Automobile Customer Segmentation: Unsupervised Machine Learning Investigation

Deirdre Boland

21 Feb 25

Dataset, Objective and Approach

Dataset

- Sales team at an automobile company has classified all customers into 4 segments (A, B, C, D). A strategy of performing segmented outreach and communication has worked exceptionally well for them.

Objective

- Can an unsupervised machine learning (ML) model predict the 4 segments?

or

- Is domain knowledge and further data required?

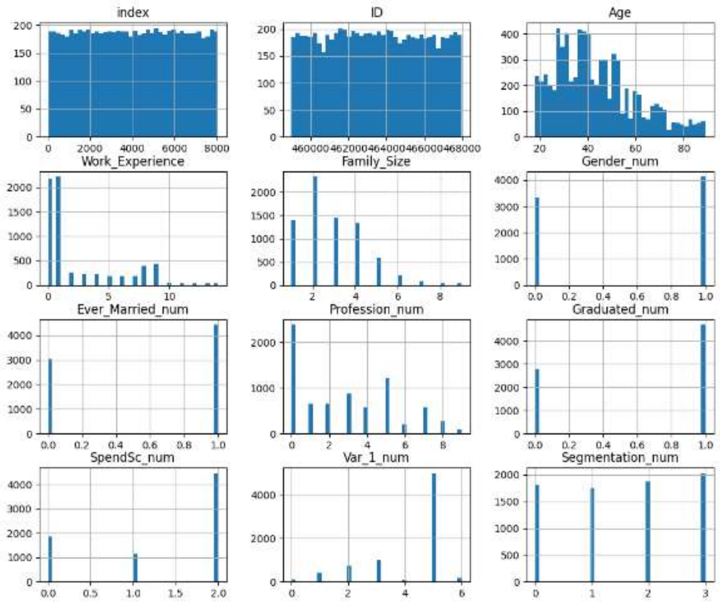
Approach

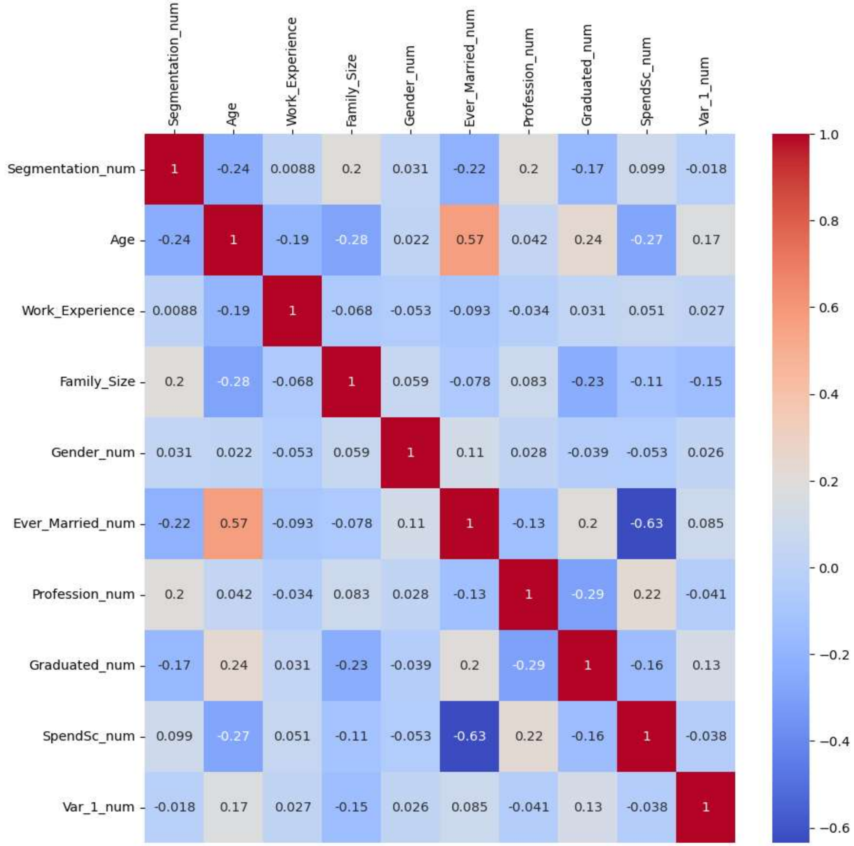
- Took training dataset and removed A-D segmentation to review against unsupervised ML results

