

# **CUSTOMER SEGMENTATION ANALYSIS**

**Project report in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted By**

Nirbhay Kumar Jha  
Sourav Kumar Das  
Chinmayi  
Akash Kumar Gupta  
Prithviraj Iyer  
Soumili Chatterjee

University Roll No. 12018009019062  
University Roll No. 12018009019400  
University Roll No. 12018009019646  
University Roll No. 12018009019596  
University Roll No. 12018009019092  
University Roll No. 12018009019697

**Under the guidance of**

**Prof. Arunabha Tarafdar**

**&**

**Prof. (Dr.) Sayan Sikdar**

**Department of Computer Science**



**UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA**

**University Area, Plot No. III – B/5, New Town, Action Area – III, Kolkata – 700160.**

## CERTIFICATE

This is to certify that the project titled **CUSTOMER SEGMENTATION ANALYSIS** submitted by **Nirbhay Kumar Jha (University Roll No. 12018009019062)**, **Sourav Kumar Das (University Roll No. 12018009019400)**, **Chinmayi (University Roll No. 12018009019646)**, **Akash Kumar Gupta (University Roll No. 12018009019596)**, **Prithviraj Iyer (University Roll No. 12018009019092)** and **Soumili Chatterjee (University Roll No. 1201800901697)** students of **UNIVERSITY OF ENGINEERING & MANAGEMENT, KOLKATA**, in partial fulfillment of requirement for the degree of Bachelor of Computer Science is a bonafide work carried out by them under the supervision and guidance of **Prof. Arunabha Tarafdar & Prof. (Dr.) Sayan Sikdar** during 8<sup>th</sup> Semester of academic session of 2018- 2022. The content of this report has not been submitted to any other university or institute. We are glad to inform that the work is entirely original and its performance is found to be quite satisfactory.

---

Signature of Guide

---

Signature of Guide

---

Signature of Head of the Department

## **ACKNOWLEDGEMENT**

We would like to take this opportunity to thank everyone whose cooperation and encouragement throughout the ongoing course of this project remains invaluable to us.

We are sincerely grateful to our guide Prof. Arunabha Tarafdar and Prof.(Dr.) Sayan Sikdar of the Department of Computer Science, UEM, Kolkata, for their wisdom, guidance and inspiration that helped us to go through with this project and take it to where it stands now.

Last but not the least, we would like to extend our warm regards to our families and peers who have kept supporting us and always had faith in our work.

**Nirbhay Kumar Jha**

**Sourav Kumar Das**

**Chinmayi**

**Akash Kumar Gupta**

**Prithviraj Iyer**

**Soumili Chatterjee**

## **TABLE OF CONTENTS**

**ABSTRACT..... <<PAGE NO 5>>**

**CHAPTER – 1: INTRODUCTION..... <<PAGE NO 6>>**

**CHAPTER – 2: LITERATURE SURVEY**

**<< Customer Segmentation>>..... <<PAGE NO 7>>**

**<< Clustering,K-Means Algorithm and Gaussian Mixture Model clusteringalgorithm.>>PAGE  
NO 7>>**

**CHAPTER – 3: PROBLEM STATEMENT.....<<PAGE NO 8>>**

**CHAPTER – 4: PROPOSED SOLUTION.....<< PAGE NO 8>>**

**CHAPTER – 5 : EXPERIMENTAL SETUP AND RESULT ANALYSIS..<<PAGE NO 9-19>>**

**CHAPTER – 6 : CONCLUSION & FUTURE SCOPE.....<<PAGE NO 20>>**

**BIBLIOGRAPHY..... <<PAGE NO 21>>**

## **ABSTRACT**

We live in a world where large and vast amount of data is collected daily. Analysing such data is an important need. In the modern era of innovation, where there is a large competition to be better than everyone, the business strategy needs to be according to the modern conditions. The business done today runs on the basis of innovative ideas as there are large number of potential customers who are confounded to what to buy and what not to buy. The companies doing the business are also not able to diagnose the target potential customers. This is where the machine learning comes into picture, the various algorithms are applied to identify the hidden patterns in the data for better decision making. The concept of which customer segment to target is done using the customer segmentation process using the clustering technique. In this project we have used K-Means clustering algorithm as well as Gaussian Mixture Model clustering algorithm.

## **INTRODUCTION**

Over the years, the competition amongst businesses is increased and the large historical data that is available has resulted in the widespread use of data mining techniques in extracting the meaningful and strategic information from the database of the organisation. Data mining is the process where methods are applied to extract data patterns in order to present it in the human readable format which can be used for the purpose of decision support. According to, Clustering techniques consider data tuples as objects. They partition the data objects into groups or clusters, so that objects within a cluster are similar to one another and dissimilar to objects in other clusters. Customer Segmentation is the process of division of customer base into several groups called as customer segments such that each customer segment consists of customers who have similar characteristics. The segmentation is based on the similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits. The customer segmentation has the importance as it includes, the ability to modify the programs of market so that it is suitable to each of the customer segment, support in business decision; identification of products associated with each customer segment and to manage the demand and supply of that product; identifying and targeting the potential customer base, and predicting customer defection, providing directions in finding the solutions. The thrust of this paper is to identify customer segments using the data mining approach, using the partitioning algorithm called as K-means clustering algorithm and Gaussian Mixture Model clustering algorithm.

## LITERATURE SURVEY

### **Customer Segmentation**

Over the years, as there is very strong competition in the business world, the organizations have to enhance their profits and business by satisfying the demands of their customers and attract new customers according to their needs. The identification of customers and satisfying the demands of each customer is a very complex and tedious task. This is because customers may be different according to their demands, tastes, preferences and so on. Instead of “one-size-fits-all” approach, customer segmentation clusters the customers into groups sharing the same properties or behavioural characteristics. According to, customer segmentation is a strategy of dividing the market into homogenous groups. The data used in customer segmentation technique that divides the customers into groups depends on various factors like, data geographical conditions, economic conditions, demographical conditions as well as behavioural patterns. The customer segmentation technique allows the business to make better use of their marketing budgets, gain a competitive edge over their rival companies, demonstrating the better knowledge of the needs of the customer. It also helps an organization in, increasing their marketing efficiency, determining new market opportunities, making better brand strategy, identifying customers retention.

