CLUSTER ANALYSIS

2017 NEVADA STATE HOME LOANS

DANA-4840-001 Classification II

TEAM MEMBERS

NAME	ID
Leo Liang	100244775
Qirui Cao	100290055
Young Lee	100366055
Latesh Subramanyam	100365556

TABLE OF CONTENTS

PROBLEM DEFINE AND OBJECTIVE

- According to the research conducted by U.S. Department of Housing and Urban Development in 2016, Nevada state was the worst state for affordable housing in the United States.
- For each record in the original data, there is information related to loan, the property characteristics, the applicant and lender demographics.
- Some of the information (variables) are removed and modified to reduce irrelevant information for the classification analysis.
- Our team implemented the classification analysis, including k-means, k-medoid, hierarchical and density-based clustering method, to find the characteristics of the consumers who applied for the mortgage in Nevada state.
- Target Audience: Anyone interested in how various customer information affects mortgage applications
- Our recommended audience for this cluster analysis may specifically help lawmakers who need to identify patterns and bias or prejudice in mortgage approvals for .

PRE-PROCESSING

ORIGINAL DATA

- The original data contains 178,587 rows and 79 columns (variables)
- 5,201,754 Nulls and 796 Duplicates
- Outlier detection is only conducted on the cleaned data set

CLEANED DATA

- The cleaned data contains 2,000 rows (Sampled) and 16 columns (variables)
- 0 Nulls and 0 Duplicates (Completed cases)
- 85 Outliers detected by Mahalanobis Distance
- 75 Outliers detected by Density-Based Spatial Clustering of Applications with Noise

DATA OVERVIEW

loan_type_name	property_type_name	loan_purpose_name	owner_occupancy_name	loan_amount_000s	applicant_ethnicity_name	applicant_race_name_1	applicant_sex_name	applicant_income_000s	lien_status_name	population
Conventional	One-to-four family dwelling (other than manufactured housing)	Home improvement	Owner-occupied as a principal dwelling	35	Not Hispanic or Latino	Black or African American	Female	37	Secured by a first lien	4083
FHA-insured	One-to-four family dwelling (other than manufactured housing)	Refinancing	Owner-occupied as a principal dwelling	191	Not Hispanic or Latino	Black or African American	Female	65	Secured by a first lien	6791
Conventional	One-to-four family dwelling (other than manufactured housing)	Refinancing	Owner-occupied as a principal dwelling	90	Not Hispanic or Latino	White	Female	32	Secured by a first lien	3835
Conventional	Manufactured housing	Home purchase	Owner-occupied as a principal dwelling	24	Hispanic or Latino	White	Female	29	Secured by a first lien	3360
Conventional	One-to-four family dwelling (other than manufactured housing)	Home purchase	Owner-occupied as a principal dwelling	138	Not Hispanic or Latino	Asian	Male	33	Secured by a first lien	10021

minority_population	tract_to_msamd_income	number_of_owner_occupied_units	number_of_1_to_4_family_units	co_applicant
65.169998	81.54	622	983	no
69.139999	106.51	1347	2332	no
32.619999	118.91	965	1313	no
44.939999	93.24	863	1258	no
58.320000	87.81	1280	2002	no

VARIABLE TABLE

Name	Class	Values
loan_type_name	factor	'Conventional' 'FHA-insured' 'FSA/RHS-guaranteed' 'VA-guaranteed'
property_type_name	factor	'Manufactured housing' 'One-to-four family dwelling (other than manufactured housing)'
loan_purpose_name	factor	'Home improvement' 'Home purchase' 'Refinancing'
owner_occupancy_name	factor	'Not owner-occupied as a principal dwelling' 'Owner-occupied as a principal dwelling'
loan_amount_000s	integer	Num: 1 to 5700
applicant_ethnicity_name	factor	'Hispanic or Latino' 'Not Hispanic or Latino'
applicant_race_name_1	factor	'American Indian or Alaska Native' 'Asian' 'Black or African American' 'Native Hawaiian or Other Pacific Islander' 'White'
applicant_sex_name	factor	'Female' 'Male'
applicant_income_000s	integer	Num: 1 to 1597
lien_status_name	factor	'Not secured by a lien' 'Secured by a first lien' 'Secured by a subordinate lien'
population	integer	Num: 562 to 10078
minority_population	numeric	Num: 4.2 to 95.95
tract_to_msamd_income	numeric	Num: 0 to 289.61
number_of_owner_occupied_units	integer	Num: 29 to 2874
number_of_1_to_4_family_units	integer	Num: 23 to 4247
co_applicant	factor	'no' 'yes'

CLUSTERING ANALYSIS – HOW MANY CLUSTERS?

