**Meets Specifications**

The report is clean and organized and the code is neat. The arguments used in the discussions are solid. Congratulations!

**Preprocessing**

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

All missing values have been re-encoded in a consistent way as NaNs.  
Convert data that matches a 'missing' or 'unknown' value code into a numpy NaN value. You might want to see how much data takes on a 'missing' or 'unknown' code, and how much data is naturally missing, as a point of interest.

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

Well done! Columns with a large amount of missing values have been removed. the six outlier columns you found below TITEL\_KZ  
AGER\_TYP  
KK\_KUNDENTYP  
KBA05\_BAUMAX  
GEBURTSJAHR  
ALTER\_HH  
is right selection.  
There are a few columns that are outliers in terms of the proportion of values that are missing. You will want to use matplotlib's hist() function to visualize the distribution of missing value counts to find these columns.

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

Well done !! Mixed-type features have been explored, resulting in re-engineered features. There are a handful of features that are marked as "mixed" in the feature summary that require special treatment in order to be included in the analysis. There are two in particular that deserve attention; the handling of the rest are up to your own choices.

**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.**

The data are correctly split off in data points that have at least nine to thirty-two missing values.

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

Well done !! Categorical features have been explored and handled based on if they are binary or multi-level.  
There is one binary variable that takes on non-numeric values. For this one, you need to re-encode the values as numbers or create a dummy variable.

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

Nice to see you dropped features like PRAEGENDE\_JUGENDJAHRE and CAMEO\_INTL\_2015.

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

A function applying pre-processing operations has been created. Nice work !! it's important to look ahead to the future and realize that you'll need to perform the same cleaning steps on the customer demographics data.

**Feature Transformation**

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

Feature scaling has been properly applied to the demographics data. Nice work !! we need to perform feature scaling so that the principal component vectors are not influenced by the natural differences in scale for features.

**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.**

Weights on at least three principal components are used to make inferences. Good work !!

**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.**

Well done !! Principal component analysis has been applied to the data to create transformed features. each principal component is a unit vector that points in the direction of highest variance (after accounting for the variance captured by earlier principal components). The further a weight is from zero, the more the principal component is in the direction of the corresponding feature.

**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.**

Nice work !! choice in clusters has come from a turning point or elbow in the curve. To find a good amount of centroids, the K-Means classifier from scikit learn can be fitted for several possible cluster sizes. The score for each cluster size was taken after the fitting process from KMeans.inertia\_ attribute which the documentation describes as the Sum of squared distances of samples to their closest cluster center and thus can be used as score.

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

Wonderful !! Cleaning, feature transformation, dimensionality reduction, and clustering models are applied properly.

**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.**

Comparison is made between the general population and customers to identify segments. Good work !! Consider the proportion of persons in each cluster for the general population, and the proportions for the customers. If we think the company's customer base to be universal, then the cluster assignment proportions should be fairly similar between the two. If there are only particular segments of the population that are interested in the company's products, then we should see a mismatch from one to the other.