information we can retain while reducing the dimensionality of the data. As we add more principal components, the cumulative variance explained increases, and at some point, we may find that adding more components doesn't significantly increase the amount of variance explained. By examining the cumulative variance explained, we can determine the optimal number of principal components to keep in order to capture the majority of the information in the original dataset.



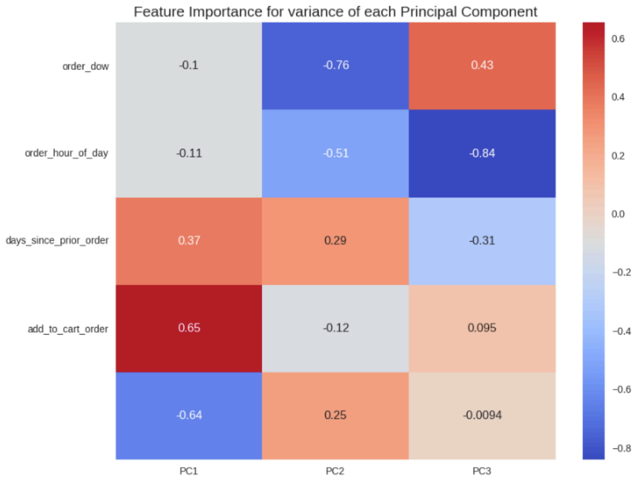
As you can see in above plot, after two principal component in PCA dimension reduction, the cumulative variance explained is more than 80% which means we have 80% distribution of our data with losing three features.

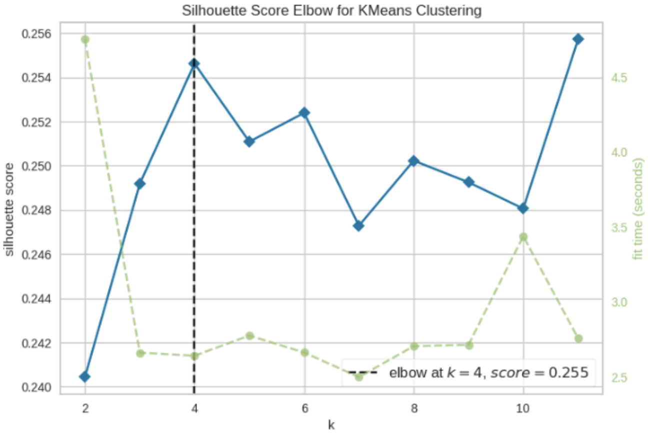
Further more, I’ve tested two different PCAs with number of components = 2 and 3 and studied their affect on clustering procedure. First we will see the affect of two principal components.





As you can see, the overall silhouette score has been increased over second order PCA. Now let’s see results for three principal components.

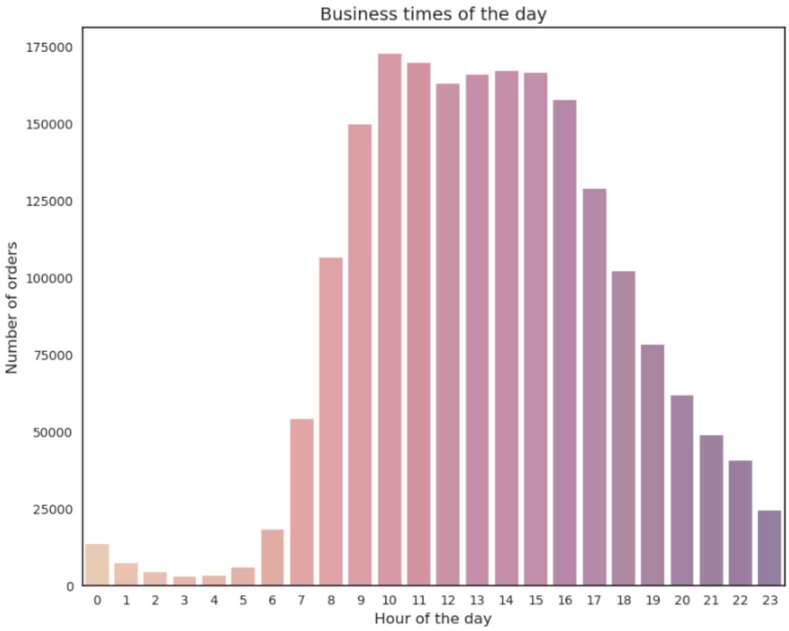




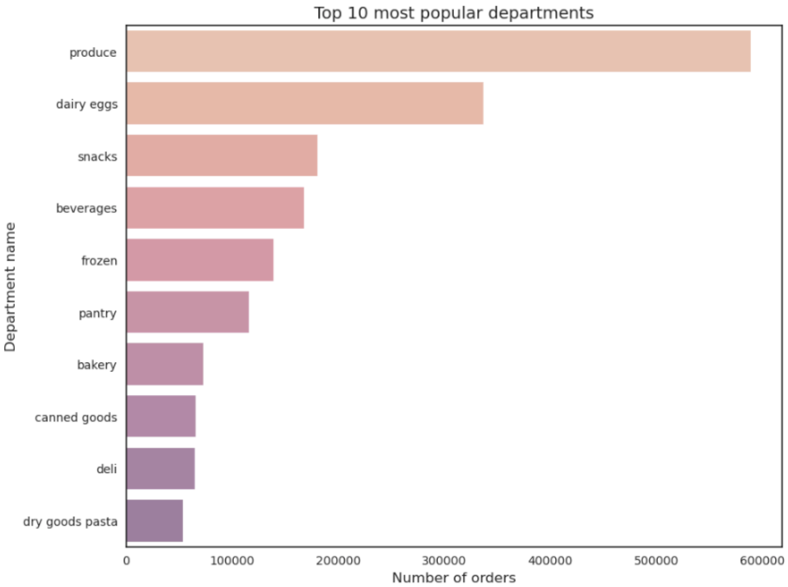
As you can see, the silhouette score has been decreased due to this fact, the three principal components can’t be a good choice for this task.