### **Clustering, K-Means Algorithm and Gaussian Mixture Model clustering algorithm.**

**Clustering algorithms** generates clusters such that within the clusters are similar based on some characteristics. Similarity is defined in terms of how close the objects are in space.

**K-means algorithm** in one of the most popular centroid based algorithm. Suppose data set,  $D$ , contains  $n$  objects in space. Partitioning methods distribute the objects in  $D$  into  $k$  clusters,  $C_1, \dots, C_k$ , that is,  $C_i \subset D$  and  $C_i \cap C_j = \emptyset$  for  $(1 \leq i, j \leq k)$ . A centroid-based partitioning technique uses the centroid of a cluster,  $C_i$ , to represent that cluster. Conceptually, the centroid of a cluster is its center point. The difference between an object  $p \in C_i$  and  $c_i$ , the representative of the cluster, is measured by  $\text{dist}(p, c_i)$ , where  $\text{dist}(x, y)$  is the Euclidean distance between two points  $x$  and  $y$ .

**Gaussian Mixture Models (GMMs)** assume that there are a certain number of Gaussian distributions, and each of these distributions represent a cluster. Hence, a Gaussian Mixture Model tends to group the data points belonging to a single distribution together.

Let's say we have three Gaussian distributions (more on that in the next section) –  $GD_1$ ,  $GD_2$ , and  $GD_3$ . These have a certain mean ( $\mu_1, \mu_2, \mu_3$ ) and variance ( $\sigma_1, \sigma_2, \sigma_3$ ) value respectively. For a given set of data points, our GMM would identify the probability of each data point belonging to each of these distributions.

Gaussian Mixture Models are probabilistic models and use the soft clustering approach for distributing the points in different clusters.

## **PROBLEM STATEMENT**

One of the biggest issues with customer segmentation is data quality, inaccurate data in source systems will usually result in poor grouping. For example, customers who are individuals having attributes like age, gender, and marital status are frequently used. If these attributes are not maintained properly, the segments will be inaccurate and as a result, the information likely to be less useful. If the business owner does not feel comfortable with the quality of the data, they are not likely to use the segmentation as well as the quality of segmentation will not fulfill their demand, they will not get proper insight of their customers' segment. Data quality issues also arise from a lack of maintenance and regular cleansing to ensure accuracy.

## **PROPOSED SOLUTION**

There are some processes that can be implemented to provide improved data quality for customer segmentation maintenance. One of the important aspects of data quality is the concept of assigning resources to manage attributes for customers. This resource, usually called data stewards, is responsible for managing the set up of a new customer, making sure all critical attributes are provided before the customer is set up and maintenance begins.

Business intelligence tools are often used to analyze customer segmentation along with other information such as sales, marketing, and trends. As stated before, using data integration tools and defining the rules for data transformations creates a single process that is consistent and easy to manage. End users must be adequately trained in not only the tool, but also the metadata that represents the attributes that are being used for analytics. When the users are comfortable with the information, advanced analytics can be used to predict future behaviors and perform what-if scenarios with the data.



## EXPERIMENTAL SETUP AND RESULT ANALYSIS

### Getting Started With Data:

To start, we'll get need some orders to evaluate. If you'd like to follow along, we will be using the **wholesale** data set, which has already been retrieved .Next, we'll get the data into a usable format, typical of an SQL query.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	12669	9656	7561	2142	2674	1338
1	7057	9810	9568	1762	3293	1776
2	6353	8808	7684	2405	3516	7844
3	13265	1196	4221	6404	5073	1788
4	22615	5410	7198	3915	1777	5185

