# CREATE A CUSTOMER SEGMENTATION REPORT FOR ARVATO FINANCIAL SERVICES

*Udacity* - MACHINE LEARNING ENGINEER NANODEGREE
CAPSTONE PROJECT

**Radu L. Enuca**
Bucharest, Romania
https://github.com/raduenuca
https://www.linkedin.com/in/raduenuca

August 14, 2019

**ABSTRACT**

In this project, we analyze demographic data for customers of a mail-order sales company [1] in Germany, comparing it against demographics information for the general population. EDA [2] is performed to understand and clean the data. Unsupervised learning techniques are used to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, we'll apply what we've learned on a third dataset with demographic information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company.

***K*eywords** Exploratory Data Analysis · Unsupervised Learning · Supervised Learning

## 1 Definition

### 1.1 Project Overview

In this project, a mail-order sales company in Germany is interested in identifying segments of the general population to target with their marketing, to grow their customer base. Demographics information is available for both the general population as well as for prior customers of the company. We use this information to build a model of the customer base of the company. The target dataset contains demographic information for targets of a mailout marketing campaign. The objective is to identify which individuals are most likely to respond to the campaign and become customers of the mail-order company.

The data has been provided by Bertelsmann Arvato Analytics and consists of four data files:

- *Udacity_AZDIAS_052018.csv*: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).
- *Udacity_CUSTOMERS_052018.csv*: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).
- *Udacity_MAILOUT_052018_TRAIN.csv*: Demographics data for individuals who were targets of a marketing campaign; 42 982 persons (rows) x 367 (columns).
- *Udacity_MAILOUT_052018_TEST.csv*: Demographics data for individuals who were targets of a marketing campaign; 42 833 persons (rows) x 366 (columns).

[1] The data is the property of Bertelsmann Arvato Analytics and represents a real-life data science task.
[2] Exploratory Data Analysis

The "CUSTOMERS" file contains three extra columns ('CUSTOMER_GROUP', 'ONLINE_PURCHASE', and 'PRODUCT_GROUP'), which provide broad information about the customers depicted in the file. Each row of the demographics files represents a single person, but also includes information outside of individuals, including information about their household, building, and neighborhood.

### 1.2 Problem Statement

The goal is to identify segments of the population that form the core customer base for the company. These segments can then be used to direct marketing campaigns towards audiences that have the highest expected rate of returns.

The information from the first two files is used to figure out how customers ("CUSTOMERS") are similar to or differ from the general population at large ("AZDIAS"), then use this analysis to make predictions on the other two files ("MAILOUT"), predicting which recipients are most likely to become a customer for the mail-order company.

The original "MAILOUT" file included one additional column, "RESPONSE," which indicated whether or not each recipient became a customer of the company. For the "TRAIN" subset, this column is present, but in the "TEST" subset it has been removed; it is against that withheld column that we will asses the final predictions. The higher the score obtained, the better the model is at predicting customers.

### 1.3 Metrics

We are dealing with an imbalanced classification problem, and we consider using metrics beyond accuracy such as recall, precision, and AUROC.

The evaluation metric chosen for this project is AUC for the ROC curve (see figure 1), relative to the detection of customers from the mail campaign. A ROC, or receiver operating characteristic, is a graphic used to plot the true positive rate (TPR, the proportion of actual customers that are labeled as so) against the false positive rate (FPR, the proportion of non-customers labeled as customers).

The line plotted on these axes depicts the performance of an algorithm as we sweep across the entire output value range. We start by accepting no individuals as customers (thus giving a 0.0 TPR and FPR) then gradually increase the threshold for accepting customers until all individuals are accepted (thus giving a 1.0 TPR and FPR). The AUC, or area under the curve, summarizes the performance of the model. If a model does not discriminate between classes at all, its curve should be approximately a diagonal line from (0, 0) to (1, 1), earning a score of 0.5. A model that identifies most of the customers first, before starting to make errors, has its curve start with a steep upward slope towards the upper-left corner before making a shallow slope towards the upper-right. The maximum score possible is 1.0 if all customers are perfectly captured by the model first.

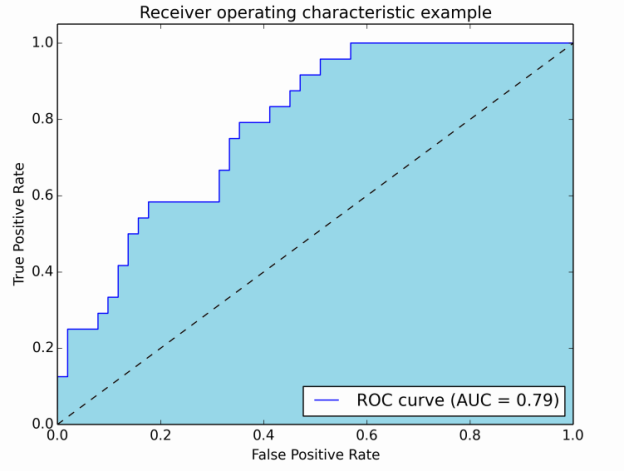



Figure 1: Example of how AUROC looks graphically

# 2 Analysis

## 2.1 Data Exploration

The demographic data for the general population of Germany contains 366 features, and to display it here, we have to transpose it. Below are the first 5 samples from the first 10 features (for all the features, please see table 4 in the ANNEX section):

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| AGER_TYP | -1 | -1 | -1 | 2 | -1 |
| AKT_DAT_KL | NaN | 9.00 | 9.00 | 1.00 | 1.00 |
| ALTER_HH | NaN | 0.00 | 17.00 | 13.00 | 20.00 |
| ALTER_KIND1 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND2 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND3 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND4 | NaN | NaN | NaN | NaN | NaN |

Table 1: Small subsample from the general population dataset

The feature names are not very explanatory, but fortunately for us along with the datasets, we have two other files:

- *DIAS Information Levels - Attributes 2017.xlsx*: Describes each feature.
- *DIAS Attributes - Values 2017.xlsx*: Describes the type of each feature along with possible values along with values that represent missing or unknown information.

As we can see the dataset contains a lot of missing values (represented by NaN in table 1 and 4) and some of the values, like (-1, 0 or 9) also indicate missing or unknown information.

Based on the *DIAS Attributes - Values 2017.xlsx* we've built a new dataset *AZDIAS_Feature_Summary.csv* that contains a summary of properties for each demographics data column, as follows (for the full list please see table 5 in the ANNEX section):

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 0 | AGER_TYP | categorical | [-1,0] | person |
| 1 | ALTERSKATEGORIE_GROB | ordinal | [-1,0,9] | person |
| 2 | ALTER_HH | interval | [0] | household |
| 3 | ANREDE_KZ | categorical | [-1,0] | person |
| 4 | ANZ_HAUSHALTE_AKTIV | numeric | [] | building |
| 5 | ANZ_HH_TITEL | numeric | [] | building |
| 6 | ANZ_PERSONEN | numeric | [] | household |
| 7 | ANZ_TITEL | numeric | [] | household |
| 8 | BALLRAUM | ordinal | [-1] | postcode |
| 9 | CAMEO_DEUG_2015 | categorical | [-1,X] | microcell_rr4 |

Table 2: Feature summary subsample

We use this file to help us make cleaning decisions for the project.

**Missing Values** The third column (*missing_or_unknown* of the feature attributes summary, documents the codes from the data dictionary that indicate missing or unknown data. Before converting data that matches a 'missing' or 'unknown' value code into a NaN value, we first have a look how much data takes on a 'missing' or 'unknown' code, and how much data is naturally missing, as a point of interest.

**Select and Re-Encode Features** Checking for missing data isn't the only way in which we can prepare a dataset for analysis. Since the unsupervised learning techniques we use only work on data that is encoded numerically, we need to

make a few encoding changes or additional assumptions to be able to make progress. While almost all of the values in the dataset are encoded using numbers, not all of them represent numeric values.

- For numeric and interval data, these features can be kept without changes.
- Most of the variables in the dataset are ordinal. While ordinal values may technically be non-linear in spacing, we make the simplifying assumption that the ordinal variables can be treated as being an interval in nature (that is, kept without any changes).
- Special handling may be necessary for the remaining two variable types: categorical, and 'mixed.'

**Categorical Features** For categorical features, we encode the levels as dummy variables. Depending on the number of categories, we can perform one of the following:

- For binary (two-level) categorical features that take numeric values, we can keep them without needing to do anything.
- If there are binary variables that take on non-numeric values, we need to re-encode the values as numbers or create a dummy variable.
- For multi-level categorical features (three or more values), we can choose to encode the values using multiple dummy variables (e.g., via *OneHotEncoder*)

**Mixed-Type Features** There are a handful of features that are marked as "mixed" in the feature summary that require special treatment before we can include then in the analysis. There are two in particular that deserve attention:

- *PRAEGENDE_JUGENDJAHRE* combines information on three dimensions: generation by decade, movement (mainstream vs. avantgarde), and nation (east vs. west). While there aren't enough levels to disentangle east from west, we create two new variables to capture the other two dimensions: an interval-type variable for the decade, and a binary variable for movement.
- *CAMEO_INTL_2015* combines information on two axes: wealth and life stage. We break up the two-digit codes by their 'tens'-place and 'ones'-place digits into two new ordinal variables (which, for this project, is equivalent to just treating them as their raw numeric values).

**Ordinal and Interval Features** Nothing special here, we need to decide what to do with the missing values

**Numerical Features** Nothing special here, we need to decide what to do with the missing values and perhaps perform some scaling.

### 2.2 Exploratory Visualization

Figure 2a, obtained using the *missingno* package shows the missing values of the provided data and figure 2b same information but after we replaced the unknown or missing value codes with NaN.

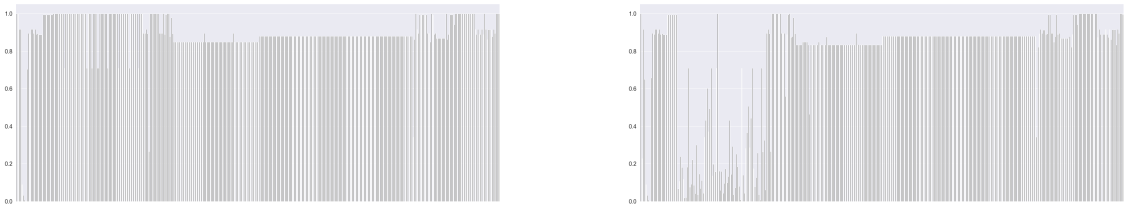
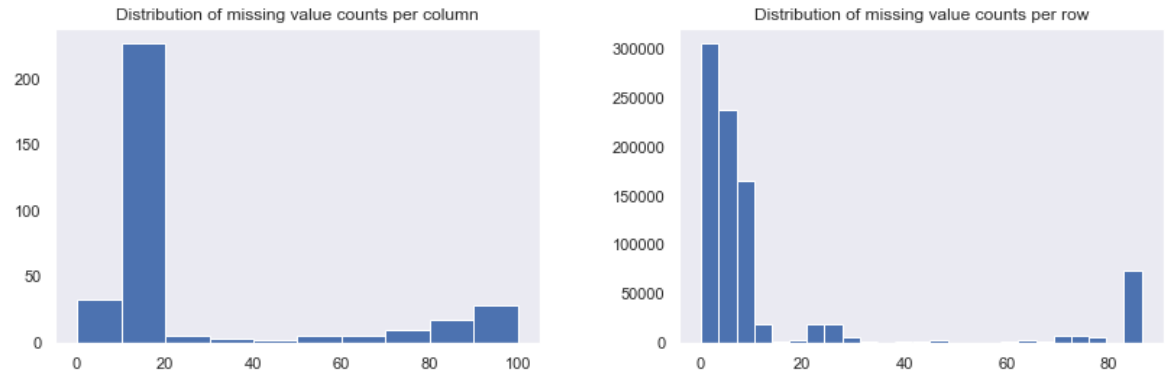


Figure 2: Missing values per column before and after replacing unknown value codes

In figure 3a we have the distribution of missing value counts where we can see that that there are a few columns that are outliers in terms of the proportion of values that are missing. We also perform a similar assessment for the rows of the dataset to asses how much data is missing in each row (see figure 3b). As with the columns, we see some groups of points that have a very different number of missing values.




(a) Distribution of missing value counts per column

(b) Distribution of missing value counts per rows

Figure 3: Distribution of missing value counts per column and row

### 2.3 Algorithms and Techniques

There are 2 parts of this project:

**Unsupervised modeling** The main bulk of our analysis is in this part of the project. Here, we use unsupervised learning techniques to describe the relationship between the demographics of the company's existing customers and the general population of Germany. By the end of this part, we should be able to describe parts of the general population that are more likely to be part of the mail-order company's primary customer base, and which parts of the general population are less so.

We first apply PCA (Principal Component Analysis) [Bui+13b] to reduce the dimensionality of the dataset [Shl05]. We decide on a number of components that explain at least 90% of the variance in data.

After PCA, we create KMeans[Bui+13a] models with clusters from 2 to 15. Clustering is a method of unsupervised learning, where each data point or cluster is grouped into a subset or a cluster, which contains similar kind of data points. We decide on the best number of clusters to take based on the Elbow[PDN04] method.

**Supervised modeling** To predict the probability of a person to reply to the mailing campaign, we create an XGBoost-Classifier model which we use to predict this probability.

Before training the model, we decide on a resampling technique for the data, and after we start by searching the best hyperparameters for models using all available features by using a Bayesian search.

Once we have a list of optimized hyperparameters, we use them for training a model on the resampled data. After training the model is used to predict on the TEST dataset.

### 2.4 Benchmark

Model evaluation is the process of objectively measuring how well machine learning models perform the specific tasks they were designed to do.

We use the AUC scores to benchmark the performance of the models. A model with the highest AUC is considered as the best performer. We train a LogisticRegression model with default parameters and use it to predict the probabilities for the train and test datasets. The obtained AUC score is our base score used for comparison.