K MEANS

1. GOWER

SW Score: 2
CH Score: 2

2. EUCLIDEAN

SW Score: 2
CH Score: 2

K MEDOIDS

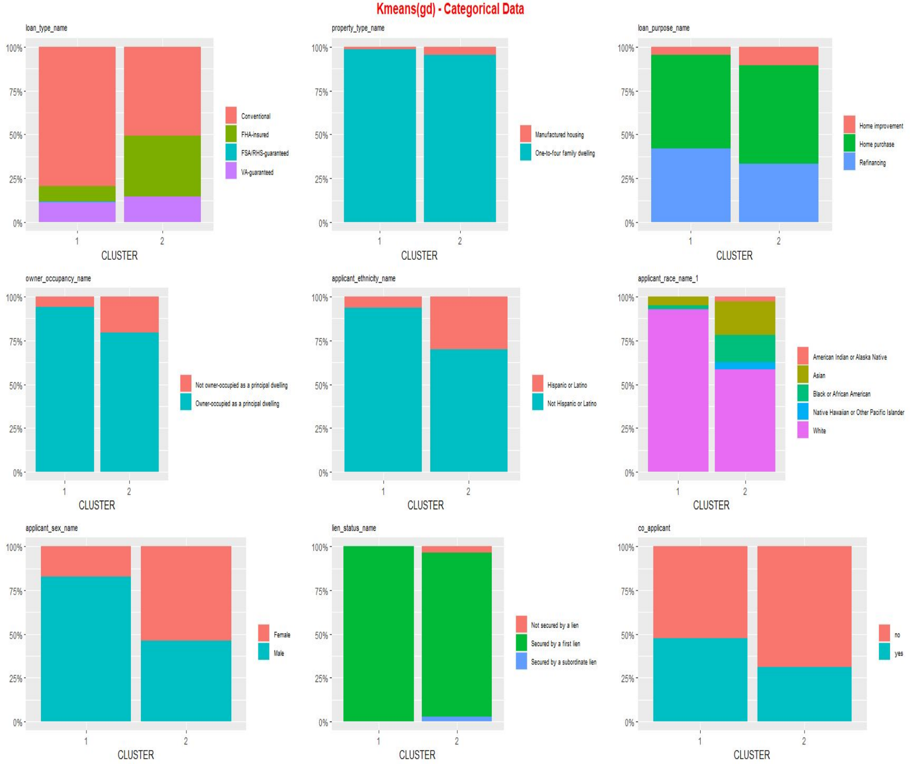
1. GOWER

SW Score: 5

2. EUCLIDEAN

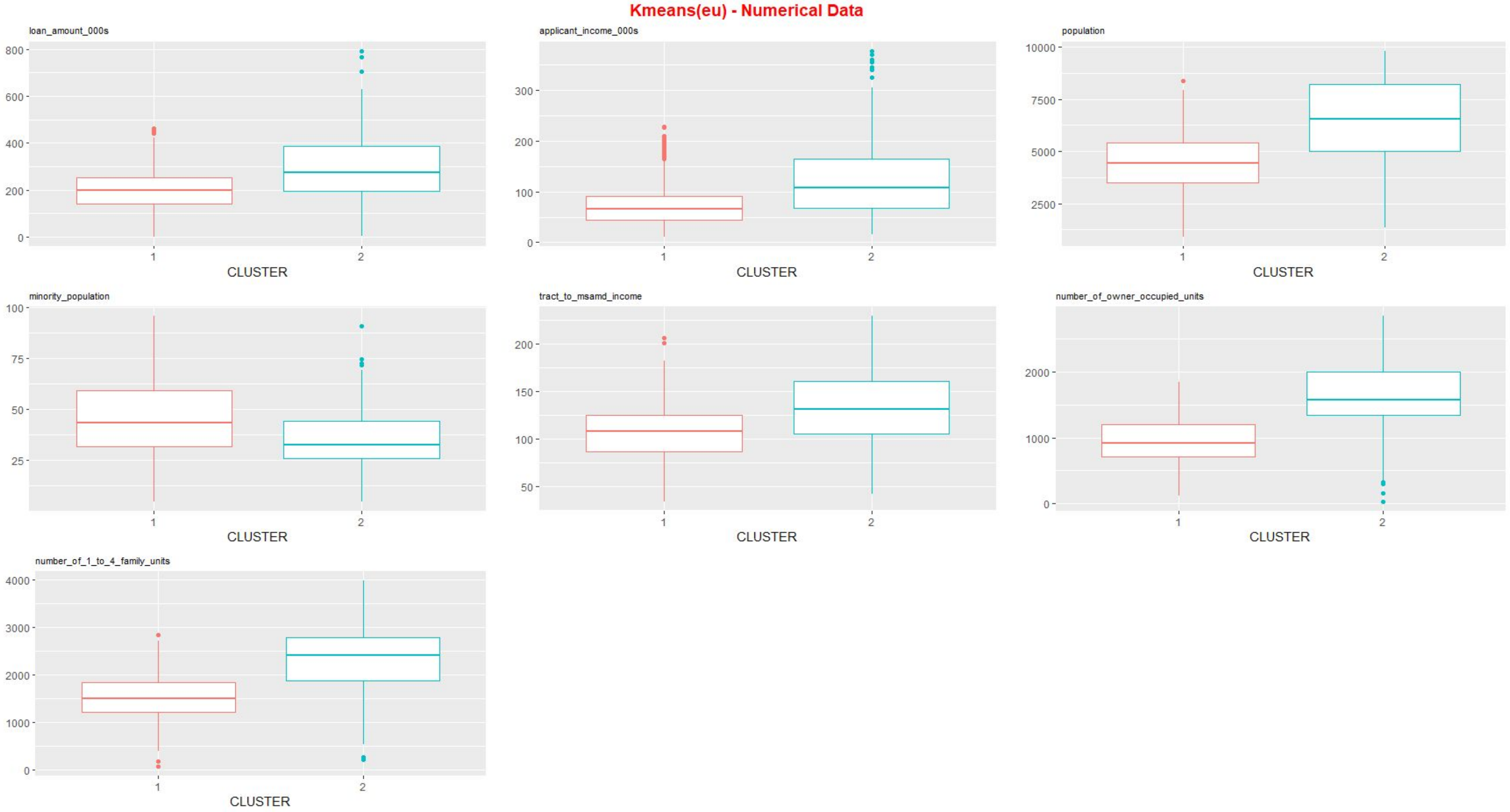
SW Score: 2

CLUSTERING ANALYSIS – K MEANS (GD)

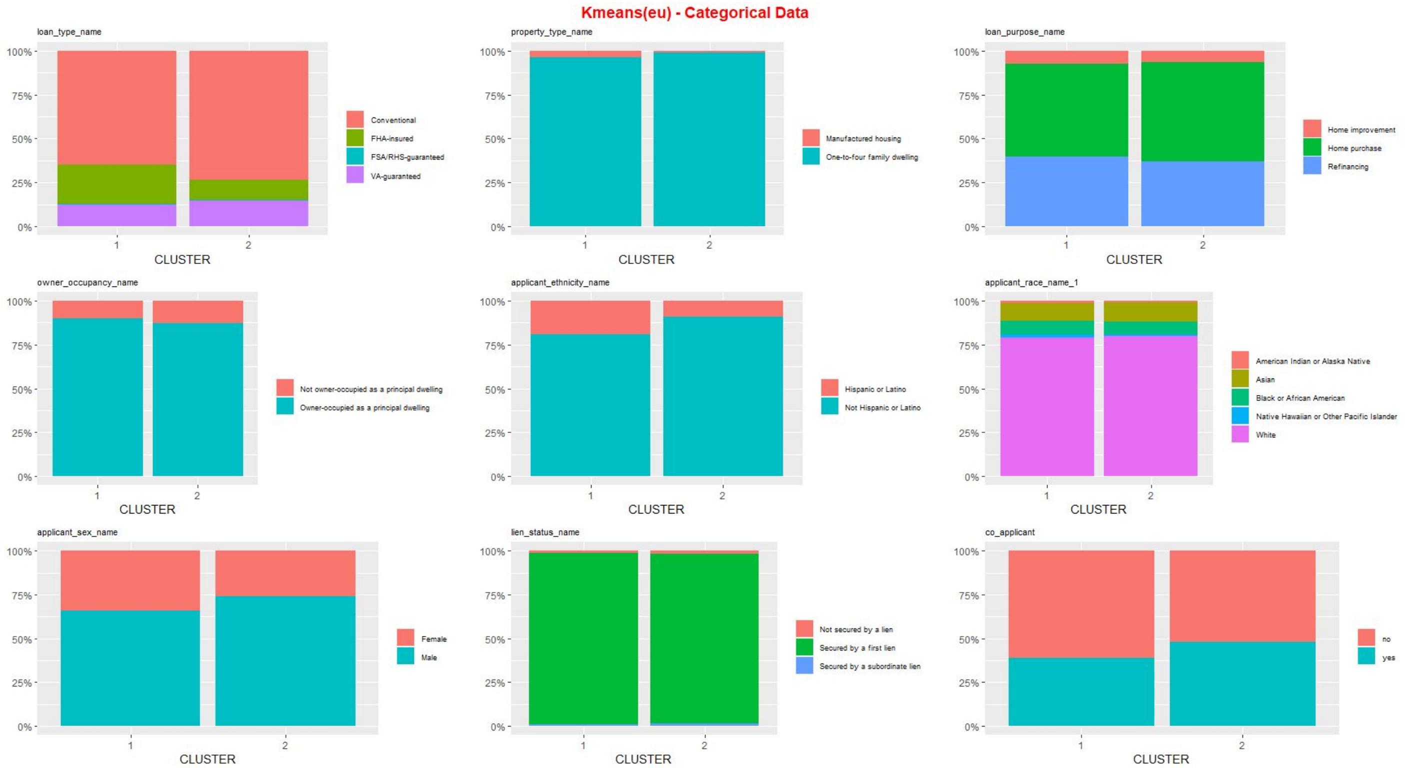


- No significant difference can be seen from the **gower distance model**

CLUSTERING ANALYSIS – K MEANS (EU)



