UNCOVERING RETAIL CUSTOMER SEGMENTATION FROM LARGE TRANSACTION RECORDS

A NUANCED COMPARISON OF CLUSTERING ALGORITHMS USING ROUGH SET REDUCED DATASET

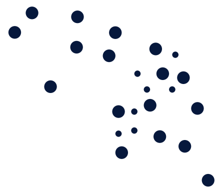
SYED AHMAD ZAKI BIN SYED SAKAF AL-ATTAS

SINGAPORE MANAGEMENT UNIVERSITY

×

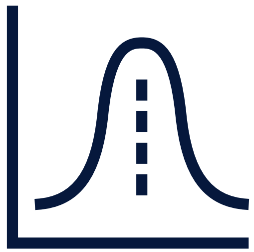
PROBLEM STATEMENT

Traditional clustering methods does not take into account inherent inconsistencies in data,
hence the need to explore and compare alternative clustering algorithms



TRADITIONAL
***K*-MEANS**
Clustering Comparison

FUZZY
***K*-MEANS**
Clustering Comparison



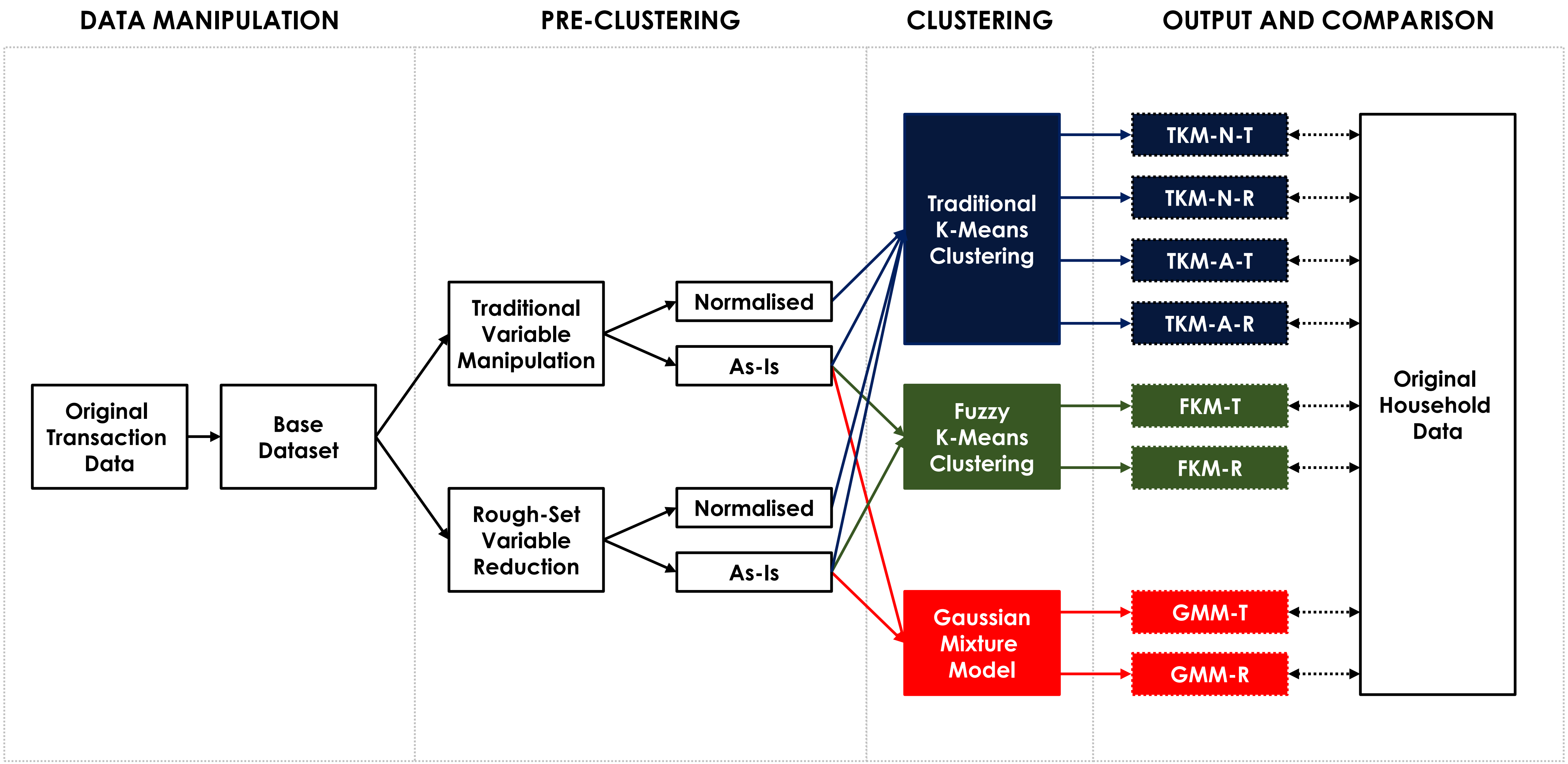
GAUSSIAN
MIXTURE MODEL
Clustering Comparison

ROUGH
SET
Feature Reduction

OPEN-SOURCE INTEGRATION (R)

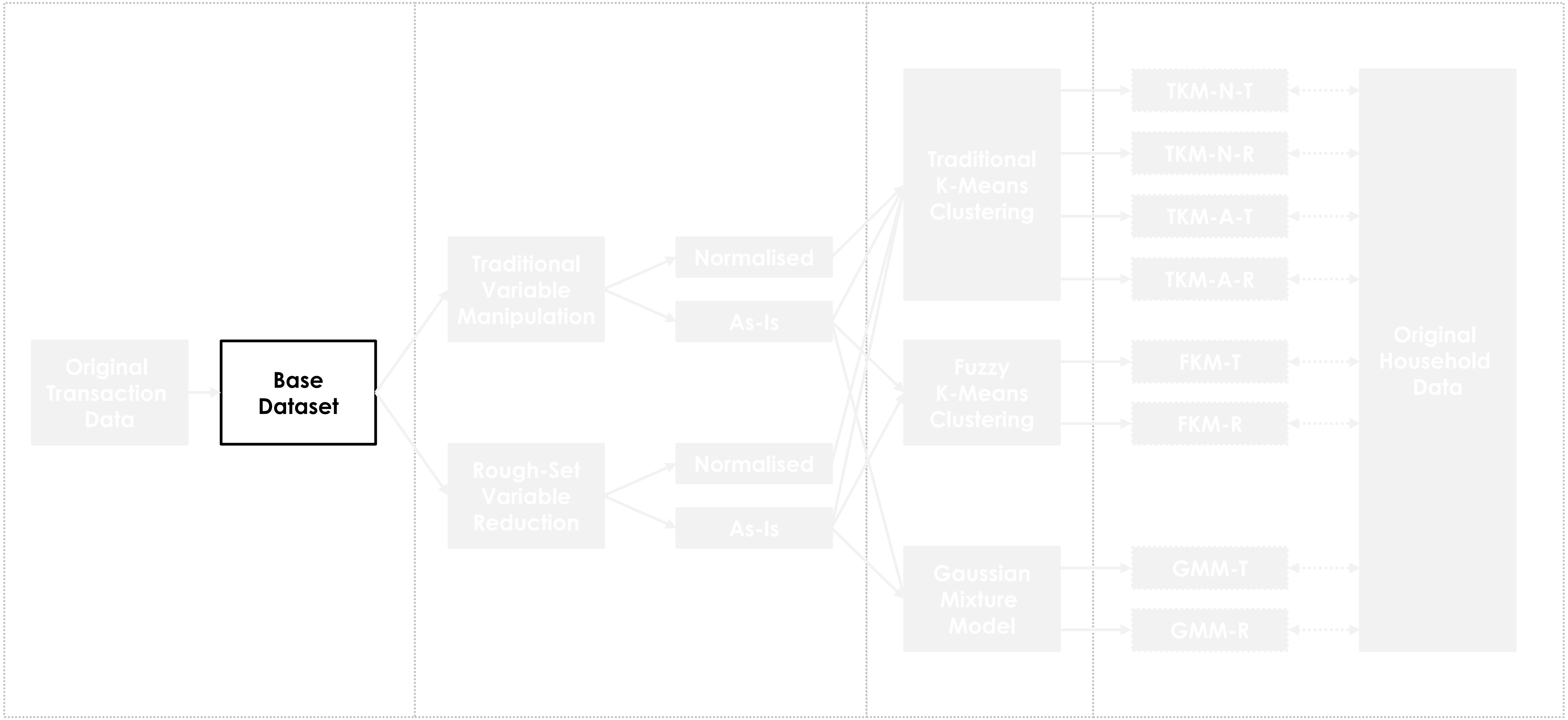
SAS ENTERPRISE MINER 14.1

CAPSTONE WORKFLOW



CAPSTONE WORKFLOW

DATA MANIPULATION | PRE-CLUSTERING | CLUSTERING | OUTPUT AND COMPARISON



BASE DATASET (RFM MODEL INSPIRED)

DUNNHUMBY'S RETAIL SHOPPING DATASET (TXNS)

RECENCY
(How long ago were their purchase?)

- Days Since First Order
- Days Since Last Order

FREQUENCY
(How often were the purchases made?)

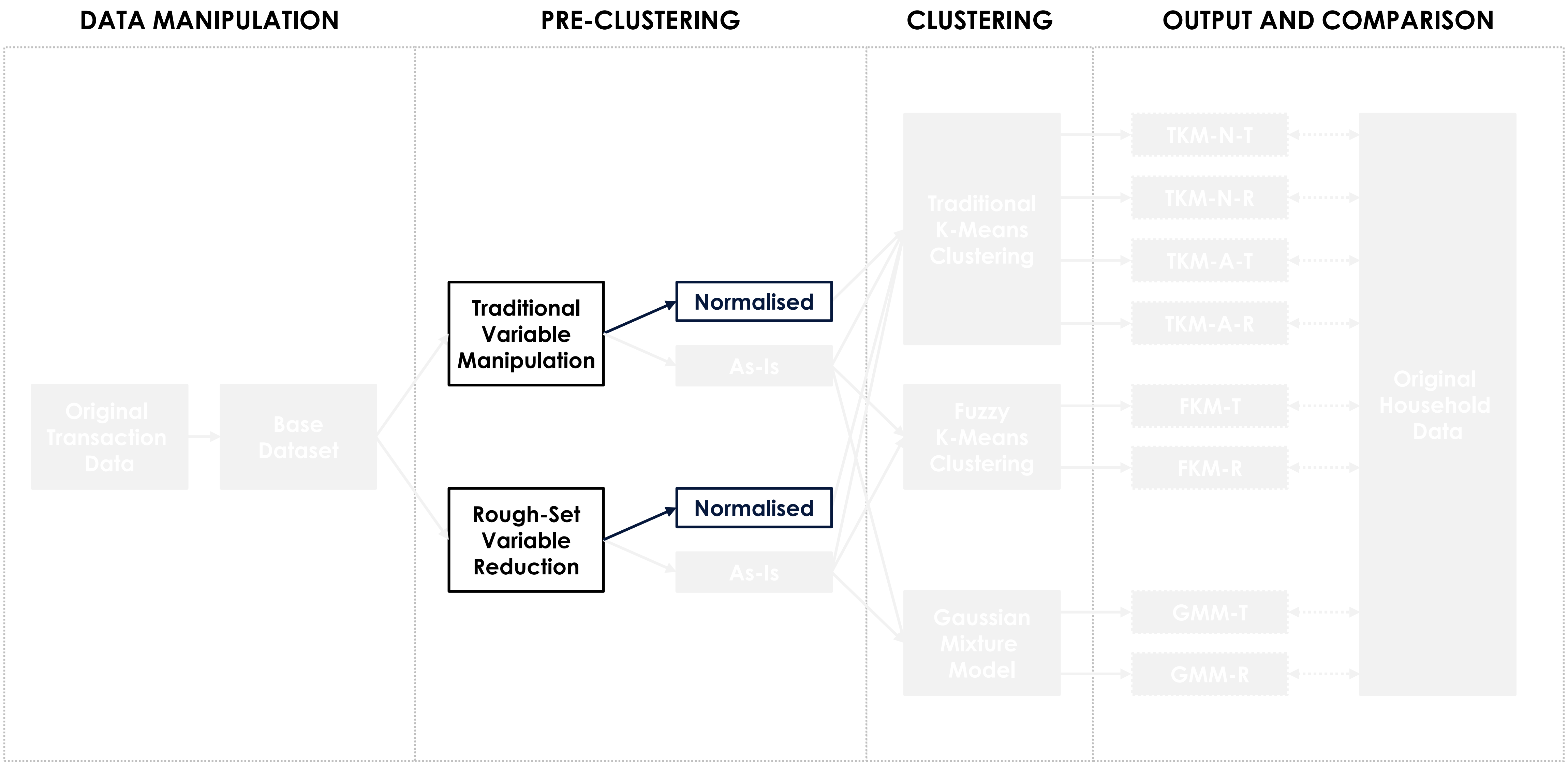
- Active Weekday Count
- Active Weekend Count
- Active Morning ($6 \le x < 12$) Count
- Active Afternoon ($12 \le x < 18$) Count
- Active Evening ($18 \le x < 0$) Count
- Active Late-Night ($0 \le x < 6$) Count
- Average Basket Count Per Active Weekday
- Average Basket Count Per Active Weekend
- Average Basket Count Per Active Morning
- Average Basket Count Per Active Afternoon
- Average Basket Count Per Active Evening
- Average Basket Count Per Active Late-Night

MONETARY
(How much money was spent?)

