Creating Customer Segments

Rafael A. Moreno Contreras
Introduction to Data Science and Python DA501
Catholic University of America

Introduction

The focus of this project is to analyze a dataset containing data on various customers' annual spending amounts (reported in *monetary units*) of diverse product categories for internal structure.

One goal of this project 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.

Problem Statement

Throughout unsupervised learning of purchasing behavior from different types of customers, we can find clusters that can categorize the types of customers in a different and more comprehensive way than the categories in the ‘Channel’ variable.

Dataset

The customer segments data is included as a selection of 440 data points collected on data found from clients of a wholesale distributor in Lisbon, Portugal. More information can be found on the UCI Machine Learning Repository.

Note (m.u.) is shorthand for *monetary units*.

Features

1. **Fresh**: annual spending (m.u.) on fresh products (Continuous);
2. **Milk**: annual spending (m.u.) on milk products (Continuous);
3. **Grocery**: annual spending (m.u.) on grocery products (Continuous);
4. **Frozen**: annual spending (m.u.) on frozen products (Continuous);
5. **Detergents_Paper**: annual spending (m.u.) on detergents and paper products (Continuous);
6. **Delicatessen**: annual spending (m.u.) on and delicatessen products (Continuous);
7. **Channel**: {Hotel/Restaurant/Cafe - 1, Retail - 2} (Nominal)
8. **Region**: {Lisbon - 1, Oporto - 2, or Other - 3} (Nominal)

For the purposes of this project, the features 'Channel' and 'Region' will be excluded in the analysis, with focus instead on the six product categories recorded for customers.

Exploring the data

In this section, we will begin exploring the data through visualizations and code to understand how each feature is related to the others. We will observe a statistical description of the dataset, consider the relevance of each feature, and select three sample data points from the dataset which we will track through the course of this project, in this case indexes 85,181, and 338.

Considering the total purchase cost of each product category and the statistical description of the dataset above for our sample customers. We make a prediction on what kind of establishment (customer) could each of the three samples we've chosen represent?

The mean values for the whole data set are as follows:

- Fresh: 12000.2977
- Milk: 5796.2
- Grocery: 3071.9
- Detergents_paper: 2881.4
- Delicatessen: 1524.8

The values for the 3 chosen samples are:

Index	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
85	16117	46197	92780	1026	40827	2944
181	112151	29627	18148	16745	4948	8550
338	3	333	7021	15601	15	550

Knowing this, how do our samples compare?

1) Index 85: posible retailer
- Largest spending on detergents and paper and groceries of the entire dataset, which usually are products for houses.

- Higher than average spending on milk.

- Lower than average spending on frozen products.

2) Index 181: posible large market
- High spending on almost every product category.

- Highest spending on fresh products of the entire dataset. Likely to be a large market.

- Low spending on detergents.

3) Index 338: posible restaurant

- The amount of every product is significantly lower than the previous two customers considered.

- The spending on fresh products is the lowest of the entire dataset.

- It may be a small and cheap restaurant which needs groceries and frozen food to serve the meals.

Feature Relevance

One interesting thought to consider is if one (or more) of the six product categories is actually relevant for understanding customer purchasing behavior. 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?