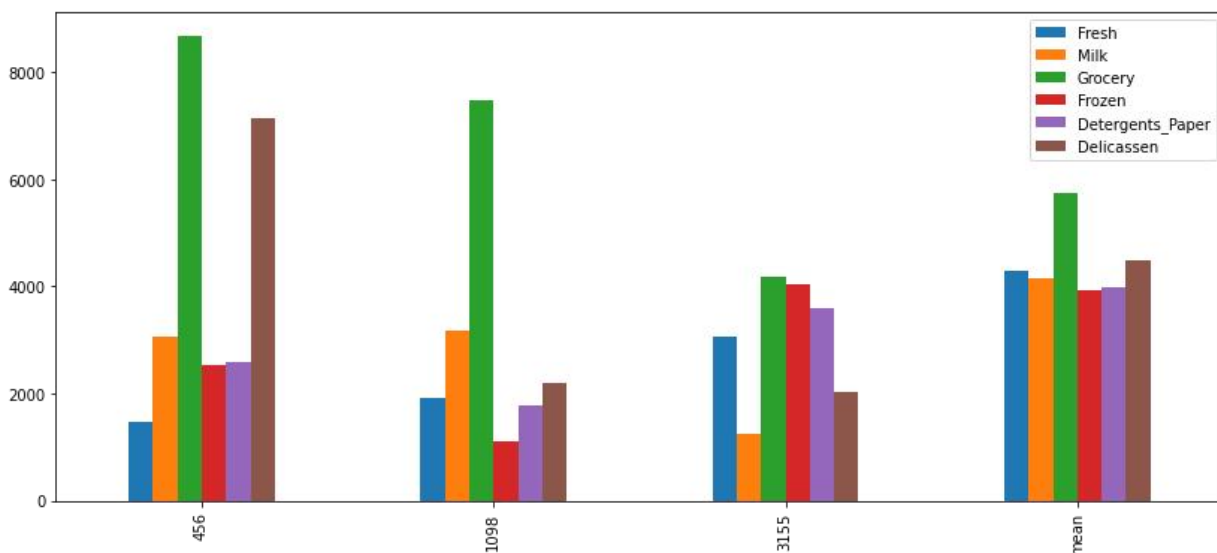
### Selecting samples

To get a better understanding of the customers and how their data will transform through the analysis, we will select a few sample data points according to the above common sense segmentation and explore them in more detail. The selected samples should vary significantly from one another.

```
fresh_q1 = 3750.75
milk_q1 = 1533
grocery_q1 = 2153
frozen_q1 = 7423.25
deter_q1 = 256.75
deli_q1 = 4068.25

fresh_q3 = 1693.75
milk_q3 = 7190.25
grocery_q3 = 10655.75
frozen_q3 = 3554.25
deter_q3 = 3922
deli_q3 = 2670.25
```

### Visualizing samples



## Feature relevance

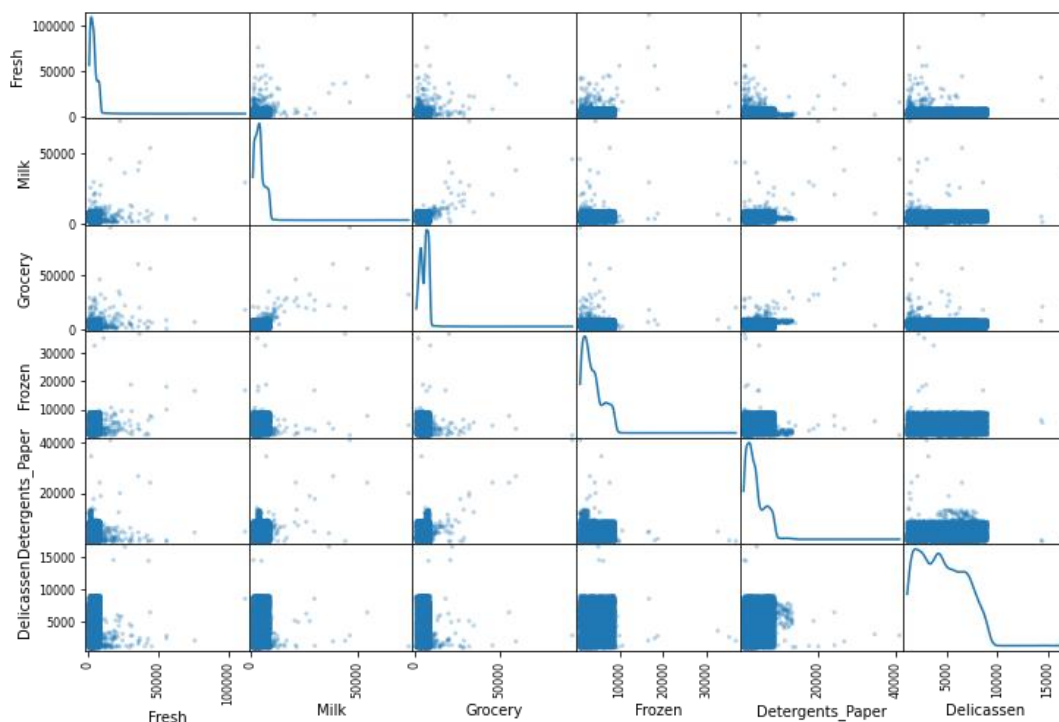
One interesting thought to consider is if one (or more) of the six product categories is actually relevant for understanding customer purchasing. That is to say, is it possible to determine whether customers purchasing some amount of one category of products will necessarily purchase some proportional amount of another category of products? We can make this determination quite easily by training a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature.

```
➞ R2 score for Fresh as dependent variable: -0.24781311471015233
   R2 score for Milk as dependent variable: -0.21167454516701256
   R2 score for Grocery as dependent variable: 0.0627272689106666
   R2 score for Frozen as dependent variable: -0.22414346260845996
   R2 score for Detergents_Paper as dependent variable: -0.3711439217311465
   R2 score for Delicassen as dependent variable: -0.7178921998441712
```

We used a loop and predicted every single feature as a dependent variable with the results shown above. As you can see, "Fresh", "Frozen" and "Delicassen" as dependent variables have negative R2 scores. Their negative scores imply that they are necessary for identifying customers' spending habits because the remaining features cannot explain the variation in them. Similarly, "Milk" and "Detergents\_Paper" have very low R2 scores. Their low scores also imply that they are necessary for identifying customers' spending habits. However, "Grocery" has a R2 score of 0.62. It is relative to the others it is much higher. It may be not as necessary, compared to the other features, for identifying customers' spending habits. We will explore this further.