- Reference – Kaggle: Customer Segmentation Vetrivel-PS

Data exploration and cleaning

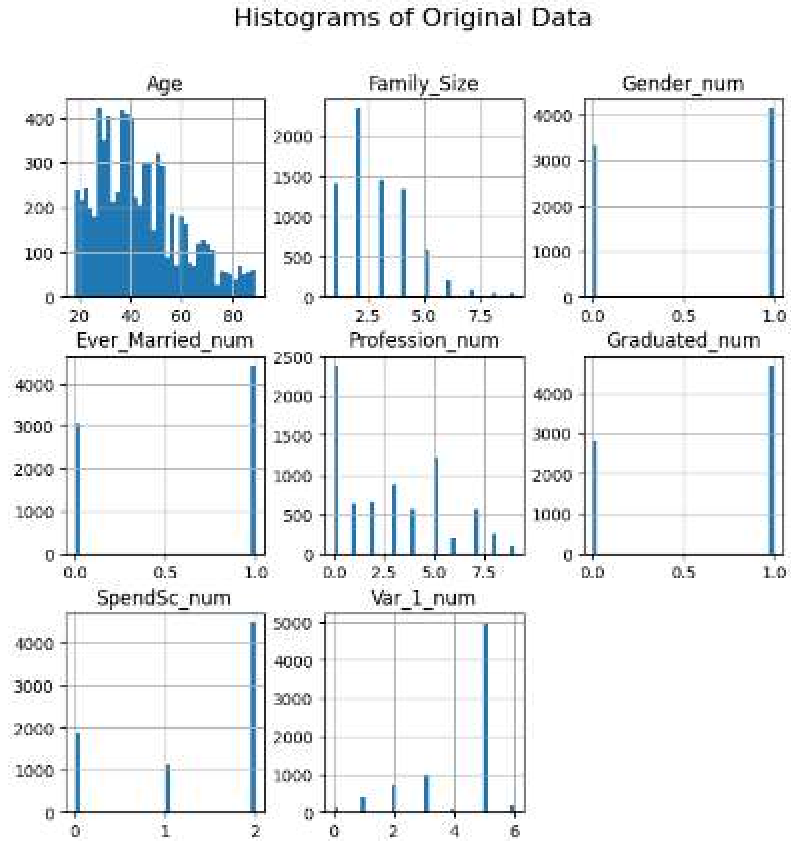
- Only ID, gender, age, spending score and segmentation have full data
- Removed null values from Ever Married (cat), Graduated (cat), Family Size (num), Var_1 (cat)
  - 7.3% data dropped – 7477 from 8068 (index = original dataset row index)
  - If not removed or replaced with string values will create new variable on encoding
- Label encoded categorical features
  - Gender – male = 1, female = 0
  - Ever married and Graduated – yes = 1, no = 0
  - Profession – Artist = 0, Doctor = 1, Engineer = 2, Entertainment = 3, Executive = 4, Healthcare = 5, Homemaker = 6, Lawyer = 7, Marketing = 8, *null= 9 **
    - **Potential loss further 1.2% if dropped null, categories seemed limited so allowing place for other/null*
  - Spend score – Average = 0, high = 1, low = 2
    - **Low has highest count not average – average may be based on external factor**
  - Var_1 – Cat 1 – 7, 0-6
  - Segmentation – A-D, 0-3
- Work experience further 10% loss if dropped nulls
  - **High count 0/1s despite wide adult age range**
  - Low correlation
  - dropped column from dataset