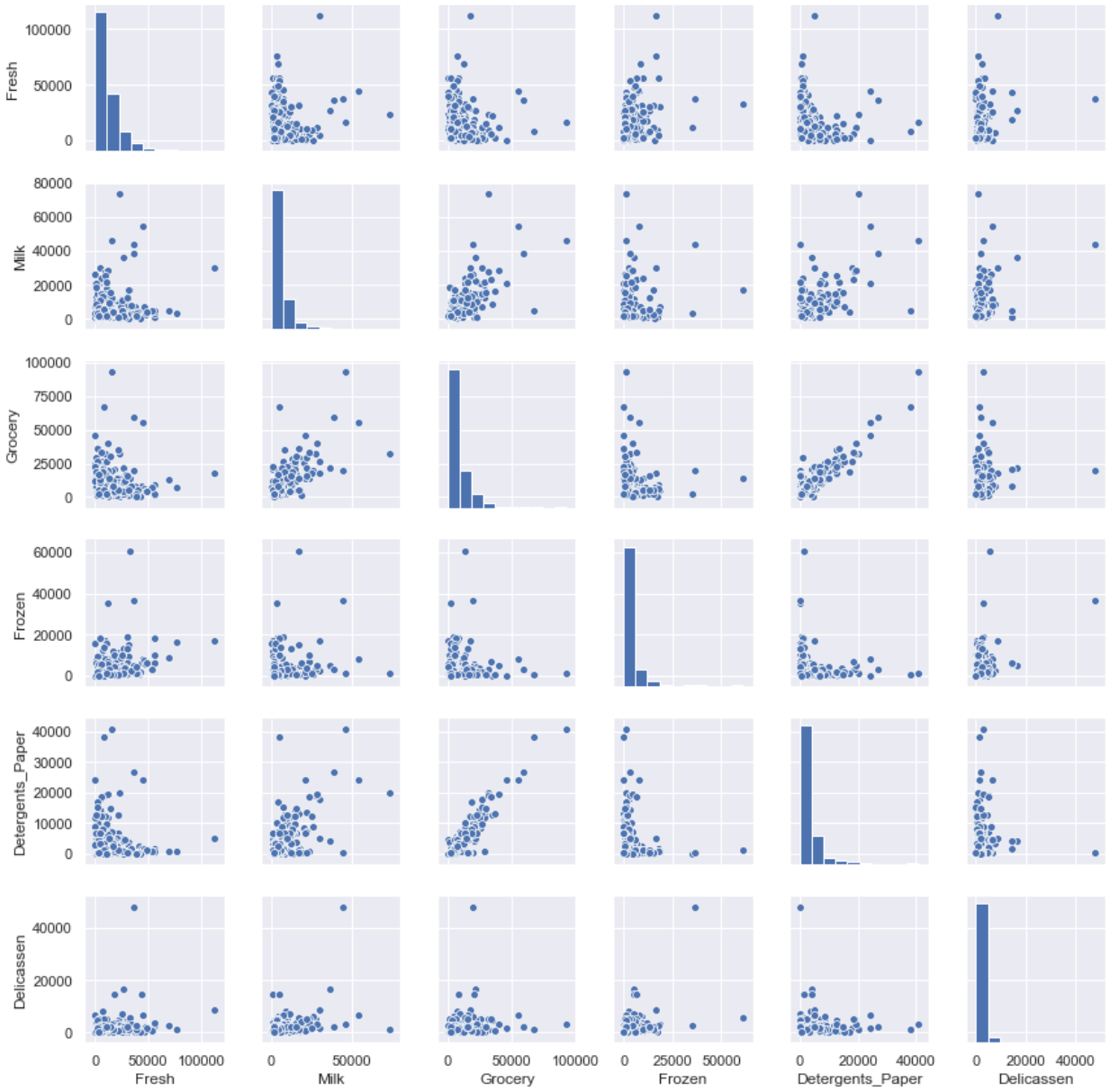
We can make this determination quite easily by training a supervised regression learner on a subset of the data with one feature removed, and then score how well that model can predict the removed feature.

We have that the prediction score for the Grocery feature is of 67.25% As we obtained a somehow high score, it as indicator of a very good fit. So this feature is easy to predict considering the rest of spending habits, and

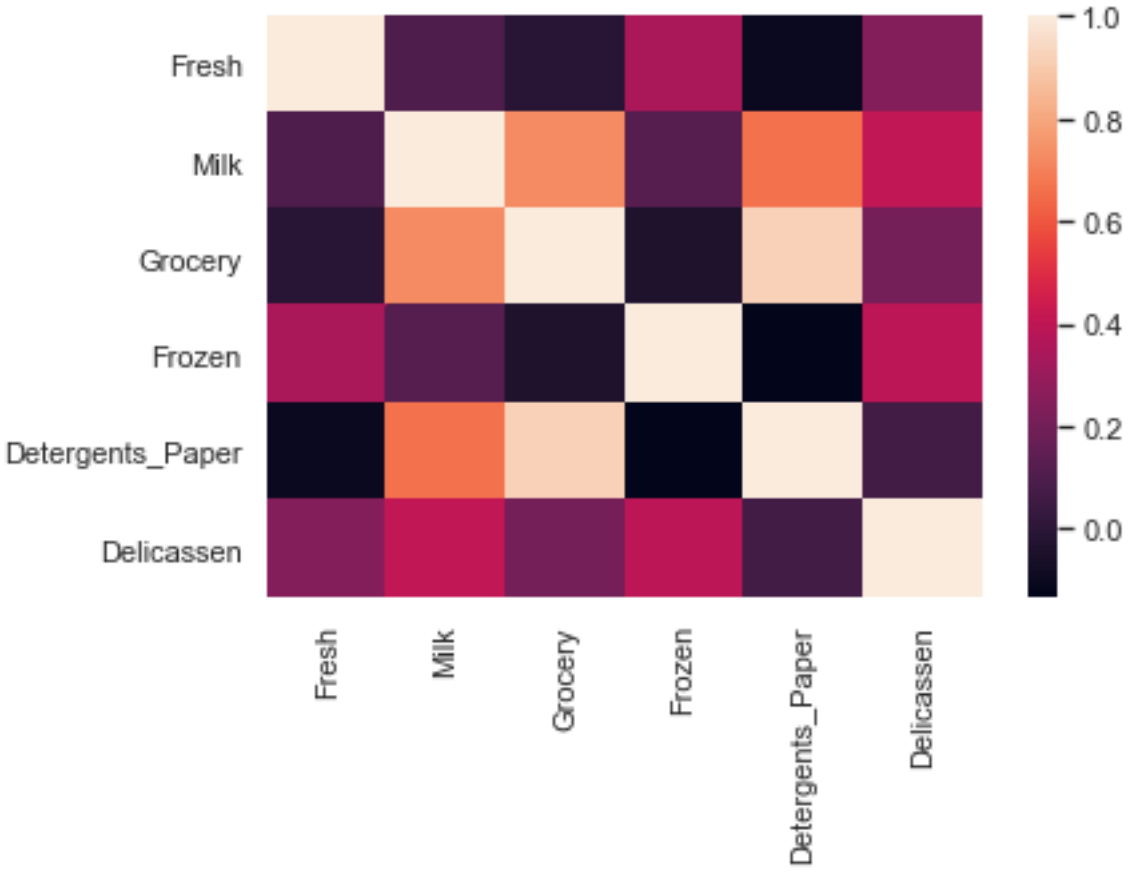
therefore, not very necessary for identifying customers' spending habits.

