Machine Learning Engineer Nanodegree

Capstone Project Report

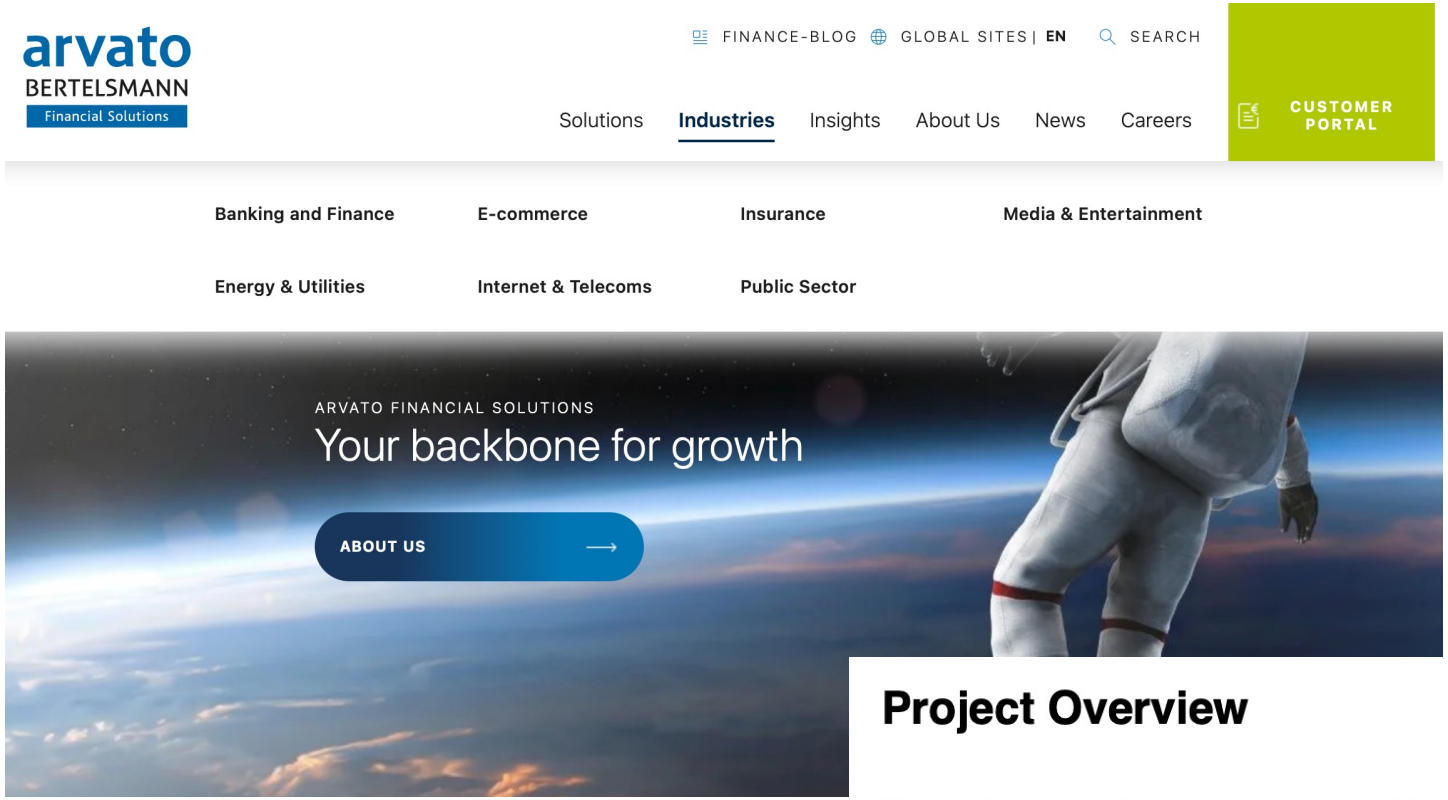
Customer segmentation report for Arvato Financial Services

Juanita Smith, Dec 2021

Machine Learning Engineer Nanodegree

Capstone Project Report

Customer segmentation report for Arvato Financial Services



https://finance.arvato.com

Project Overview

The project selected were sponsored by Udacity in partnership with Arvato Financial Solutions.
The project are within the marketing domain, with main objective to help one of Arvato's clients, a mail-order company in Germany, acquire more clients easier and smarter.
The mail-out company are running mail-out campaigns, with their objective to increase efficiency in their customer acquisition process by targeting the right people.

The was data was provided by Arvato and is protected under Terms and Conditions.
This is a real data-science project with real data.

Project Goals

Age, Vacation habits, Gender, Buying channels
Financial type (saver, house owner, investor, etc)

Unemployment
Size of community
Employment in relation to country

Main age, number of persons, income, number of children, length at residence
Type of customer in last 12 months (regular, active, new, inactive)
Transaction activities, Structure (single, couple, family)
in last 12 and 24 months
Last transactions in several segments/categories

Sub-postcode (about 8 PLZ8 make up one PLZ) and
new microcell level (about 500 households)
Use vehicle characterises to classify consumers by comparing speed, engine power, segment, age of vehicle and car owners

Which type of building do they live in:
Number of households in building
Residential, Business, Near city centre,
Neighbourhood-area (very good -> rather poor, rural), etc

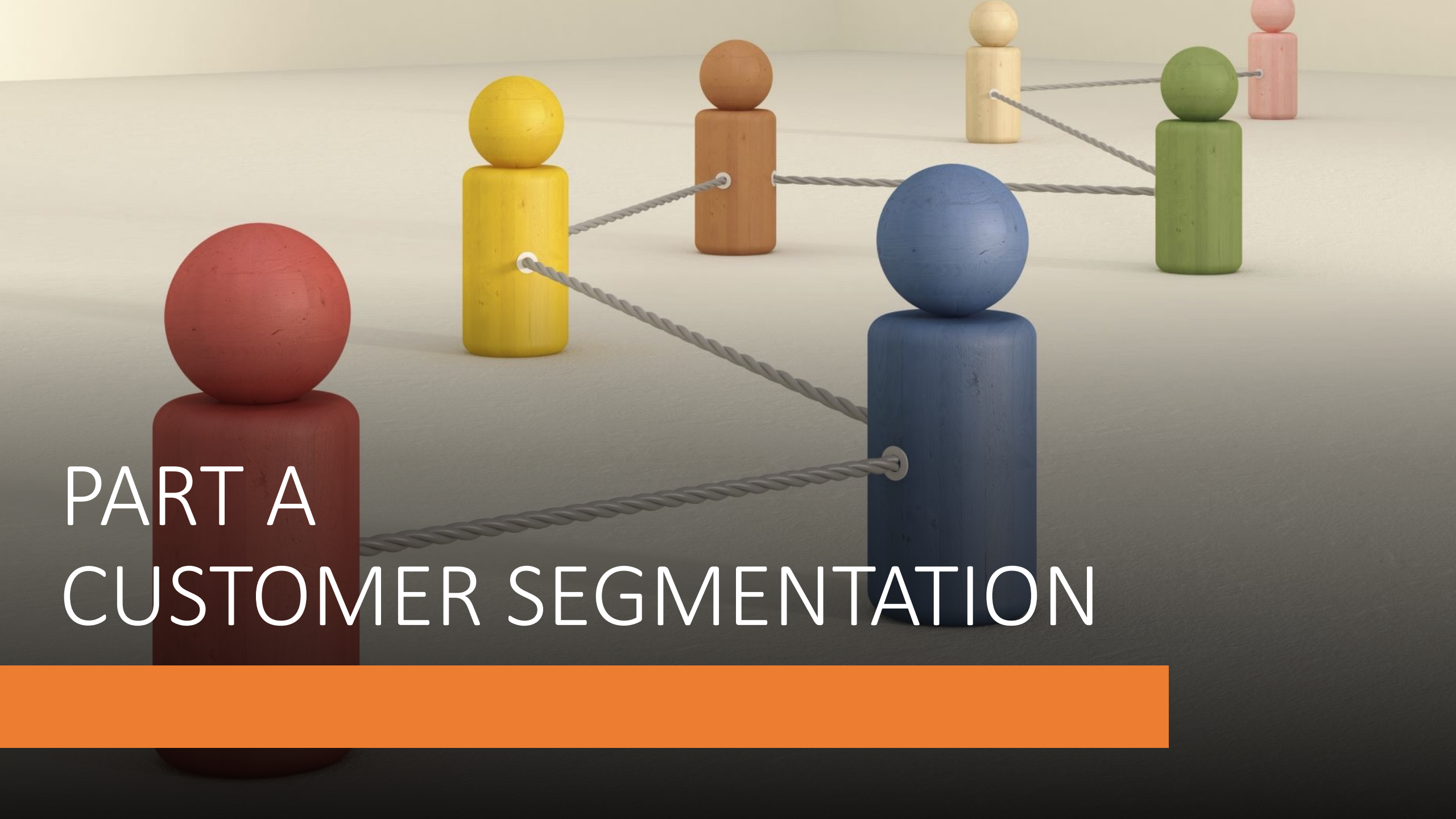
Residential vs business cells
Purchasing power
Online Affinity
upper to lower class

