The Capstone project proposal for Arvato Financial Solutions project

My capstones project's domain background is a machine learning project that is able to differentiate and highlight possible customers for a financing company based on Customer Segmentation Machine Learning model and the input data describing a population of candidates using multiple features.

Arvato Financial Solutions the company which provided the dataset for this project is looking for possible candidates that will likely be seeking the services provided by the company. To reach the target population of customers Arvato cannot and would not be efficient to reach all the population of the region to offer their services.

- Every dataset consists of numerous features describing every entry or person in the respective datasets.
  There are four data files associated with this project:
- `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).

- Some exploratory data analysis using the `df.describe()` method show us the descriptive statistics relating to the dataframes showing parameters like the mean, median and the standard deviation for every column or feature.

```
In [11]: df_azdias.describe()
```

Out[11]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_
count	8.912210e+05	891221.000000	817722.000000	817722.000000	81058.000000	29499.000000	6170.000000	1205.000000	628274.000000	
mean	6.372630e+05	-0.358435	4.421928	10.864126	11.745392	13.402658	14.476013	15.089627	13.700717	
std	2.572735e+05	1.198724	3.638805	7.639683	4.097660	3.243300	2.712427	2.452932	5.079849	
min	1.916530e+05	-1.000000	1.000000	0.000000	2.000000	2.000000	4.000000	7.000000	0.000000	
25%	4.144580e+05	-1.000000	1.000000	0.000000	8.000000	11.000000	13.000000	14.000000	11.000000	
50%	6.372630e+05	-1.000000	3.000000	13.000000	12.000000	14.000000	15.000000	15.000000	14.000000	
75%	8.600680e+05	-1.000000	9.000000	17.000000	15.000000	16.000000	17.000000	17.000000	17.000000	
max	1.082873e+06	3.000000	9.000000	21.000000	18.000000	18.000000	18.000000	18.000000	25.000000	

8 rows × 360 columns

```
In [12]: df_customers.describe()
```

Out[12]:

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ
count	191652.000000	191652.000000	145056.000000	145056.000000	11766.000000	5100.000000	1275.000000	236.000000	139810.000000	
mean	95826.500000	0.344359	1.747525	11.352009	12.337243	13.672353	14.647059	15.377119	10.331579	
std	55325.311233	1.391672	1.966334	6.275026	4.006050	3.243335	2.753787	2.307653	4.134828	
min	1.000000	-1.000000	1.000000	0.000000	2.000000	2.000000	5.000000	8.000000	0.000000	
25%	47913.750000	-1.000000	1.000000	8.000000	9.000000	11.000000	13.000000	14.000000	9.000000	
50%	95826.500000	0.000000	1.000000	11.000000	13.000000	14.000000	15.000000	16.000000	10.000000	
75%	143739.250000	2.000000	1.000000	16.000000	16.000000	16.000000	17.000000	17.000000	13.000000	
max	191652.000000	3.000000	9.000000	21.000000	18.000000	18.000000	18.000000	18.000000	25.000000	

8 rows × 361 columns

- We use `df.<column name>.unique()` to show the values present in that column.

This project goal is to provide, using customer segmentation techniques the recommended population that are likely be future customers. This can be achieved using the machine learning algorithms and models, applied to the dataset provided. In the project the current clients of the company will be compared with the Germany population data to highlight key features that may predict the future customers.

The model that will be applied in this project relies on customer segmentation model to be able to target the suitable, prospective customers who would use the financial services provided by Arvato Finsncial solutions.
There are several benefits of implementing customer segmentation including informing marketing strategy, promotional strategy, product development, budget management, and delivering relevant content to your customers or prospective customers.

Example for customer segmentation model paper:
“Customer Segmentation Using Unsupervised Learning on Daily Energy Load Profiles” by J. du Toit, R. Davimes, A. Mohamed, K. Patel, and J. M. Nye
http://www.jait.us/uploadfile/2016/0505/20160505105403530.pdf

The evaluation metric that will be used to measure the clustering model is the Adjusted Rand Index (ARI)
where the pairs of every 2 datapoints are compared between the predicted clusters and

the true clusters.

Project design:

1. Explore and preprocess data:
    1. data preprocessing: for example:
        - managing missing data
        - normalization for wide range data
        - categorical data transformation into numerical features
    2. choosing key features and values for some of these features from the current customers using Principal component analysis (PCA) to create principal component features that cover a high variance of the data and the most significant features with the highest weights, and most importantly reducing the dimensionality of the data
2. Modeling:
    1. using K-Means unsupervised learning to create a model, to create customer clusters
    2. training the model
3. Evaluate the model:
    1. apply the created model on the population of Germany dataset, to predict future customers using the Adjusted Rand Index (ARI)

student submission notes Capstone proposal: https://review.udacity.com/#!/reviews/2102097