



---

## PROJECT

### Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

---

## PROJECT REVIEW

NOTES
-------


SHARE YOUR ACCOMPLISHMENT!  

## Requires Changes

8 SPECIFICATIONS REQUIRE CHANGES

Dear student,

Even though you need to revise answers to a few questions, this is a very good first attempt! I hope that the hints and the reading material given in this review will help you meet all the specifications in your next submission.

Keep up the hard work! 

## Data Exploration

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

<b>Required:</b>
------------------

To predict the establishments represented, please compare explicitly the features for all the sample points to the statistical measures of the dataset, like mean, median, etc.
---

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

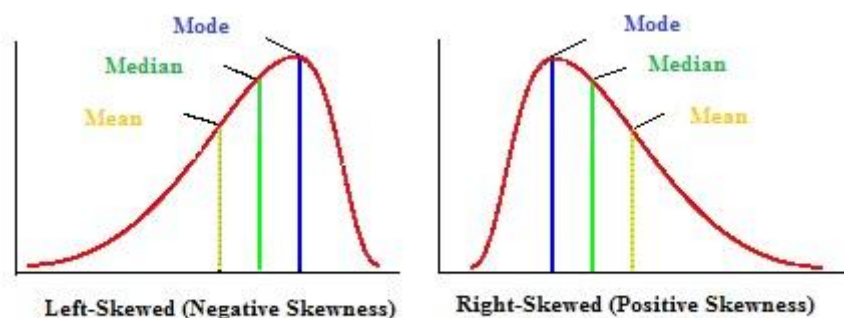
### Required:

- You do need to interpret the relevance of `Delicatessen` for categorising customers' spending habits, based on the  $R^2$ -score obtained.  
**Hint:** For the task of clustering or classification, do we gain much by adding a feature which can be predicted by other features, or the one which cannot be?
- Good job fixing the `random_state` while splitting the dataset, but please do the same for `Regressor` as well, so that we obtain the same score for every run of the program.
- (Optional) To mitigate the impact of a particular choice of `random_state(s)`, you can average the prediction scores over many values of `random_state(s)`, say, from 0 to 100.

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

### Remarks:

- The most significant correlation is definitely between `Grocery` and `Detergents_Paper`. `Milk` is also correlated with both these features, but the correlation is relatively mild. For the exact values, you can use `data.corr()` to get a matrix of correlations for all feature pairs.
- Well done remarking that the features' distribution is not normal, but long-tailed! Another way to describe the distribution would be that it is skewed to the right, as in the following graph:



Clustering algorithms discussed in this project work under the assumption that the data features are (roughly) normally distributed. Significant deviation from zero skewness indicates that we must apply some kind of normalisation to make the features normally distributed.

- Although I'm not making it obligatory, but please do interpret the relevance of `Delicatessen` in view of these correlations, after interpreting in the previous question, and comment upon how the two interpretations align.

## 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.

### Code issue

Your code,

```
Q1 = np.percentile(log_data, 25)
Q3 = np.percentile(log_data, 75)
```

calculates percentiles with respect to the all the features, while to find the outliers for a specific feature, we would need these percentiles for that particular feature. For example, to find the outliers for 'Grocery', the corresponding code should look like:

```
# This calculates the percentiles for just `Grocery`
Q1 = np.percentile(log_data['Grocery'], 25)
Q3 = np.percentile(log_data['Grocery'], 75)
```

How would this piece of code look within a `for` loop?

Once you fix your code, you need to identify all the outliers for more than one features (**Hint:** there are 5 of them).

Instead of manually looking for them, you can use the concept of [counter](#).

### Suggestions:

- You are free to keep/remove whatever outliers you like, but you must discuss the impact of including these on the variance of the dataset, and on the PCA and clustering algorithms performed later in the project. In particular, you might find that your decision here could have a huge impact later on the optimal number of clusters.

- You could also check this [article](#) for an excellent discussion on this topic, and among the four cases discussed, try to identify which case best characterises the outliers in our dataset.

## 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.

Good work getting the correct values of cumulative explained variance for the first two and four dimensions. You could also use the following code to compute these values:

```
print pca_results['Explained Variance'].cumsum()
```

### Required:

*"The third combines Fresh and Delicatessen and the Fourth Frozen and Delicatessen. "*

There seems to be some misconception here.