Visualize Feature Distributions

To get a better understanding of the dataset, we can construct a scatter matrix of each of the six products



Another useful visual representation of the correlation among the variables is a correlation matrix.



Some basic takeaways from the scatter matrix and correlation matrix are as follows:

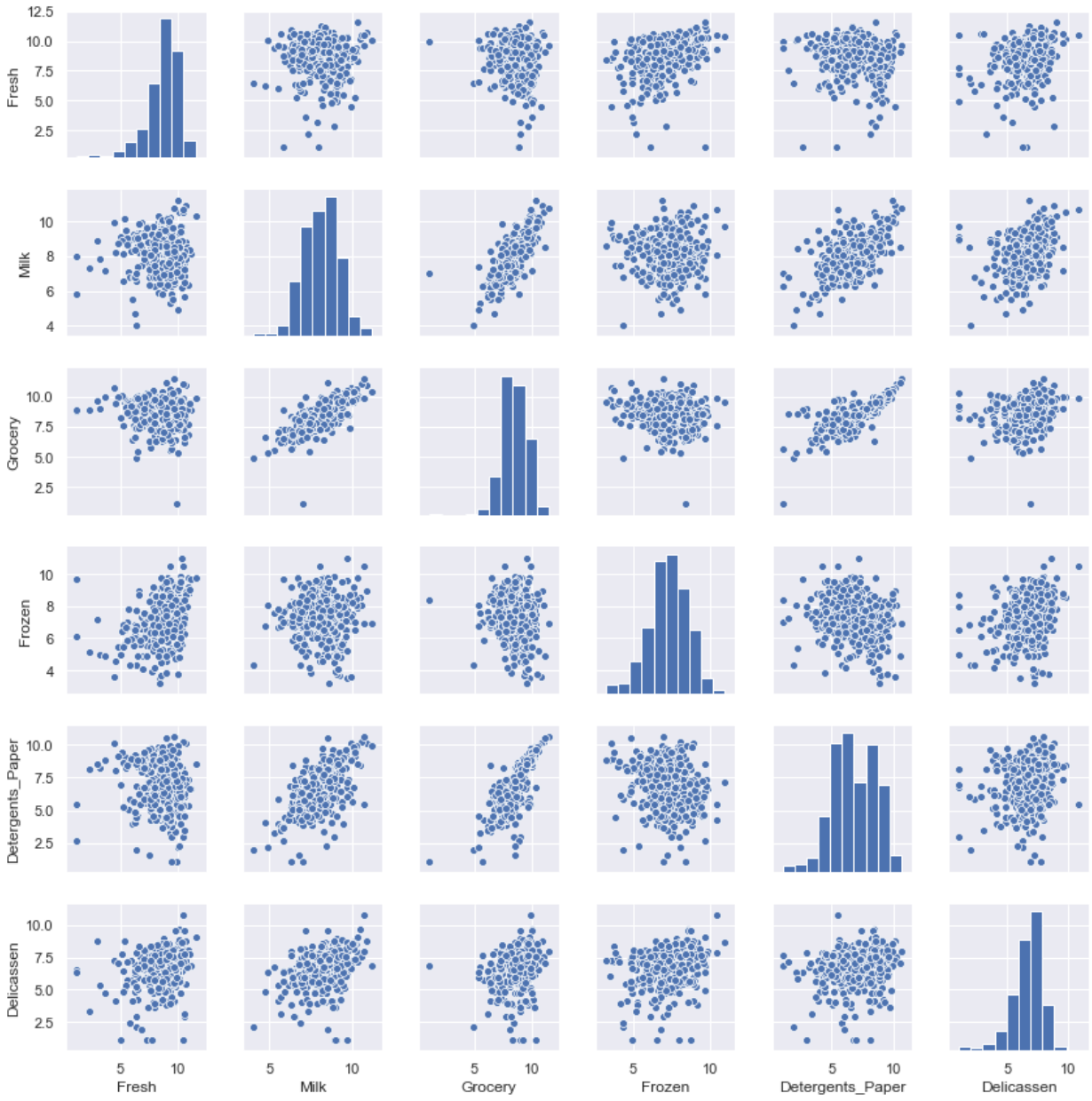
- Data is not normally distributed, it is positively skewed and it resemble the log-normal distribution.
- From the scatter plots and the correlation heatmap, we can see that there is a strong correlation between the ‘Grocery’ and ‘Detergent_paper’ features. The features ‘Grocery’ and ‘Milk’ also show a good degree of correlation.