## 3 Methodology

### 3.1 Data Preprocessing

Using the analysis and data exploration above, we've built the following preprocessing pipeline (the same cleaning process will also be applied on the training and testing datasets):

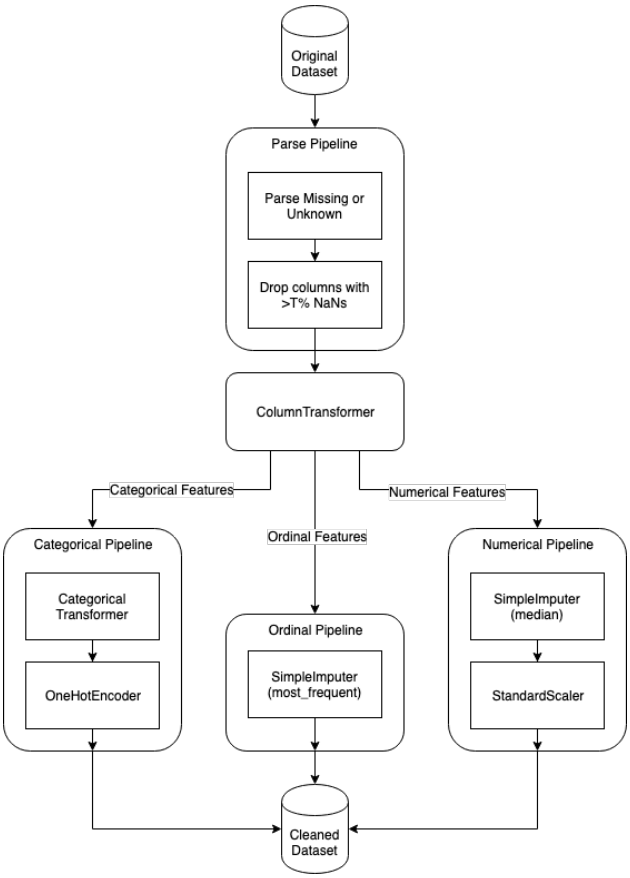



Figure 4: Preprocessing Pipeline

The preprocessing pipeline has the following components:

**Parse pipeline** composed of two steps:

- *Parse Missing or Unknown* - custom pipeline that uses the feature summary constructed dataset (column missing_or _unknown) to recode values as NaNs.
- *Drop columns with T% NaNs* - Custom pipeline that drops all features that have more than T% (75% in our case) missing values

**Column Transform Pipeline** uses the *ColumnTransform* pipeline from sci-kit learn to combine the following pipelines:

- **Categorical Pipeline** applies the following steps to all categorical features:
  - *Categorical Transformer* - Custom pipeline transformer (see the paragraph: "Categorical Transformer: below for more details).
  - *OneHotEncoder* - Encode categorical integer features as a one-hot numeric array.
- **Oridnal Pipeline** applies the following steps to all ordinal features:
  - *SimpleImputer* - Imputation transformer for completing missing values by using the most frequent value along each column.
- **Numerical Pipeline** applies the following steps to all numeric features:
  - *SimpleImputer* - Imputation transformer for completing missing values by using the mean along each column.
  - *StandardScaler* - Standardize features by removing the mean and scaling to unit variance

**Categorical Transformer** Is a custom pipeline transformer that processes the categorical and mixed-type features by applying the following steps:

- Fill in the missing values using the most frequent value along each column
- Re-engineer the *PRAEGENDE_JUGENDJAHRE* mixed-type feature into two additional categorical features *DECADE* and *MOVEMENT* and then drop the original column
- Re-engineer the *CAMEO_INTL_2015* mixed-type feature into two additional categorical features *WEALTH* and *LIFE_STAGE* and then drop the original column

In figure 5, we can see the distribution of feature types before and after dropping features with more than 75% missing values.

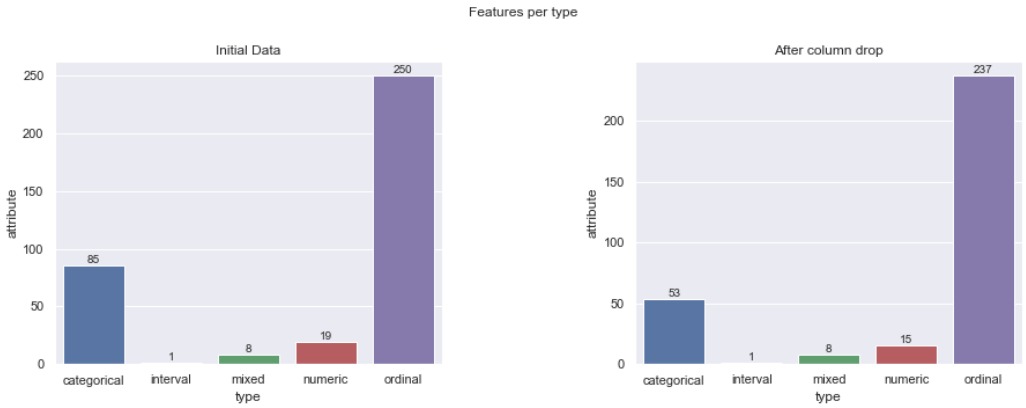



Figure 5: Feature per type

We also perform an analysis for rows with a lot of missing values. As we can see in Figure 3b, there are some rows with more than 85% missing values per row.

To know what to do with the outlier rows, we look the distribution of data values in columns that are not missing data (or are missing very little data) are similar or different between the two groups.

As we can see in Figure 13 in Section 6, the data with many missing values looks very different from the data with few or no missing values. We decide not to remove these rows.

### 3.2 Implementation

#### 3.2.1 Perform Dimensionality Reduction

On our preprocessed data, we are now ready to apply dimensionality reduction techniques.

We use sklearn's PCA class to apply principal component analysis on the data, thus finding the vectors of maximal variance in the data.

We start by fitting a PCA on 685 dimensions (our initial dataset has 366 dimensions but increases to 685 after one-hot encoding the categorical features). You can find the results for the PCA in figure 6 below.

We check out the ratio of variance explained by each principal component as well as the cumulative variance explained.

Based on the results from the PCA fitted previously, we decide to keep the first 150 reduced dimensions, that explain 90% cumulative variance in data.

Now that we have our transformed principal components, we check out the weight of each variable on the first few components to see if we can interpret them some fashion.

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. If two features have large weights of the same sign (both positive or both negative), then increases in one tend to expect to be associated with increases in the other. To contrast, features with different signs can be expected to show a negative correlation: increases in one variable should result in a decrease in the other.

To investigate the features, we map each weight to their corresponding feature name, then sort the features according to weight. The most interesting features for each principal component, are those at the beginning and end of the sorted list.

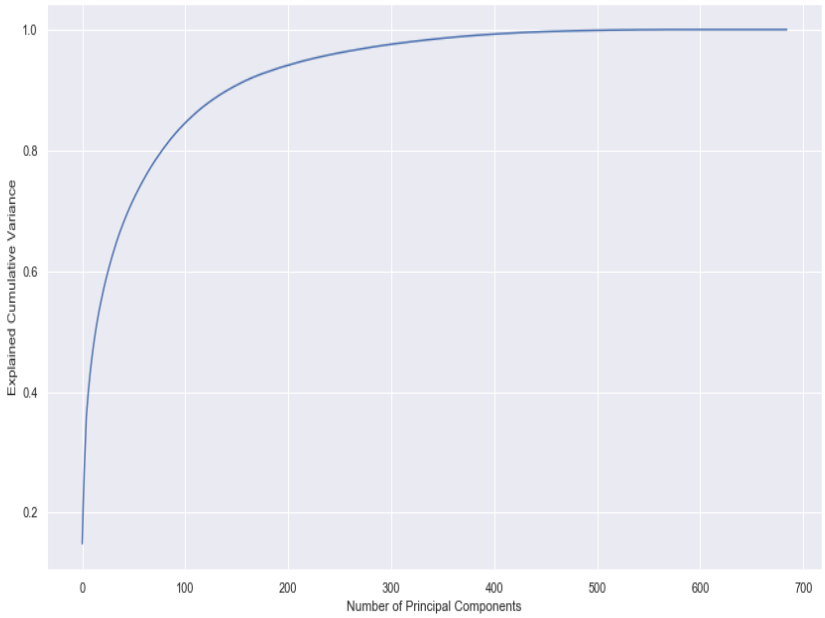



Figure 6: Principal Component Analysis

We present here the investigation of feature associations from the first three principal components:

**Top 5 weights for PC1**

| | |
|---|---|
| D19_GESAMT_ONLINE_QUOTE_12 | 0.4903 |
| D19_VERSAND_ONLINE_QUOTE_12 | 0.4737 |
| ONLINE_AFFINITAET | 0.1393 |
| D19_BANKEN_ONLINE_QUOTE_12 | 0.0913 |
| D19_GESAMT_ANZ_12 | 0.0885 |

First component is all about online affinity and online transactions in the last 12 months

**Top 5 weights for PC2**

| | |
|---|---|
| ALTER_HH | 0.3310 |
| SEMIO_REL | 0.2465 |
| SEMIO_PFLICHT | 0.2112 |
| FINANZ_SPARER | 0.2055 |
| ORTSGR_KLS9 | 0.1711 |

Second component describes the number and age of inhabitants, affinity to religion, being traditional minded and money saver financial topology

**Top 5 weights for PC3**

| | |
|---|---|
| ORTSGR_KLS9 | 0.3036 |
| EWDICHTE | 0.2092 |
| SEMIO_ERL | 0.1452 |
| FINANZ_HAUSBAUER | 0.1156 |
| SEMIO_LUST | 0.1153 |

Third component describes the number and density per square kilometer of inhabitants, affinity to events and being sensual minded, as well as having the house as the main financial focus

Next, we see how the data clusters in the principal components space. We apply k-means clustering to the dataset and use the average within-cluster distance to decide the number of clusters to keep. We use sklearn's KMeans class to perform k-means clustering on the PCA-transformed data.

We fit a KMeans model on the 150 reduced dimensions, and we investigate the change within-cluster distance across for a range of clusters between 2 and 16.

Based on Figure 7, we can see that a good number of clusters is 8. We refit the k-means model with the selected number of clusters and obtain cluster predictions for the general population demographics data.


Figure 7: Average within-cluster distances

We compare the proportion of data in each cluster for the customer data to the proportion of data in each cluster for the general population:

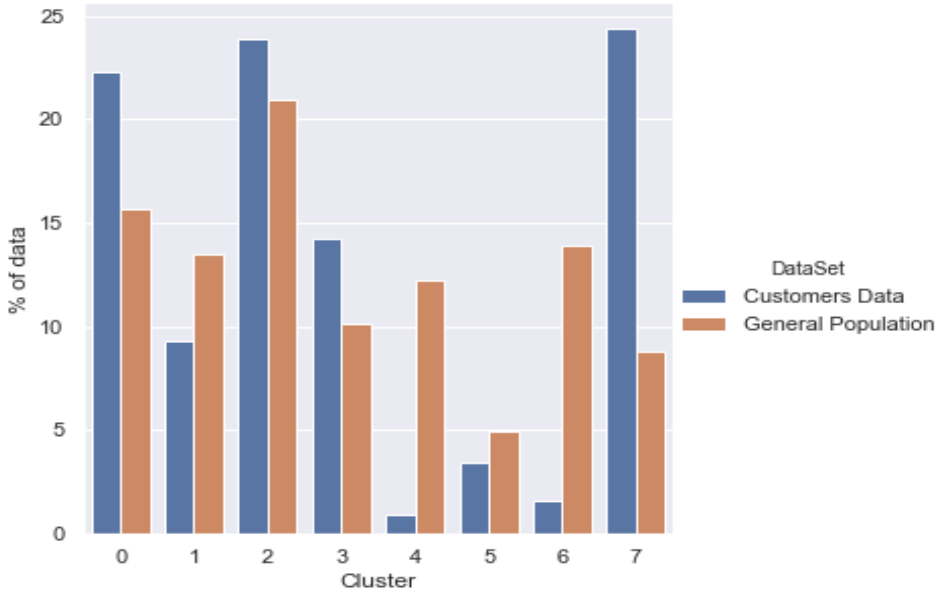



Figure 8: Proportion of data points in each cluster for the general population and the customer data.