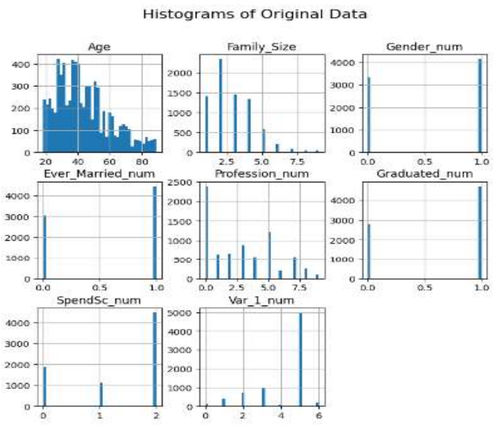
Modelling approach

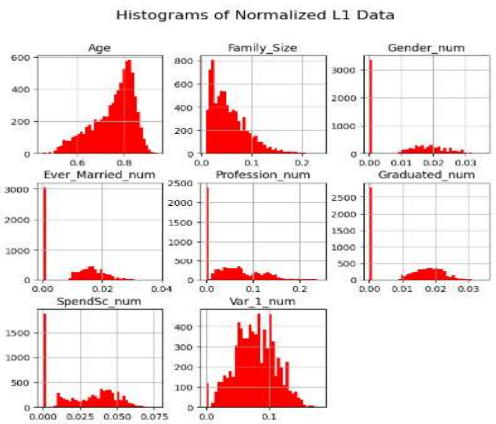
- 8 numeric features after dropping segmentation (target)
- Data preparation
  - Regularisation
    - Scaled by Z scale, Robust scale and Min max scale
    - Normalised by L1, L2 or Max normalisation
  - PCA of original, scaled and normalised
  - Unsupervised modelling
    - Forcing to 4 categories (k = 4) to match target set even if not optimal
    - K means (Km) clustering
      - Default settings, random state set to limit variation from setting initial centroids
      - K assessment using elbow method and silhouette score
    - Agglomerative (Aggl)/Hierarchical clustering
      - metric= "Euclidean", linkage = "ward"
      - Deterministic, random state not required
      - K assessment using silhouette score

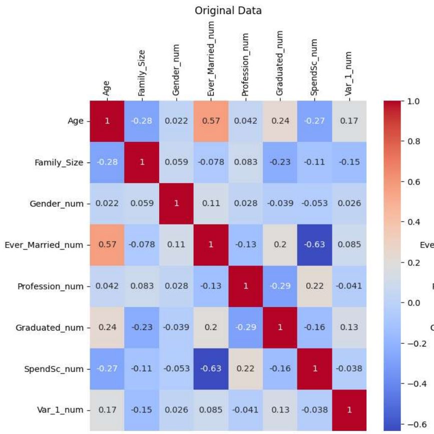


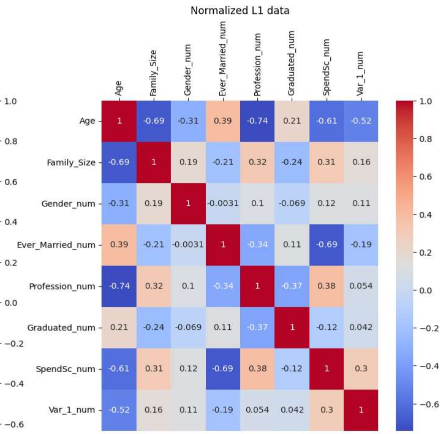
Effects of normalisation

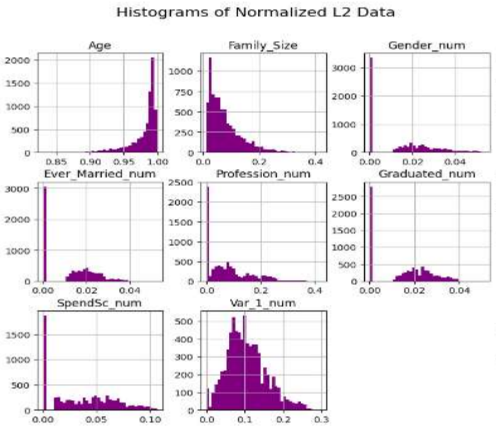
- Normalizer works on each sample (row) independently
  - calculates the norm of each sample and then scales the individual features by dividing them by the computed norm.
  - A regularization method, e.g. a method to keep the coefficients of the model small, and in turn, the model less complex.
- L1 Norm: aka Manhattan norm – sum of absolute values
- L2 Norm: aka Euclidean norm – square root of the sum of squared values – most common
- Max Norm: Scales each feature by the maximum absolute value in the sample – used neural network weights
- References – https://www.pythonprog.com/sklearn-preprocessing-normalizer/#:~:text=The%20Normalizer%20in%20Scikit-Learn%20focuses%20on%20transforming%20individual,values%20so%20they%20fall%20within%20a%20certain%20range. https://machinelearningmastery.com/vector-norms-machine-learning/

