UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

1 SPECIFICATION REQUIRES CHANGES

Getting close here! I only ask you to review your answer on the meaning of the principal components (check below for more detail on what is to be changed). Keep up the good work and your next submission will certainly meet all specifications!

Data Exploration

Three separate samples of the data are chosen and their establishment representations are proposed based on the statistical description of the dataset.

Comment

Remember that some of the distributor's customers can be retailers! Particularly when the spending is above the median for many items, this seems like a plausible hypothesis, as a retailer would buy lots of products to resell to its customers.

A prediction score for the removed feature is accurately reported. Justification is made for whether the removed feature is relevant.

Awesome

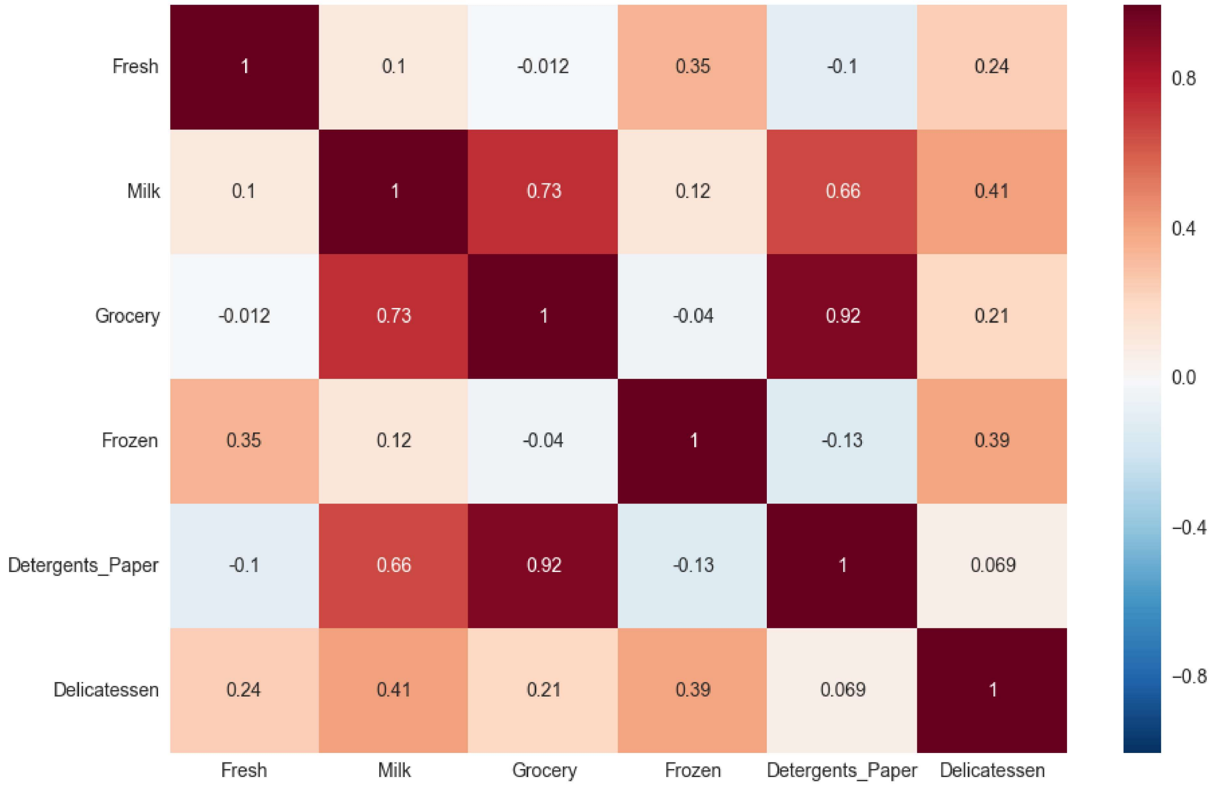
Great answer! Since most of the variation in `Detergents_Paper` can be explained by other variables, we could arguably remove `Detergents_Paper` from our features and still have enough information to model the target variable.

Student identifies features that are correlated and compares these features to the predicted feature. Student further discusses the data distribution for those features.

Awesome

Nice catch that the spending pattern seems to follow a log-normal distribution for all products. You could also produce a heatmap of the features' correlations:

```
import seaborn as sns
sns.heatmap(data.corr(), annot=True, annot_kws={'fontsize': 14});
```



Data Preprocessing

Feature scaling for both the data and the sample data has been properly implemented in code.

Student identifies extreme outliers and discusses whether the outliers should be removed. Justification is made for any data points removed.

