Data

# Cleaning

* Dealing with nulls, some columns etc
* Dropping car-specific columns after looking at information levels data dictionary

# Preparation

* For MCA all continuous numeric columns are dropped, and categorical columns are one-hot encoded. This is due to a bug in the implementation that means categories missing in unseen data cause the .transform() method to fail, so dummy features were created and edited to exist (albeit filled with 0s), and the train and test dataframes forced to match one another. The prince mca implementation deals with scaling etc, but cannot handle negative values so all -1 categories (often used for unknown) were moved to 100.
* For PCA all categorical columns are dropped
* Samples were taken of all datasets due to their size. For the general population a random sample of 5% is used (44561 rows), and for the customers dataset 15% (28748 rows)

Customer segmentation report

# Approaches

Due to mix of categorical and continuous data, FAMD was trialled, but due to a bug in the implementation, it cannot handle ‘categorical’ type data (only ‘object’ type), and so PCA and MCA were utilised separately for the continuous and categorical features, respectively. The general population dataset was used as a training set, and the customers dataset as test. See links for explanations of approaches.

# Results

## Missing/new categories in test set

It is possible to draw some conclusions based on simply the categories that:

in train but not test:

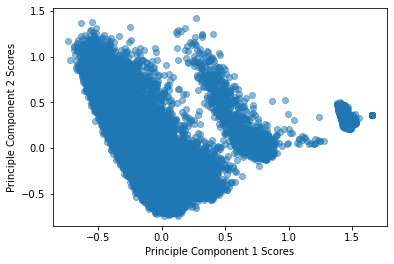
{'D19\_TELKO\_ANZ\_24\_6', 'D19\_VERSI\_ONLINE\_QUOTE\_12\_3.0', 'D19\_TELKO\_ANZ\_12\_6', 'D19\_BANKEN\_ONLINE\_QUOTE\_12\_1.0', 'ANZ\_KINDER\_6.0', 'ORTSGR\_KLS9\_0.0', 'D19\_BANKEN\_LOKAL\_1', 'D19\_LOTTO\_4.0', 'ANZ\_KINDER\_9.0', 'D19\_TELKO\_ONLINE\_QUOTE\_12\_5.0', 'GEBAEUDETYP\_5.0', 'TITEL\_KZ\_2.0', 'D19\_VERSI\_ONLINE\_DATUM\_2'}

in test but not train:

{'D19\_LOTTO\_2.0', 'D19\_BANKEN\_ONLINE\_QUOTE\_12\_6.0', 'D19\_BANKEN\_OFFLINE\_DATUM\_3', 'D19\_VERSI\_ANZ\_12\_6', 'TITEL\_KZ\_3.0', 'ANZ\_KINDER\_8.0'}