Upper/Lower Class

Distance to city centre, next metropole
Density of inhabitants, etc.

Use vehicle details to classify consumers

Transaction activity by different product groups
(insurance, technology, mobile communications, food, clothing, etc



PART A
CUSTOMER SEGMENTATION

PART 1 – CUSTOMER SEGMENTATION

Datasets and Input

2 datasets were provided by Arvato/Udacity:

1. Demographics and own attributes of the mail-order company's **existing customers**
   191 652 persons (rows) x 369 features (columns).

2. A bigger dataset containing the same demographics and attributes for the wider **general population of Germany**
   891 211 persons (rows) x 366 features (columns)

Demographic data were obtained using open public data sources in Germany and were provided to students to downloaded from Udacity

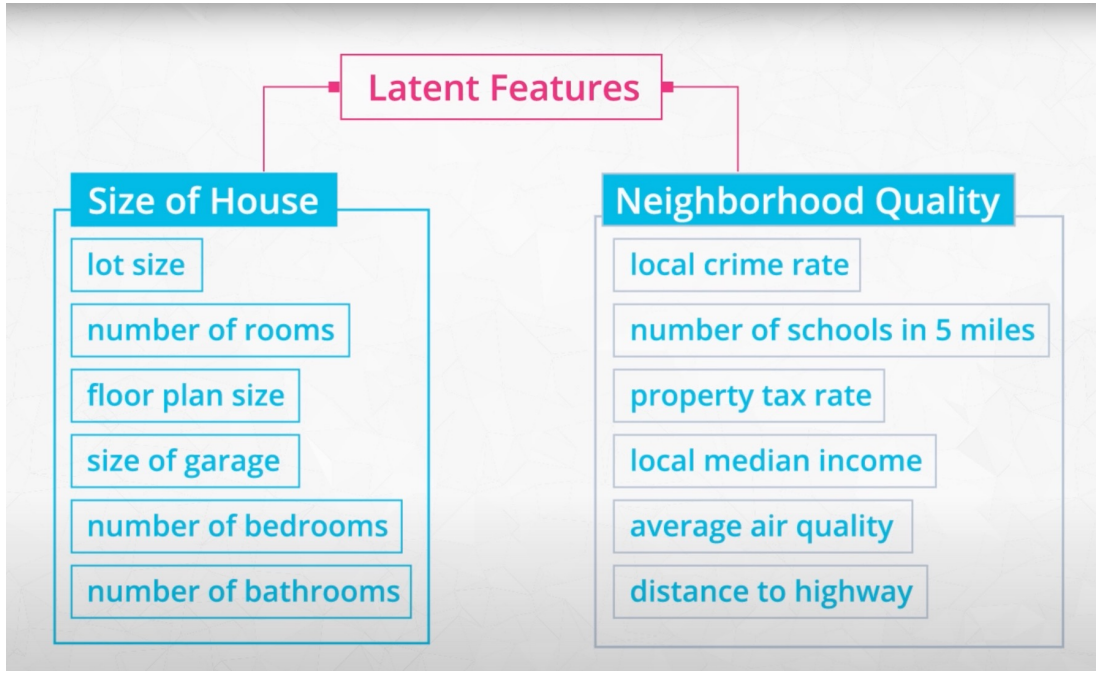
PART 1 – CUSTOMER SEGMENTATION

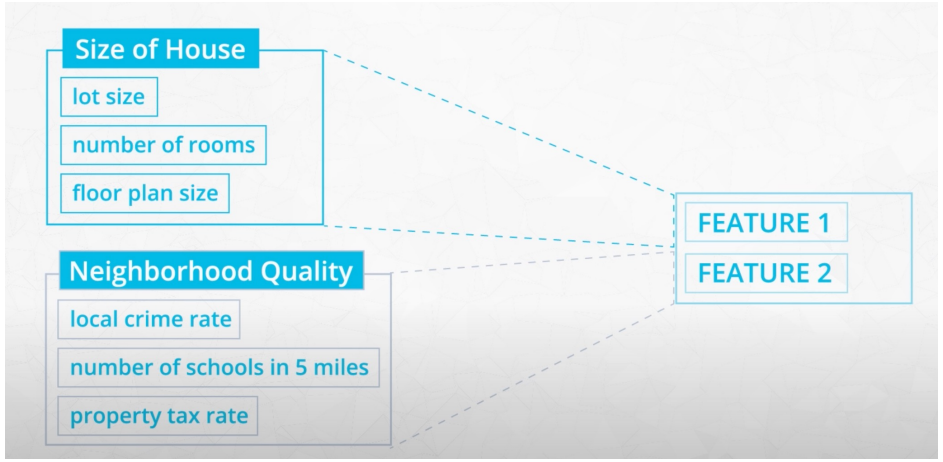
Steps

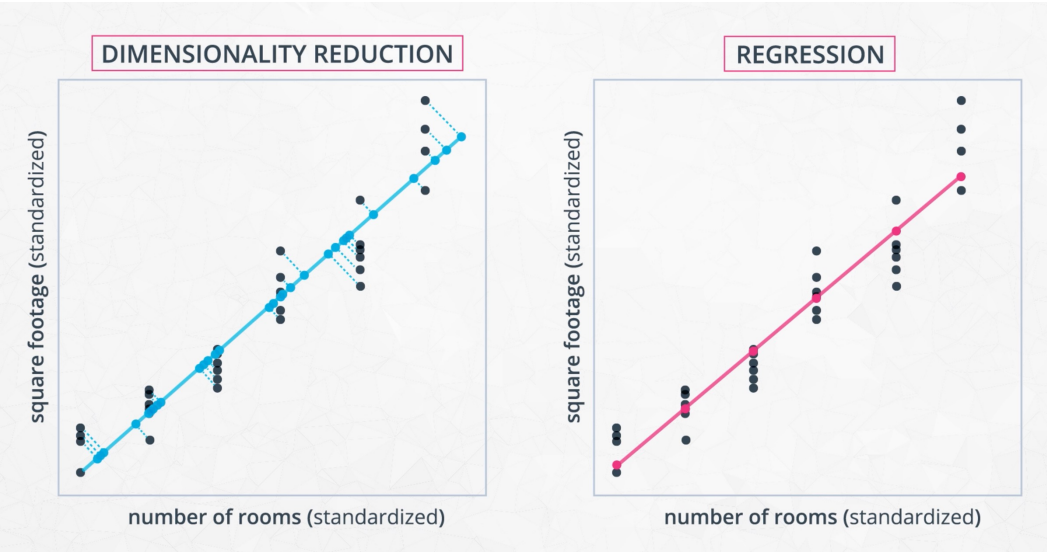
1. PCA (Principal Component Analysis) technique to summarize the data
   -> Build persona's