We inspect what kind of people are part of a cluster that is over-represented in the customer data compared to the general population (cluster 7)

**Popular with the company - Cluster 7** By using PCA's inverse transform we obtain the following values

- SEMIO_VERT (1) # affinity indicating in what way the person is dreamily (highest affinity)

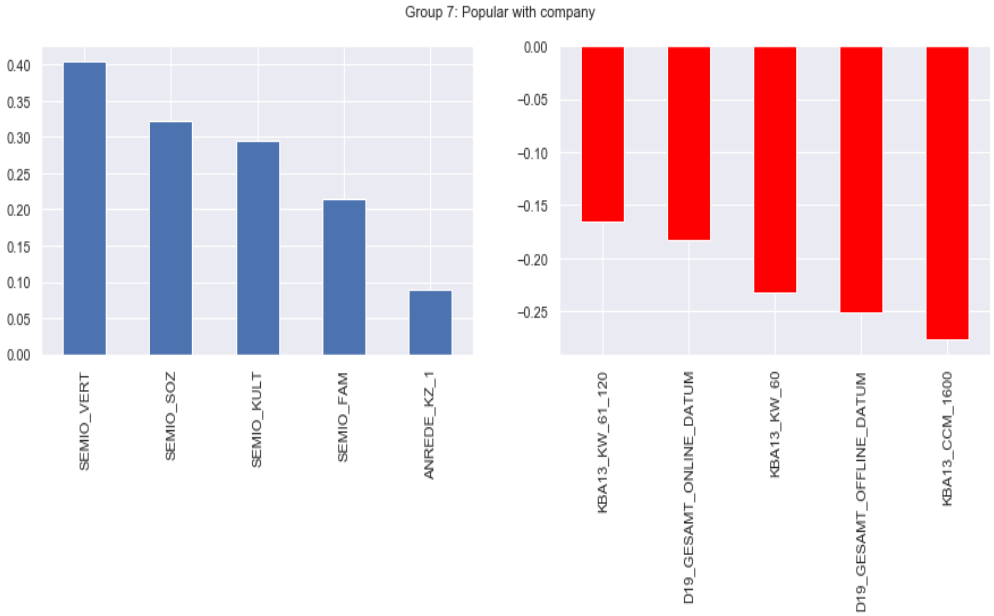



Figure 9: Cluster 7 - People popular with the company

- SEMIO_SOZ (2) # affinity indicating in what way the person is social-minded (very high affinity)
- SEMIO_KULT (3) # affinity indicating in what way the person is cultural minded (high affinity)
- SEMIO_FAM (6) # affinity indicating in what way the person is familiar minded (very low affinity)
- ANREDE_KZ_1 (0) # gender: female

We inspect what kind of people are part of a cluster with low representation in the customer data compared to the general population (cluster 4)

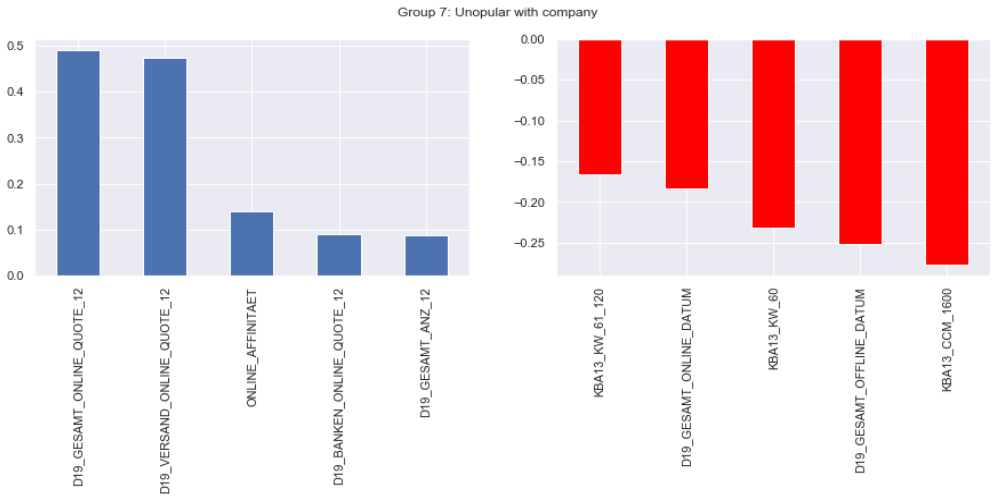



Figure 10: Cluster 4 - People unpopular with the company

**Unpopular with the company - Cluster 4** By using PCA's inverse transform we obtain the following values

- D19_GESAMT_ONLINE_QUOTE_12 (0) # amount of online transactions within all transactions in the complete file: no Online-transactions within the last 12 months

- D19_VERSAND_ONLINE_QUOTE_12 (0) # amount of online transactions within all transactions in the segment mail-order: no Online-transactions within the last 12 months
- ONLINE_AFFINITAET (2) # online affinity: (average affinity)
- D19_BANKEN_ONLINE_QUOTE_12 (0) # amount of online transactions within all transactions in the segment bank: no Online-transactions within the last 12 months
- D19_GESAMT_ANZ_12 (1) # transaction activity TOTAL POOL in the last 12 months: very low activity

### 3.2.2 Supervised Learning Model

To implement the supervised model, we start by preprocessing the training dataset using the same pipeline as above.

We use the unsupervised model fitted in the previous step to predict the cluster in which the observations are present.

We start by analysing the training dataset and especially the distribution between the two typs of responses: 0-non customer and 1-customer:

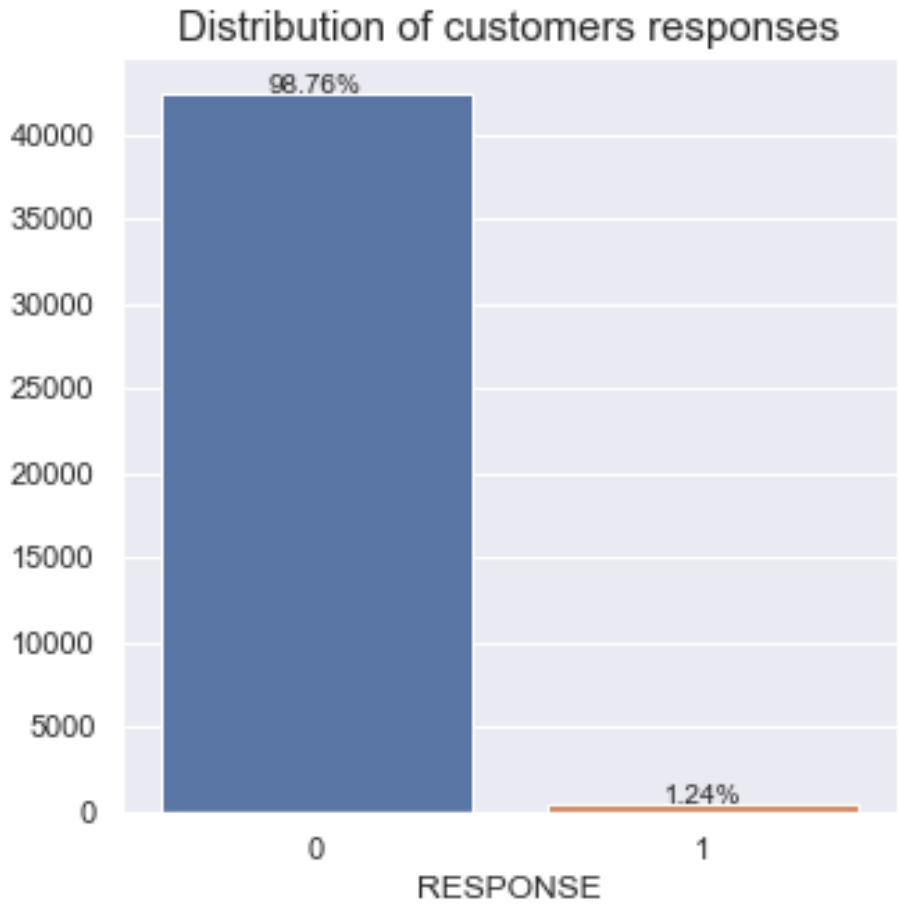



Figure 11: Distribution of customer responses shows that we are dealing with an imbalanced dataset

There are several techniques to deal with an imbalanced dataset:

- **Use the right evaluation metrics** - Evaluation metrics that can be applied in this case:
  - Precision/Specificity: how many selected instances are relevant
  - Recall/Sensitivity: how many relevant instances are selected
  - F1 score: harmonic mean of precision and recall

  - MCC: correlation coefficient between the observed and predicted binary classifications
  - AUC: the relation between true-positive rate and false positive rate (**This is what we use in this project**)
- **Re-sample the training set** - Apart from using different evaluation criteria, one can also work on getting a different dataset. Two approaches to making a balanced dataset out of an imbalanced one are under-sampling and over-sampling.
  - Under-sampling balances the dataset by reducing the size of the abundant class. This method is used when the quantity of data is sufficient. By keeping all samples in the rare class and randomly selecting an equal number of samples in the abundant class, a balanced new dataset can be retrieved for further modeling.
  - Oversampling is used when the quantity of data is insufficient. It tries to balance dataset by increasing the size of rare samples. Rather than getting rid of abundant samples, new rare samples are generated by using, e.g., repetition, bootstrapping, or SMOTE (Synthetic Minority Over-Sampling Technique) [Cha+02].

We split the dataset in train and test datasets by specifying that the two datasets are to be stratified using the target and keep the same weight for the classes. The proportion of data after splitting is 80% for training and 20% for validation.

We create a benchmark model using LogisticRegression with the default parameters, and then we fine-tune an XGBoostClasifier model using Bayesian optimization [Koe18].

### 3.3 Refinement

We use several resampling techniques:

- No resampling
- SMOTE - Synthetic Minority Over-sampling Technique
- ADASYN - Adaptive Synthetic (ADASYN) sampling approach for imbalanced datasets
- ClusterCentroids - Method that under samples the majority class by replacing a cluster of majority samples by the cluster centroid of a KMeans algorithm.
- TomekLinks - Under-sampling by removing Tomek's links.
- SMOTETomek - Over-sampling using SMOTE and cleaning using Tomek links.

We also define a search hyperspace that we use to do a Bayesian parameter search. We define a search space for the following hyper-parameters for XGBoost:

- 'n_estimators': hp.quniform('n_estimators', 100, 1000, 1),
- 'eta': hp.quniform('eta', 0.025, 0.5, 0.025),
- 'max_depth': hp.choice('max_depth', np.arange(1, 14, dtype=int)),
- 'min_child_weight': hp.quniform('min_child_weight', 1, 6, 1),
- 'subsample': hp.quniform('subsample', 0.5, 1, 0.05),
- 'gamma': hp.quniform('gamma', 0.5, 1, 0.05),
- 'colsample_bytree': hp.quniform('colsample_bytree', 0.5, 1, 0.05),
- 'eval_metric': 'auc',
- 'objective': 'binary:logistic',
- 'nthread': 4,
- 'booster': 'gbtree',
- 'tree_method': 'gpu_hist',
- 'silent': 1,
- 'seed': random_state

# 4 Results

## 4.1 Model Evaluation and Validation

We obtain the following results (in bold the best results for each case):

| Resampling/Algorithm | Logistic Regression | Voting Classifier | Tuned XGBoostClassifier |
|---|---|---|---|
| No resampling | 0.5 | 0.5 | N/A |
| SMOTE | 0.6564 | 0.9357 | N/A |
| ADASYN | 0.6584 | **0.9365** | N/A |
| ClusterCentroids | **0.6955** | 0.8785 | N/A |
| TomekLinks | 0.5 | 0.5 | N/A |
| SMOTETomek | 0.6564 | 0.9357 | **0.9467** |

Table 3: Predictions for the validation set

The VotingClassifier uses the following classifiers with default parameters:

- SVC - C-Support Vector Classification
- MLPClassifier - Multi-layer Perceptron classifier.
- KNeighborsClassifier - Classifier implementing the k-nearest neighbors vote.
- RandomForestClassifier - A random forest classifier

The tuned XGBoostClassifier has the following parameters after the Bayesian search:

| | |
|---|---|
| colsample_bytree | 0.75 |
| eta | 0.25 |
| gamma | 0.5 |
| max_depth | 8 |
| min_child_weight | 1.0 |
| n_estimators | 673.0 |
| subsample | 0.8 |

## 4.2 Justification

The final tuned XGBoostClassifier performs better not only on the validation set but also on the test set held for the Kaggle competition. Although the difference between the final model and the VotingClassifier with default parameters is small on the validation dataset, the VotingClassifier gets a result under 0.5 (or worse than random guessing) on the Kaggle competition.

# 5 Conclusion

## 5.1 Reflection

For a direct marketing campaign, it is essential to correctly identify the customers who will respond to a particular campaign.

In this project, we analyzed demographic data for customers of a mail-order sales company in Germany, comparing it against demographics information for the general population. Exploratory Data Analysis was performed to understand and clean the data. Unsupervised learning techniques were used to perform customer segmentation, identifying the parts of the population that best describe the core customer base of the company. Then, we applied what we've learned on a third dataset with demographic information for targets of a marketing campaign for the company, and use a model to predict which individuals are most likely to convert into becoming customers for the company (all of this is reflected in Figure 12).

This project was an excellent opportunity to apply and learn new techniques primarily related to imbalanced data problems. Also going beyond simple grid search for hyper-parameter tuning was both a new tool to learn and also a time saver.

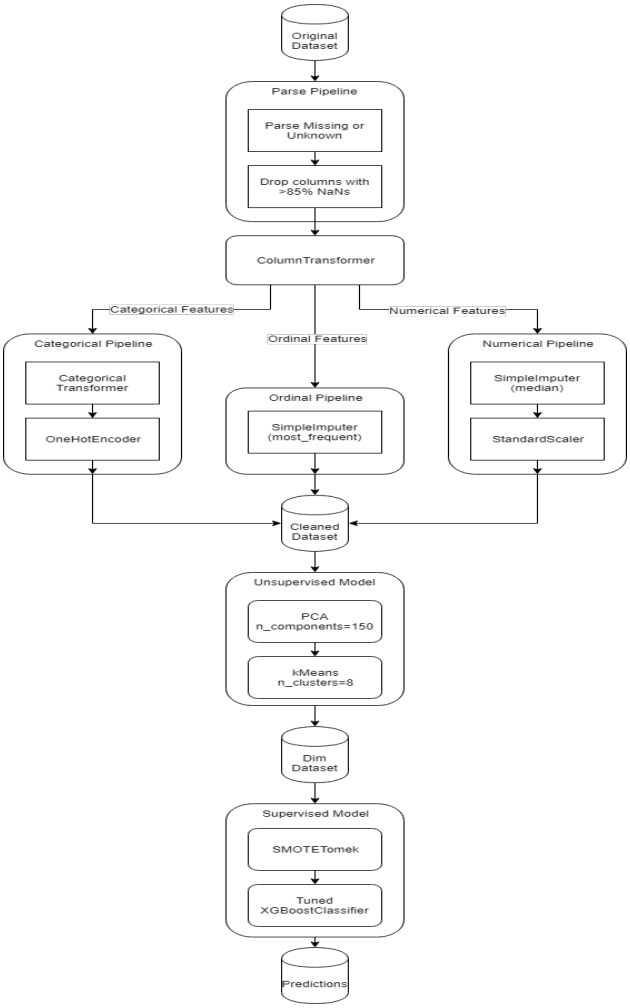



Figure 12: Entire project pipeline

### 5.2 Improvement

Reflecting on the steps taken in this project, we can identify some areas where improvements can be made:

- Data preprocessing - Engineer more categorical features: We believe that better results can be obtained if more categorical features are treated like mixed-type features and re-engineered.
- Data preprocessing - Missing Data:
  - Analyze if there is data missing at random or there are patterns.
  - Try to find correlations between missing values and use PCA to remove some of them
  - Use a supervised model to predict the values for NaN instead of just filling with the median or mode.
- Dimensionality Reduction - Use FAMD (Factor Analysis of Mixed Data) instead of applying PCA on both numerical and Categorical features

## References

Buitinck, Lars et al. *sklearn.cluster.KMeans*. 2013. URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.
– *sklearn.decomposition.PCA*. 2013. URL: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html.
Chawla, Nitesh V. et al. "SMOTE: Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (June 2002). https://arxiv.org/pdf/1106.1813.pdf, pp. 321–357.
Koehrsen, Will. *Automated Machine Learning Hyperparameter Tuning in Python*. July 2018. URL: https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a.
Pham, D T, S S Dimov, and C D Nguyen. "Selection of K in K-means clustering". In: *Columbia Engineering - Eletrical Engineering* (May 2004). https://www.ee.columbia.edu/ dpwe/papers/PhamDN05-kmeans.pdf, pp. 1–17.
Shlens, Jonathon. "A Tutorial on Principal Component Analysis". In: *Carnegie Mello University - School of Computer Science* (Dec. 2005). https://www.cs.cmu.edu/ elaw/papers/pca.pdf, pp. 1–13.

