ARVATO CUSTOMER SEGMENTATION AND CAMPAIGN CONVERSION

Customer Segmentation

Md Ahsanur Rashid
Udacity ML NanoDegree

Data Exploration

Exploring the data, I looked at the data sets and different columns & the values they hold. I also looked at the distribution of NaN values.

I had two datasets for this part of the project:

a. Azdias dataset

   This is the dataset for the general population in Germany. There were data for 891221 persons, each with 366 attributes.

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHAL
0	910215	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
1	910220	NaN	9.0	NaN	NaN	NaN	NaN	NaN	21.0	
2	910225	NaN	9.0	17.0	NaN	NaN	NaN	NaN	17.0	
3	910226	2.0	1.0	13.0	NaN	NaN	NaN	NaN	13.0	
4	910241	NaN	1.0	20.0	NaN	NaN	NaN	NaN	14.0	
5	910244	3.0	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
6	910248	NaN	9.0	NaN	NaN	NaN	NaN	NaN	NaN	
7	910261	NaN	1.0	14.0	NaN	NaN	NaN	NaN	14.0	
8	645145	NaN	9.0	16.0	NaN	NaN	NaN	NaN	16.0	
9	645153	NaN	5.0	17.0	NaN	NaN	NaN	NaN	17.0	
10	645165	0.0	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
11	645169	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
12	612558	NaN	5.0	21.0	NaN	NaN	NaN	NaN	14.0	
13	612561	NaN	8.0	20.0	NaN	NaN	NaN	NaN	20.0	
14	612565	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	

Table 1: Brief view of Azdias dataset

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FE
count	8.912210e+05	213718.000000	817722.000000	580954.000000	81058.000000	29499.000000	6170.000000	1205.000000	628274.0000
mean	6.372630e+05	1.675376	4.421928	15.291805	11.745392	13.402658	14.476013	15.089627	13.7007
std	2.572735e+05	0.742250	3.638805	3.800536	4.097660	3.243300	2.712427	2.452932	5.0798
min	1.916530e+05	0.000000	1.000000	1.000000	2.000000	2.000000	4.000000	7.000000	0.0000
25%	4.144580e+05	1.000000	1.000000	13.000000	8.000000	11.000000	13.000000	14.000000	11.0000
50%	6.372630e+05	2.000000	3.000000	16.000000	12.000000	14.000000	15.000000	15.000000	14.0000
75%	8.600680e+05	2.000000	9.000000	18.000000	15.000000	16.000000	17.000000	17.000000	17.0000
max	1.082873e+06	3.000000	9.000000	21.000000	18.000000	18.000000	18.000000	18.000000	25.0000

Table 2: Azdias dataset descriptive statistics

b. Customers dataset

This is the dataset for Avrato customers. There were data for 191652 customers, each with 369 attributes.

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FEIN	ANZ_HAUSHAL
0	9626	2.0	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
1	9628	NaN	9.0	11.0	NaN	NaN	NaN	NaN	NaN	
2	143872	NaN	1.0	6.0	NaN	NaN	NaN	NaN	0.0	
3	143873	1.0	1.0	8.0	NaN	NaN	NaN	NaN	8.0	
4	143874	NaN	1.0	20.0	NaN	NaN	NaN	NaN	14.0	
5	143888	1.0	1.0	11.0	NaN	NaN	NaN	NaN	10.0	
6	143904	2.0	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
7	143910	1.0	1.0	10.0	NaN	NaN	NaN	NaN	9.0	
8	102160	2.0	3.0	5.0	NaN	NaN	NaN	NaN	4.0	
9	102173	1.0	1.0	20.0	NaN	NaN	NaN	NaN	13.0	
10	102184	NaN	7.0	14.0	NaN	NaN	NaN	NaN	14.0	
11	102185	1.0	1.0	10.0	NaN	NaN	NaN	NaN	10.0	
12	102227	NaN	1.0	21.0	NaN	NaN	NaN	NaN	14.0	
13	102230	NaN	1.0	15.0	8.0	NaN	NaN	NaN	14.0	
14	102239	2.0	1.0	6.0	NaN	NaN	NaN	NaN	6.0	