Comment

There are actually 5 data points that are outliers for more than one feature. You can return them programatically in some ways - this is one of them:

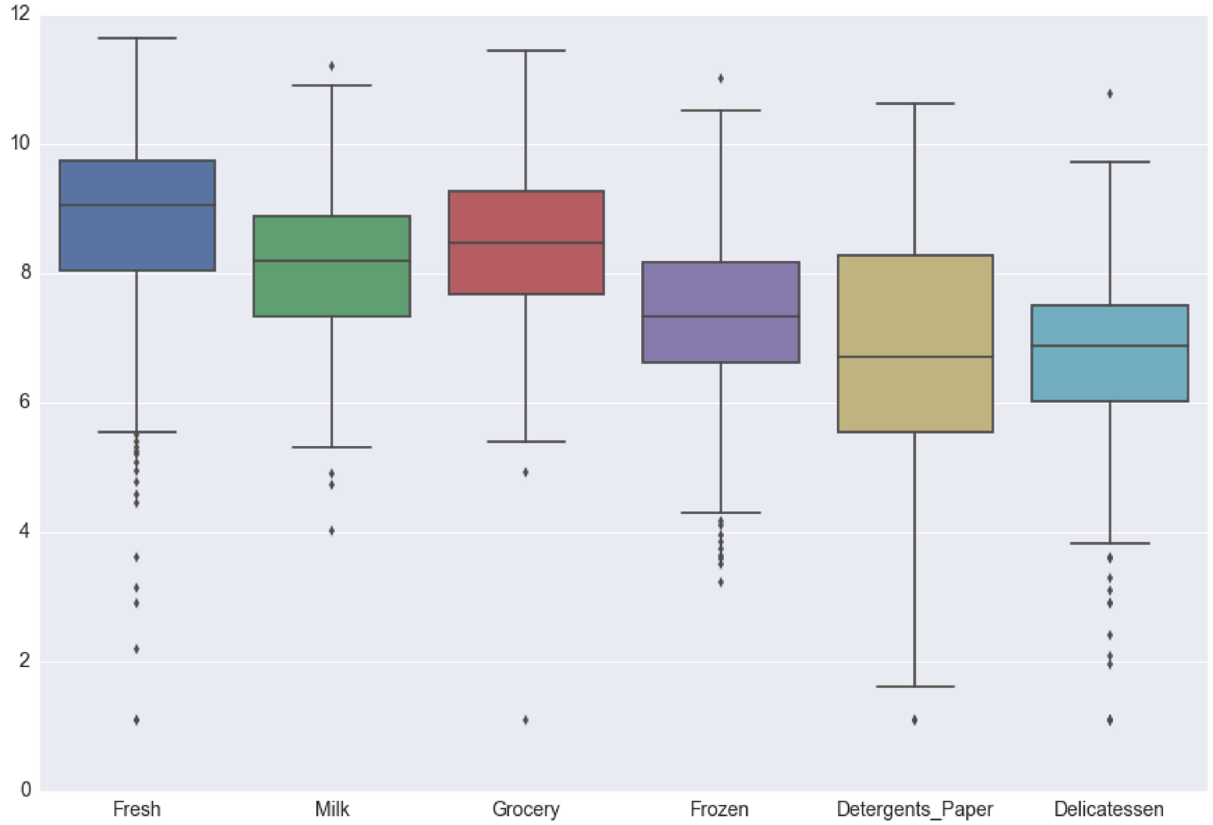
```
# dict with number of features in which each outlier is an outlier
outlier_dict = {index: outlier_indices.count(index) for index in outlier_indices}

# show outliers that are outliers in more than one feature
[outlier for outlier in outlier_dict if outlier_dict[outlier]>1]
```

> We are not removing the outliers because we want to build a robust model that will handles outliers and classify those as well. There is no reason for not classifying outliers for this example, once we are segmenting customers and there is no reason for not classifying a customer, even if he is an outlier.

This is a good justification, but note that outliers can heavily influence the PCA analysis you perform later on: the principal components are the directions in which the data varies the most, and an outlier will mean a large variation for the feature it's in, so outliers may mean some features will be overrepresented in the principal components.

Suggestion

One way to visualize outliers is with boxplots, such as the one below:



Based on these boxplots, I'd argue that the lowest value for `Grocery` is indeed a glaring outlier that should be removed from our data set, because it is not close to any other value - unlike the other outliers, which are close to the "outlier boundary" or to other data points. What do you think?

Feature Transformation

The total variance explained for two and four dimensions of the data from PCA is accurately reported. The first four dimensions are interpreted as a representation of customer spending with justification.

Required

Here each component does *not* represent a different customer segment; rather, they represent correlations between the features that are found across *all* customers.

You are correct to say, for example, that the first principal component indicates a correlation between `Detergents_Paper`, `Milk`, and `Grocery` - as we had previously seen in the scatter matrix from the notebook or the heatmap above. But this is not a kind of customer! Rather, it's a trait of the data set: most customers who spend a lot on one of these products will tend to spend a lot on the other two products as well.

PCA has been properly implemented and applied to both the scaled data and scaled sample data for the two-dimensional case in code.

Clustering

The Gaussian Mixture Model and K-Means algorithms have been compared in detail. Student's choice of algorithm is justified based on the characteristics of the algorithm and data.

