@ NYC Data Science Academy

(15th September 2017)

Sito Mobile Big Data Analysis

by

Team Entropy

Yadi Li, Kumar Nathan, Wei Liu, Janet Hu and Andre Toujas

I Overview

II Exploratory Data Analysis

III Unsupervised Learning

IV Supervised Learning

V Conclusions and Future Work

I Overview

- **Goal:** Customer Segmentation to provide business insights
- **Data:** ~150MM observations in 500+ Parquets
- **Parquet:** ~250K observations & 962 features (551 numerical)
- **Strategy/Workflow:**
  a) Performed EDA on data set

  b) Assigned verticals to features

  c) Analyzed aggregate results in several parquets

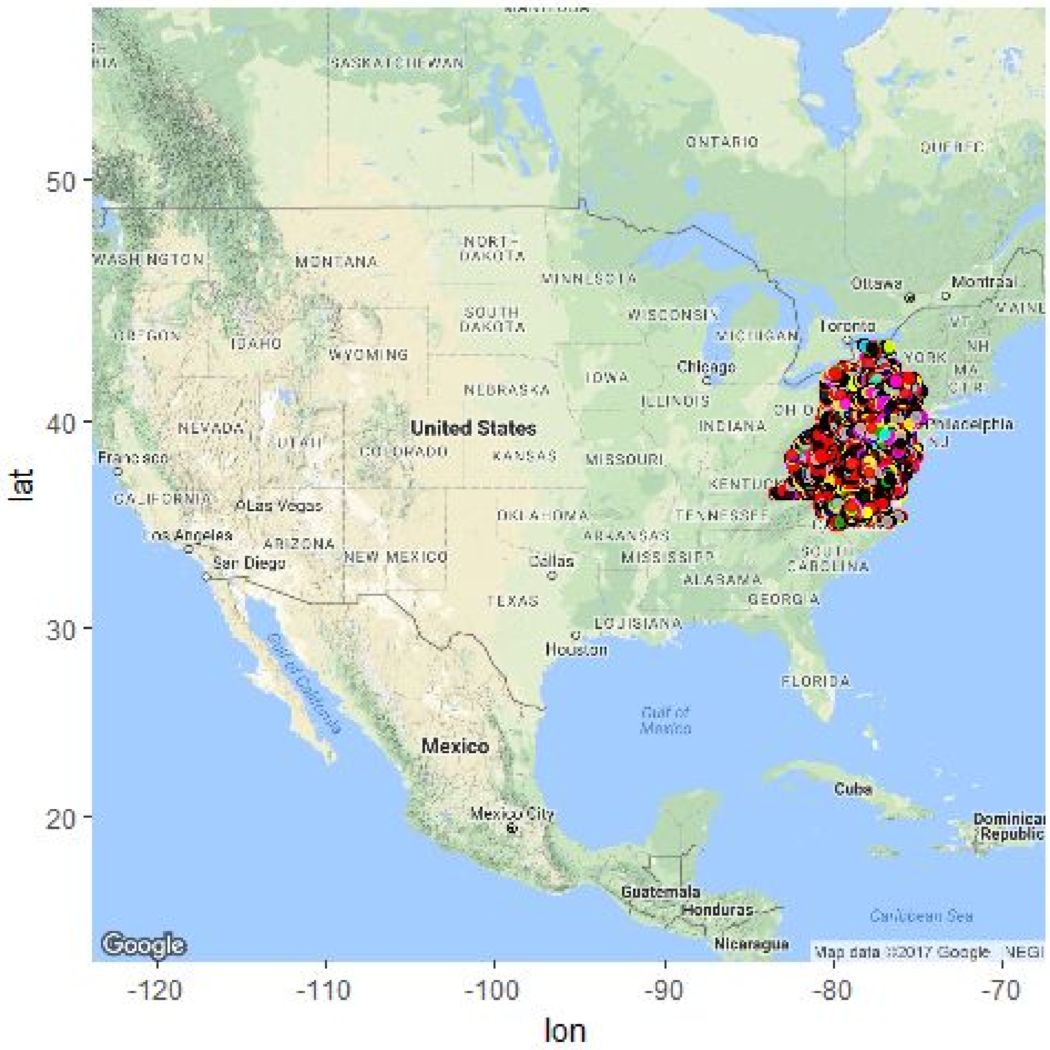
  d) Performed dimension reduction and clustering on entire parquets

  e) Performed dimension reduction and clustering on each vertical category

  f) Performed supervised learning to answer specific business-related questions

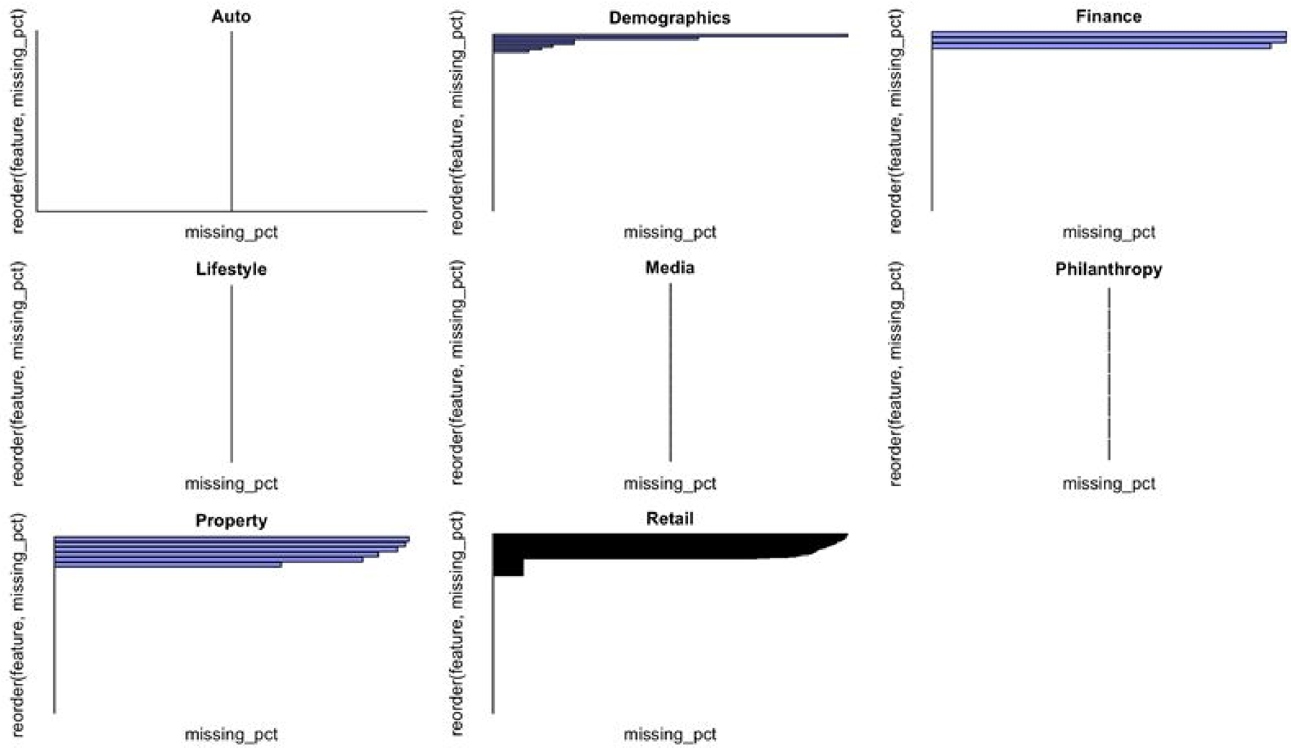
1 Exploratory Data Analysis

- Geospatial location of observations in Parquet 222 [similar pattern for other Parquets]



II Exploratory Data Analysis

- Missingness in the Dataset [Parquet 222]

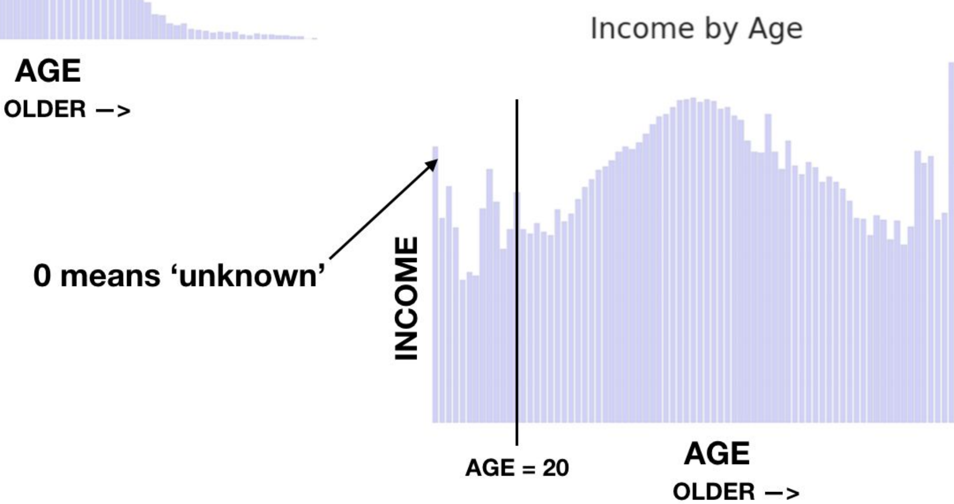


551 Numeric columns categorized into 8 verticals

II Exploratory Data Analysis

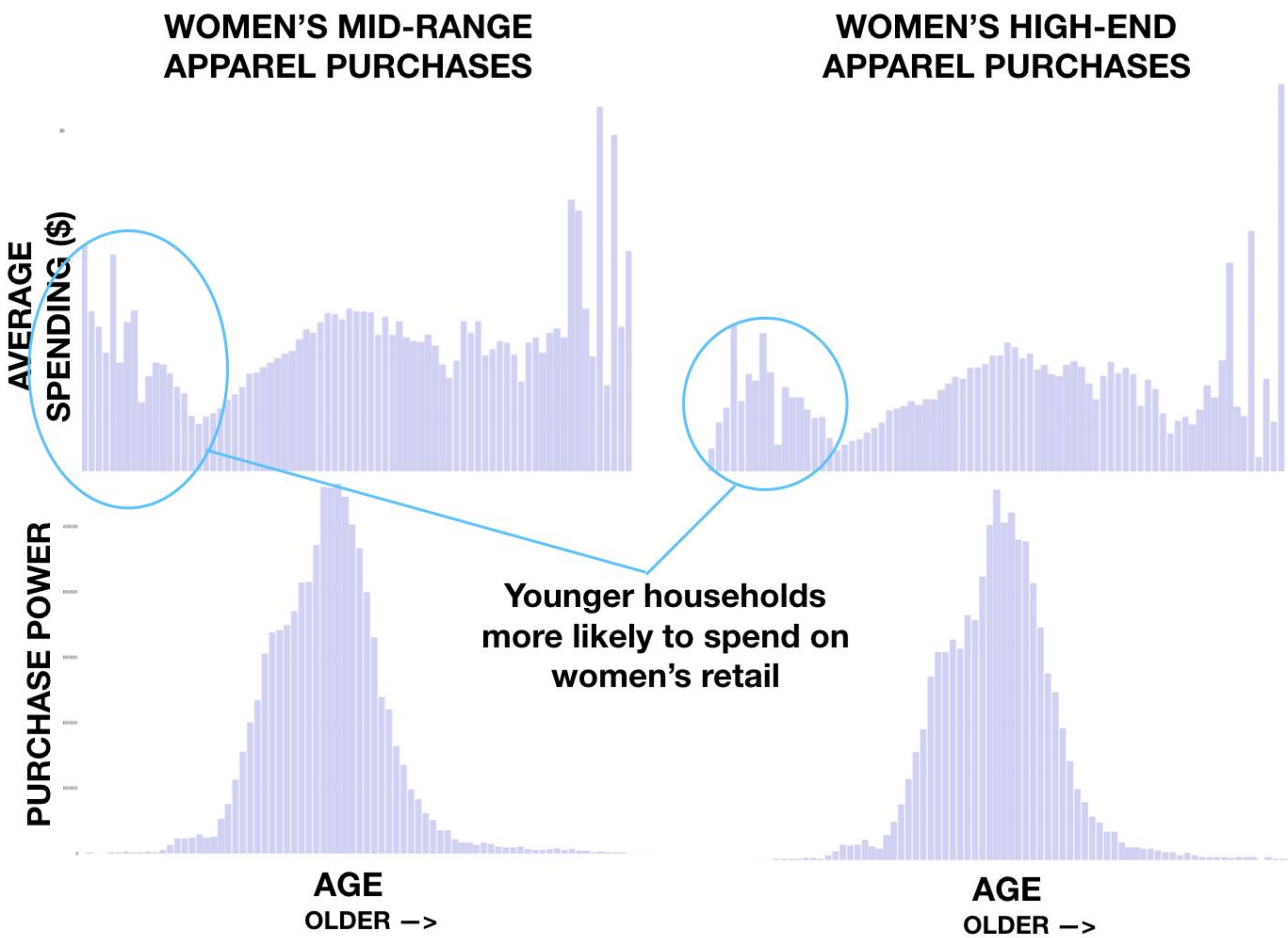
- Used grouping by household median age for example
- Example answers business-related questions
- Plotting functions created in Databricks for quick visualizations for future analysis

Key Takeaways:
- **More middle-aged households in data set**
- **Households with age < 20 have higher median income**
  - **Data needs scrubbing**