Table 3: Brief view of Customers dataset

	LNR	AGER_TYP	AKT_DAT_KL	ALTER_HH	ALTER_KIND1	ALTER_KIND2	ALTER_KIND3	ALTER_KIND4	ALTERSKATEGORIE_FE
count	191652.000000	99545.000000	145056.000000	122905.000000	11766.000000	5100.000000	1275.000000	236.000000	139810.0000
mean	95826.500000	1.588267	1.747525	13.397966	12.337243	13.672353	14.647059	15.377119	10.3315
std	55325.311233	0.713589	1.966334	4.365868	4.006050	3.243335	2.753787	2.307653	4.1348
min	1.000000	0.000000	1.000000	2.000000	2.000000	2.000000	5.000000	8.000000	0.0000
25%	47913.750000	1.000000	1.000000	10.000000	9.000000	11.000000	13.000000	14.000000	9.0000
50%	95826.500000	2.000000	1.000000	13.000000	13.000000	14.000000	15.000000	16.000000	10.0000
75%	143739.250000	2.000000	1.000000	17.000000	16.000000	16.000000	17.000000	17.000000	13.0000
max	191652.000000	3.000000	9.000000	21.000000	18.000000	18.000000	18.000000	18.000000	25.0000

Table 4: Azdias dataset descriptive statistics

ANDREDE_KZ or gender column caught my attention since it seemed like there was either male or unknown in the values. No females. In the dataset:

- -1, 0 : unknown
- 1 : male
- 2 : female

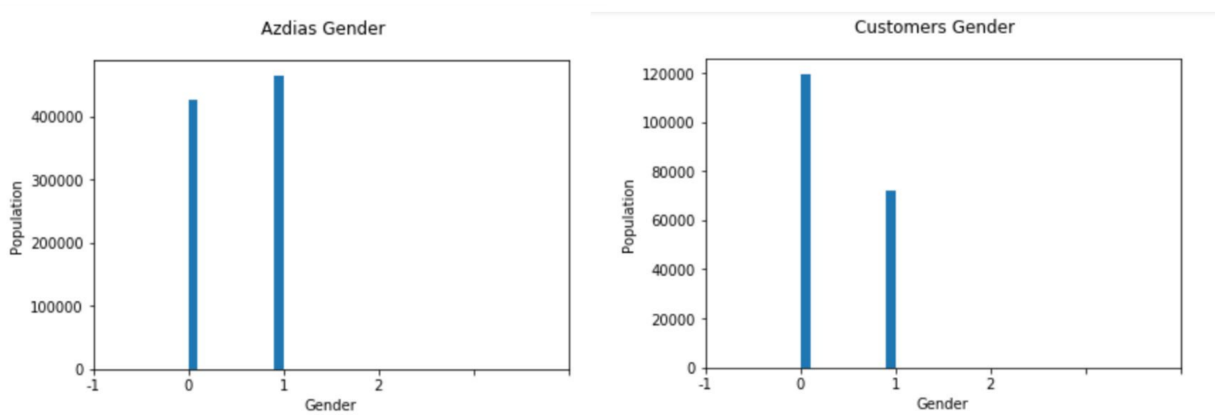

Fig 1: Value distribution for ANDREDE_KZ (gender) in Azdias dataset and Customers dataset

Another interesting one was LNR which had a different value for every row. I ended up dropping both LNR and ANDREDE_KZ columns.

```
0 ColumnName:  LNR UniqueLength 191652
Unique Values:  [  9626   9628 143872 ... 148813 148852 148883]

18 ColumnName:  CAMEO_DEUG_2015 UniqueLength 11
Unique Values:  ['1' nan '5' '4' '7' '3' '9' '2' '6' '8' 'X']
19 ColumnName:  CAMEO_INTL_2015 UniqueLength 23
Unique Values:  ['13' nan '34' '24' '41' '23' '15' '55' '14' '22' '43' '51
' '33' '25' '44' '54' '32' '12' '35' '31' '45' '52' 'XX']

367 ColumnName:  ANREDE_KZ UniqueLength 2
Unique Values:  [1 2]
```

Fig 2: Different possible and total values for the columns LNR, CAMEO_DEUG_2015, CAMEO_INTL_2015 and ANDREDE_KZ

