UDACITY

‹ Return to "Machine Learning Engineer Nanodegree" in the classroom

DISCUSS ON STUDENT HUB

Creating Customer Segments

REVIEW

HISTORY

Meets Specifications

Dear Student,

Overall, excellent work!. You demonstrate a good understanding of the concepts and methods deployed.

Congratulations on passing your exam and stay Udacious!

Data Exploration

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

Nice work using the mean to describe each of the samples. Another option is to **use the median instead of the mean** as it is less sensitive to outliers. For example:

Without Outlier	With Outlier
4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7	4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7,300

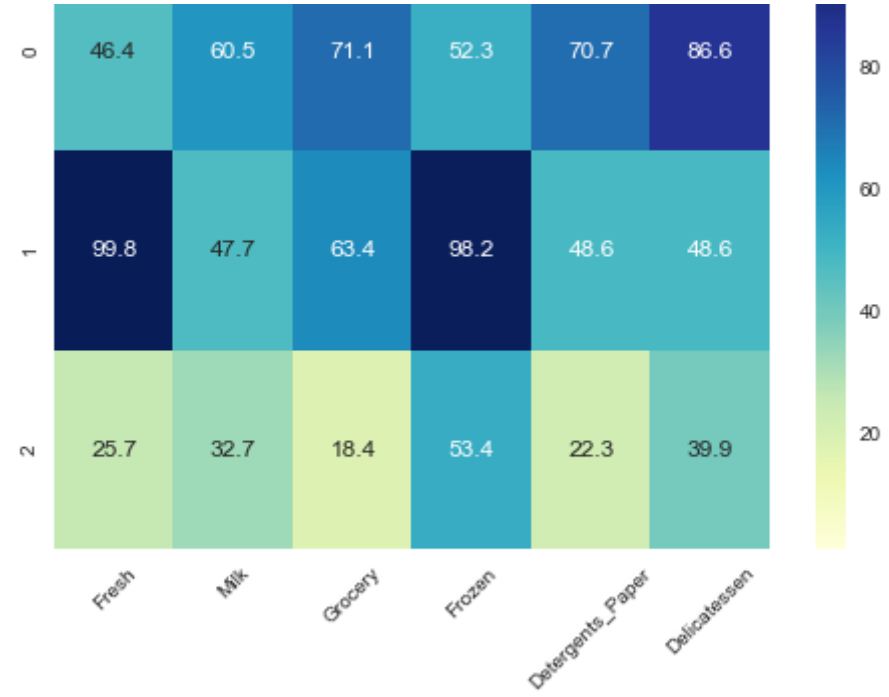
Mean = 5.45	Mean = 30.00
Median = 5.00	Median = 5.50
Mode = 5.00	Mode = 5.00
Standard Deviation = 1.04	Standard Deviation = 85.03

As a side comment, you could use pandas rank function to visualize which customers are above/below the average for each feature:

```
# get indices:
indices = [0, 1, 2]

# look at percentile ranks
pcts = 100. * data.rank(axis=0, pct=True).iloc[indices].round(decimals=3)

# visualize percentiles with heatmap
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
sns.heatmap(pcts.reset_index(drop=True), annot=True, vmin=1, vmax=99, fmt='.1f',
 cmap='YlGnBu')
plt.title('Percentile ranks of\nsamples\' category spending')
plt.xticks(rotation=45, ha='center');
```



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

Well done training a supervised regression learner for the selected features and reasoning that a positive coefficient of determination indicates this feature is "predictable" using other features and vice-versa.

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

Well done recognizing from the scatter matrix which features are significantly correlated, this visualization support your previous conclusion. Also, the diagonal plots can be used to determine the skewness of the different columns. As we can see, the data is significantly right-skewed!, here is a nice article about skewed distributions.

Data Preprocessing

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

Nice work transforming your data.