- Total Spend
- Min Spend Per Basket
- Average Spend Per Basket
- Max Spend Per Basket
- Average Sales Value Per Qty
- Min Spend Per Active Week
- Average Spend Per Active Week
- Max Spend Per Active Week
- Average Spend Per Active Morning
- Average Spend Per Active Afternoon
- Average Spend Per Active Evening
- Average Spend Per Active Late Night
- Discount Used Per Active Weekday
- Discount Used Per Active Weekend

**Active refers to time or periods when the transaction occurred*

CAPSTONE WORKFLOW



PRE-CLUSTERING

Traditional Variable Manipulation

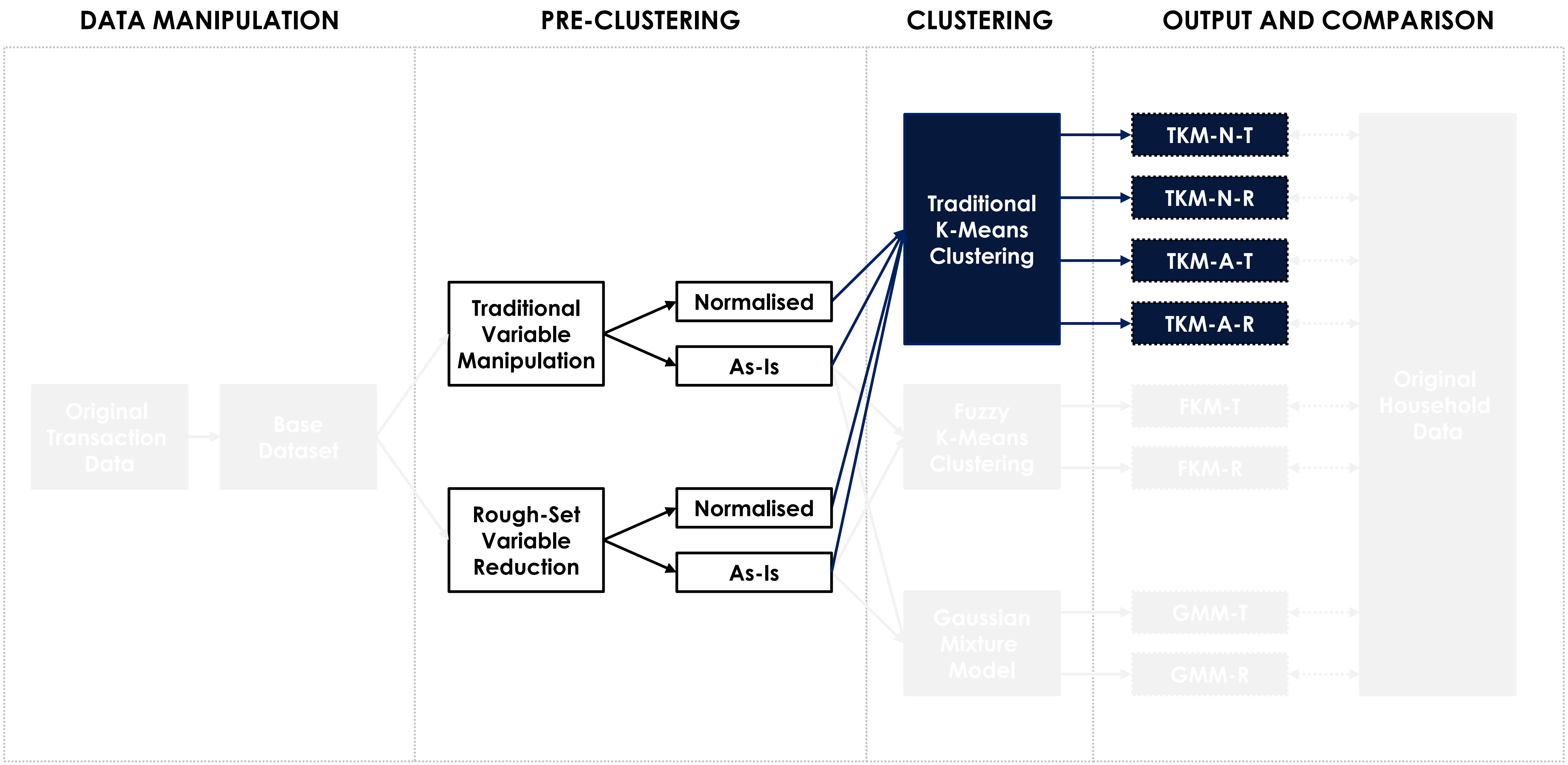
28 Base Dataset Variables → 16 Non-Collinear Variables → 16 <u>Normalised</u> Non-Collinear Variables → 16 <u>Normalised & Scaled</u> Non-Collinear Variables

Rough-Set Variable Reduction

28 Base Dataset Variables → 2 & 3 Reduct Variables → 2 & 3 <u>Normalised</u> Reduct Variables → 2 & 3 <u>Normalised & Scaled</u> Reduct Variables

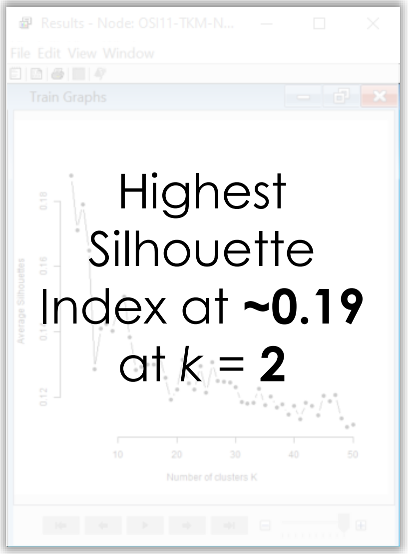
**Identifying non-collinearity and normalisation was done using standard SAS Enterprise Miner 14.1 nodes. Rest of pre-clustering used R code.*

CAPSTONE WORKFLOW



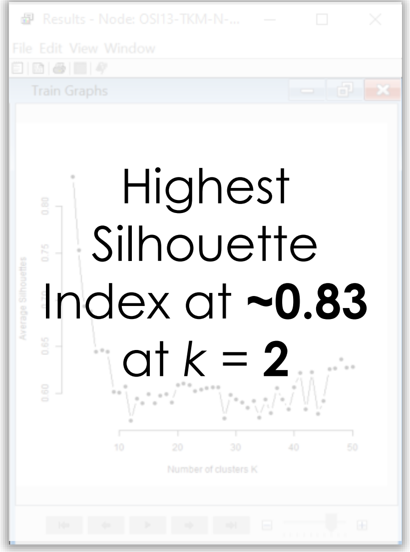
ANALYSIS & RESULTS (TRAD. K-MEANS)

TKM-N-T



Average Silhouette Index (k = 2 to 50)

TKM-N-R



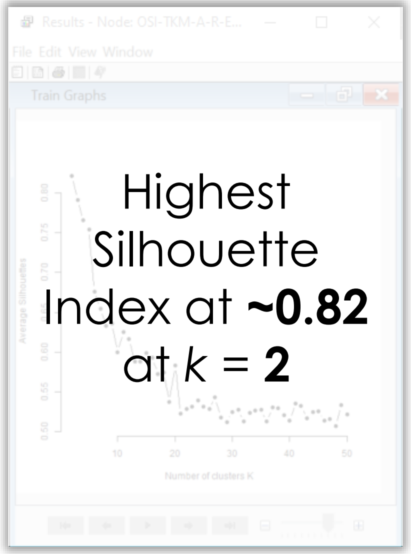
TKM-A-T

Average Silhouette Index (k = 2 to 50)

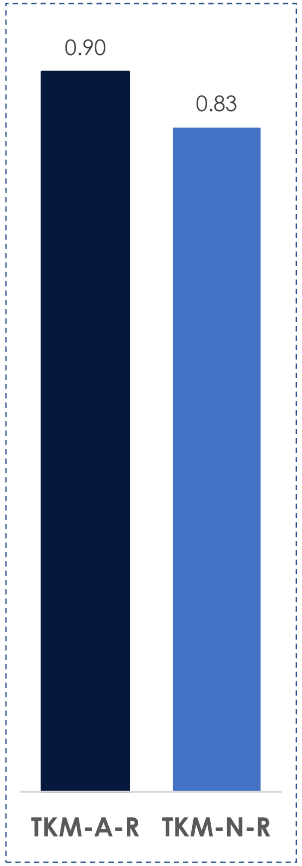
TKM-A-R

TKM-N-R3

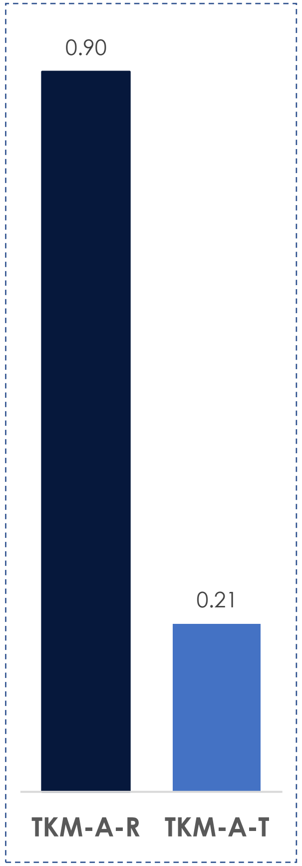
TKM-A-R3

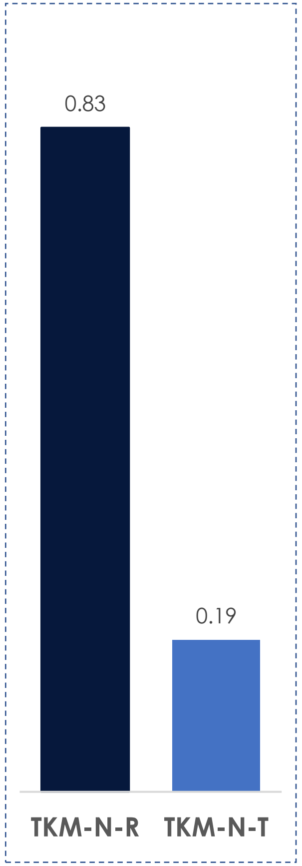


SILHOUETTE COMPARISON (TRAD. K-MEANS)

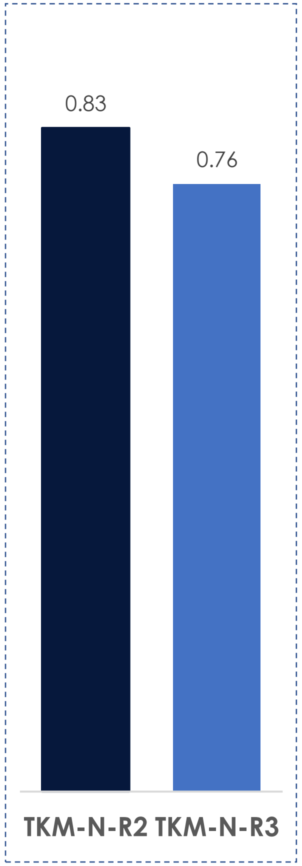


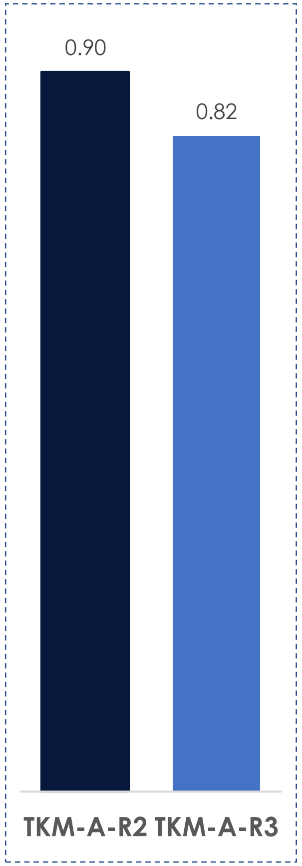
AS-IS DATA HIGHER THAN NORMALISED DATA





REDUCT DATA HIGHER THAN NON-REDUCT DATA





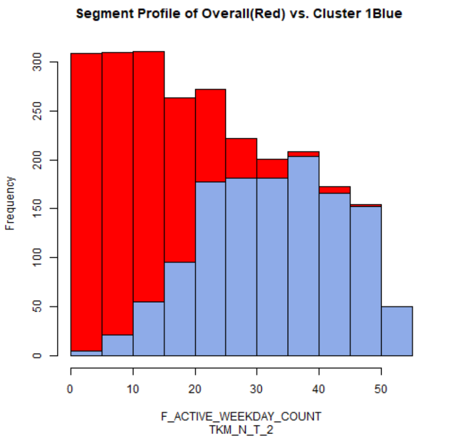
REDUCT 2 DATA HIGHER THAN REDUCT 3 DATA

CLUSTER COMPARISON (TRAD. K-MEANS)