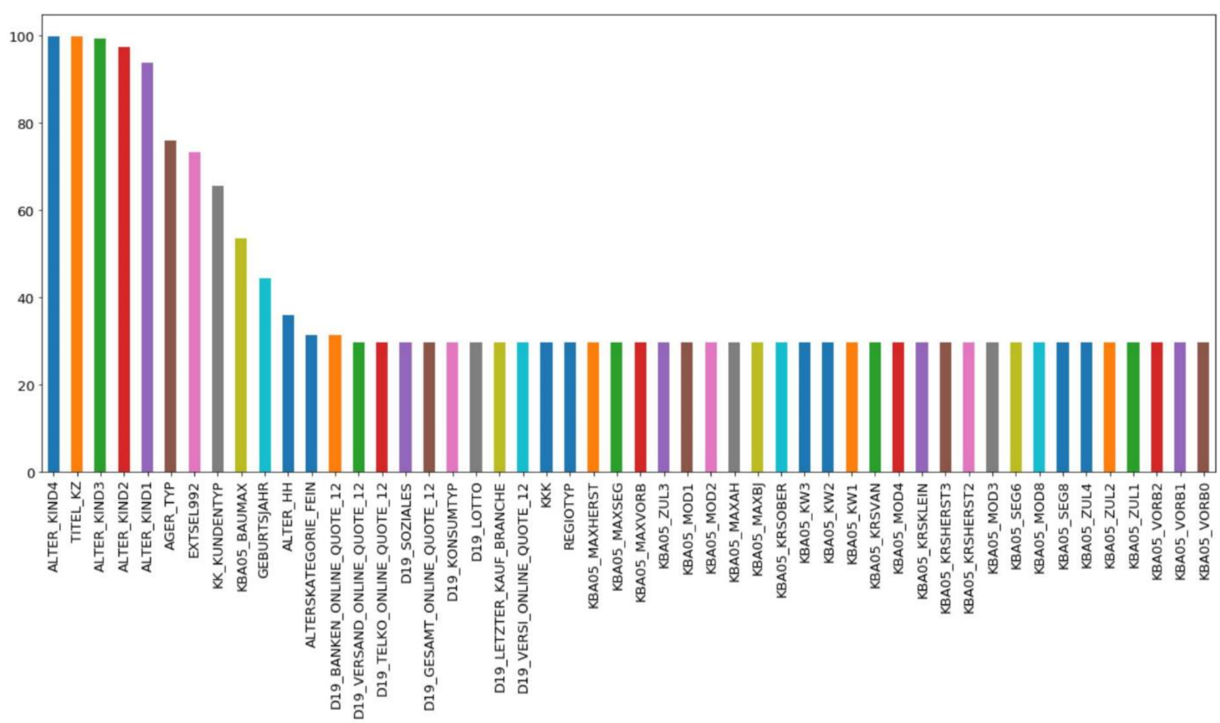

Fig 3: Azdias dataset NaN distribution (percentage of NaN values for each column vs all columns)

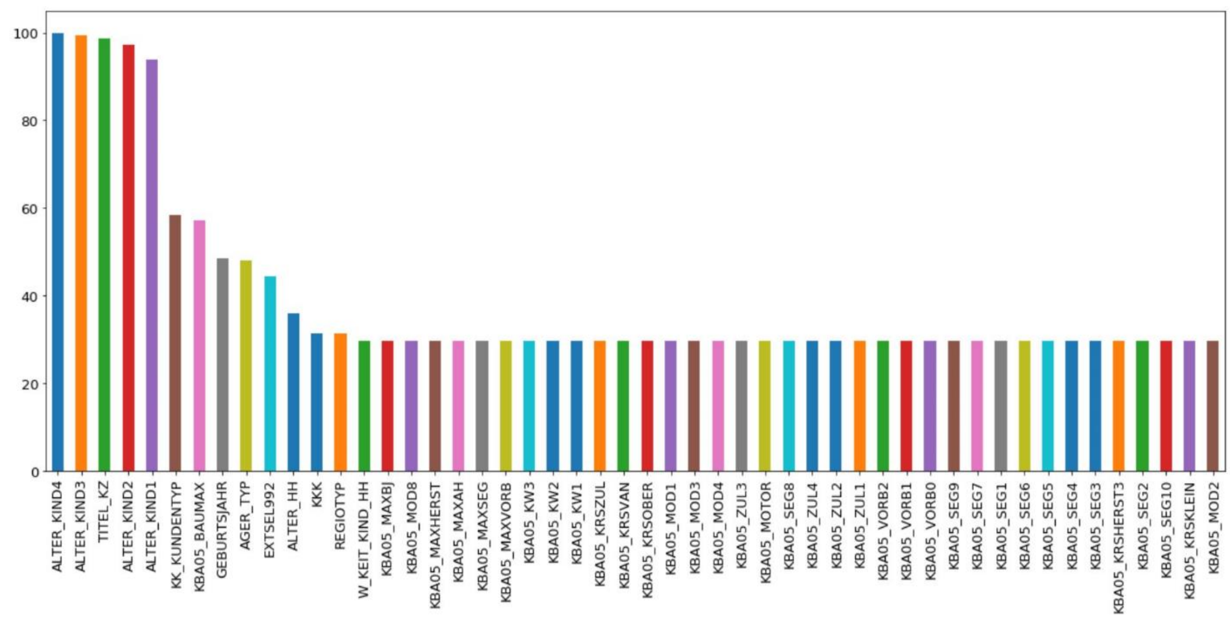

Fig 4: Customers dataset NaN distribution (percentage of NaN values for each column vs all columns)

Data Cleanup

Data cleanup is the most important & time-consuming part of any ML work and this was no different. I had to go through the data, explore it and figure out different techniques for cleanup. Clean up was done on both Azdias and Customers data set. I did the following, in order, to clean the data before doing PCA analysis and KMeans clustering:

- Marked some data fields as NaN

  Based on the provided attributes and probable values documented, many fields had -1, 0, 9 etc. for unknown values and were marked down as NaN for further processing later.