- This correlation also reflects the how truly relevant the ‘Grocery’ feature is, which can be accurately predicted with the ‘Detergent_paper’ feature. And, therefore, is not an absolutely necessary feature in the dataset.

Data Preprocessing

As mentioned before, we can observe a positive skewness from the scatter matrix, therefore log transformation is a good option. Applying a log function on these values should give a linear trend and convert the set of values into ‘normally distributed’ values.

After applying log transformation to scale the data, on the new scatter matrix bellow be can see the distribution of each feature appear much more normal. For any pairs of features we have identified earlier as being correlated, we observe here that correlation is still present in a more clear way.



Outlier 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.
Here, we will use Tukey's Method for identfying 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.

K-Means is heavily influenced by the presence of outliers as they increase significantly the loss function that the algorithm tries to minimize. This loss function is the squared sum of the distances of each datapoint to the centroid, so, if the outlier is far enough, the centroid will be incorrectly situated.

Because of this persistent outliers across variables should be removed. The data points considered outliers that are present in more than one feature are: 65, 66, 75, 128, 154.

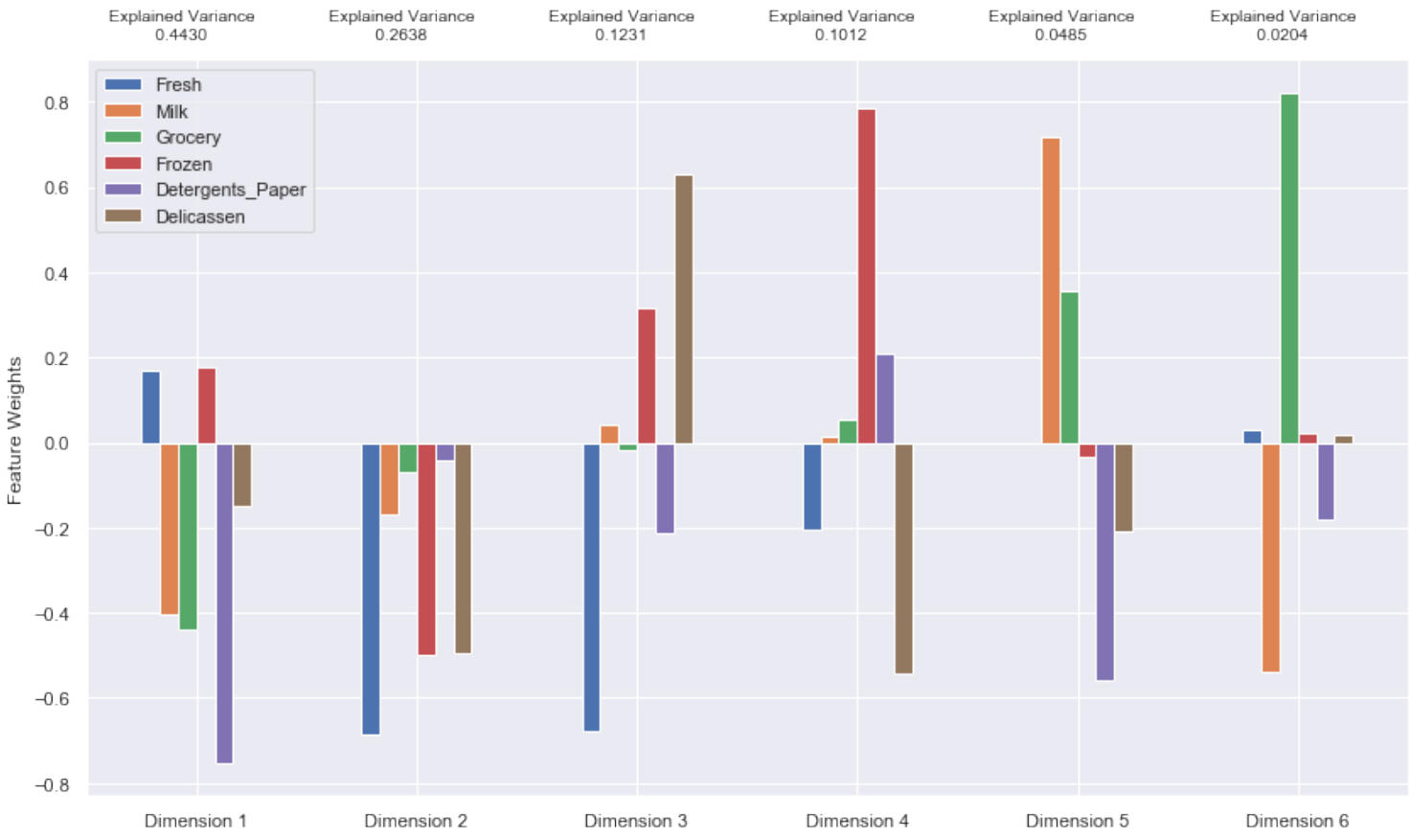
Feature Engineering

Now we will use Principal Component Analysis (PCA) to extract conclusions about the hidden structure of the dataset. PCA is used to calculate those dimensions that maximize variance, so we will find the combination of features that describe best each customer.