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

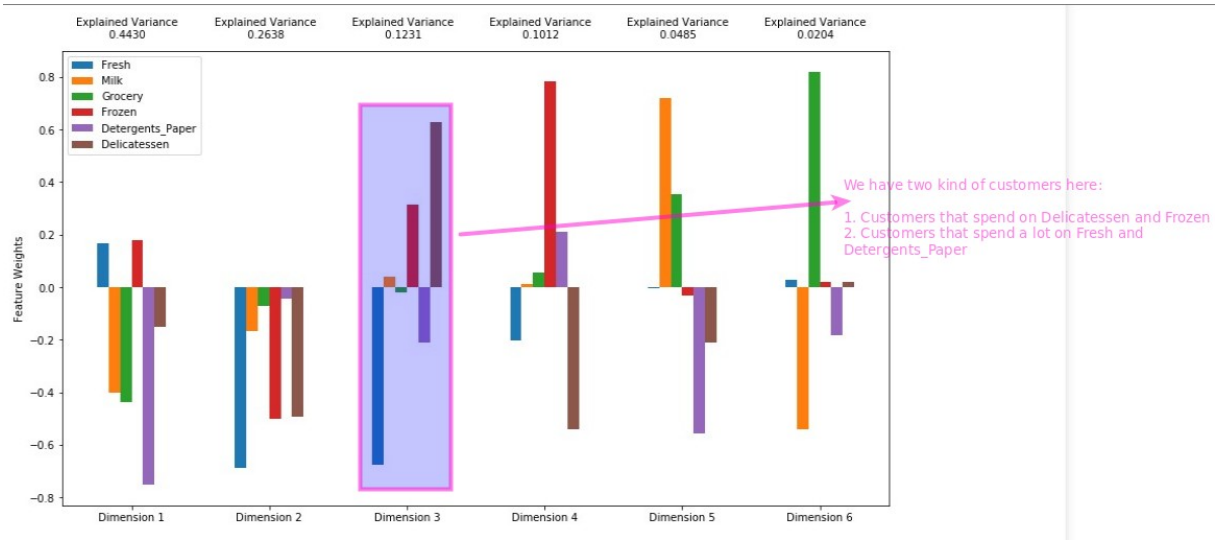
Great work removing outliers!

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.

Nice work giving a good understanding of the accumulated variance and describing the different dimensions, well done!. However, the dimensions' interpretation provided is not as accurate as required.

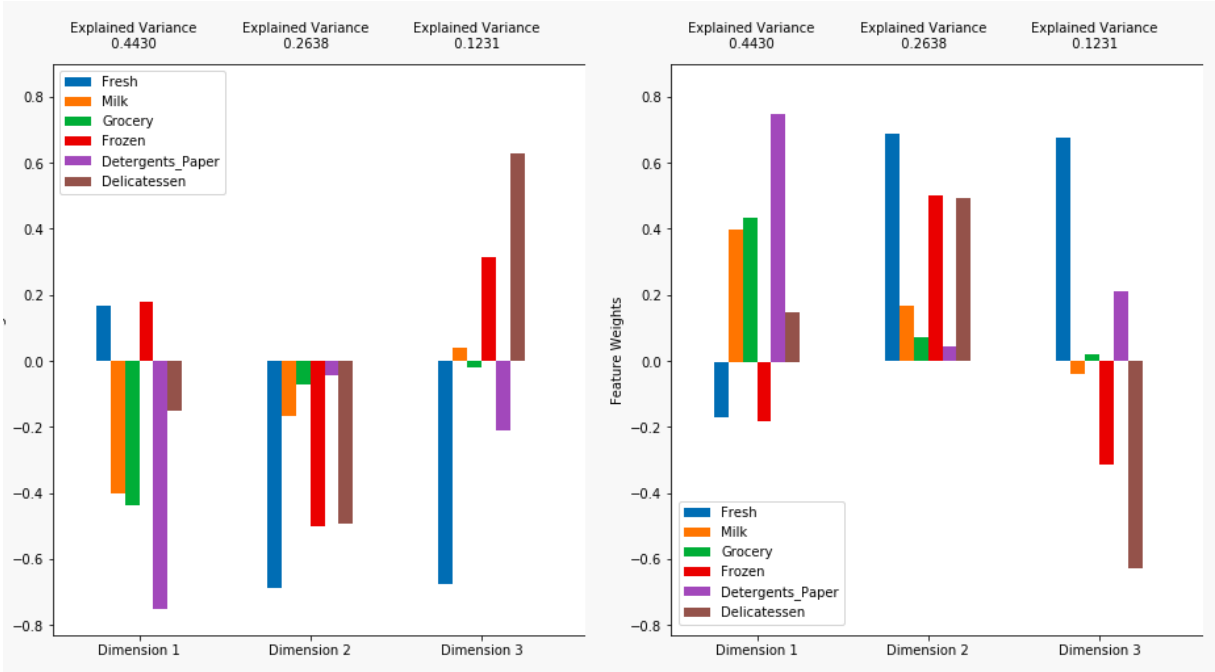
For example, for the third dimension, it is mentioned: "Fresh, Frozen and Delicatessen are the main contributors to the Dimension 3", but note that this dimension does not represent customers that expand a lot on Delicatessen **and** Frozen, but customers that spend on Delicatessen and **not** on fresh and **vice-versa**, basically, there are two different spending patterns. So, it is important to mention this dimension reflects two patterns of spending:



As a general comment, a principal component with feature weights that have **opposite** directions can reveal how customers *buy more* in one category while they *buy less* in the other category. Note also, that components' sign is arbitrary

For example, for any dimension, when we transform our data: customers are likely buying more of the **positive**-weight features while buying less of the **negative**-weight features. But note the **reverse is also true**, so there are customers that more likely buy *less* of the **positive**-weight features while buying *more* of the **negative**-weight features.

