Clustering Project

Asmaa Chraibi & Andrew Weber

Project Topics & Goals

Goal: Create two customer segmentations on demographic & financial customer information and visualize the clusters

Topics:

- Data Wrangling
- Data Visualization
- Data Preparation and Feature Engineering
- Dimensionality Reduction
- Unsupervised Learning

Data Engineering

Standard cleaning & EDA on datasets

Demographics: Customer information dataset

- Binning (e.g. Geography)
- Dummy Coding
- Min-max Scaler for specific columns

Financial: Combination of variables from transactions and account datasets

- Recorded transactions into categories (small, medium, high) based on distance from mean (sd)
- Removed categorical variables
- Standard Scalar
- Tried other combinations of variables and recoding. This was the final iteration

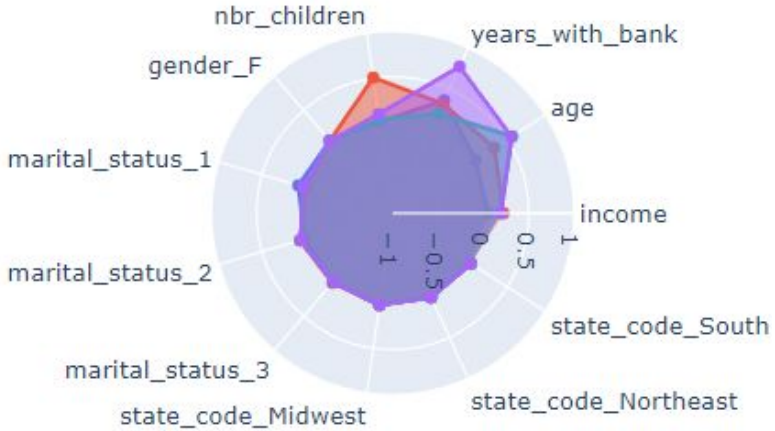
Demographic Segmentation

7 scenarios:

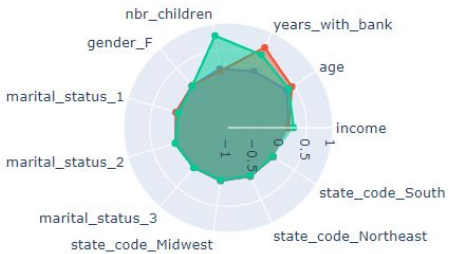
- Standard scaler with all features: segmentation based on marital status
- Standard scaler with all features: segmentation based on state bins
- Standard scaler with all features: big influence with gender feature
- Standard scaler with all features: big influence without dummies
- Standard scaler only 'income', 'age','years_with_bank','nbr_children' and multiplying the dummies with 0.5: income, nbre_children
- MinMax scaler only 'income', 'age','years_with_bank','nbr_children' and multiplying the dummies with 0.5: influence of marital status
- MinMax scaler only 'income', 'age','years_with_bank','nbr_children' and multiplying the dummies with 0.1 less influence

Demographic Segmentation

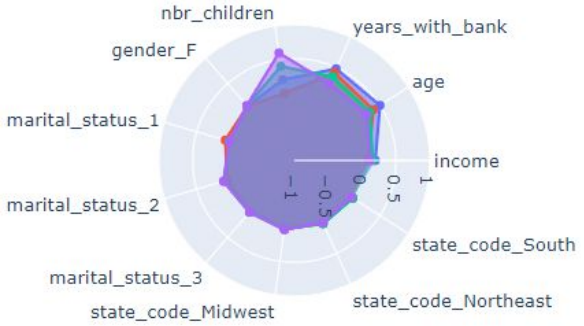
Kmeans (4)



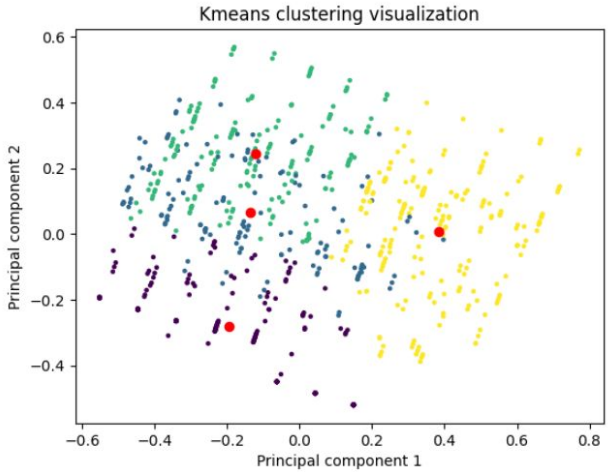
Hierarchical(3)

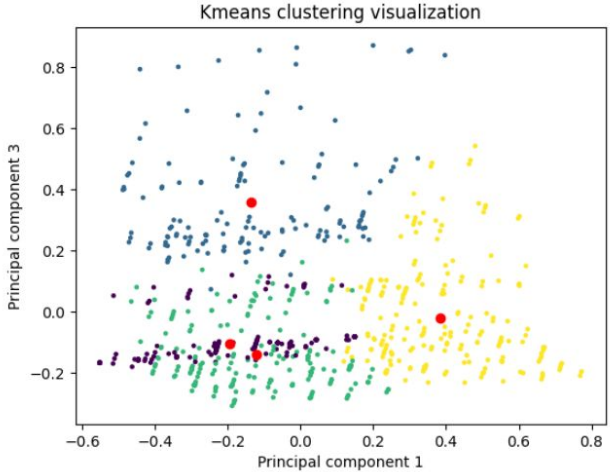


DBSCAN(4)-170



Demographic Segmentation





	income	age	years_with_bank	nbr_children	gender_F	marital_status_1	marital_status_2	marital_status_3	state_code_Midwest	state_code_Northeast	state_code_South
0	6684.148515	19.504950	3.287129	0.094059	0.564356	0.836634	0.123762	0.014851	0.202970	0.188119	0.252475
1	31925.317241	38.220690	2.924138	2.537931	0.565517	0.000000	0.655172	0.144828	0.213793	0.186207	0.248276
2	26749.351064	55.994681	1.877660	0.212766	0.558511	0.281915	0.579787	0.058511	0.202128	0.218085	0.239362
3	28159.334906	55.297170	6.971698	0.504717	0.551887	0.254717	0.584906	0.061321	0.212264	0.198113	0.240566