Principal Component Analysis (PCA)

Once the data has been scaled to a normal distribution and the necessary outliers have been removed, we can 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.



- The variance explained by the first two Principal Components is the 70.68% of the total.
- The variance explained by the first three Principal Components is the 93.11% of the total.

Dimensionality Reduction

When using principal component analysis, one of the main goals is to reduce the dimensionality of the data, but 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 signifiant amount of variance is explained by only two or three dimensions, the reduced data can be visualized afterwards.

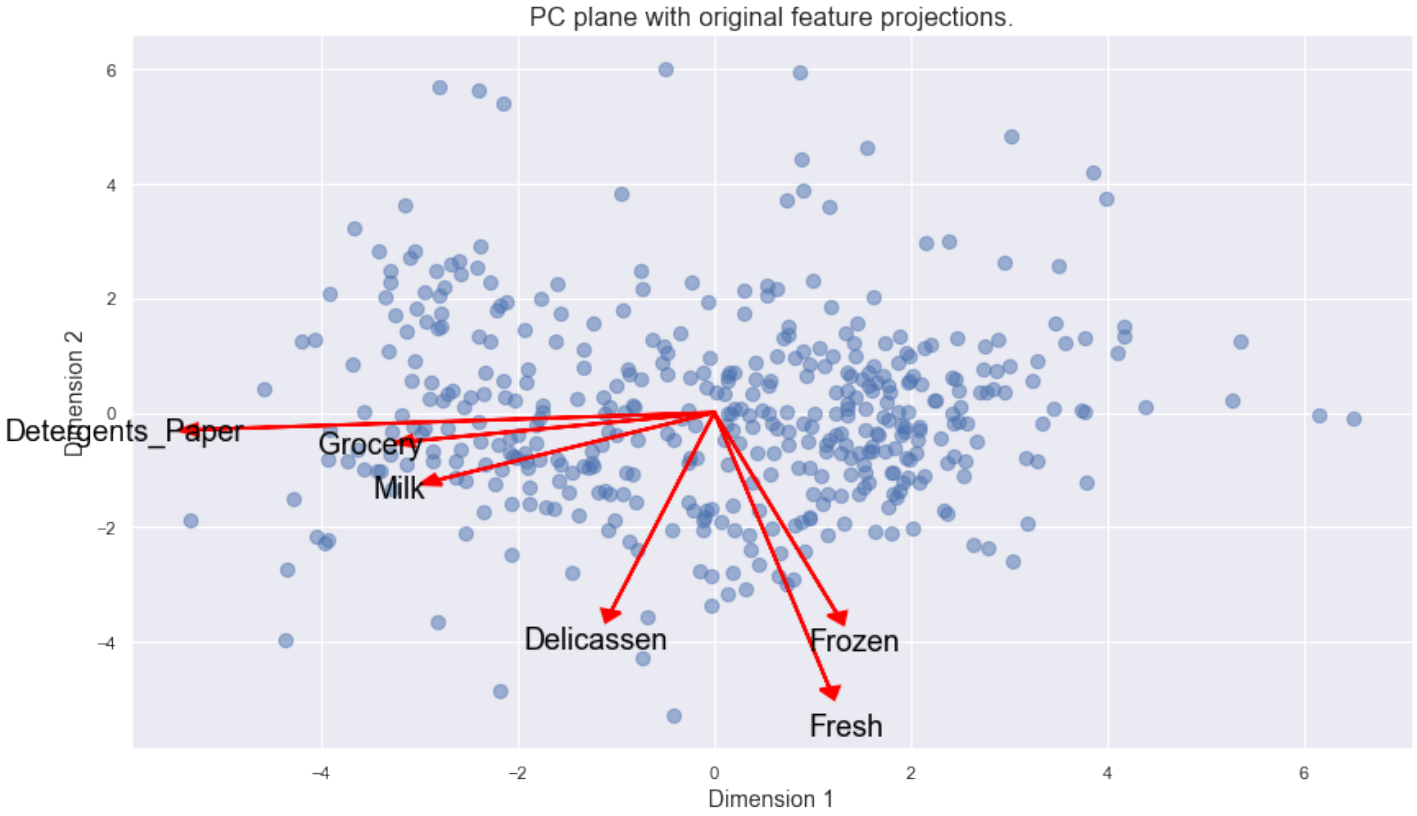
The cell below show how the log-transformed sample data has changed after having a PCA transformation applied to it using only two dimensions. Observe how the values for the first two dimensions remains unchanged when compared to a PCA transformation in six dimensions.

