UDACITY

Logout

PROJECT SPECIFICATION

Identify Customer Segments

Preprocessing

CRITERIA	MEETS SPECIFICATIONS
Have data values representing missing values been encoded consistently?	All missing values have been re-encoded in a consistent way as NaNs.
Has the quality of the data been assessed by looking at missing values by columns?	Columns with a large amount of missing values have been removed from the analysis. Patterns in missing values have been identified between other columns.
Has the quality of the data been assessed by looking at missing values by rows?	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.
Have categorical features been properly processed?	Categorical features have been explored and handled based on if they are binary or multi-level.
Have mixed-type features been properly processed?	Mixed-type features have been explored, resulting in re-engineered features.
Has the dataset been cleaned to include only relevant columns?	Dataset includes all original features with appropriate data types and re-engineered features. Features that are not formatted for further analysis have been excluded.
Has a function performing cleaning operations been created?	A function applying pre-processing operations has been created, so that the same steps can be applied to the general and customer demographics alike.

Feature Transformation

CRITERIA	MEETS SPECIFICATIONS

Has feature scaling been applied to the data?	Feature scaling has been properly applied to the demographics data. Imputation has been performed to remove remaining missing values.
Have an appropriate number of features been selected following dimensionality reduction?	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.
Have relationships between original features been investigated for the data?	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.

Clustering

CRITERIA	MEETS SPECIFICATIONS
Have an appropriate number of clusters been set?	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.
Have the customer demographics data been handled in a fashion consistent with the general demographics data?	Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly to the customer demographics data.
What customer segments are popular or unpopular with the mail-order sales company?	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.

Suggestions to Make Your Project Stand Out!

- Use the data dictionary to understand the meaning of variables in the dataset.
- Use plotting packages like matplotlib and seaborn as necessary to visualize trends in the data.
- Add additional cells to your notebook to organize your code in manageable pieces and make notes on your observations.