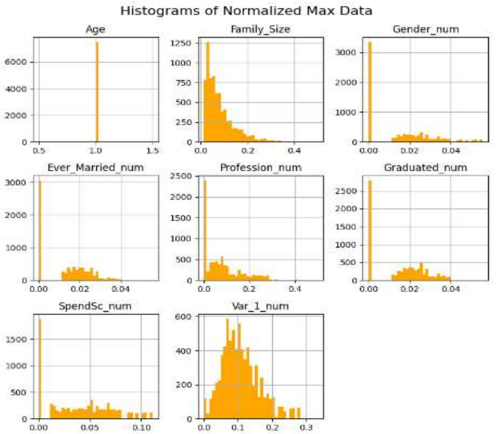


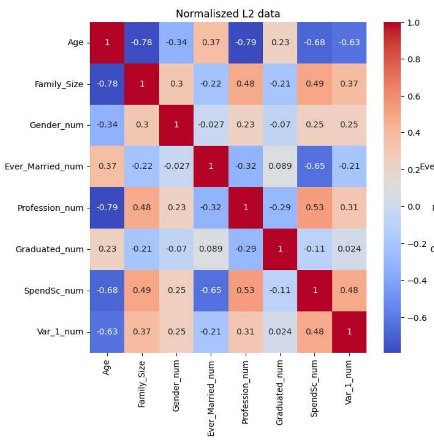


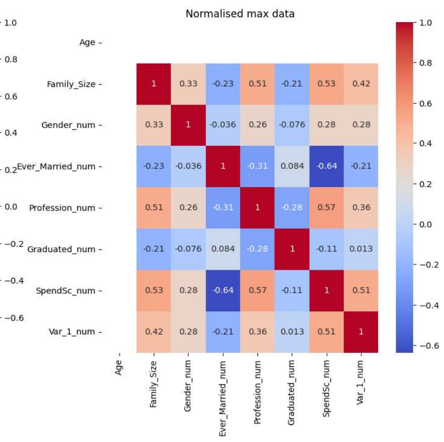






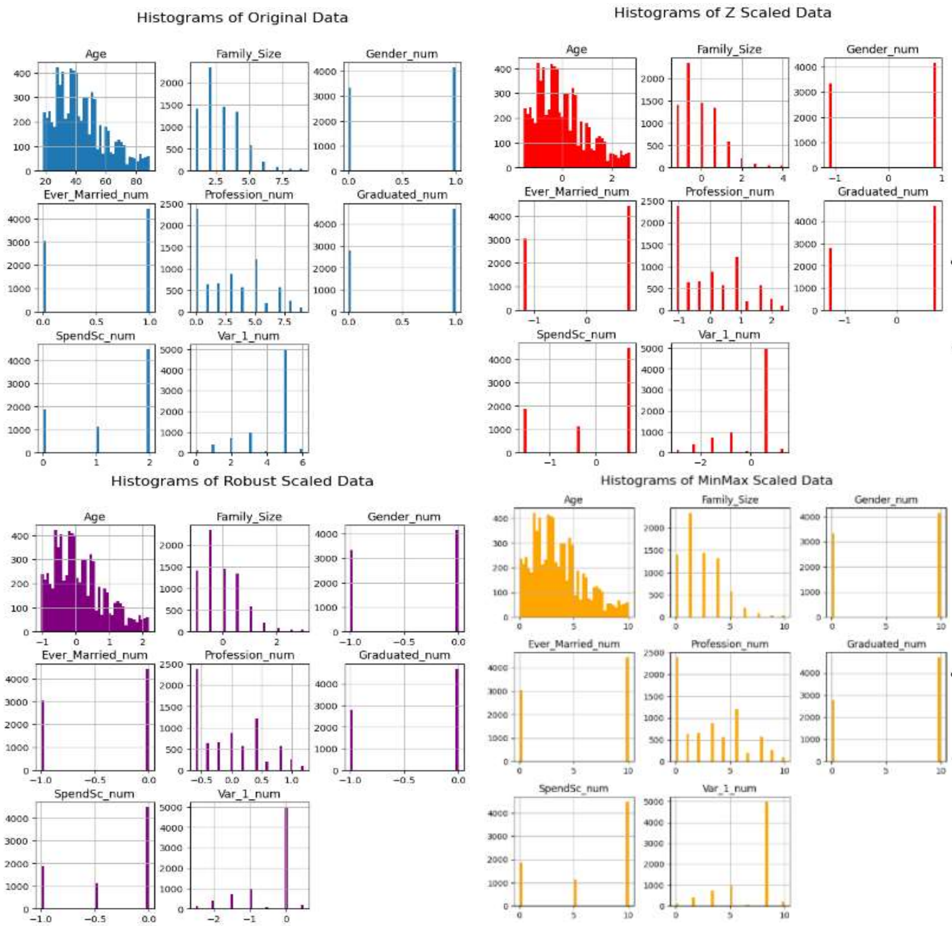


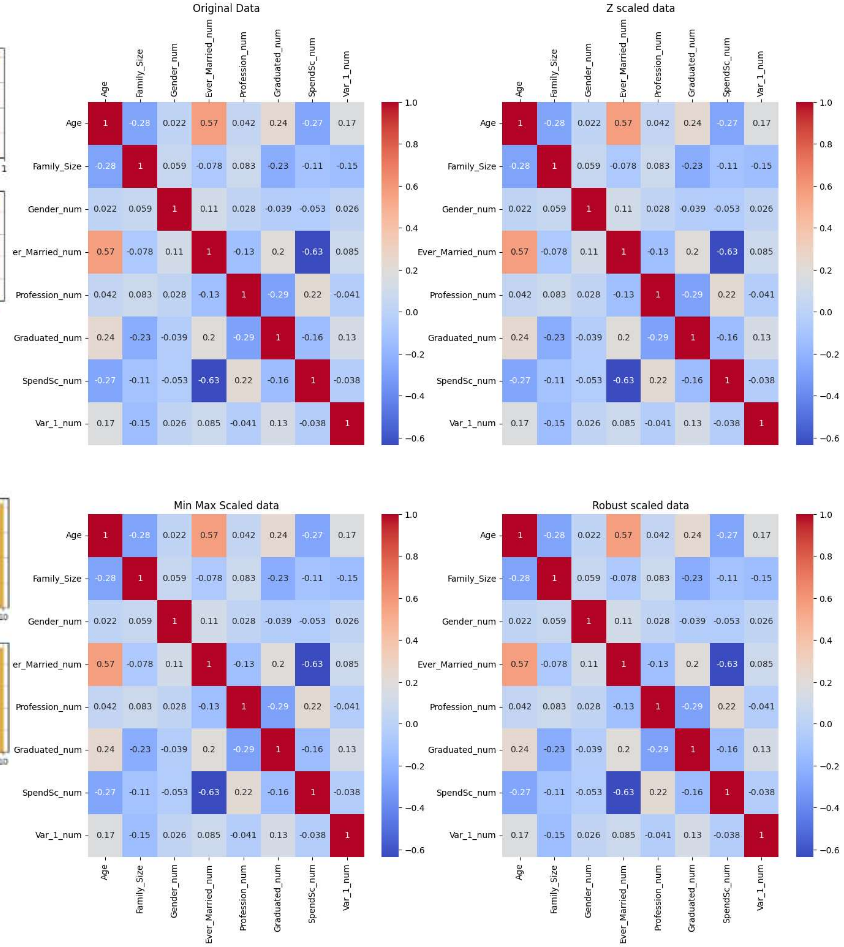




Effects of scaling

- X axis changes but no major change to distribution or correlation



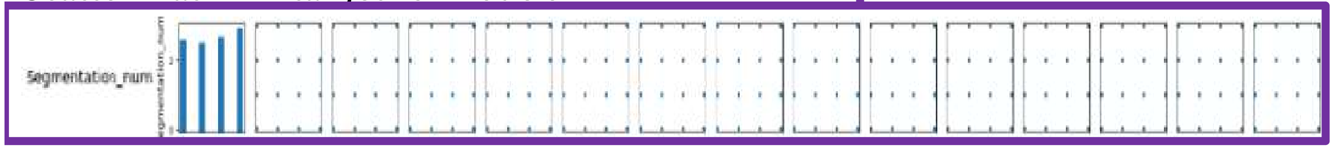


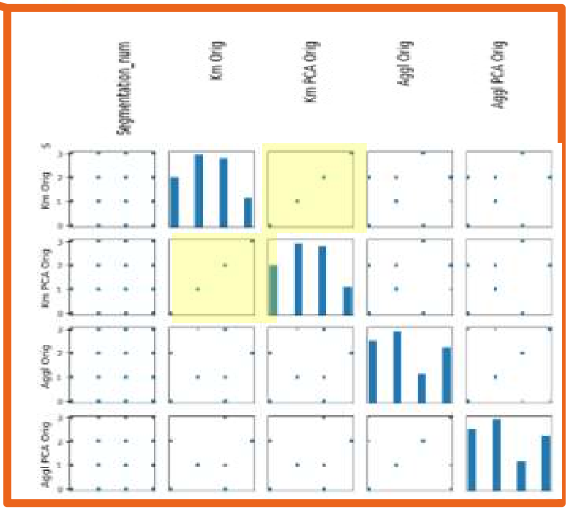
Results – 8 features

