Steps in the analysis:

0 Get to Know the Data
1 Customer Segmentation Report
2 Supervised Learning Model
3 Kaggle Competition

Datasets used:	columns	
AZDIAS	366	demographic data
CUSTOMERS	369	demographic data + customer_group + online_purchase + product_group
MAILOUT TEST	366	demographic data
MAILOUT TRAIN	367	demographic data + target

1. Customer Segmentation Report

CUSTOMERS
(demographics data for customers)
369 x 191k

Preprocessing → Univariate Analysis → Feature Scaling → Transform CUSTOMERS data with AZDIAS_PCA object

0. Get to Know the Data

AZDIAS
(Germany demographic data)
366 x 891k

Check integrity adn explore metadata → Preprocessing → Univariate Analysis → Feature Scaling → Build Functions

Principal Component Analysis (150 components) → Select optimal K value with AZDIAS_PCA data → Fit K-means with optimal K and AZDIAS_PCA data → Get AZDIAS_PCA clusters and CUSTOMERS_PCA clusters → Get ratio between customers and general population in each cluster → Analyze feature weights in the clusters with the highest percentage of customers.

2. Supervised Learning Model

MAILOUT_TRAIN
(demographics data for customers with target)
367 x 43k

Preprocessing and Univariate Analysis → Split MAILOUT data in: train (70%) / validation (30%) → Set a benchmark: fitting a unbalanced simple estimator with unscaled data (LogisticRegression) → Try with different classification estimators with scaled data (LR, DT, RFC, GBC, ABC, XGBC..) and compare AUCRoc → Select the estimator that give the best result → Fit hyperparameter tunning → Select the best model

3. Kaggle Competition

MAILOUT_TEST
(demographics data for customers without target)
366 x 43k

Preprocessing and Univariate Analysis → Predict with MAILOUT data for predict