# Customer Segmentation Project for Arvato Financial Services

Capstone project in the Machine Learning Nanodegree

## Project Overview

Customer acquisition process is vital for companies to grow their customer base. Bertelsmann Arvato Analytics challenge us with a real-life data science task to analyse demographics data of a mail-order sales company’s core customers in Germany and the demographics data of the general population to identify customer segments to target with their marketing campaigns. Furthermore, we also build a model to predict the probability of an individual to respond to the marketing campaigns and become a new customer.

## Problem Statement

To fulfil the project’s requirements. There are 4 parts we need to complete

* Data pre-processing: the provided datasets are data collected in production from different sources, it needs to be pre-processed and clean before it can be used for any analysis tasks.
* Customer segmentation: finding the similarities between the general population and the mail-order sales company’s core customers will help use identify the customer segments to target with marketing campaigns. Unsupervised learning models help us achieve this task
* Customer Prediction: after identifying the targeted customers, we are going to use the demographic features and the ground truths provided by Bertelsmann Arvato Analytics to predict whether a person will become a new customer after a mailout campaign or not.
* Kaggle competition: we will enter a Kaggle competition to evaluate our prediction results

## Metrics

## Data Input

4 datasets with demographics features are provided.

* Udacity\_AZDIAS\_052018.csv - demographics data of the German general population; 891 211 (rows) x 366 features.
* Udacity\_CUSTOMERS\_052018.csv - demographics data of the customers; 191 652 (rows) x 369 features.
* Udacity\_MAILOUT\_052018\_TRAIN.csv - demographics data of individuals who were targeted by a marketing campaign; 42 982 (rows) x 367 features. This dataset has a RESPONSE column which indicates if the targeted person has replied and become a new customer or not.
* Udacity\_MALOUT\_052018\_TEST.csv - demographics data of individuals who were targeted by a marketing campaign; 42 833 (rows) x 366 features.

In addition to the above datasets, there are 2 metadata files:

* DIAS Information Levels – Attributes 2017.xlsx - a top level list of attributes and descriptions, organised by informational category
* DIAS attributes – Values 2017.xlsx - a detailed mapping of data values for each feature in alphabetic order

## Data Exploration and Pre-processing

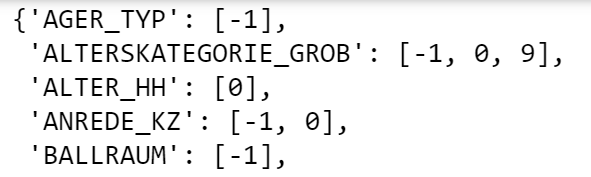
We will first load demographics data in file Udacity\_AZDIAS\_052018.csv into a dataframe for analysis.

### Deal with Missing/Unknown/NaN values

Missing/Unknown/NaN values are marked differently depending on the feature e.g missing and unknown values of ‘ALTERSKATEGORIE\_GROB’ are marked with any of these values [0, -1, 9] and NaN in the dataset.

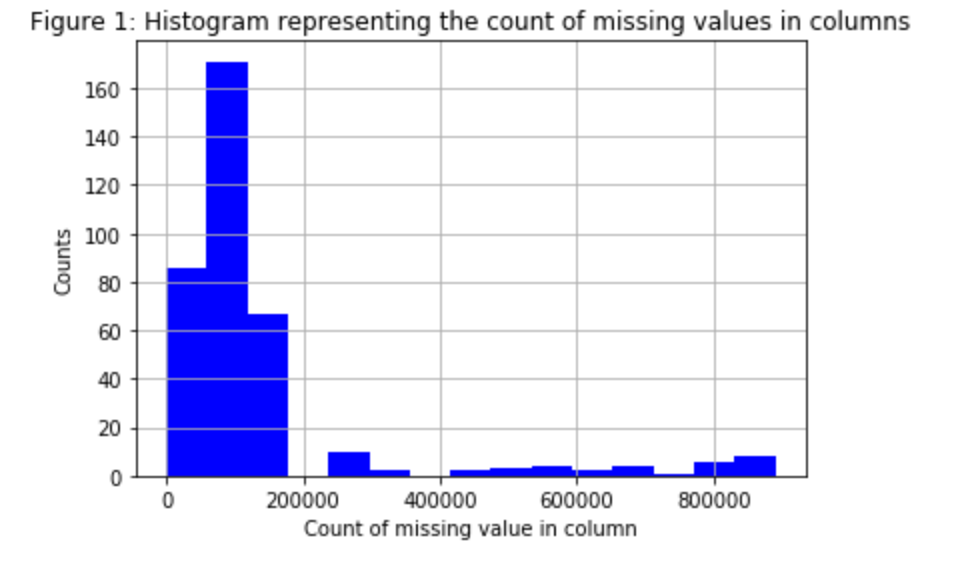
From DIAS attributes – Values 2017.xlsx, we can construct a dictionary of all unknown values of given attributes. Then we are going to re-encode all missing/unknown/NaN values with np.nan in the AZDIAS dataframe.

