UDACITY

PROJECT

Creating Customer Segments

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW

NOTES

SHARE YOUR ACCOMPLISHMENT!

Requires Changes

3 SPECIFICATIONS REQUIRE CHANGES

This was a truly exceptional submission! Your coding skills are excellent and your understanding of key clustering concepts is top-notch. I also greatly appreciate the many citations you provide throughout the discussion.

We just need a little more clarity and detail in analysis at some places, but nothing that requires more than some further reading on the topic, which you would probably enjoy anyways ;)
I'm sure your next submission will be close to perfect.

Keep up the hard work and have fun learning!

Data Exploration

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

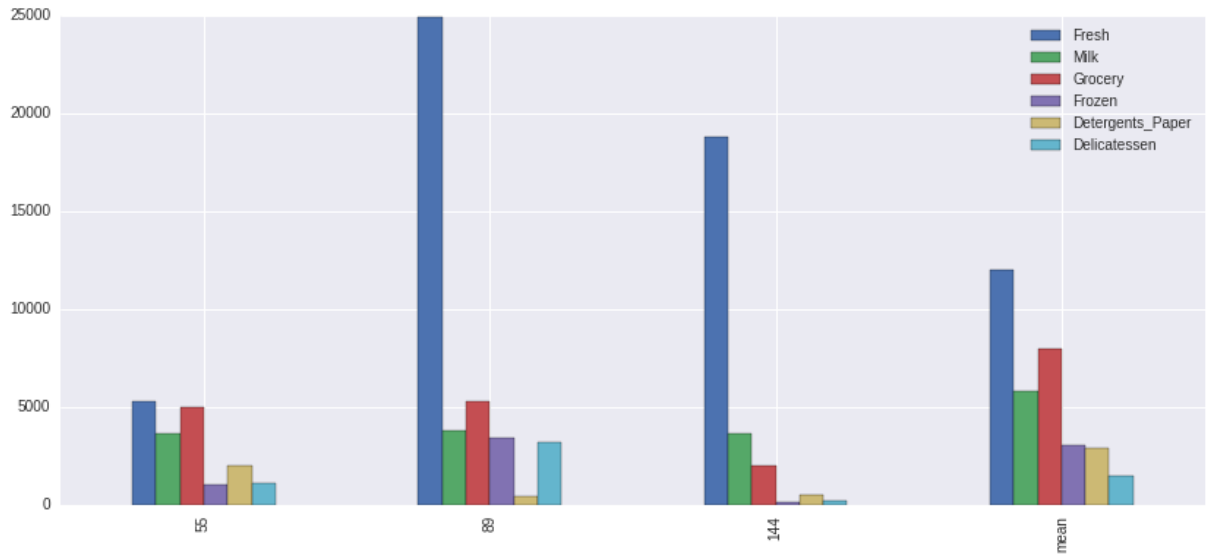
Good work choosing three sample points which differ by a wide margin across at least a few features (intriguing use of Fibonacci :)). Nice comparison of sample points to some of the dataset statistics, in order to obtain a fairer judgement of which establishments they represent.
However, a few observations are slightly off the mark, for instance, when you say that the first point *"is pretty close to the mean in every value and in the 50% or 75% percentile."* You could see this easily using visual representations below, which I strongly encourage you to implement.

Pro Tips:

COMPARING TO DATASET AVERAGE

The following code would draw a bar plot to visualise the amount of each product purchased for each sample, together with the dataset mean.

```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns
samples_bar = samples.append(data.describe().loc['mean'])
samples_bar.index = indices + ['mean']
_ = samples_bar.plot(kind='bar', figsize=(14,6))
```



This will make comparing the three different sample points with each other much easier.

ADVANCED: COMPARING THE PERCENTILES

You could also produce a simple heat map of the percentiles from the dataset. This will make comparison of the different relative values of the purchased units for each product much easier. Below is a simple implementation:

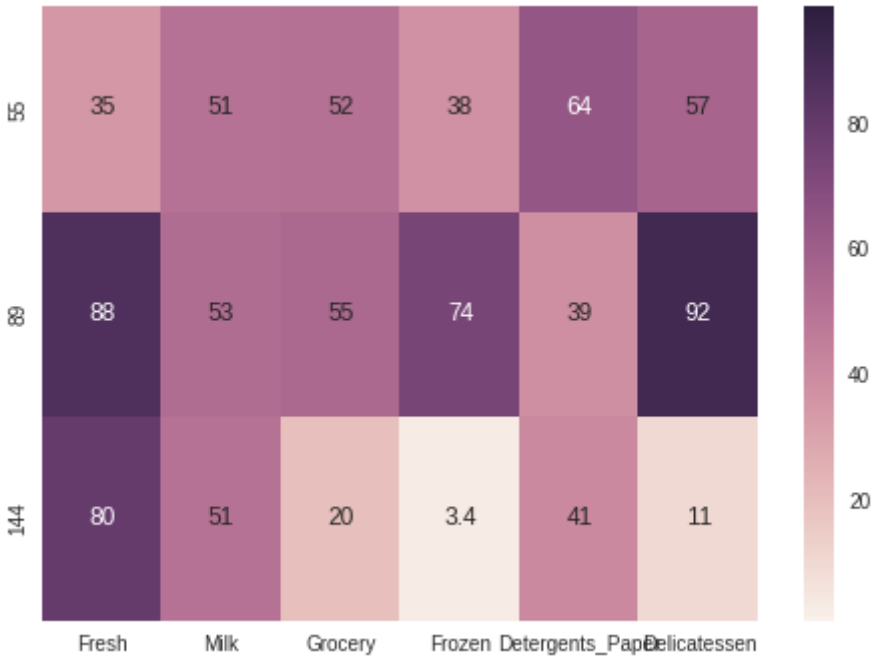
```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns

# First, calculate the percentile ranks of the whole dataset.
percentiles = data.rank(pct=True)

# Then, round it up, and multiply by 100
percentiles = 100*percentiles.round(decimals=3)

# Select the indices you chose from the percentiles dataframe
percentiles = percentiles.iloc[indices]

# Now, create the heat map using the seaborn library
_ = sns.heatmap(percentiles, vmin=1, vmax=99, annot=True)
```



Hope this is helpful for you to make an unbiased explanation!

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

Correct, indeed! A feature that can be predicted from other features would not really give us any additional information and thus, would be a fit candidate for removal. The fact that `Delicatessen` cannot be predicted implies that it is relevant.

Also, good coding job! In particular, fixing the random states at appropriate places is a good, though often neglected, practice.

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

Excellent work correctly identifying the correlations as well as the skew of the distribution! The latter is very important to determine the normalization that we want to apply to the data.
And your finding from the previous question on the `Delicatessen` feature nicely aligns with your finding from this section.
Well done!

Data Preprocessing

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

Suggestion :
It is good to be cautious and work with copies of data whenever there is any fear of aliasing), but it is also

important to keep in mind the memory inefficiencies this frequent copying would introduce for very large datasets.

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

Awesome coding prowess on display here!

Comments and Suggestions

- An alternative, perhaps a tad simpler way to find the outliers for more than one features is:
  Inside the `for` loop, modify

  ```
  outlier_indices_list.append(feature_outliers.index)
  ```

  to

  ```
  outlier_indices_list.extend(feature_outliers.index)
  ```

  and once out of the `for` loop, use

  ```
  outliers_multiple_set = set([x for x in outlier_indices_list if outlier_ind
  ices_list.count(x) > 1])
  outliers_multiple_list = list(outliers_twice)
  ```

- Overall, a great job with identifying the outliers. However, the justification to keep some and remove some of the multiply-counted outliers is slightly off. From the point#4 of the source that you yourself provided:
  "More commonly, the outlier affects both results and assumptions. In this situation, it is not legitimate to simply drop the outlier. "
  But you seemed to basing your decision of choosing outliers solely on how they align with your assumptions regarding the "correlation" among various features. Ideally, as mentioned in your excellent source, you should *"run the analysis both with and without it,"* and see *"how the results changed"*.
- It would also be nice to discuss the impact of including these outliers on PCA and clustering algorithms performed later in the project.
- Bravo for including the source, a very healthy practice!

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.

Fantastic work here! Indeed, the first four principal components explain most of the variance in this dataset. Also, a great job in interpreting the first four dimensions as a representation of customer spending.

Suggestions:

- PCA is an important and fascinating topic. You might find these links helpful in advancing your understanding:
  https://onlinecourses.science.psu.edu/stat505/node/54
  http://setosa.io/ev/principal-component-analysis/
- It is a good practice to code even for seemingly trivial tasks. For example, you could use

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

  to obtain the total variances for the first four components.

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.

This section exceeds expectations! It seems that you have gone quite deep into these two models, and I would probably learn a thing or two from the many citations that you provided. While I go through them, I would share some more resources, seeing your evident enthusiasm for the topic :)

SUGGESTED READING

- *Gaussian Mixture Model*
  http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html
  http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier
- *K-Means*
  http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html
  https://sites.google.com/site/dataclusteringalgorithms/k-means-clustering-algorithm
  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

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.