CLUSTERING ANALYSIS – K MEANS (EU)



- No significant difference can be seen from the **categorical variables**

CLUSTER INTERPRETATION - K MEANS (EU)

CLUSTER 1 - SMALL TOWN, MORE MINORITY, LESS NUCLEAR FAMILY

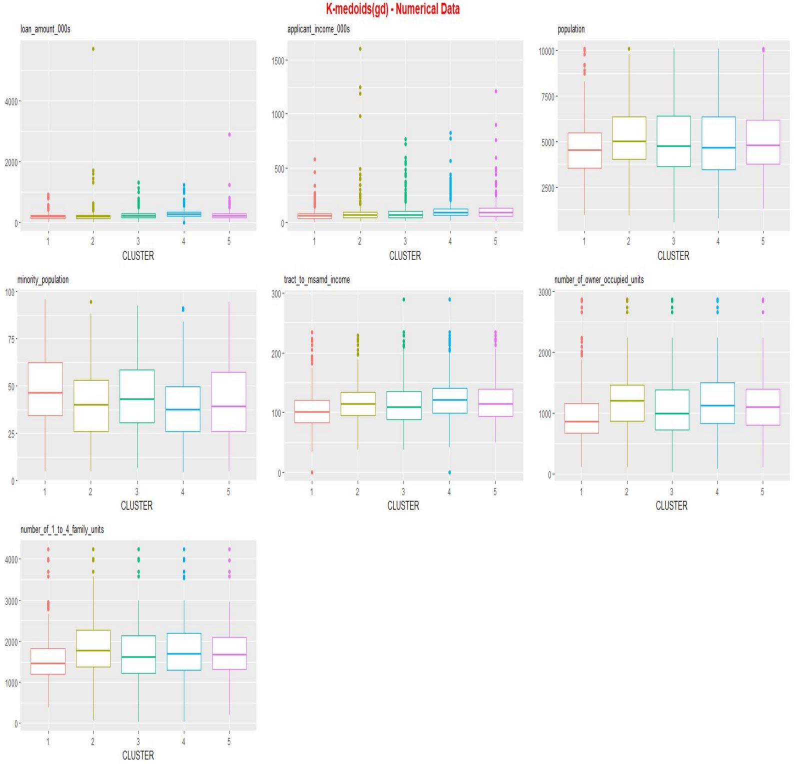
- Less loan amount
- Less applicant income
- **Less population of city/town/area**
- More minority population
- Less Metropolitan Statistical Area/Metropolitan Division income
- **Less number of owner occupied units in the area**
- **Less number of family of 1 to 4 units in the area**

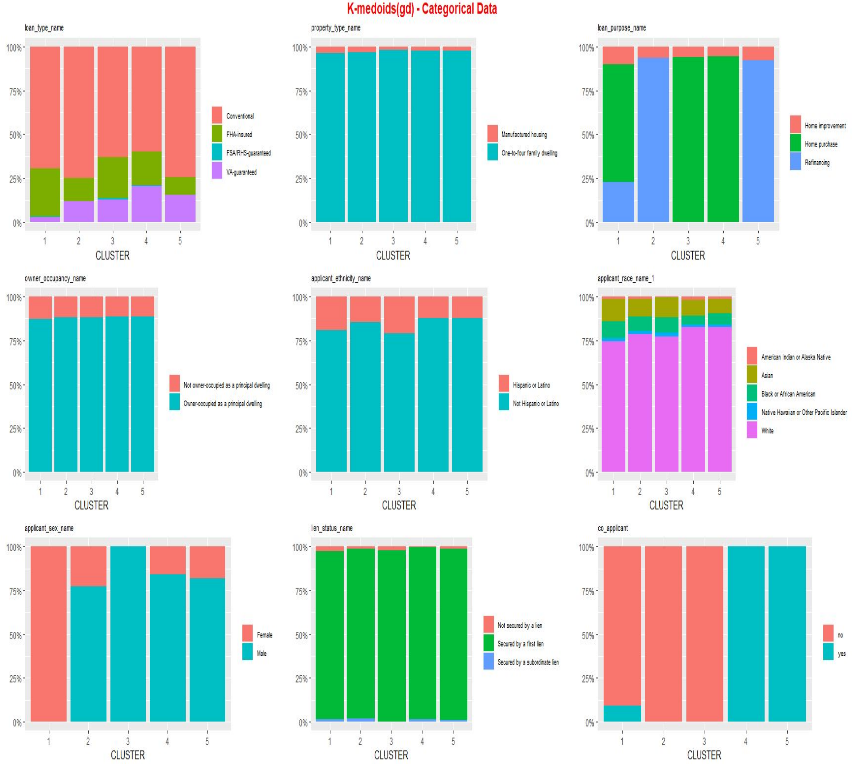
CLUSTER 2 - LARGER TOWN, MORE MAJORITY, MORE NUCLEAR FAMILY

- More loan amount
- More applicant income
- **More population of city/town/area**
- Less minority population
- More Metropolitan Statistical Area/Metropolitan Division income
- **More number of owner occupied units in the area**
- **More number of family of 1 to 4 units in the area**