	Dimension 1	Dimension 2
0	-5.3316	-1.8845
1	-2.1899	-4.8605
2	3.0206	4.8169

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).

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.



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.

Modeling

Clustering

In this section, we will perform a comparative between K-Means clustering algorithm and Gaussian Mixture Model (GMM) clustering algorithm to identify which one fits better the various customer segments hidden in the data.

We will then recover specific data points from the clusters to understand their significance by transforming them back into their original dimension and scale.

K-Means vs GMM

1) The main advantages of using **K-Means** as a cluster algorithm are:

- It is easy to implement.
- With large number of variables, if (K is small), it may be computationally faster than hierarchichal clustering.

- Consistent and scale-invariant.
- It is guaranteed to converge.

2) The main advantages of using **Gaussian Mixture Models** as a cluster algorithm are:

- It is much more flexible in terms of cluster covariance. Which means that each cluster can have unconstrained covariance structure. In other words, whereas K-means assumes that every cluster have spherical estructure, GMM allows eliptical.
- Points can belong to different clusters, with different level of memebership. This level of membership is the probability of each point to belong to each cluster.

Creating Clusters

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.

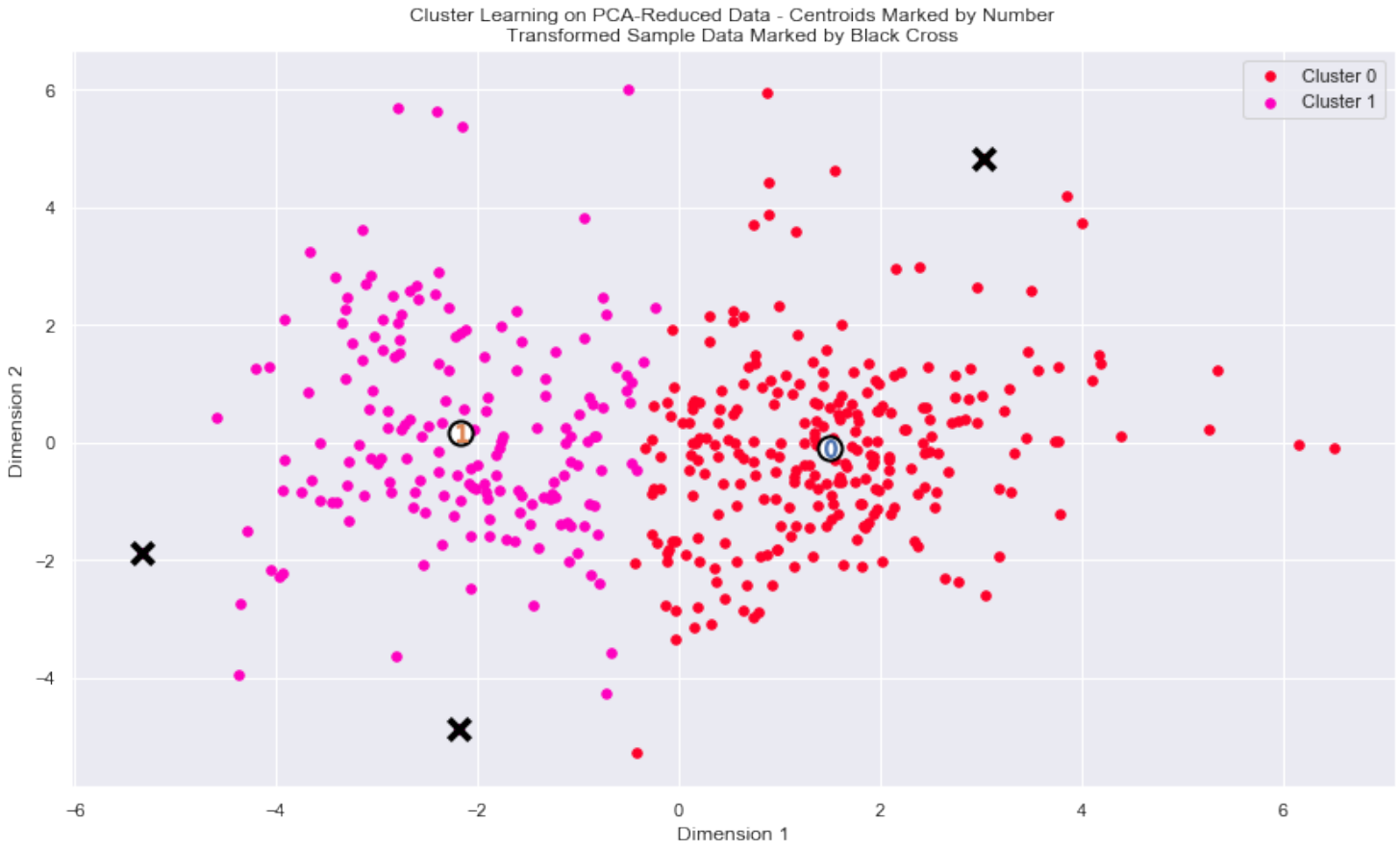
However, we can quantify the “goodness” of a clustering by calculating each data point’s **silhouette coefficient**. The silhouette coefficient for a data point measures how similar it is to its assigned cluster from -1 (dissimilar) to 1 (similar). Calculating the mean silhouette coefficient provides for a simple scoring method of a given clustering.