Several silhouette scores are accurately reported, and the optimal number of clusters is chosen based on the best reported score. The cluster visualization provided produces the optimal number of clusters based on the clustering algorithm chosen.

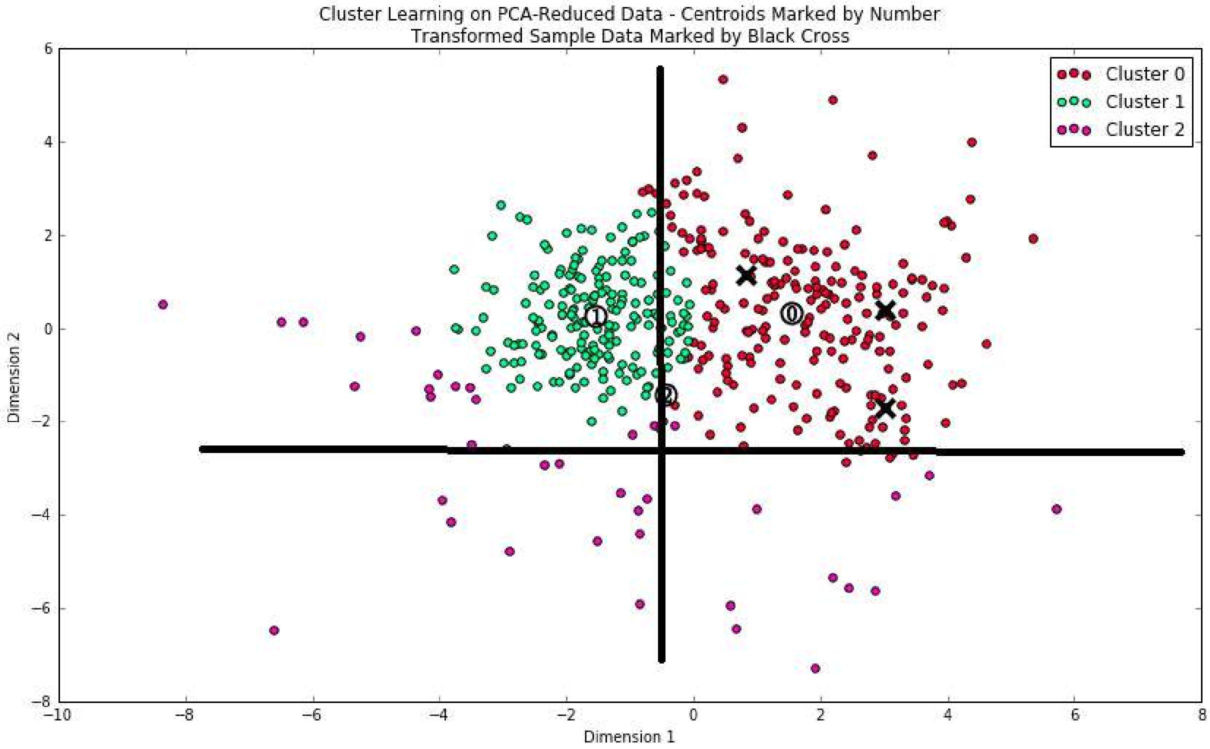
Comment

This may have to do with the train/test split you got, or (more likely) with the fact that you didn't remove any outliers, but most of the times the best number of clusters is 2 (which makes sense given what follows in the notebook!). However, I see no mistakes in your code, so you meet specifications here. :)

The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.

Suggestion

I like your answer here, but I would have tackled this question somewhat differently. Note in the poorly paintbrushed image below that two straight lines can pretty much define the boundaries in your clusters:



Cluster 1 and cluster 2 have similar values of the second principal component, but different values for the first one. And cluster 2 has smaller values for the second principal component.

Based on these observations, and on the meaning of each principal component, what would you say about each group of customers?

Sample points are correctly identified by customer segment, and the predicted cluster for each sample point is discussed.

Conclusion

Student correctly identifies how an A/B test can be performed on customers after a change in the wholesale distributor's service.

Awesome

Nice revision here! The fundamental thing here is that the segments should be tested *separately*, to avoid the contamination between segments that could happen when performing the test treating all customers as a single segment.

Student discusses with justification how the clustering data can be used in a supervised learner for new predictions.

Comment

I'm following the previous reviewer here, but I think the main point of this question is that the clusters *you have already defined* can be used as **features** in later analyses.

Comparison is made between customer segments and customer 'Channel' data. Discussion of customer segments being identified by 'Channel' data is provided, including whether this representation is consistent with previous results.

Comment

As I mention, in most cases two clusters will yield the best silhouette score, and they will approximate the separation between retailers and hotels/restaurants/cafes quite well. But you are right that your clustering also produced very good results!

☑ RESUBMIT

DOWNLOAD PROJECT

Learn the best practices for revising and resubmitting your project.

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Rate this review

Student FAQ