Meets Specifications

Hi Student,

Congrats on finishing this project. I can understand the pain you have gone through for submitting this project. I have to appreciate the effort that you kept in this project.

Student the effort that you have kept in this project never goes in vain. This project gives an opportunity to explore more because there is no straightforward solution for this. I can say the knowledge you gain by doing this project is really invaluable. Udacity is more of knowledge sharing and you can be master in a specific domain by completing the project, I want to see you out stand with the skills that you learn here.

**Resources:**

I recommend the classic Harvard [CS109](http://cs109.github.io/2015/pages/videos.html) course and studying good Kaggle kernels like [this one](https://www.kaggle.com/arthurtok/principal-component-analysis-with-kmeans-visuals?scriptVersionId=1543947) and good blogs like [this one](https://machinelearningmastery.com/?s=clustering&post_type=post&submit=Search). You can also look at some more segmentation techniques like latent class segmentation which will be helpful in your career. Want to see you have a good career in Data Science...

Finally, Excellent job implementing and submitting a great project  I can sense that you have spend hours doing this, that inspired me to take some time for giving you some valuable suggestions.

Share your comments on this review..

**Preprocessing**

**All missing values have been re-encoded in a consistent way as NaNs.**

Nice job,

**Advice**: Adding function for removing Nan's will make your code look modular

**Columns with a large amount of missing values have been removed from the analysis. Patterns in missing values have been identified between other columns.**

Outlier columns have been removed successfully with detailed discussion.

**The data has been split into two parts based on how much data is missing from each row. The subsets have been compared to see if they are qualitatively different from one another.**

Good job plotting the two different discrete distributions.

**Categorical features have been explored and handled based on if they are binary or multi-level.**

**Mixed-type features have been explored, resulting in re-engineered features.**

Good job exploring those two mixed type features.

**Dataset includes all original features with appropriate data types and re-engineered features. Features that are not formatted for further analysis have been excluded.**

All original features with appropriate data types are re-engineered features. Awesome.

**A function applying pre-processing operations has been created, so that the same steps can be applied to the general and customer demographics alike.**

Well done.

**Feature Transformation**

**Feature scaling has been properly applied to the demographics data. Imputation has been performed to remove remaining missing values.**

Awesome job

**Advice**: There is no perfect solution here. There can be many no of solutions

**Principal component analysis has been applied to the data to create transformed features. A variability analysis has been performed to justify a decision on the number of features to retain.**

**Complement**: Good job plotting the scree plot to visualize the variance. I have to appreciate this work.

**Weights on at least three principal components are used to make inferences on correlations between original features of the data. General meanings are ascribed to principal components where applicable.**

Good work on mapping the weights and explaining the relative spread.

**Clustering**

**Multiple cluster counts have been tested on the general demographics data, and the average point-centroid distances have been reported. A decision on the number of clusters to use is made and justified.**

Awesome Job.

score = np.abs(model.score(data))

**Complement**: You have used the absolute function. Many students made mistake by not considering the sign

**Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data.**

Great job.

This is the most important section and glad to see that you have not refitted the customers data again.

Project theme is of identifying relative spread of latent factors across two different datasets. You got this right

**A comparison is made between the general population and customers to identify segments of the population that are central to the sales company's base as well as those that are not.**

Glad to see that you have mentioned this here

**segments of the population that are relatively popular with the mail-order company(cluster 3): most of them are middle aged and elderly people（between 30 and 60 years old),typical house owners,household income level is average,likely to have children in household,wealth / Life Stage is likely to be middle class, unemployment level in the comunnity is relatively low.**