Calculated the silhouette coefficient both models converge on 2 clusters, with values of: 0.42628 for K-Means and 0.42192 for GMM.

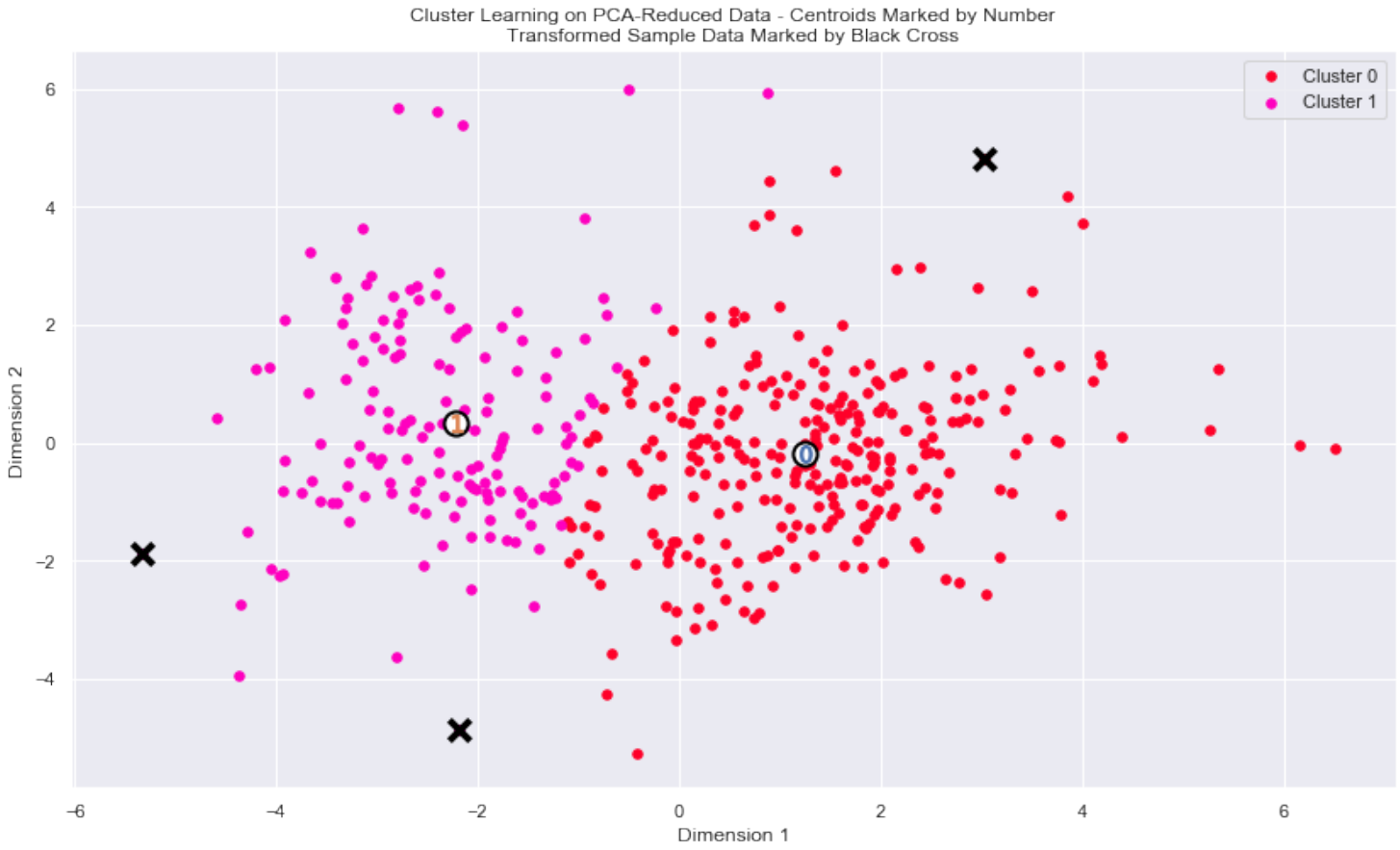
Cluster Visualization

Once we've chosen the optimal number of clusters for the clustering algorithm using the scoring metric above, we can now visualize the results in the plots below.