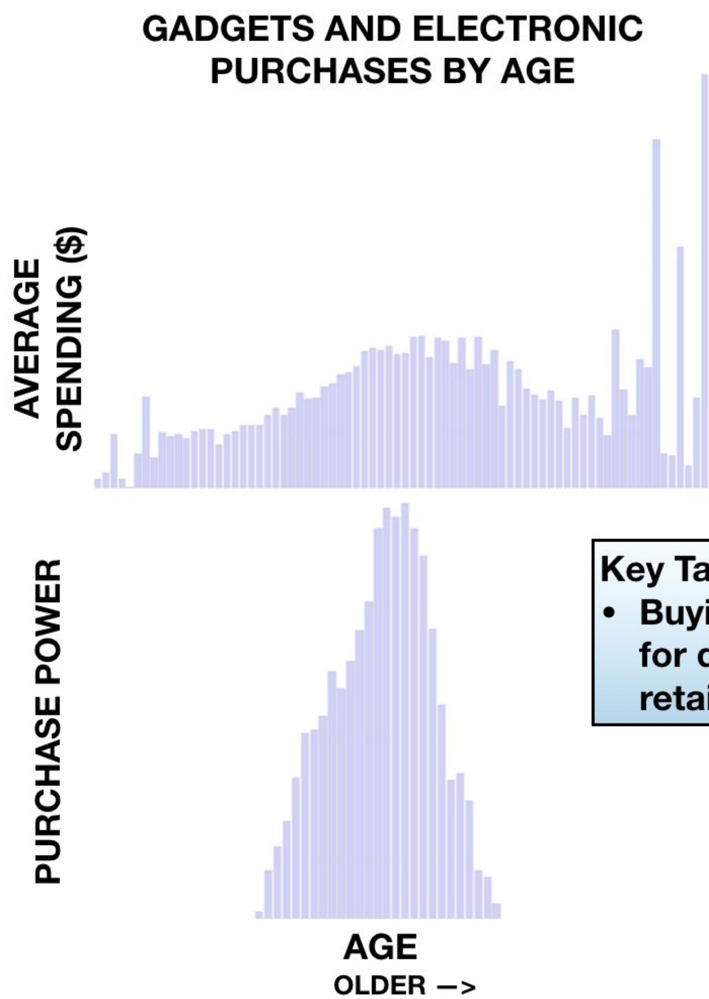
II Exploratory Data Analysis

- Women's Retail



II Exploratory Data Analysis

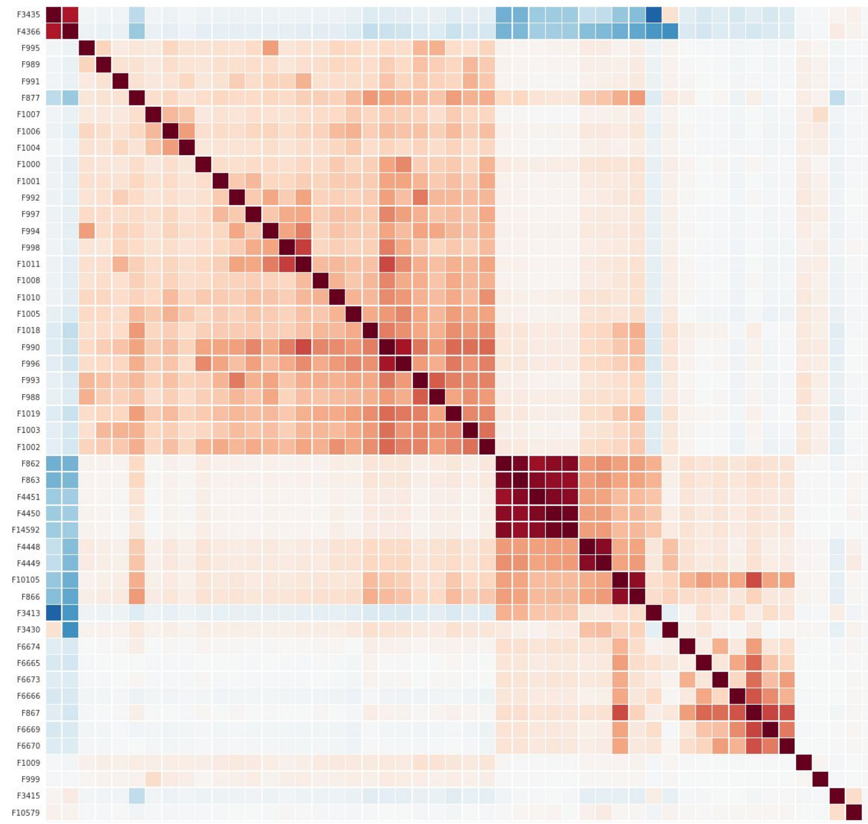
- Electronics



Key Takeaways:
- **Buying trends are similar for different categories of retail**

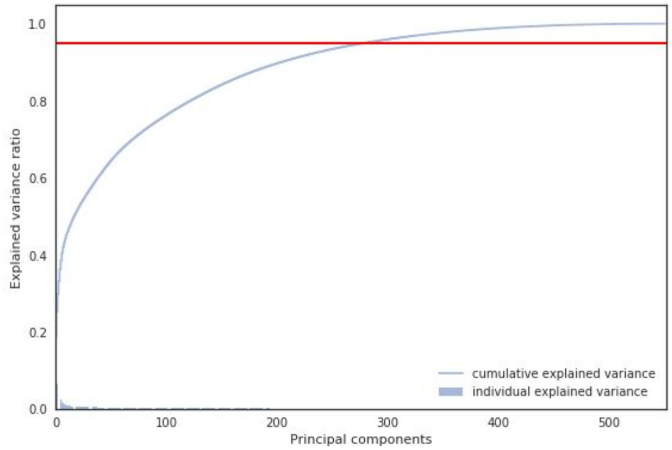
II Exploratory Data Analysis

- Correlation Analysis [Parquet 222]: Only the first 50 columns displayed

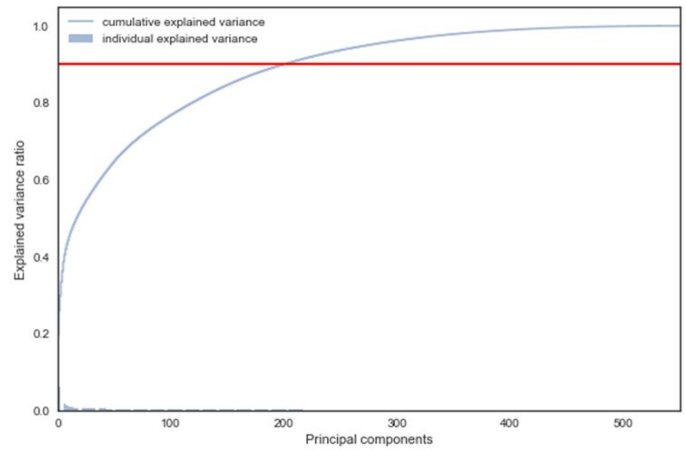


III Unsupervised Learning

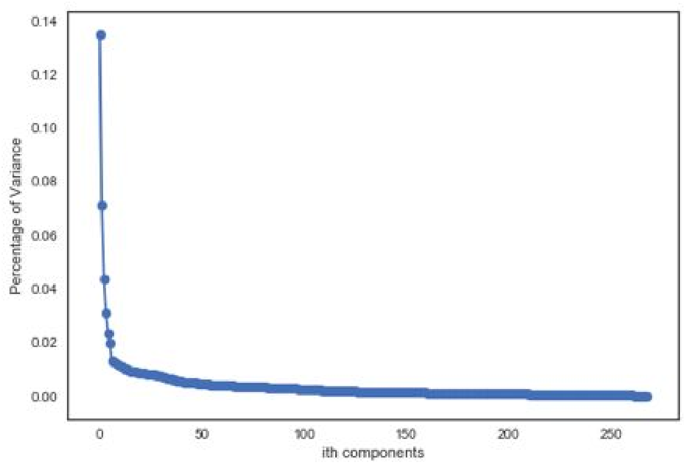
- Principal Component Analysis on all the numerical features in the Parquet