2. Clustering using KMEANS algorithm
   -> Use persona's to cluster

PCA (PRINCIPAL COMPONENT ANALYSIS)
DIMENSION REDUCTION – A CLEVER WAY TO SUMMARIZE / GROUP DATA BEFORE WE CLUSTER





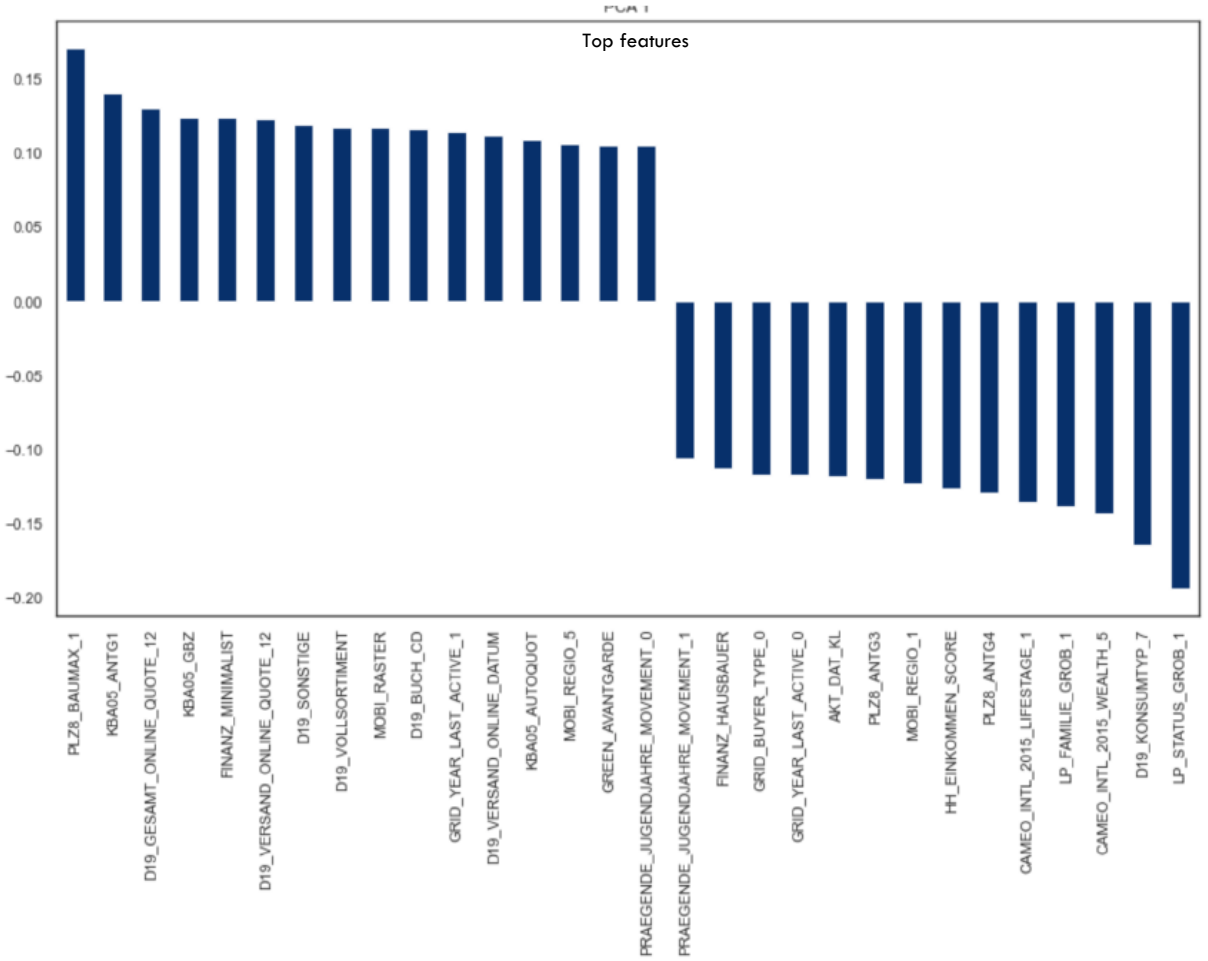


Summarize features to build new features
looking for properties (characteristics) that strongly differ across the data, e.g. age describe wine better than color

PCA component/PERSONA 1 – TOP EARNER FAMILY

PCA component 1 uses mainly building type and geographical location, mobility and financial status to describe this persona and explain 8.1% of the variance when we select 180 components

Analysing the strongest positive and negative linear relationships this component is describing top earner families living in area's with mostly 1-2 family houses which they own, outside of the city but in larger residential area's. They own multiple cars per household. They are active online customers, yet minimalistic, and are part of the green avantgarde.



Types of building

Strong positive correlations:
PLZ8_BAUMAX_1 / KBA05_BAUMAX_1 = mainly 1-2 family houses in PLZ / microcell
KBA05_GBZ – Number of buildings in the microcell is going up which suggest it's a large residential area
WOHNLAGE_RURAL = Living outside of the city
GEBAEUDETYP_1 = residential building

Strong Negative correlations:
PLZ8_ANTG3 / ANTG4 = > 6 family houses

Vehicles and mobility

Strong positive correlations:
KBA05 - AUTOQUOT - Share of cars per household high
MOBI_RASTER / MOBI_REGIO_5 = very low mobility is going up

Strong Negative correlations:
MOBI_REGIO_1 = very high mobility is going down

Financial status and family orientation

Positive correlations
Finance minimalist - moving towards low means they are not minimalistic
CAMEO_INTL_2015_WEALTH_2 = Prosperous Households
LP_STATUS_GROB 4 and 5 = top earners, house owners
LP_STATUS_FEIN_10 = top earners
CAMEO_INTL_2015_WEALTH_LIFESTAGE_1 = prefamily and single people going down
LP_FAMILIE_GROB_5 = multi familie households

Strong Negative correlations:
Hausbauer – moving towards high means they are house owners
HH_EINKOMMEN_SCORE – organized from 1 = high, 5 = low so I assume it moves towards high income
LP_STATUS_GROB_1 = not low-income earners
CAMEO_INTL_2015_WEALTH_5 = not poorer households