➢ No significant difference can be seen from the **categorical variables**

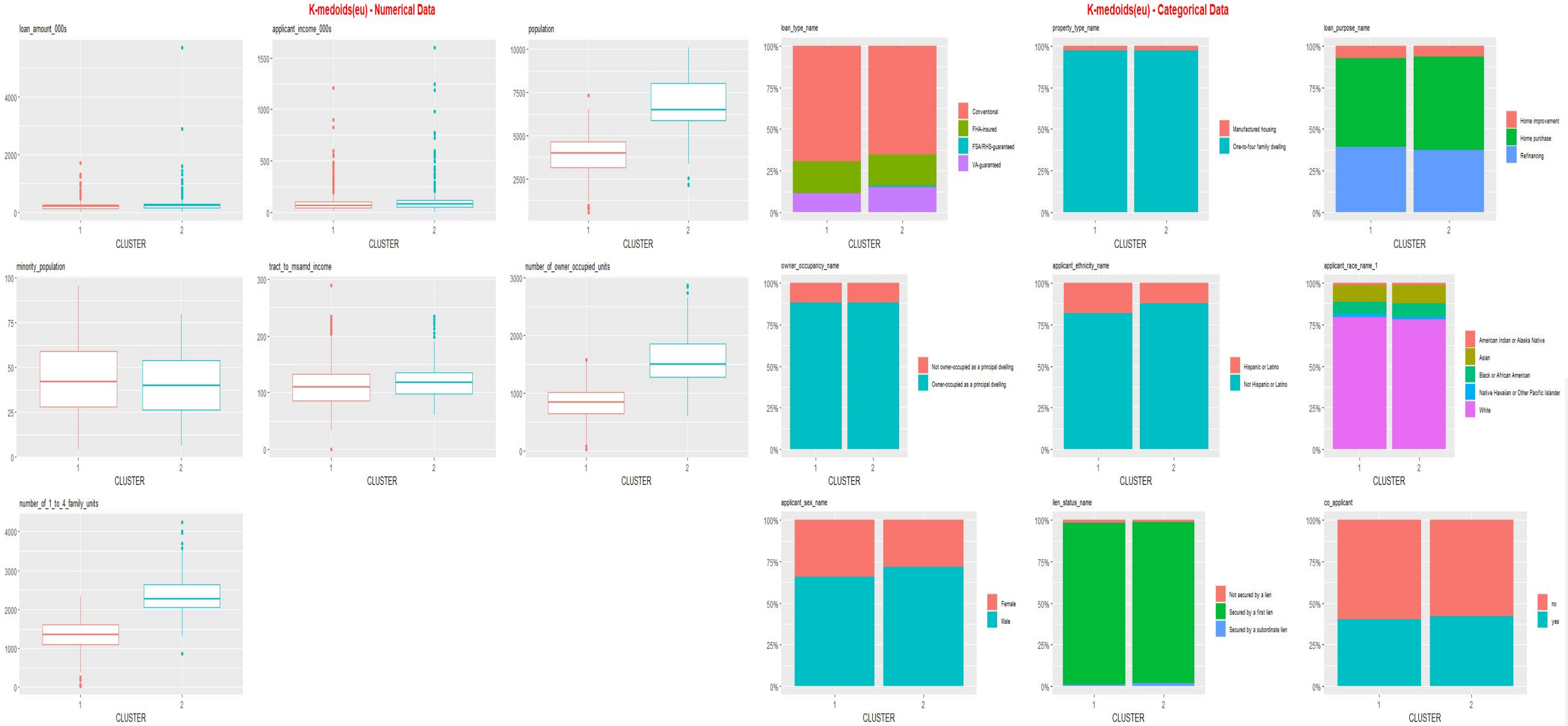
CLUSTERING ANALYSIS – K MEDOIDS (GD)





- No significant difference can be seen from the **gower distance model**

CLUSTERING ANALYSIS – K MEDOIDS (EU)



- No significant difference can be seen from the **categorical variables**

CLUSTER INTERPRETATION – K MEDOIDS (EU)

CLUSTER 1 - SMALL TOWN, LESS NUCLEAR FAMILY

- Less population of city/town/area
- Less number of owner occupied units
- Less number of family of 1 to 4 units

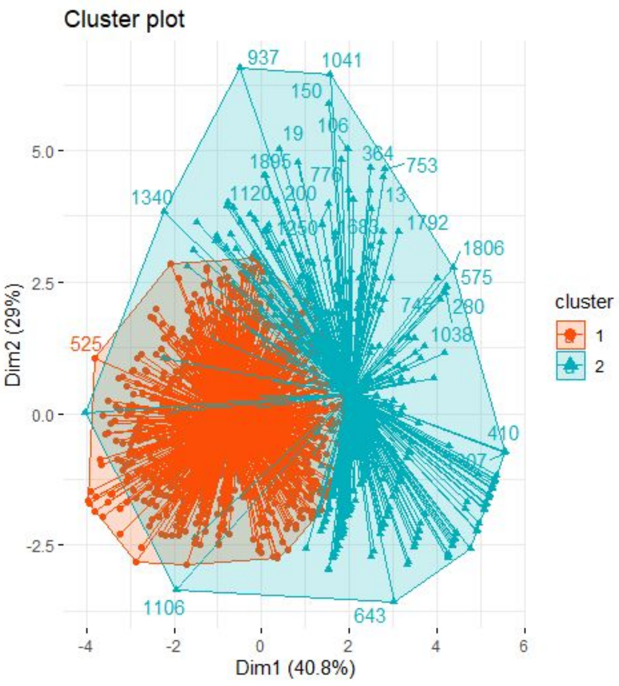
CLUSTER 2 - LARGER TOWN, MORE NUCLEAR FAMILY

- More population of city/town/area
- More number of owner occupied units
- More number of family of 1 to 4 units

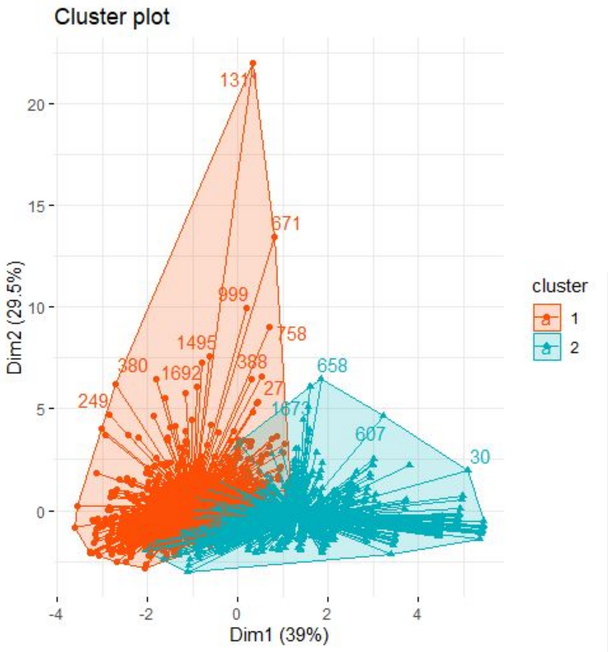
➢ No significant difference can be seen from the **categorical variables**

CLUSTERING ANALYSIS – QUALITY OF CLUSTERS

CLUSTER PLOTS - NUMERICAL VARIABLES ONLY



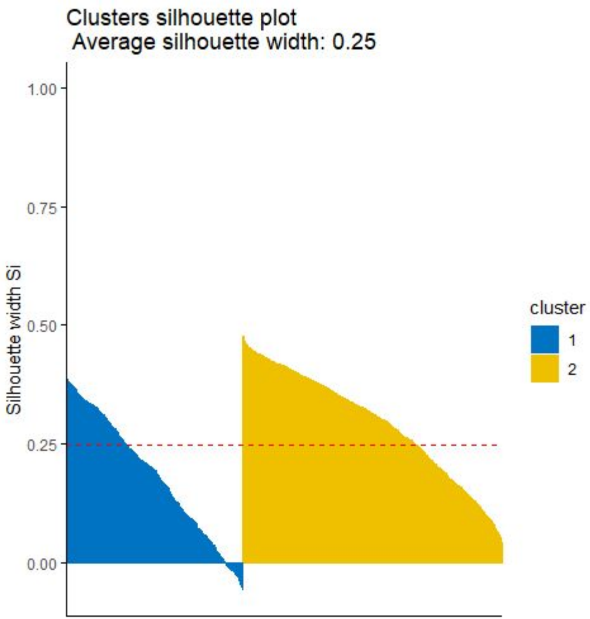
k means



k medoids

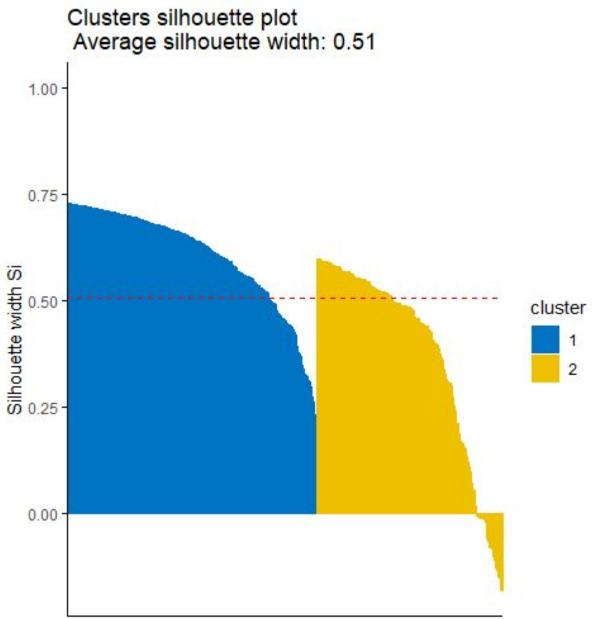
K medoids have better performance but not the best clustering because of the overlap