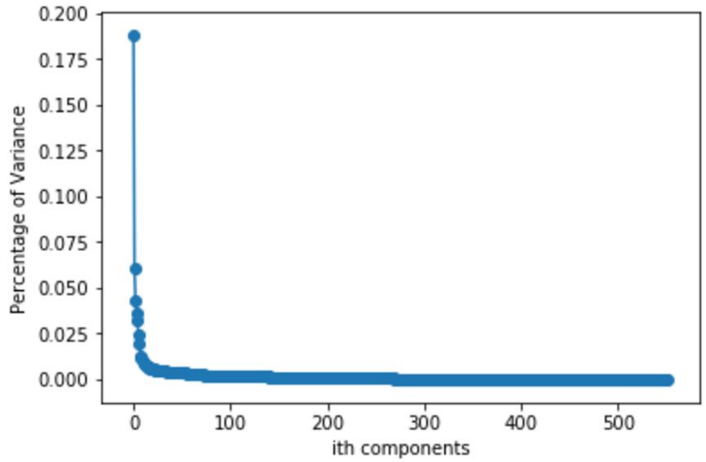
Parquet 11
204 out of the original 551



Parquet 99
201 Principal Components out of 551



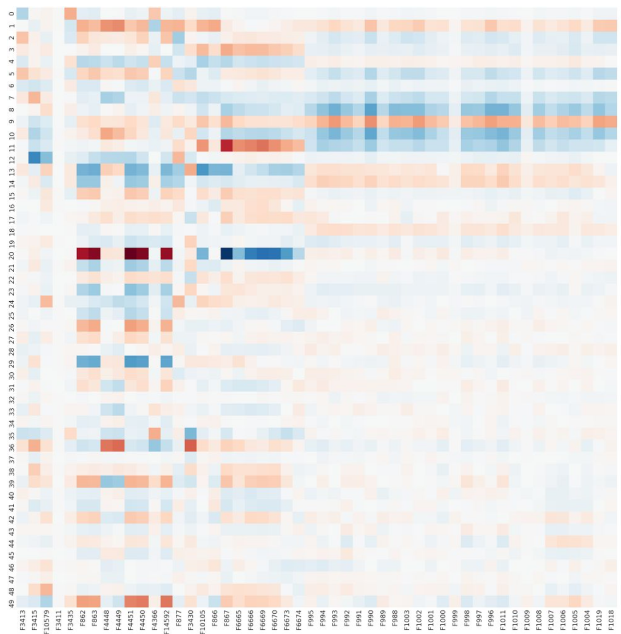
Parquet 20
204 out of the original 562



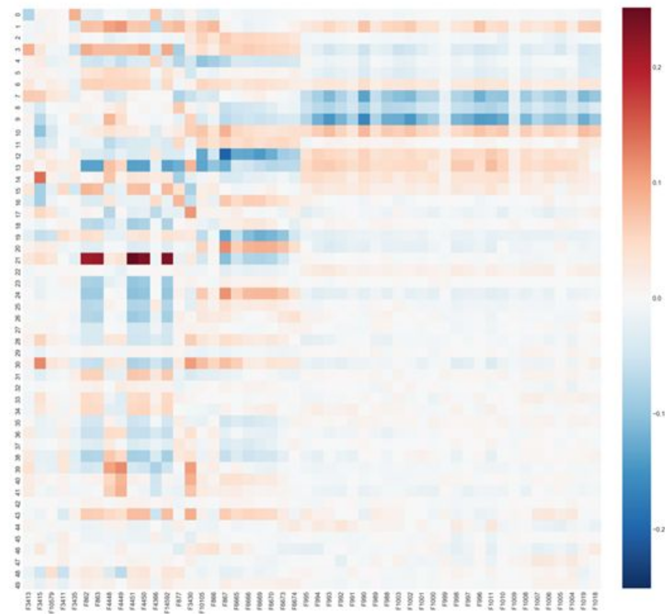
Parquet 490
194 out of the original 562

III Unsupervised Learning

- Principal Component Analysis on all the numerical features in the Parquet



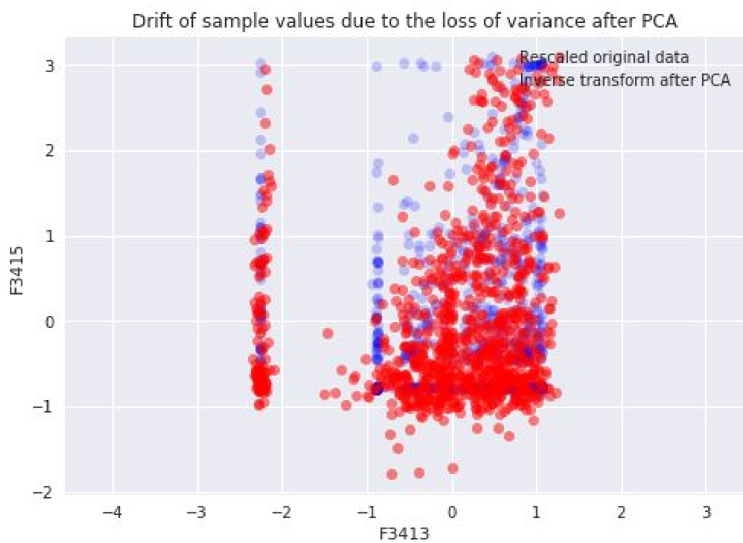
Parquet 11



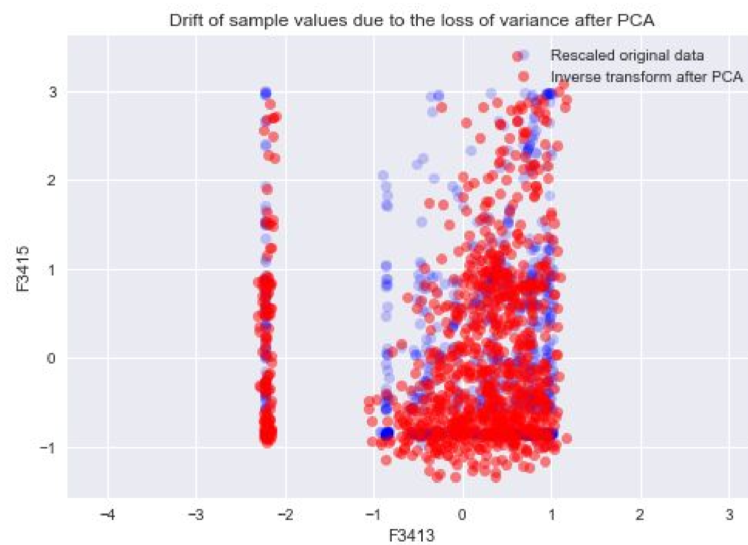
Parquet99

III Unsupervised Learning

- Principal Component Analysis



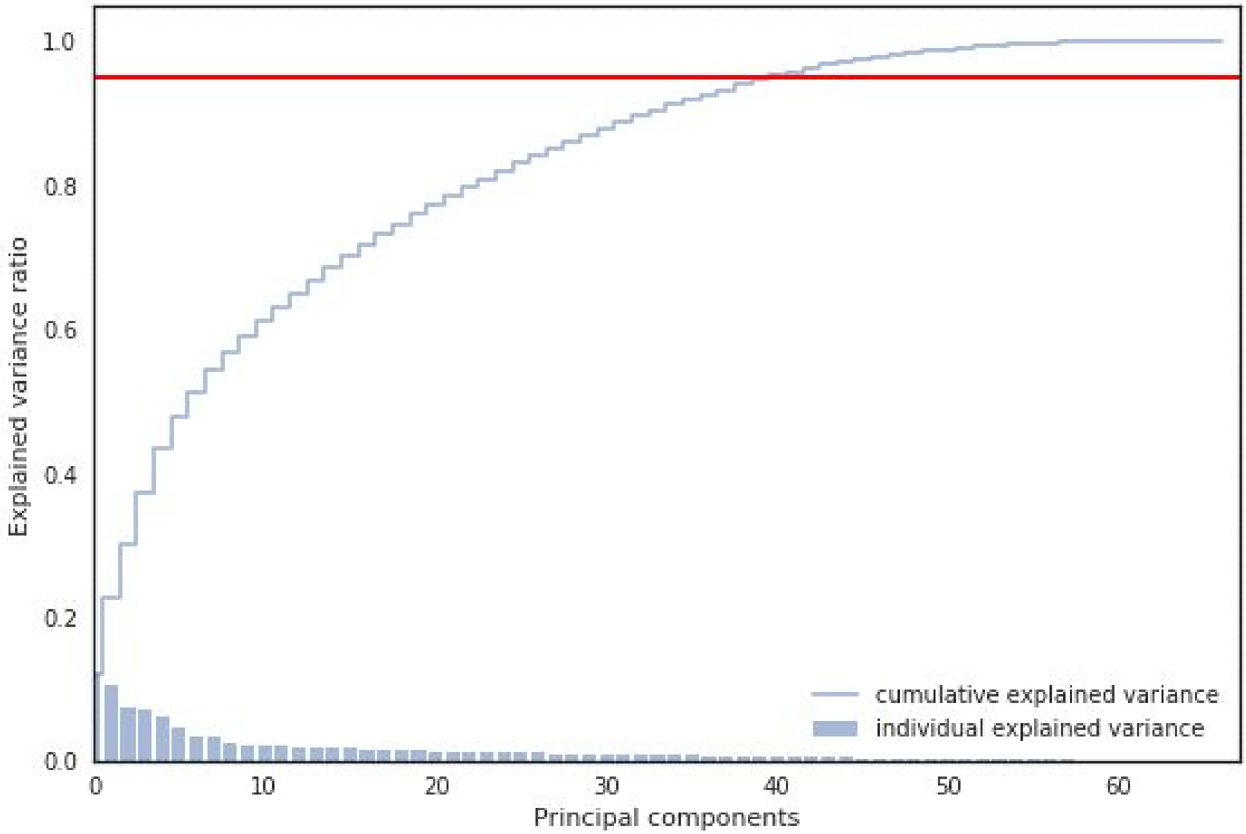
Parquet 11



Parquet 99

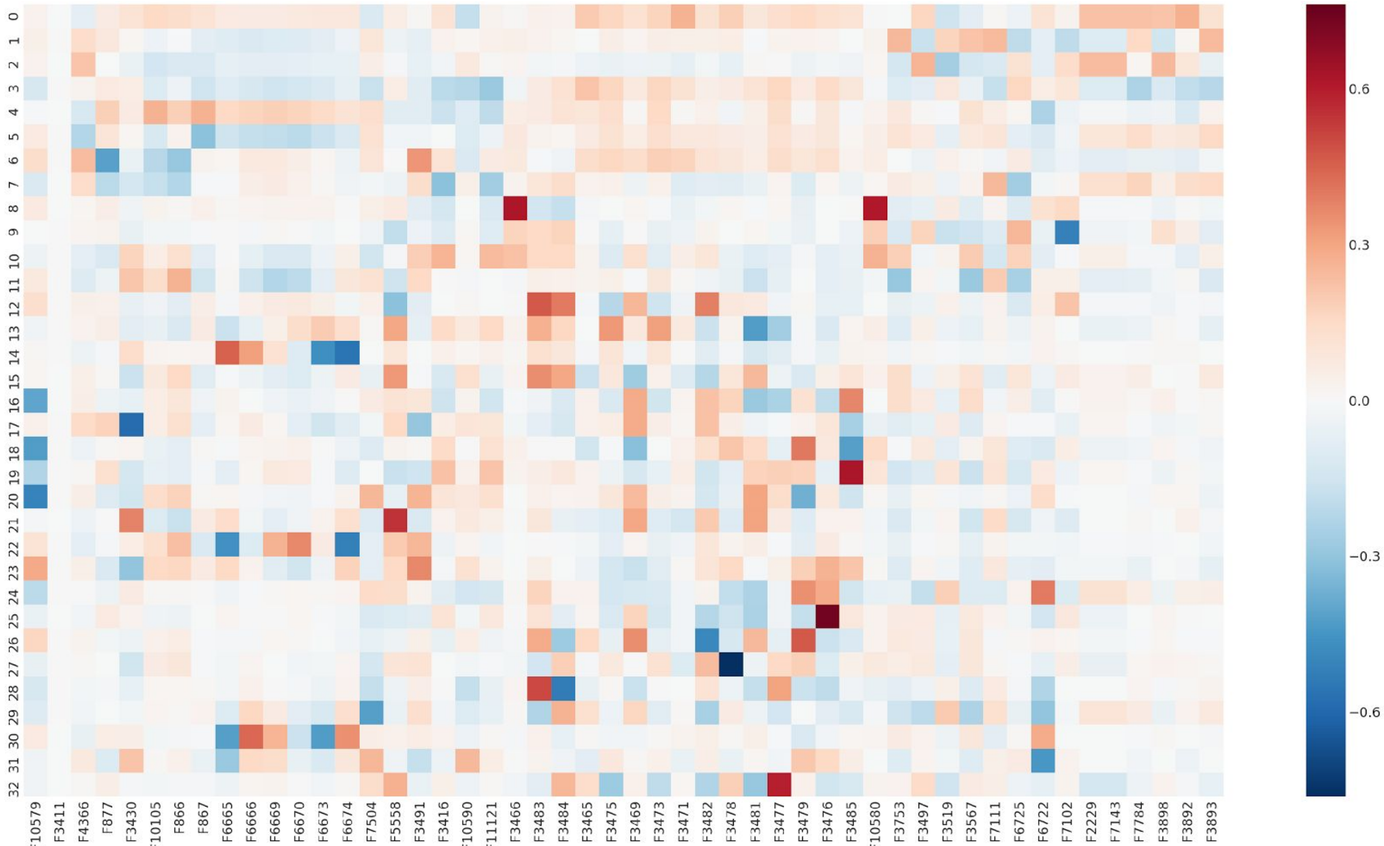
III Unsupervised Learning [Vertical--Demographics, Paquet 222]

- Principal Component Analysis [Partition 222]



III Unsupervised Learning [Vertical--Demographics, Paquet 222]

- Principal Component Analysis



1st pc:

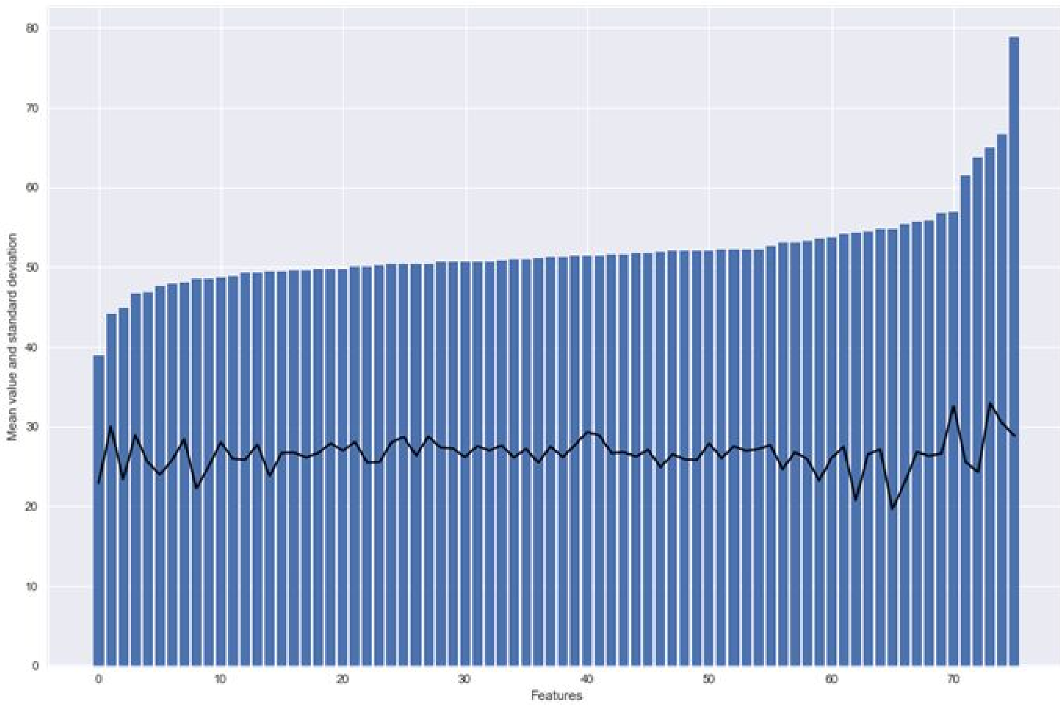
- Number of Bathrooms
- Percentage of Households that are Married Couple Family Households with Presence of Persons Under 18

Lifestyle Features

- **Model predicted likelihood of households interests/activities**
  - 63 columns
  - Numeric features, 0-99
  - 1: highest likelihood, 99: lowest likelihood
  - 0: unknown, replaced with 50
- **Real data of household interests/activities**
  - 202 columns
  - Apparently from two different sources: self reported and survey data
  - Categorical features
  - Y: Yes(6.5%), U: unknown(66.6%), NA(26.9%)

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

EDA on numeric features



On average more likely:

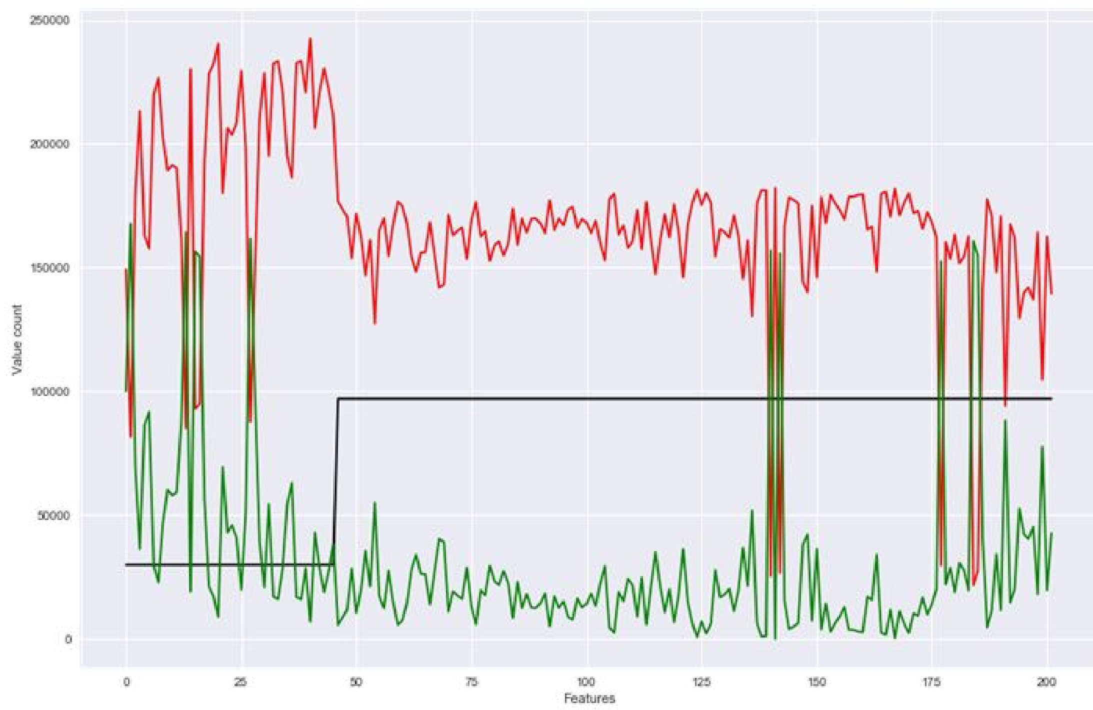
- Hunting Enthusiasts
- Sweepstakes/Lottery
- Do-it-yourselfers