Below are a few values of the dictionary.



### Assess missing values in each column

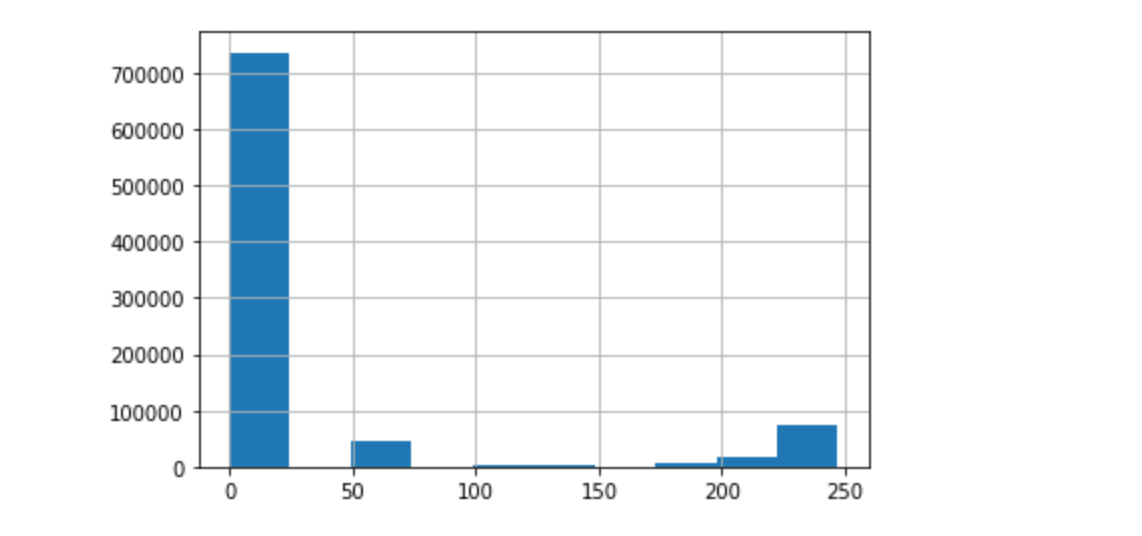
After re-encoding all the missing/unknown values to NaN, we now can assess the NaN value in each column.



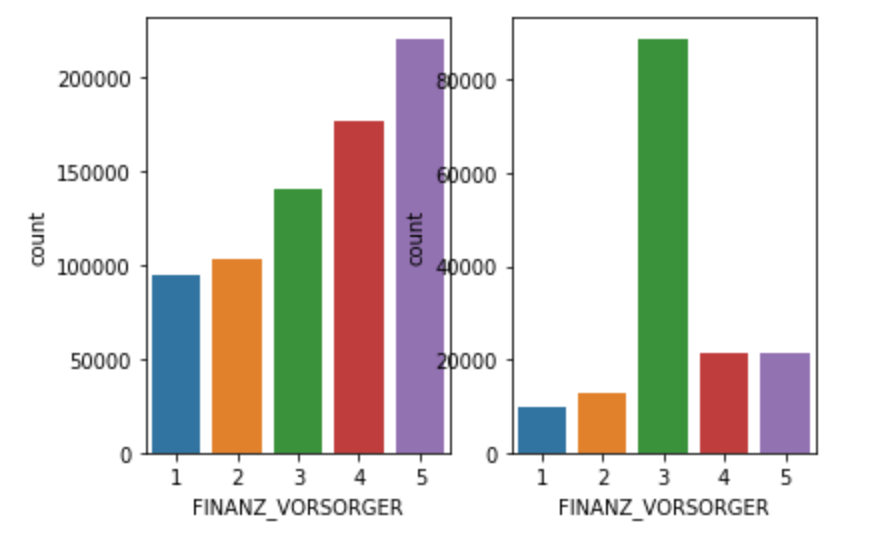
There is a total of 891221 rows in the general population dataset. Out of 366 columns, 302 columns have missing values and 42 columns with more than 200k missing values. We are going to drop columns with more than 200k missing values and deal with the missing values in the remaining columns later

