Arvato Financial Services, Customer Segmentation Use Case

Domain Background: Financial Services, Arvato is an IT company that has different products lines. We will be focusing on a use case that will help their financial group to identify potential customers through their mail order campaign.

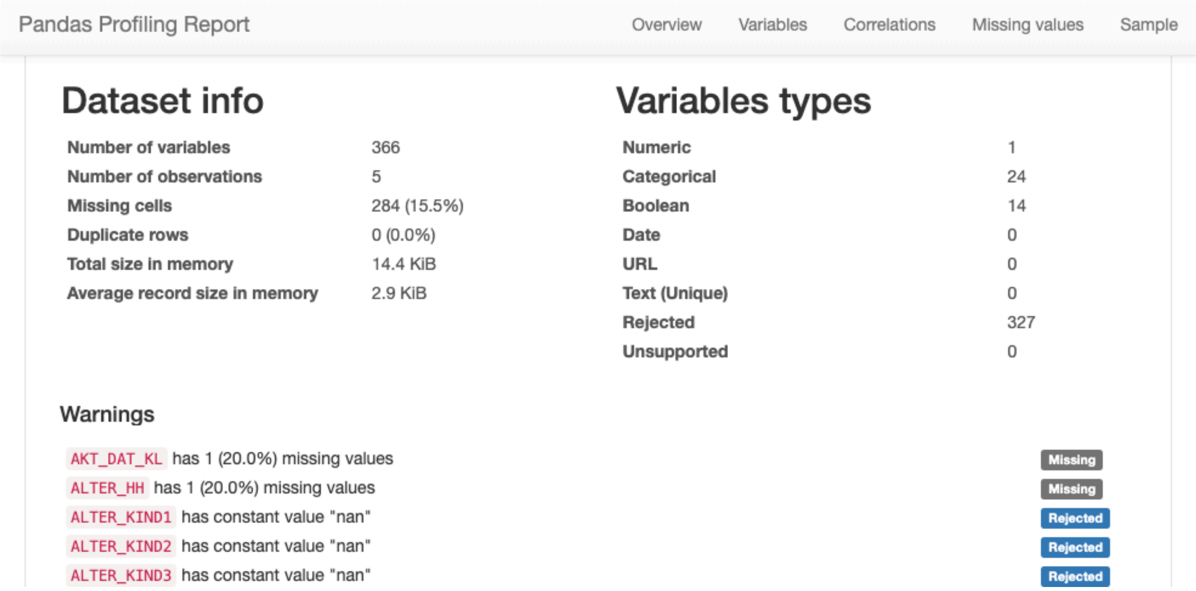
Problem Statement: Analyze demographics data for customers of a mail-order sales company (Arvato) in Germany, comparing it against demographics information for the general population. We will use unsupervised learning techniques to perform customer segmentation and identifying parts of the population that best describe the core customer base of the company. Then, we'll apply what we've learned on a dataset, with demographics 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.

Key Project Milestones:

- Customer Segmentation using Demographics Data
- Develop a Model to predict potential customer for the company through ad campaign
- Upload the results to Kaggle competition to share and score the performance and analysis

Datasets and Inputs: Arvato is graceful enough to provide us with real demographic and mail order camping dataset to work on the project use case. There are four data files provided for the use case. Pandas Profiling API was used to run a high-level diagnostic on the dataset and have listed my findings below

- Udacity_AZDIAS_052018.csv: Demographics data for the general population of Germany; 891 211 persons (rows) x 366 features (columns).



Dataset info

Number of variables	366
Number of observations	5
Missing cells	284 (15.5%)
Duplicate rows	0 (0.0%)
Total size in memory	14.4 KiB
Average record size in memory	2.9 KiB

Variables types

Numeric	1
Categorical	24
Boolean	14
Date	0
URL	0
Text (Unique)	0
Rejected	327
Unsupported	0