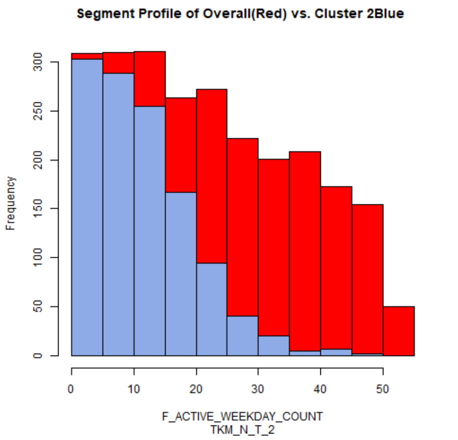
Chosen One of 28 Transactional Attributes (F_ACTIVE_WEEKDAY_COUNT) For Apple-to-Apple Comparison

TKM-N-T

Cluster 1 (1,287 IDs)

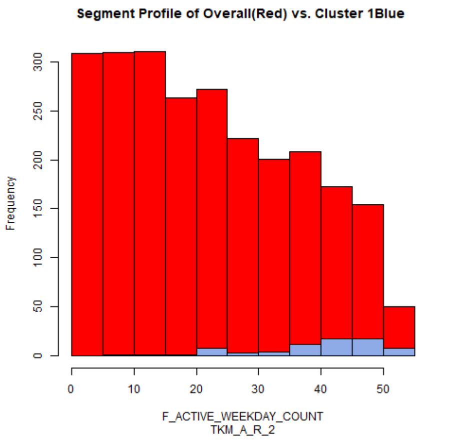


Cluster 2 (1,183 IDs)

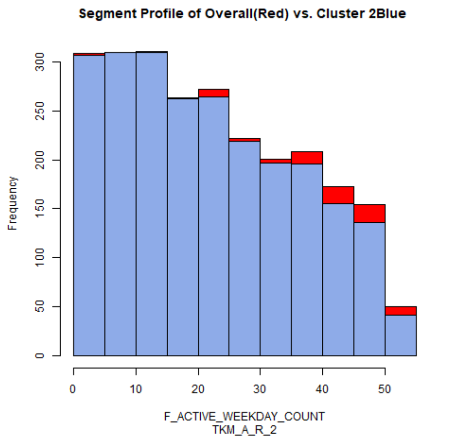


TKM-A-R

Cluster 1 (74 IDs)

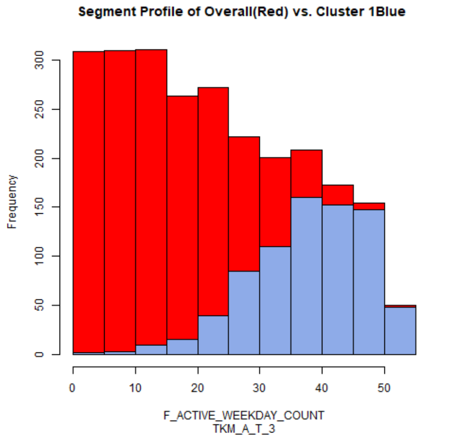


Cluster 2 (2,396 IDs)

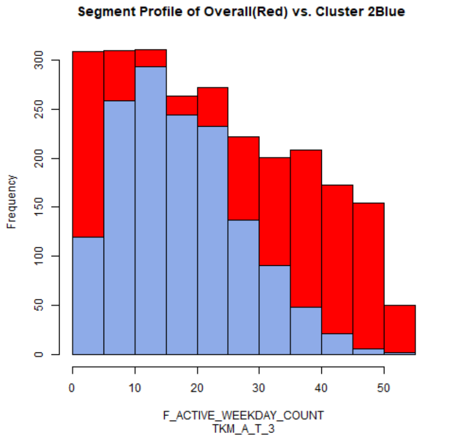


TKM-A-T

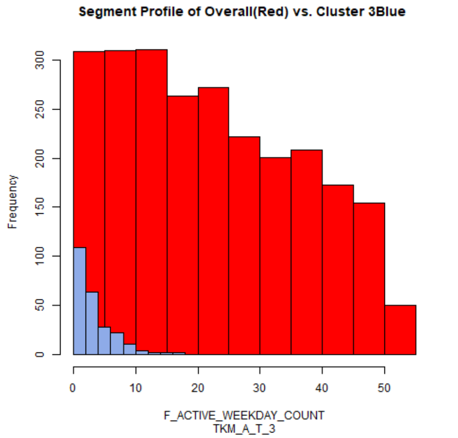
Cluster 1 (774 IDs)



Cluster 2 (1,452 IDs)



Cluster 3 (244 IDs)

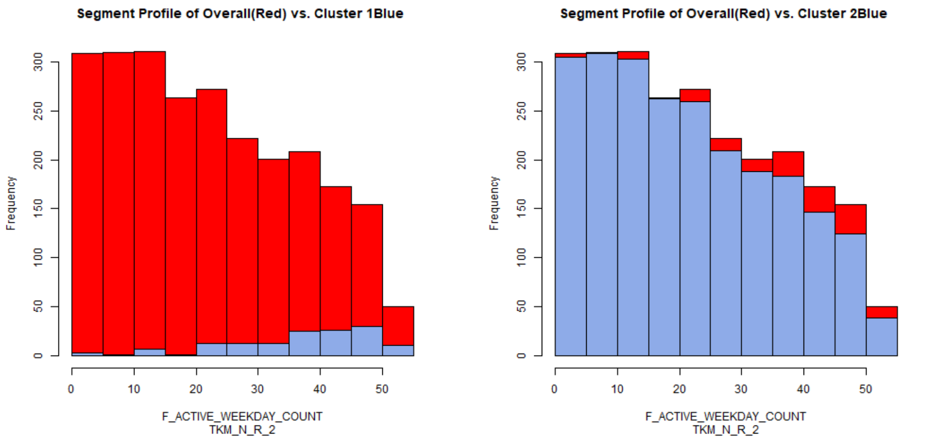


CLUSTER COMPARISON (TRAD. K-MEANS)

Chosen One of 28 Transactional Attributes (F_ACTIVE_WEEKDAY_COUNT) For Apple-to-Apple Comparison

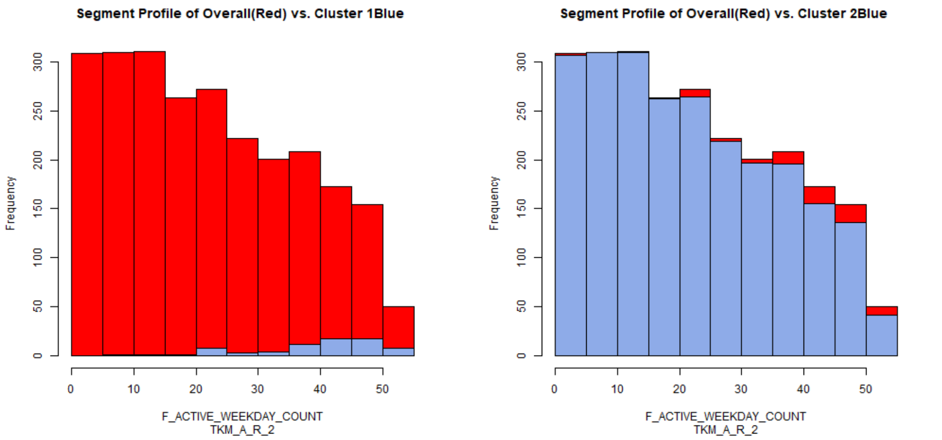
TKM-N-R

Cluster 1 (143 IDs) | Cluster 2 (2,327 IDs)



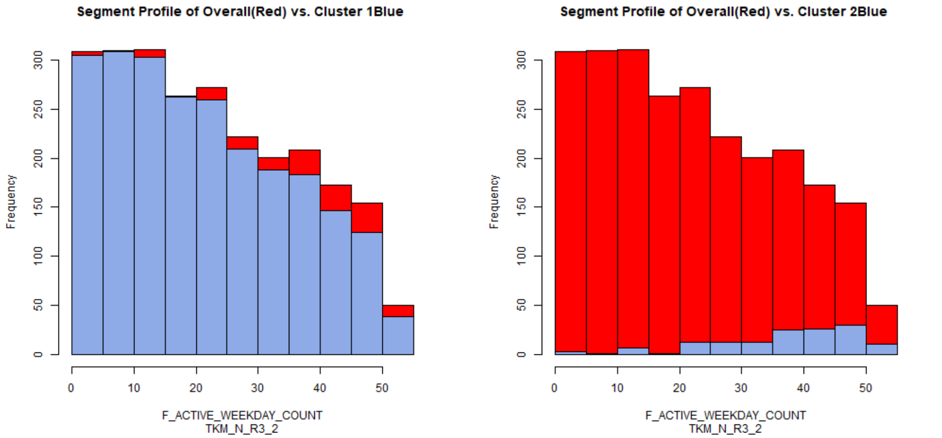
TKM-A-R

Cluster 1 (74 IDs) | Cluster 2 (2,396 IDs)



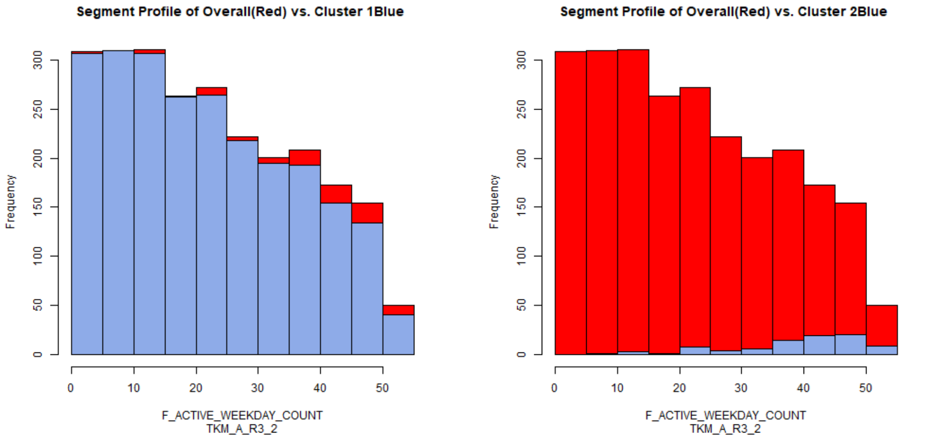
TKM-N-R3

Cluster 1 (2,327 IDs) | Cluster 2 (143 IDs)

