UDACITY

PROJECT

# Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

CODE REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

## Requires Changes

2 SPECIFICATIONS REQUIRE CHANGES

Hi there, it's Cláudio! Thanks for sending all the required files for the review process and for all code executing without any problem.
Congratulations for your project submission and for the quality presented in this challenge. You really did a great job.
However there are some areas which has opportunities to improve and to fix to match the requirements in this project rubric. I don't think you will have any problem to implement that as seems you have mastered these concepts very well.
I hope you had enjoyed doing this project and put in practice important concepts from machine learning. I will leave my contact below in case you have any doubt about this review as well to stay connected.
That's all. Enjoy machine learning and keep it up the great work. I will look forward for your next project submission.

Ask:
Are you going to evaluate this review? This is very important to me and Udacity. We take so seriously the student's experience that we are continuing learning from you.
Not going to give a 5 stars? Please let me know what opportunities to improve I can make for the next reviews. Let me know what you didn't like so I can always improve.
5 stars? WOW! Super thanks! That would be great to hear from you. Please, also let me know what you liked the most so I can keep it.

Email: cgimenest@uol.com.br
Linkedin: https://www.linkedin.com/in/claudiogimenestoledo/

## Data Exploration

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

Great job. You have selected correctly three separate samples:

> indices = [5,10,75]

This is a fundamental step in order to understand a little bit more your dataset during the exploratory analysis.

### Suggestion:

- Also a great start would be you getting statistics from the sample data. For instance using the describe from pandas method:

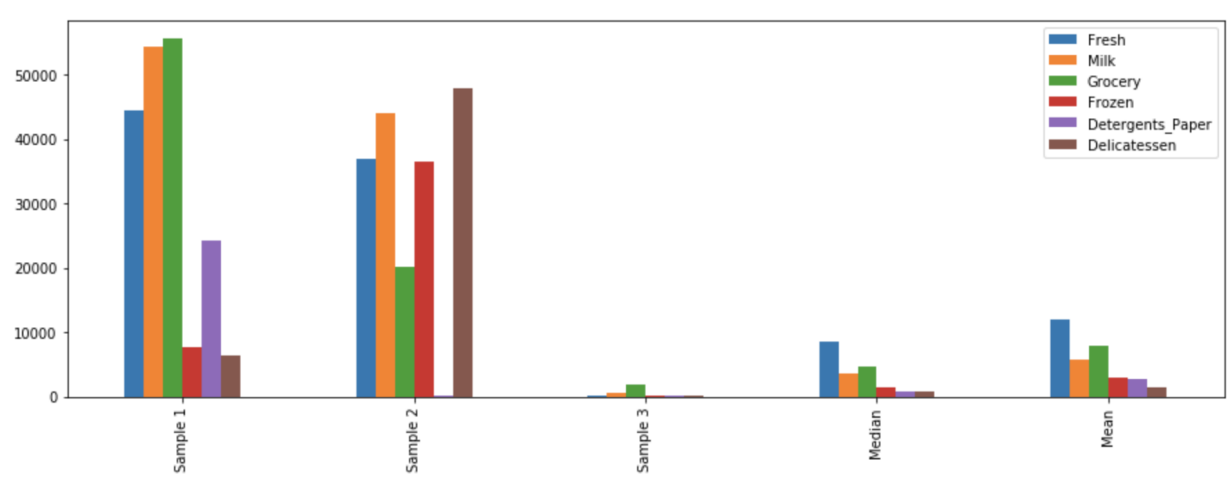
```
pd.sample_dataframe.describe()
```

- Alternatively you also may want to plot their correlations as well visual. For instance:

```
import matplotlib.pyplot as plt
import seaborn as sns
samples_for_plot = "sample values here"
samples_for_plot.loc[3] = data.median()
samples_for_plot.loc[4] = data.mean()

labels = ['Sample 1', 'Sample 2', 'Sample 3', 'Median', 'Mean']
samples_for_plot.plot(kind='bar', figsize=(15, 5))
plt.xticks(range(5), labels)
plt.show()
```

It will create something similar to this:




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

Almost there.

Actually in this part of exercise you should remove irrelevant variable as it won't affect the predictions. The objective of this section is to find out the relevance of features for the unsupervised clustering model—whether it's necessary for us to have that feature or if we can simply delete it without losing much information about our customers.

## Bonus:

- Here I will leave some good articles about feature selection: https://www.kdnuggets.com/2017/06/practical-importance-feature-selection.html

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

Great job.