Online and consumer behavior

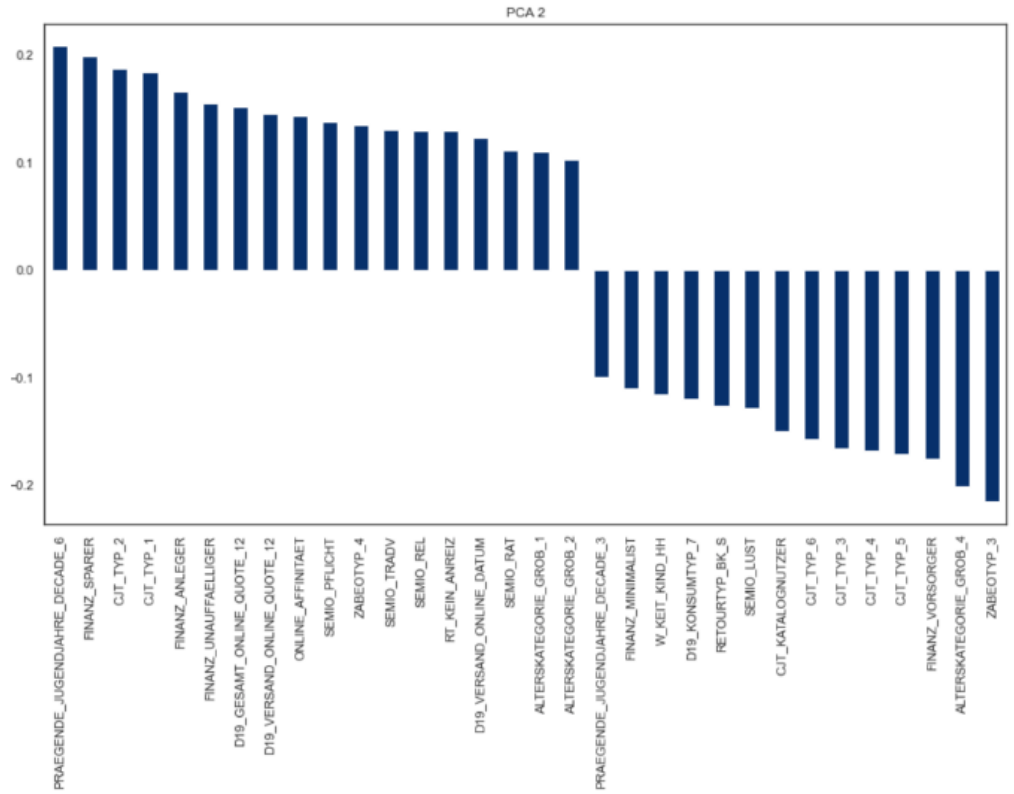
Strong positive correlations:
High online transactions especially in category MAIL-ORDER category
Buying mostly in categories: other, books, cd, house decor, technical, cloths, insurance, complete mail-order offers
Online affinity is increasing
Activity in the last year
D_KONSUMTYP_1 – Consumer type Minimalist

Personal behavior
Green Avantgarde = Belongs to the green avantgarde
ZABEOTYP_1 = green

PCA component/PERSONA 2 – DIGITAL MEDIA KIDS

PCA component 2 is focussing on youth movement and age, financial status and online shopping behaviour and explain 5.6% of the variance when we use 180 components

This component describe digital media kids, adults who's youth was in their 90's. Younger low income families with high probability of having children. They have a super strong online presence, seems to shop online frequently and react well to advertising campaigns and catalogues. They don't save or invest, but in contrast wants to be financially prepared.



Financial status and family orientation

Positive correlations

PRAEGENDE_JUGENDJAHRE_DECADE_6 = youth in 90's is going up suggesting these are younger people

FINANZ_SPARER, FINANZ_ANLEGER, FINANZ_UNAUFFAELLIGER moving towards low means they don't save or invest

Negative correlations

FINANZ_VORSORGER, FINANZ_MINIMALIST moving towards high so means they like to be financially prepared, and they are minimalists

W_KEIT_KIND_HH - likelihood of a child present in this household are likely

Online and consumer behavior

Positive correlations

CJT 1, 2 – moving towards low means NOT advertising and consumption minimalist and traditionalist

On-line transactions are high, great mail-order client

Multi buyers

Buy cloths and insurance

D19_KONSUMTYP_1 = Universal

Negative correlations

Negative CJT 3-6 = moving towards high means a big shopper (store, online shopper, cross channel)

RETOURTYP_BK_S – is this return customer - going down means it's a return customer possibly crazy shopper

D19_KONSUMTYP_7 = inactive (but means opposite actually suggest it's an active customer)

Personal behavior

Positive correlations

NOT Dutiful, Traditional, Rational, Religious, Cultural

ZABEOTYP_4 Energy consumer – price driven

ALTERSKATEGORIE_GROB_1 = < 30 years

ALTERSKATEGORIE_GROB_2 = between 30 and 60 , but with stronger relationship

LP_STATUS_GROB_1 – low-income earners

Negative correlations

PRAEGENDE_JUGENDJAHRE_DECADE_3: 46 – 60 going down, means it's younger people

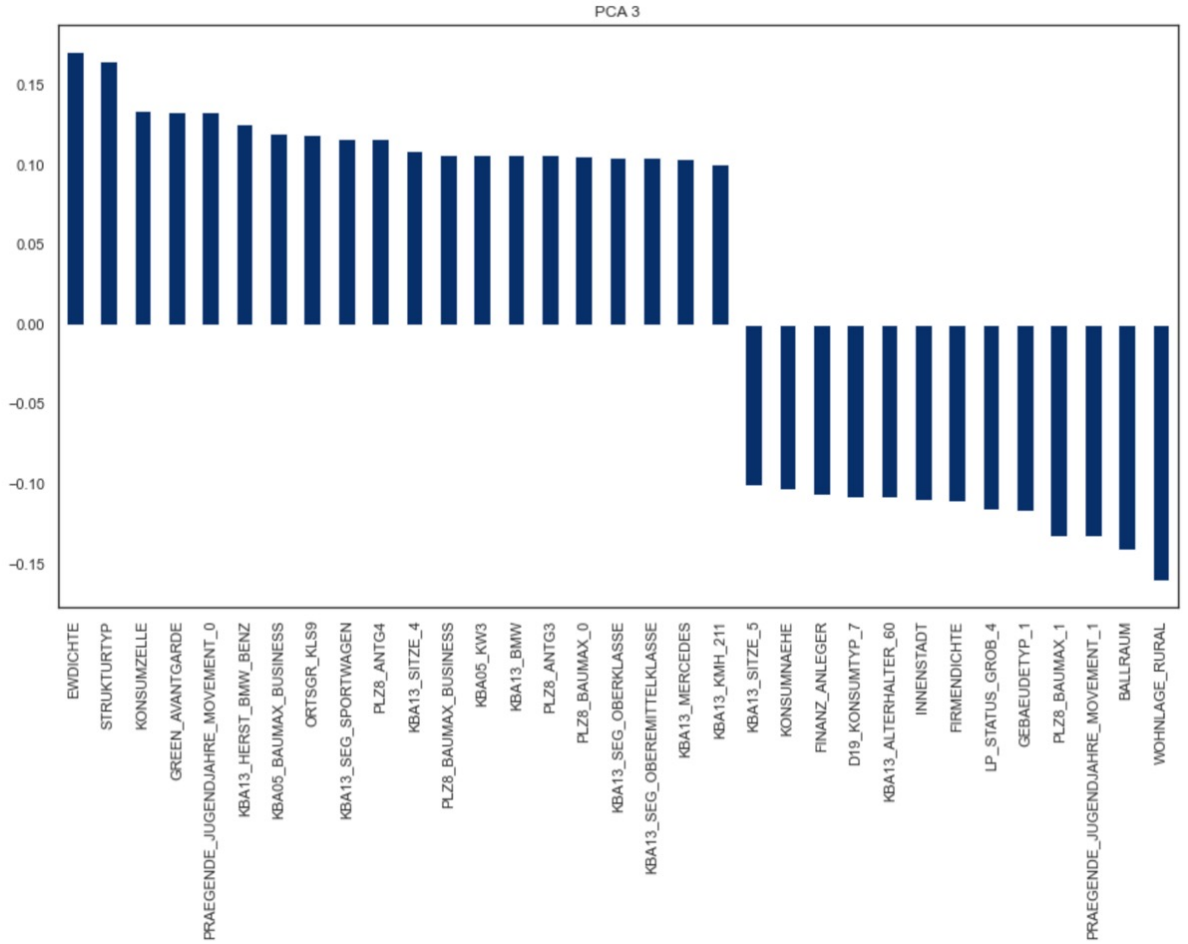
ALTERSKATEGORIE_GROB_4 – Over 60 is going down

ZABEOTYP_3 Energy – fair supplied

They are sensual or dutiful

PCA component/PERSONA 3 – DRIVING GERMAN CARS IN WEST-GERMANY

PCA component 3 persona describes individuals that are living in dense populated area's in West-Germany near the city centre. They live in area's where there are high number of > 6 family houses but also lots of businesses present. They drive top and middle class German manufacturer car brands mostly smaller BMW and Mercedes with <= 4 seats.They are part of the green avantgarde. These individuals are not in age group 46-60, and are most likely to be pre-family and single younger people. They were active online in the last year. The component explain 3.9% of the variance.