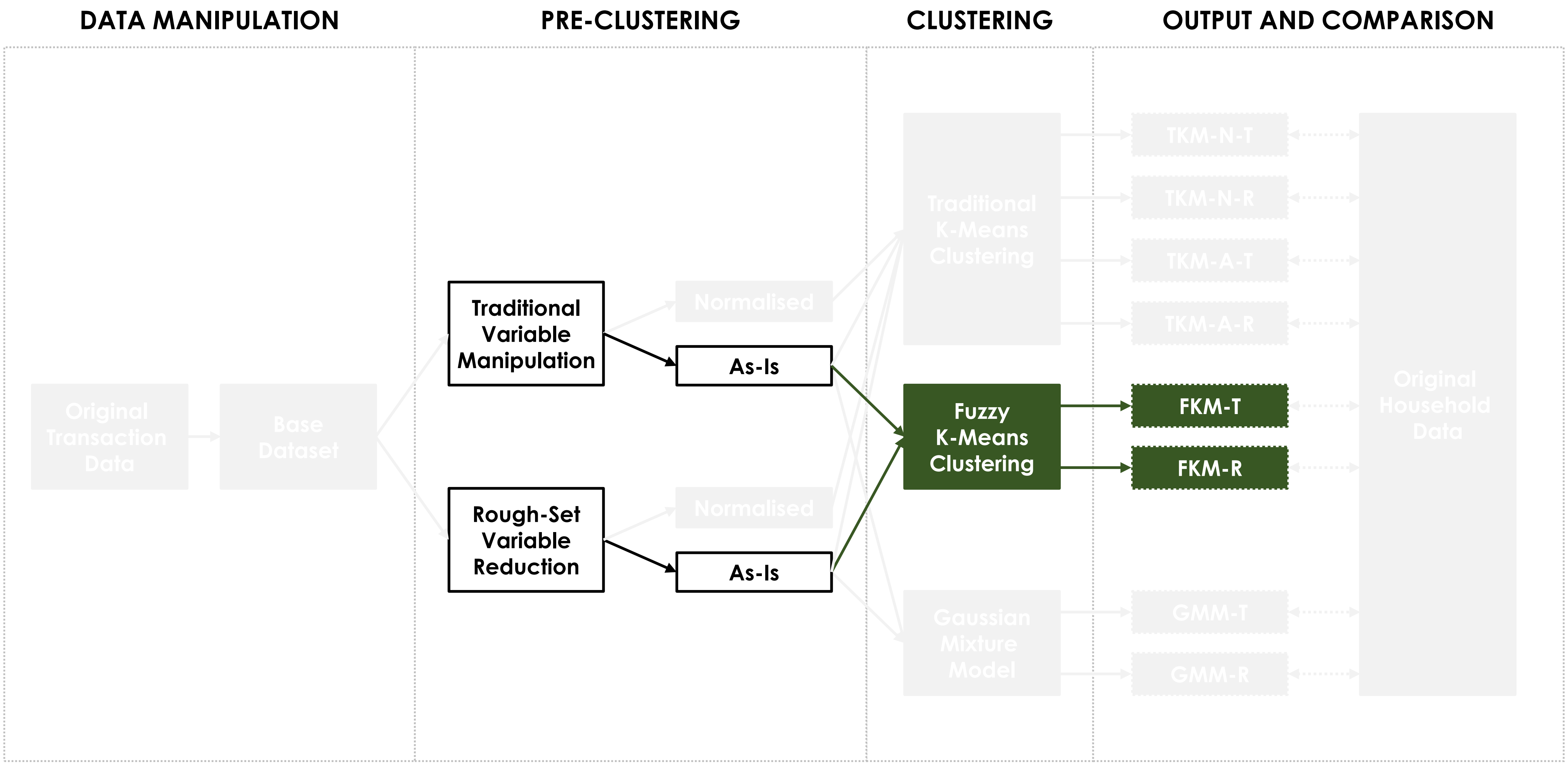


TKM-A-R3

Cluster 1 (2,384 IDs) | Cluster 2 (86 IDs)



CAPSTONE WORKFLOW

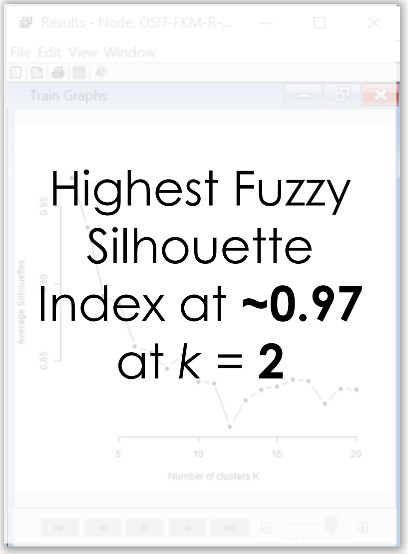


ANALYSIS & RESULTS (FUZZY K-MEANS)

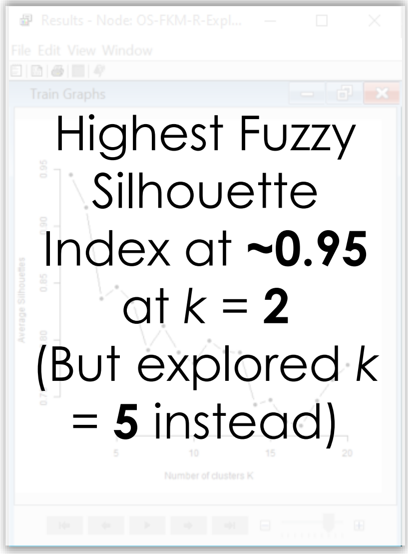
FKM-T

Average Silhouette Index (k = 2 to 20)

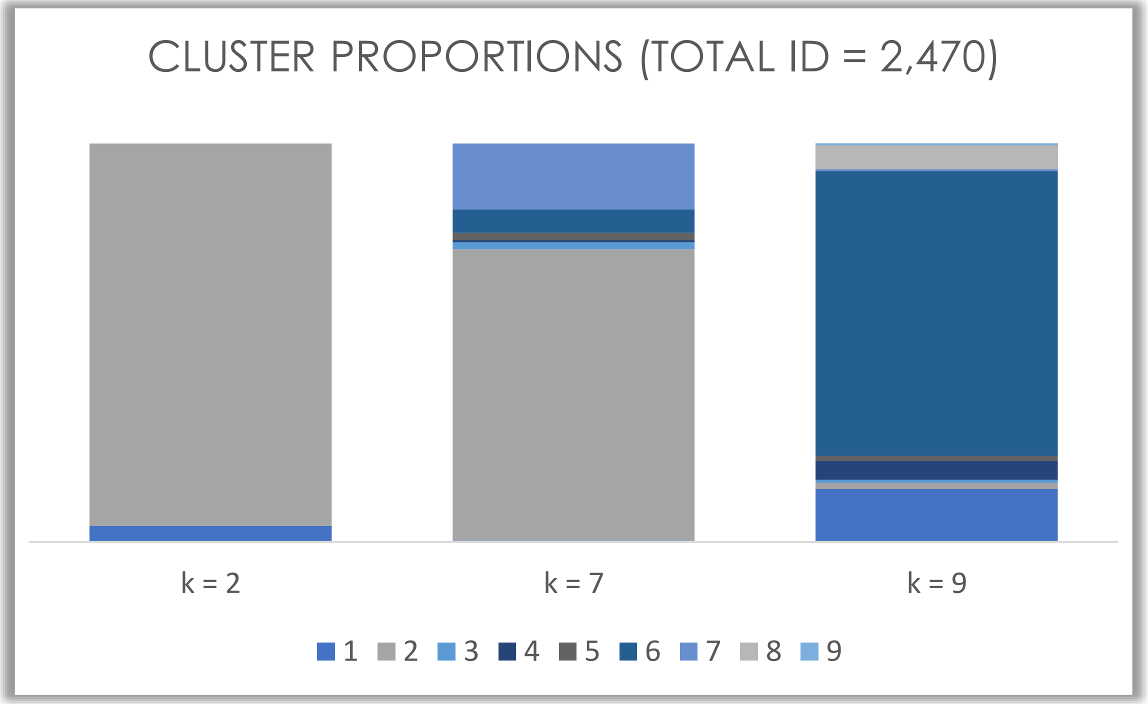
FKM-R



FKM-R3



FKM-R had the highest silhouette index values amongst TKM and FKM. Possibility of exploring other cluster counts than just the one with the highest silhouette value



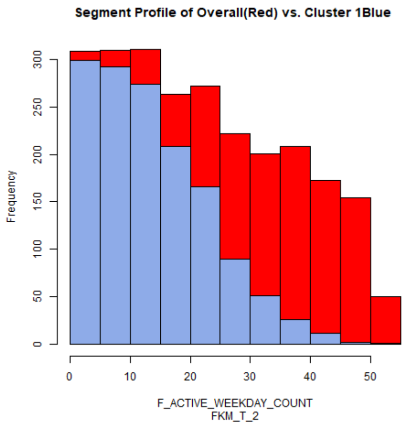
Highly uneven distribution, despite the high silhouette values

CLUSTER COMPARISON (FUZZY K-MEANS)

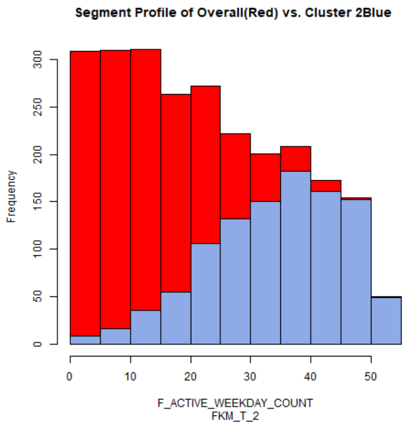
Chosen One of 28 Transactional Attributes (F_ACTIVE_WEEKDAY_COUNT) For Apple-to-Apple Comparison

FKM-T

Cluster 1 (1,421 IDs)

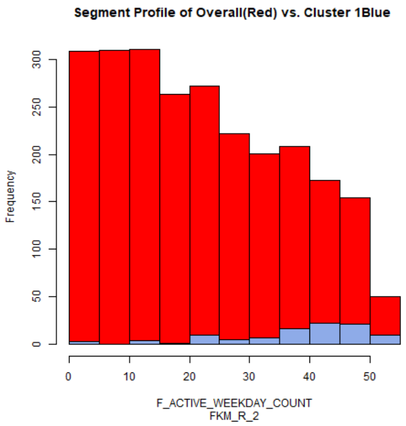


Cluster 2 (1,049 IDs)

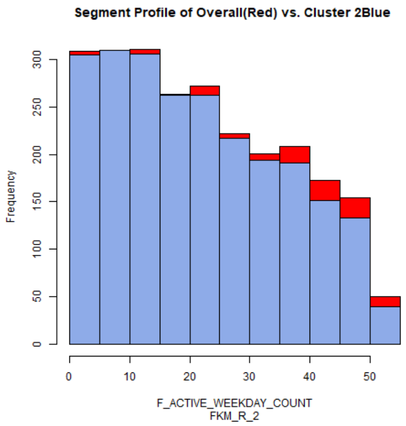


FKM-R

Cluster 1 (100 IDs)

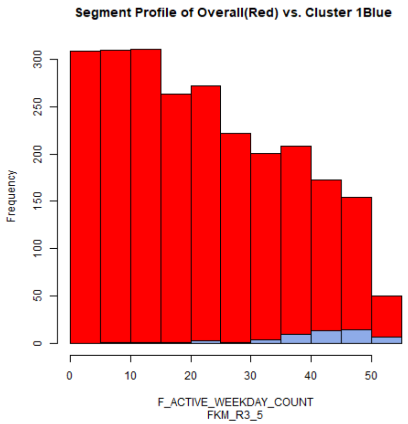


Cluster 2 (2,370 IDs)

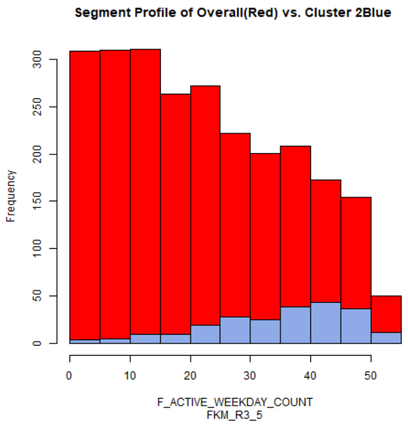


FKM-R3

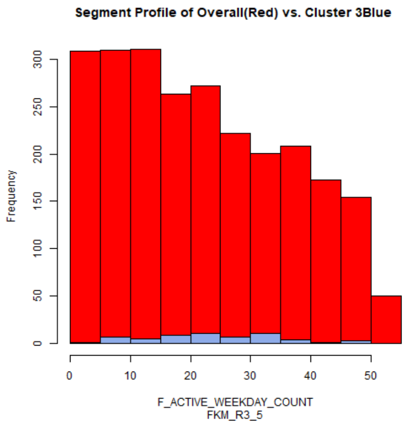
Cluster 1 (57 IDs)



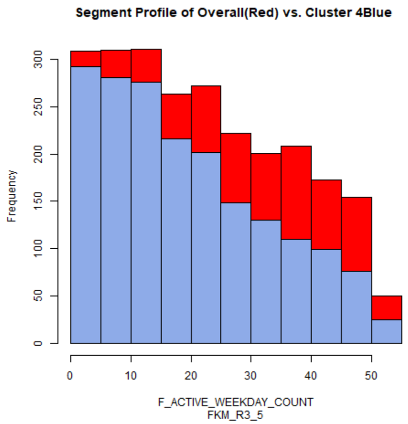
Cluster 2 (233 IDs)



