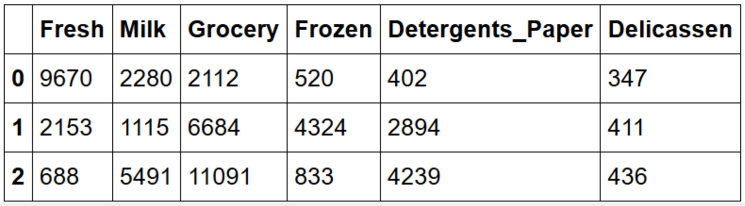
Creating Customer Segments

OVERVIEW

In this project I showcase a use case for unsupervised learning techniques on product spending data collected from customers of a wholesale distributor in Lisbon, Portugal to identify customer segments hidden in the data. The goal is to best describe the variation in the different types of customers that a wholesale distributor interacts with. Doing so would equip the distributor with insight into how to best structure their delivery service to meet the needs of each customer.

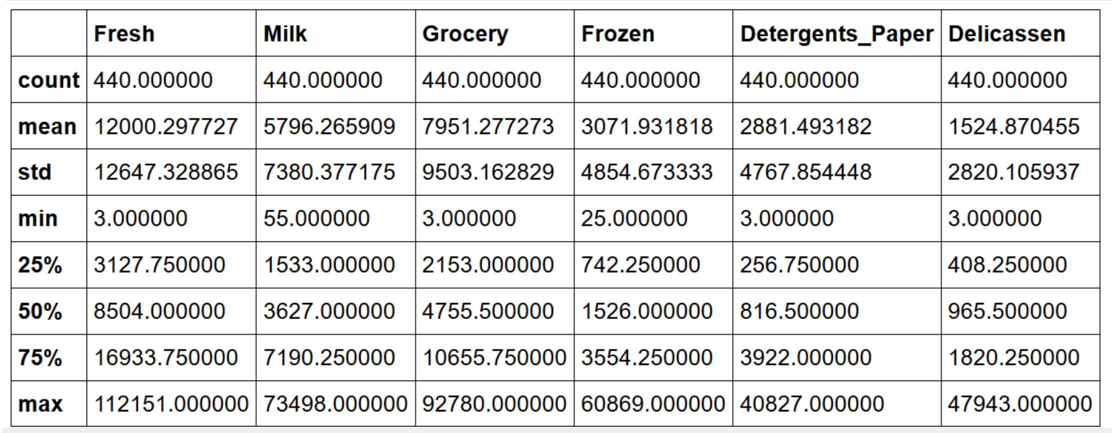
DATA EXPLORATION

To get a better understanding of the customers and how their data will transform through the analysis, it would be best to select a few sample data points(index: 199,263,35) and explore them in more detail. Following the Data sample there is a brief statistical description table using the .describe() method.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	9670	2280	2112	520	402	347
1	2153	1115	6684	4324	2894	411
2	688	5491	11091	833	4239	436

```
Wholesale customers dataset
has 440 samples with 6
features each.
```



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
count	440.000000	440.000000	440.000000	440.000000	440.000000	440.000000
mean	12000.297727	5796.265909	7951.277273	3071.931818	2881.493182	1524.870455
std	12647.328865	7380.377175	9503.162829	4854.673333	4767.854448	2820.105937
min	3.000000	55.000000	3.000000	25.000000	3.000000	3.000000
25%	3127.750000	1533.000000	2153.000000	742.250000	256.750000	408.250000
50%	8504.000000	3627.000000	4755.500000	1526.000000	816.500000	965.500000
75%	16933.750000	7190.250000	10655.750000	3554.250000	3922.000000	1820.250000
max	112151.000000	73498.000000	92780.000000	60869.000000	40827.000000	47943.000000

Looking at the dataset sample leads me to assume the following:

- *Index 0:* This establishment has low quantity in all categories besides **Fresh,** so my guess is that it's some type restaurant.
- *Index 1:* Just shy of the average in **Grocery**, **Frozen**, and **Detergents Paper** so my guess would be that it's a convenience store
- *Index 2:* Due to the average and above quantities of **Milk**, **Grocery**, and **Detergents Paper**, I would guess that index 2is a Market.