## MCA

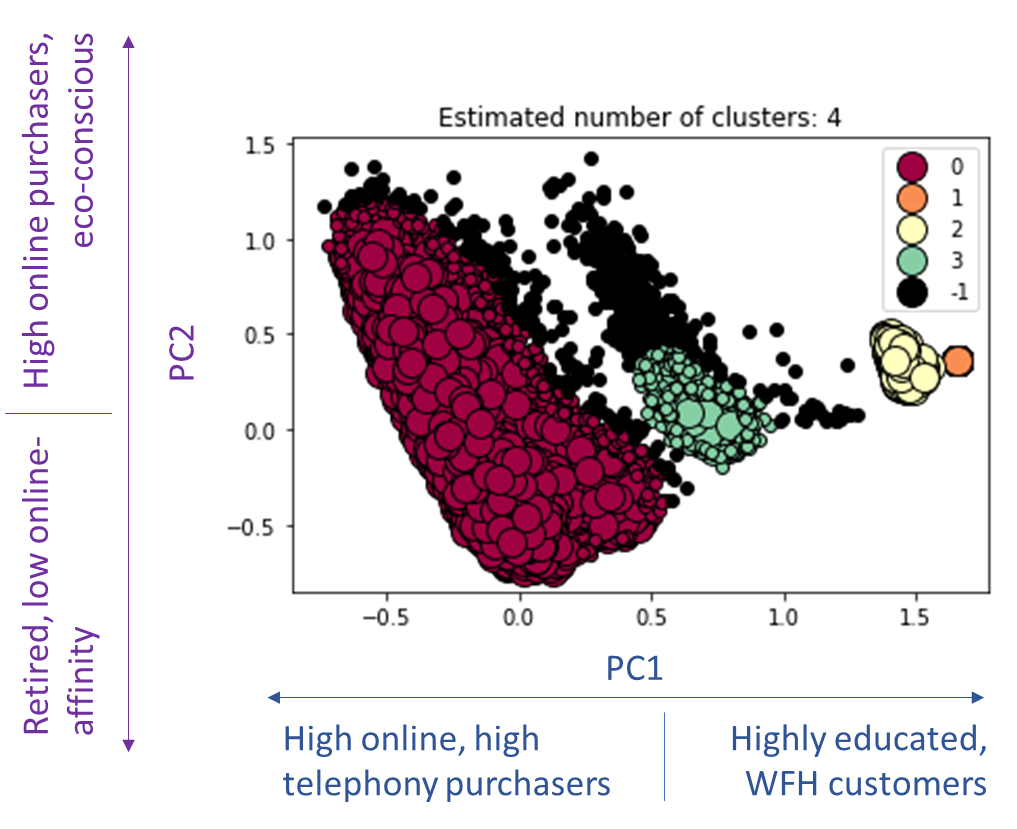
Fitting the MCA to the general population dataset yields the following distribution of customers:



First two components explain: [0.03720371795640544, 0.02184419941411917], total inertia = 7.180232558139535

It is worth noting that it looks like the explained variance here is relatively low for these first two components, possibly due to how sparse the dataframe used is as all categories had to be one-hot-encoded in order to be analysed. Therefore conclusions drawn from this will not represent the full picture.

Analysing the loadings for each of the top two components, we can see that component 1 varies from high-online, high-telephony purchasers at low values, to highly educated, WFH (work-from-home) customers at high values. Component 2 varies from retired, low-online affinity customers at low values, to high-online purchasers and possible eco-conscious customers at high values.

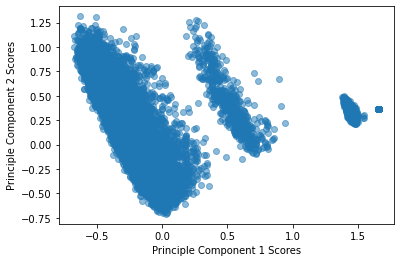


See loadings for each component:

|  |  |  |  |
| --- | --- | --- | --- |
| Driving highly negative PC1 values | Meaning | Driving highly positive PC1 values | Meaning |
| D19\_VERSI\_ONLINE\_DATUM\_1 | actuality of the last transaction for the segment insurance ONLINE, possibly 'very low' | **WOHNDAUER\_2008\_1.0** | length of residence below 1 year |
| D19\_TELKO\_ANZ\_12\_5 | transaction activity TELCO in the last 12 months, 'high activity' | **TITEL\_KZ\_1.0** | Customer holds 'Dr.' title |
| D19\_BANKEN\_ONLINE\_QUOTE\_12\_9.0 | amount of online transactions within all transactions in the segment bank, '90% Online-transactions within the last 12 months' | **SOHO\_KZ\_1.0** | customer has small office/home office |
| D19\_BANKEN\_ONLINE\_QUOTE\_12\_2.0 | amount of online transactions within all transactions in the segment bank, '20% Online-transactions within the last 12 months' | **OST\_WEST\_KZ\_1.0** | unknown whether customer is former east/west germany |
| D19\_TELKO\_ONLINE\_QUOTE\_12\_5.0 | ? | **EWDICHTE\_-1.0** | Unknown density of inhabitants per square km |