## 6 ANNEX

### 6.1 First 5 rows from the AZDIAS dataset

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| AGER_TYP | -1 | -1 | -1 | 2 | -1 |
| AKT_DAT_KL | NaN | 9.00 | 9.00 | 1.00 | 1.00 |
| ALTER_HH | NaN | 0.00 | 17.00 | 13.00 | 20.00 |
| ALTER_KIND1 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND2 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND3 | NaN | NaN | NaN | NaN | NaN |
| ALTER_KIND4 | NaN | NaN | NaN | NaN | NaN |
| ALTERSKATEGORIE_FEIN | NaN | 21.00 | 17.00 | 13.00 | 14.00 |
| ANZ_HAUSHALTE_AKTIV | NaN | 11.00 | 10.00 | 1.00 | 3.00 |
| ANZ_HH_TITEL | NaN | 0.00 | 0.00 | 0.00 | 0.00 |
| ANZ_KINDER | NaN | 0.00 | 0.00 | 0.00 | 0.00 |
| ANZ_PERSONEN | NaN | 2.00 | 1.00 | 0.00 | 4.00 |
| ANZ_STATISTISCHE_HAUSHALTE | NaN | 12.00 | 7.00 | 2.00 | 3.00 |
| ANZ_TITEL | NaN | 0.00 | 0.00 | 0.00 | 0.00 |
| ARBEIT | NaN | 3.00 | 3.00 | 2.00 | 4.00 |
| BALLRAUM | NaN | 6.00 | 2.00 | 4.00 | 2.00 |
| CAMEO_DEU_2015 | NaN | 8A | 4C | 2A | 6B |
| CAMEO_DEUG_2015 | NaN | 8 | 4 | 2 | 6 |
| CAMEO_INTL_2015 | NaN | 51 | 24 | 12 | 43 |
| CJT_GESAMTTYP | 2.00 | 5.00 | 3.00 | 2.00 | 5.00 |
| CJT_KATALOGNUTZER | 5.00 | 1.00 | 2.00 | 3.00 | 3.00 |
| CJT_TYP_1 | 1.00 | 5.00 | 4.00 | 2.00 | 3.00 |
| CJT_TYP_2 | 1.00 | 5.00 | 4.00 | 2.00 | 3.00 |
| CJT_TYP_3 | 5.00 | 2.00 | 1.00 | 4.00 | 3.00 |
| CJT_TYP_4 | 5.00 | 3.00 | 3.00 | 4.00 | 4.00 |
| CJT_TYP_5 | 5.00 | 1.00 | 2.00 | 5.00 | 3.00 |
| CJT_TYP_6 | 5.00 | 1.00 | 2.00 | 3.00 | 3.00 |
| D19_BANKEN_ANZ_12 | 0 | 0 | 0 | 0 | 3 |
| D19_BANKEN_ANZ_24 | 0 | 0 | 0 | 0 | 5 |
| D19_BANKEN_DATUM | 10 | 10 | 10 | 10 | 5 |
| D19_BANKEN_DIREKT | 0 | 0 | 0 | 0 | 1 |
| D19_BANKEN_GROSS | 0 | 0 | 0 | 0 | 2 |
| D19_BANKEN_LOKAL | 0 | 0 | 0 | 0 | 0 |
| D19_BANKEN_OFFLINE_DATUM | 10 | 10 | 10 | 10 | 10 |
| D19_BANKEN_ONLINE_DATUM | 10 | 10 | 10 | 10 | 5 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| D19_BANKEN_ONLINE_QUOTE_12 | NaN | NaN | 0.00 | 0.00 | 10.00 |
| D19_BANKEN_REST | 0 | 0 | 0 | 0 | 6 |
| D19_BEKLEIDUNG_GEH | 0 | 0 | 0 | 0 | 6 |
| D19_BEKLEIDUNG_REST | 0 | 0 | 0 | 0 | 1 |
| D19_BILDUNG | 0 | 0 | 6 | 0 | 6 |
| D19_BIO_OEKO | 0 | 0 | 0 | 0 | 0 |
| D19_BUCH_CD | 0 | 0 | 0 | 6 | 6 |
| D19_DIGIT_SERV | 0 | 0 | 0 | 0 | 0 |
| D19_DROGERIEARTIKEL | 0 | 0 | 0 | 0 | 1 |
| D19_ENERGIE | 0 | 0 | 0 | 0 | 5 |
| D19_FREIZEIT | 0 | 0 | 0 | 0 | 0 |
| D19_GARTEN | 0 | 0 | 0 | 0 | 0 |
| D19_GESAMT_ANZ_12 | 0 | 0 | 0 | 0 | 6 |
| D19_GESAMT_ANZ_24 | 0 | 0 | 0 | 0 | 6 |
| D19_GESAMT_DATUM | 10 | 10 | 10 | 10 | 1 |
| D19_GESAMT_OFFLINE_DATUM | 10 | 10 | 10 | 10 | 6 |
| D19_GESAMT_ONLINE_DATUM | 10 | 10 | 10 | 10 | 1 |
| D19_GESAMT_ONLINE_QUOTE_12 | NaN | NaN | 0.00 | 0.00 | 10.00 |
| D19_HANDWERK | 0 | 0 | 0 | 0 | 0 |
| D19_HAUS_DEKO | 0 | 0 | 0 | 0 | 5 |
| D19_KINDERARTIKEL | 0 | 0 | 0 | 0 | 0 |
| D19_KONSUMTYP | NaN | NaN | 9.00 | 9.00 | 1.00 |
| D19_KONSUMTYP_MAX | 9 | 9 | 8 | 8 | 1 |
| D19_KOSMETIK | 0 | 0 | 6 | 0 | 0 |
| D19_LEBENSMITTEL | 0 | 0 | 0 | 0 | 0 |
| D19_LOTTO | NaN | NaN | 0.00 | 0.00 | 0.00 |
| D19_NAHRUNGSERGAENZUNG | 0 | 0 | 0 | 0 | 0 |
| D19_RATGEBER | 0 | 0 | 0 | 0 | 0 |
| D19_REISEN | 0 | 0 | 0 | 6 | 0 |
| D19_SAMMELARTIKEL | 0 | 0 | 0 | 6 | 0 |
| D19_SCHUHE | 0 | 0 | 0 | 0 | 1 |
| D19_SONSTIGE | 0 | 0 | 6 | 6 | 4 |
| D19_SOZIALES | NaN | NaN | 0.00 | 0.00 | 0.00 |
| D19_TECHNIK | 0 | 0 | 6 | 6 | 5 |
| D19_TELKO_ANZ_12 | 0 | 0 | 0 | 0 | 0 |
| D19_TELKO_ANZ_24 | 0 | 0 | 0 | 0 | 1 |
| D19_TELKO_DATUM | 10 | 10 | 10 | 10 | 6 |
| D19_TELKO_MOBILE | 0 | 0 | 0 | 0 | 6 |
| D19_TELKO_OFFLINE_DATUM | 10 | 10 | 10 | 10 | 8 |
| D19_TELKO_ONLINE_DATUM | 10 | 10 | 10 | 10 | 10 |
| D19_TELKO_ONLINE_QUOTE_12 | NaN | NaN | 0.00 | 0.00 | 0.00 |
| D19_TELKO_REST | 0 | 0 | 0 | 0 | 5 |
| D19_TIERARTIKEL | 0 | 0 | 0 | 0 | 0 |
| D19_VERSAND_ANZ_12 | 0 | 0 | 0 | 0 | 6 |
| D19_VERSAND_ANZ_24 | 0 | 0 | 0 | 0 | 6 |
| D19_VERSAND_DATUM | 10 | 10 | 10 | 10 | 1 |
| D19_VERSAND_OFFLINE_DATUM | 10 | 10 | 10 | 10 | 9 |
| D19_VERSAND_ONLINE_DATUM | 10 | 10 | 10 | 10 | 1 |
| D19_VERSAND_ONLINE_QUOTE_12 | NaN | NaN | 0.00 | 0.00 | 10.00 |
| D19_VERSAND_REST | 0 | 0 | 0 | 0 | 2 |
| D19_VERSI_ANZ_12 | 0 | 0 | 0 | 0 | 1 |
| D19_VERSI_ANZ_24 | 0 | 0 | 0 | 0 | 3 |
| D19_VERSI_DATUM | 10 | 10 | 10 | 10 | 2 |
| D19_VERSI_OFFLINE_DATUM | 10 | 10 | 10 | 10 | 7 |
| D19_VERSI_ONLINE_DATUM | 10 | 10 | 10 | 10 | 10 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| D19_VERSI_ONLINE_QUOTE_12 | NaN | NaN | 0.00 | 0.00 | 0.00 |
| D19_VERSICHERUNGEN | 0 | 0 | 0 | 0 | 3 |
| D19_VOLLSORTIMENT | 0 | 0 | 7 | 0 | 0 |
| D19_WEIN_FEINKOST | 0 | 0 | 0 | 0 | 0 |
| DSL_FLAG | NaN | 1.00 | 1.00 | 1.00 | 1.00 |
| EINGEZOGENAM_HH_JAHR | NaN | 2004 | 2000 | 1998 | 2004 |
| EWDICHTE | NaN | 3.00 | 4.00 | 2.00 | 5.00 |
| EXTSEL992 | NaN | NaN | 14.00 | 31.00 | NaN |
| FINANZ_ANLEGER | 5 | 5 | 2 | 2 | 1 |
| FINANZ_HAUSBAUER | 3 | 5 | 5 | 2 | 2 |
| FINANZ_MINIMALIST | 3 | 1 | 1 | 4 | 4 |
| FINANZ_SPARER | 4 | 5 | 4 | 2 | 3 |
| FINANZ_UNAUFFAELLIGER | 5 | 4 | 3 | 1 | 3 |
| FINANZ_VORSORGER | 3 | 2 | 1 | 5 | 4 |
| FINANZTYP | 4 | 1 | 1 | 6 | 5 |
| FIRMENDICHTE | NaN | 2.00 | 4.00 | 4.00 | 5.00 |
| GEBAEUDETYP | NaN | 8.00 | 1.00 | 1.00 | 1.00 |
| GEBAEUDETYP_RASTER | NaN | 3.00 | 4.00 | 4.00 | 5.00 |
| GEBURTSJAHR | 0 | 1996 | 1979 | 1957 | 1963 |
| GEMEINDETYP | NaN | 22.00 | 22.00 | 40.00 | 21.00 |
| GFK_URLAUBERTYP | 10.00 | 10.00 | 10.00 | 1.00 | 5.00 |
| GREEN_AVANTGARDE | 0 | 0 | 1 | 0 | 0 |
| HEALTH_TYP | -1 | 3 | 3 | 2 | 3 |
| HH_DELTA_FLAG | NaN | 0.00 | 0.00 | NaN | 0.00 |
| HH_EINKOMMEN_SCORE | 2.00 | 6.00 | 4.00 | 1.00 | 5.00 |
| INNENSTADT | NaN | 8.00 | 4.00 | 6.00 | 1.00 |
| KBA05_ALTER1 | NaN | 3.00 | 2.00 | 2.00 | 0.00 |
| KBA05_ALTER2 | NaN | 4.00 | 3.00 | 5.00 | 4.00 |
| KBA05_ALTER3 | NaN | 1.00 | 3.00 | 3.00 | 4.00 |
| KBA05_ALTER4 | NaN | 4.00 | 3.00 | 0.00 | 3.00 |
| KBA05_ANHANG | NaN | 0.00 | 0.00 | 1.00 | 0.00 |
| KBA05_ANTG1 | NaN | 0.00 | 1.00 | 4.00 | 1.00 |
| KBA05_ANTG2 | NaN | 0.00 | 3.00 | 1.00 | 4.00 |
| KBA05_ANTG3 | NaN | 0.00 | 1.00 | 0.00 | 1.00 |
| KBA05_ANTG4 | NaN | 2.00 | 0.00 | 0.00 | 0.00 |
| KBA05_AUTOQUOT | NaN | 1.00 | 3.00 | 4.00 | 3.00 |
| KBA05_BAUMAX | NaN | 5.00 | 0.00 | 1.00 | 0.00 |
| KBA05_CCM1 | NaN | 1.00 | 5.00 | 2.00 | 4.00 |
| KBA05_CCM2 | NaN | 5.00 | 2.00 | 3.00 | 1.00 |
| KBA05_CCM3 | NaN | 1.00 | 3.00 | 5.00 | 4.00 |
| KBA05_CCM4 | NaN | 4.00 | 0.00 | 1.00 | 2.00 |
| KBA05_DIESEL | NaN | 2.00 | 0.00 | 3.00 | 1.00 |
| KBA05_FRAU | NaN | 4.00 | 3.00 | 4.00 | 5.00 |
| KBA05_GBZ | NaN | 1.00 | 3.00 | 4.00 | 3.00 |
| KBA05_HERST1 | NaN | 5.00 | 2.00 | 4.00 | 2.00 |
| KBA05_HERST2 | NaN | 5.00 | 2.00 | 3.00 | 2.00 |
| KBA05_HERST3 | NaN | 2.00 | 3.00 | 3.00 | 4.00 |
| KBA05_HERST4 | NaN | 2.00 | 2.00 | 2.00 | 1.00 |
| KBA05_HERST5 | NaN | 0.00 | 5.00 | 3.00 | 4.00 |
| KBA05_HERSTTEMP | NaN | 4.00 | 4.00 | 3.00 | 3.00 |
| KBA05_KRSAQUOT | NaN | 1.00 | 3.00 | 4.00 | 3.00 |
| KBA05_KRSHERST1 | NaN | 5.00 | 3.00 | 4.00 | 3.00 |
| KBA05_KRSHERST2 | NaN | 4.00 | 2.00 | 2.00 | 3.00 |
| KBA05_KRSHERST3 | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA05_KRSKLEIN | NaN | 1.00 | 3.00 | 1.00 | 2.00 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| KBA05_KRSOBER | NaN | 2.00 | 2.00 | 2.00 | 2.00 |
| KBA05_KRSVAN | NaN | 1.00 | 2.00 | 2.00 | 3.00 |
| KBA05_KRSZUL | NaN | 2.00 | 3.00 | 3.00 | 2.00 |
| KBA05_KW1 | NaN | 1.00 | 3.00 | 3.00 | 3.00 |
| KBA05_KW2 | NaN | 3.00 | 2.00 | 4.00 | 2.00 |
| KBA05_KW3 | NaN | 4.00 | 2.00 | 1.00 | 3.00 |
| KBA05_MAXAH | NaN | 2.00 | 3.00 | 3.00 | 2.00 |
| KBA05_MAXBJ | NaN | 1.00 | 4.00 | 4.00 | 2.00 |
| KBA05_MAXHERST | NaN | 2.00 | 5.00 | 3.00 | 3.00 |
| KBA05_MAXSEG | NaN | 4.00 | 1.00 | 2.00 | 1.00 |
| KBA05_MAXVORB | NaN | 3.00 | 1.00 | 2.00 | 1.00 |
| KBA05_MOD1 | NaN | 3.00 | 0.00 | 2.00 | 3.00 |
| KBA05_MOD2 | NaN | 2.00 | 2.00 | 4.00 | 1.00 |
| KBA05_MOD3 | NaN | 2.00 | 5.00 | 4.00 | 1.00 |
| KBA05_MOD4 | NaN | 0.00 | 1.00 | 2.00 | 4.00 |
| KBA05_MOD8 | NaN | 0.00 | 1.00 | 1.00 | 2.00 |
| KBA05_MODTEMP | NaN | 1.00 | 4.00 | 3.00 | 3.00 |
| KBA05_MOTOR | NaN | 3.00 | 1.00 | 3.00 | 4.00 |