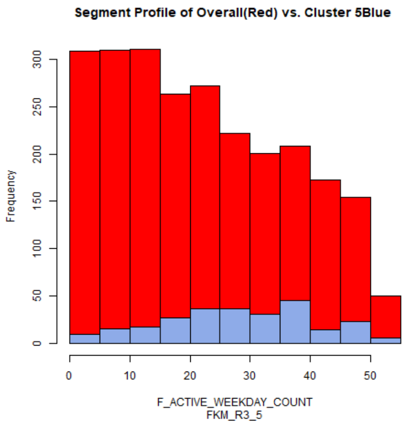
Cluster 3 (59 IDs)



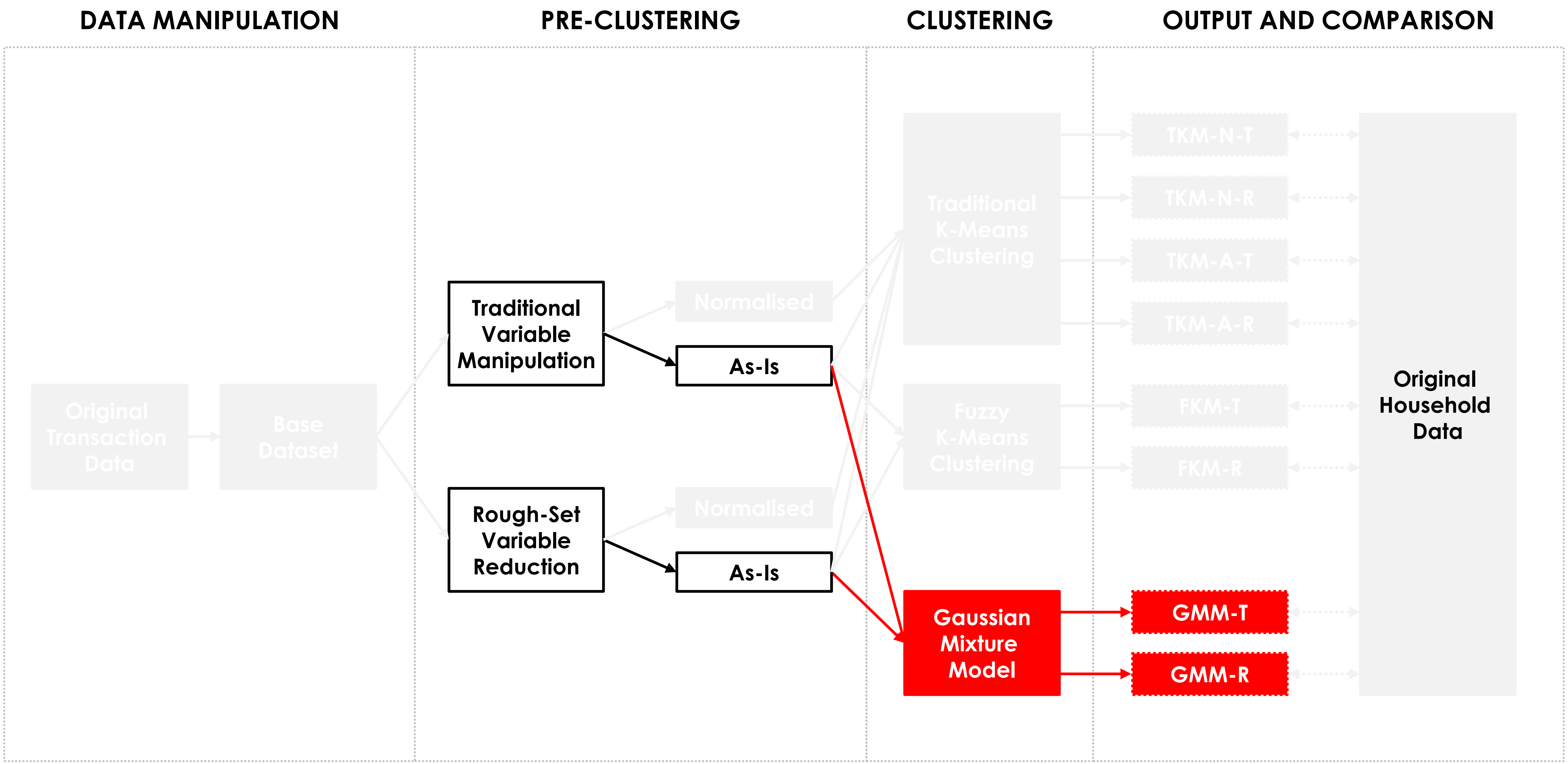
Cluster 4 (1,856 IDs)



Cluster 5 (265 IDs)

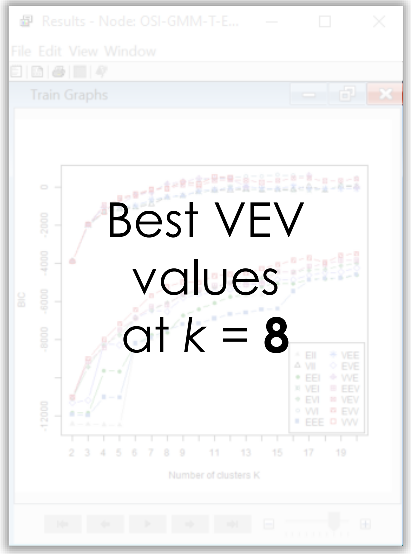


CAPSTONE WORKFLOW



ANALYSIS & RESULTS (GAUSSIAN)

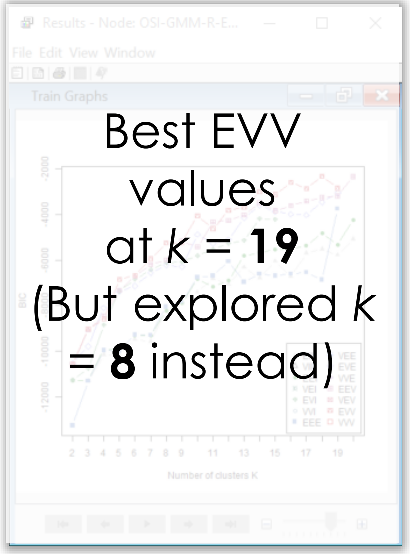
GMM-T



Best VEV values at k = **8**

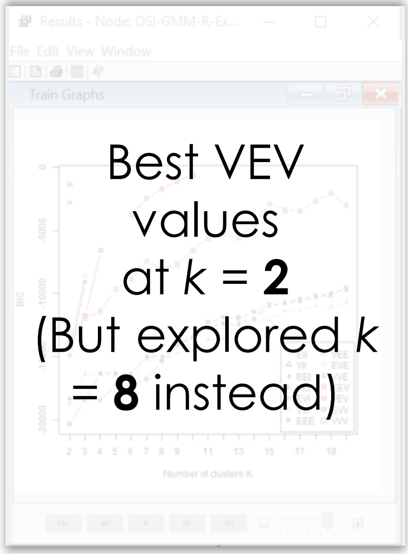
Average Silhouette Index
(k = 2 to 20)

GMM-R



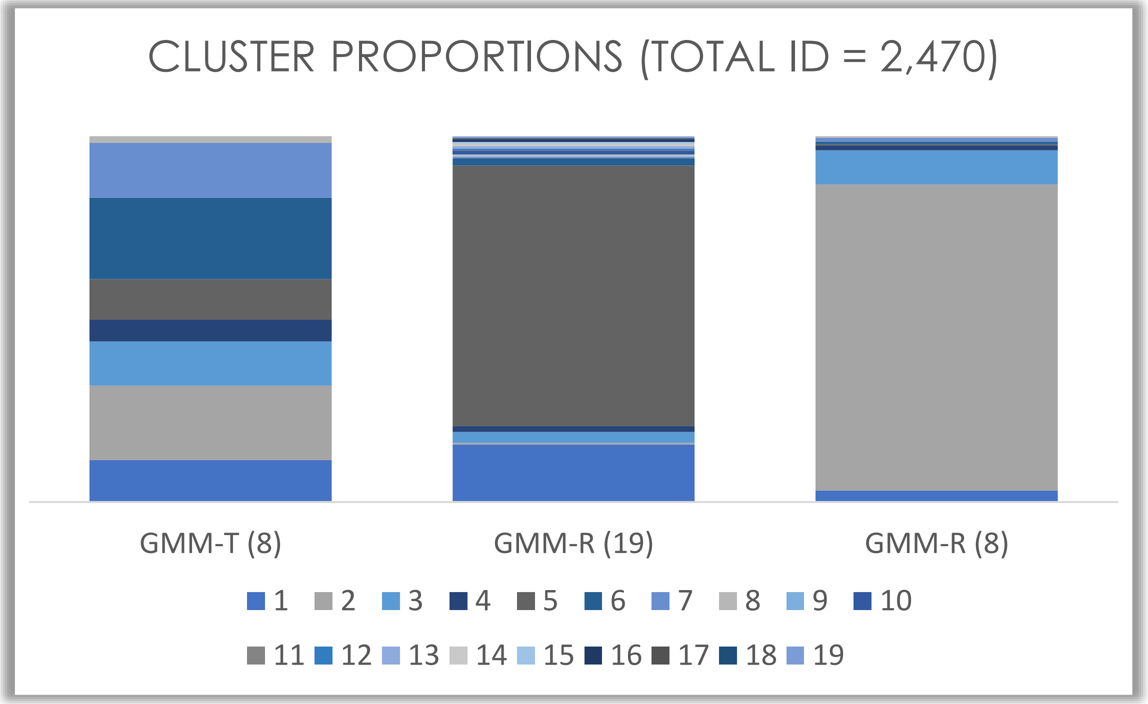
Best EVV values at k = **19** (But explored k = **8** instead)

GMM-R3



Best VEV values at k = **2** (But explored k = **8** instead)

GMM-R had optimal cluster count of 19, but it is too granular. Though sub-optimal, capstone super-imposed k=8 onto GMM-R



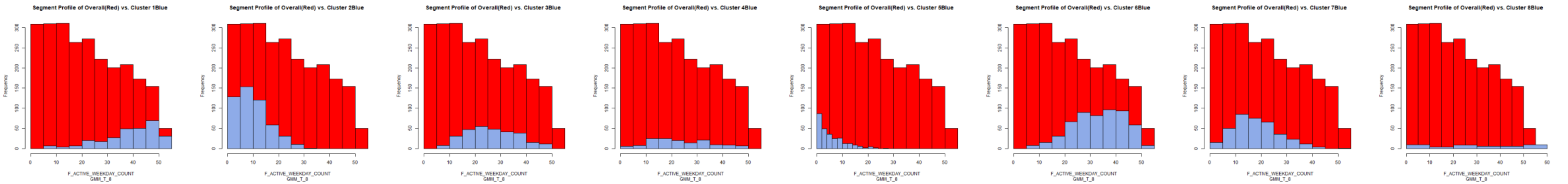
Reduct dataset had highly uneven distribution

CLUSTER COMPARISON (GAUSSIAN)

Chosen One of 28 Transactional Attributes (F_ACTIVE_WEEKDAY_COUNT) For Apple-to-Apple Comparison

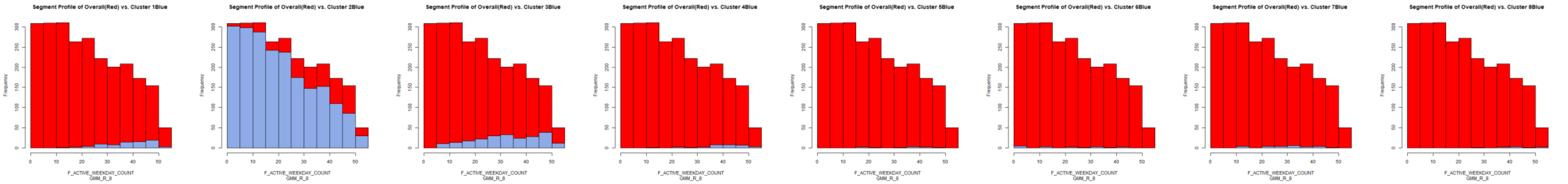
GMM-T

Cluster 1 (283 IDs)	Cluster 2 (504 IDs)	Cluster 3 (298 IDs)	Cluster 4 (146 IDs)	Cluster 5 (274 IDs)	Cluster 6 (552 IDs)	Cluster 7 (368 IDs)	Cluster 8 (45 IDs)



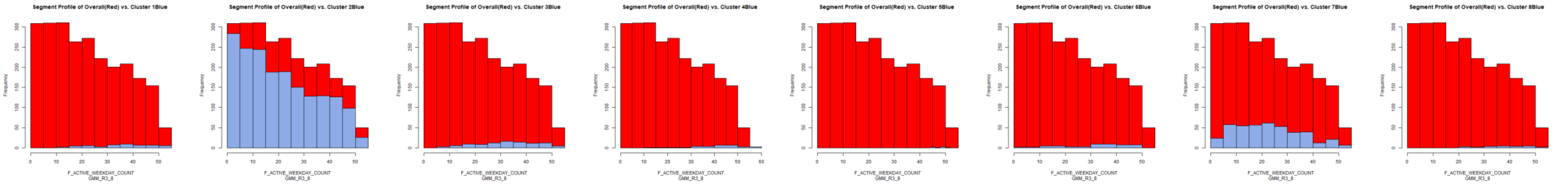
GMM-R (8)