- Cleaned up mixed type (float and string) columns
  - CAMEO_INTL_2015
  - CAMEO_DEUG_2015

  The values ‘X’ and ‘XX’ were replaced by NaN.

- Removed differences between the Customers and Azdias data set columns
  - CUSTOMER_GROUP
  - ONLINE_PURCHASE
  - PRODUCT_GROUP

  Removed the extra columns from the Customers data set.

- Dropped columns where 33.33% of the data were NaN
  - ALTER_KIND4
  - ALTER_KIND3
  - TITEL_KZ
  - ALTER_KIND2
  - ALTER_KIND1
  - KK_KUNDENTYP

    - KBA05_BAUMAX
    - GEBURTSJAHR
    - AGER_TYP
    - EXTSEL992
    - ALTER_HH

  This was decided by looking at the percentage of missing data for each column. Removing the top 11 columns with the most missing data. This was selected based on how many columns need to be dropped and what percentage of missing data seemed feasible. This also happened to overall drop the same columns for both Azdias and Customers data set.

- Removed columns with mostly unique values
    - LNR

  Looked at the different types of values for each column and noticed LNR column had unique values for each row. Deleted the column from both Customers and Azdias data sets.

- Dropped columns where data is clearly skewed
    - ANREDE_KZ

  While exploring the data set it was noticed that the values for the column were either male or unknown. Dropped from both data sets.

- Dropped undocumented columns
    - AKT_DAT_KL
    - ANZ_STATISTISCHE_HAUSHALTE
    - ARBEIT
    - CJT_KATALOGNUTZER
    - CJT_TYP_1
    - CJT_TYP_2
    - CJT_TYP_3

- CJT_TYP_4
- CJT_TYP_5
- CJT_TYP_6
- D19_KONSUMTYP_MAX
- D19_LETZTER_KAUF_BRANCHE
- D19_SOZIALES
- D19_TELKO_ONLINE_QUOTE_12
- D19_VERSI_DATUM
- D19_VERSI_OFFLINE_DATUM
- D19_VERSI_ONLINE_DATUM
- D19_VERSI_ONLINE_QUOTE_12
- DSL_FLAG
- EINGEFUEGT_AM
- EINGEZOGENAM_HH_JAHR
- EXTSEL1992
- FIRMENDICHTE
- GEMEINDETYP
- HH_DELTA_FLAG
- KBA13_ANTG1
- KBA13_ANTG2
- KBA13_ANTG3
- KBA13_ANTG4
- KBA13_BAUMAX
- KBA13_GBZ
- KBA13_HHZ
- KBA13_KMH_210
- KK_KUNDENTYP
- KOMBIALTER
- KONSUMZELLE
- MOBI_RASTER
- RT_KEIN_ANREIZ

    - RT_SCHNAEPPCHEN
    - RT_UEBERGROESSE
    - SOHO_KZ
    - STRUKTURTYP
    - UMFELD_ALT
    - UMFELD_JUNG
    - UNGLEICHENN_FLAG
    - VERDICHTUNGSRAUM
    - VHA
    - VHN
    - VK_DHT4A
    - VK_DISTANZ
    - VK_ZG11

  Some columns from the list above were dropped already as part of other dropping criteria. Some undocumented columns were not removed, since they were easy to understand and seemed important to keep:

    - ANZ_KINDER

- Dropped columns with too many values
    - CAMEO_DEU_2015
    - D19_LETZTER_KAUF_BRANCHE

  These has 44 and 36 types of values respectively.

- Dropped columns deemed unnecessary
    - MIN_GEBAEUDEJAHR

  Since this represents the year the building was first mentioned in the database, it seemed unnecessary data to analyze.

- Dropped additional columns for Grob vs Fein scenarios.

There were 4 pairs of columns that had remarkably similar data. One was the FEIN or Fine column and the other one was the GROB or rough column. The fine column had more possible values or buckets and more finely sorted the data. Whereas the rough column had bigger buckets or less number of probable values. Since the fine columns had quite a lot of probable values, decided to drop FEIN columns and keep GROB columns.