| KBA05_MOTRAD | NaN | 0.00 | 1.00 | 3.00 | 1.00 |
| KBA05_SEG1 | NaN | 0.00 | 2.00 | 1.00 | 3.00 |
| KBA05_SEG10 | NaN | 4.00 | 1.00 | 1.00 | 3.00 |
| KBA05_SEG2 | NaN | 1.00 | 5.00 | 2.00 | 4.00 |
| KBA05_SEG3 | NaN | 2.00 | 3.00 | 5.00 | 1.00 |
| KBA05_SEG4 | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA05_SEG5 | NaN | 2.00 | 1.00 | 2.00 | 2.00 |
| KBA05_SEG6 | NaN | 1.00 | 0.00 | 0.00 | 0.00 |
| KBA05_SEG7 | NaN | 3.00 | 0.00 | 0.00 | 1.00 |
| KBA05_SEG8 | NaN | 3.00 | 0.00 | 0.00 | 0.00 |
| KBA05_SEG9 | NaN | 0.00 | 1.00 | 1.00 | 2.00 |
| KBA05_VORB0 | NaN | 1.00 | 4.00 | 2.00 | 5.00 |
| KBA05_VORB1 | NaN | 1.00 | 2.00 | 5.00 | 1.00 |
| KBA05_VORB2 | NaN | 5.00 | 3.00 | 1.00 | 5.00 |
| KBA05_ZUL1 | NaN | 5.00 | 2.00 | 3.00 | 3.00 |
| KBA05_ZUL2 | NaN | 1.00 | 3.00 | 3.00 | 4.00 |
| KBA05_ZUL3 | NaN | 0.00 | 4.00 | 3.00 | 2.00 |
| KBA05_ZUL4 | NaN | 2.00 | 4.00 | 3.00 | 2.00 |
| KBA13_ALTERHALTER_30 | NaN | 3.00 | 3.00 | 2.00 | 3.00 |
| KBA13_ALTERHALTER_45 | NaN | 2.00 | 2.00 | 3.00 | 3.00 |
| KBA13_ALTERHALTER_60 | NaN | 3.00 | 3.00 | 5.00 | 3.00 |
| KBA13_ALTERHALTER_61 | NaN | 4.00 | 3.00 | 2.00 | 3.00 |
| KBA13_ANTG1 | NaN | 2.00 | 2.00 | 2.00 | 1.00 |
| KBA13_ANTG2 | NaN | 4.00 | 3.00 | 3.00 | 4.00 |
| KBA13_ANTG3 | NaN | 2.00 | 1.00 | 1.00 | 2.00 |
| KBA13_ANTG4 | NaN | 1.00 | 0.00 | 0.00 | 1.00 |
| KBA13_ANZAHL_PKW | NaN | 963.00 | 712.00 | 596.00 | 435.00 |
| KBA13_AUDI | NaN | 4.00 | 3.00 | 5.00 | 4.00 |
| KBA13_AUTOQUOTE | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_BAUMAX | NaN | 2.00 | 1.00 | 1.00 | 2.00 |
| KBA13_BJ_1999 | NaN | 3.00 | 2.00 | 2.00 | 3.00 |
| KBA13_BJ_2000 | NaN | 3.00 | 2.00 | 2.00 | 3.00 |
| KBA13_BJ_2004 | NaN | 3.00 | 4.00 | 3.00 | 3.00 |
| KBA13_BJ_2006 | NaN | 3.00 | 5.00 | 3.00 | 2.00 |
| KBA13_BJ_2008 | NaN | 3.00 | 3.00 | 4.00 | 0.00 |
| KBA13_BJ_2009 | NaN | 2.00 | 1.00 | 3.00 | 5.00 |
| KBA13_BMW | NaN | 3.00 | 4.00 | 4.00 | 2.00 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| KBA13_CCM_0_1400 | NaN | 2.00 | 1.00 | 3.00 | 3.00 |
| KBA13_CCM_1000 | NaN | 0.00 | 1.00 | 4.00 | 5.00 |
| KBA13_CCM_1200 | NaN | 0.00 | 2.00 | 2.00 | 1.00 |
| KBA13_CCM_1400 | NaN | 4.00 | 2.00 | 3.00 | 2.00 |
| KBA13_CCM_1401_2500 | NaN | 3.00 | 3.00 | 2.00 | 1.00 |
| KBA13_CCM_1500 | NaN | 1.00 | 4.00 | 3.00 | 4.00 |
| KBA13_CCM_1600 | NaN | 2.00 | 3.00 | 3.00 | 1.00 |
| KBA13_CCM_1800 | NaN | 2.00 | 4.00 | 2.00 | 3.00 |
| KBA13_CCM_2000 | NaN | 5.00 | 3.00 | 3.00 | 3.00 |
| KBA13_CCM_2500 | NaN | 3.00 | 3.00 | 4.00 | 3.00 |
| KBA13_CCM_2501 | NaN | 3.00 | 4.00 | 4.00 | 5.00 |
| KBA13_CCM_3000 | NaN | 0.00 | 3.00 | 3.00 | 5.00 |
| KBA13_CCM_3001 | NaN | 5.00 | 5.00 | 5.00 | 5.00 |
| KBA13_FAB_ASIEN | NaN | 2.00 | 4.00 | 3.00 | 3.00 |
| KBA13_FAB_SONSTIGE | NaN | 3.00 | 3.00 | 2.00 | 2.00 |
| KBA13_FIAT | NaN | 4.00 | 3.00 | 3.00 | 3.00 |
| KBA13_FORD | NaN | 2.00 | 4.00 | 3.00 | 4.00 |
| KBA13_GBZ | NaN | 4.00 | 4.00 | 4.00 | 3.00 |
| KBA13_HALTER_20 | NaN | 3.00 | 3.00 | 2.00 | 3.00 |
| KBA13_HALTER_25 | NaN | 3.00 | 3.00 | 2.00 | 3.00 |
| KBA13_HALTER_30 | NaN | 3.00 | 2.00 | 2.00 | 3.00 |
| KBA13_HALTER_35 | NaN | 3.00 | 2.00 | 3.00 | 3.00 |
| KBA13_HALTER_40 | NaN | 3.00 | 2.00 | 3.00 | 3.00 |
| KBA13_HALTER_45 | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_HALTER_50 | NaN | 2.00 | 3.00 | 5.00 | 4.00 |
| KBA13_HALTER_55 | NaN | 3.00 | 3.00 | 5.00 | 4.00 |
| KBA13_HALTER_60 | NaN | 3.00 | 3.00 | 4.00 | 2.00 |
| KBA13_HALTER_65 | NaN | 3.00 | 4.00 | 3.00 | 3.00 |
| KBA13_HALTER_66 | NaN | 4.00 | 3.00 | 2.00 | 3.00 |
| KBA13_HERST_ASIEN | NaN | 1.00 | 3.00 | 3.00 | 3.00 |
| KBA13_HERST_AUDI_VW | NaN | 4.00 | 2.00 | 4.00 | 4.00 |
| KBA13_HERST_BMW_BENZ | NaN | 4.00 | 4.00 | 3.00 | 3.00 |
| KBA13_HERST_EUROPA | NaN | 4.00 | 3.00 | 2.00 | 4.00 |
| KBA13_HERST_FORD_OPEL | NaN | 2.00 | 3.00 | 2.00 | 3.00 |
| KBA13_HERST_SONST | NaN | 3.00 | 3.00 | 2.00 | 2.00 |
| KBA13_HHZ | NaN | 5.00 | 4.00 | 3.00 | 3.00 |
| KBA13_KMH_0_140 | NaN | 3.00 | 1.00 | 5.00 | 5.00 |
| KBA13_KMH_110 | NaN | 1.00 | 1.00 | 1.00 | 1.00 |
| KBA13_KMH_140 | NaN | 3.00 | 1.00 | 5.00 | 5.00 |
| KBA13_KMH_140_210 | NaN | 3.00 | 2.00 | 1.00 | 1.00 |
| KBA13_KMH_180 | NaN | 2.00 | 2.00 | 2.00 | 1.00 |
| KBA13_KMH_210 | NaN | 4.00 | 4.00 | 2.00 | 3.00 |
| KBA13_KMH_211 | NaN | 3.00 | 4.00 | 5.00 | 5.00 |
| KBA13_KMH_250 | NaN | 3.00 | 4.00 | 5.00 | 5.00 |
| KBA13_KMH_251 | NaN | 1.00 | 1.00 | 1.00 | 1.00 |
| KBA13_KRSAQUOT | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_KRSHERST_AUDI_VW | NaN | 4.00 | 3.00 | 4.00 | 4.00 |
| KBA13_KRSHERST_BMW_BENZ | NaN | 3.00 | 3.00 | 4.00 | 3.00 |
| KBA13_KRSHERST_FORD_OPEL | NaN | 3.00 | 2.00 | 2.00 | 2.00 |
| KBA13_KRSSEG_KLEIN | NaN | 2.00 | 2.00 | 2.00 | 2.00 |
| KBA13_KRSSEG_OBER | NaN | 2.00 | 3.00 | 2.00 | 2.00 |
| KBA13_KRSSEG_VAN | NaN | 2.00 | 2.00 | 2.00 | 1.00 |
| KBA13_KRSZUL_NEU | NaN | 1.00 | 1.00 | 2.00 | 3.00 |
| KBA13_KW_0_60 | NaN | 3.00 | 1.00 | 3.00 | 3.00 |
| KBA13_KW_110 | NaN | 4.00 | 3.00 | 1.00 | 3.00 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| KBA13_KW_120 | NaN | 4.00 | 4.00 | 3.00 | 1.00 |
| KBA13_KW_121 | NaN | 3.00 | 4.00 | 5.00 | 5.00 |
| KBA13_KW_30 | NaN | 1.00 | 1.00 | 1.00 | 1.00 |
| KBA13_KW_40 | NaN | 2.00 | 1.00 | 4.00 | 5.00 |
| KBA13_KW_50 | NaN | 4.00 | 2.00 | 3.00 | 3.00 |
| KBA13_KW_60 | NaN | 0.00 | 1.00 | 2.00 | 0.00 |
| KBA13_KW_61_120 | NaN | 3.00 | 5.00 | 2.00 | 2.00 |
| KBA13_KW_70 | NaN | 1.00 | 4.00 | 3.00 | 2.00 |
| KBA13_KW_80 | NaN | 2.00 | 4.00 | 2.00 | 1.00 |
| KBA13_KW_90 | NaN | 3.00 | 2.00 | 3.00 | 3.00 |
| KBA13_MAZDA | NaN | 2.00 | 3.00 | 2.00 | 3.00 |
| KBA13_MERCEDES | NaN | 4.00 | 4.00 | 3.00 | 3.00 |
| KBA13_MOTOR | NaN | 3.00 | 3.00 | 3.00 | 4.00 |
| KBA13_NISSAN | NaN | 2.00 | 3.00 | 5.00 | 4.00 |
| KBA13_OPEL | NaN | 3.00 | 2.00 | 2.00 | 3.00 |
| KBA13_PEUGEOT | NaN | 4.00 | 3.00 | 3.00 | 3.00 |
| KBA13_RENAULT | NaN | 3.00 | 3.00 | 2.00 | 4.00 |
| KBA13_SEG_GELAENDEWAGEN | NaN | 2.00 | 5.00 | 3.00 | 3.00 |
| KBA13_SEG_GROSSRAUMVANS | NaN | 3.00 | 3.00 | 4.00 | 3.00 |
| KBA13_SEG_KLEINST | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_SEG_KLEINWAGEN | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_SEG_KOMPAKTKLASSE | NaN | 5.00 | 1.00 | 4.00 | 3.00 |
| KBA13_SEG_MINIVANS | NaN | 4.00 | 3.00 | 4.00 | 2.00 |
| KBA13_SEG_MINIWAGEN | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_SEG_MITTELKLASSE | NaN | 3.00 | 2.00 | 4.00 | 2.00 |
| KBA13_SEG_OBEREMITTELKLASSE | NaN | 3.00 | 4.00 | 3.00 | 4.00 |
| KBA13_SEG_OBERKLASSE | NaN | 3.00 | 3.00 | 1.00 | 4.00 |
| KBA13_SEG_SONSTIGE | NaN | 2.00 | 2.00 | 2.00 | 5.00 |
| KBA13_SEG_SPORTWAGEN | NaN | 3.00 | 4.00 | 3.00 | 4.00 |
| KBA13_SEG_UTILITIES | NaN | 3.00 | 5.00 | 2.00 | 2.00 |
| KBA13_SEG_VAN | NaN | 4.00 | 3.00 | 4.00 | 2.00 |
| KBA13_SEG_WOHNMOBILE | NaN | 2.00 | 2.00 | 2.00 | 5.00 |
| KBA13_SITZE_4 | NaN | 3.00 | 4.00 | 3.00 | 3.00 |
| KBA13_SITZE_5 | NaN | 3.00 | 2.00 | 3.00 | 3.00 |
| KBA13_SITZE_6 | NaN | 4.00 | 3.00 | 3.00 | 3.00 |
| KBA13_TOYOTA | NaN | 2.00 | 3.00 | 3.00 | 3.00 |
| KBA13_VORB_0 | NaN | 3.00 | 3.00 | 4.00 | 4.00 |
| KBA13_VORB_1 | NaN | 3.00 | 4.00 | 3.00 | 2.00 |
| KBA13_VORB_1_2 | NaN | 3.00 | 4.00 | 2.00 | 2.00 |
| KBA13_VORB_2 | NaN | 3.00 | 2.00 | 3.00 | 3.00 |
| KBA13_VORB_3 | NaN | 3.00 | 2.00 | 2.00 | 4.00 |
| KBA13_VW | NaN | 4.00 | 2.00 | 4.00 | 3.00 |
| KK_KUNDENTYP | NaN | NaN | NaN | NaN | 1.00 |
| KKK | NaN | 2.00 | 2.00 | 0.00 | 3.00 |
| KOMBIALTER | 9 | 1 | 2 | 4 | 3 |
| KONSUMNAEHE | NaN | 1.00 | 5.00 | 4.00 | 4.00 |
| KONSUMZELLE | NaN | 1.00 | 0.00 | 0.00 | 0.00 |
| LP_FAMILIE_FEIN | 2.00 | 5.00 | 1.00 | 0.00 | 10.00 |
| LP_FAMILIE_GROB | 2.00 | 3.00 | 1.00 | 0.00 | 5.00 |
| LP_LEBENSPHASE_FEIN | 15.00 | 21.00 | 3.00 | 0.00 | 32.00 |
| LP_LEBENSPHASE_GROB | 4.00 | 6.00 | 1.00 | 0.00 | 10.00 |
| LP_STATUS_FEIN | 1.00 | 2.00 | 3.00 | 9.00 | 3.00 |
| LP_STATUS_GROB | 1.00 | 1.00 | 2.00 | 4.00 | 2.00 |
| MIN_GEBAEUDEJAHR | NaN | 1992 | 1992 | 1997 | 1992 |
| MOBI_RASTER | NaN | 1.00 | 2.00 | 4.00 | 1.00 |