PCA component 4 persona is describing older people > 60 years living in East-Germany from poorer households and lower income and not part of green avantgarde. They are driving smaller mid-class vehicles like Ford, Masda and of Asien origin with 5 seats. They are frequent online shoppers buying insurance, technology and respond to complete mailout offers and were active in the last year. They live in area's where there are a high share of 6-10 family houses in the area.

PCA Component 5 is using mainly personal interest to describe the persona.
This persona represent males, fight-full, dominant, critical and rational, but not dreamy, cultural, family orientated, social or religious. They hate shopping in general and are frequent users of advertising catalogues. They are unlikely to be 46-60 years old. Online shopping are a great convenience for those who hate going to the shops

First top PCA components describe most of the data

For the project I selected the first 200 components which described 90% of the data

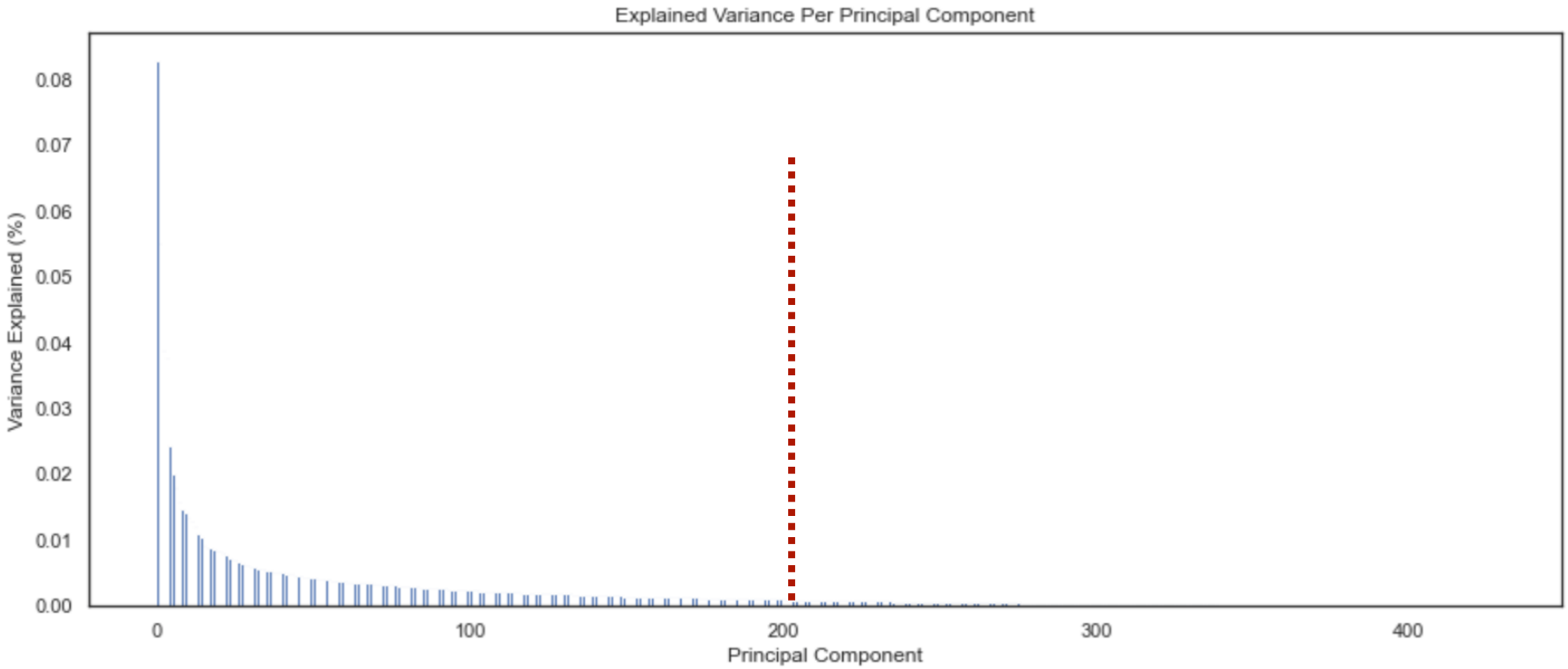


Fig 3.7

CLUSTERING PCA COMPONENTS USING KMEANS

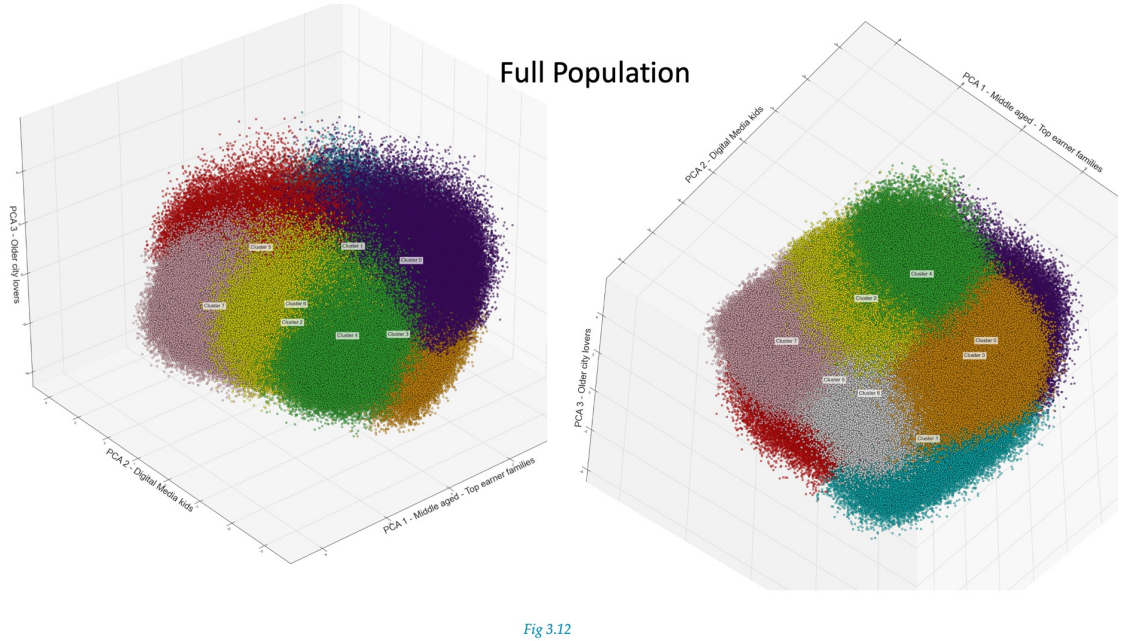
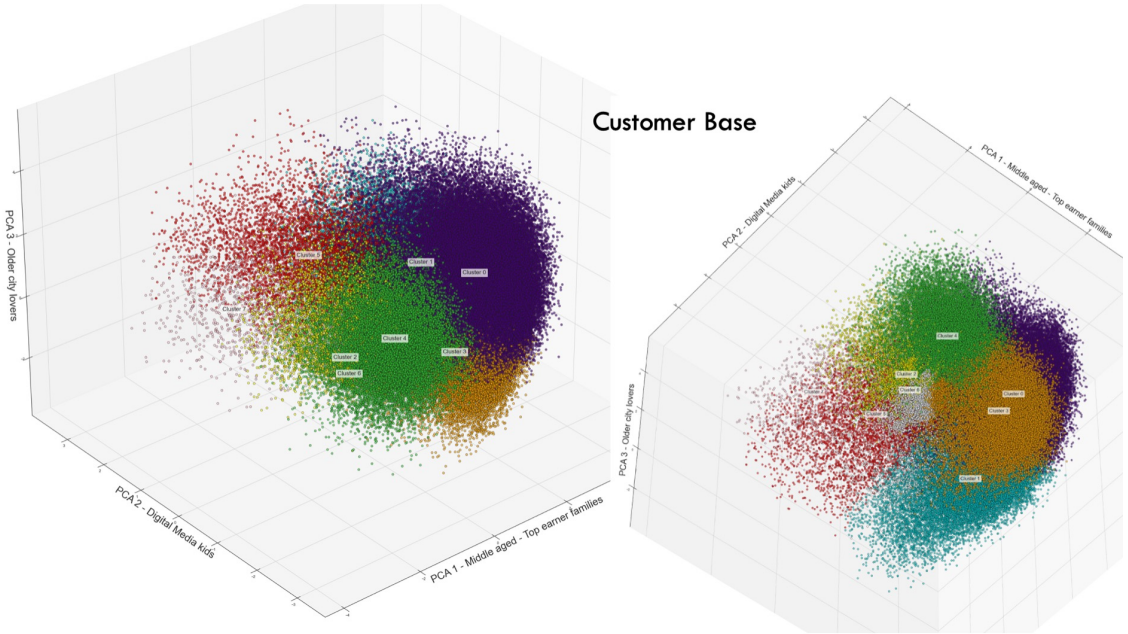
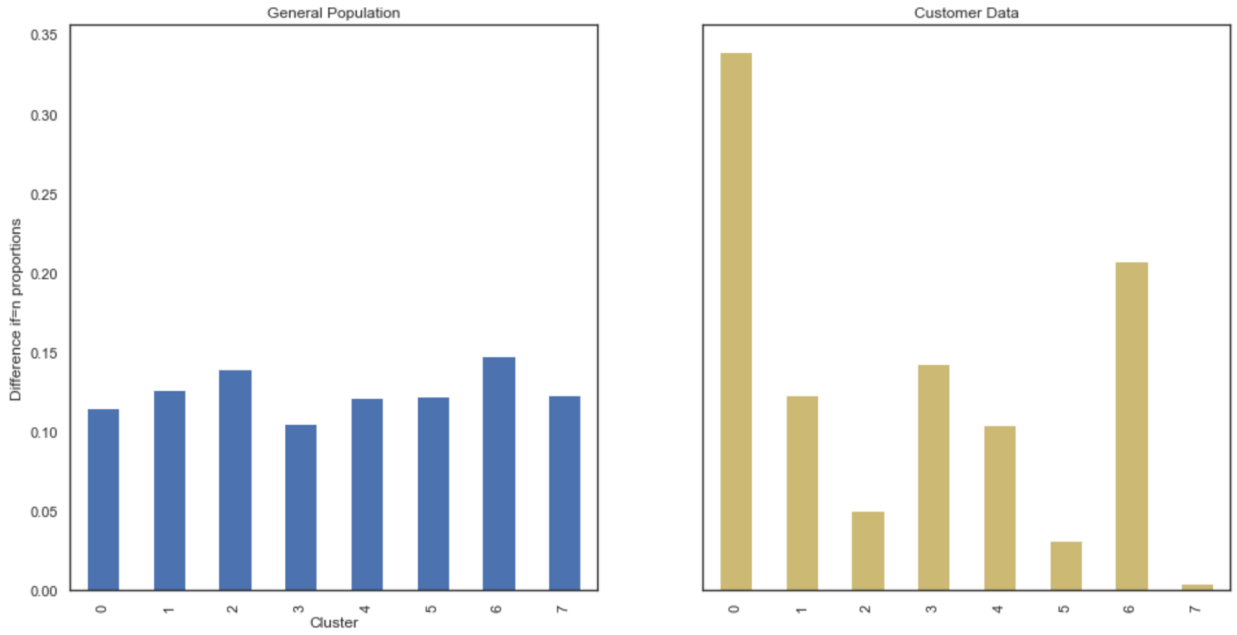


Fig 3.12



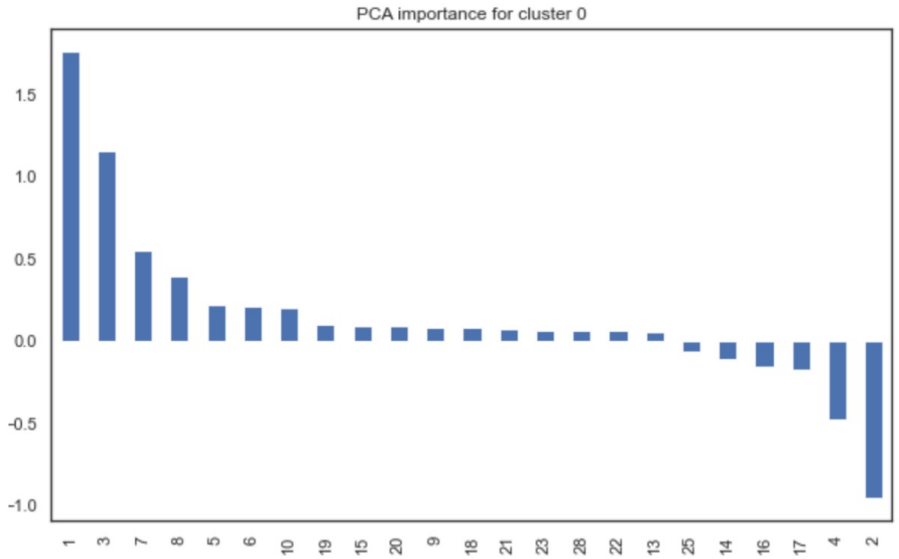