These are equal



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

Nice work! 💪

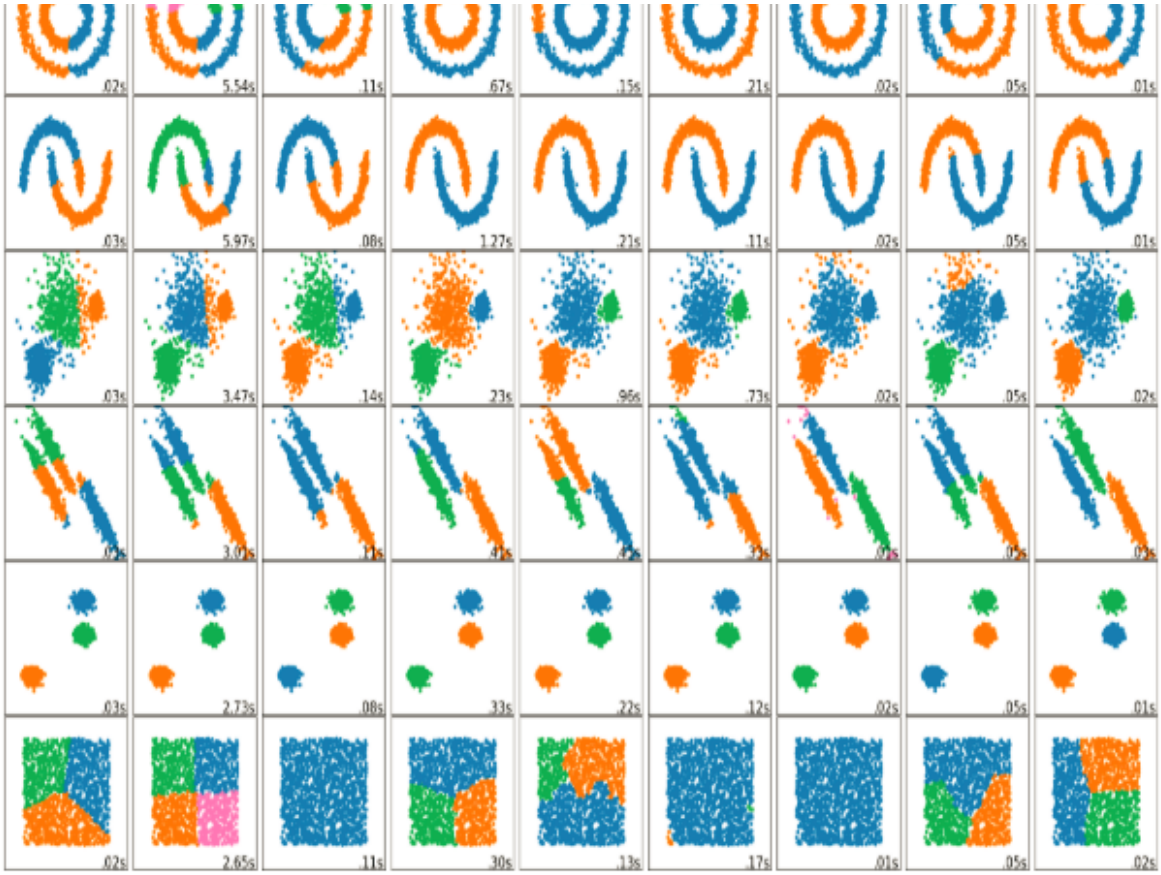
Clustering

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.

Nice work comparing both algotihms. Basically, there are two key considerations here:

- K-means is much faster than GMM as well as scalable
- GMM is richer as it is a soft clustering that provides information related to the cluster's wide

As a side comment, in this exercise, we used KMeans and GMM but note there are plenty of clustering algorithms out there. Of course, Scikit-Learn includes many of them and actually, there is a great overview:



As you can see in the image above, different clustering techniques offer different results for different datasets shapes. So, whenever you are choosing one of them, you need to identify some dataset characteristics that might be beneficial for the clustering technique to use.

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.

Excellent discussion to decide 3 as the number of clusters. 👌

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.

Well done explaining what kind of customer each cluster represents.

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

Well done describing your samples in terms of the clusters identified.

Conclusion

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

Good discussion on how this information could be used in an A/B test in order to test actions.

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

You could use the clustering-classification approach to assign a new customer to a segment group.

However, what we want here are the students to focus and recognize the output of a clustering method could be used for input in **any other supervised problem**. For example, in Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records. Probably we would use the clustering-classification approach to assign groups to new customers but would also use the already identified groups as an input feature for another supervised method.

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.

Nice work!

DOWNLOAD PROJECT

RETURN TO PATH