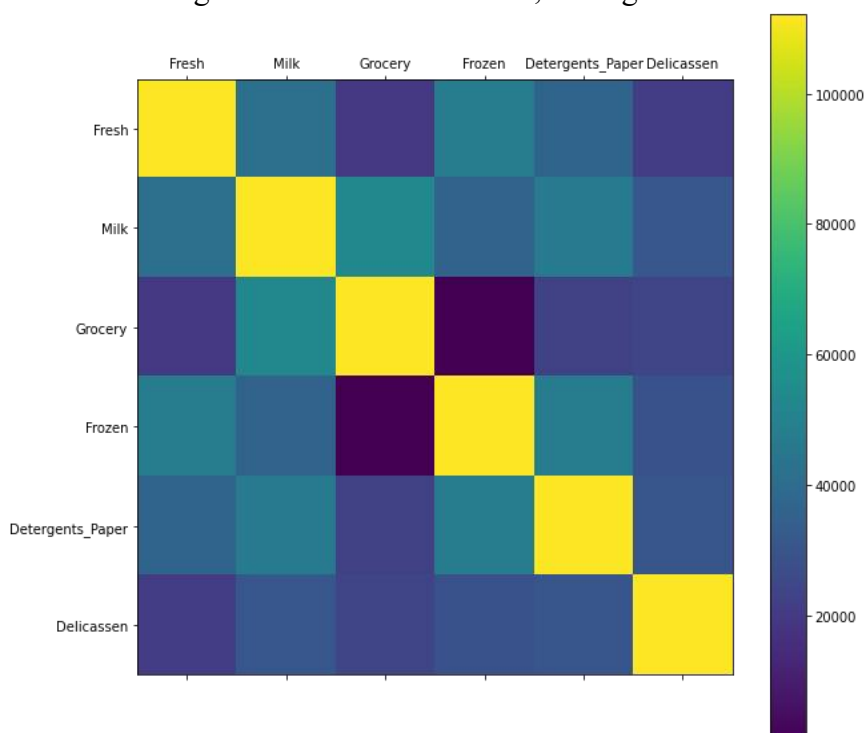
## Visualize Feature Distributions¶

To get a better understanding of the dataset, we can construct a scatter matrix of each of the six product features present in the data. If you found that the feature you attempted to predict above is relevant for identifying a specific customer, then the scatter matrix below may not show any correlation between that feature and the others. Conversely, if you believe that feature is not relevant for identifying a specific customer, the scatter matrix might show a correlation between that feature and another feature in the data. Run the code block below to produce a scatter matrix.



## Correlation Matrix

- This is to cross-reference with the scatter matrix above to draw more accurate insights from the data.
- The higher the color is on the bar, the higher the correlation.



We have plotted a correlation matrix to compare with the scatter matrix to ensure this answer is as accurate as possible.

The follow pairs of features seem to have some correlation as observed from the scatter plot showing a linear trend and the correlation plot showing a high correlation between the two features. I have ranked them in order of correlation from strongest to weakest:

\* **Grocery and Milk.**

\* **Grocery and Detergents\_Paper.**

\* **Grocery and Frozen (not too strong).**

These features that are strongly correlated does lend credence to our initial claim that Grocery may not be necessary for identifying customers' spending habits.

Grocery has a high correlation with Detergents\_Paper and Milk that corresponds to a relatively high R2 score when we regress Grocery on all other features.

The data are not normally distributed due to the presence of many outliers. Evidently,

most are skewed to the left where most of the data points lie.

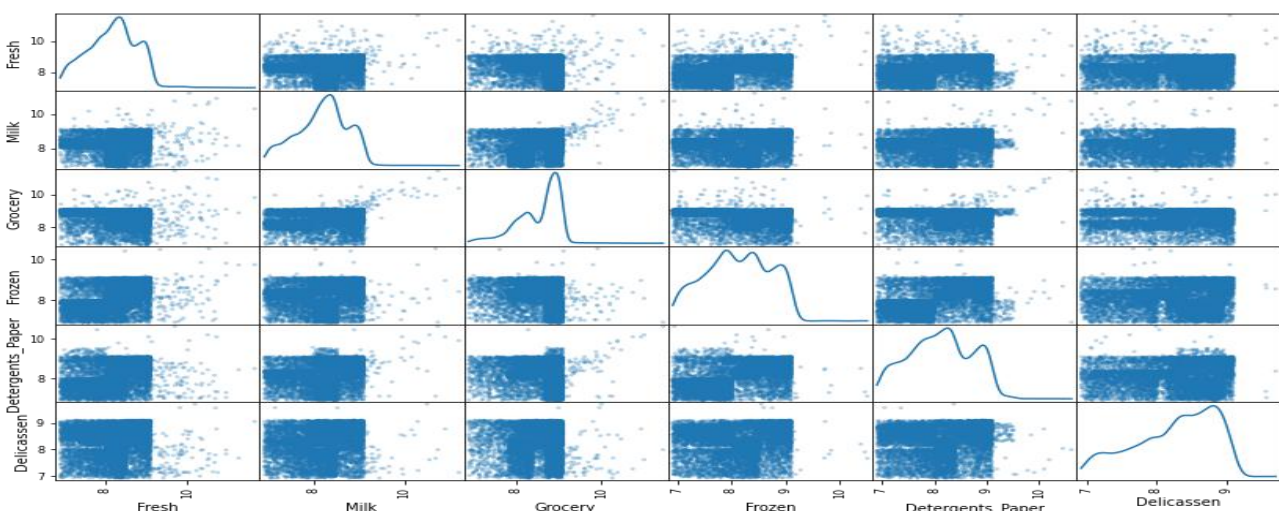
This indicates how normalization is required to make the data features normally distributed as clustering algorithms require them to be normally distributed.

## Data Preprocessing

Now we will preprocess the data to create a better representation of customers by performing a scaling on the data and detecting (and optionally removing) outliers. Preprocessing data is often times a critical step in assuring that results you obtain from your analysis are significant and meaningful.

### Implementation: Feature Scaling

If data is not normally distributed, especially if the mean and median vary significantly (indicating a large skew), it is most often appropriate to apply a non-linear scaling — particularly for financial data. One way to achieve this scaling is by using a Box-Cox test, which calculates the best power transformation of the data that reduces skewness. A simpler approach which can work in most cases would be applying the natural logarithm.



## Outliers Detection

Detecting outliers in the data is extremely important in the data preprocessing step of any analysis. The presence of outliers can often skew results which take into consideration these data points. There are many "rules of thumb" for what constitutes an outlier in a dataset. Here, we will use Tukey's Method for identifying outliers: An outlier step is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal.