Warnings

AKT_DAT_KL has 1 (20.0%) missing values Missing

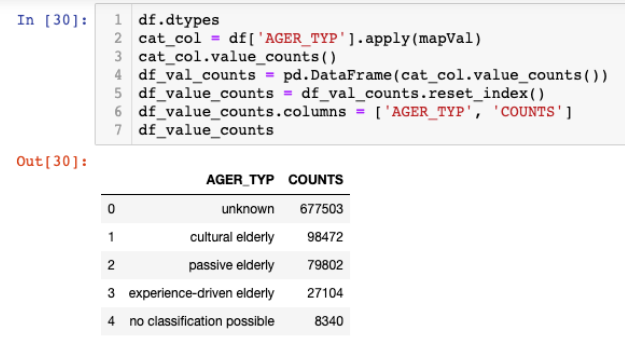
ALTER_HH has 1 (20.0%) missing values Missing

ALTER_KIND1 has constant value "nan" Rejected

ALTER_KIND2 has constant value "nan" Rejected

ALTER_KIND3 has constant value "nan" Rejected

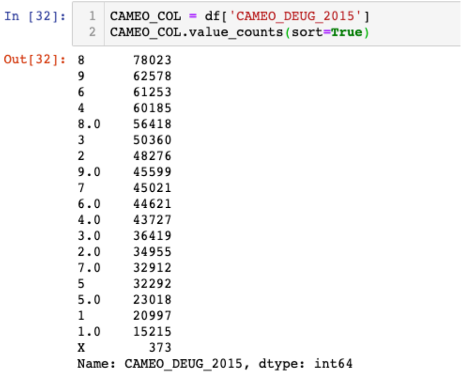
Sampling few categorical columns suggest the data is not balanced and requires cleanup as we could see the distribution in the case of people age and income categories in the general population demographics data. We may need to handle the imbalance by over/under sampling some categorical data based on model performance.



In [30]:
```
df.dtypes
cat_col = df['AGER_TYP'].apply(mapVal)
cat_col.value_counts()
df_val_counts = pd.DataFrame(cat_col.value_counts())
df_value_counts = df_val_counts.reset_index()
df_value_counts.columns = ['AGER_TYP', 'COUNTS']
df_value_counts
```

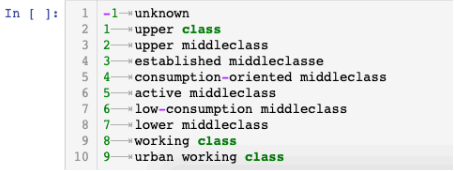
Out[30]:

	AGER_TYP	COUNTS
0	unknown	677503
1	cultural elderly	98472
2	passive elderly	79802
3	experience-driven elderly	27104
4	no classification possible	8340



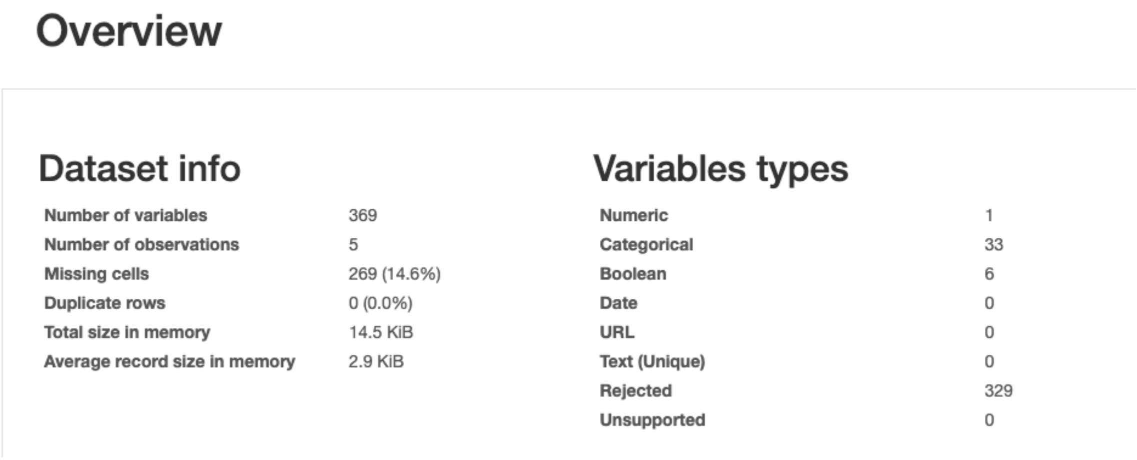
In [32]:
```
CAMEO_COL = df['CAMEO_DEUG_2015']
CAMEO_COL.value_counts(sort=True)
```