Results Interpretation:

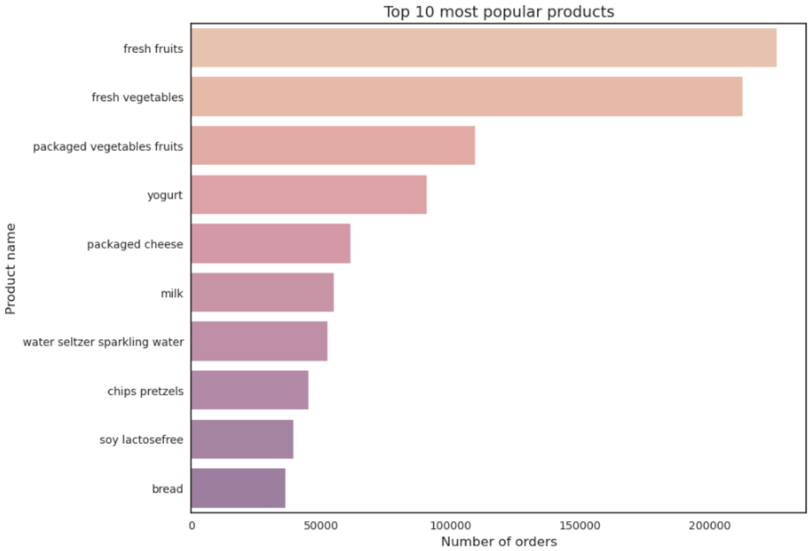
Data science results interpretation involves analyzing the output from data models such as regression, classification or clustering to make informed decisions. It is important to understand the assumptions and limitations of the model, as well as the accuracy and statistical significance of the results. Visualizations such as graphs and charts can help in communicating insights and patterns in the data. It is also critical to consider the context in which the model was created and how it may or may not generalize to new situations. Finally, data science results interpretation must be communicated effectively to stakeholders to enable informed decision-making.



According to this plot, the busiest time of this grocery store is between 9 and 17 o’clock. That means if they want to get hourly wage employers they can set them between 9 and 17 for the best performance.



As you can see in this plot, the top 10 most popular departments are shown. That means this departments are playing very crucial role in this store's financial terms. These departments should be sharp and always ready to serve.

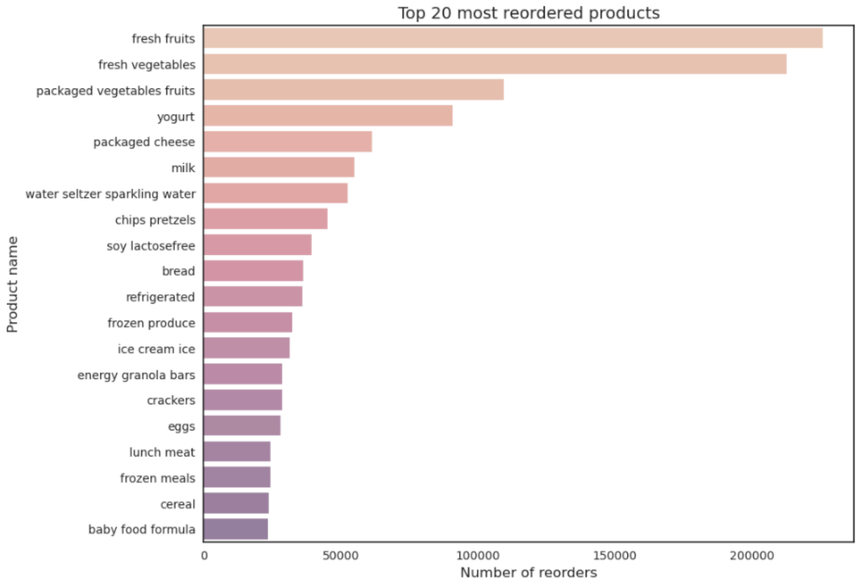


In this plot, the top 10 most popular products has been shown. These products are very important for this business and if store can't fulfil customers needs in these products, it will lose so much money. That means store should always reordered these products and try to provide vast number of companies for each product.

Department	Most Popular Item
alcohol	beers coolers
babies	baby food formula
bakery	bread
beverages	water seltzer sparkling water
breakfast	cereal
bulk	bulk grains rice dried goods
canned goods	soup broth bouillon
dairy eggs	yogurt
deli	lunch meat
dry goods pasta	dry pasta

Department	Most Popular Item
frozen	frozen produce
household	paper goods
international	asian foods
meat seafood	hot dogs bacon sausage
pantry	baking ingredients
personal care	oral hygiene
pets	cat food care
produce	fresh fruits
snacks	chips pretzels

This is the list of most popular product in each department. Departments should consider these products as the most profitable product in their department and invest a large amount of money on them.



In this plot, the top 20 most reordered products has been shown. These products must always be available in store and recharged every day due to high demand to them. Lack of products in this list will lose a lot of customers and money for this store.