K-Means



GMM



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	8953.0	2114.0	2765.0	2075.0	353.0	732.0
Segment 1	3552.0	7837.0	12219.0	870.0	4696.0	962.0

- Segment 0 may represent a fresh food market as every feature except Frozen and Fresh are below the median.
- Segment 1 may represent a supermarket as every feature except fresh and frozen are above the median.

```
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
```

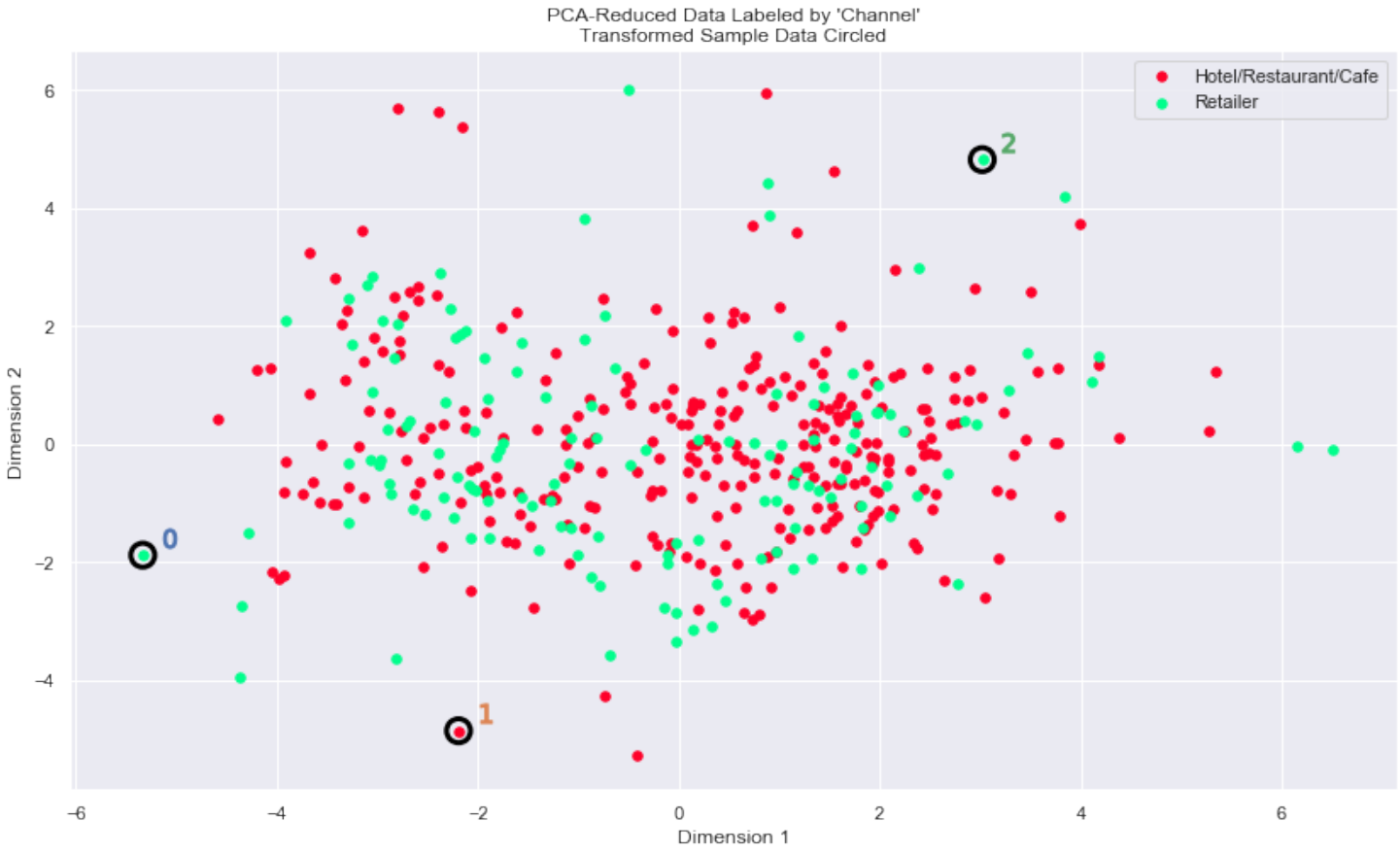
- Sample point 0 → Supermarket and the original guess was a retailer. This difference may be explained because of the size of the cluster (which is pretty big)
- Sample point 1 → Supermarket and the original guess was the same.
- Sample point 2 → Fresh food market and the original guess was a restaurant which is reasonable considering the amount of the spending of the features.

Results

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 plot below shows how each data point is labeled either 'HoReCa' (Hotel/Restaurant/Cafe) or 'Retail' the reduced space.



We can observe that the clusters algorithms did a somehow good job of clustering the data to the underlying distribution as the cluster 0 can be associated perfectly with a retailer and the cluster 1 to the Ho/Re/Ca.

Also we can appreciate that the difference between both clustering models is not very significant, since both came to the conclusion of two clusters fairly similar, and their silhouette coefficients are just different by 0.00436.

Recommendations

A supervised learning algorithm could be used with the estimated product spending as attributes and the customer segment as the target variable, making it a classification problem (we would have 2 possible labels).

As there is not a clear mathematical relationship between the customer segment and the product spending KNN could be a good algorithm to work with.