### Assessing missing values in each row

### Now we will perform a similar assessment for missing values in each row.



The chart above shows the number of missing values in each row. We are going to divide the dataset into 2 subsets and investigate the distribution of data values on columns are similar or different in both groups



The data with many missing values looks different from the one with few or no missing values. To continue with our analysis, we are going to use the subset of the data with few or no missing values.

We are going to assess the missing values in each row

* Drop all columns with missing value - No because we will be left with 40 columns out of 366
* Drop all rows with missing values - No because we will lose ~30% of the original dataset
* Drop rows with more than 150 missing values and impute missing values with mode model

### Select & Re-encode features

Since the unsupervised learning technique only works with numeric data, we need to make a few encoding changes for un-numerical data in the dataset.

* For numeric and interval data, these features can be kept with no changes
* Most of the features in the dataset are ordinal in nature. We assume the ordinal features can be treated as interval in nature so they will be kept unchanged
* Special handling will be done for mixed and categorical features.

The feature ‘OST\_WEST\_KZ’ has values of ‘W’ and ‘O’. I re-encoded it to 0 and 1.

A list of multilevel categorical features ['CJT\_GESAMTTYP', 'FINANZTYP', 'GFK\_URLAUBERTYP', 'NATIONALITAET\_KZ', 'SHOPPER\_TYP', 'ZABEOTYP', 'GEBAEUDETYP'] were converted into dummy variables.

Features such as [‘CAMEO\_DEUG\_2015’, ‘CAMEO\_INTL\_2015’ ] has value ‘X’ and ‘XX’ which were converted to np.nan for further analysis

### Assessing features with mixed values

We are going to re-engineer features with mixed values

* CAMEO\_INTL\_2015 is split into WEALTH\_INDEX (1: Wealthy Household, 2: Prosperous Households, 3: Comfortable Households, 4: Less Affluent Households, 5: Poorer Households) and LIFE\_STAGE\_INDEX (1: Pre-Family Couples & Singles, 2: Young Couples With Children, 3: Families With School Age Children, 4: Older Families & Mature Couples, 5: Elders In Retirement)
* PLZ8\_BAUMAX is re-engineer to PLZ8\_BAUMAX\_FAMILY (0: 0 families, 1: mainly 1–2 family homes, 2: mainly 3–5 family homes, 3: mainly 6–10 family homes, 4: mainly 10+ family homesand PLZ8\_BAUMAX\_BUSINESS (0: Not Business, 1: Business)
* PRAEGENDE\_JUGENDJAHRE is split into MOVEMENT (1: Mainstream, 2: Avantgarde) and DECADES (4: 40s, 5: 50s, 6: 60s, 7: 70s, 8: 80s, 9: 90s)

### Drop features

There are features we are going to drop because of various reasons

* LNR -unique identifier
* EINGEFUEGT\_AM - insertion timestamp - not very suitable for this analysis
* GEBURTSJAHR - too many nan/0 values
* CAMEO\_DEU\_2015 - very similar to CAMEO\_DEUG\_2015 -drop for similarity
* 'LP\_LEBENSPHASE\_GROB' & 'LP\_LEBENSPHASE\_FEIN' indicate lifestage and income of a person which is done in 'CAMEO\_INTL\_2015'
* 'CAMEO\_INTL\_2015', 'PLZ8\_BAUMAX', 'PRAEGENDE\_JUGENDJAHRE' are dropped because they have been reengineered

### Fill the NaN values in each row with a random value in the list of existing values

If we are going to drop all rows with one or more missing values, we are going to drop 30% of the dataset so we have decided to impute the remaining rows which have missing values with a random value in the list of existing values of the column in question. This is also our last step in the cleaning process