|  |  |  |  |
| --- | --- | --- | --- |
| Driving highly negative PC2 values | Meaning | Driving highly positive PC2 values | Meaning |
| GEBAEUDETYP\_5.0 | Type of building 'company building w/o known company' | **D19\_TELKO\_ONLINE\_QUOTE\_12\_5.0** | ? |
| ALTER\_HH\_2.0 | main age within the household '01.01.1900 bis 31.12.1904' ?? | **D19\_BANKEN\_ONLINE\_QUOTE\_12\_9.0** | amount of online transactions within all transactions in the segment bank, '90% Online-transactions within the last 12 months' |
| ALTER\_HH\_3.0 | main age within the household '01.01.1905 bis 31.12.1909' ?? | **D19\_BIO\_OEKO\_2** | transactional activity based on the product group ECOLOGICALS 'Doublebuyer 0-12 months' |
| LP\_LEBENSPHASE\_FEIN\_6.0 | Lifestage fine 'single low-income earners at retirement age' | **D19\_BANKEN\_ONLINE\_QUOTE\_12\_7.0** | amount of online transactions within all transactions in the segment bank, '70% Online-transactions within the last 12 months' |
| ONLINE\_AFFINITAET\_0.0 | online affinity 'none' | **D19\_BANKEN\_ONLINE\_QUOTE\_12\_8.0** | amount of online transactions within all transactions in the segment bank, '80% Online-transactions within the last 12 months' |

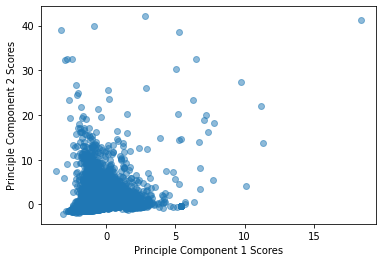
Transforming the customer data using the same components yields the following distribution:



In comparison to the general population , this dataset seems to have lower density in the region of high PC1 combined with low PC2, indicating that, relatively fewer highly educated, WFH (work-from-home) customers and retired, low-online affinity customers are present in this sample

## PCA

Fitting PCA to the general population dataset yields the following distribution of customers:



First two components explain: [0.37621404 0.25678651], total inertia = 7.999999999999999

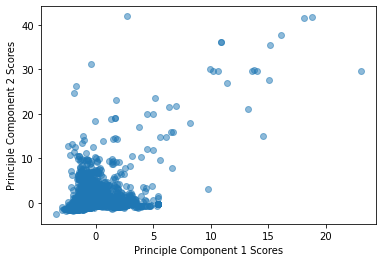
Here much more of the variance is explained, though there are only 6 continuous columns used here from the dataset.

PC1 goes from younger customers with more cars at low values to more highly educated customers at higher values. PC2 possibly indicates higher density of housing at high values, and differentiates between professional title holders (low values) and academic title holders (high values).

|  |  |  |  |
| --- | --- | --- | --- |
| Driving highly negative PC1 values | Meaning | Driving highly positive PC1 values | Meaning |
| EINGEZOGENAM\_HH\_JAHR | not found in data dictionaries | **ANZ\_TITEL** | number of professional title holder in household |
| KBA13\_ANZAHL\_PKW | number of cars in the PLZ8 | **ANZ\_HH\_TITEL** | number of academic title holder in building |
| GEBURTSJAHR | Birth year | **ANZ\_STATISTISCHE\_HAUSHALTE** | not found in data dictionaries |
| ANZ\_PERSONEN | number of adult persons in the household | **ANZ\_HAUSHALTE\_AKTIV** | number of households in the building |
| ANZ\_HAUSHALTE\_AKTIV | number of households in the building | **ANZ\_PERSONEN** | number of adult persons in the household |