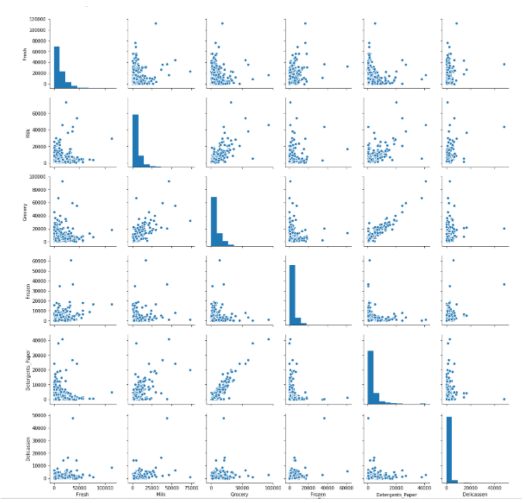
FEATURE RELEVANCE

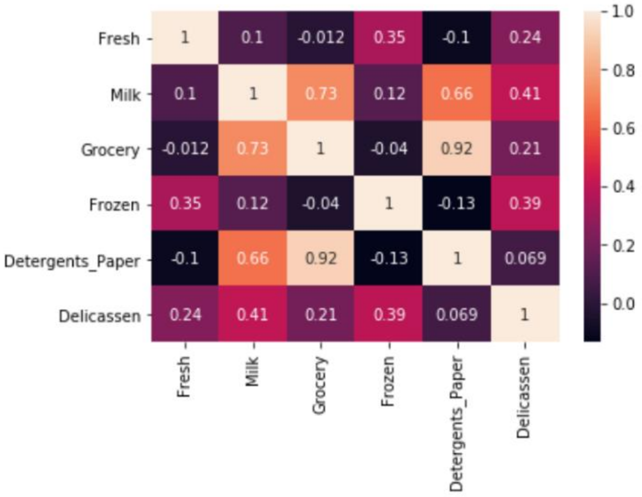
One interesting thought to consider is if one (or more) of the six product categories is actually relevant for understanding customer purchasing. Is it possible to determine whether customers purchasing some amount of one category of products will necessarily purchase some proportional amount of another 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.

I attempted to predict the "Fresh" feature and the score was -0.36% which means that this feature is necessary for identifying customers since the correlation is negative. The coefficient of determination, `R^2`, is scored between 0 and 1, with 1 being a perfect fit. A negative `R^2` implies the model fails to fit the data. If you get a low score for a particular feature, that lends us to believe that that feature point is hard to predict using the other features, thereby making it an important feature to consider when considering relevance.

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. I found that the feature I attempted to predict above is relevant for identifying a specific customer, which means the scatter matrix below may not show any correlation between that feature and the others.





Correlation Matrix Analysis

The distribution seems to be skewed to the right with a large amount of data points near 0(The origin). The "Fresh" confirms my findings above since the highest correlation score is .35 with the "Frozen " category. Additionally, I found the features listed below to have a mid to high correlation:

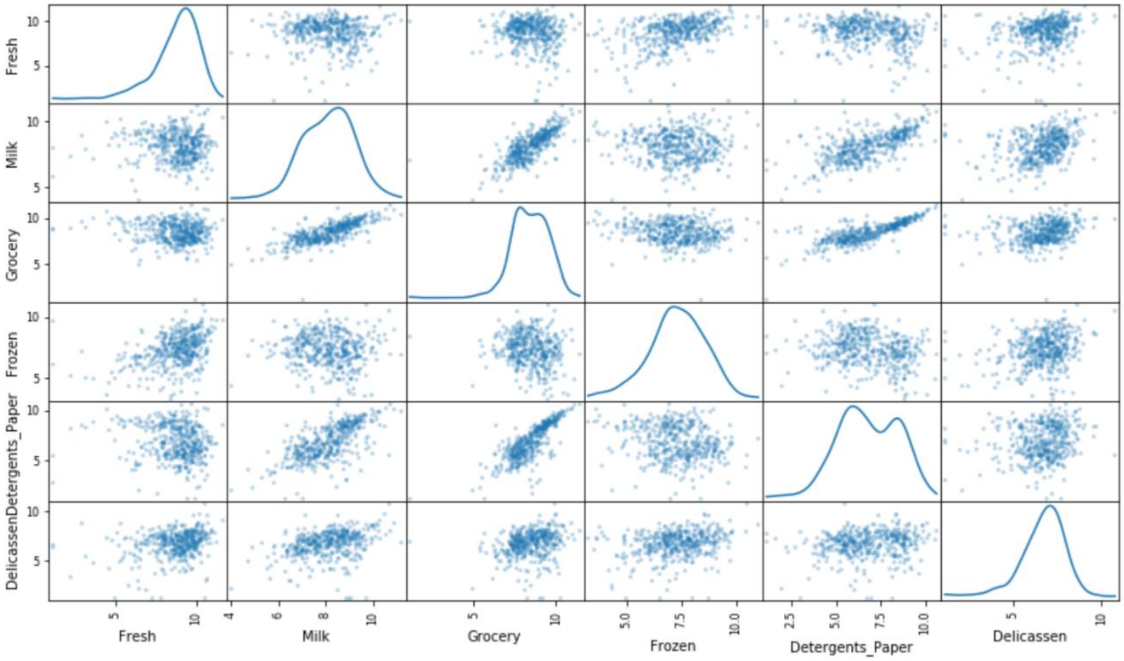
- Detergents_Paper and Grocery: 92%
- Milk and Grocery: 73%
- Milk and Detergents_Paper: 66%