CLUSTERING ANALYSIS – QUALITY OF CLUSTERS



	cluster	size	ave.sil.width
1	1	735	0.18
2	2	1080	0.29

k means



	cluster	size	ave.sil.width
1	1	1148	0.60
2	2	852	0.37

k medoids

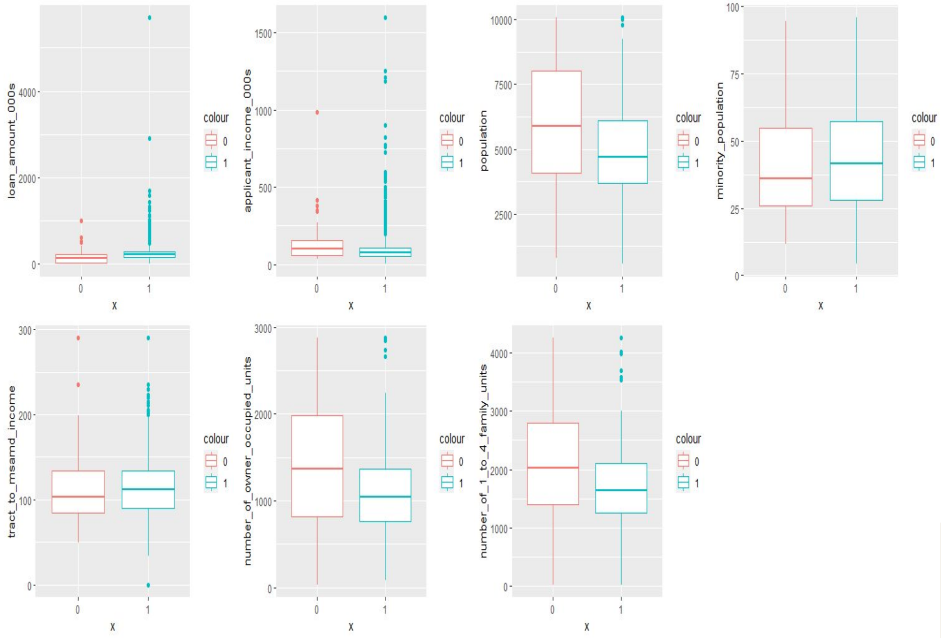
The average of silhouette width also suggests that K medoids performs better than k means because it measures how similar a data point is to its own cluster compared to the other clusters.
(This is also known as cohesion and separation)

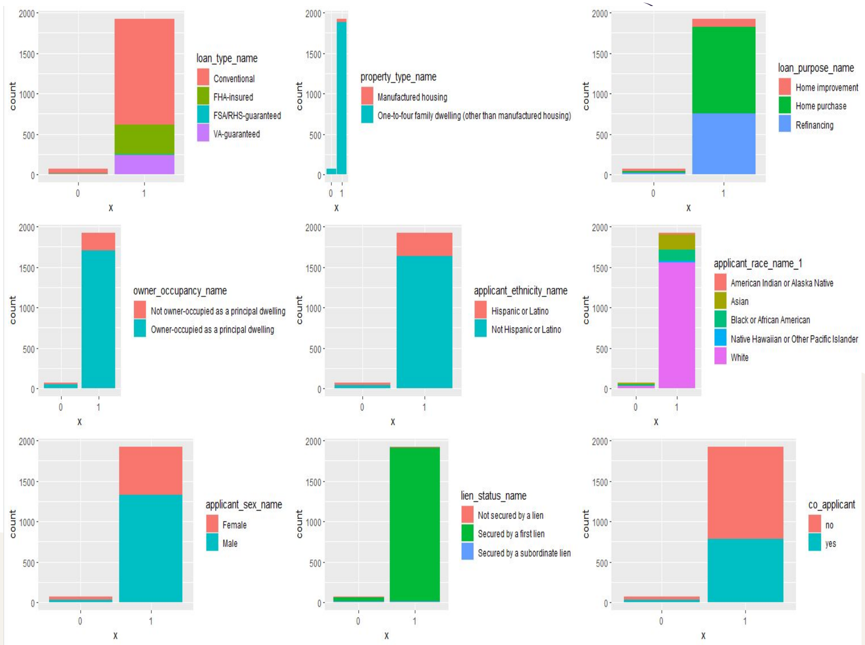
Next, hierarchical clustering with density-based clustering are implemented for comparison because numerical variables dominated in k means and k medoids clustering.

DENSITY-BASED CLUSTERING - OUTLIER DETECTION

- K-Means and K-Medoids methods are sensitive towards noise and outliers
- To address this concern we can apply DBSCAN algorithm to separate noise and outliers.
- We use KNN displot to identify the elbow and select appropriate epsilon and we opted for minpts to be 50
- We use the Gower Distance Dissimilarity matrix as the input for the algorithm and table the classification against (of numerical only) mahalanobis outliers.

DENSITY-BASED CLUSTERING - OUTLIER DETECTION





DENSITY-BASED CLUSTERING - OUTLIER DETECTION

```
table(df_sample$action_taken_name, df_main$db)

     0    1
0   42  636
1   33 1289
```

As we can observe from the boxplots above, outliers (DB=0) have relatively lower loan amount and higher income.

Meanwhile the frequency table suggests that outliers are more likely to be rejected applications (42/75) rather than (636/1925), which contradicts to the common sense that higher income and lower loan, more likely the application is approved. But this maybe due to other factors such as lower return on investment or credit issues .

Thus, removing the DBSCAN outliers can reduce the impact of the corner cases. This will also help better clustering output reducing the effect of outliers.

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING

Linkage (sw score)	k = 2	k = 3	k = 4	k = 5	k = 6	k = 7	k = 8
single	0.1793	0.1078	0.0382	-0.0616	-0.0837	-0.1122	-0.1396
complete	0.1883	0.1593	0.1349	0.1413	0.1162	0.1266	0.1269
average	0.1793	0.0748	0.1067	0.0808	0.0817	0.1360	0.1477
ward.d2	0.1396	0.1613	0.1832	0.1783	0.1643	0.1503	0.1547

Complete outperforms single and average for each k, while k = 2 gives complete linkages the maximized silhouette width.