Data points considered outliers for the feature 'Fresh':

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
4	10.026369	8.596004	8.881558	8.272571	7.482682	8.553525
12	10.364514	9.418898	9.372204	7.160069	8.263848	7.983099
13	9.962558	8.733594	9.614605	8.037543	8.810907	8.703507
14	10.112654	9.155356	9.400217	7.986505	8.528726	7.681560
22	10.350606	7.558517	8.404920	9.149316	7.775276	8.374246
23	10.180096	10.502956	9.999661	8.547528	8.374938	9.712509
24	10.027783	9.187686	9.531844	7.977625	8.407825	8.661813

Data points considered outliers for the feature 'Milk':

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
23	10.180096	10.502956	9.999661	8.547528	8.374938	9.712509
28	8.321908	9.927399	10.164197	7.054450	9.059982	8.557567
45	8.552753	10.000796	9.977249	7.461640	8.902864	8.514189
47	10.702480	10.901524	10.925417	8.959569	10.092909	8.774158
49	8.510571	9.971707	10.272323	7.494430	9.516574	7.058758
56	8.318254	10.305346	10.198617	7.869402	9.783577	7.200425
61	10.489662	10.555005	10.995377	8.087640	10.192456	7.609367
65	9.054037	9.950323	10.732651	8.380915	10.095388	7.260523
85	9.687630	10.740670	11.437986	6.933423	10.617099	7.987524

Data points considered outliers for the feature 'Grocery':

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
28	8.321908	9.927399	10.164197	7.054450	9.059982	8.557567
39	10.935942	8.621733	7.021976	9.210540	7.058758	7.977968
43	8.748464	9.314250	10.085726	8.170751	9.162095	8.086718
47	10.702480	10.901524	10.925417	8.959569	10.092909	8.774158
49	8.510571	9.971707	10.272323	7.494430	9.516574	7.058758
...	...	...	...	...	...	...
6841	8.199464	8.954028	6.973543	8.988755	8.665096	7.473069
6881	7.178545	8.408048	6.998510	8.250359	8.900822	8.425078
6883	8.124743	8.884472	6.984716	8.805975	9.085117	8.975124
6885	8.799360	8.946245	6.955593	8.147867	9.083529	9.053920
6906	9.069353	8.461046	7.057037	7.180831	7.970049	8.437717

Data points considered outliers for the feature 'Frozen':

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
93	9.333796	8.035926	7.631432	10.463360	7.334329	7.900266
166	8.480944	8.812992	8.123693	10.386901	8.511779	8.198914
183	10.514529	10.680808	9.911952	10.505999	7.123673	7.659643

Data points considered outliers for the feature 'Detergents\_Paper':

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
47	10.702480	10.901524	10.925417	8.959569	10.092909	8.774158
61	10.489662	10.555005	10.995377	8.087640	10.192456	7.609367
65	9.054037	9.950323	10.732651	8.380915	10.095388	7.260523
68	7.802209	8.890135	8.292298	8.677610	10.449960	8.037543
85	9.687630	10.740670	11.437986	6.933423	10.617099	7.987524

Data points considered outliers for the feature 'Delicassen':

Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
[4, 12, 13, 14, 22, 23, 23, 24, 28, 28, 29, 32, 33, 36, 39, 39, 40, 41, 43, 45, 47, 47, 47, 47, 49, 49, 52, 54, 56, 56, 61, 61, 61, 61, 65, 65, 65]					
[23, 28, 39, 47, 49, 56, 61, 65, 75, 85, 86, 92, 109, 145, 181, 183]					
(6999, 6)					
(6983, 6)					

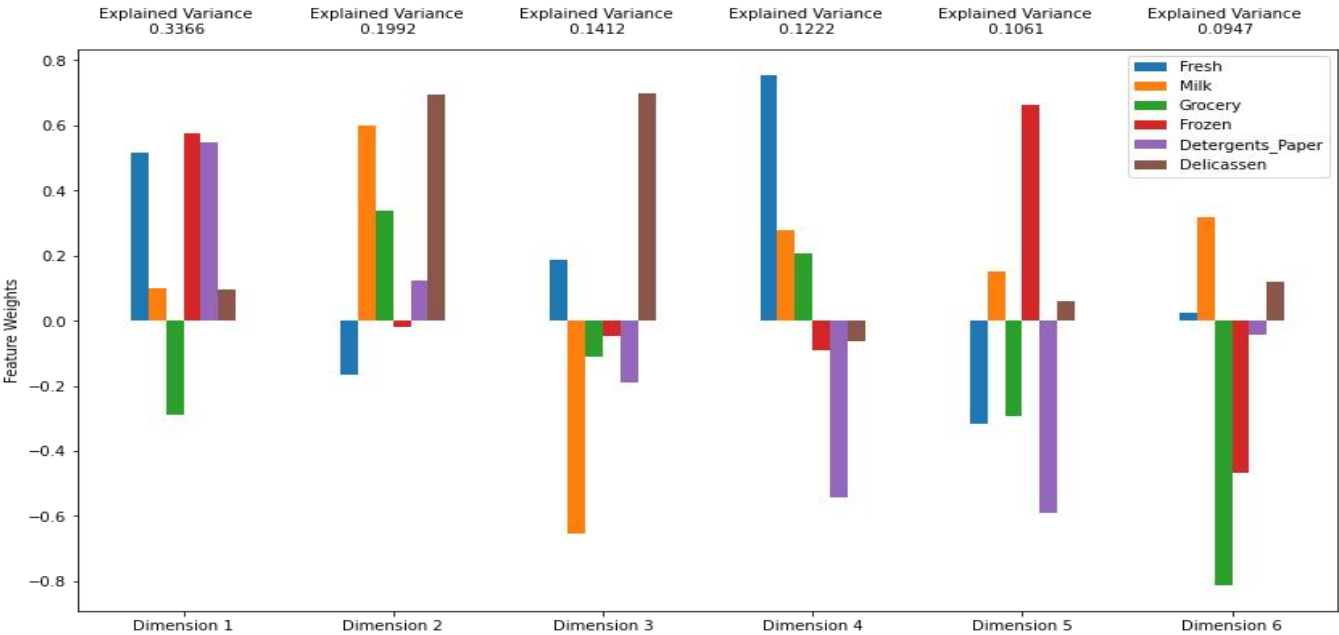
We can see that with this method 1/2 of the points are considered as outliers and I don't like to remove so many data points so I have implemented another method Z score (based on standard deviation), I have considered the data outside the 2sigma as outliers this will remove 54 outlier points from dataset. We have to sometime careful as Outliers, being the most extreme observations, may include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum are not always outliers because they may not be unusually far from other observations.