### Data pre-processing summary

The 2 datasets AZDIAS and CUSTOMERS undergo the same cleaning process which includes

1. Drops columns and rows
2. Reengineer features
3. Transformation using dummy values

Overall, AZDIAS dataset was transformed to 791241 rows and 356 features, CUSTOMERS was transformed to

## Customer Segmentation

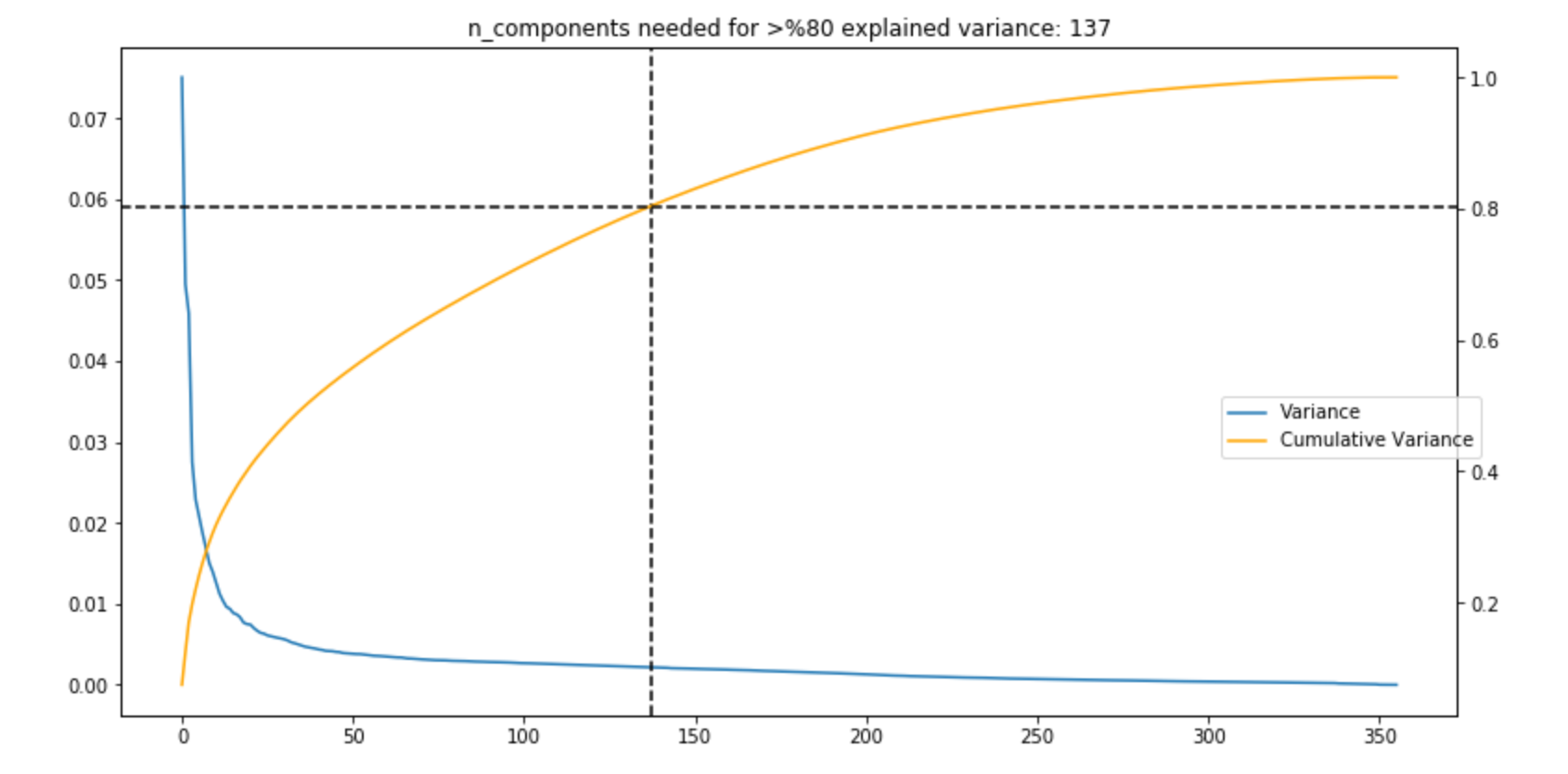
### Feature transformation

Before we apply features reduction techniques to the datasets. We need to perform feature scaling so that the principal component vectors are not influenced by the natural differences in scale for features using a standardscaler

### Perform Dimensionality Reduction

On the scaled AZDIAS dataset, I use Principal Component Analysis (PCA) for dimensionality reduction.