Plotting the clusters we can observe the clusters in the full population are quite evenly distributed, where in the customer base the we have a lot of variation.



We can observe cluster 0 (INDIGO) are the most over represented, and cluster 7 (PINK) are the most under represented

Analysis cluster 0 (most over represented in customer base)
PCA component 1 and 3 contain the biggest positive weight in cluster 0, whilst component 2 and 4 have the largest negative weight.

This suggest that cluster 0 are focussing on the high-income earners, which are either settled top earning families living outside of the city, or individuals from West-Germany driving top class german manufacture vehicles living near or in the city centre.

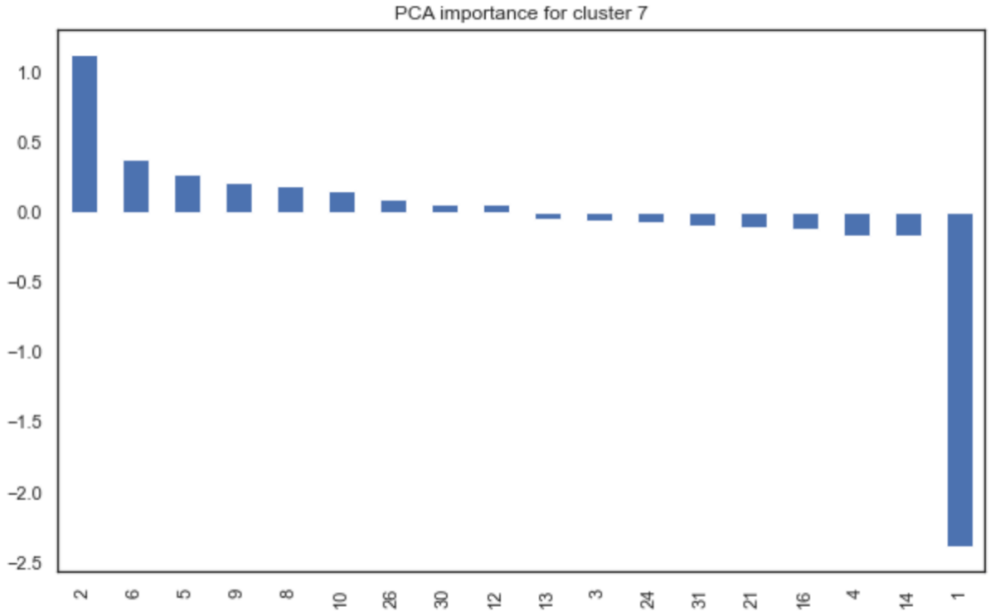


Recap of top 5 persona's

- PCA component 1 persona describes TOP EARNER FAMILIES
- PCA component 2 persona describes DIGITAL MEDIA KIDS in their youth (young family, poor income)
- PCA component 3 persona describes people in West Germany driving German cars
- PCA component 4 persona is describing older people > 60 years living in East-Germany from poorer households
- PCA component 5 person describes males, fight-full, dominant that hates shopping

Analysis of cluster 7 (most under represented in customer base)
Cluster 7 have a very strong negative weight for PCA component 1 which indicates that top earners families are totally excluded.