DATA PREPROCESSING

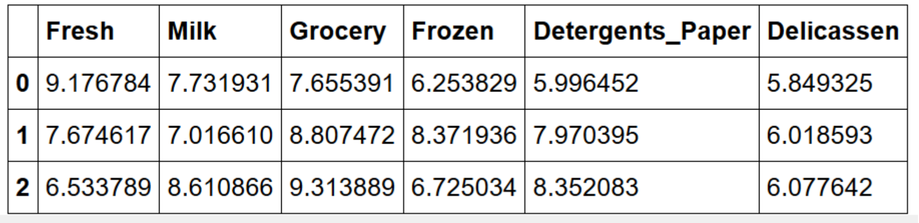
In this section, I will preprocess the data to create a better representation of customers by performing scaling and detecting (and optionally removing) outliers.

Feature Scaling

Since the data is not normally distributed, especially with the mean and median varying significantly (indicating a large skew), it is most often appropriate to apply a non-linear scaling. 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. After applying the natural logarithm scaling to the data, the distribution of each feature should appear much more normal.



Here is the same sample from earlier after the natural logarithm scaling has been applied.



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
0	9.176784	7.731931	7.655391	6.253829	5.996452	5.849325
1	7.674617	7.016610	8.807472	8.371936	7.970395	6.018593
2	6.533789	8.610866	9.313889	6.725034	8.352083	6.077642

OUTLIER DETECTION

According to Tuckey'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. Here are the steps of my process in identifying outliers:

- Assign the value of the 25th percentile for the given feature to Q1 using np.percentile.
- Assign the value of the 75th percentile for the given feature to Q3. Again, using np.percentile.
- Assign the calculation of an outlier step for the given feature to step.
- Optionally remove data points from the dataset by adding indices to the outliers list.

If I choose to remove any outliers, I must ensure that the sample data does not contain any of these points. Once I performed this implementation, the dataset will be stored in the variable good_data.

```
The is Normalized Dataset with outliers removed: (398, 6)
Normalized Data including outliers: (440, 6)
```

Below is the mean of the log_data.