We’ve plotted cumulative explained variance and numbers of principal components to obtain at least 85% explained variance with ~ 137 components.

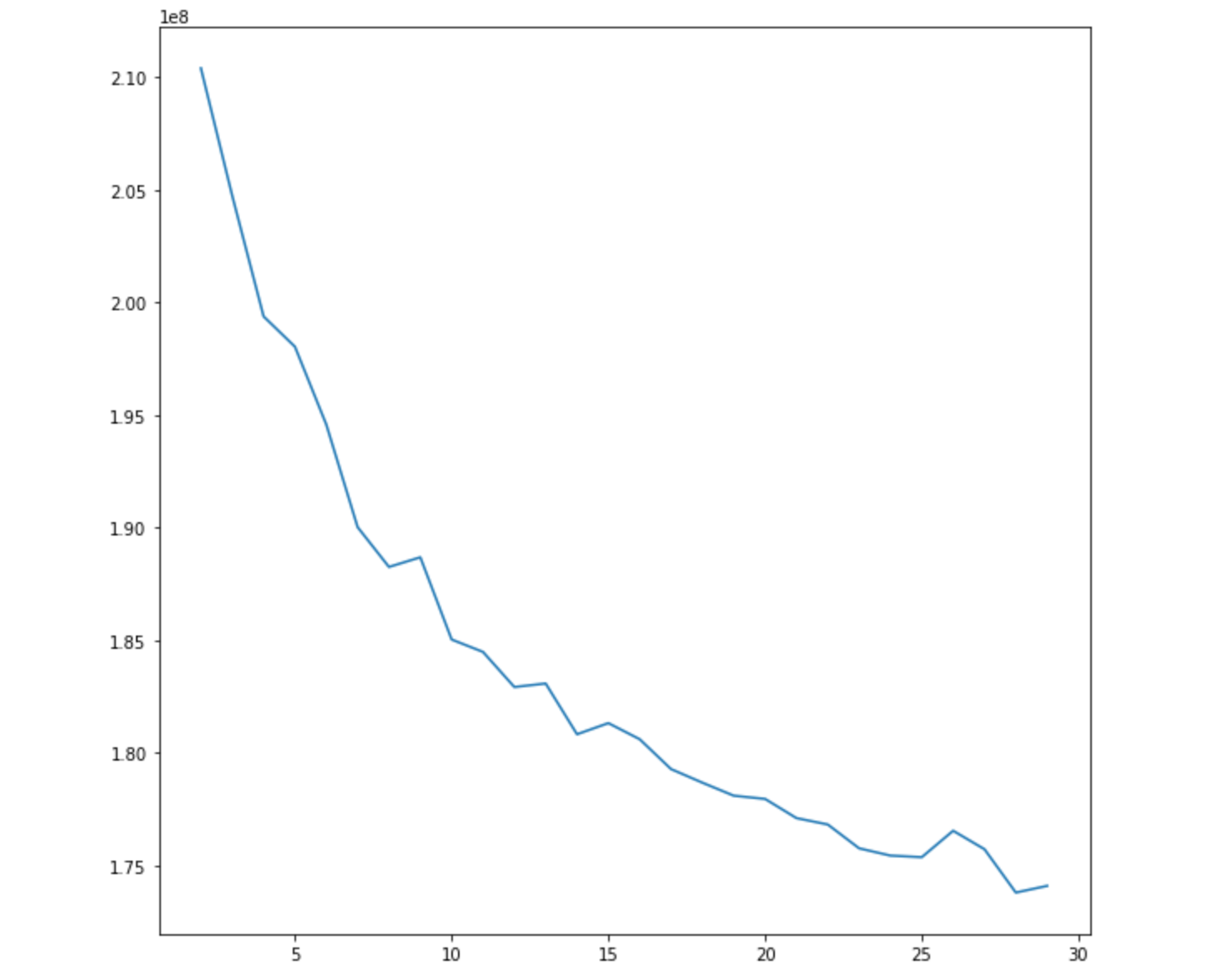


## Unsupervised Learning Model – K-means clustering

### Using the PCA features-reduced dataset, we are going to use unsupervised learning algorithm K-Means to apply clustering in the principal components space. We will apply k-means clustering to the dataset and use the average within-cluster distances from each point to their assigned clusters' centroid to decide on a few clusters to keep