Great job, not only with correctly identifying that the best Silhouette score is given by 2 clusters, but also for playing around with many different parameters such as `covar_type`, `metric` etc.

NITPICKING :)

It is evident from your `for` loop that you must have checked Silhouette score for various number of clusters, but apparently forgot to explicitly report them in the Question 7.

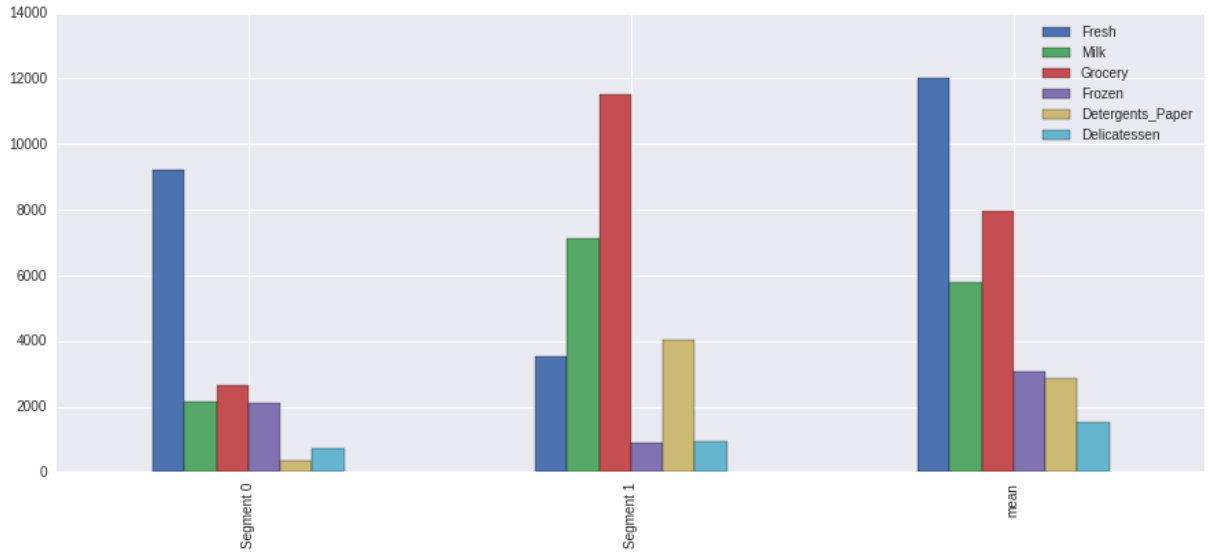
The establishments represented by each customer segment are proposed based on the statistical description of the dataset. The inverse transformation and inverse scaling has been properly implemented and applied to the cluster centers in code.

No issues with coding or the prediction of potential customers, but for this section we need a bit more analysis on the purchasing behavior of the cluster segments. Therefore make sure you compare both centroids to the statistical description of the dataset, similar to the way it was done in Question 1. This would also be a good opportunity to implement my code suggestions for Question 1 ;)
To remind,

COMPARING TO DATASET AVERAGE

The following code would draw a bar plot to visualise the amount of each product purchased for each `true_center`, together with the dataset mean.

```
# Import Seaborn, a very powerful library for Data Visualisation
import seaborn as sns
true_centers = true_centers.append(data.describe().loc['mean'])
_ = true_centers.plot(kind='bar', figsize=(15,6))
```



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

Again, this question needs some more details.
Required: Before running the `predictions`, please compare the `true_centers` of the algorithm to the

sample points (using the above visualisation tools, for example) , then run the `predictions` and briefly discuss whether the predictions agree with your intuition or not.

Conclusion

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

"Firstly, I would recommend to select customers for an A/B test from the same cluster only to minimize differences between them."
Indeed! The key is to conduct the A/B test on only one segment at the time, as for an A/B test to be effective, the experiment group has to be highly similar to the control group, 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.

Suggestion:

Your discussion following the statement that I cited here is somewhat muddled and can benefit from a clearer rewrite. I would suggest going through these links before you attempt that.

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.

"The supervised learner could use the posterior probability and of course the classification from the trained GMM."
Again, you start with an excellent first statement, but you must follow it up with a discussion of how using the `cluster_labels` or `posterior_probabilities` as features would help in the task of supervised classification. Consider, for example, the preference for a particular delivery time as something we want to predict. If we do indeed run the A/B testing above, what is the best way to use it for supervised learning?

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.

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)

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

RETURN TO PATH

Student FAQ