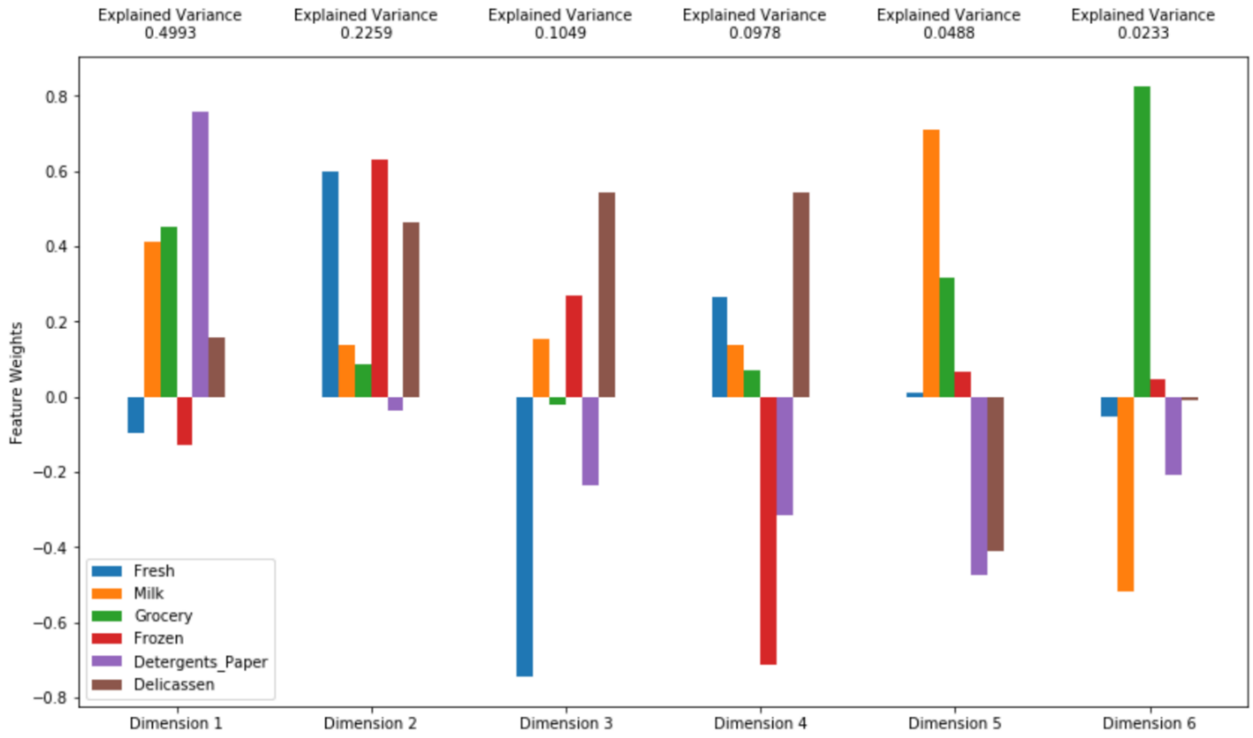
```
Fresh               8.730544
Milk                8.121047
Grocery             8.441169
Frozen              7.301396
Detergents_Paper    6.785972
Delicassen          6.665133
```

FEATURE TRANSFORMATION

In this section I 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, I will find which compound combinations of features best describe customers. 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.

PCA Implementation

- Import sklearn.decomposition.PCA and assign the results of fitting PCA in six dimensions with good_data to pca.
- Apply a PCA transformation of log_samples using pca.transform and assign the results to pca_samples.

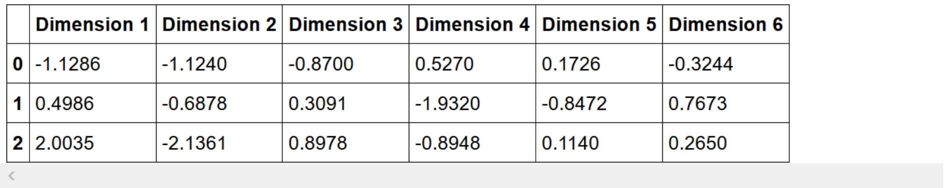


Plot Analysis

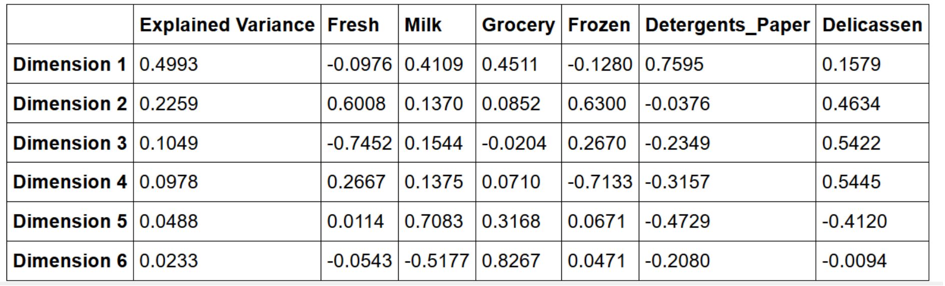
The PCA result from the first two dimensions is 0.7252, while the result from the first four dimensions is 0.9279. Meaning, that 72.5% of the variance is explained by the first two principal components 92.7 is explained by the first four.

- **Dimension 1:** In this dimension we can see that it covers many of the features including *Milk*, *Grocery*, _Detergents *Paper* and *Delicassen*. It makes sense why this dimension alone is responsible for 49% of the variance in the data. This Dimension refers to the retail customers due to the high *Detergents Paper* weight.
- **Dimension 2:**This dimension includes describes a positive variance in the *Fresh*, *Frozen*, and *Delicassen* features. Delicassen had the least variance in the previous **Dimension 1**. Dimension 2 refers to the Non-Retail customers such as restaurants and Hotels due to substantial negative variance in the *Detergents Paper* feature.
- **Dimension 3:** In this dimension although we get a slight positive increase in variance for *Milk* and *Delicassen*, we lose it on the other features such as *Fresh* and *Detergents Paper*. Due to the similarity with Dimension 2 and the large negative *Fresh* weight, Dimension 3 could refer to fast food establishments.
- **Dimension 4:** In this dimension there a large negative variance in the *Frozen* and *Detergents Paper* features while having a positive increase in the *Fresh* feature. *Milk* and Grocery show a slight positive variance with *Delicassen* staying at same level as Dimension 3. There are some similarities with Dimension 1 which leads me to believe this dimension refers to a Deli or small convenience shops.