- ALTERSKATEGORIE_FEIN
- LP_FAMILIE_FEIN
- LP_LEBENSPHASE_FEIN
- LP_STATUS_FEIN

- Dropped all rows with 30% or more NaN values.

  51,281 rows were dropped for Customers dataset and 105,800 rows were dropped for Azdias dataset.

- Binary encoded OST_WEST_KZ and VERS_TYP columns

Column Name	Old Value	New Value
OST_WEST_KZ	W	1
	O	0
VERS_TYP	1	1
	2	0

Table 5: Value mapping for binary encoding of OST_WEST_KZ and VERS_TYP

- Replaced NaN values with median or most frequently used values
  - For binary columns, used most frequently used value to replace NaNs
  - For all other columns used median value to replace NaNs

- Split some columns into multiple columns
  - CAMEO_INTL_2015
  - PLZ8_BAUMAX

- PRAEGENDE_JUGENDJAHRE
- WOHNLAGE

Except for the last two, all of them were dropped and new ones created to replace them.

Old Column Name	Old Value	Meaning	New Column Name	New Value	Meaning
CAMEO_INTL_2015	11	Wealthy Households-Pre-Family Couples & Singles	CI2_Family Type	1	Pre Family Couples & Singles
	12	Wealthy Households-Young Couples With Children		2	Young Couples with Children
	13	Wealthy Households-Families With School Age Children		3	Families with school age children
	14	Wealthy Households-Older Families & Mature Couples		4	Older families & Mature couples
	15	Wealthy Households-Elders In Retirement		5	Elders in retirement
	21	Prosperous Households-Pre-	CI2_Wealth Type	1	Wealthy Households

		Family Couples & Singles			
	22	Prosperous Households-Young Couples With Children		2	Prosperous Households
	23	Prosperous Households-Families With School Age Children		3	Comfortable Households
	24	Prosperous Households-Older Families & Mature Couples		4	Less Affluent Households
	25	Prosperous Households-Elders In Retirement		5	Poorer Households
	31	Comfortable Households-Pre-Family Couples & Singles			
	32	Comfortable Households-Young Couples With Children			
	33	Comfortable Households-Families With School Age Children			

	34	Comfortable Households-Older Families & Mature Couples	
	35	Comfortable Households-Elders In Retirement	
	41	Less Affluent Households-Pre-Family Couples & Singles	
	42	Less Affluent Households-Young Couples With Children	
	43	Less Affluent Households-Families With School Age Children	
	44	Less Affluent Households-Older Families & Mature Couples	
	45	Less Affluent Households-Elders In Retirement	
	51	Poorer Households-Pre-Family Couples & Singles	

	52	Poorer Households-Young Couples With Children			
	53	Poorer Households-Families With School Age Children			
	54	Poorer Households-Older Families & Mature Couples			
	55	Poorer Households-Elders In Retirement			
PLZ8_BAUMAX	1	mainly 1-2 family homes	PB_Family	0	Not mainly family home
	2	mainly 3-5 family homes		1	Mainly family home
	3	mainly 6-10 family homes	PB_Business	0	Not maintly business building
	4	mainly >10 family homes		1	Mainly business building
	5	mainly business building			
PRAEGENDE_JUGENDJAHRE	1	40's - war years (Mainstream, O+W)	PJ_Movement	0	Mainstream
	2	40's - reconstruction years (Avantgarde, O+W)		1	Avantgarde
	3	50's - economic miracle (Mainstream, O+W)	PJ_Generation	1	40's

	4	50’s - milk bar / Individualisation (Avantgarde, O+W)		2	50’s
	5	60’s - economic miracle (Mainstream, O+W)		3	60’s
	6	60’s - generation 68 / student protestors (Avantgarde, W)		4	70’s
	7	60’s - opponents to the building of the Wall (Avantgarde, O)		5	80’s
	8	70’s - family orientation (Mainstream, O+W)		6	90’s
	9	70’s - peace movement (Avantgarde, O+W)			
	10	80’s - Generation Golf (Mainstream, W)			
	11	80’s - ecological awareness (Avantgarde, W)			
	12	80’s - FDJ / communist party youth organisation (Mainstream, O)			

	13	80’s - Swords into ploughshares (Avantgarde, O)			
	14	90’s - digital media kids (Mainstream, O+W)			
	15	90’s - ecological awareness (Avantgarde, O+W)			
WOHNLAGE	1	very good neighbourhood	WL_Rural	0	Not rural
	2	good neighbourhood		1	Rural
	3	average neighbourhood			
	4	poor neighbourhood			
	5	very poor neighbourhood			
	7	rural neighbourhood			
	8	new building in rural neighbourhood			

Table 6: Value mapping of newly split columns

- Converted to integer values

  Converted all the values to integer values for both the data sets.