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING

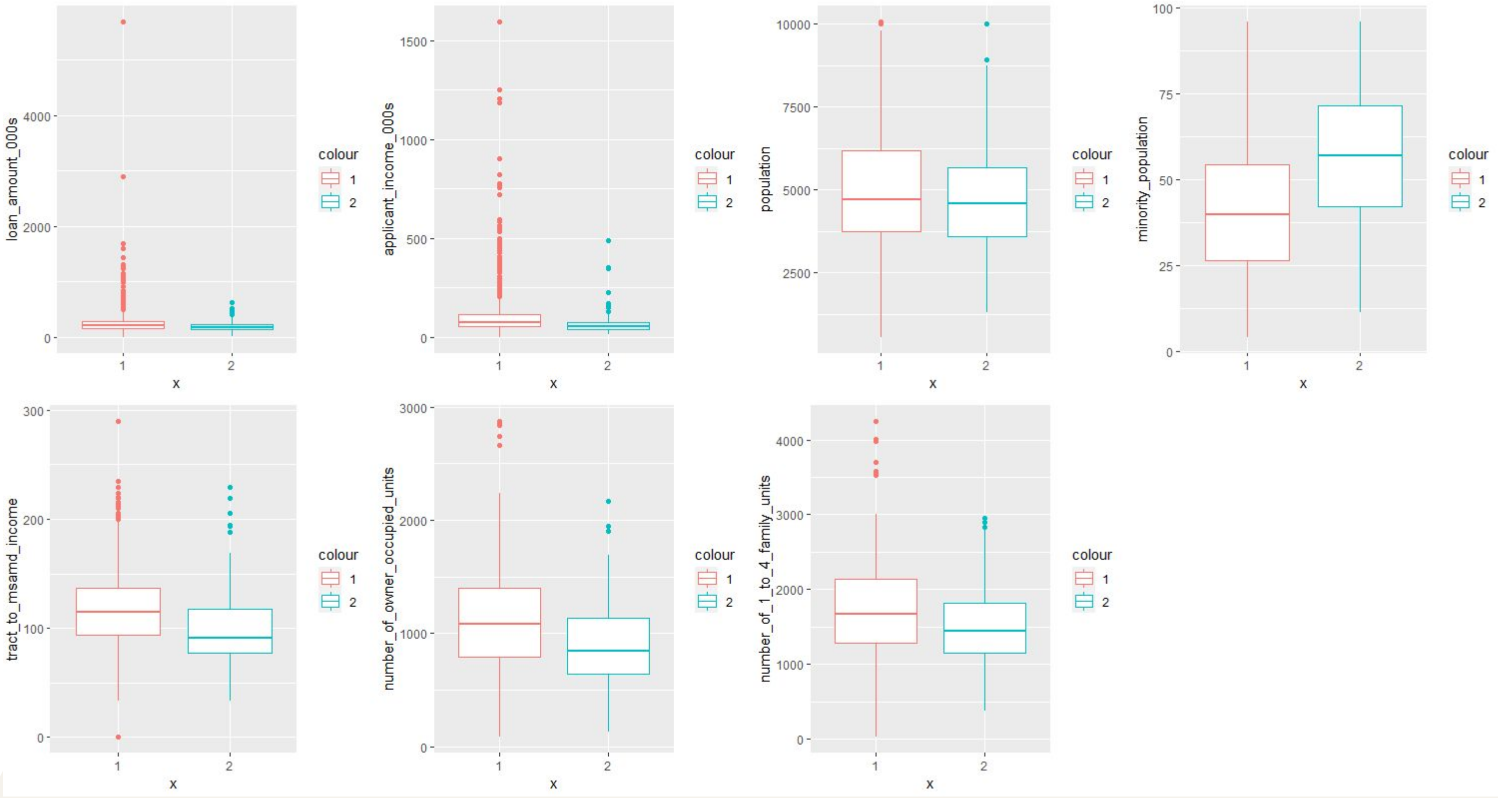
Let’s start with k = 2, complete linkages. Cluster 2 has a higher chance to get an approval .

```
> table(group = df_work$action_taken_name, cluster = df_work$hc2)
     cluster
group    1    2
    0  572   64
    1 1102  187
```

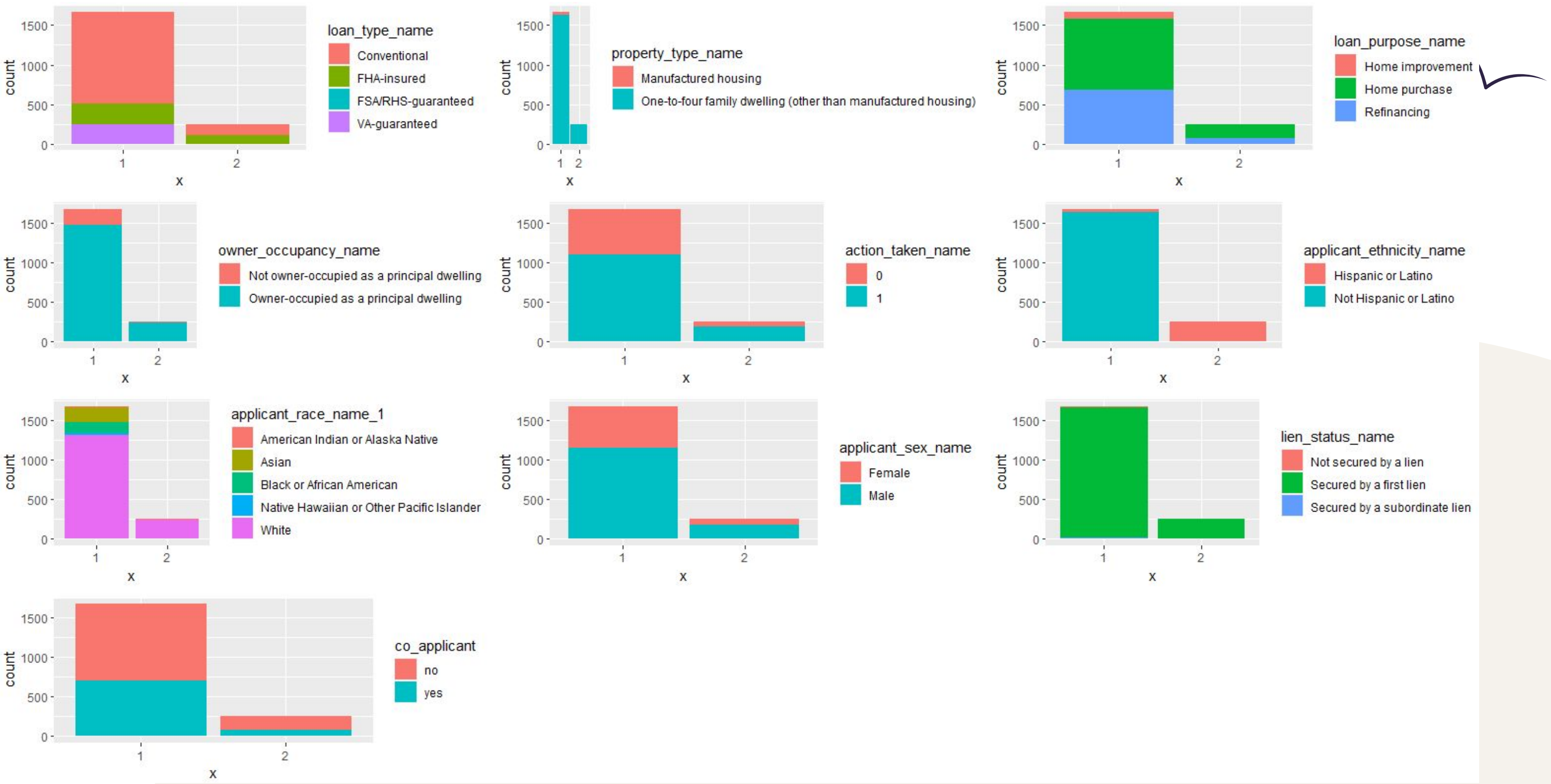
From the boxplots and frequency plots in the next two slides, we can tell that cluster 2 has a lower loan amount (in terms of average and extreme values) and a lower level of variety in terms of loan amount and income.