In the tables below you will find that the log-transformed sample data has changed after having a PCA transformation applied to it in six dimensions



	Dimension 1	Dimension 2	Dimension 3	Dimension 4	Dimension 5	Dimension 6
0	-1.1286	-1.1240	-0.8700	0.5270	0.1726	-0.3244
1	0.4986	-0.6878	0.3091	-1.9320	-0.8472	0.7673
2	2.0035	-2.1361	0.8978	-0.8948	0.1140	0.2650



	Explained Variance	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Dimension 1	0.4993	-0.0976	0.4109	0.4511	-0.1280	0.7595	0.1579
Dimension 2	0.2259	0.6008	0.1370	0.0852	0.6300	-0.0376	0.4634
Dimension 3	0.1049	-0.7452	0.1544	-0.0204	0.2670	-0.2349	0.5422
Dimension 4	0.0978	0.2667	0.1375	0.0710	-0.7133	-0.3157	0.5445
Dimension 5	0.0488	0.0114	0.7083	0.3168	0.0671	-0.4729	-0.4120
Dimension 6	0.0233	-0.0543	-0.5177	0.8267	0.0471	-0.2080	-0.0094

Dimensionality Reduction

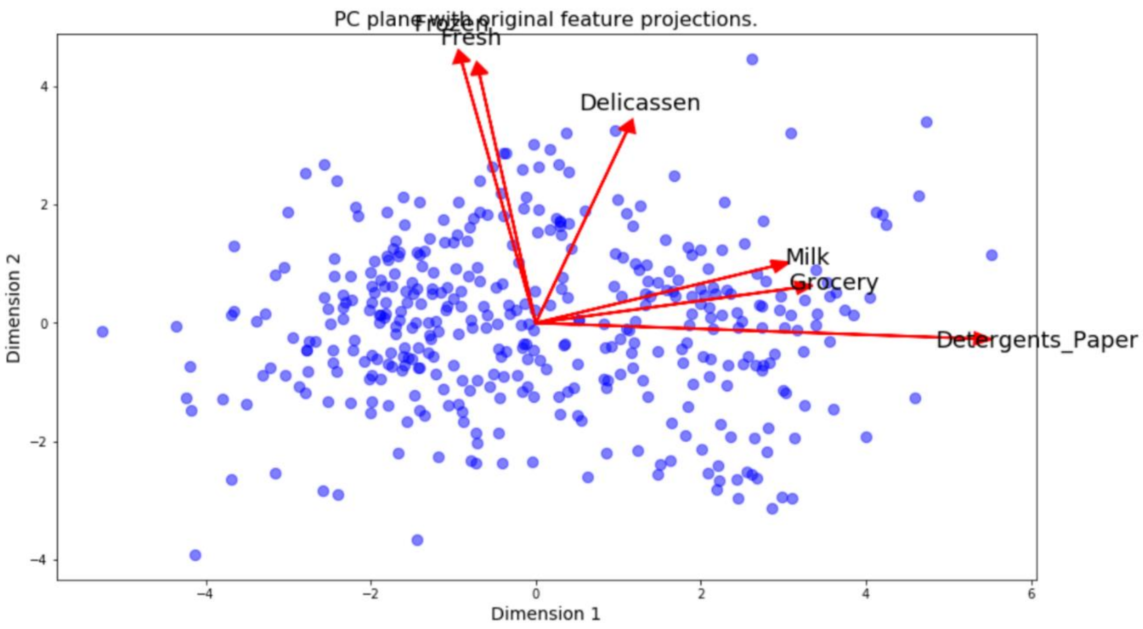
When using principal component analysis, one of the main goals is to reduce the dimensionality of the data — in effect, reducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of 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.

I proceeded to implement the following:

- Assign the results of fitting PCA in two dimensions with good_data to pca.
- Apply a PCA transformation of good_data using pca.transform and assign the results to reduced data.
- Apply a PCA transformation of log_samples using pca.transform and assign the results to pca_samples.