- Removed dataset outliers

  For all rows, for each columns with non-binary value, values outside a +/- 6 difference with standard deviation were dropped.

17,353 rows were dropped from Customers data set and 82,904 rows were dropped from Azdias data set.

- Scaled dataset values

  Performed standard scaler on both Customers and Azdias data sets.

Post Cleanup Data Processing

- **Principal Component Analysis**

Principal Component Analysis or PCA is a dimensionality reduction method that is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set. [1] PCA is done either by singular value decomposition of a design matrix or by calculating the correlation or covariance matrix and performing eigenvalue decomposition on that. [2]

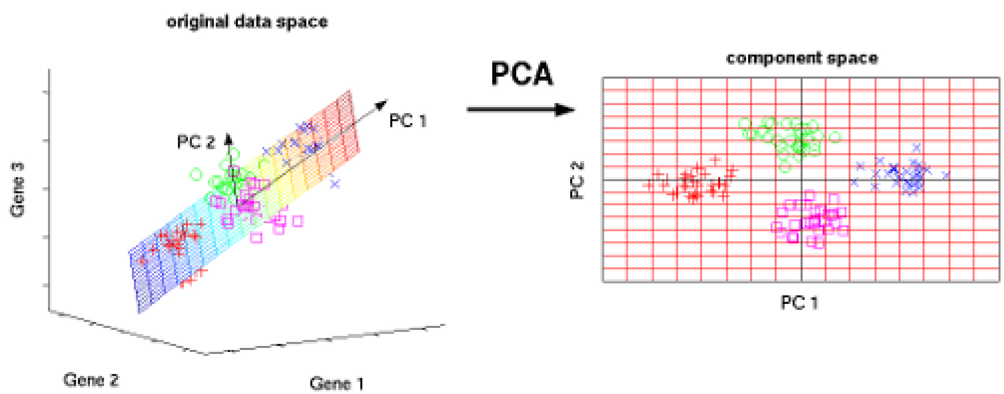

Fig 5: Principal Component Analysis [4]

I performed PCA on the cleaned Azdias data set and calculated the cumulative variance for the principal components such that we achieve a certain variance % (in this case 95%). The same principal components were used to transform the Customers data set.