Note that the relative signs of features forming the PCA dimension are important, even though the sign of a PCA dimension itself is not. In fact, if you run the PCA code again, you might get the PCA dimensions with the signs inversed. For an intuition about this, think about a vector and its negative in 3-D space - both are essentially representing the same direction in space. You might find this [exchange](#) informative in this context.

The important thing to remark here is that a high/low (absolute) value along the PCA dimension can help differentiate between different types of customers. For example, a dimension giving relatively high (positive or negative) weights to `Fresh`, `Milk`, `Frozen` and `Delicatessen` would likely separate out the restaurants from the other types of customers.

The following links might be of some help in answering this question:

<https://onlinecourses.science.psu.edu/stat505/node/54>

<http://setosa.io/ev/principal-component-analysis/>

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

Good coding 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.

Good job comparing GMM and KMeans, but it would have been nice to discuss the difference in speed of the two algorithms.

From a practical standpoint, the main criteria for deciding between these two algorithms are the speed v/s second order information (confidence levels) desired and the underlying structure of our data.

### Regarding your choice of algorithm:

Your decision to use KMeans is absolutely fine. In my opinion, it is a good strategy to go with the faster KMeans for preliminary analysis, and if you later think that the results could be significantly improved, use GMM in the next step while using the cluster assignments and centres obtained from KMeans as the initialisation for GMM. In fact, many implementations of GMM automatically perform this preliminary step for initialisation.

I provide below some citations which might prove useful, if you would like to go deeper into the dynamics of these algorithms:

[http://home.deib.polimi.it/matteucc/Clustering/tutorial\\_html/mixture.html](http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html)

<http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier>

<http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html>

[http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means\\_Clustering\\_Overview.htm](http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm)

<http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means>

<http://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/>

<http://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/>

<https://shapeofdata.wordpress.com/2013/07/30/k-means/>

<http://mlg.eng.cam.ac.uk/tutorials/06/cb.pdf>

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.

*"However, there is also a local maximum at 6 Clusters, which represents a more detailed classification than just 2 clusters, which is rather trivial."*

If you check the Silhouette scores for number of clusters greater than 8, you would get many more local maxima. Local maximum is not a valid reason to choose the optimal number of clusters. Rather, we are looking for a global maximum while the number of clusters is within a reasonable bound. Evidently when the number of clusters start approaching the number of points in the dataset, the Silhouette score would start shooting up, because every point would

tend to be considered as a cluster by itself, therefore, such cluster-numbers should be considered unreasonable for finding global maximum.

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.

### Required:

Please revisit this after updating your choice for the optimal number of clusters.

And as suggested for Question 1, please make sure here as well that you make an explicit reference to the statistical description of the dataset when proposing the establishments represented by `true_centers`.

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

### Required:

Please revisit this after updating your choice for the optimal number of clusters.

And the ideal way of going about this question would be:

- Compare the features of the sample points to those of the `cluster_centers` and thus guess the cluster to which each sample point belong *before* running the code for `predictions`.
- Then, run the code and briefly discuss whether the predictions agree with your intuition or not.
- Lastly, compare to the conjectures made in Question 1 as well.

## Conclusion

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

Excellent! You have correctly identified the key point here which is to conduct the A/B test on each segment independently, since for an A/B test to be effective, the experiment group (A) has to be highly similar to the control group (B), before the treatment is applied to the experiment

group. If they are dissimilar to each other, then the result of the A/B test might be due to some variable other than the variable being tested.

Here are a few links for further reading on A/B testing:

<https://www.quora.com/When-should-A-B-testing-not-be-trusted-to-make-decisions/answer/Edwin-Chen-1>

<http://multithreaded.stitchfix.com/blog/2015/05/26/significant-sample/>

<http://techblog.netflix.com/2016/04/its-all-about-testing-netflix.html>

<https://vwo.com/ab-testing/>

<http://stats.stackexchange.com/questions/192752/clustering-and-a-b-testing>

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

Excellent suggestions!

**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.**

Please revisit this after updating your choice for the optimal number of clusters.

 RESUBMIT

 DOWNLOAD PROJECT





## Best practices for your project resubmission

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

[📺 Watch Video](#) (3:01)

RETURN TO PATH

Rate this review



---

[Student FAQ](#)