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.



Biplot Analysis

Once we have the original feature projections (in red), it is easier to interpret the relative position of each data point in the scatterplot. For instance, a point the lower right corner of the figure will likely correspond to a customer that spends a lot on 'Milk', 'Grocery' and 'Detergents_Paper', but not so much on the other product categories.
The features are most strongly correlated with the first component are *Milk*, *Grocery*, and *Detergents Paper*. The features are most strongly correlated with the second component are *Frozen*, *Fresh*, and *Delicassen*.

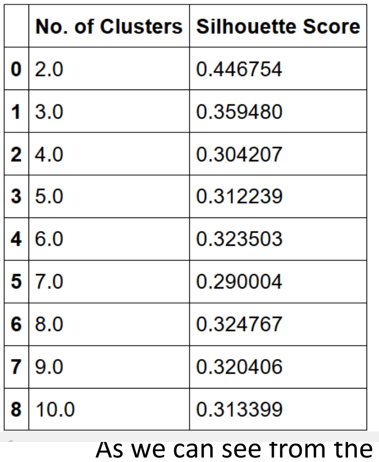
CLUSTERING

K-Means Clustering	Gaussian Mixture Model
Pros: • It's computationally faster due to the random initialization • Great for large variables • Produces tight spherical clusters	Pros: • soft clustering(sample membership of multiple clusters) • cluster shape flexibility

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 model 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.

Because the data seems to be bunched together it might be difficult to create clear clusters therefore, it will be easier to have an algorithm like GMM which finds the probability of which Gaussian distribution the point most likely come from.