|  |  |  |  |
| --- | --- | --- | --- |
| Driving highly negative PC2 values | Meaning | Driving highly positive PC2 values | Meaning |
| ANZ\_PERSONEN | number of adult persons in the household | **ANZ\_STATISTISCHE\_HAUSHALTE** | not found in data dictionaries |
| GEBURTSJAHR | Birth year | **ANZ\_HAUSHALTE\_AKTIV** | number of households in the building |
| KBA13\_ANZAHL\_PKW | number of cars in the PLZ8 | **ANZ\_HH\_TITEL** | number of academic title holder in building |
| ANZ\_TITEL | number of professional title holder in household | **EINGEZOGENAM\_HH\_JAHR** | not found in data dictionaries |
| EINGEZOGENAM\_HH\_JAHR | not found in data dictionaries | ANZ\_TITEL | number of professional title holder in household |

Transforming the customer data using the same components yields the following distribution:

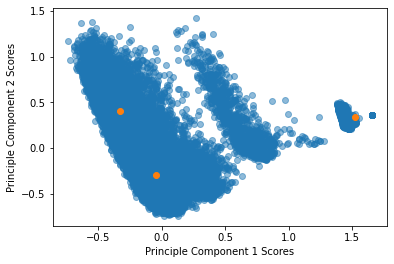


In comparison to the general population , this dataset seems to have lower density in the region of high PC2 combined with low PC1, and higher density in the region of high PC1 and high PC2.

This indicates that there are relatively more customers who are highly educated, and with fewer cars in comparison to the general population, and perhaps suggests that of people who live in denser accommodation and hold academic titles, the ones who have fewer cars are over-represented in the customer dataset.

## Clustering

Kmeans did not do so well:



Due to unusual shape (long and thin) DBSCAN was instead trialled. DBSCAN has no cluster centres and therefore no ‘predict’ method to apply cluster centres to a new dataset. Therefore we must compare clusters fitted to each of the datasets:

|  |  |  |
| --- | --- | --- |
|  | General Population | Customers |
| MCA (categorical variables): |  |  |
| PCA (continuous variables): |  |  |

Please note that black data points (in cluster -1) are noise according to DBSCAN (i.e. they are not reachable from any other cluster point using the set epsilon value). The datasets are similar enough to produce similar clusters that can be compared.

Looking at the percentage of each population (general population and customer population) that makes up each of the clusters identified by DBSCAN, we can see which groups of individuals are either over or under represented in the customer dataset.

| **MCA Cluster** | **Percentage of general population** | **Percentage of customer population** | **Percentage difference between general population and customers** |
| --- | --- | --- | --- |
| **-1** | 1.117569 | 1.694031 | 51.6 |
| **0** | 89.277619 | 73.705997 | -17.4 |
| **1** | 4.995400 | 21.003200 | 320.5 |
| **2** | 3.179911 | 2.960206 | -6.9 |
| **3** | 1.429501 | 0.636566 | -55.5 |

