Creating Customer Segments

* [**REVIEW**](https://review.udacity.com/)
* [**HISTORY**](https://review.udacity.com/)

Meets Specifications

Excellent, your work is impressive. You really did a professional work here, congrats

Keep doing this to go on with your great trajectory.

Good luck with your next projects!

If you want to add me on the [Linkedin (Rafael Buck)](https://www.linkedin.com/in/rafaelbuck/" \t "_blank) feel free.

**Data Exploration**

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

Excellent description of selected examples. It was really cool. Good job selecting 3 sample observations.

**Suggested**: The part of translating data into concrete examples for a more business audience is one of the most important in the area of ​​data science and machine learning. I always recommend training this: smile:

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

Excellent, properly implemented. Note that lower scoring attributes are more relevant to the model since they can not be predicted. They are less affected by the collinearity effect. Attributes with higher scores are easily predicted by the other attributes, so they are not bringing any new information to the analysis.

**Suggested**: Here an [article that covers the question of multi-collinearity](https://www.kdnuggets.com/2015/03/machine-learning-data-science-common-mistakes.html/2) . And follows [another reference on parameter removal techniques before making the predictions](https://www.knime.com/blog/seven-techniques-for-data-dimensionality-reduction)

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

Very good results. The cool thing about them is that you can verify the relationship you found earlier. Freso is unrelated to anything, whereas Grocery and Detergents\_Paper are practically redundant. Another nice thing to note is that the data is positive skewed, more focused on the source, so the pre-processing of the next item is done in the data.

**Suggested**: Below is an example of the types of distribution that is important we note at this stage of the analysis:

And here's a [very cool article on how outliers affect distributions](https://www.quora.com/How-of-outliers-affect-normal-distribution-in-statistics).

**Data Preprocessing**

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

Great! An alternative would be the use of Box-Cox or even the preprocessing.scale of Sklearn: smile:

**Suggested**: Here is a nice link to [some techniques of how to do a normal transformation of a distribution](https://machinelearningmastery.com/how-to-transform-data-to-fit-the-normal-distribution/).

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

Your analysis of how to identify outliers was excellent. The explanation for the removal was good also.

**Suggested**: [here an article](https://www.kdnuggets.com/2017/01/3-methods-deal-outliers.html) on dealing with dealing with outliers. And [this article](http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/) also discusses whether or not to remove outliers. Another thing that is cool is to test our model’s performance with and without the presence of the outliers in the dataset. So, we can see the practical effect of them in the final results.

**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.**

Perfect, properly implemented.

**Suggested**: [follows an article](http://setosa.io/ev/principal-component-analysis/) that illustrates well a visual example of PCA. [This article here](https://onlinecourses.science.psu.edu/stat505/node/54/) also discusses how to analyze each dimension.

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

Perfect, the explained variances are correct. Note that with only 2 principal components it is already possible to make a representative clustering. It is also important to note the relevance of the absolute value of each category in that dimension.

**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.**

Excellent explanation of K-means and GMM and the justification for choosing the algorithm was also excellent.

**Suggested**: [here an article](https://ultraquartz.wordpress.com/2016/04/22/difference-between-k-means-and-gmm/) which discusses exactly that, quite interesting. If you want to test the GMM, I implemented my work (the text is in Portuguese, sorry about that): <https://github.com/rafaelmartinsbuck/machine-learning/blob/master/creating-customer-segments/customer_segments_PT.ipynb> if you're going to have fun follow as reference

**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.**

Customer review is good. Congratulations!

**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.**

Perfect, proposed groups the right way!

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

Awesome!!! It is where we return to a more business language, explaining what the respective cluster represents in more practical terms.

**Suggested**: Again, it's an excellent opportunity to go coaching this, which is one of the most important skills in the field of data science and machine learning

**Conclusion**

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

Great! Excellent analysis of how to run an A/B test for this problem. I have some suggestions to your test.

**Suggested**: It is important to note that if you test only one cluster, you may get an entirely different result on the other. In an A/B test it would be essential to understanding each cluster as a distinct, unmixed type of customer, to have a better knowledge of its customers. [Here an interesting article from the Netflix blog about this](https://medium.com/netflix-techblog/its-all-a-bout-testing-the-netflix-experimentation-platform-4e1ca458c15). There is also a [reference to how to conduct an A/B test in practice](https://blog.hubspot.com/marketing/how-to-do-ab-testing), what to take into account when planning it, and run it

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

Excellent answer to how to use the data for supervised learning. With this segmentation, it is possible to understand customers better and offer better strategies to approach each segment.

**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.**

Excellent analysis. Congratulations on the work and hope you had fun!

[**DOWNLOAD PROJECT**](https://review-api.udacity.com/api/v1/submissions/1679154/archive)

RETURN TO PATH