Out[32]:
```
8      78023
9      62578
6      61253
4      60185
8.0    56418
3      50360
2      48276
9.0    45599
7      45021
6.0    44621
4.0    43727
3.0    36419
2.0    34955
7.0    32912
5      32292
5.0    23018
1      20997
1.0    15215
X        373
Name: CAMEO_DEUG_2015, dtype: int64
```



In [ ]:
```
-1	unknown
1	upper class
2	upper middleclass
3	established middleclasse
4	consumption-oriented middleclass
5	active middleclass
6	low-consumption middleclass
7	lower middleclass
8	working class
9	urban working class
```

- Udacity_CUSTOMERS_052018.csv: Demographics data for customers of a mail-order company; 191 652 persons (rows) x 369 features (columns).



Overview

Dataset info

Number of variables	369
Number of observations	5
Missing cells	269 (14.6%)
Duplicate rows	0 (0.0%)
Total size in memory	14.5 KiB
Average record size in memory	2.9 KiB

Variables types

Numeric	1
Categorical	33
Boolean	6
Date	0
URL	0
Text (Unique)	0
Rejected	329
Unsupported	0

- 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).

Based on the high-level analysis the dataset is not clean and has a good number of NaNs and missing values. We will use Pandas and Numpy APIs to fill in/remove NaN's by each column basis.

Solution Statement: The proposed solution will use the given dataset and potentially identify demographics of the customer base for the company and will use machine learning algorithms and other techniques to identify potential customers for the company from the mail order marketing campaign.

The following High Level Steps will be taken:

- Analyze and Clean-up data by handling NaNs, outliers and other anomalies
- Feature engineering to identify valid inputs
- Part 1: Unsupervised Learning Section
  - For customer segmentation, K-Means algorithm will be used to identify various customer base
  - Elbow and Silhouette methods will be used to identify and validate optimum clusters (K)
- Part 2: Supervised Learning Section
  - For Predicting potential customers, Logistic Regression and XGBoost algorithms will be used
  - GridSearchCV APIs will be used for hyper parameter tuning
  - Accuracy and ROC-AOC score metrics will be used to validate the performance of the algorithms

Benchmark Model: The models built for the use case will use the demographics and customer data as a ground truth for evaluation and bench mark of the model and analysis. Given the dataset size, Logistic algorithm will be used as the benchmark model to evaluate and tweak the solution.

Evaluation Metrics: We will use the real user data provided by Arvato for the project. Numpy and Pandas libraries will be used to perform data cleanup, EDA and pre-processing activities. PCA will be used to reduce the dimensionality of the data identifying as part of the feature engineering to identify ideal input features.

K-Means algorithm will be applied on the processed data to identity different population clusters, while fine tuning the algorithm parameters. The second part of the project involves creating a supervised model to predict potential customers for the company using the given dataset. Binary classification will be used to classify the outcome. Logistic Regression, XGBoost algorithms will be used.

As we are trying to predict the potential customers for the ad campaign, we can use the metrics like Accuracy and ROC-AOC curve to evaluate our model. We will take a liberal approach by using Accuracy as a metric rather than Precision as our goal is to identify potential candidates. Along with accuracy we will use the ROC-AOC score as another metric which provides outcome as a probability which will helpful to fine tune the model by tweaking the range.

Project Design: We will follow the high-level process flow below for processing data, identifying population cluster and creating a supervised model.

High Level Process Flow:

- Collect Data for the use case
- Clean up and Pre-process the demographics (General & Customer) data
- Feature Engineering

- PCA, reduce input features to identify effective input features
- Split the dataset into test, train and validation
- Generate the shoulder curve to identify the cluster size
- Apply K-means algorithm to the processed dataset to identify clusters
- Identify potential customer base clusters by comparing general vs customer cluster
- Clean up and Pre-process the given train and test customer data
- Identify Input Features
- Create a model
- Train the model
- Test the model with test customer data
- Validate output data and model performance
- Infer insights from the final predicted output