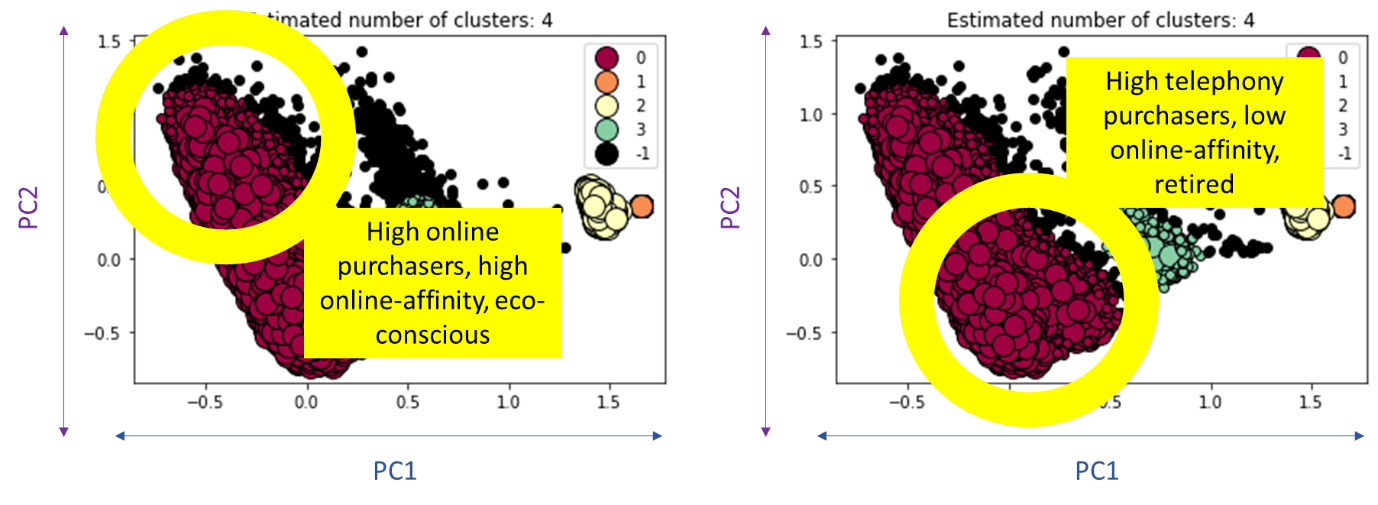
We can see that that for the MCA, clusters 0, 2 and 3 are under-represented in the customer population, while cluster 1 is significantly over-represented.

The main differences are in clusters 1 and 3, with clusters 0 and 2 showing less of a difference.

Cross referencing this back to the loadings of each of the components, we can say that cluster 1 (orange), with high values for component 1 and medium values for component 2 represents highly educated, possibly working-from-home customers, with not-so-high online affinity.

Cluster 3 (green) sits between clusters 0 and 1, with around 0 for component 2, and moderately positive component 1 scores. This cluster may represent customers who are similar to clusters 1 and 3 but less likely to work from home. Their under-representation in the customers set may signify their inability to be home to receive orders by mail.

Interestingly, cluster 0 (maroon), the largest cluster in each case, with lower component 1 values and the whole range of component 2 values, represents possibly two groups of customers (see below), one with high online affinity and high online purchase rates, and one with low online affinity, high telephony purchase rates and likely to be retired. Although these groups are not differentiated by DBSCAN, from visually assessing the clusters, it may be that the customers dataset has fewer individuals in the second group (retirees), compared to the overall population dataset.



Summarising this, it may be said that the mail-order company might be better focussing on attracting customers that fall into cluster 1 – highly educated (possibly high disposable income proxy), often at home, but not high online purchasers (as you might expect) – rather than retired customers, or those who already purchase items online often.

K-means clustering could also be revisited, and then cluster centre coordinates used to discover PC values at those points.

| **PCA Cluster** | **Percentage of general population** | **Percentage of customer population** | **Percentage difference between general population and customers** |
| --- | --- | --- | --- |
| **-1** | 1.864859 | 1.801864 | -3.4 |
| **0** | 89.953098 | 74.231251 | -17.5 |
| **1** | 8.182043 | 23.966885 | 193 |

In PCA, there are only two clusters, and the largest change is seen in cluster 1, which is over-represented in the customer datset.

Again looking back to the loadings of each of the components, we can see that cluster 1 – with higher component 1 values than cluster 0 – may represent customers with fewer cars and higher education levels. This cluster also has less positive component 2 values, indicating that perhaps these customers live in housing with lower densities, but in professional roles.

Summarising this, it could be that this cluster represents individuals who are not in city centres, but are less likely to have cars to travel and shop. Higher education and professional roles may indicate a likelihood to be working from home. It may be worth combining the PCA PC1 with the MCA results to see if this coincides with the clusters with high values for the small office/home office flag.