Also, cluster 2 is more likely to be caucasian, or mortgage for primary residence, or the mortgage is secured.

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING (K=2)



CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING (K=2)



CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING (K=2)

CLUSTER 1 - DIVERSIFIED RACE, MORE INVESTMENT MORTGAGE

- More racial diversity
- More likely to be an investment loan
- More cases not secured by a lien
- More income/loan amount variation

CLUSTER 2 - MONO-RACE, MORE PRINCIPAL RESIDENCE MORTGAGE

- Less racial diversity, mostly white and hispanic
- More likely to be a principal residence
- More cases secured by a first lien or a second lien
- Less income/loan amount variation

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING EXPAND K FROM 2 TO 3

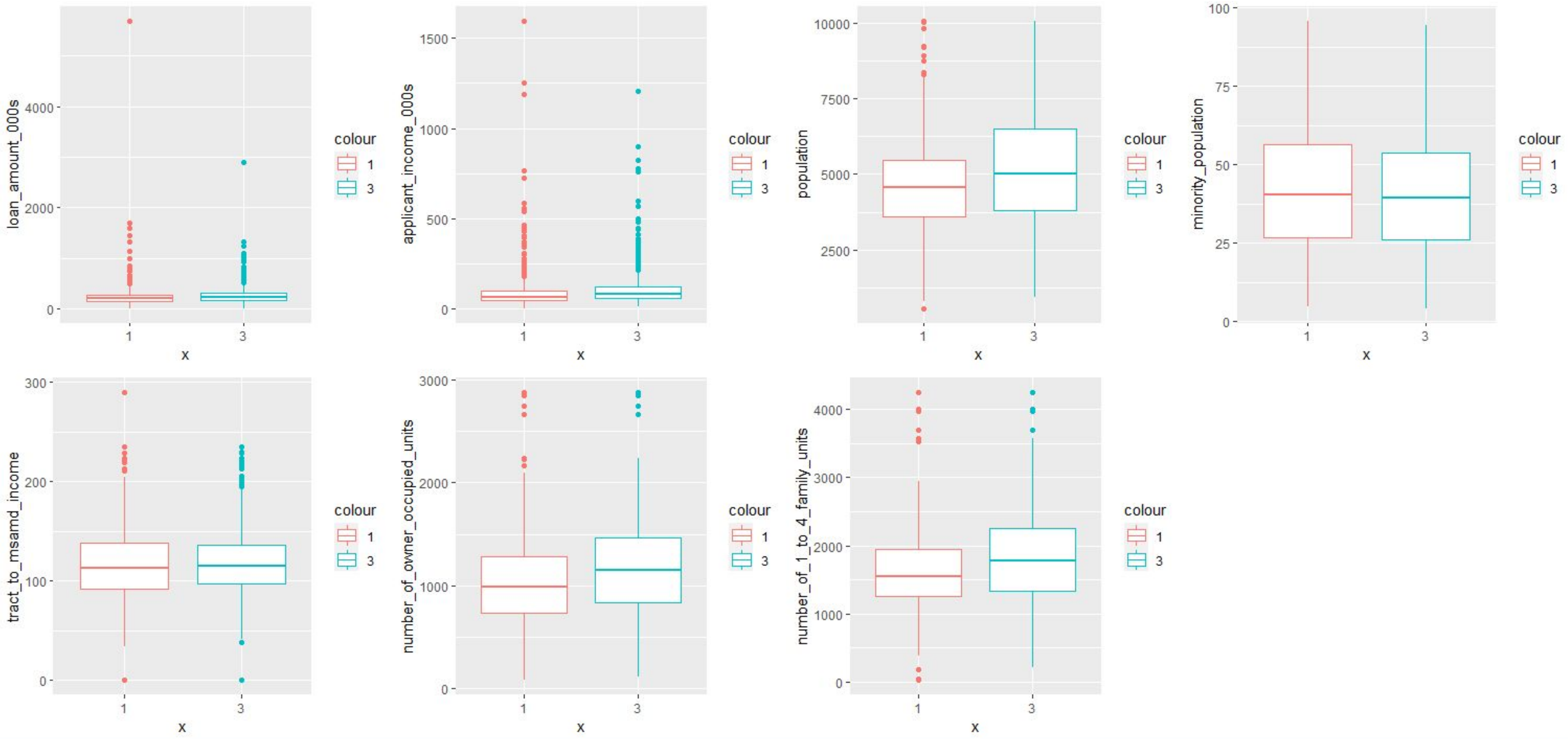
Expand k to 3, and the extra clustering spilt cluster 1 into two clusters. Cluster 3 is more likely to get an approval.

```
> table(df_work$hc2, df_work$hc3)

      1   2   3
  1 684   0 990
  2   0 251   0
```