On average Less likely:

- Medicare Policy Holders
- High Frequency Business Traveler
- Have Grandchildren
- Military – Active
- Military - Inactive

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

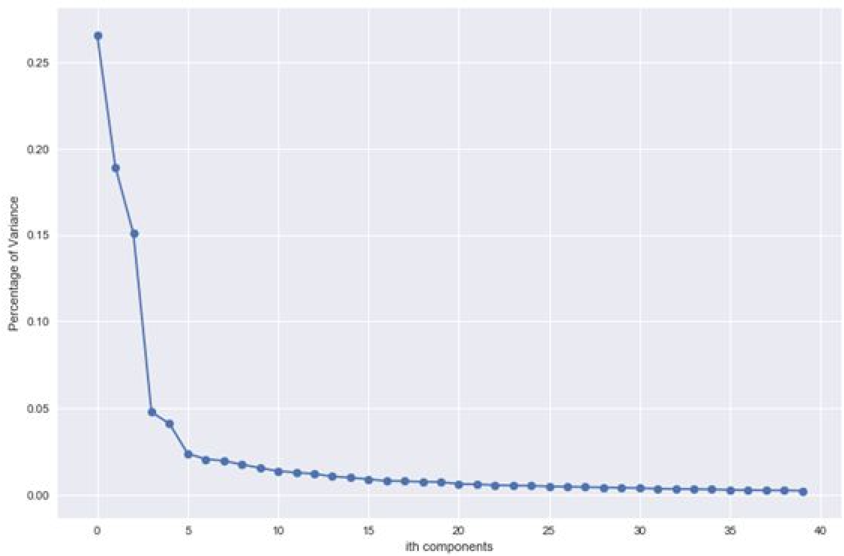
EDA on categorical features

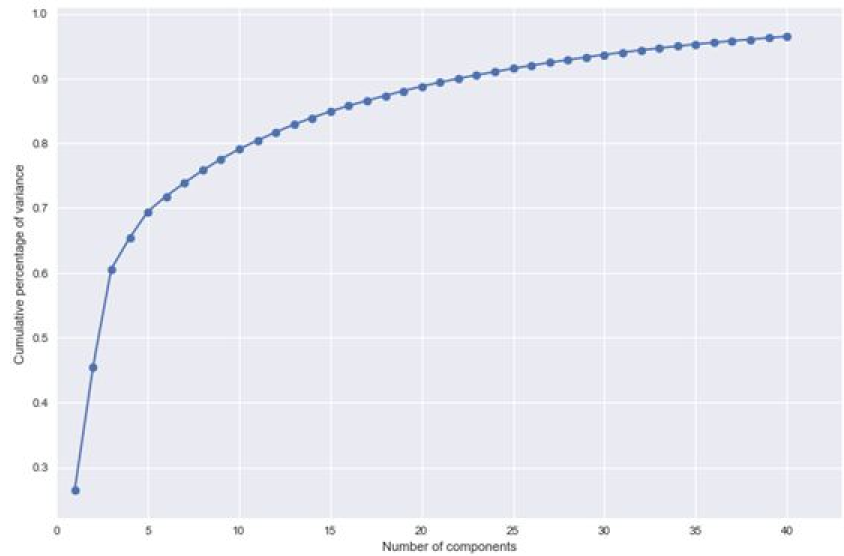


- Data clearly from two sources
- Features that have more "Yes":
  - Computers/peripherals
  - Purchased through the mail
  - Internet/online subscriber
  - Purchase via online
  - Hi-tech owner
  - PC & Internet:Own Computer
  - PC & Internet:Use Internet Servi
  - Info/Buying:Purc By Internet
  - Buying:By Catalog
  - Buying:By Internet

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

PCA on numeric features

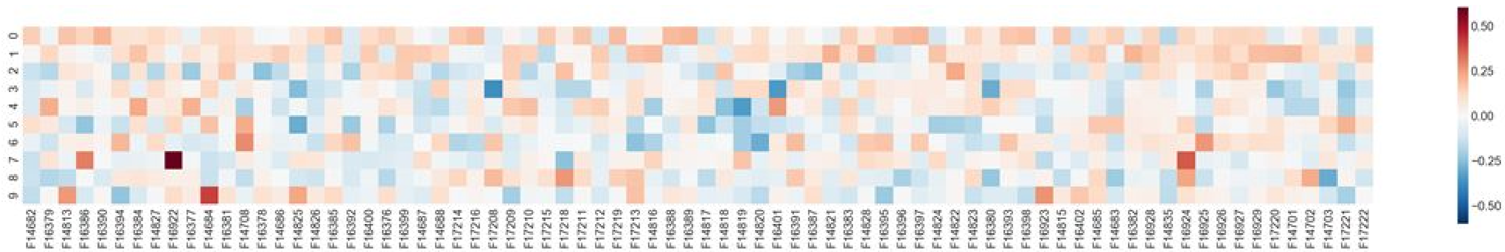




The first 5 principal components seem to capture significant amount of variance

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

Interpretation of principal components



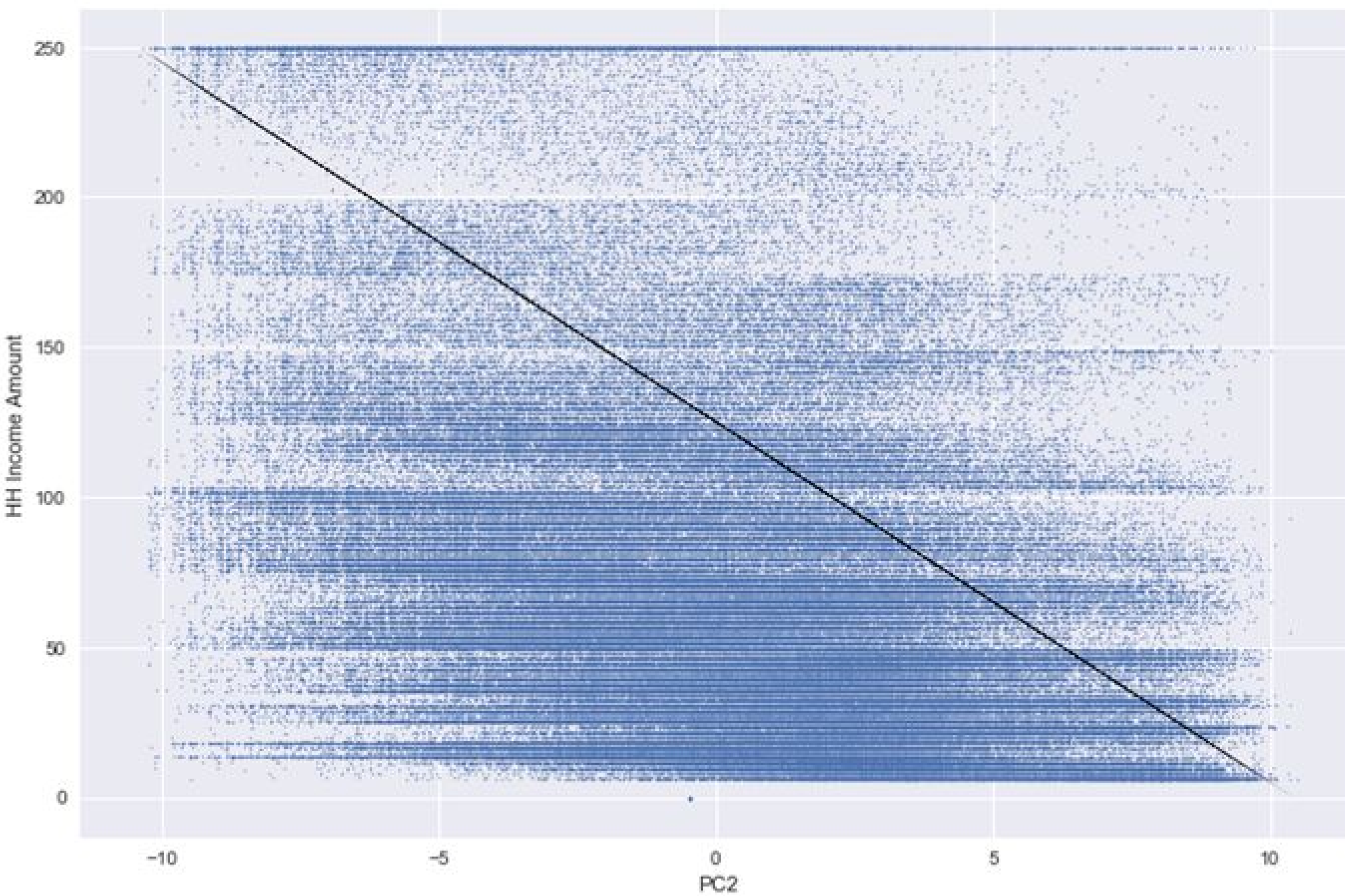
- **PC1: pop music, sports, education, young people**

  Attend/Order Educational Programs(F14813), Avid Runners(F16390), E-Book Reader(F16385), Listens to Alternative Music(F17216), Listens to Music(F17211), Listens to Pop Music(F17219), Music Download(F16388), Music Streaming(F16389), Plays Soccer(F16396), Plays Tennis(F16397), Snow Sports (F16393), Sports Enthusiast(F16398), Video Gamer(F14815), - Have Grandchildren (F14835)
- **PC2: travel, golf, rock music, middle age rich**

  Book Reader(F16384), Healthy Living(F16399), Listens to Rock Music(F17213), MLB Enthusiast(F14816), PGA Tour Enthusiast(F14821), Play Golf(F14828), Political Viewing on TV - Conservative(F14824), Gardening(F16382), Hotel Guest Loyalty Program(F16929), Interest in Religion(F17220), Life Insurance Policy Holders (F14701)
- **PC3: pet, outdoor, country music vs travel around the world**
- **PC4: religion, Christian music**
- **PC5: man vs woman**

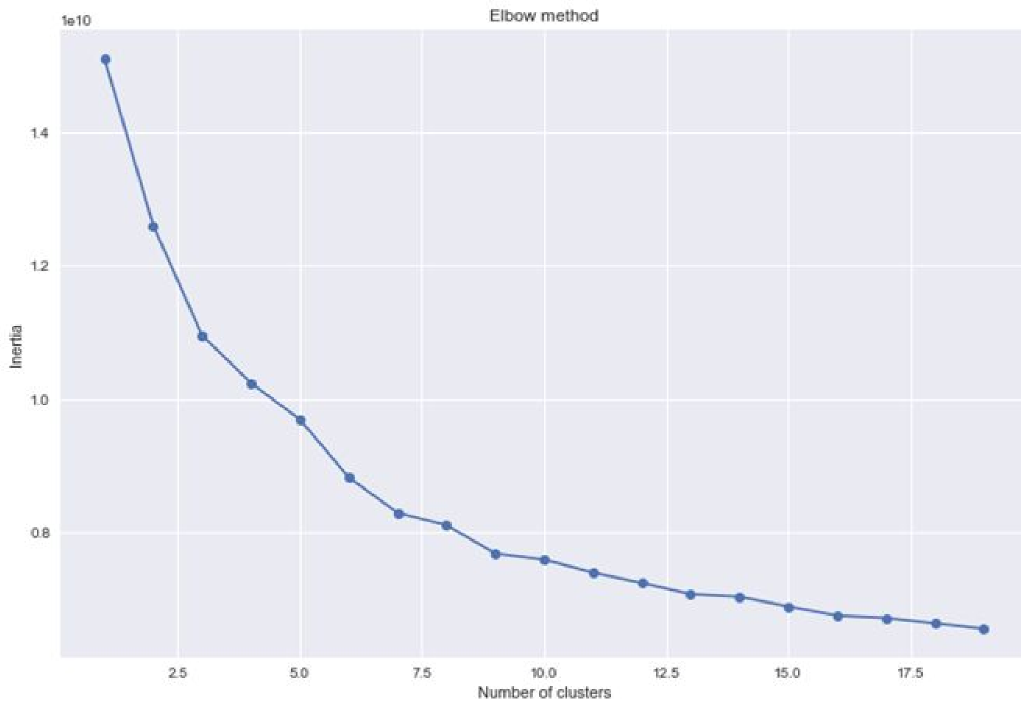
III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

PC2 vs household income



III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

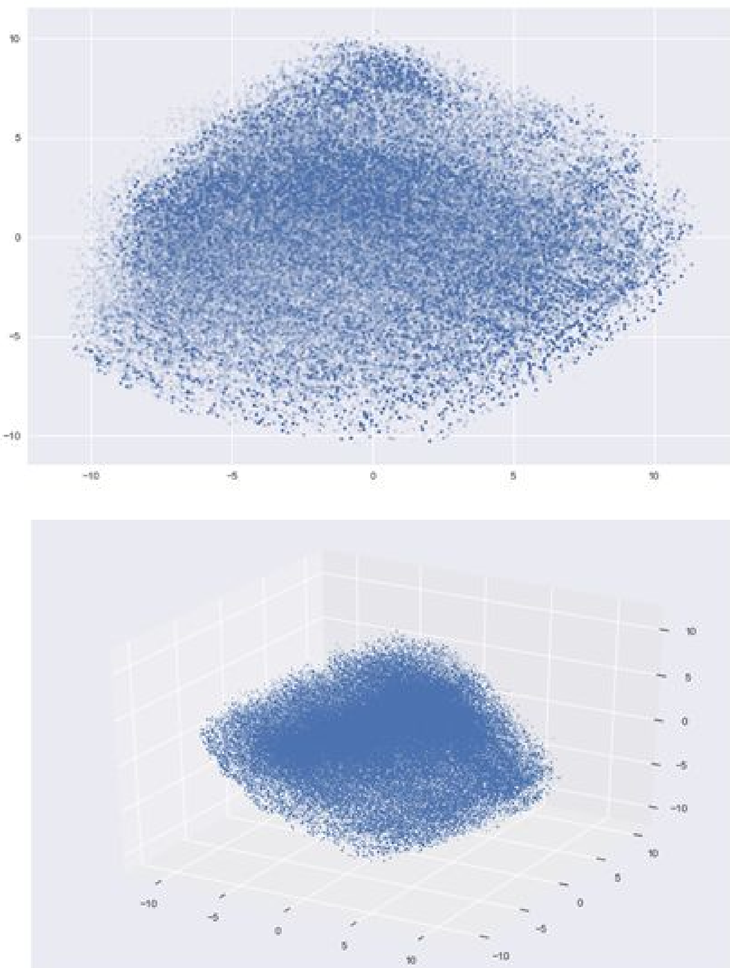
K-means Clustering using MiniBatchKMeans