Continued on next page

| | Row_1 | Row_2 | Row_3 | Row_4 | Row_5 |
|---|---|---|---|---|---|
| MOBI_REGIO | NaN | 1.00 | 3.00 | 4.00 | 3.00 |
| NATIONALITAET_KZ | 0 | 1 | 1 | 1 | 1 |
| ONLINE_AFFINITAET | 1.00 | 3.00 | 2.00 | 1.00 | 5.00 |
| ORTSGR_KLS9 | NaN | 5.00 | 5.00 | 3.00 | 6.00 |
| OST_WEST_KZ | NaN | W | W | W | W |
| PLZ8_ANTG1 | NaN | 2.00 | 3.00 | 2.00 | 2.00 |
| PLZ8_ANTG2 | NaN | 3.00 | 3.00 | 2.00 | 4.00 |
| PLZ8_ANTG3 | NaN | 2.00 | 1.00 | 2.00 | 2.00 |
| PLZ8_ANTG4 | NaN | 1.00 | 0.00 | 0.00 | 1.00 |
| PLZ8_BAUMAX | NaN | 1.00 | 1.00 | 1.00 | 2.00 |
| PLZ8_GBZ | NaN | 4.00 | 4.00 | 4.00 | 3.00 |
| PLZ8_HHZ | NaN | 5.00 | 4.00 | 3.00 | 3.00 |
| PRAEGENDE_JUGENDJAHRE | 0 | 14 | 15 | 8 | 8 |
| REGIOTYP | NaN | 3.00 | 2.00 | 0.00 | 5.00 |
| RELAT_AB | NaN | 4.00 | 2.00 | 3.00 | 5.00 |
| RETOURTYP_BK_S | 5.00 | 1.00 | 3.00 | 2.00 | 5.00 |
| RT_KEIN_ANREIZ | 1.00 | 5.00 | 5.00 | 3.00 | 3.00 |
| RT_SCHNAEPPCHEN | 4.00 | 3.00 | 4.00 | 2.00 | 5.00 |
| RT_UEBERGROESSE | 1.00 | 5.00 | 5.00 | 3.00 | 5.00 |
| SEMIO_DOM | 6 | 7 | 7 | 4 | 2 |
| SEMIO_ERL | 3 | 2 | 6 | 7 | 4 |
| SEMIO_FAM | 6 | 4 | 1 | 1 | 4 |
| SEMIO_KAEM | 6 | 4 | 7 | 5 | 2 |
| SEMIO_KRIT | 7 | 4 | 7 | 4 | 3 |
| SEMIO_KULT | 3 | 3 | 3 | 4 | 6 |
| SEMIO_LUST | 5 | 2 | 4 | 4 | 4 |
| SEMIO_MAT | 5 | 3 | 3 | 1 | 2 |
| SEMIO_PFLICHT | 5 | 7 | 3 | 4 | 4 |
| SEMIO_RAT | 4 | 6 | 4 | 3 | 2 |
| SEMIO_REL | 7 | 4 | 3 | 2 | 4 |
| SEMIO_SOZ | 2 | 5 | 4 | 5 | 6 |
| SEMIO_TRADV | 3 | 6 | 3 | 4 | 2 |
| SEMIO_VERT | 1 | 1 | 4 | 4 | 7 |
| SHOPPER_TYP | -1 | 3 | 2 | 1 | 2 |
| SOHO_KZ | NaN | 1.00 | 0.00 | 0.00 | 0.00 |
| STRUKTURTYP | NaN | 2.00 | 3.00 | 1.00 | 3.00 |
| TITEL_KZ | NaN | 0.00 | 0.00 | 0.00 | 0.00 |
| UMFELD_ALT | NaN | 3.00 | 2.00 | 4.00 | 4.00 |
| UMFELD_JUNG | NaN | 3.00 | 5.00 | 5.00 | 3.00 |
| UNGLEICHENN_FLAG | NaN | 1.00 | 0.00 | 0.00 | 0.00 |
| VERDICHTUNGSRAUM | NaN | 0.00 | 1.00 | 0.00 | 1.00 |
| VERS_TYP | -1 | 2 | 1 | 1 | 2 |
| VHA | NaN | 0.00 | 0.00 | 1.00 | 0.00 |
| VHN | NaN | 4.00 | 2.00 | 0.00 | 2.00 |
| VK_DHT4A | NaN | 8.00 | 9.00 | 7.00 | 3.00 |
| VK_DISTANZ | NaN | 11.00 | 9.00 | 10.00 | 5.00 |
| VK_ZG11 | NaN | 10.00 | 6.00 | 11.00 | 4.00 |
| W_KEIT_KIND_HH | NaN | 3.00 | 3.00 | NaN | 2.00 |
| WOHNDAUER_2008 | NaN | 9.00 | 9.00 | 9.00 | 9.00 |
| WOHNLAGE | NaN | 4.00 | 2.00 | 7.00 | 3.00 |
| ZABEOTYP | 3 | 5 | 5 | 3 | 4 |
| ANREDE_KZ | 1 | 2 | 2 | 2 | 1 |
| ALTERSKATEGORIE_GROB | 2 | 1 | 3 | 4 | 3 |