Cluster 1 (78 IDs)	Cluster 2 (2,067 IDs)	Cluster 3 (231 IDs)	Cluster 4 (32 IDs)	Cluster 5 (10 IDs)	Cluster 6 (17 IDs)	Cluster 7 (26 IDs)	Cluster 8 (9 IDs)

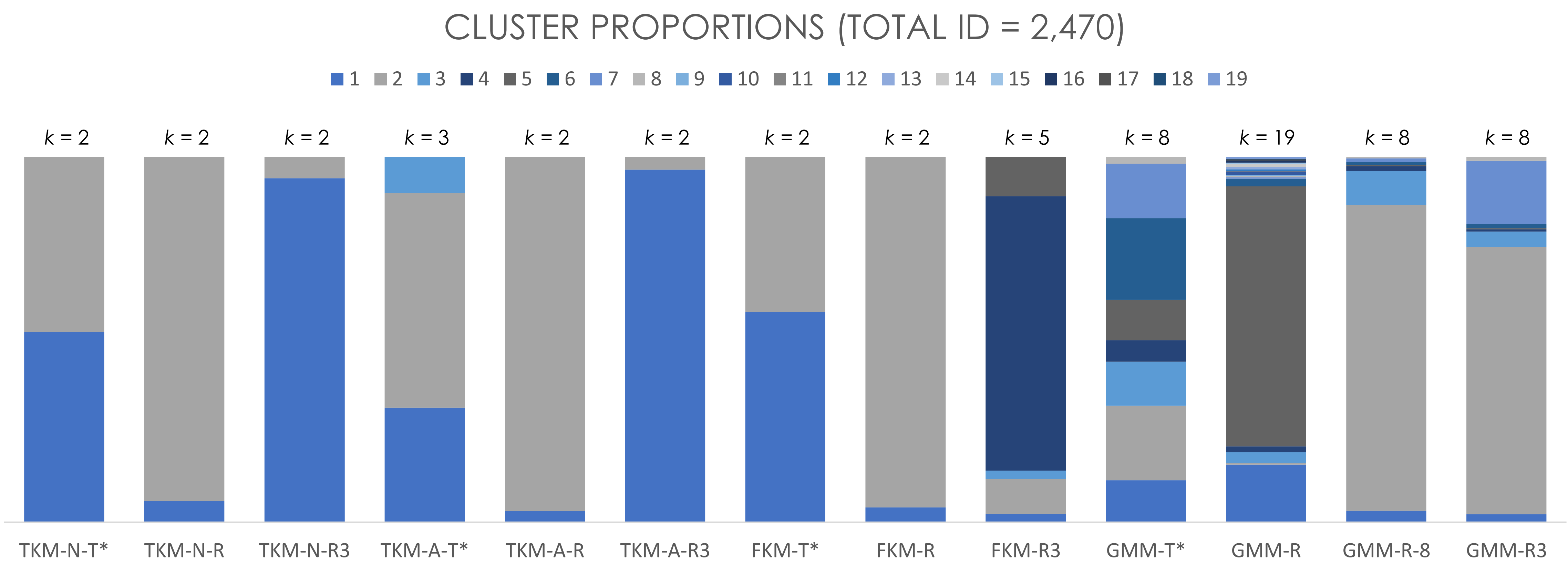


GMM-R3 (8)

Cluster 1 (54 IDs)	Cluster 2 (1,808 IDs)	Cluster 3 (103 IDs)	Cluster 4 (16 IDs)	Cluster 5 (7 IDs)	Cluster 6 (28 IDs)	Cluster 7 (429 IDs)	Cluster 8 (25 IDs)

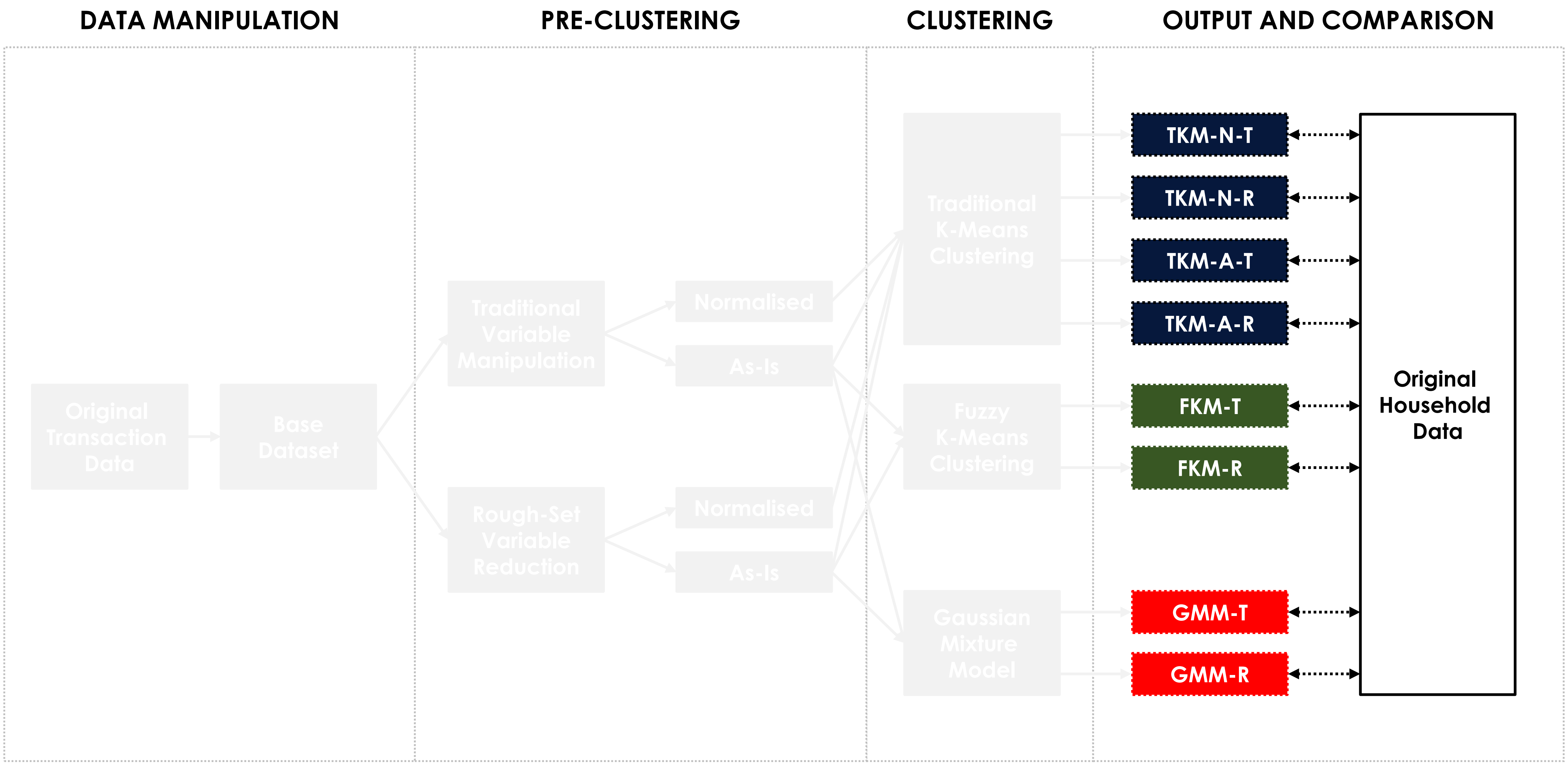


ANALYSIS & RESULTS (OVERVIEW)



- Optimised Cluster Counts Were Higher For GMM Than Other Clustering Algos (TKM And FKM)
- Non-Reduct Data* Had More Evenly Spread Cluster Proportions Than Reduct Data
- Little Clustering Difference Between Reduct-2 and Reduct-3 Dataset

CAPSTONE WORKFLOW

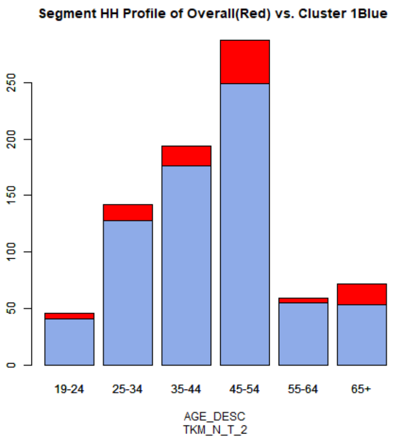


CLUSTER COMPARISON (TRAD. K-MEANS)

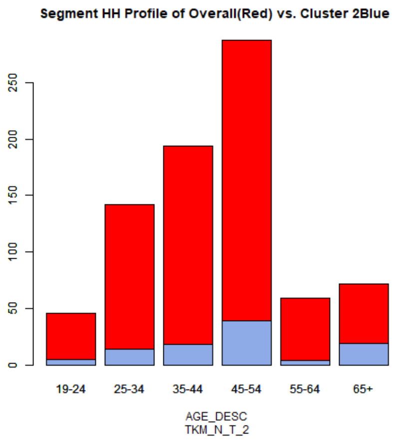
Chosen One of 6 Demographic Attributes (AGE_DESC) For Apple-to-Apple Comparison

TKM-N-T

Cluster 1 (1,287 IDs)

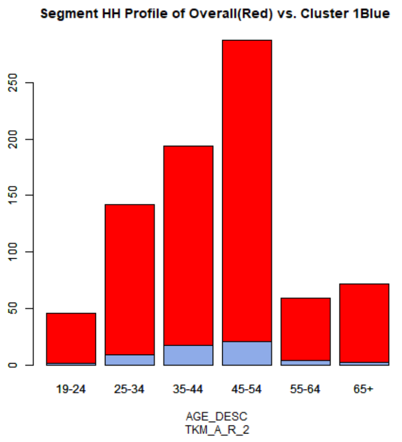


Cluster 2 (1,183 IDs)

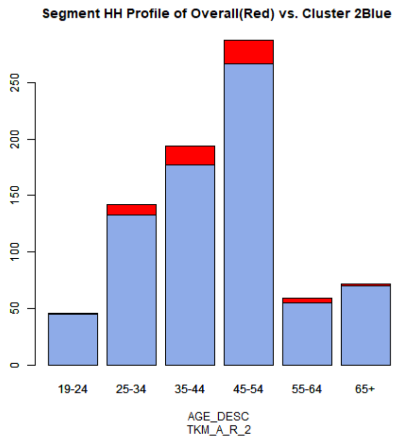


TKM-A-R

Cluster 1 (74 IDs)

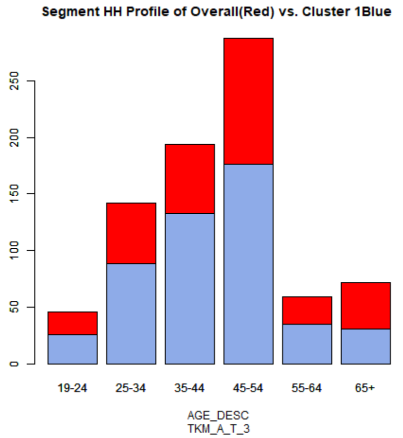


Cluster 2 (2,396 IDs)

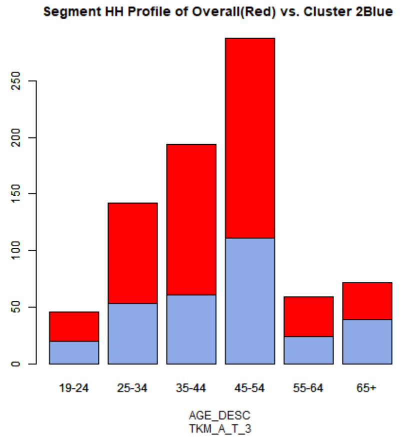


TKM-A-T

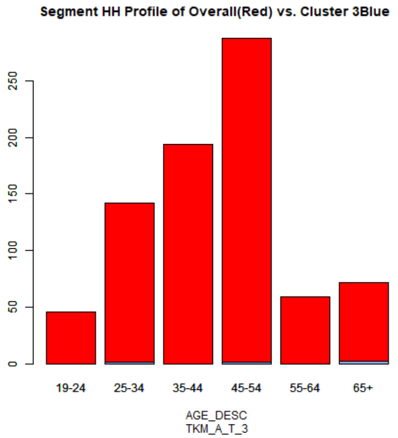
Cluster 1 (774 IDs)