Correlation matrix – target vs. models

Original data Variance Ratio: [0.958 0.024] 2 component PCA covers 98.17 %
L2 normalised data Variance Ratio: [0.62 0.18 0.128] 3 component PCA covers 93.35 %
Min max Scaled data Varaiance Ratio: [0.35 0.226 0.182] 3 component PCA covers 75.82 %
Z Scaled data Varaiance Ratio: [0.28 0.168 0.145] 3 component PCA covers 59.47 %
L1 normalised data Varaiance Ratio: [0.778 0.112 0.069] 3 component PCA covers 95.86 %
Max normalised data Varaiance Ratio: [0.635 0.179 0.125] 3 component PCA covers 93.89 %

Scatter matrix – target vs. models



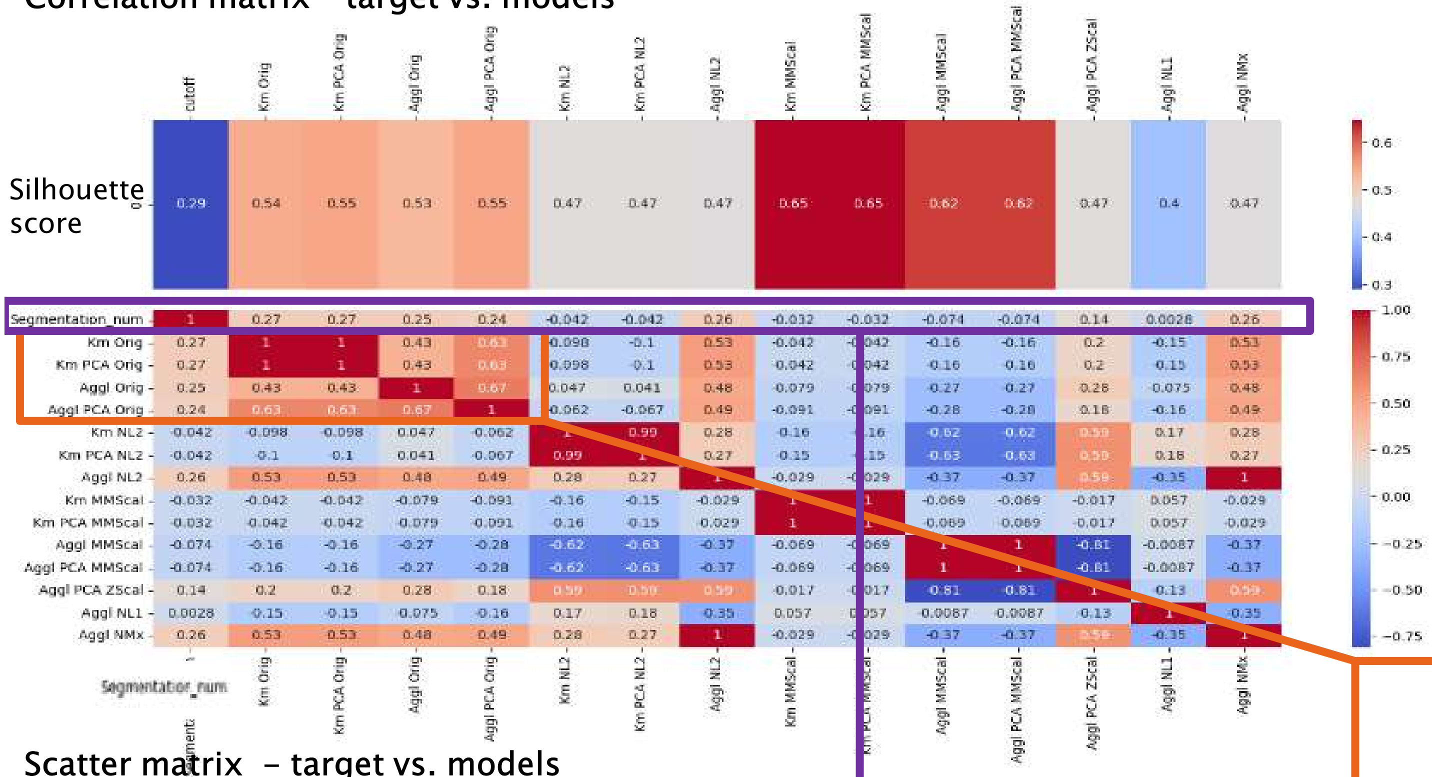


- minimal alignment
-Desired scatter matrix output is 4/minimum points only on scatter as per Km Orig vs. Km PCA orig (on right)
- Actual outcome = a 0 in segmentation has a 0,1,2 or 3 in model and same for all categories