Table 4: First 5 rows from the general population dataset

### 6.2 First 5 rows from the Feature Summary file

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 0 | AGER_TYP | categorical | [-1,0] | person |
| 1 | ALTERSKATEGORIE_GROB | ordinal | [-1,0,9] | person |
| 2 | ALTER_HH | interval | [0] | household |
| 3 | ANREDE_KZ | categorical | [-1,0] | person |
| 4 | ANZ_HAUSHALTE_AKTIV | numeric | [] | building |
| 5 | ANZ_HH_TITEL | numeric | [] | building |
| 6 | ANZ_PERSONEN | numeric | [] | household |
| 7 | ANZ_TITEL | numeric | [] | household |
| 8 | BALLRAUM | ordinal | [-1] | postcode |
| 9 | CAMEO_DEUG_2015 | categorical | [-1,X] | microcell_rr4 |
| 10 | CAMEO_DEU_2015 | categorical | [XX] | microcell_rr4 |
| 11 | CAMEO_INTL_2015 | mixed | [-1,XX] | microcell_rr4 |
| 12 | CJT_GESAMTTYP | categorical | [0] | person |
| 13 | D19_BANKEN_ANZ_12 | ordinal | [0] | household |
| 14 | D19_BANKEN_ANZ_24 | ordinal | [0] | household |
| 15 | D19_BANKEN_DATUM | ordinal | [10] | household |
| 16 | D19_BANKEN_DIREKT | categorical | [0] | grid_125_125 |
| 17 | D19_BANKEN_GROSS | categorical | [0] | grid_125_125 |
| 18 | D19_BANKEN_LOKAL | categorical | [0] | grid_125_125 |
| 19 | D19_BANKEN_OFFLINE_DATUM | ordinal | [10] | household |
| 20 | D19_BANKEN_ONLINE_DATUM | ordinal | [10] | household |
| 21 | D19_BANKEN_ONLINE_QUOTE_12 | ordinal | [] | household |
| 22 | D19_BANKEN_REST | categorical | [0] | grid_125_125 |
| 23 | D19_BEKLEIDUNG_GEH | categorical | [0] | grid_125_125 |
| 24 | D19_BEKLEIDUNG_REST | categorical | [0] | grid_125_125 |
| 25 | D19_BILDUNG | categorical | [0] | grid_125_125 |
| 26 | D19_BIO_OEKO | categorical | [0] | grid_125_125 |
| 27 | D19_BUCH_CD | categorical | [0] | grid_125_125 |
| 28 | D19_DIGIT_SERV | categorical | [0] | grid_125_125 |
| 29 | D19_DROGERIEARTIKEL | categorical | [0] | grid_125_125 |
| 30 | D19_ENERGIE | categorical | [0] | grid_125_125 |
| 31 | D19_FREIZEIT | categorical | [0] | grid_125_125 |
| 32 | D19_GARTEN | categorical | [0] | grid_125_125 |
| 33 | D19_GESAMT_ANZ_12 | ordinal | [0] | household |
| 34 | D19_GESAMT_ANZ_24 | ordinal | [0] | household |
| 35 | D19_GESAMT_DATUM | ordinal | [10] | household |
| 36 | D19_GESAMT_OFFLINE_DATUM | ordinal | [10] | household |
| 37 | D19_GESAMT_ONLINE_DATUM | ordinal | [10] | household |
| 38 | D19_GESAMT_ONLINE_QUOTE_12 | ordinal | [] | household |
| 39 | D19_HANDWERK | categorical | [0] | grid_125_125 |
| 40 | D19_HAUS_DEKO | categorical | [0] | grid_125_125 |
| 41 | D19_KINDERARTIKEL | categorical | [0] | grid_125_125 |
| 42 | D19_KONSUMTYP | categorical | [] | household |
| 43 | KK_KUNDENTYP | categorical | [-1] | household |
| 44 | D19_KOSMETIK | categorical | [0] | grid_125_125 |
| 45 | D19_LEBENSMITTEL | categorical | [0] | grid_125_125 |
| 46 | D19_LOTTO | categorical | [0] | grid_125_125 |
| 47 | D19_NAHRUNGSERGAENZUNG | categorical | [0] | grid_125_125 |
| 48 | D19_RATGEBER | categorical | [0] | grid_125_125 |
| 49 | D19_REISEN | categorical | [0] | grid_125_125 |
| 50 | D19_SAMMELARTIKEL | categorical | [0] | grid_125_125 |
| 51 | D19_SCHUHE | categorical | [0] | grid_125_125 |
| 52 | D19_SONSTIGE | categorical | [0] | grid_125_125 |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 53 | D19_TECHNIK | categorical | [0] | grid_125_125 |
| 54 | D19_TELKO_ANZ_12 | ordinal | [0] | household |
| 55 | D19_TELKO_ANZ_24 | ordinal | [0] | household |
| 56 | D19_TELKO_DATUM | ordinal | [10] | household |
| 57 | D19_TELKO_MOBILE | categorical | [0] | grid_125_125 |
| 58 | D19_TELKO_OFFLINE_DATUM | ordinal | [10] | household |
| 59 | D19_TELKO_ONLINE_DATUM | ordinal | [10] | household |
| 60 | D19_TELKO_REST | categorical | [0] | grid_125_125 |
| 61 | D19_TIERARTIKEL | categorical | [0] | grid_125_125 |
| 62 | D19_VERSAND_ANZ_12 | ordinal | [0] | household |
| 63 | D19_VERSAND_ANZ_24 | ordinal | [0] | household |
| 64 | D19_VERSAND_DATUM | ordinal | [10] | household |
| 65 | D19_VERSAND_OFFLINE_DATUM | ordinal | [10] | household |
| 66 | D19_VERSAND_ONLINE_DATUM | ordinal | [10] | household |
| 67 | D19_VERSAND_ONLINE_QUOTE_12 | ordinal | [] | household |
| 68 | D19_VERSAND_REST | categorical | [0] | grid_125_125 |
| 69 | D19_VERSICHERUNGEN | categorical | [0] | grid_125_125 |
| 70 | D19_VERSI_ANZ_12 | ordinal | [0] | household |
| 71 | D19_VERSI_ANZ_24 | ordinal | [0] | household |
| 72 | D19_VOLLSORTIMENT | categorical | [0] | grid_125_125 |
| 73 | D19_WEIN_FEINKOST | categorical | [0] | grid_125_125 |
| 74 | EWDICHTE | ordinal | [-1] | postcode |
| 75 | FINANZTYP | categorical | [-1] | person |
| 76 | FINANZ_ANLEGER | ordinal | [-1] | person |
| 77 | FINANZ_HAUSBAUER | ordinal | [-1] | person |
| 78 | FINANZ_MINIMALIST | ordinal | [-1] | person |
| 79 | FINANZ_SPARER | ordinal | [-1] | person |
| 80 | FINANZ_UNAUFFAELLIGER | ordinal | [-1] | person |
| 81 | FINANZ_VORSORGER | ordinal | [-1] | person |
| 82 | GEBAEUDETYP | categorical | [-1,0] | building |
| 83 | GEBAEUDETYP_RASTER | ordinal | [] | rr1_id |
| 84 | GEBURTSJAHR | numeric | [0] | person |
| 85 | GFK_URLAUBERTYP | categorical | [] | person |
| 86 | GREEN_AVANTGARDE | categorical | [] | person |
| 87 | HEALTH_TYP | ordinal | [-1,0] | person |
| 88 | HH_EINKOMMEN_SCORE | ordinal | [-1,0] | household |
| 89 | INNENSTADT | ordinal | [-1] | postcode |
| 90 | KBA05_ALTER1 | ordinal | [-1,9] | rr3_id |
| 91 | KBA05_ALTER2 | ordinal | [-1,9] | rr3_id |
| 92 | KBA05_ALTER3 | ordinal | [-1,9] | rr3_id |
| 93 | KBA05_ALTER4 | ordinal | [-1,9] | rr3_id |
| 94 | KBA05_ANHANG | ordinal | [-1,9] | rr3_id |
| 95 | KBA05_ANTG1 | ordinal | [-1] | rr3_id |
| 96 | KBA05_ANTG2 | ordinal | [-1] | rr3_id |
| 97 | KBA05_ANTG3 | ordinal | [-1] | rr3_id |
| 98 | KBA05_ANTG4 | ordinal | [-1] | rr3_id |
| 99 | KBA05_AUTOQUOT | ordinal | [-1,9] | rr3_id |
| 100 | KBA05_BAUMAX | mixed | [-1,0] | rr3_id |
| 101 | KBA05_CCM1 | ordinal | [-1,9] | rr3_id |
| 102 | KBA05_CCM2 | ordinal | [-1,9] | rr3_id |
| 103 | KBA05_CCM3 | ordinal | [-1,9] | rr3_id |
| 104 | KBA05_CCM4 | ordinal | [-1,9] | rr3_id |
| 105 | KBA05_DIESEL | ordinal | [-1,9] | rr3_id |
| 106 | KBA05_FRAU | ordinal | [-1,9] | rr3_id |
| 107 | KBA05_GBZ | ordinal | [-1,0] | rr3_id |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 108 | KBA05_HERST1 | ordinal | [-1,9] | rr3_id |
| 109 | KBA05_HERST2 | ordinal | [-1,9] | rr3_id |
| 110 | KBA05_HERST3 | ordinal | [-1,9] | rr3_id |
| 111 | KBA05_HERST4 | ordinal | [-1,9] | rr3_id |
| 112 | KBA05_HERST5 | ordinal | [-1,9] | rr3_id |
| 113 | KBA05_HERSTTEMP | ordinal | [-1,9] | building |
| 114 | KBA05_KRSAQUOT | ordinal | [-1,9] | rr3_id |
| 115 | KBA05_KRSHERST1 | ordinal | [-1,9] | rr3_id |
| 116 | KBA05_KRSHERST2 | ordinal | [-1,9] | rr3_id |
| 117 | KBA05_KRSHERST3 | ordinal | [-1,9] | rr3_id |
| 118 | KBA05_KRSKLEIN | ordinal | [-1,9] | rr3_id |
| 119 | KBA05_KRSOBER | ordinal | [-1,9] | rr3_id |
| 120 | KBA05_KRSVAN | ordinal | [-1,9] | rr3_id |
| 121 | KBA05_KRSZUL | ordinal | [-1,9] | rr3_id |
| 122 | KBA05_KW1 | ordinal | [-1,9] | rr3_id |
| 123 | KBA05_KW2 | ordinal | [-1,9] | rr3_id |
| 124 | KBA05_KW3 | ordinal | [-1,9] | rr3_id |
| 125 | KBA05_MAXAH | ordinal | [-1,9] | rr3_id |
| 126 | KBA05_MAXBJ | ordinal | [-1,9] | rr3_id |
| 127 | KBA05_MAXHERST | ordinal | [-1,9] | rr3_id |
| 128 | KBA05_MAXSEG | ordinal | [-1,9] | rr3_id |
| 129 | KBA05_MAXVORB | ordinal | [-1,9] | rr3_id |
| 130 | KBA05_MOD1 | ordinal | [-1,9] | rr3_id |
| 131 | KBA05_MOD2 | ordinal | [-1,9] | rr3_id |
| 132 | KBA05_MOD3 | ordinal | [-1,9] | rr3_id |
| 133 | KBA05_MOD4 | ordinal | [-1,9] | rr3_id |
| 134 | KBA05_MOD8 | ordinal | [-1,9] | rr3_id |
| 135 | KBA05_MODTEMP | categorical | [-1,9] | building |
| 136 | KBA05_MOTOR | ordinal | [-1,9] | rr3_id |
| 137 | KBA05_MOTRAD | ordinal | [-1,9] | rr3_id |
| 138 | KBA05_SEG1 | ordinal | [-1,9] | rr3_id |
| 139 | KBA05_SEG10 | ordinal | [-1,9] | rr3_id |
| 140 | KBA05_SEG2 | ordinal | [-1,9] | rr3_id |
| 141 | KBA05_SEG3 | ordinal | [-1,9] | rr3_id |
| 142 | KBA05_SEG4 | ordinal | [-1,9] | rr3_id |
| 143 | KBA05_SEG5 | ordinal | [-1,9] | rr3_id |
| 144 | KBA05_SEG6 | categorical | [-1,9] | rr3_id |
| 145 | KBA05_SEG7 | ordinal | [-1,9] | rr3_id |
| 146 | KBA05_SEG8 | ordinal | [-1,9] | rr3_id |
| 147 | KBA05_SEG9 | ordinal | [-1,9] | rr3_id |
| 148 | KBA05_VORB0 | ordinal | [-1,9] | rr3_id |
| 149 | KBA05_VORB1 | ordinal | [-1,9] | rr3_id |
| 150 | KBA05_VORB2 | ordinal | [-1,9] | rr3_id |
| 151 | KBA05_ZUL1 | ordinal | [-1,9] | rr3_id |
| 152 | KBA05_ZUL2 | ordinal | [-1,9] | rr3_id |
| 153 | KBA05_ZUL3 | ordinal | [-1,9] | rr3_id |
| 154 | KBA05_ZUL4 | ordinal | [-1,9] | rr3_id |
| 155 | KBA13_ALTERHALTER_30 | ordinal | [-1] | plz8 |
| 156 | KBA13_ALTERHALTER_45 | ordinal | [-1] | plz8 |
| 157 | KBA13_ALTERHALTER_60 | ordinal | [-1] | plz8 |
| 158 | KBA13_ALTERHALTER_61 | ordinal | [-1] | plz8 |
| 159 | KBA13_ANZAHL_PKW | numeric | [] | plz8 |
| 160 | KBA13_AUDI | ordinal | [-1] | plz8 |
| 161 | KBA13_AUTOQUOTE | ordinal | [-1] | plz8 |
| 162 | KBA13_BJ_1999 | ordinal | [-1] | plz8 |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 163 | KBA13_BJ_2000 | ordinal | [-1] | plz8 |
| 164 | KBA13_BJ_2004 | ordinal | [-1] | plz8 |
| 165 | KBA13_BJ_2006 | ordinal | [-1] | plz8 |
| 166 | KBA13_BJ_2008 | ordinal | [-1] | plz8 |
| 167 | KBA13_BJ_2009 | ordinal | [-1] | plz8 |
| 168 | KBA13_BMW | ordinal | [-1] | plz8 |
| 169 | KBA13_CCM_1000 | ordinal | [-1] | plz8 |
| 170 | KBA13_CCM_1200 | ordinal | [-1] | plz8 |
| 171 | KBA13_CCM_1400 | ordinal | [-1] | plz8 |
| 172 | KBA13_CCM_0_1400 | ordinal | [-1] | plz8 |
| 173 | KBA13_CCM_1500 | ordinal | [-1] | plz8 |
| 174 | KBA13_CCM_1600 | ordinal | [-1] | plz8 |
| 175 | KBA13_CCM_1800 | ordinal | [-1] | plz8 |
| 176 | KBA13_CCM_2000 | ordinal | [-1] | plz8 |
| 177 | KBA13_CCM_2500 | ordinal | [-1] | plz8 |
| 178 | KBA13_CCM_2501 | ordinal | [-1] | plz8 |
| 179 | KBA13_CCM_3000 | ordinal | [-1] | plz8 |
| 180 | KBA13_CCM_3001 | ordinal | [-1] | plz8 |
| 181 | KBA13_FAB_ASIEN | ordinal | [-1] | plz8 |
| 182 | KBA13_FAB_SONSTIGE | ordinal | [-1] | plz8 |
| 183 | KBA13_FIAT | ordinal | [-1] | plz8 |
| 184 | KBA13_FORD | ordinal | [-1] | plz8 |
| 185 | KBA13_HALTER_20 | ordinal | [-1] | plz8 |
| 186 | KBA13_HALTER_25 | ordinal | [-1] | plz8 |
| 187 | KBA13_HALTER_30 | ordinal | [-1] | plz8 |
| 188 | KBA13_HALTER_35 | ordinal | [-1] | plz8 |
| 189 | KBA13_HALTER_40 | ordinal | [-1] | plz8 |
| 190 | KBA13_HALTER_45 | ordinal | [-1] | plz8 |
| 191 | KBA13_HALTER_50 | ordinal | [-1] | plz8 |
| 192 | KBA13_HALTER_55 | ordinal | [-1] | plz8 |
| 193 | KBA13_HALTER_60 | ordinal | [-1] | plz8 |
| 194 | KBA13_HALTER_65 | ordinal | [-1] | plz8 |
| 195 | KBA13_HALTER_66 | ordinal | [-1] | plz8 |
| 196 | KBA13_HERST_ASIEN | ordinal | [-1] | plz8 |
| 197 | KBA13_HERST_AUDI_VW | ordinal | [-1] | plz8 |
| 198 | KBA13_HERST_BMW_BENZ | ordinal | [-1] | plz8 |
| 199 | KBA13_HERST_EUROPA | ordinal | [-1] | plz8 |
| 200 | KBA13_HERST_FORD_OPEL | ordinal | [-1] | plz8 |
| 201 | KBA13_HERST_SONST | ordinal | [-1] | plz8 |
| 202 | KBA13_KMH_110 | ordinal | [-1] | plz8 |
| 203 | KBA13_KMH_140 | ordinal | [-1] | plz8 |
| 204 | KBA13_KMH_180 | ordinal | [-1] | plz8 |
| 205 | KBA13_KMH_0_140 | ordinal | [-1] | plz8 |
| 206 | KBA13_KMH_140_210 | ordinal | [-1] | plz8 |
| 207 | KBA13_KMH_211 | ordinal | [-1] | plz8 |
| 208 | KBA13_KMH_250 | ordinal | [-1] | plz8 |
| 209 | KBA13_KMH_251 | ordinal | [-1] | plz8 |
| 210 | KBA13_KRSAQUOT | ordinal | [-1] | plz8 |
| 211 | KBA13_KRSHERST_AUDI_VW | ordinal | [-1] | plz8 |
| 212 | KBA13_KRSHERST_BMW_BENZ | ordinal | [-1] | plz8 |
| 213 | KBA13_KRSHERST_FORD_OPEL | ordinal | [-1] | plz8 |
| 214 | KBA13_KRSSEG_KLEIN | ordinal | [-1] | plz8 |
| 215 | KBA13_KRSSEG_OBER | ordinal | [-1] | plz8 |
| 216 | KBA13_KRSSEG_VAN | ordinal | [-1] | plz8 |
| 217 | KBA13_KRSZUL_NEU | ordinal | [-1] | plz8 |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 218 | KBA13_KW_30 | ordinal | [-1] | plz8 |
| 219 | KBA13_KW_40 | ordinal | [-1] | plz8 |
| 220 | KBA13_KW_50 | ordinal | [-1] | plz8 |
| 221 | KBA13_KW_60 | ordinal | [-1] | plz8 |
| 222 | KBA13_KW_0_60 | ordinal | [-1] | plz8 |
| 223 | KBA13_KW_70 | ordinal | [-1] | plz8 |
| 224 | KBA13_KW_61_120 | ordinal | [-1] | plz8 |
| 225 | KBA13_KW_80 | ordinal | [-1] | plz8 |
| 226 | KBA13_KW_90 | ordinal | [-1] | plz8 |
| 227 | KBA13_KW_110 | ordinal | [-1] | plz8 |
| 228 | KBA13_KW_120 | ordinal | [-1] | plz8 |
| 229 | KBA13_KW_121 | ordinal | [-1] | plz8 |
| 230 | KBA13_MAZDA | ordinal | [-1] | plz8 |
| 231 | KBA13_MERCEDES | ordinal | [-1] | plz8 |
| 232 | KBA13_MOTOR | ordinal | [-1] | plz8 |
| 233 | KBA13_NISSAN | ordinal | [-1] | plz8 |
| 234 | KBA13_OPEL | ordinal | [-1] | plz8 |
| 235 | KBA13_PEUGEOT | ordinal | [-1] | plz8 |
| 236 | KBA13_RENAULT | ordinal | [-1] | plz8 |
| 237 | KBA13_SEG_GELAENDEWAGEN | ordinal | [-1] | plz8 |
| 238 | KBA13_SEG_GROSSRAUMVANS | ordinal | [-1] | plz8 |
| 239 | KBA13_SEG_KLEINST | ordinal | [-1] | plz8 |
| 240 | KBA13_SEG_KLEINWAGEN | ordinal | [-1] | plz8 |
| 241 | KBA13_SEG_KOMPAKTKLASSE | ordinal | [-1] | plz8 |
| 242 | KBA13_SEG_MINIVANS | ordinal | [-1] | plz8 |
| 243 | KBA13_SEG_MINIWAGEN | ordinal | [-1] | plz8 |
| 244 | KBA13_SEG_MITTELKLASSE | ordinal | [-1] | plz8 |
| 245 | KBA13_SEG_OBEREMITTELKLASSE | ordinal | [-1] | plz8 |
| 246 | KBA13_SEG_OBERKLASSE | ordinal | [-1] | plz8 |
| 247 | KBA13_SEG_SONSTIGE | ordinal | [-1] | plz8 |
| 248 | KBA13_SEG_SPORTWAGEN | ordinal | [-1] | plz8 |
| 249 | KBA13_SEG_UTILITIES | ordinal | [-1] | plz8 |
| 250 | KBA13_SEG_VAN | ordinal | [-1] | plz8 |
| 251 | KBA13_SEG_WOHNMOBILE | ordinal | [-1] | plz8 |
| 252 | KBA13_SITZE_4 | ordinal | [-1] | plz8 |
| 253 | KBA13_SITZE_5 | ordinal | [-1] | plz8 |
| 254 | KBA13_SITZE_6 | ordinal | [-1] | plz8 |
| 255 | KBA13_TOYOTA | ordinal | [-1] | plz8 |
| 256 | KBA13_VORB_0 | ordinal | [-1] | plz8 |
| 257 | KBA13_VORB_1 | ordinal | [-1] | plz8 |
| 258 | KBA13_VORB_1_2 | ordinal | [-1] | plz8 |
| 259 | KBA13_VORB_2 | ordinal | [-1] | plz8 |
| 260 | KBA13_VORB_3 | ordinal | [-1] | plz8 |
| 261 | KBA13_VW | ordinal | [-1] | plz8 |
| 262 | KKK | ordinal | [-1,0] | rr1_id |
| 263 | KONSUMNAEHE | ordinal | [] | building |
| 264 | LP_FAMILIE_FEIN | categorical | [0] | person |
| 265 | LP_FAMILIE_GROB | categorical | [0] | person |
| 266 | LP_LEBENSPHASE_FEIN | mixed | [0] | person |
| 267 | LP_LEBENSPHASE_GROB | mixed | [0] | person |
| 268 | LP_STATUS_FEIN | categorical | [0] | person |
| 269 | LP_STATUS_GROB | categorical | [0] | person |
| 270 | MIN_GEBAEUDEJAHR | numeric | [] | building |
| 271 | MOBI_REGIO | ordinal | [] | rr1_id |
| 272 | NATIONALITAET_KZ | categorical | [-1,0] | person |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 273 | ONLINE_AFFINITAET | ordinal | [] | rr1_id |
| 274 | ORTSGR_KLS9 | ordinal | [-1] | community |
| 275 | OST_WEST_KZ | categorical | [-1] | building |
| 276 | PLZ8_ANTG1 | ordinal | [-1] | plz8 |
| 277 | PLZ8_ANTG2 | ordinal | [-1] | plz8 |
| 278 | PLZ8_ANTG3 | ordinal | [-1] | plz8 |
| 279 | PLZ8_ANTG4 | ordinal | [-1] | plz8 |
| 280 | PLZ8_BAUMAX | mixed | [] | plz8 |
| 281 | PLZ8_GBZ | ordinal | [-1] | plz8 |
| 282 | PLZ8_HHZ | ordinal | [-1] | plz8 |
| 283 | PRAEGENDE_JUGENDJAHRE | mixed | [-1,0] | person |
| 284 | REGIOTYP | ordinal | [-1,0] | rr1_id |
| 285 | RELAT_AB | ordinal | [-1,9] | community |
| 286 | RETOURTYP_BK_S | ordinal | [0] | person |
| 287 | SEMIO_DOM | ordinal | [-1,9] | person |
| 288 | SEMIO_ERL | ordinal | [-1,9] | person |
| 289 | SEMIO_FAM | ordinal | [-1,9] | person |
| 290 | SEMIO_KAEM | ordinal | [-1,9] | person |
| 291 | SEMIO_KRIT | ordinal | [-1,9] | person |
| 292 | SEMIO_KULT | ordinal | [-1,9] | person |
| 293 | SEMIO_LUST | ordinal | [-1,9] | person |
| 294 | SEMIO_MAT | ordinal | [-1,9] | person |
| 295 | SEMIO_PFLICHT | ordinal | [-1,9] | person |
| 296 | SEMIO_RAT | ordinal | [-1,9] | person |
| 297 | SEMIO_REL | ordinal | [-1,9] | person |
| 298 | SEMIO_SOZ | ordinal | [-1,9] | person |
| 299 | SEMIO_TRADV | ordinal | [-1,9] | person |
| 300 | SEMIO_VERT | ordinal | [-1,9] | person |
| 301 | SHOPPER_TYP | categorical | [-1] | person |
| 302 | TITEL_KZ | categorical | [-1,0] | person |
| 303 | VERS_TYP | categorical | [-1] | person |
| 304 | WOHNDAUER_2008 | ordinal | [-1,0] | household |
| 305 | WOHNLAGE | mixed | [-1] | building |
| 306 | W_KEIT_KIND_HH | ordinal | [-1,0] | household |
| 307 | ZABEOTYP | categorical | [-1,9] | person |
| 308 | ARBEIT | ordinal | [-1,9] | community |
| 309 | AKT_DAT_KL | categorical | [] | unknown |
| 310 | ALTERSKATEGORIE_FEIN | categorical | [-1,0] | person |
| 311 | ALTER_KIND1 | numeric | [] | unknown |
| 312 | ALTER_KIND2 | numeric | [] | unknown |
| 313 | ALTER_KIND3 | numeric | [] | unknown |
| 314 | ALTER_KIND4 | numeric | [] | unknown |
| 315 | ANZ_KINDER | numeric | [] | unknown |
| 316 | ANZ_STATISTISCHE_HAUSHALTE | numeric | [] | unknown |
| 317 | CJT_KATALOGNUTZER | categorical | [] | unknown |
| 318 | CJT_TYP_1 | categorical | [] | unknown |
| 319 | CJT_TYP_2 | categorical | [] | unknown |
| 320 | CJT_TYP_3 | categorical | [] | unknown |
| 321 | CJT_TYP_4 | categorical | [] | unknown |
| 322 | CJT_TYP_5 | categorical | [] | unknown |
| 323 | CJT_TYP_6 | categorical | [] | unknown |
| 324 | D19_KONSUMTYP_MAX | categorical | [] | household |
| 325 | D19_SOZIALES | categorical | [0] | grid_125_125 |
| 326 | D19_TELKO_ONLINE_QUOTE_12 | ordinal | [10] | household |
| 327 | D19_VERSI_DATUM | ordinal | [10] | household |