## Feature Transformation

In this section we will use principal component analysis (PCA) to draw conclusions about the underlying structure of the wholesale customer data. Since using PCA on a dataset calculates the dimensions which best maximize variance, we will find which compound combinations of features best describe customers.

### Implementation: PCA

Now that the data has been scaled to a more normal distribution and has had any necessary outliers removed, we can now apply PCA to the good\_data to discover which dimensions about the data best maximize the variance of features involved. In addition to finding these dimensions, PCA will also report the explained variance ratio of each dimension — how much variance within the data is explained by that dimension alone. Note that a component (dimension) from PCA can be considered a new "feature" of the space, however it is a composition of the original features present in the data.



	Explained Variance	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Dimension 1	0.3366	0.5155	0.1000	-0.2900	0.5749	0.5479	0.0971
Dimension 2	0.1992	-0.1675	0.6004	0.3390	-0.0196	0.1252	0.6931
Dimension 3	0.1412	0.1866	-0.6534	-0.1105	-0.0453	-0.1908	0.6983
Dimension 4	0.1222	0.7556	0.2787	0.2052	-0.0926	-0.5447	-0.0635
Dimension 5	0.1061	-0.3161	0.1513	-0.2927	0.6622	-0.5910	0.0612
Dimension 6	0.0947	0.0254	0.3194	-0.8130	-0.4689	-0.0420	0.1215

pandas.core.frame.DataFrame

Dimension 1 0.3366

Dimension 2 0.5358

Dimension 3 0.6770

Dimension 4 0.7992

Dimension 5 0.9053

Dimension 6 1.0000

Name: Explained Variance, dtype: float64



53.58% of the variance in the data is explained by the first and second principal components. 79.92% of the variance in the data is explained by the first four principal components.

Components breakdown:

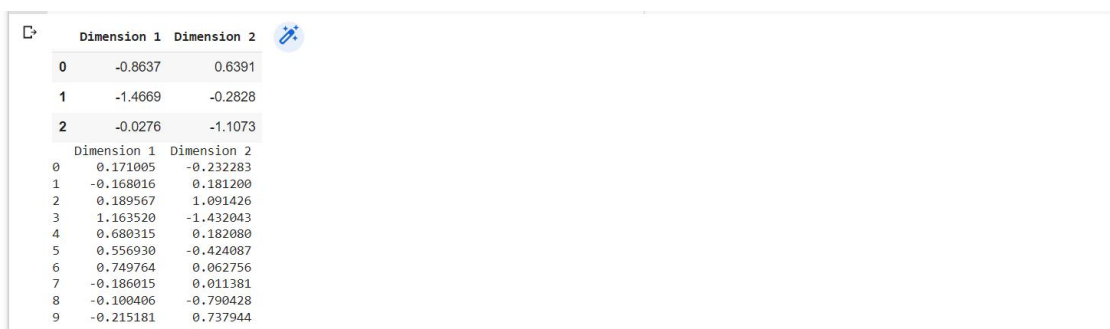
- The first principal component (PC1):
  1. An increase in PC1 is associated with large increases in "Milk", "Grocery" and "Detergents\_Paper" spending.
  2. These features best represent PC1.
  3. This is in line with our initial findings where the 3 features are highly correlated.
- The second principal component (PC2):
  1. An increase in PC2 is associated with large increases in "Fresh", "Frozen" and "Delicatessen" spending.
  2. These features best represent PC2.
  3. This makes sense as PC1 represents different features. And in PC2, the features in PC1 have very small positive weights.
- The third principal component (PC3):
  1. An increase in PC3 is associated with a large increase in "Delicatessen" and a large decrease in "Fresh" spending.
  2. These features best represent PC3.
- The fourth principal component (PC4):
  1. An increase in PC4 is associated with a large increasing in "Frozen" and a large decrease in "Delicatessen" spending.
  2. These features best represent PC4.

### Implementation: Dimensionality Reduction

When using principal component analysis, one of the main goals is to reduce the dimensionality of the data — in effect, reducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of this, the cumulative explained variance ratio is extremely important for knowing how many dimensions are necessary for the problem. Additionally, if a significant amount of variance is explained by only two or three dimensions, the reduced data can be visualized afterwards.



	Dimension 1	Dimension 2
0	-0.8664	0.6173
1	-1.4683	-0.2378
2	-0.0522	-1.1239



	Dimension 1	Dimension 2
0	-0.8637	0.6391
1	-1.4669	-0.2828
2	-0.0276	-1.1073

	Dimension 1	Dimension 2
0	0.171005	-0.232283
1	-0.168016	0.181200
2	0.189567	1.091426
3	1.163520	-1.432043
4	0.680315	0.182080
5	0.556930	-0.424087
6	0.749764	0.062756
7	-0.186015	0.011381
8	-0.100406	-0.790428
9	-0.215181	0.737944

We can see how the log-transformed sample data has changed after having a PCA transformation applied to it using only two dimensions. Observe how the values for the first two dimensions remains unchanged when compared to a PCA transformation in six dimensions.