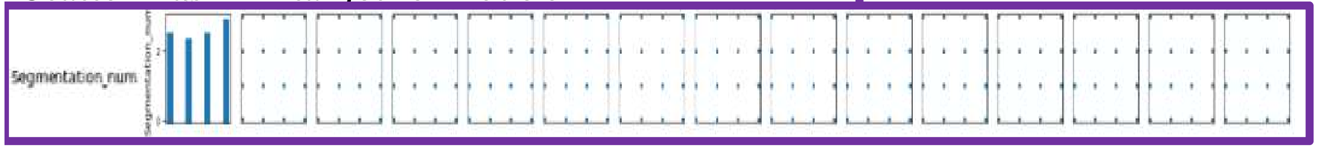
Results – 3 features

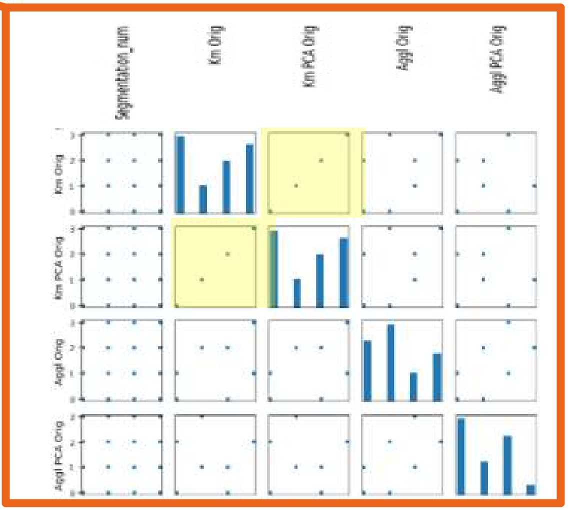
- Only ID, gender, age, spending score and segmentation have full data – questions around spending score and profession
- Copied notebook, remodelled data for just gender, age and spending score and re-ran

Correlation matrix – target vs. models



Scatter matrix – target vs. models





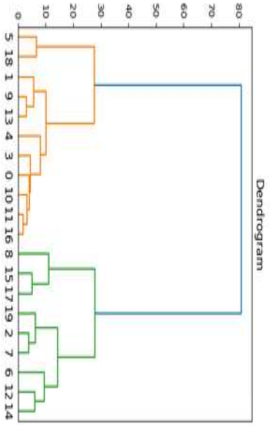
- minimal alignment
-Desired scatter matrix output is 4/minimum points only on scatter as per Km Orig vs. Km PCA orig (on right)
- Actual outcome = a 0 in segmentation has a 0,1,2 or 3 in model and same for all categories

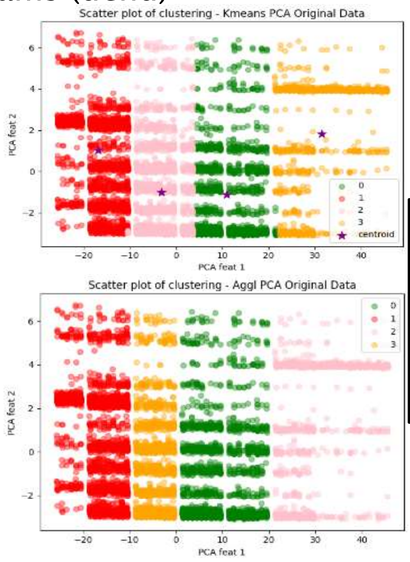
Cluster visualisation

- Kmeans (Km) and Agglomerative (Aggl) scatter plots of data use whole dataset
- 20 row subset (fixed sampling using numpy seed) used to plot dendrograms (dend)
  - rough idea of categorisation compared to full data