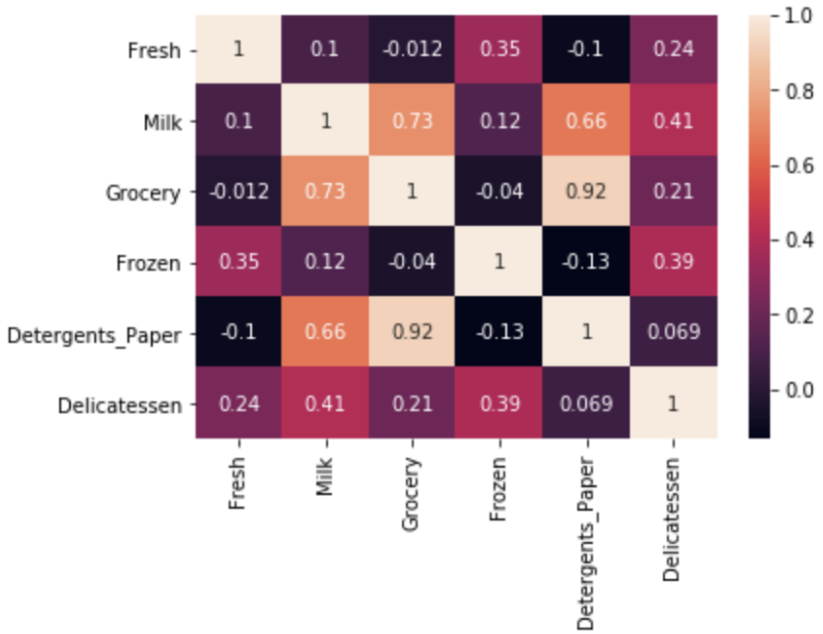
> Grocery <-> Detergents_Paper are correlated. Milk <-> Detergents_Paper, and Milk <-> Grocery are also correlated but not to the same degree.

## Suggestion:

- I would recommend you to implement the heat map which makes this job easier, take a look into an example below:

In [6]:
```
#Adding graphs for correlation to help out throught the process
import seaborn as sns
sns.heatmap(data.corr(), annot=True)
```

Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x1a1a732790>




## Data Preprocessing

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

Great job. You have correctly implemented the feature scaling as expected:

```
# TODO: Scale the data using the natural logarithm
log_data = np.log(data)

# TODO: Scale the sample data using the natural logarithm
log_samples = np.log(samples)
```

### Bonus:

- Here I will leave some good articles about why feature scaling matters:
  http://sebastianraschka.com/Articles/2014_about_feature_scaling.html
  http://scikit-learn.org/stable/auto_examples/preprocessing/plot_scaling_importance.html
  https://stackoverflow.com/questions/26225344/why-feature-scaling

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

Almost there.

Actually there are only 5 data points which are double-counted. Try to identify from the tables you have provided where a indice is also counted at another table.

For example:
65, 66, 75, etc...

### Bonus:

- Here is an great article discussing about the strategy of drop or not outliers, definitely check it out: http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

## 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 job. You have correctly identified the total variance for cumulative. You have find out this precisely:

> The cumulative explained variance for two and four dimensions is approximately 71% and 93%, respectively. This can change by a few percent based on what outliers are removed.

### Bonus:

- Here I will leave good articles about PCA analysis:
https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32
https://stats.stackexchange.com/questions/132976/does-pca-mean-selecting-most-important-features-and-ignoring-the-others
http://abhijitannaldas.com/dimensionality-reduction-and-principal-component-analysis-pca-explained.html

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

Good implementing this correctly:

```
# TODO: Apply PCA by fitting the good data with only two dimensions
pca = PCA(n_components=2)
pca.fit(good_data)

# TODO: Transform the good data using the PCA fit above
reduced_data = pca.transform(good_data)

# TODO: Transform log_samples using the PCA fit above
pca_samples = pca.transform(log_samples)
```

## 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 answer and concise one. Good choice on GMM.

### Suggestion:

- It's always a good idea to link references, images an business cases in order to convey a clearer message to all type of audience dealing with the problem and it's solution. Try to elaborate more on these type of discussions.

**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.**

Almost there. Usually the best silhouettes isn't 3. I would check if the outliers were removed properly.

> Two clusters will almost always give the best Silhouette score of approximately 0.42 (depending on what outliers were removed).

Awesome job plotting the scores you have got.

**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.**

Very well implemented:

```
# TODO: Inverse transform the centers
log_centers = pca.inverse_transform(centers)

# TODO: Exponentiate the centers
true_centers = np.exp(log_centers)
```

**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.**

Great explanation.

### Bonus:

- Here I will leave some articles about A/B testing in AI and how their are complementing each other:
  https://hackernoon.com/ai-as-complement-to-a-b-test-design-e8f4b5e28d92
  https://www.dynamicyield.com/ab-testing/
  https://www.mediapost.com/publications/article/305837/ab-testing-vs-ai-conversions.html

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

Good job. Your answer is accurate.

> The 'customer segment' labels can be used as an additional input feature, which a supervised learner could train on and then make predictions for the new customers.

**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.**

Good job. Answering to your question it might be too. Check it out these great discussion about it:

https://stats.stackexchange.com/questions/316199/pca-and-visualization-using-biplots-on-data-with-mixed-types
https://stackoverflow.com/questions/16705229/reversing-the-axis-range-in-3d-graph-in-python
https://sukhbinder.wordpress.com/2015/08/05/biplot-with-python/

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

## Student FAQ