### Choosing a good K

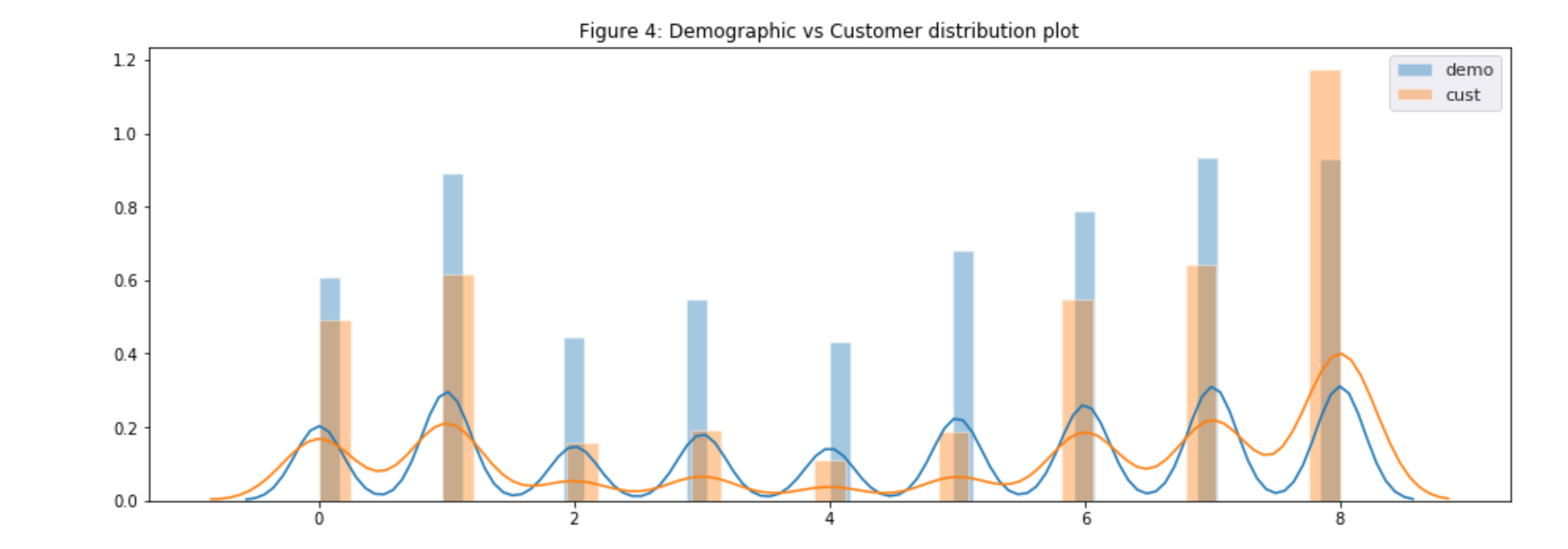
Elbow method was used to identify an ideal number of clusters for k-means clustering on the PCA-transformed data. Average of sum of squared errors within-cluster distances was plotted against number of clusters from 2 to 30. The plot below shows that the score decreased for the first 8 clusters and increased for the 9th clusters and then continued to decrease for higher number of clusters. So 9 was selected as an ideal K number