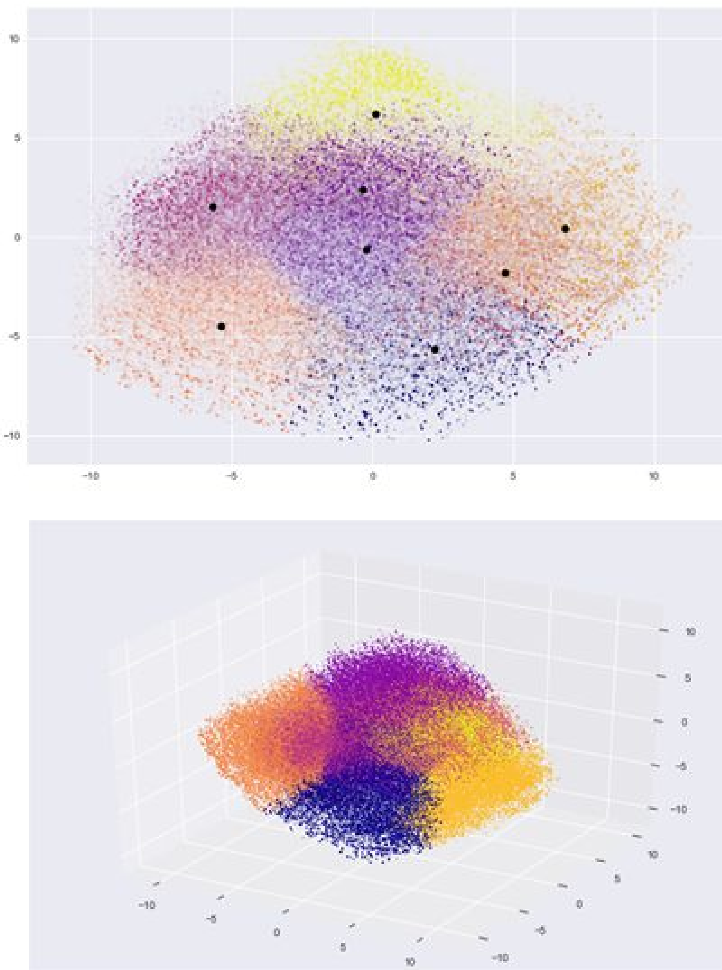
- MiniBatchKmeans is a variant of Kmeans
- uses mini-batches to reduce the computation time while still attempting to optimise the same objective function
- the quality of the results is reduced. In practice this difference in quality can be quite small

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

Clusters (K = 8)



Before



Clusters

III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

Visualization of clusters

- Households in cluster 1 and 2 have young people who lives In parents house or has moved out
- Cluster 3,4,5 are households of people finished education and just started working
- Cluster 6,7,8 are households of parents and grandparents

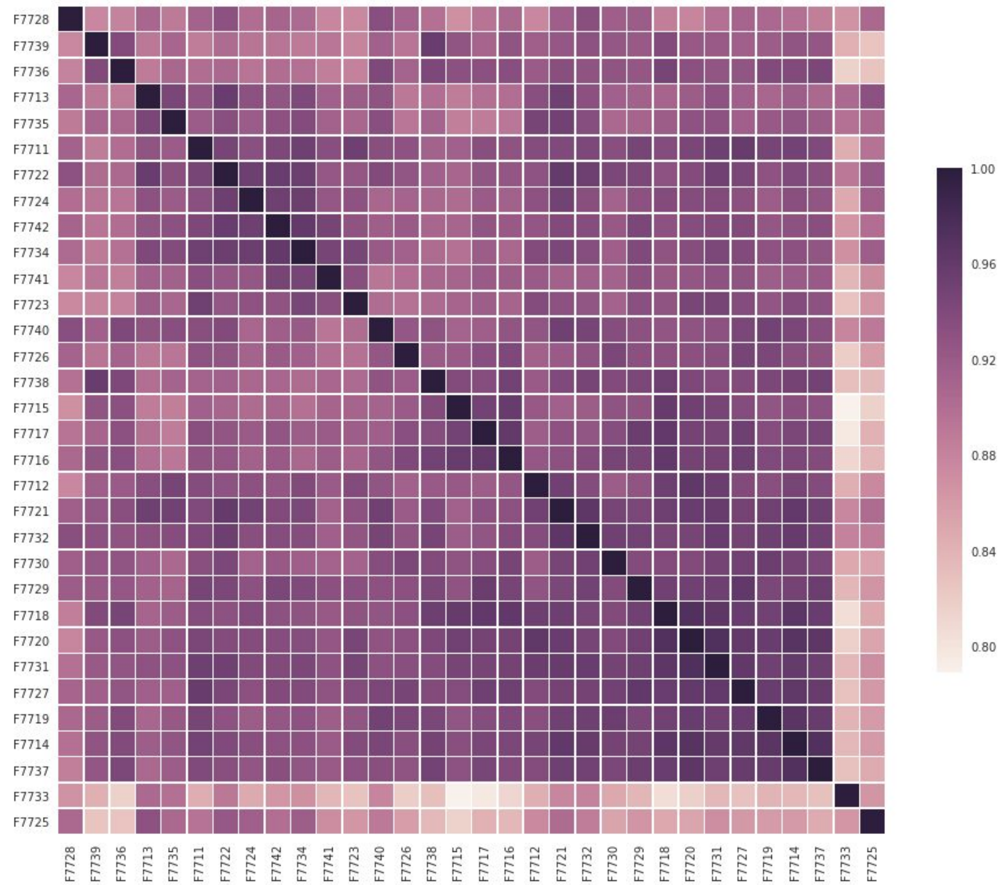
III Unsupervised Learning [Vertical--Lifestyle, Paquet 222]

Checking the clusters

clusters	F4810	F4850	F4855	F4871	F4874	F8975	F8977	F9253	F9269	F9273	pc1	pc2	pc3
5	0.866611	0.895821	0.862190	0.855124	0.851912	0.896501	0.894234	0.888641	0.893213	0.898201	-5.816368	-4.656116	0.561528
0	0.508644	0.529893	0.513614	0.497599	0.494694	0.702783	0.699974	0.695460	0.700766	0.705712	-5.605243	1.524159	-1.607674
1	0.560203	0.571352	0.555362	0.546069	0.541811	0.612880	0.610512	0.603604	0.608463	0.618799	-0.264535	-0.088037	0.030486
3	0.452824	0.473940	0.453136	0.420885	0.412682	0.645442	0.639734	0.627632	0.638673	0.661317	0.339158	6.311155	-1.778416
7	0.578416	0.601865	0.577011	0.557001	0.546512	0.726357	0.721778	0.704288	0.718023	0.729288	0.444676	1.731120	5.011385
6	0.886030	0.898659	0.865793	0.854616	0.848419	0.892854	0.888579	0.878539	0.884618	0.909130	1.226509	-5.884211	-2.491926
2	0.832396	0.832759	0.811532	0.784789	0.768860	0.850484	0.841776	0.817464	0.830056	0.867611	5.536839	-2.086107	2.810786
4	0.690555	0.659001	0.643330	0.612164	0.603324	0.738026	0.731018	0.718481	0.726194	0.797788	6.664705	-0.054646	-3.993693

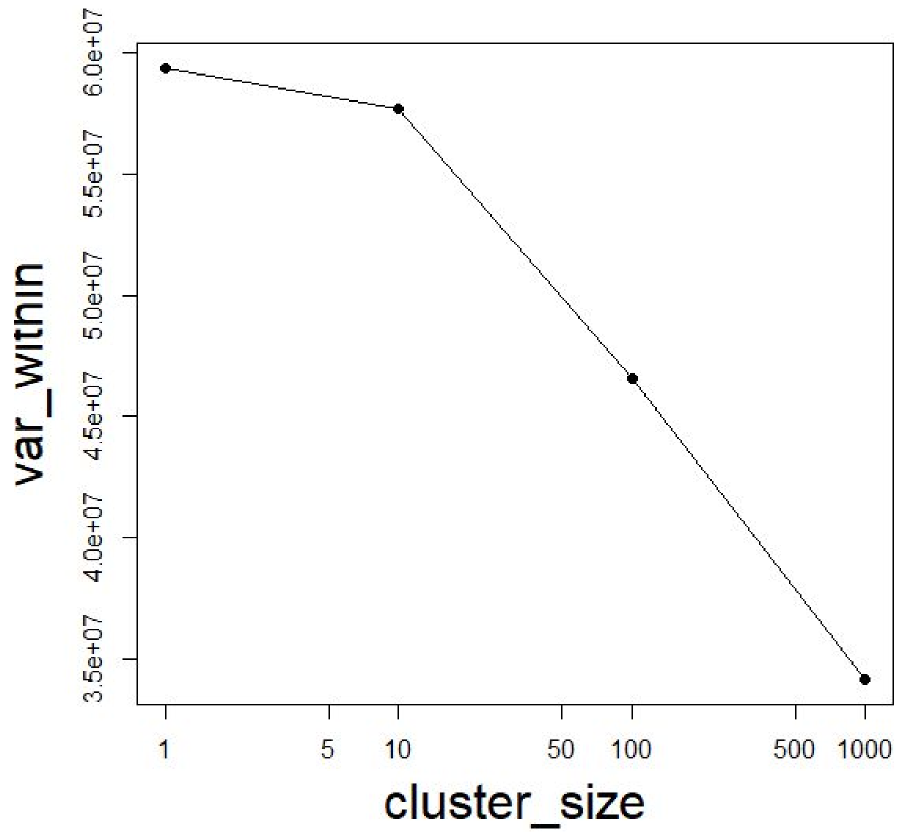
III Unsupervised Learning [Vertical--Auto, Paquet 222]

- Correlation Analysis



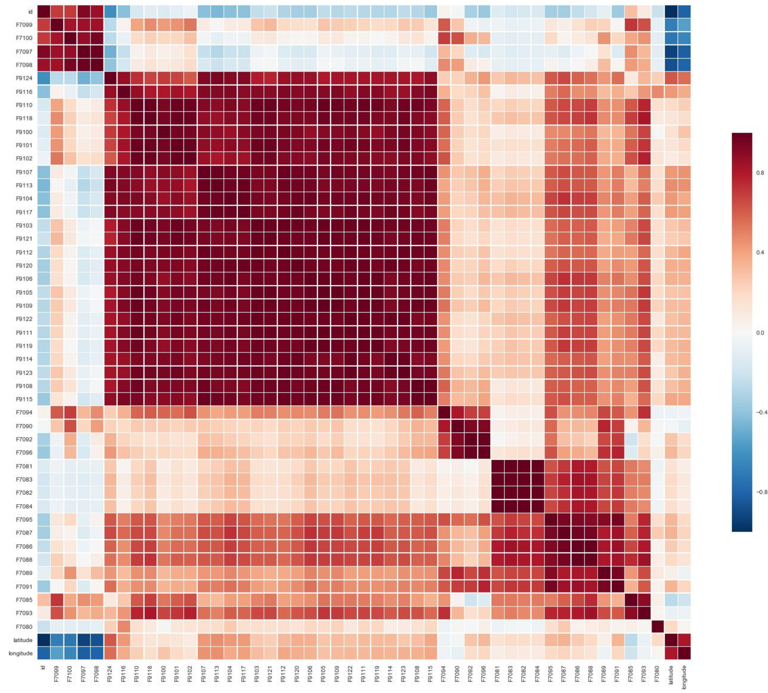
III Unsupervised Learning [Parquet 11]

- Cluster Analysis

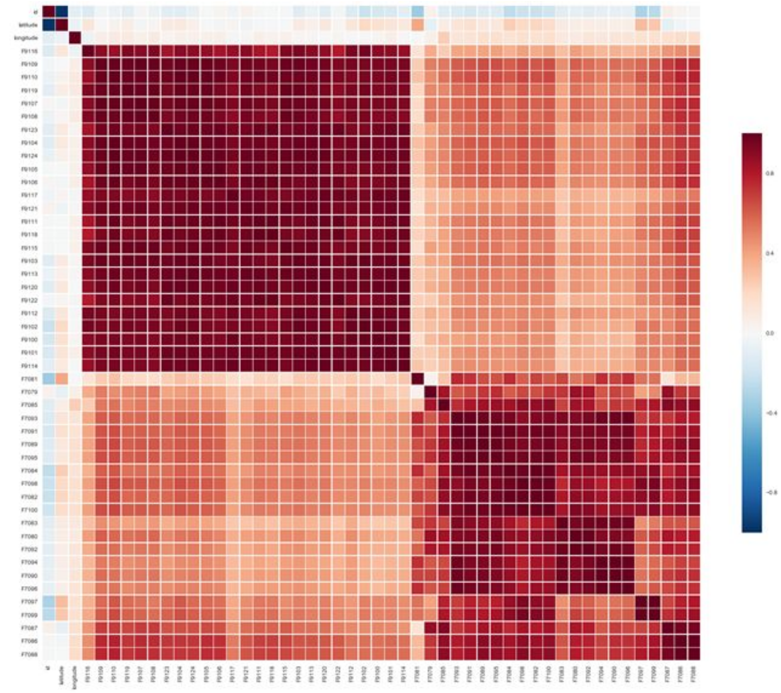


III Unsupervised Learning

- Cluster Analysis [k=14]