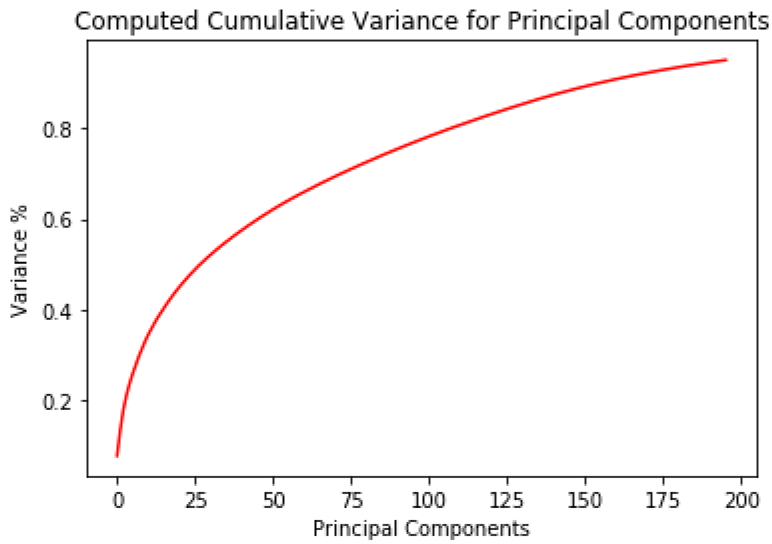

Fig 6: Computed Cumulative Variance for Principal Components

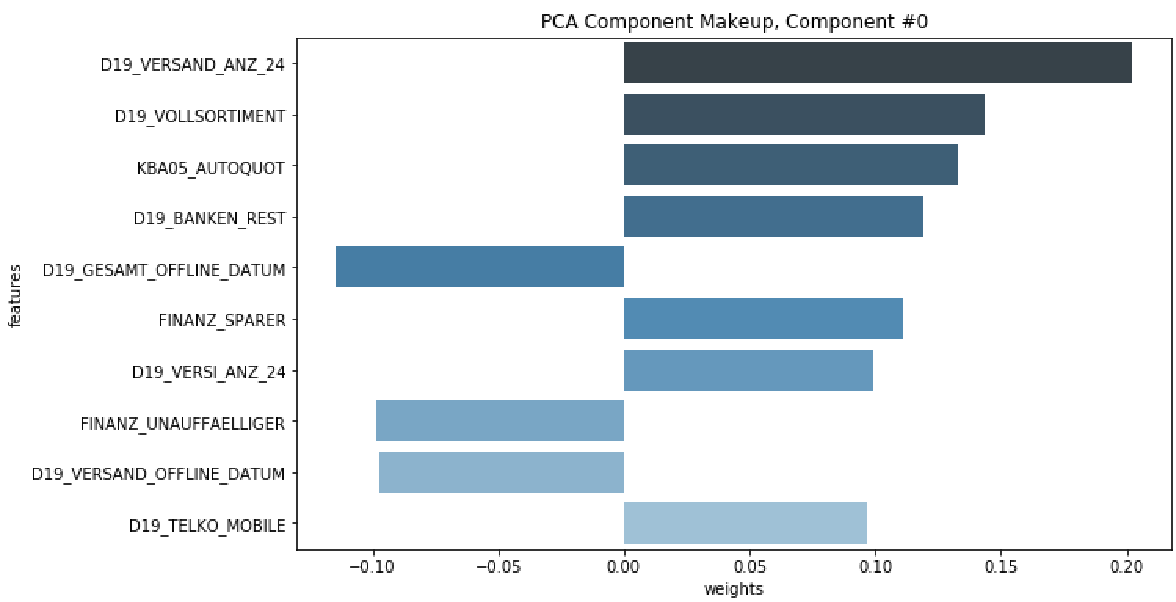

Fig 7: PCA Component Makeup, Component #0

- K-Means Clustering

K-means clustering is an iterative algorithm that aims to partition the dataset into k pre-defined, distinct, non-overlapping clusters where each data point belongs to only one group. [3] It also tries to make the intra-cluster data points as similar as possible while keeping the clusters as different (far) as possible.

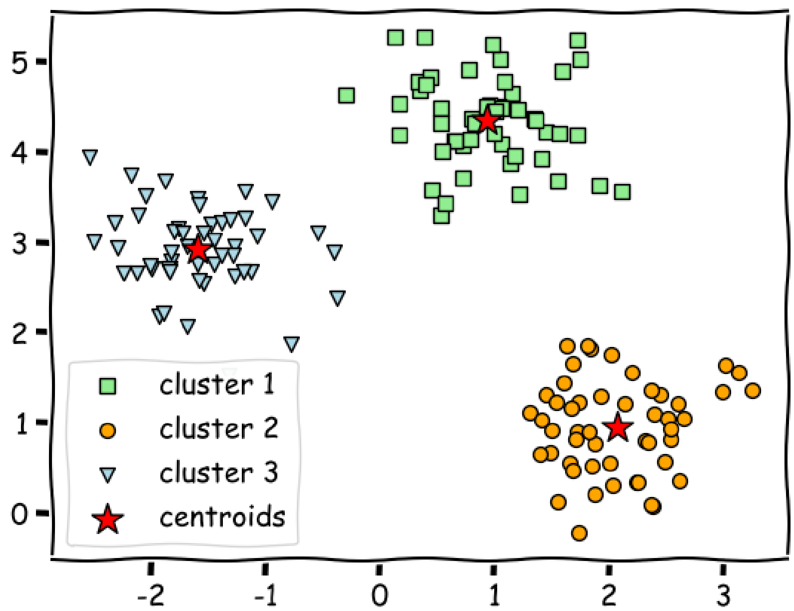

Fig 8: K-Means Clustering [5]

I did k-means clustering for the Azdias data set using different cluster sizes from 1 to 20. Plotting the cluster size against the model score, I got the elbow chart which helped me choose the cluster size of 10.