## Clustering

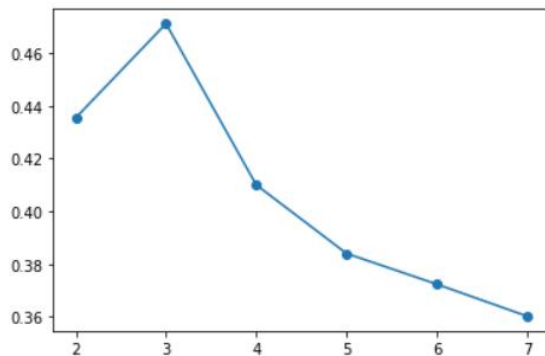
In this section, we will use both K-Means clustering algorithm and Gaussian Mixture Model clustering algorithm to identify the various customer segments hidden in the data. You will then recover specific data points from the clusters to understand their significance by transforming them back into their original dimension and scale.

### Implementation: Creating Clusters

Depending on the problem, the number of clusters that you expect to be in the data may already be known. When the number of clusters is not known a priori, there is no guarantee that a given number of clusters best segments the data, since it is unclear what structure exists in the data — if any. However, we can quantify the "goodness" of a clustering by calculating each data point's silhouette coefficient. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). Calculating the mean silhouette coefficient provides for a simple scoring method of a given clustering.

### USING K-Means Clustering Algorithm:

```
7 clusters: 0.36025
6 clusters: 0.37236
5 clusters: 0.384
4 clusters: 0.40998
3 clusters: 0.47108
2 clusters: 0.43542
```



Silhouette refers to a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object lies within its cluster. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from -1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

The silhouette score (SS) comes out to be 7 clusters: 0.36025 6 clusters: 0.37236 5 clusters: 0.384 4 clusters: 0.40998 3 clusters: 0.47108 2 clusters: 0.43542 The best silhouette score is for 3 clusters which is obvious as separating any data points with a plane is the easiest but our underlying data in question is more complex than the 2 clusters and if you read carefully on the grouping (G1, G2, G3, G4, G5 and G6) are the most logical groups based on total spending profile of individual customers.



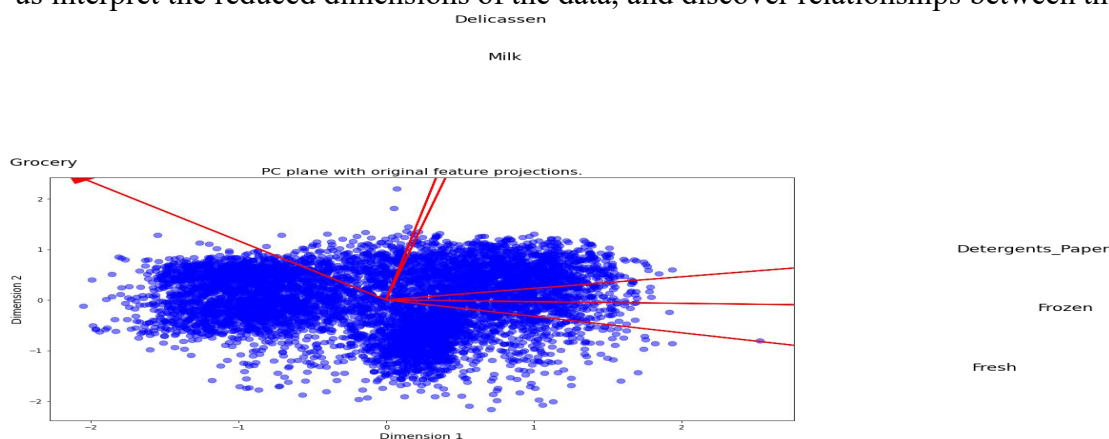
## Cluster Visualization



Once you've chosen the optimal number of clusters for your clustering algorithm using the scoring metric above, you can now visualize the results by executing the code block below. Note that, for experimentation purposes, you are welcome to adjust the number of clusters for your clustering algorithm to see various visualizations. The final visualization provided should, however, correspond with the optimal number of clusters

## Visualizing a Biplot

A biplot is a scatterplot where each data point is represented by its scores along the principal components. The axes are the principal components (in this case Dimension 1 and Dimension 2). In addition, the biplot shows the projection of the original features along the components. A biplot can help us interpret the reduced dimensions of the data, and discover relationships between the principal



components and original features.

## Observations:

Delicassen, Grocery and Milk are the features most strongly correlated with the Dimension 1.

Fresh, Frozen, Detergents\_Paper are the features most strongly correlated with the Dimension 2.

## Gaussian Mixture Model clustering algorithm:

```
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score

clusterer = GaussianMixture(n_components=2, covariance_type='full')
clusterer.fit(reduced_data)

# Predict the cluster for each data point
preds = clusterer.predict(reduced_data)

# Find the cluster centers
centers = clusterer.means_

# Predict the cluster for each transformed sample data point
sample_preds = clusterer.predict(pca_samples)

# Calculate the mean silhouette coefficient for the number of clusters chosen
score = silhouette_score(reduced_data, preds)
print("Silhouette coefficient for {} clusters: {:.3f}".format(2, score))
```

/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but GaussianMixture was fitted with  
"X does not have valid feature names, but"  
Silhouette coefficient for 2 clusters: 0.431