Parquet 11

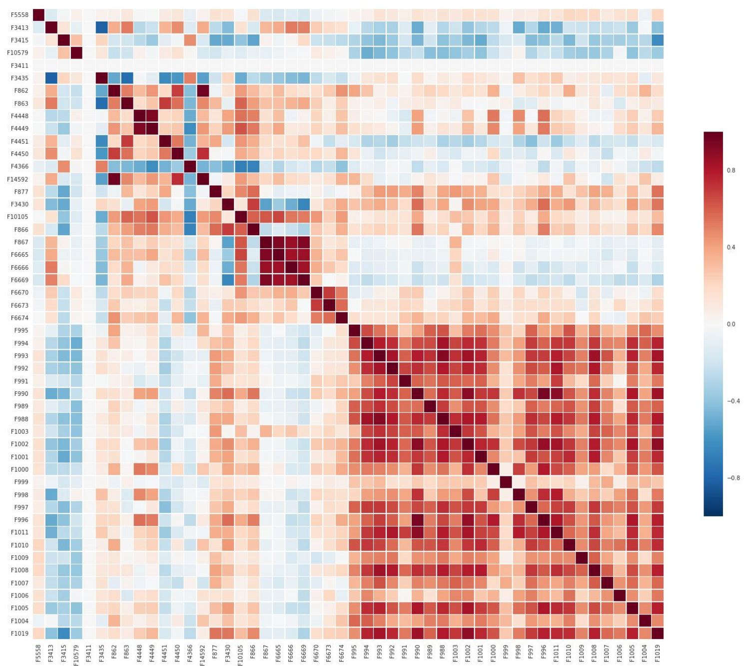


Parquet 99

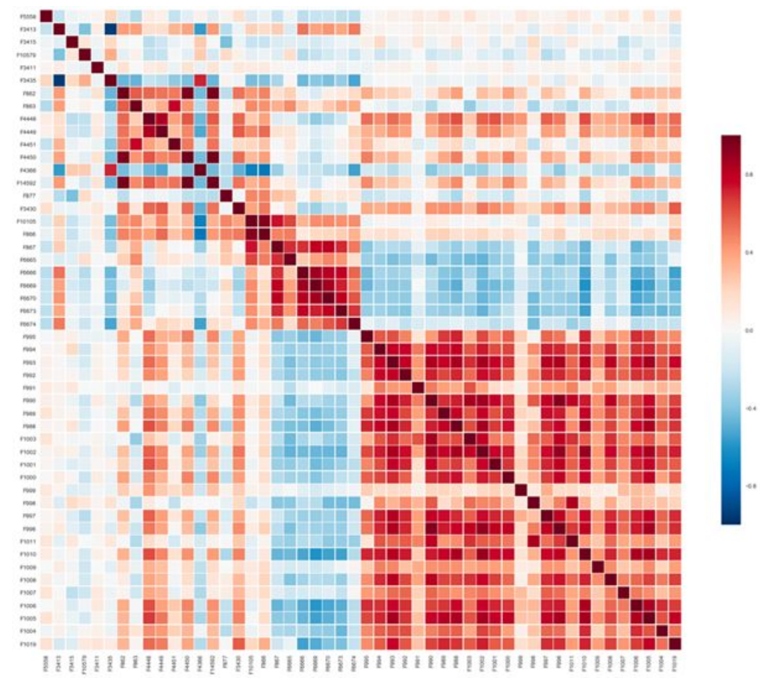
F9100 - F9124 Probability of purchase (clothes, furniture, pet supply, electronics)
F7079 - F7093 Total amount spent/ purchase frequency

III Unsupervised Learning

- Cluster Analysis [group by city, 102 distinct cities]



Parquet 11



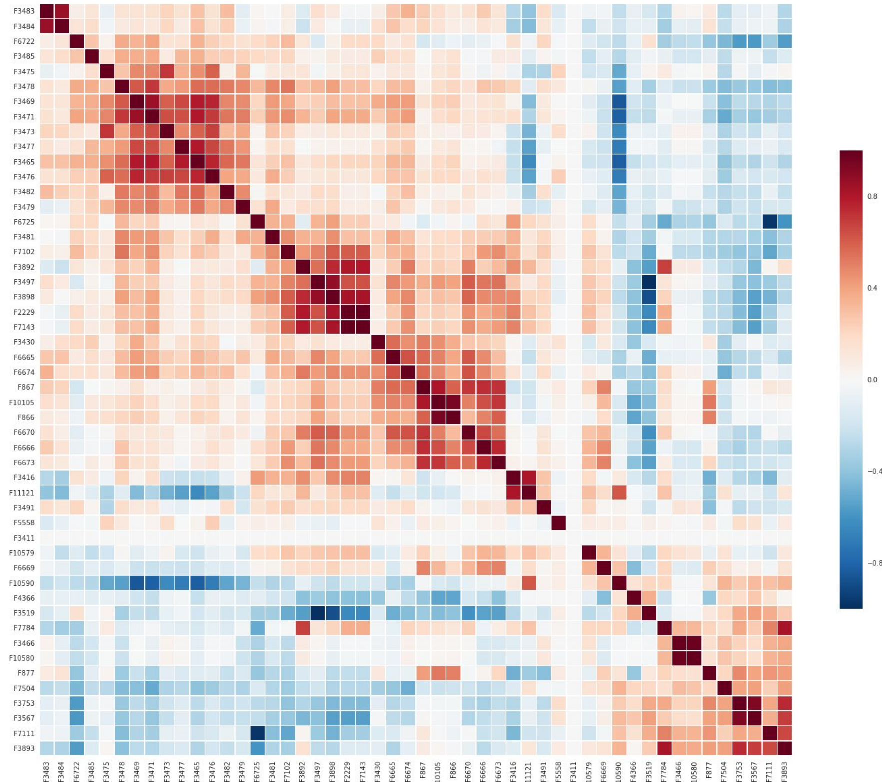
Parquet 99

F995 - F1019: Number of times different categories are purchased

- Different cities have different income level (NYC-Upscale merchandise)
- Can be expanded in more detailed geospatial segmentation(zip code, communities)

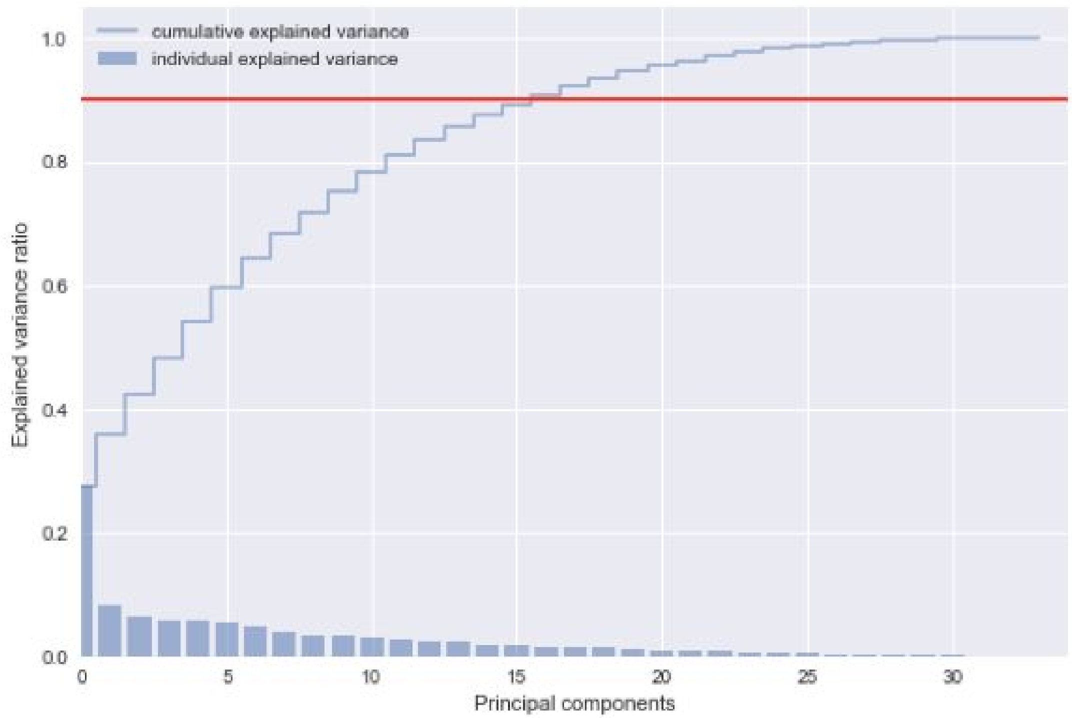
III Unsupervised Learning

- Cluster Analysis [correlation of mean demographics grouped by city]



III Unsupervised Learning [Vertical--Property, Paquet 222]

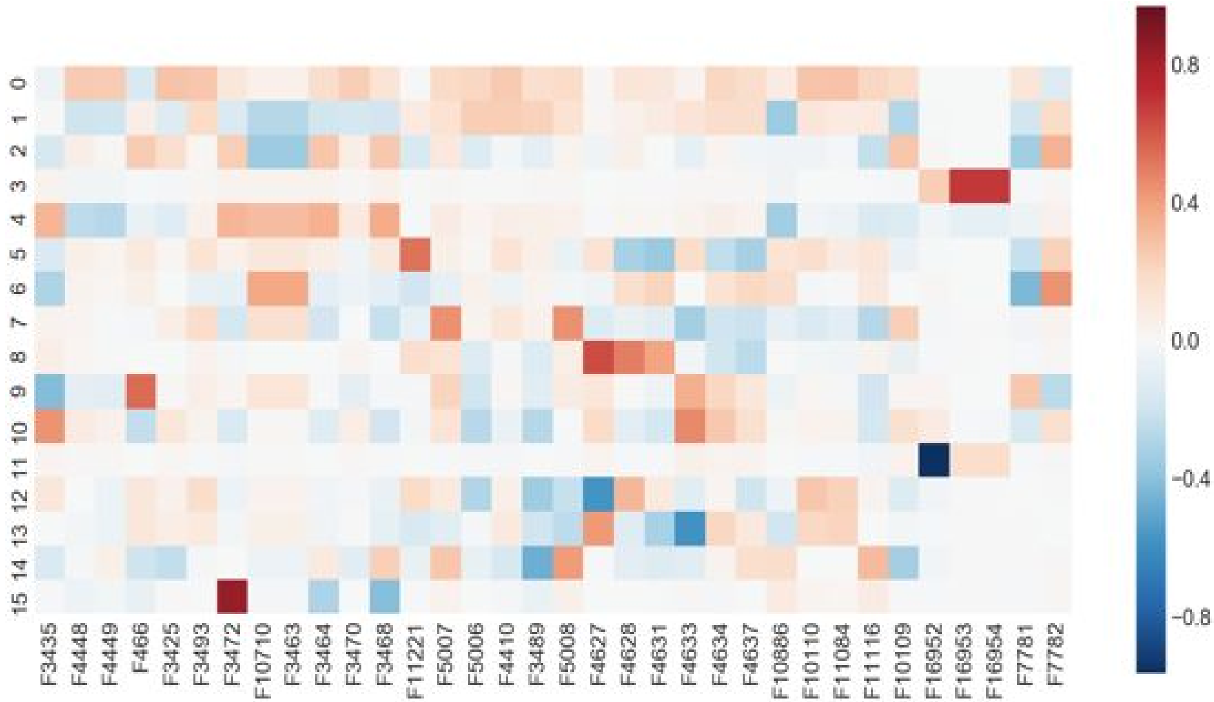
- Principal Component Analysis [Partition 222]



16 out of 34 explain 90% of variance

III Unsupervised Learning [Vertical--Property, Paquet 222]

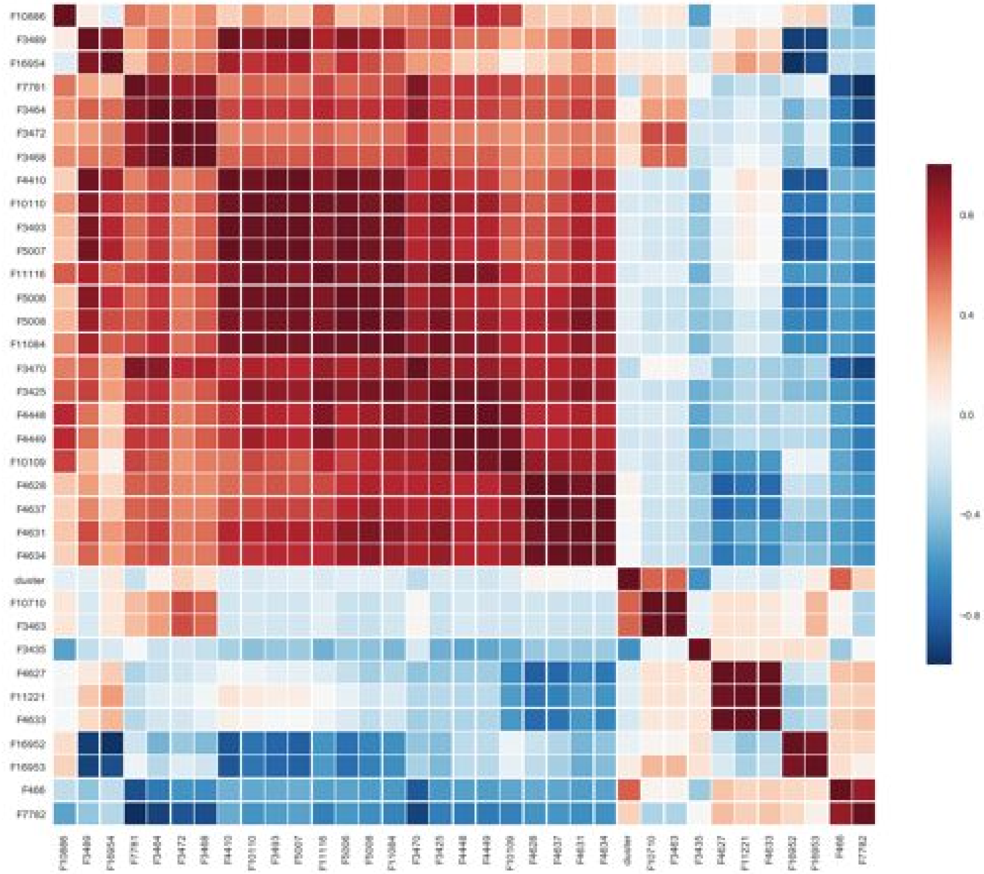
- Principal Component Analysis



1st pc: monthly spending(mortgage, payment)
2nd pc: percentage(percentage of loan, percentage of total value)