Gaussian Mixture Clustering Model Implementation

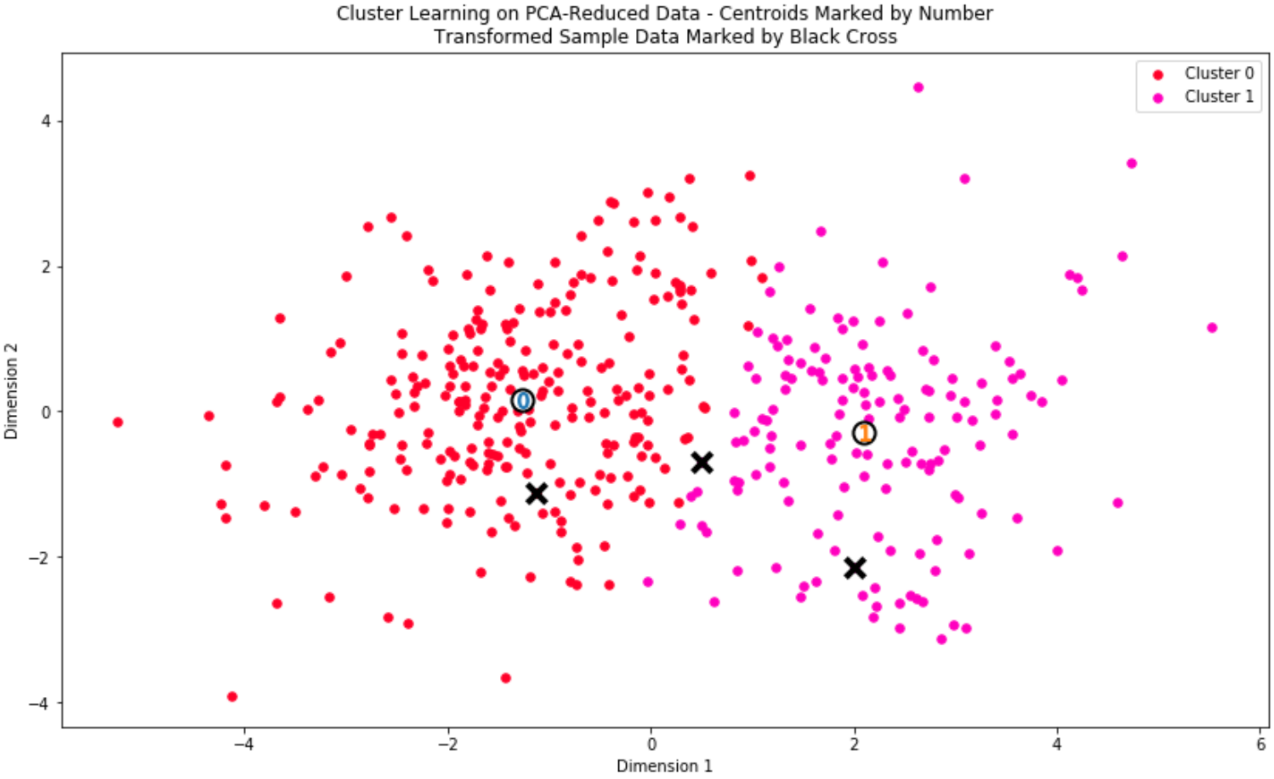


	No. of Clusters	Silhouette Score
0	2.0	0.446754
1	3.0	0.359480
2	4.0	0.304207
3	5.0	0.312239
4	6.0	0.323503
5	7.0	0.290004
6	8.0	0.324767
7	9.0	0.320406
8	10.0	0.313399

- Fit a clustering algorithm to the reduced_data and assign it to clusterer.
- Predict the cluster for each data point in reduced_data using clusterer.predict and assign them to preds.
- Find the cluster centers using the algorithm's respective attribute and assign them to centers.
- Predict the cluster for each sample data point in pca_samples and assign them sample_preds.
- Import sklearn.metrics.silhouette_score and calculate the silhouette score of reduced_data against preds.
- Assign the silhouette score to score and print the result.

As we can see from the table above 2 clusters have the highest Silhouette score which corresponds to the best number of segments.

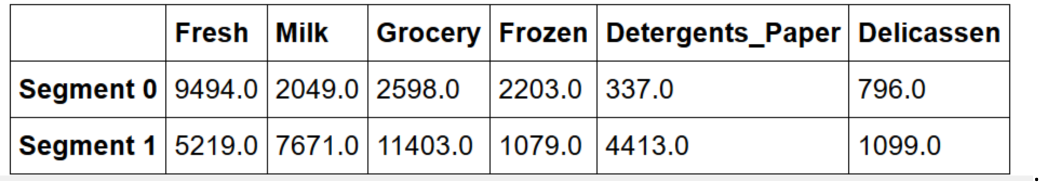
Cluster Visualization



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.

Data Recovery

- Apply the inverse transform to centers using pca.inverse_transform and assign the new centers to log centers.
- Apply the inverse function of np.log to log centers using np.exp and assign the true centers to true_centers



	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
Segment 0	9494.0	2049.0	2598.0	2203.0	337.0	796.0
Segment 1	5219.0	7671.0	11403.0	1079.0	4413.0	1099.0

.

RESULTS

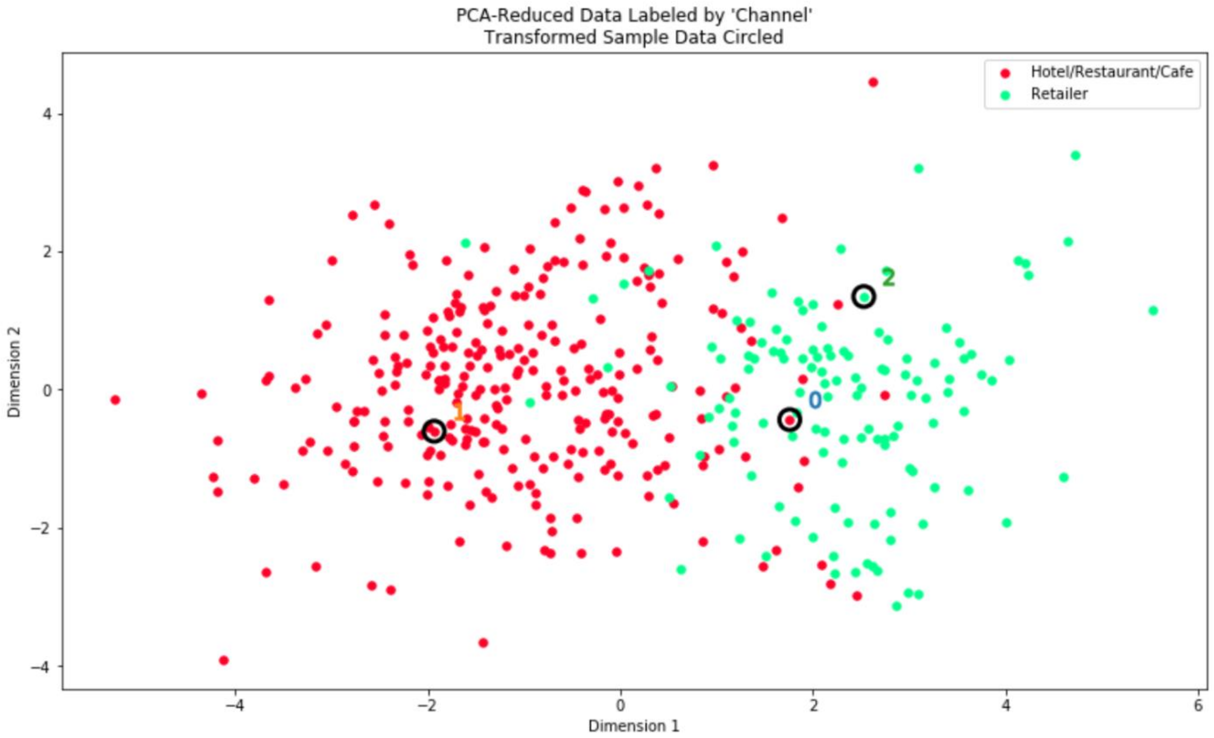
I believe the Detergents_Paper feature is a key indicator as to what kind of establishments each segment represents.
Segment 0: Due to the below average quantity of Detergents_Paper and just shy from the average in Fresh and Frozen. This segment represents *Restaurant* type of establishments
Segment 1: Due to the high volume of Milk, Grocery, and Detergents_paper, this segment represents *Convenience Stores and Markets* type of establishments. To support my case here are the results from the samples:

index 199: segment 0 index 263: segment 1 index 35: segment 1

CONCLUSION

At the beginning of this project, it was discussed that the 'Channel' and 'Region' features would be excluded from the dataset so that the customer product categories were emphasized in the analysis. By reintroducing the 'Channel' feature to the dataset, an interesting structure emerges when considering the same PCA dimensionality reduction applied earlier to the original dataset.

The image below will show how each data point is labeled either 'HoReCa' (Hotel/Restaurant/Cafe) or 'Retail' the reduced space. In addition, you will find the sample points are circled in the plot, which will identify their labeling.



This plot solidifies the accuracy of the unsupervised learning model with a clean distinction between the two clusters. Customers left from **-2** (on dimension 1)can be classified as purely `Retailers` and customers on the right of **2** (on dimension 1) can be classified as `Hotels/Restaurants/Cafes`.

Questions

Below you will find some questions that might be asked from the Distributor on the use case of this project.

1) *How can the wholesale distributor use the customer segments to determine which customers, if any, would react positively to the change in delivery service?*

- The first step here is to see how the A/B Test was conducted as there are some key criteria. The A/B test is a way of comparing the difference between two outcomes in a selected variable. In this case, this it would be the measure of how positive the customers' reaction with the new delivery schedule was. Please find the steps to a complete A/B test below:

  Conduct some research: look at the current satisfaction of the customers and if there is a true need to change the schedule. Take into account the current company costs of having items delivered daily and the types of customers who the change of schedule will not impact their business.

  Create a Hypothesis: This will help the company focus on the exact question they want to answer as every good analytics process requires. In this case, the wholesaler can serve some customers on a 3 day a week basis and see if there is a difference in the satisfaction. Once the Null Hypothesis is ready usually questioning that changing the schedule will have no impact in the satisfaction then you proceed to the next step.

A/B Split: Split the customers into two groups, A being the control group(customers on 5 day schedule) and B the customers on the new 3 day schedule. It's important to note that these groups are created randomly and there is no form of bias when it comes to splitting them.

Testing: Collect the results from A and B during the designed test period. for example: one week, two weeks, one month etc.

Analysis: Here is where you find out if there is a real difference in the results you received from the testing. The way to do that is by using the T-test and P-Value. The T-test will give you the difference coefficient between the two groups and the P-value will explain the probability of that coefficient being driven by random noise in the data vs general affect. For example, if the P-value is **.47** it means that there is a **47%** chance that T-test coefficient came from the noise in the data. In order to have a "statistical Significance" result the P-value must be equal to or less than **0.05**.

I would suggest that customers in Segment 1 with large quantities for Fresh items would like to receive daily deliveries due to the nature of their business. If the quality standard lowers, they will lose sales and begin to look for other distributors. Customers in Segment 0 inquire items with a longer expiration dates so they can afford to have them delivered less frequently.

2) *How can the wholesale distributor label the new customers using only their estimated product spending and the **customer segment** data?*

   - The wholesaler can you a Supervised Learning Model such as Random Forest with **Customer Segment** as the target variable. Once the training is complete the wholesaler will be able to predict the types of segments the new customers belong to.