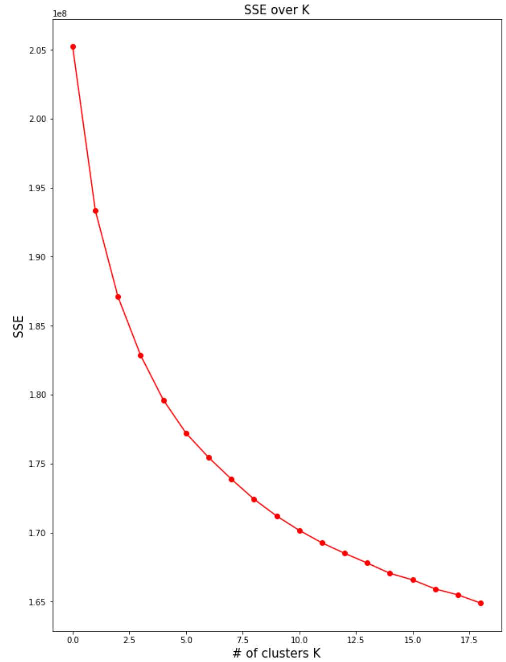

Fig 9: Elbow Chart

	Azdias	Customers	% of processed Azdias	% of processed Customers
0	71778	1819	10.031095	1.439253
1	90890	15403	12.702028	12.187364
2	47323	2806	6.613468	2.220200
3	66218	5344	9.254076	4.228350
4	64821	27365	9.058842	21.652095
5	28607	3336	3.997876	2.639554
6	89864	1127	12.558643	0.891720
7	53235	2453	7.439680	1.940895
8	103936	31124	14.525229	24.626340
9	98883	35608	13.819064	28.174230

Fig 10: Final cluster list

- From K-Means Clustering to Interpretable Data

  Given the customer and population distribution from Fig 11 (below), I chose a couple of clusters to investigate. For these clusters, I got the cluster centers. Then, I inverse transformed the principal component of the cluster centers to the original components. Then I performed another inverse transformation using the standard scaler. These mapped the cluster center in the k-means clustering algorithm output to the the original column values in the cleaned dataset.

Analysis

Fig 11: Azdias vs Customers (% of total) per cluster

Some of the provided data is not defined and some are not well understood. Since this is just the centroid for the clusters the actual results will not match exactly but have a small delta difference. Additionally, some of the attributes may not be in agreement with each other. Mostly there is a sort of loose definition for what group of people make up a higher percentage of customers and what group of people do not.

I looked at cluster 9 where the % of customers were much larger than the % of general population. I also looked at cluster 6 where % of customers were much smaller than the % of general population.

People in cluster 9, with high % of customers, tend to have the following characteristics:

- Older, smaller family
    - Smaller family
    - Mature couple
    - No children
    - Relatively older (46-60 years)

    - Golden ager
    - Low mobility
    - Have pet(s)
    - Takes care of themselves (Luxury clothing, education, gardening, food, dietary supplements, wine, medicine, leisure, travel, shoes etc)
    - 3 HH/Building
- Financially prosperous
    - High earner
    - Average investor
    - Average money saver
    - Consumption oriented
- Mail Order history:
    - Has quite a bit of MO history
    - Double buy (12 months) of further mail order
    - Actuality of last transaction for MO total is high
    - Actuality of last transaction for MO online increased
    - % of online transactions within all in MO is 80%
- Cars
    - Avg cars
    - Not much of a difference than the other cluster
    - High % of car in HH
    - Low % of upper- & middle-class cars
- Shopping habits
    - Advertisement interested online shopper
    - Gourmet shopper
    - Multi and double buys of luxury clothing, education, gardening, food, dietary supplements, wine, medicine, leisure, travel, shoes and the like
- Social
    - Average affinity for: Family, Materials, Tradition, Religion, Rational, Tradition, Lust

People in cluster 6, with low % of customers, tend to have the following characteristics:

- Young family or single
    - HH size of ~1
    - Young couple
    - Relatively younger (80's, 30-45 years)
    - Homeland connected vacationist
    - 6 HH/Building
- Financially comfortable
    - Active Middle Class
    - Comfortable
    - Low income
    - Low investment earner
    - Low money saver
- Mail Order history
    - Not much of a history with Mail Orders
- Cars
    - Average cars
    - Not that much of a difference than the other cluster
    - Average % of car in HH
    - Very low % of upper- & middle-class cars
- Shopping habits
    - Advertisement interested online shopper
    - Stressed shopper
    - Multi buys of Books/CDs, technology, mobile, further clothing
- Social
    - Very low affinity for: Religion
    - Low affinity for: Traditions, family, rational, material and culture
    - Average affinity for: Social
    - Very high affinity for: Lust

To summarize, Arvato customers seem to be older, financially better, traditional/religious, have pets and some history of mail orders. Non customers seem to be younger, financially comfortable, less traditional/religious, no pets and very less history of mail orders.

References

1. https://builtin.com/data-science/step-step-explanation-principal-component-analysis
2. https://en.wikipedia.org/wiki/Principal_component_analysis
3. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a
4. http://www.nlpca.org/pca_principal_component_analysis.html
5. https://heartbeat.fritz.ai/understanding-the-mathematics-behind-k-means-clustering-40e1d55e2f4c