

< [Return to Classroom](#)

Identify Customer Segments

REVIEW

HISTORY

Meets Specifications

OVERALL COMMENTS

Congratulations on finishing the project 🎉

This was a brilliant submission. You did a great job and should be proud of yourself. After reviewing this submission, I am impressed and satisfied with the effort and understanding put in to make this project a success. All the requirements have been met successfully 100 %

I have tried to provide you a detailed review by adding :-

1. Few Suggestions which you can try and improve your project.
2. Appreciation where you did great
3. Some learning opportunities for knowledge beyond coursework

I hope you find the complete review informative 😊👍

Keep doing the great work and all the best for future project.

Few Links to Refer

- [DRY Principle](#)
- [PCA](#)
- [K means](#)
- [Why is elbow important in K Means](#)

Preprocessing

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

Importance of this Rubric

- It is important to clean the data and remove the missing values to perform correct analysis
- There can be different patterns by which missing datas are represented. In this project we cover few of them like

x

xx

, etc.

How you did

- Outlier columns in the dataset is correctly identified and removed.
- Correctly identified the columns that have the same or similar counts of missing values.

[How to handle missing values in data](#)

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

Importance of this Rubric

Since data can have inconsistency in terms of missing value representation, it is important for us to identify them and encode it in consistent way. In this case we encode using Nans

How you did

- Good job encoding the missing values.
- Missing value codes given in feat_info's last column have been used to convert all codes to NaNs.
- Taken care of the 'X' and 'XX'

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

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

How you did

- Well done on selecting the binary feature for your data
- You have correctly re-encoded the `OST_WEST_KZ` to numerical binary feature for using it.
- You have dropped the multi-level feature which is absolutely fine and acceptable. In most cases, high cardinality makes it difficult for the model to identify such patterns; hence, the model doesn't generalize well to examples outside the training set.

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

Two mixed-type features, PRAEGENDE_JUGENDJAHRE and CAMEO_INTL_2015, engineered into two new features. -- Well done on this

How you did

- Well done on creating dictionary for mapping the values.
- Well done on removing the mixed features after generating new separate columns.

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.



Importance of this Rubric

This rubric helps us to analyse how two different splits of data with various missing values perform on feature distribution.
You can learn the importance of removing missing data which might be noise by observing the distinctive plots of two different sets

How you did

- Well done on splitting the data into two by taking the threshold value as nominal, which is neither too high or two low
 - It is essential to look at the distribution over specific columns for rows with more than a certain number of missing values, with the rows more petite than a certain number of missing values.
- These bar charts show us a summary of all the values in a single picture
• Good visuals and comparisons of the features provided in the notebook.

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

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

How you did

- Well done on retaining the original features which didn't need any modification.
- Having reengineered the mixed value features, it is a right choice to drop the original features.

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

Importance of the rubric.

You should carefully create this function with reapplying exactly the same transformation which you did above.
This helps in ensuring that the test data passes through exact same transformation as the original data which we want to learn cluster from.

How you did

- Well done on using all the preprocessing work into this function.
- Order of applying the transformation is important and should be same as the used during training time.

Suggestions

- You could have actually followed [DRY Principle](#) by creating functions for the transformation which you used and calling the function here. This will ensure that you have all the transformation done with exact same steps and would ensure non repeating of codes and avoid human error

Feature Transformation




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

Importance of the rubric



What if you don’t feature scale the data?

- Suppose a feature’s variance is orders of magnitude more than the variance of other features. In that case, that feature might dominate other features in the dataset, which is not something we want to happen.
- This skews the PCA towards high magnitude features when fitted on such data.

How you did

Feature scaling has been properly applied to the demographics data. 
Imputation has been performed to remove the remaining missing values 
Valid justifications are provided for handling the missing values and scaling operation, in the report. 

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.

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. 

Importance of the rubric

- You need to first perform principal component analysis and then run the variability analysis to determine the number of components which you want to use and which explains the maximum variability in the data.
- This is a tradeoff which you need to make between selecting the number of features and loosing on the variability

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.

- Student has been able to analyse the distribution of weights in atleast 3 different principal components.
- Its a great idea to write a function that you can call at any time to print the sorted list of feature weights, for the i-th principal component

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.

- Multiple cluster counts have been tested on the general demographics data, and the average point-centroid distances have been reported.
- Well done on selecting the elbow i.e. the turning point.
- Your selected range for k is good given the dataset and you have evenly spread them to get a good plot for selecting k.
- Justification for elbow selection has been provided

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

- Cleaning, feature transformation, dimensionality reduction, and clustering models have been applied
- Student uses the existing sklearn objects for transforming data and doesn't create any new object.
- Objects from general data has been used in demographics data and cluster assignments have been made accordingly.
- It is important that the demographics data goes through exact same transformation as the general data so that we don't see any deviation in cluster assignment

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.

[↓](#) DOWNLOAD PROJECT

RETURN TO PATH