Cluster 2 (1,452 IDs)



Cluster 3 (244 IDs)

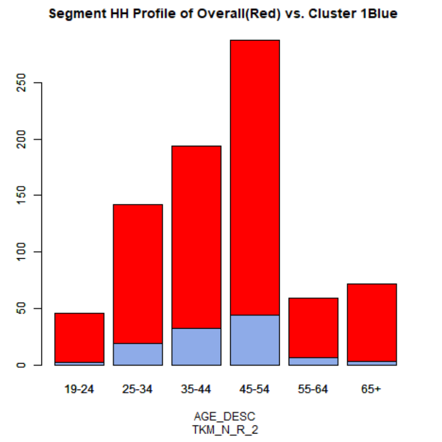


CLUSTER COMPARISON (TRAD. K-MEANS)

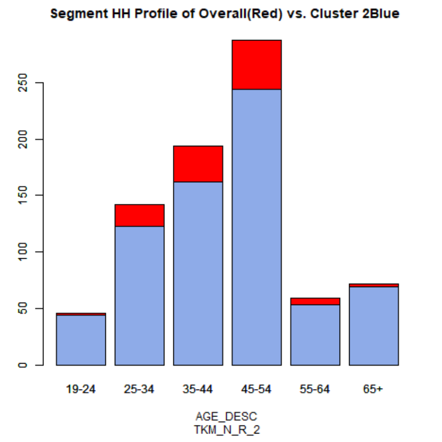
Chosen One of 6 Demographic Attributes (AGE_DESC) For Apple-to-Apple Comparison

TKM-N-R

Cluster 1 (143 IDs)

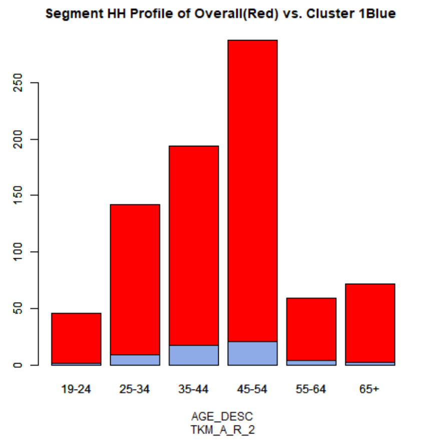


Cluster 2 (2,327 IDs)

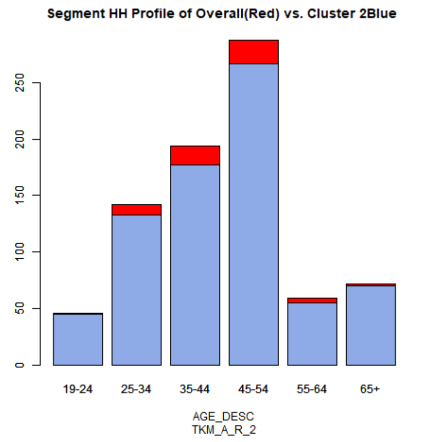


TKM-A-R

Cluster 1 (74 IDs)

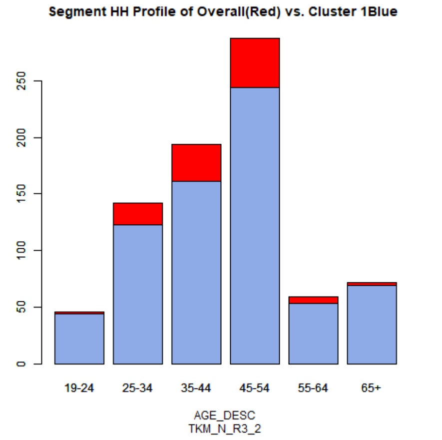


Cluster 2 (2,396 IDs)

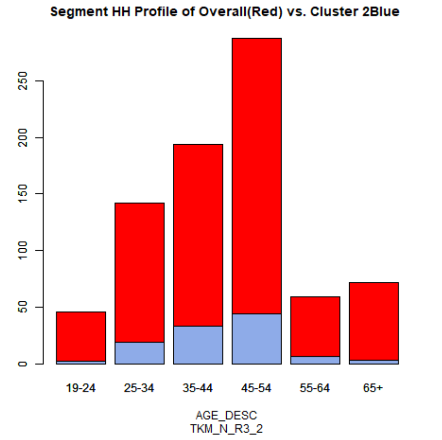


TKM-N-R3

Cluster 1 (2,327 IDs)

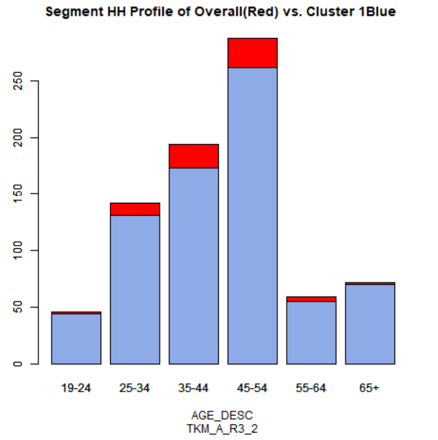


Cluster 2 (143 IDs)

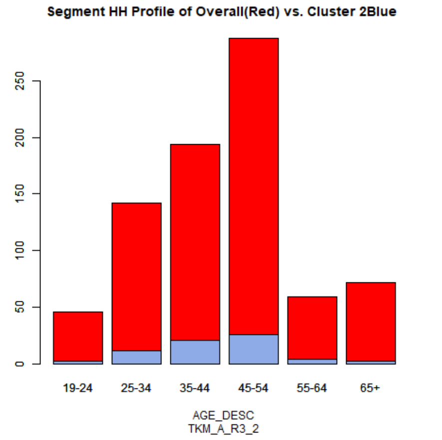


TKM-A-R3

Cluster 1 (2,384 IDs)



Cluster 2 (86 IDs)

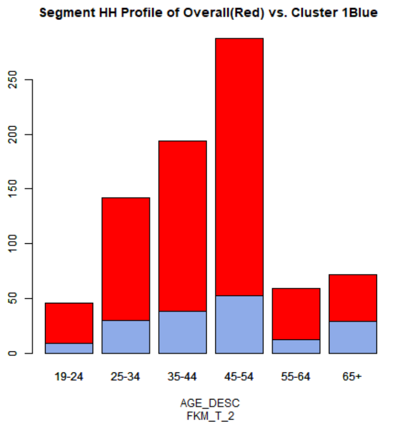


CLUSTER COMPARISON (FUZZY K-MEANS)

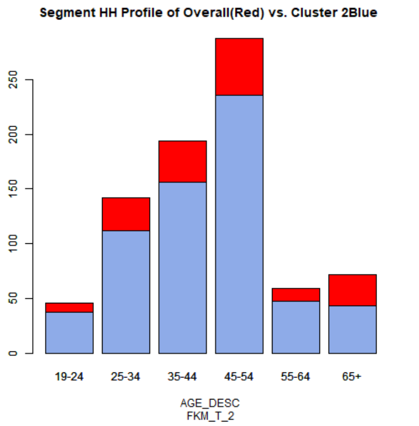
Chosen One of 6 Demographic Attributes (AGE_DESC) For Apple-to-Apple Comparison

FKM-T

Cluster 1 (1,421 IDs)

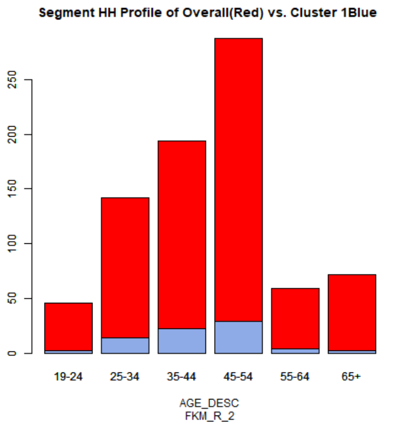


Cluster 2 (1,049 IDs)

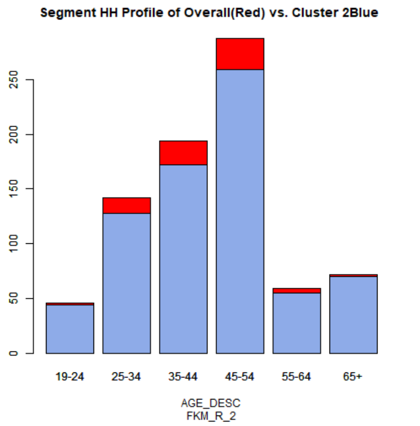


FKM-R

Cluster 1 (100 IDs)

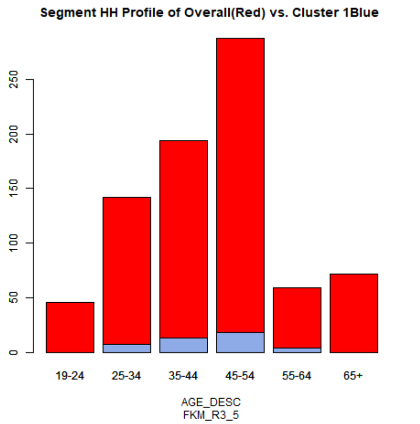


Cluster 2 (2,370 IDs)

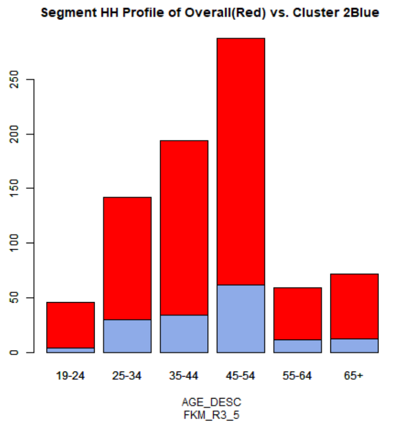


FKM-R3

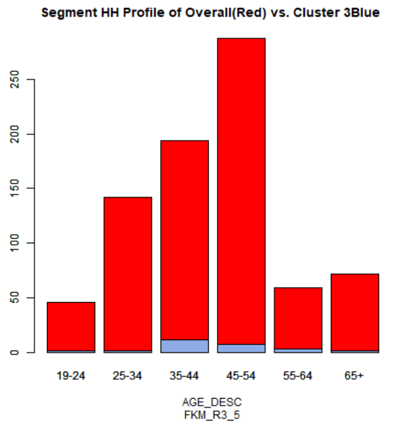
Cluster 1 (57 IDs)



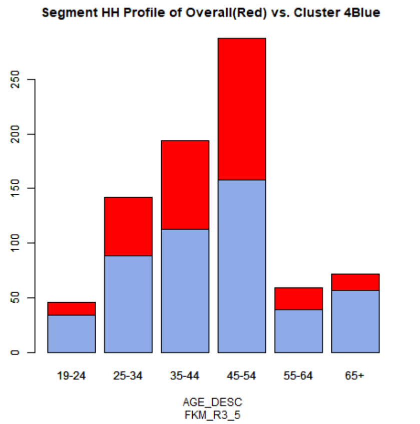
Cluster 2 (233 IDs)



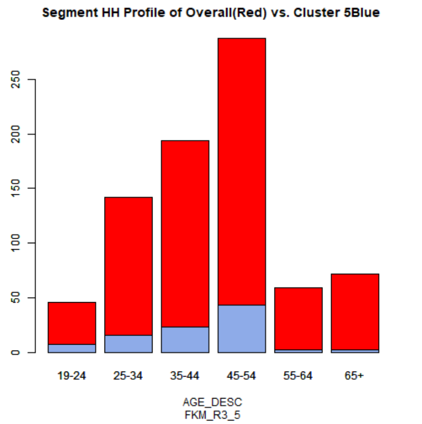
Cluster 3 (59 IDs)



Cluster 4 (1,856 IDs)



Cluster 5 (265 IDs)

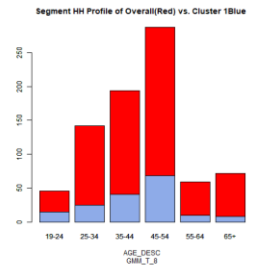


CLUSTER COMPARISON (GAUSSIAN)

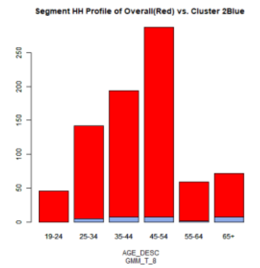
Chosen One of 6 Demographic Attributes (AGE_DESC) For Apple-to-Apple Comparison