Continued on next page

| | attribute | type | missing_or_unknown | information_level |
|---|---|---|---|---|
| 328 | D19_VERSI_OFFLINE_DATUM | ordinal | [10] | household |
| 329 | D19_VERSI_ONLINE_DATUM | ordinal | [10] | household |
| 330 | D19_VERSI_ONLINE_QUOTE_12 | ordinal | [] | household |
| 331 | DSL_FLAG | categorical | [] | unknown |
| 332 | EINGEZOGENAM_HH_JAHR | numeric | [] | unknown |
| 333 | EXTSEL992 | numeric | [] | unknown |
| 334 | FIRMENDICHTE | categorical | [] | unknown |
| 335 | HH_DELTA_FLAG | categorical | [] | unknown |
| 336 | KBA13_ANTG1 | ordinal | [-1] | plz8 |
| 337 | KBA13_ANTG2 | ordinal | [-1] | plz8 |
| 338 | KBA13_ANTG3 | ordinal | [-1] | plz8 |
| 339 | KBA13_ANTG4 | ordinal | [-1] | plz8 |
| 340 | KBA13_CCM_1401_2500 | ordinal | [-1] | plz8 |
| 341 | KBA13_GBZ | ordinal | [-1,0] | unknown |
| 342 | KBA13_HHZ | ordinal | [-1] | unknown |
| 343 | KBA13_KMH_210 | ordinal | [-1] | plz8 |
| 344 | KOMBIALTER | categorical | [9] | unknown |
| 345 | KONSUMZELLE | categorical | [] | unknown |
| 346 | MOBI_RASTER | categorical | [] | unknown |
| 347 | RT_KEIN_ANREIZ | categorical | [] | unknown |
| 348 | RT_SCHNAEPPCHEN | categorical | [] | unknown |
| 349 | RT_UEBERGROESSE | categorical | [] | unknown |
| 350 | SOHO_KZ | categorical | [] | person |
| 351 | STRUKTURTYP | categorical | [] | household |
| 352 | UMFELD_ALT | categorical | [] | unknown |
| 353 | UMFELD_JUNG | categorical | [] | unknown |
| 354 | UNGLEICHENN_FLAG | categorical | [] | unknown |
| 355 | VERDICHTUNGSRAUM | numeric | [] | unknown |
| 356 | VHA | categorical | [] | unknown |
| 357 | VHN | categorical | [] | unknown |
| 358 | VK_DHT4A | numeric | [] | unknown |
| 359 | VK_DISTANZ | numeric | [] | unknown |
| 360 | VK_ZG11 | numeric | [] | unknown |
| 361 | KBA13_BAUMAX | mixed | [-1,0] | plz8 |
| 362 | GEMEINDETYP | categorical | [] | unknown |

Table 5: Feature summary

### 6.3 Rows missing data analysis

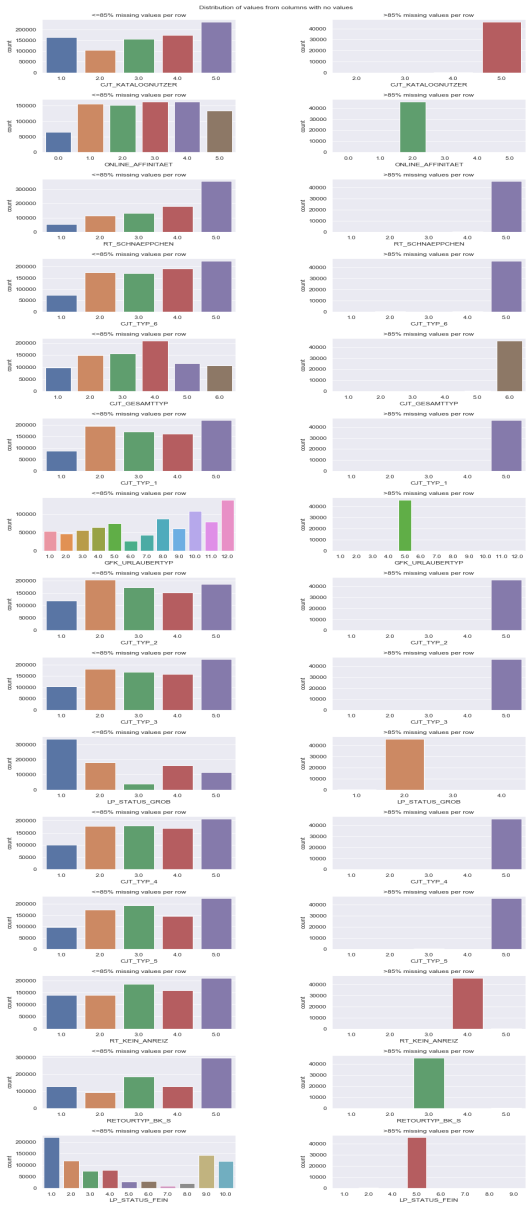



Figure 13: Distribution of data values in columns that are not missing data (or are missing very little data) for rows with less or more than 85% missing data