III Unsupervised Learning[Vertical--Property, Paquet 222]

- Cluster Analysis [k=4, Property vertical, Parquet 222]

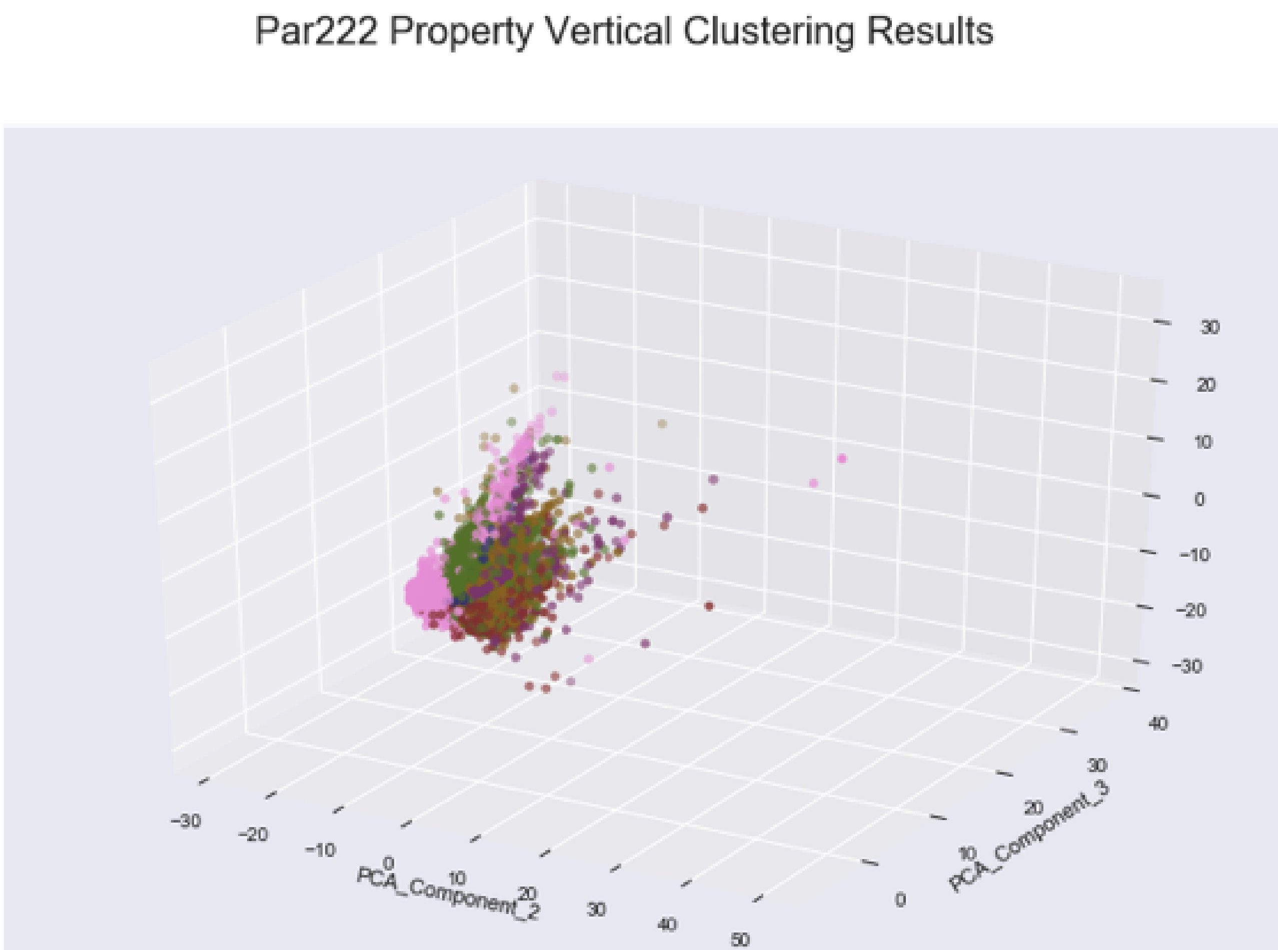


Positive: values(home value, mortgage amount)

Negative: owner vs. renter(percentage owner occupied vs. percentage renter occupied)

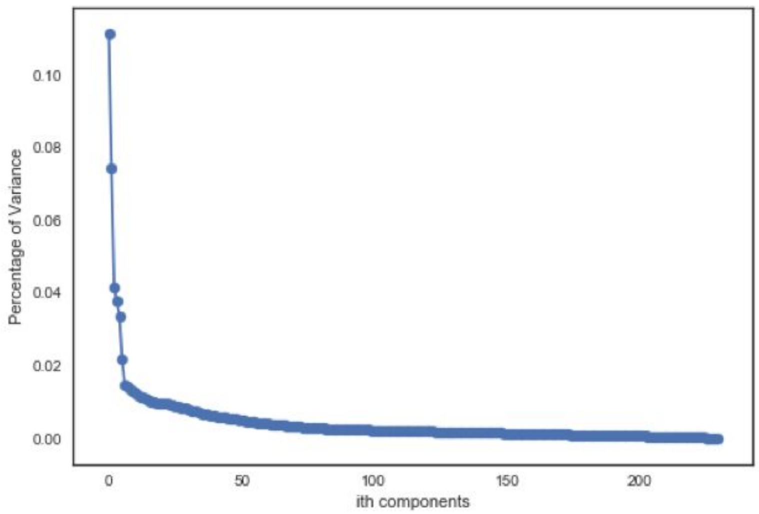
III Unsupervised Learning[Vertical--Property, Paquet 222]

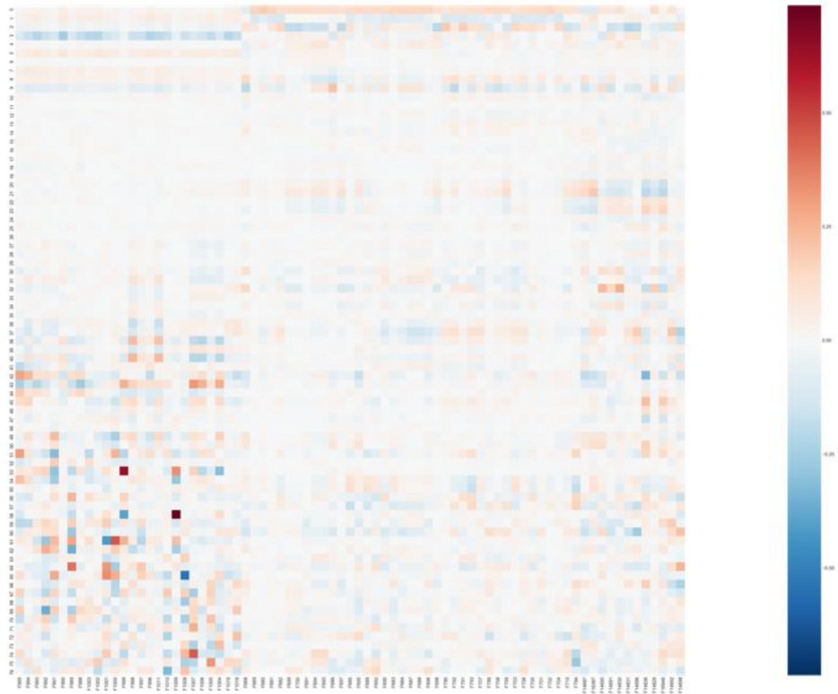
- Cluster Analysis [k=4, Property vertical, Parquet 222]



III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Principal Component Analysis [variance_ratio=.80,n_components=77]





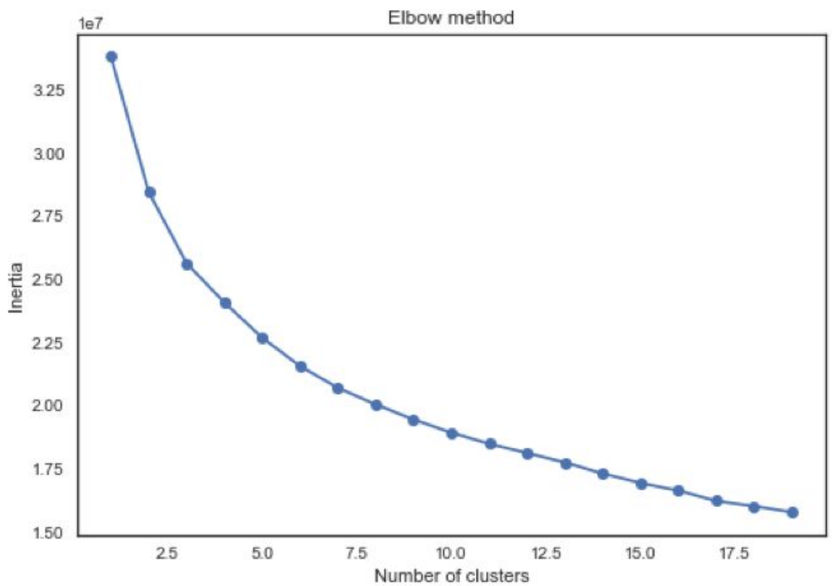
- **Used PCA to do dimension reduction, contained 80% of information and picked 77 principal components.**
- **The heatmap of 77 principal components:**

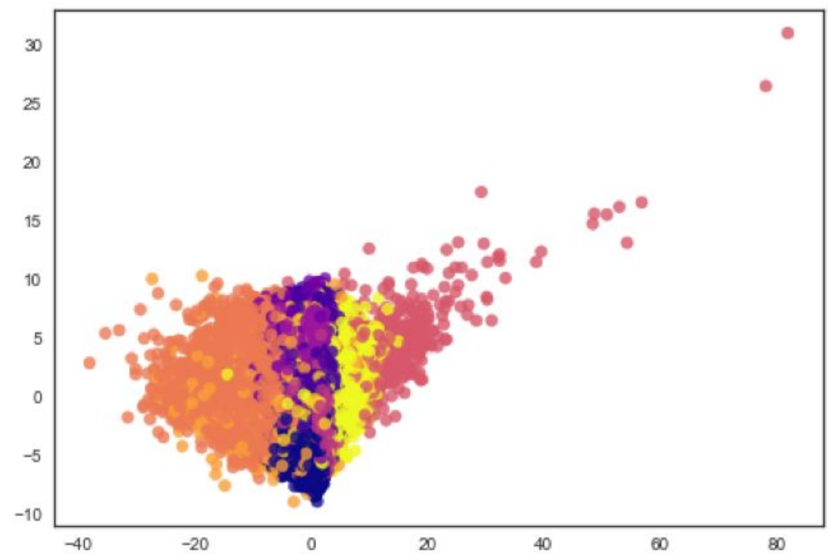
  y axis: principal components--from 0 to 77

  x axis: features of retail columns

III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Cluster Analysis [k:1~20, k=10]

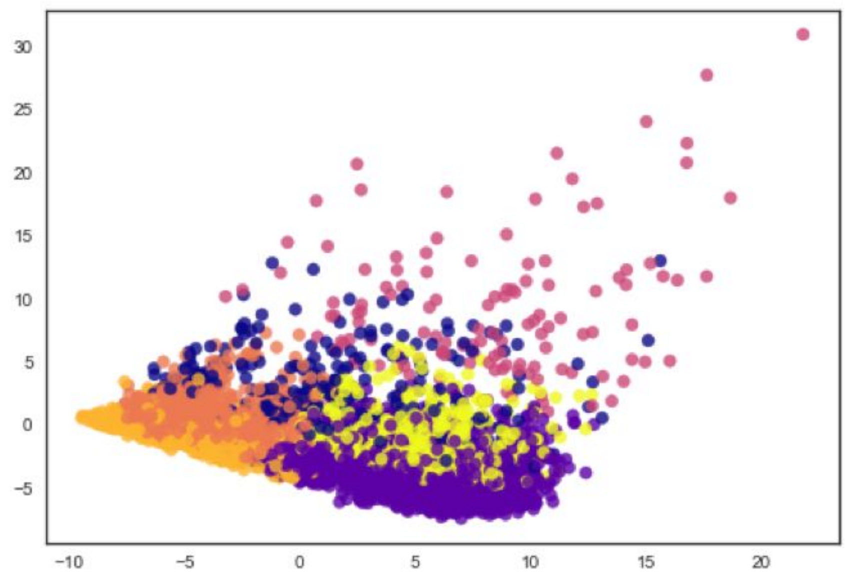


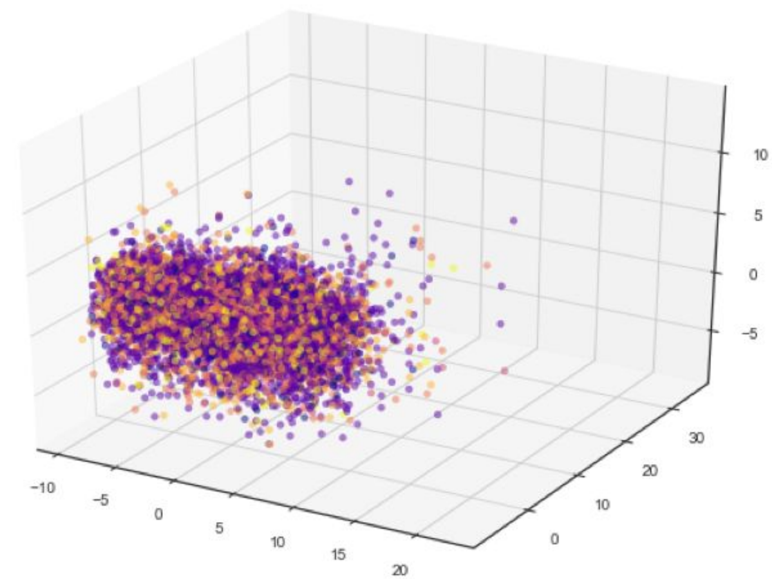


- **Cluster analysis based on the entire dataset and use principal component analysis results to visualize different group of clustering.**
- **Tested different choice of K, the right one is choosing 10 groups**