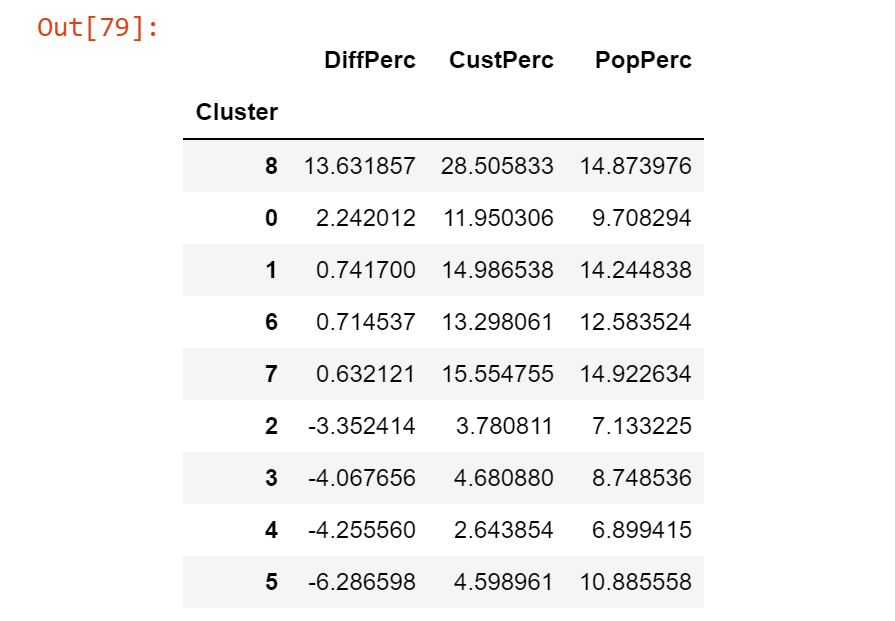


### Apply clustering for AZDIAS and CUSTOMERS datasets

We have clustered data based on demographics of the general population of Germany and seen how the customer data maps onto those demographics clusters. We are going to compare the 2 clusters distributions to see where the strongest customer base for the company is

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 any particular segments of the population that are interested in company's products, then we should see a mismatch. If there is a higher proportion of persons in a cluster for the customer data compared to the general population then that suggests the people in that cluster to be a target audience for the company. On the other hand, the proportion of the data in a cluster being larger in the general population than the customer data suggests that group of persons to be outside of the target demographics





### Compare Customer data to Demographics data

Conclusion: comparing the ratio of cluster 8 is significantly higher in the customer data relative to the demographic/population data.

* Cluster 8: Represents the 30-45 years and above, male grown up in a comfortable household typical an established middle class and owing a business in healthy relationship

## Supervised Learning Model

Now we can start analysing MAILOUT datasets to build a supervised learning model to make predictions. Each of the rows in MAILOUT datasets represents an individual that was targeted for a mailout campaign with ‘RESPONSE’ column. We have removed ‘RESPONSE’ column from TRAIN dataset so that we can add it into the result prediction dataset later. It is noteworthy that only ~1.2% of MAILOUT\_TRAIN dataset are customers. We do the cleaning process without removing any rows of the MAILOUT\_TRAIN dataset.

Generally, to build supervised model we need to split data into training and testing datasets and that’s what we are going to do with TRAIN dataset. For the TEST dataset, we are using the same model to make prediction and submit our predictions to Kaggle competition for validation.

### Classifiers

The below models were used.

* Logistic Regression
* Logistic Regression + GridSearch
* Random Forest
* Random Forest + GridSearch

Because the TRAIN dataset is unbalanced with most of the RESPONSE is 0 (~99%), the following steps are used

* Use GridSearch method for classifier parametetrization. GridSearch tests all possible combination of specified hyperparameters then choose the model with maximum score (roc\_auc) on a validation set.
* Use ‘class\_weight’ parameter
* Use stratify parameter in train\_test\_split process

## Conclusion

In this project, real-life demographics of the general population of Germany and customer segments are analysed.

* The first part of the task is to make cleaning decisions on the provided datasets. This part was challenging because there are ~400 features in the dataset, many of them miss descriptions. There are many features with correlation information as well. I believe this task is the most important one and it can be improved with experience. I think the process to choose and remove correlation features can be done better
* The unsupervised learning model using PCA for dimensionality reduction and K-means clustering to identify targeted segments is straightforward
* The last part uses Logistic Regression Classifier with GridSearch to build a supervised model to make predictions. As the MAILOUT\_TRAIN dataset is heavily unbalanced, the predictions might be improved with a better RESPONSE dataset.

### Reflection

### Improvement