Largest positive weight is PCA component 2, which represent young low income families. Also component 5 is featuring means those dominant males who hates going to the shops and prefer online shopping, are included in the customer base of the mail order company.



Recap of top 5 persona's

- PCA component 1 persona describes TOP EARNER FAMILIES
- PCA component 2 persona describes DIGITAL MEDIA KIDS in their youth (young family, poor income)
- PCA component 3 persona describes people in West Germany driving German cars
- PCA component 4 persona is describing older people > 60 years living in East-Germany from poorer households
- PCA component 5 person describes males, fight-full, dominant that hates shopping

V. Conclusion

We can now answer the questions defined in the problem statement:

1. Help the mail order company to understand who their customers are

Main customer base are coming from cluster 0, who are presenting high income families living in less dense residential area's possibly in rural area's, or individuals who drive top german car manufacturer brands living in densely populated city centres.

2. How does their client base compare to the rest of Germany ?

Rest of Germany have an even distribution between the 8 different clusters, where the client base of the mail order company are over represented by high income groups, and under represented by young families of low income.

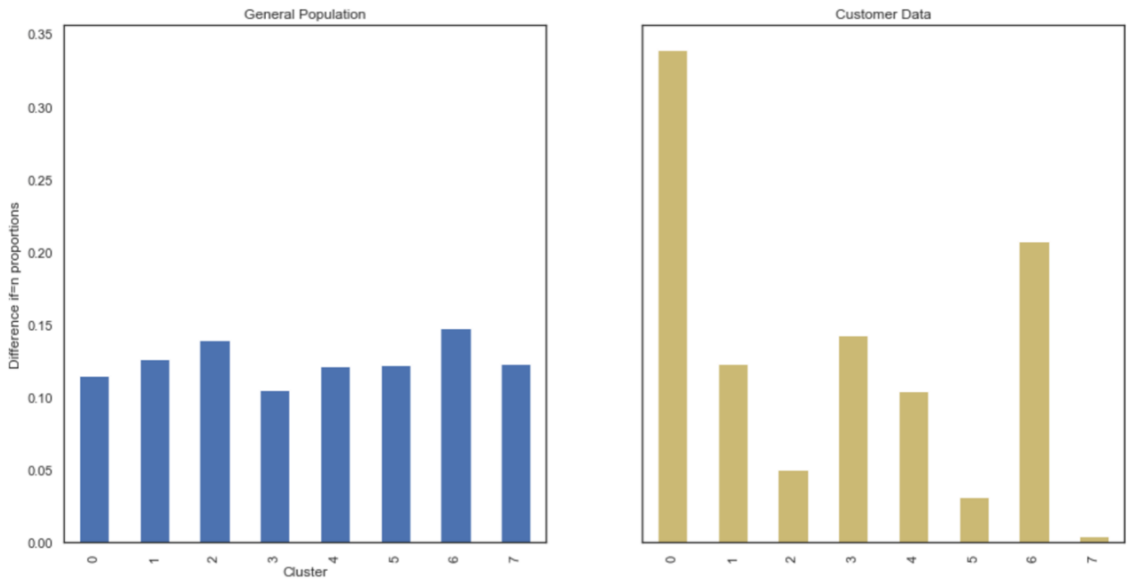


Fig 5.1

PART B
TARGET MARKETING

PART 2 – MAILOUT CAMPAIGN PREDICTION

Datasets and Input

3. Training data:

Demographics data for individuals who were targets of a marketing campaign
Subset of the customer dataset
42 982 persons (rows) x 367 (columns)
'RESPONSE' column indicate if individual responded or not. We use this information to train the model

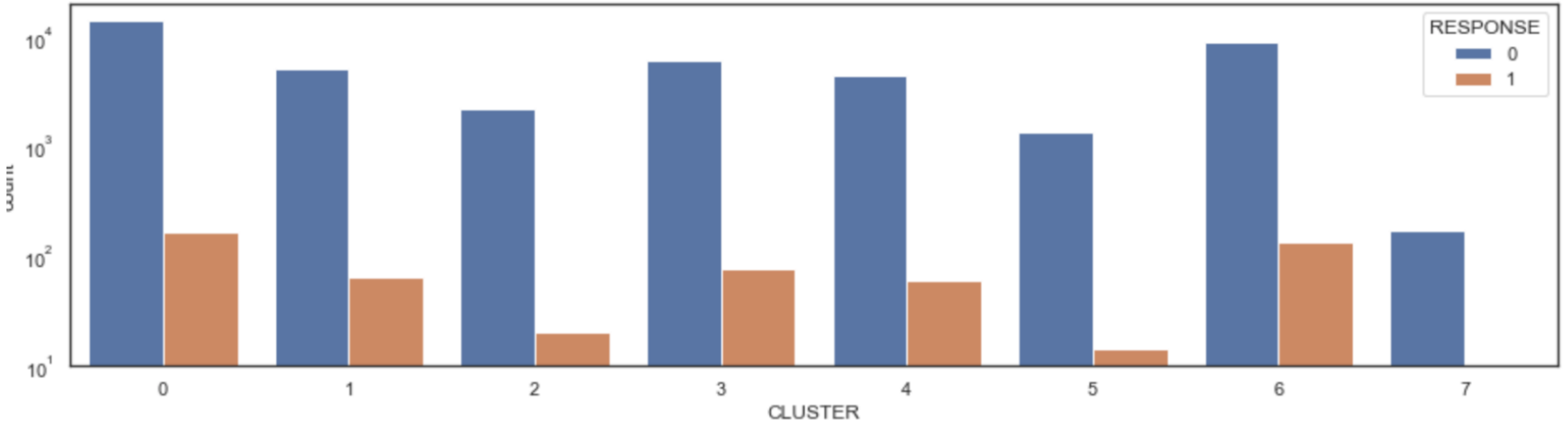
4. Testing data:

Demographics data for individuals who were targets of a marketing campaign
New subset of the customer dataset
42 833 persons (rows) x 366 (columns)
'RESPONSE' column was omitted

PART 2 – MAILOUT CAMPAIGN PREDICTION

When combining the clustering result from Part A with the mail-out campaign responses from the provided training dataset, it can be observed that no individuals from cluster 7 responded to the campaign, whilst most individuals from cluster 0 responded. This fits with the previous analysis where cluster 7 where the most under represented and cluster 0 where the most over represented. This indicate the clustering results are plausible.