8 features – Age tracks & Aggl and Kmeans close to dend (2 and 3 swapped)

Age	Family_Size	Gender_num	Ever_Married_num	Profession_num	Graduated_num	SpendSc_num	Var_1_num	Segmentation	Segmentation_num	Km Orig	Aggl Orig	Index
28	3	0	0	5	0	2	3	C	2	1	1	5
22	4	1	0	5	0	2	5	D	3	1	1	18
43	1	1	0	0	1	2	2	B	1	2	3	1
43	5	1	0	0	1	2	5	C	2	2	3	9
42	4	0	1	0	1	0	5	A	0	2	3	13
39	7	1	1	4	0	1	3	D	3	2	3	4
38	4	1	1	1	0	1	5	D	3	2	3	3
39	2	0	0	3	1	2	3	A	0	2	3	0
36	1	1	0	1	1	2	5	B	1	2	3	10
39	1	1	0	1	1	2	5	A	0	2	3	11
38	1	0	0	2	1	2	5	D	3	2	3	16
48	1	0	0	6	1	2	2	B	1	0	0	8
58	4	1	1	4	0	1	2	B	1	0	0	15
55	4	1	1	6	1	0	5	C	2	0	0	17
67	2	1	1	0	1	1	5	B	1	3	2	19
62	2	0	1	0	1	2	5	C	2	0	0	2
62	5	1	1	0	1	0	5	A	0	0	0	7
73	2	0	1	0	1	0	5	C	2	3	2	6
72	2	0	1	7	1	1	5	C	2	3	2	12
66	2	0	1	7	1	1	5	A	0	3	2	14



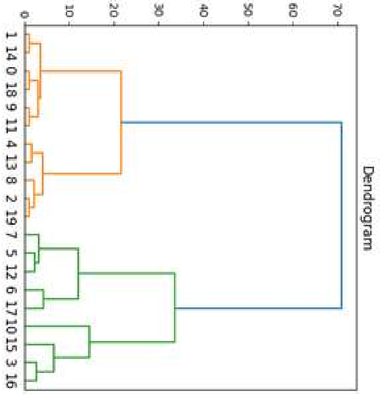


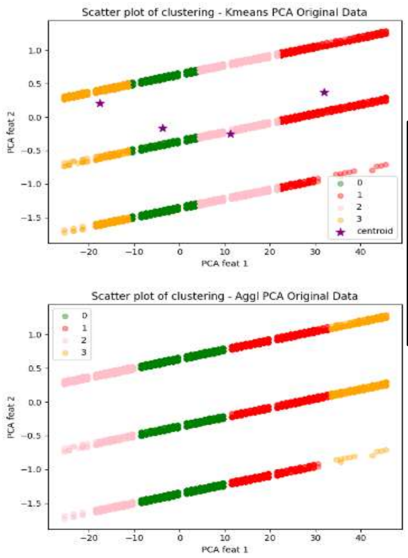
Original data Variance Ratio: [0.95767885 0.02400603] Variance by 2 component PCA covers 98.17 %

---

3 features – Age tracks & Kmeans matches and Aggl close to dend

Age	Gender_num	SpendSc_num	Segmentation	Segmentation_num	Km Orig	Aggl Orig	Index
29	0	0	A	0	3	3	1
30	0	0	A	0	3	3	14
28	1	2	D	3	3	3	0
28	0	2	C	2	3	3	18
30	1	2	C	2	3	1	9
30	0	2	D	3	3	1	11
19	0	2	D	3	3	3	4
18	1	2	D	3	3	3	13
22	0	2	D	3	3	3	8
20	1	2	D	3	3	3	2
21	1	2	D	3	3	3	19
39	1	0	A	0	0	1	7
36	1	0	C	2	0	1	5
37	1	2	D	3	0	1	12
47	1	0	B	1	0	0	6
43	0	0	B	1	0	1	17
65	1	2	A	0	2	2	10
57	1	0	C	2	2	0	15
51	1	0	C	2	2	0	3
52	0	2	A	0	2	0	16





Original data Variance Ratio: [0.99673525 0.00238464] Variance by 2 component PCA covers 99.91 %

Conclusion

- Unsupervised learning was not able to reproduce existing A–D customer classification
  - More domain knowledge required – oddities in work experience and spending score noted in data exploration
- Future work/improvements
  - Could have split dataset into training and test data with known answer
  - Iterative approach taken – function or pipeline would reduce coding lines for review