## Cluster Visualization:



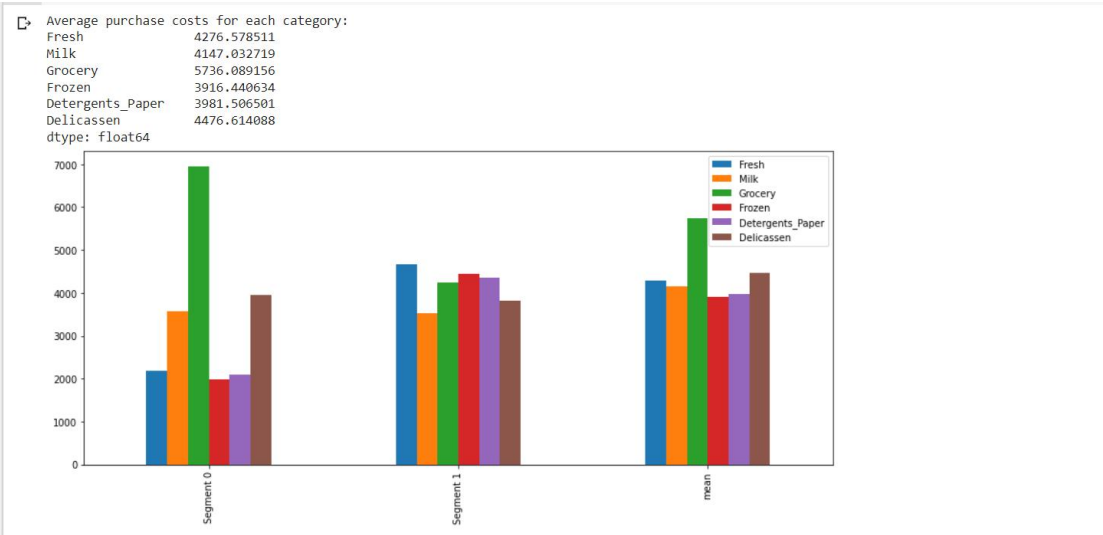
## Data recovery

Each cluster present in the visualization above has a central point. These centers (or means) are not specifically data points from the data, but rather the averages of all the data points predicted in the respective clusters. For the problem of creating customer segments, a cluster's center point corresponds to the average customer of that segment. Since the data is currently reduced in dimension and scaled by a logarithm, we can recover the representative customer spending from these data points by applying the inverse transformations.

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Segment 0	2177.0	3568.0	6953.0	1975.0	2088.0	3963.0
Segment 1	4673.0	3529.0	4254.0	4439.0	4358.0	3814.0

+ Code + Text

## Result:



**Cluster 0 has a high demand on Grocery though it has a lower than average for all other categories. It should best identify as the grocery product market.**

**Cluster 1 has a higher than average demand on Fresh, Detergents\_Paper and Frozen categories. It should best identify as the restaurants.**

Sample point 0 predicted to be in Cluster 1  
Sample point 1 predicted to be in Cluster 1  
Sample point 2 predicted to be in Cluster 0

Our initial assumption for Sample point 0 is not consistent with its predicted classification of Cluster 0. It has a very similar distribution of needs as Cluster 1, where your typical fresh produce market has Fresh foods as their primary product, and a small selection of Frozen foods. Our initial prediction for Sample point 1 is completely consistent with its predicted classification of Cluster 0. It has a very similar distribution of needs as Cluster 0, where your typical cafe or restaurant needs a good supply Milk and Grocery ingredients, and a good supply of Detergents\_Paper to clean & sanitize the place and provide paper napkins for their customers. It's hard to say whether my initial prediction for Sample point 2 is consistent with its predicted classification of Cluster 0.

Your typical deli inherently needs a lot of Delicatessen foods and some need for Grocery and Fresh foods. Neither clusters put too much importance on Delicatessen foods, and while its Grocery needs matches Cluster 0, its Fresh foods needs better matches Cluster 1. We can see its on the border between the two clusters in the cluster visualization, so we can't have much confidence in which cluster it belongs to either way.

## **Conclusion & Future Scope**

The Customer Segmentation Analysis guide us to provide a step-by-step process for identifying, prioritizing, and targeting your best current customer segments, simply following it does not guarantee success. To be effective, you must prepare and plan for the various challenges and hurdles that each step may present, and always make sure to adapt your process to any new information or feedback that might change its output. Additionally, we cannot force feed this process on our business. If the key stakeholders that will be impacted by the best current customers segmentation process do not fully buy-in, then the outputs produced from it will be relatively meaningless. If we properly manage the best current customer segmentation process, however, the impact it can have on every part of your organization — sales, marketing, product development, customer service, etc. — is immense. Your business will possess stronger customer focus and market clarity, allowing it to scale in a far more predictable and efficient manner. Ultimately, that means no longer needing to take on every customer that is willing to pay for your product or service, which will allow you to instead hone in on a specific subset of customers that present the most profitable opportunities and efficient use of resources. That is critical for every business, of course, but at the expansion stage, it can often be the difference between incredible success and certain failure.

## **BIBLIOGRAPHY**

- Dr. R. Gardener “The Essential R Reference” (2014),
- Concepts of customer segmentation [<http://www.business-science.io>]  
[<https://labs.openviewpartners.com/customer-segmentation/>]
- [<https://www.rstudio.com/>]
- Constantinou, Andreas. “Operators: service-pipes or bit-pipes ?” November 8, 2006. Retrieved March 15, 2010. [<http://www.visionmobile.com/blog/2006/11/operators-service-pipes-or-bit-pipes>]
- Jennings, Jeanne. “Customer Segmentation Analysis” July 14, 2008. Retrieved March 16, 2010. [<https://www.clickz.com/3630202>]
- [<https://www.r-project.org/>]
- “Market Segmentation”. Learn Marketing. Retrieved March 15, 2010. [<http://www.learnmarketing.net/segmentation2.htm>]
- [<https://github.com/teddyroland/python-biplot>]
- “Market Segmentation - Introduction” inclusive design toolkit. 2007-2008. Retrieved March 16, 2010. [<http://www-edc.eng.cam.ac.uk/betterdesign/knowledge/catusers/catusers2.html>]
- [<https://www.kaggle.com/>]
- Kotler, Philip. Principles of Marketing. Third European Edition 2002.