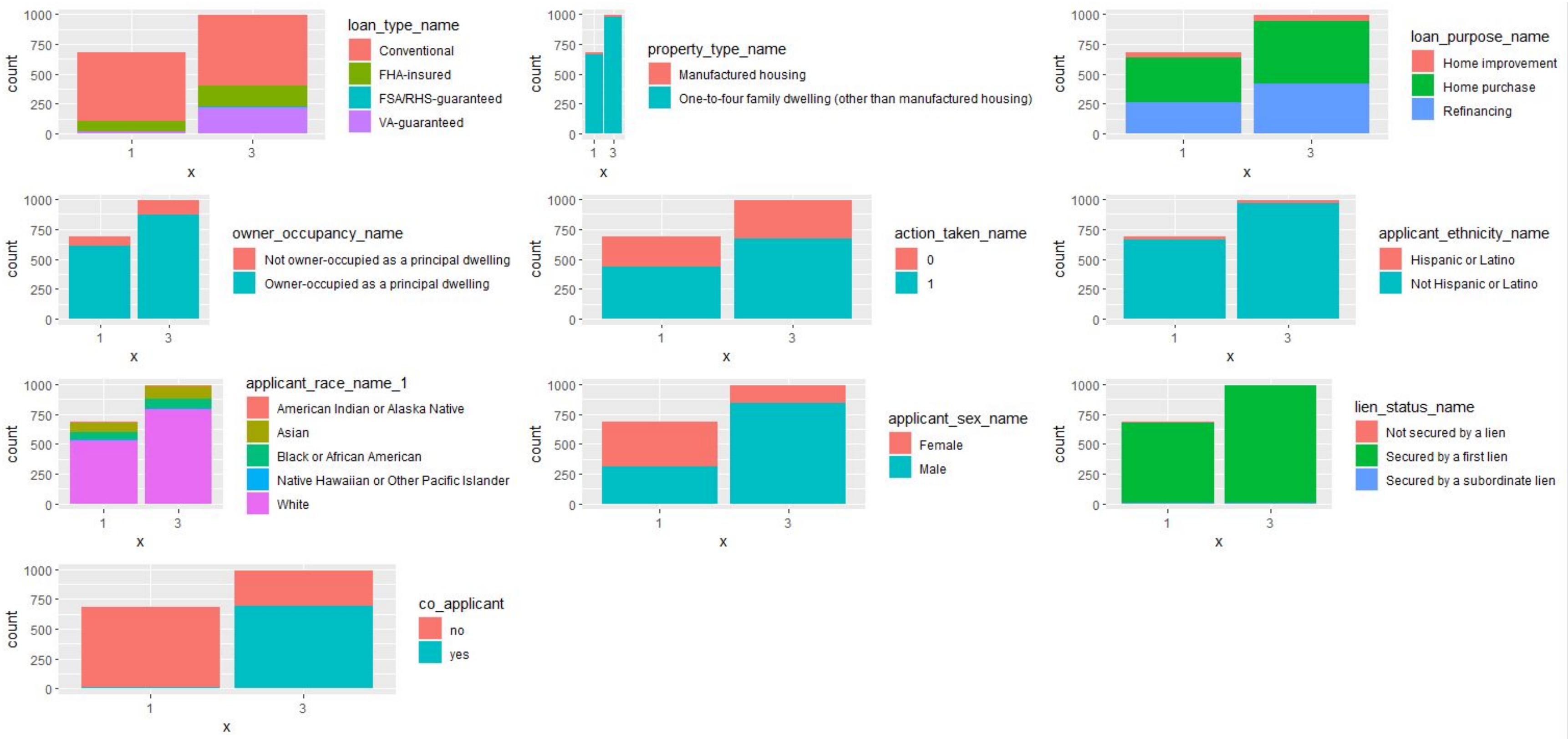
```
> table(group = df_work_2$action_taken_name, cluster = df_work_2$hc3)
     cluster
group   1   2   3
    0 252   0 320
    1 432   0 670
```

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING EXPAND K FROM 2 TO 3



CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING

EXPAND K FROM 2 TO 3



CLUSTERING ANALYSIS – DENSITY-BASED CLUSTERING EXPAND K FROM 2 TO 3

CLUSTER 1 - MORE CONVENTIONAL MORTGAGE, NO CO-APPLICANT, LESS MALE

- More Conventional mortgage
- Less likely to have a co-applicant
- Less likely to be a man

CLUSTER 3 - MORE INCENTIVE PROGRAM INVOLVED, CO-APPLICANT, MORE MALE

- More VA-Guaranteed cases (veteran related)
- More likely to have a co-applicant
- More likely to be a man

CLUSTERING ANALYSIS – HIERARCHICAL CLUSTERING EXPAND K FROM 2 TO 3 AND MORE

```
> tree$variable.importance
                co_applicant                 loan_type_name            applicant_sex_name
                   369.84433                      115.23462                      83.18347
                  population          applicant_income_000s              loan_amount_000s
                    52.67716                       52.64236                      39.95751
           loan_purpose_name  number_of_1_to_4_family_units number_of_owner_occupied_units
                    38.62509                       37.98820                      31.24320
          minority_population           applicant_race_name_1          owner_occupancy_name
                    25.96438                       21.65598                      16.50469
         tract_to_msamd_income       applicant_ethnicity_name
                    14.10952                        7.76734
>
```

co_applicant and loan type plays the most important role in hierarchical clustering when k expands from 2 to 3, which coincides with our observations.

Sequentially, loan type helps the most splitting the cluster when k expands from 3 to 4, co_applicant again splits the cluster when k expands from 4 to 5, and applicant’s gender and race helps splitting the cluster when k expands from 5 to 6.

PREDICTION OF MORTGAGE APPLICATION

Can we predict the outcome of mortgage application by the characteristics of the applicant?

```
> ## eu ##
> (km_eutab <- as.matrix(table(km_eu$cluster,df_action_noout)))
   df_action_noout
      0   1
  1 472 912
  2 139 292
> d <- diag(1, nrow(km_eutab))
> (adjusted_km_eutab <- pMatrix.min(km_eutab,d))
   df_action_noout
      0   1
  2 139 292
  1 472 912
>
> #get accuracy
> sum(diag(adjusted_km_eutab))/sum(adjusted_km_eutab)
[1] 0.5790634
```

K MEANS

Accuracy: 58%

```
> ## eu ##
> (kmed_eutab <- as.matrix(table(kmed_eu$cluster,df_sample$action_taken_name)))

     0   1
  1 438 813
  2 240 509
> d <- diag(1, nrow(kmed_eutab))
> (adjusted_kmed_eutab <- pMatrix.min(kmed_eutab,d))

     0   1
  2 240 509
  1 438 813
>
> sum(diag(adjusted_kmed_eutab))/sum(adjusted_kmed_eutab)
[1] 0.5265
```

K MEDOIDS

Accuracy: 53%

HIERARCHICAL CLUSTERING

Accuracy: 59%

```
> table(group = df_work$action_taken_name, cluster = df_work$hc2)
     cluster
group    1    2
    0  572   64
    1 1102  187
```

CONCLUSION - SUMMARY

PARTITIONING METHODS - K MEANS AND K MEDOIDS

- More significant differences in the **numeric variables** between the clusters
- Both K means and K medoids methods shows **less population of city/town/area, less number of owner occupied units, L=less number of family of 1 to 4 units** for **cluster 1** and **more population of city/town/area, more number of owner occupied units, more number of family of 1 to 4 units** for **cluster 2**.

HIERARCHICAL CLUSTERING

- More significant differences in the **categorical variables** between the clusters
- Hierarchical clustering shows **more racial diversity, more likely to be an investment loan, more cases not secured by a lien, more income/loan amount variation** for **cluster 1** and **less racial diversity, mostly white and hispanic, more likely to be a principal residence, more cases secured by a first lien or a second lien, less income/loan amount variation** for **cluster 2.**

- co_applicant and loan type plays the **most important role** in hierarchical clustering
- Hierarchical clustering shows

FURTHER RESEARCH AND IMPROVEMENT

How to improve your clustering analysis in future?

- For the mortgage specialists, they probably look into qualitative criteria before evaluating the applicant's ability to repay, such as loan type, principal residence, and other criteria such as gender, race.
  We could apply the hierarchical clustering until all qualitative variables are well-split, then apply partition clustering within each cluster. In this way, we might see different selection criteria for repayment ability for different types of mortgage.

- The clustering classification was implemented with 2,000 sampled data because of resource issue. The randomness and small volume of data might have some impact on the quality of clusters.

- Some variables, such as lien security, property type, are not balanced?. We can remove those variables for better clustering results.

QNA