GMM-T

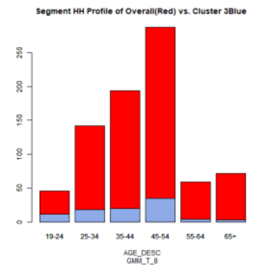
Cluster 1 (283 IDs)



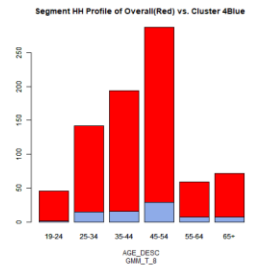
Cluster 2 (504 IDs)



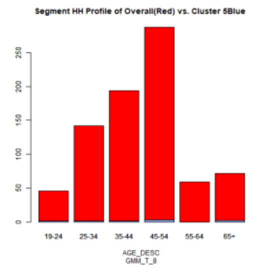
Cluster 3 (298 IDs)



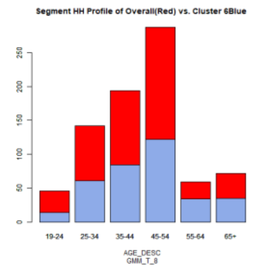
Cluster 4 (146 IDs)



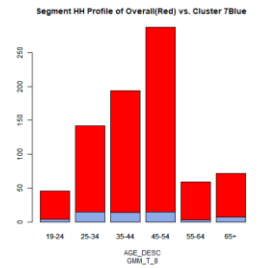
Cluster 5 (274 IDs)



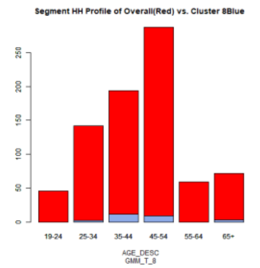
Cluster 6 (552 IDs)



Cluster 7 (368 IDs)

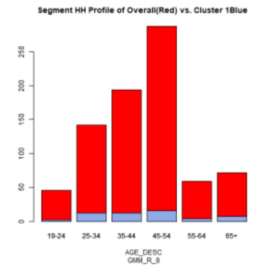


Cluster 8 (45 IDs)

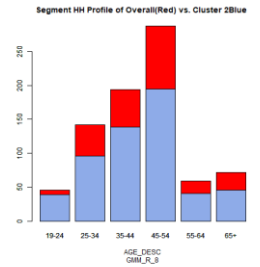


GMM-R (8)

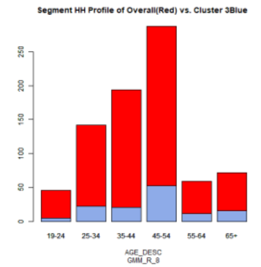
Cluster 1 (78 IDs)



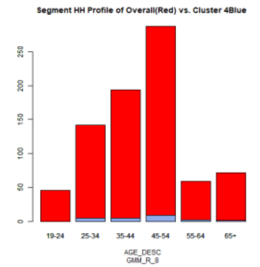
Cluster 2 (2,067 IDs)



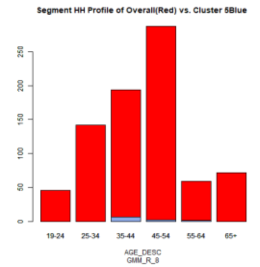
Cluster 3 (231 IDs)



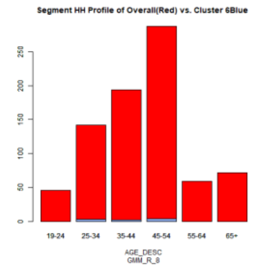
Cluster 4 (32 IDs)



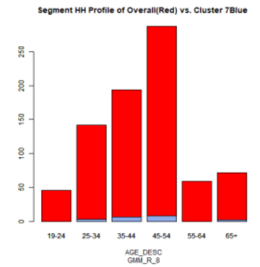
Cluster 5 (10 IDs)



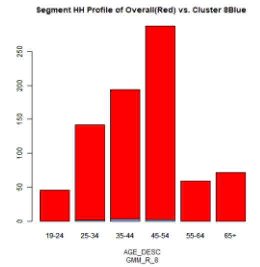
Cluster 6 (17 IDs)



Cluster 7 (26 IDs)

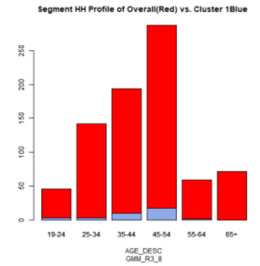


Cluster 8 (9 IDs)

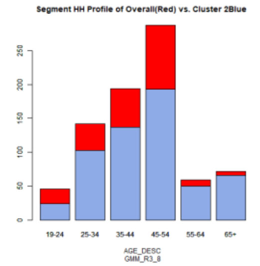


GMM-R3 (8)

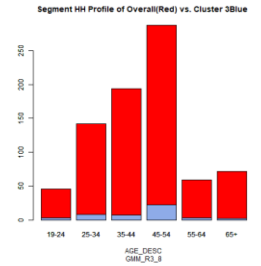
Cluster 1 (54 IDs)



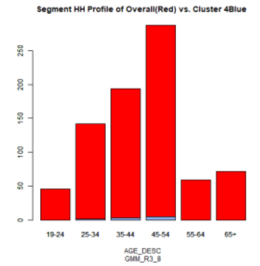
Cluster 2 (1,808 IDs)



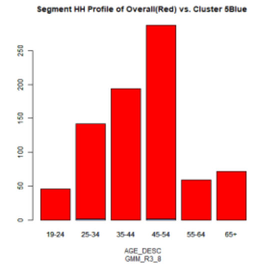
Cluster 3 (103 IDs)



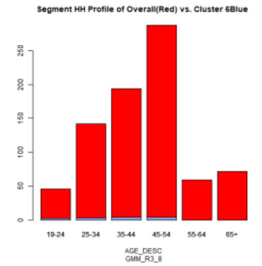
Cluster 4 (16 IDs)



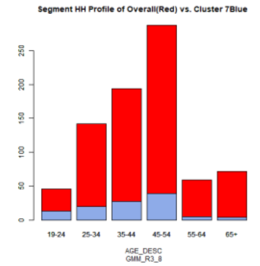
Cluster 5 (7 IDs)



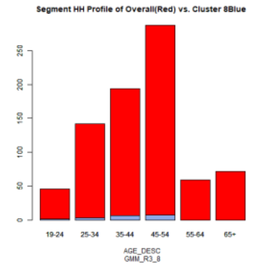
Cluster 6 (28 IDs)



Cluster 7 (429 IDs)



Cluster 8 (25 IDs)



IN SUMMARY

Data Type	Traditional k-means (TKM)	Fuzzy k-means (FKM)	Gaussian Mixture Model (GMM)
Non-Reduct Dataset	• Similar output to FKM • Low optimal cluster counts of 2 and 3 • Even distribution • Distinct cluster traits (transactional) • Distinct cluster traits (demographic)	• Similar output to TKM • Low optimal cluster count of 2 • Even distribution • Distinct cluster traits (transactional) • Distinct cluster traits (demographic)	• High optimal cluster counts of 8 • Even distribution • Somewhat distinct cluster traits (transactional) • Somewhat distinct cluster traits (demographic)
Reduct Dataset	• Low optimal cluster count of 2 • Uneven distribution • Highly distinct cluster traits (transactional) • Distinct cluster traits (demographic) • Little difference seen between reduct-2 and reduct-3	• Low optimal cluster count of 2 • Uneven distribution • Highly distinct cluster traits (transactional) • Distinct cluster traits (demographic) • Reduct-3 had more similar cluster traits than reduct-2	• High optimal cluster counts of 19 • Uneven distribution • Similar cluster traits (transactional) for $k = 8$ • Similar cluster traits (demographic) • Little difference seen between reduct-2 and reduct-3

DISCUSSION

DISTINCTIVE REDUCT DATASET

Rough Set's feature reduction generated more distinctive cluster attributes

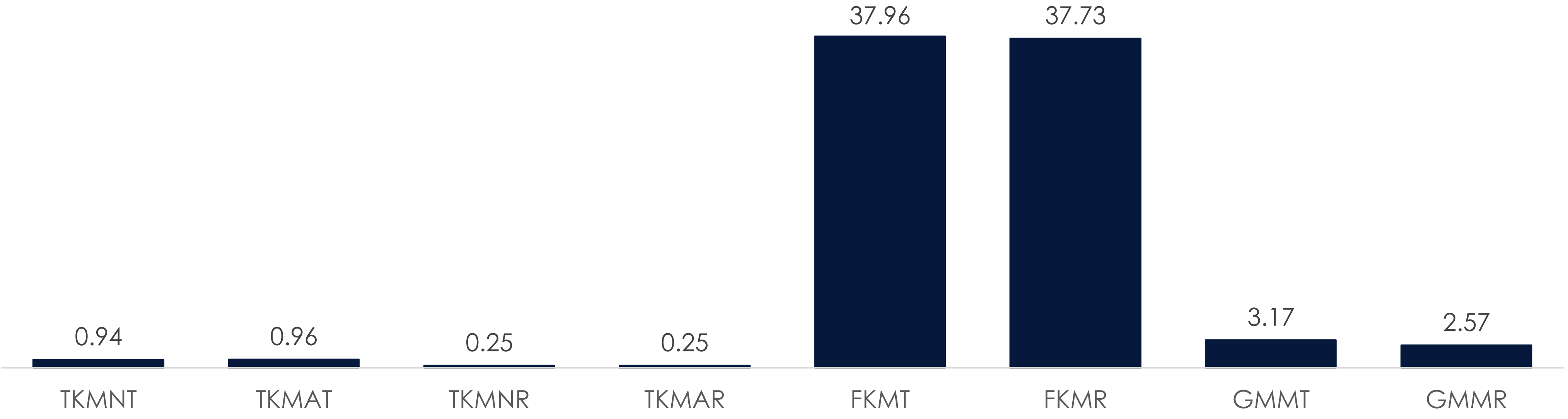
SLOWER SOFT CLUSTERING

Slower processing speed observed for FKM and GMM than TKM

HIGHER k ≠ BETTER & CLEAR CLUSTERS

Higher cluster counts (k>2) does not necessarily show distinctive clusters

Average Execution Time (Secs) of R Output with Cluster k=2 with 20 iterations



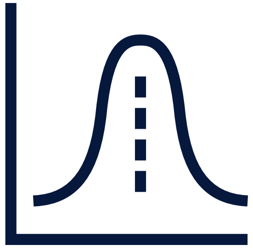
*TKM and FKM time includes silhouette calculations

CONCLUSION & CONTRIBUTIONS

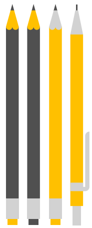
USEFUL ADDITIONS TO A DATA ANALYST's TOOLBOX

TRADITIONAL *K*-MEANS
Clustering Comparison

FUZZY *K*-MEANS
Clustering Comparison



GAUSSIAN MIXTURE MODEL
Clustering Comparison



ROUGH SET
Feature Reduction

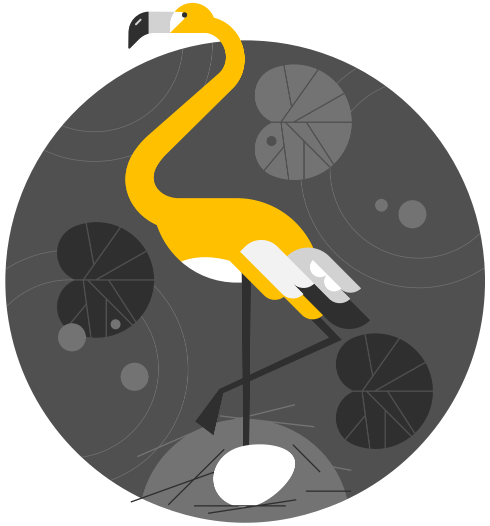
FUTURE WORK

- Exploring other Rough Set reduct method
  - Capstone used global discernibility
  - Available methods include local discernibility, quick reduct.rst and quick reduct.frst
- Explore use of Rough Set reduct on both *pre* and *post* clustering
  - Capstone used reduct *pre* clustering
  - Other journals used reduct *post* clustering to distil significant attributes for each cluster

MILESTONES & REFLECTIONS

NEVER UNDER-ESTIMATE CAPSTONE RIGOUR

Taking additional courses would've sufficed as MITB course requirements, but capstone opportunity proved invaluable



JOURNEY BEGINS WITH THE FIRST STEP

Initially fuzzy; Gained clarity at every step; Where clarity was at its best, scale of work seemed daunting

SUSTAINABLE SAS & R INTEGRATION

R packages complements existing SAS environment. Challenge is to make it sustainable, amidst compatibility concerns

THANK YOU

FOR THE OPPORTUNITY!

SMU
SINGAPORE MANAGEMENT UNIVERSITY

×

sas®