Response to mailout campaign by cluster

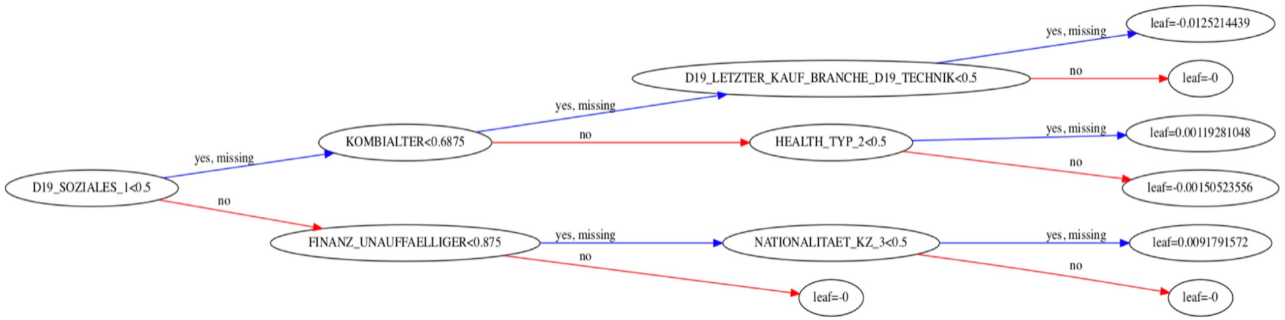


PART 2 – MAILOUT CAMPAIGN PREDICTION

XGBoost, a decision tree main algorithm, was chosen
XGBoost involves creating and adding trees to the model sequentially.
Each new tree correct the errors in the previous tree

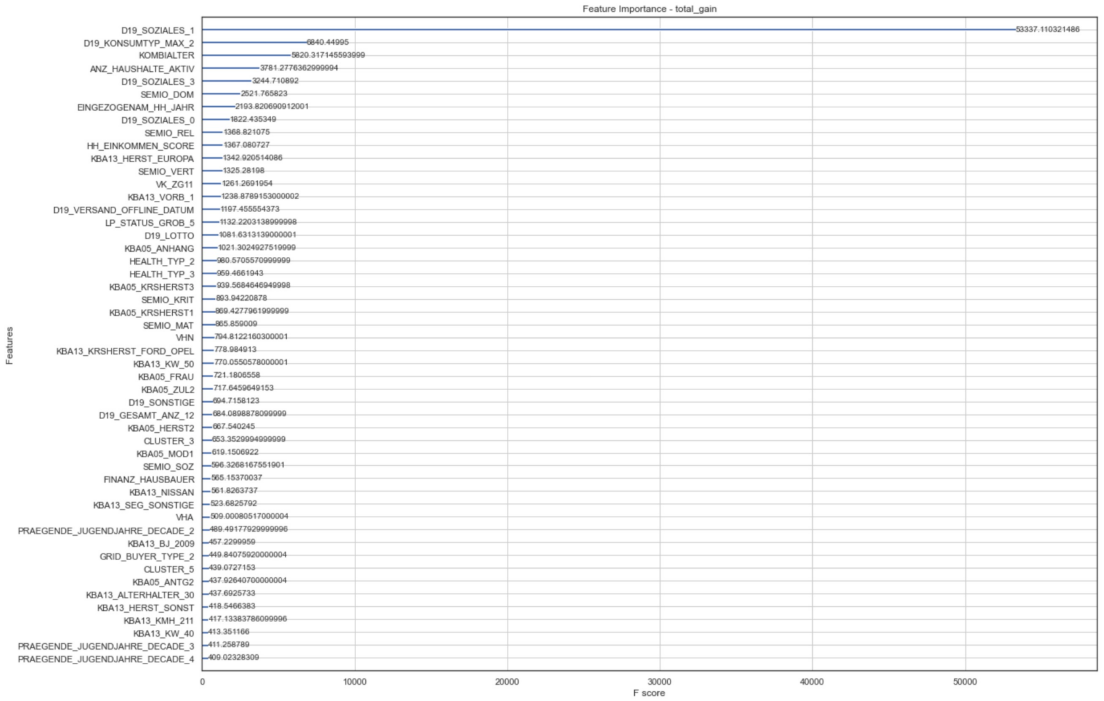
One key strength of the algorithm is its ability to apply regularisation on the features, meaning it could get rid of useless features or shrink importance of features that cause bias.

As the datasets have over 400 features in the end, this becomes extremely important. XGBoost is a competition winning algorithm on Kaggle



Visualising tree 0 of 800 gives a deeper understanding how the end-solution could look like. Over 800 such trees were created, with maximum depth of 2.
Tree complexity are quite high with 800 trees.

Feature importance, using total gain



Gain is the improvement in accuracy brought by a feature to the branches it is on. The Gain is the most relevant attribute to interpret the relative importance of each feature.

The most important features as seen in Fig 4.2 are D19_SOZIALES_1, D19_KONSUMTYP_MAX_2 and KOMBIALTIER are not described in the data provided.
D19_SOZIALES_1, using intuition, might describe if a person is active on social media.
D19_KONSUMTYP_MAX_2 can be inferred from D19_KONSUMTYP which are consumption type, of which 'Gourmet' seems to be the most dominant after 'Universal' and 'Versatile'
'ANZ_HAUSHALTE_ACTIVE' describe the households in the building
I suspect KOMBIALTIER is related to age due to it's strong correlation with GEBURSJAHR

PART 2 – MAILOUT CAMPAIGN PREDICTION – TRAINING RESULTS

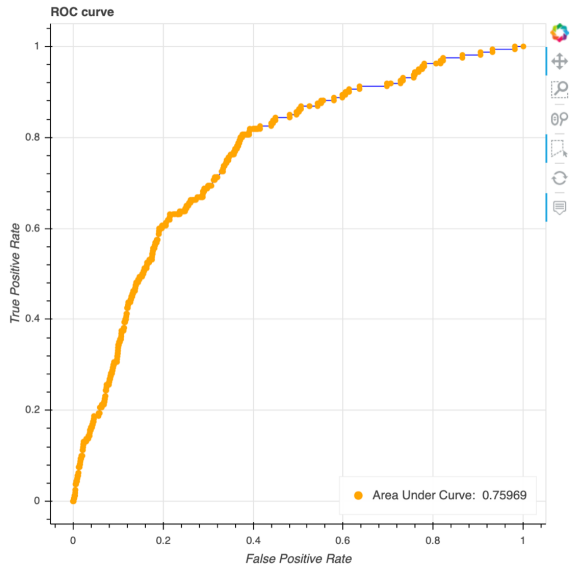
Training results from 12889 individuals

	Predicted **0**	Predicted **1**
Actual **0**	TN *9113*	FP *3616*
Actual **1**	FN *44*	TP *116*

TPR = TP / TP + FN = 115 / (115 + 45)

FPR = FP / FP + TN = (3721 / (3721 + 9008)

Performance metrics AUC (Area under the Curve)



AUC graph plots the ratio of positive errors over negative errors. Its goal is reduce the overall amount of errors made. The curve plots the ratio of True Positive rate over the False Positive rate
TRP = True Positives / All Actual Positives
FPR = False Positives / All Actual Negatives

77.1 % during training

PART 2 – MAILOUT CAMPAIGN PREDICTION – FINAL CONCLUSION

The final result of 79.9% AUC on the testing dataset in kaggle competition was achieved, and is improving the original benchmark score of 64% by 15%. As not a lot of students score over 80%, I think this is a fantastic result giving the complexity and quality of the dataset.

A private score of 76.3% AUC were achieved when only 70% of the testing set was used, which indicate the model has a consistent precision. I would have been 11th position of 438 students if I made the competition deadline.

Submission and Description	Private Score	Public Score
kaggle.csv just now by Juanita Smith	0.76349	0.79852

3. How can the mail order company acquire new clients more efficiently ?

All persona's in top 5 PCA groups explaining around 25% of the population variance are actually active online customers, even if they are from low income groups and from all age groups. We are living in a digital age, where most individuals do online shopping, just that the mail-order company is not attracting those individuals from lower income groups, especially young families just starting out as described in the most under represented cluster 7 who did not respond at all to the campaign.
The mail order company should consider to offer more competitive products and focussed targeting of individuals in cluster 7 as they do seem to love shopping after all and are searching for great offers.

Fig 5.2

Refer to PCA component 2 in Fig 3.9 as reminder of the full persona description.