III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Cluster Analysis [k=7, n_components=77 (PCA)]

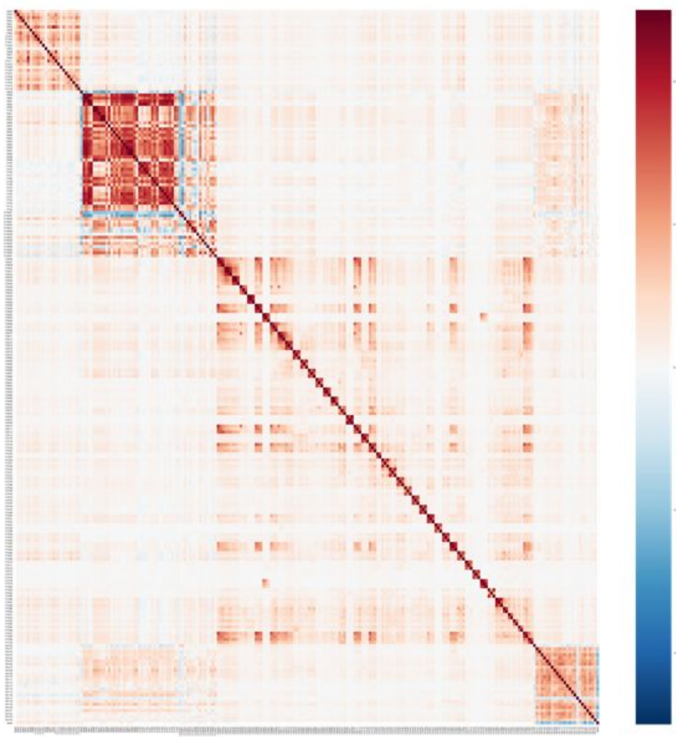


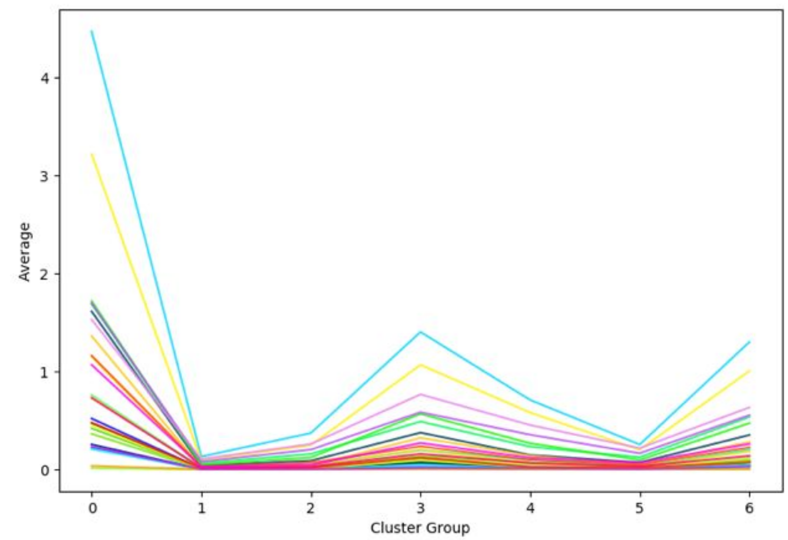


- **Choosed 7 groups, and the left one is clustering distribution using the first principal component and the second principal component.**
- **3D image of clustering distribution using the first, second and third principal component.**

III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Cluster Analysis [correlation map, k=7, mean, n_colnums=26]

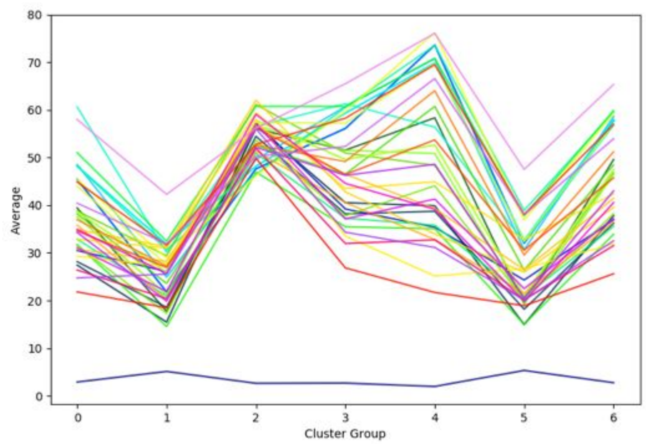


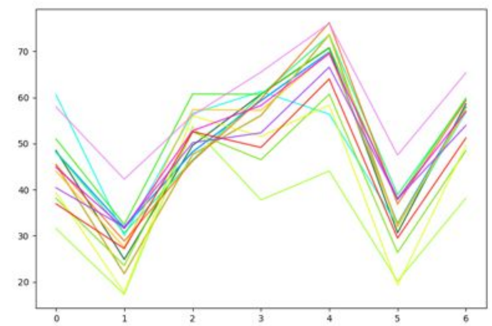


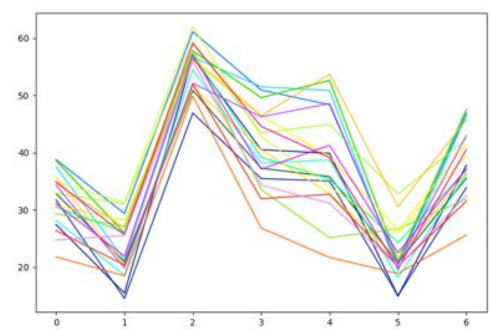
- **Correlation analysis of all the retail columns, there are 4 obvious group of columns from the correlation plot.**
- **Labeled all the observations from clustering analysis, group by clustering label and get the mean value of each features**
- **The first group of correlation columns: 26 columns; MOR (Mail Order Responders)--the number of times people bought the products.**
- **Group 0,3,6 people much more frequency buying behavior than group 1,2,4, and 5**

III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Cluster Analysis [k=7, mean, n_colnums=39(17,21)]



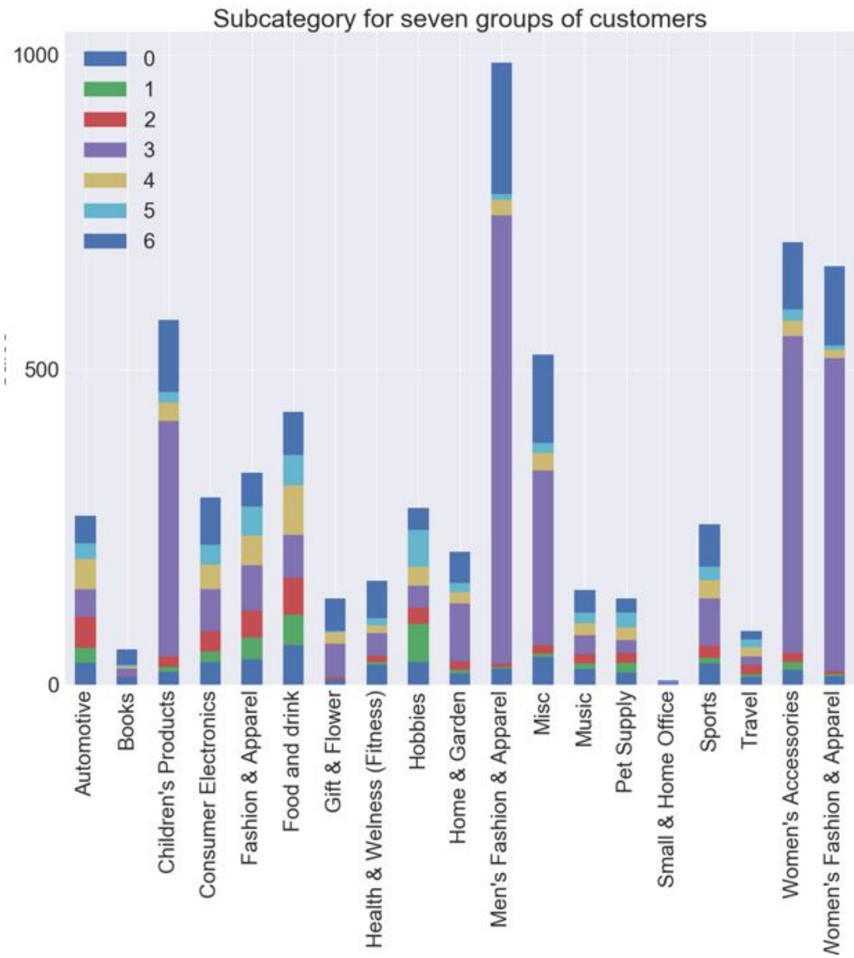




- **The second group of correlation columns; subcategory: Auto; 39 columns; Model prediction of the likelihood of buying different types of cars**
- **The number of clustering group -- the mean value of each columns**
- **Two different behavior trends: group 2/group 4 shows the most likelihood of buying cars**
- **Group 2,3,4,6 show more interest and likelihood of buying cars than other groups**

III Unsupervised Learning [Vertical--Retail, Paquet 20]

- Cluster Analysis Conclusion



- Group 3 customers have high buying power of women's and men's fashion and apparel, women's accessories and children's' products.
- Group 6 customers have high buying power of books, consumer electronics, health & wellness, music and sports.
- Group 5 customers have high buying power of hobbies and pet supply.
- Group 1 overall buying power smaller than other groups.
- Group 3 overall buying power larger than other groups.

IV Supervised Learning

- Chose subset of demographic columns
  - Columns selected based on data could be easily collected
  - Proof of concept: can add more columns easily for better results

- Make predictions to answer business-related questions
  - Predicted frequency households bought high-end women's retail
  - Classified median household income
  - Used Gradient Boosting Machine in Spark ML for regression
  - Used Random Forest Classifier for classification

GRADIENT BOOSTING MACHINE		
TREE-BASED METHOD THAT BUILDS TREES SEQUENTIALLY		
USES RESIDUAL INFO FROM PREVIOUS TREE TO FIT NEW TREE		
USED K-FOLD (K=5) CROSS-VALIDATION GRID SEARCH		
USED MEAN AVERAGE ERROR (MAE) EVALUATION METRIC		
ADVANTAGES	**DISADVANTAGES**	
Good predictive power	Needs CV to prevent overfitting	
Handles non-scaled data well	Computationally/time expensive	
RESULTS		
GBM	n_trees = 500 shrinkage = .001 Max depth = 15 Min obs in terminal node = 8	**MAE = 1.4486**

IV Supervised Learning

RANDOM FOREST CLASSIFIER		
TREE-BASED METHOD		
RANDOMLY SAMPLES BOTH OBSERVATIONS AND FEATURES		
USES ENSEMBLING PHILOSOPHY BY COMBINING WEAK LEARNERS AND AVERAGING RESULTS		
SEE CONFUSION MATRIX FOR RESULTS		

ADVANTAGES	DISADVANTAGES
Good predictive power	Need lots of data
Handles non-scaled data well	Computationally/time expensive
Over-fitting not an issue	Results can vary due to random nature

RESULTS		
RF Classifier	n_trees = 1000 Min obs in terminal node = 4 Feature Subset Strategy = ‘auto’	**See confusion matrix**

IV Supervised Learning

- Random Forest Classifier Confusion Matrix

PREDICTED MEDIAN INCOME CLASSES

ACTUAL MEDIAN INCOME CLASSES

F14593_index	0	2	3	5	6
0	58731	2041	18220	2	2
1	44804	1020	11864	2	0
2	37297	2215	17029	1	1
3	16471	1358	30783	2	1
4	30721	1098	10994	0	0
5	21252	1292	19986	6	3
6	21787	1230	17447	3	6
7	29587	637	9204	1	1
8	18607	377	4606	0	1
9	11893	166	3169	0	0
10	10645	223	3829	1	0
11	7421	132	2447	0	0
12	8115	50	1106	1	0

Key Takeaways:

- **Add more columns for better results**
- **Class 0, 2, 3 majority classes predicted**

V Conclusions

- Dimension reducing and clustering full parquet did not produce very interpretable results
- Clustering specific verticals produced more interpretable results
- Can find interesting patterns through EDA to gain business insight
- Subsetting data set to answer specific questions is a good strategy
- Findings from individual parquets similar and can be generalized to population

QUESTIONS???