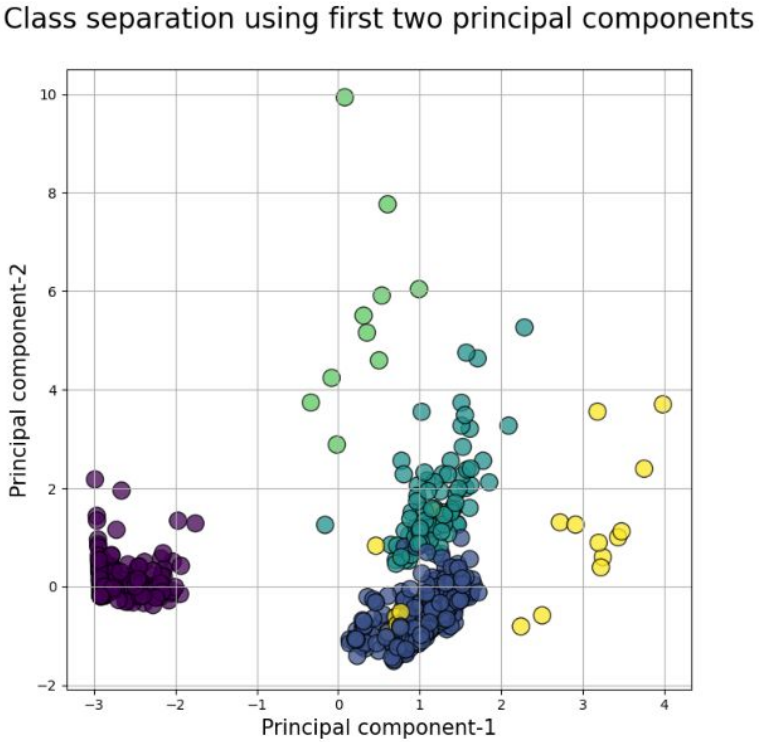
Financial Segmentation

Segmentations:

- **KMeans (selected)**
- Hierarchical
- DBScan

Selected **K = 5 clusters** based on Silhouette & Elbow

Segmentation plotted to the right on first two PCs indicates 3 main classes and then 2 ‘outlier’ / rarer classes



Financial Segmentation (continued)

- Segment 0: Few transactions of any type (n=189)
- Segment 1: Many small transactions (n=359)
- Segment 2: Many transactions of all types (n=85)
- Segment 3: Many transactions & high savings balance (n=10)*
- Segment 4: Many transactions & high credit balance (n=22)*

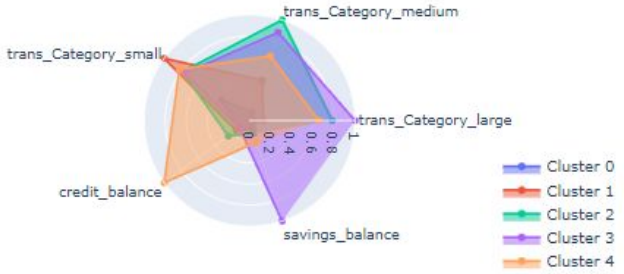
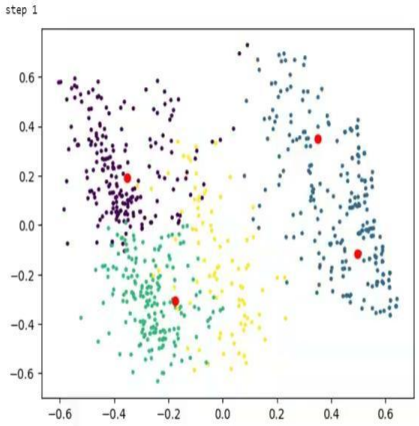


Fig: Polar chart, values rescaled to % of max means between clusters

Mean Values by Segment:

class	trans_Category_large	trans_Category_medium	trans_Category_small	credit_balance	savings_balance	years_with_bank	acct_type_CC	acct_type_CK	acct_type_SV
0	0.735849	1.358491	28.566038	715.837522	1016.999306	3.957672	0.597884	0.280423	0.761905
1	2.200557	6.637883	86.103064	939.505358	871.900046	3.699164	0.738162	1.000000	0.601671
2	13.411765	16.658824	67.752941	1476.993333	1437.573902	4.258824	0.705882	1.000000	0.482353
3	17.100000	14.600000	65.600000	703.050000	11172.385000	4.500000	0.800000	1.000000	1.000000
4	11.230769	10.692308	70.923077	5928.360455	2420.088000	3.818182	1.000000	0.590909	0.454545

KMeans Stepwise Visualization



Challenges

- Categorical / binary variables
- Running many models / checking different ways of clustering
- Data engineering challenges