Clustering Analysis of Home Loan Applicants in the State of Nevada

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

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

Introduction

According to the research conducted by the U.S. Department of Housing and Urban Development in 2016, Nevada was the worst state for affordable housing in the United States. As a result, we would like to look more into the housing issue and gather informative insights from our analysis. The data set that was used for this analysis is the HDMA Nevada State Home Loans, 2017 dataset. This data set contains information regarding demographic, area specific data, loan status, property type, loan type of the mortgagor in the state of Nevada in 2017.

The overview of this analysis consists of implementation of the clustering analysis including: k-means, k-medoid, hierarchical and density-based clustering method to find the characteristics of the individuals who applied for the mortgage in Nevada state. The target audience of this report is for those who are 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.

Data Preprocessing

The original data contains 178,587 rows and 79 columns (variables) 5,201,754 null values. The data were checked for any existing duplicates and 796 duplicate rows were removed. Variables not relevant to our analysis have been removed as many variables are coupled with a column consisting of its corresponding numerical codes. Then, a binary variable is transformed and created from the ‘action_taken_name’ to indicate whether co-applicant exist. Missing values were checked both column wise and row wise and only complete cases were kept. Finally, a random sample was taken from the cleaned for our analysis. The cleaned, sampled data contains 2,000 rows and 16 columns (variables). The Variables Table in the appendix shows the variables that were kept for our clustering analysis. The Variable Description Table in the appendix shows the description of variables with less intuitive variable names.

K-means Clustering

Using the cleaned, randomly sampled dataset, the numerical outliers were removed by using the mahalanobis method. Then, two dissimilarity matrices were calculated using the Gower’s distance and the euclidean distance for comparison. The optimal number of centroids(k) to apply to the k-means clustering were then achieved by using both the average silhouette score and the Calinski and Harabasz Score. The table below shows the result for the optimal number of k to use in the k-means analysis for the dissimilarity matrices calculated using the Gower’s distance and the euclidean distance.

Optimal # of k	Gower	Euclidean
Avg. silhouette score	2	2
CH score	2	2

Gower's

By applying the k-means clustering method on the gower's distance dissimilarity matrices using 2 centroids, we can see from the graphs that there are not any obvious differences between the clusters for k-means clustering using the gower's distance dissimilarity matrices. (Please refer to the appendix for the corresponding graphs.)

Euclidean

By applying the k-means clustering method on the euclidean distance dissimilarity matrices using 2 centroids, we can see from the graphs that there are some obvious differences in the numerical variables between the two clusters for k-means clustering using the euclidean distance dissimilarity matrices. (Please refer to the appendix for the corresponding graphs.)

Remarks

Looking at the k-means clustering that uses the euclidean distance dissimilarity matrices, we can see the following.
Cluster 1 contains the following characteristics for the loan applicants:

- 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

Thus, it appears that loan applicants for cluster one live in areas with a smaller overall population, a higher proportion of minorities and less nuclear families and the opposite applies for those in cluster 2.

K-medoid Clustering

For K-medoid clustering, outliers were kept in the data-set for calculating the dissimilarity matrices using the Gower's distance and the euclidean distance. The procedure for calculating the optimal number of centroids(k) is the same as that of K-means clustering except Calinski and Harabasz Score were not considered. The table below shows the result for the optimal number of k to use.

Optimal # of k	Gower	Euclidean
Avg. silhouette score	5	2

Gower's

By applying the k-medoid clustering method on the gower's distance dissimilarity matrices using 5 centroids, we can not see any obvious differences between the clusters from the graphs. (Please refer to the appendix for the corresponding graphs.)

Euclidean

By applying the k-medoid clustering method on the euclidean distance dissimilarity matrices using 2 centroids, we can see some obvious differences between the clusters for the numerical variables by looking at the graphs. (Please refer to the appendix for the corresponding graphs.)

Remarks

Looking at the k-medoid clustering that used the euclidean distance dissimilarity matrices, we can see the following.
Cluster 1 contains the following characteristics for the loan applicants:

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

Thus, it appears that loan applicants for cluster one live in areas with a smaller overall population and less nuclear families and the opposite applies for those in cluster 2.

K-means vs K-medoid

Numerical Variables

By applying dimension reduction to the numerical variables and projecting them to a 2-D plane, The figures represent the first two principal components. We can see that K-medoid performs better than that of K-means as there is less of an overlap between the 2 clusters. (Please refer to the appendix for the corresponding visualization and results.)

All Variables

By looking at the average silhouette width for both of the clustering methods, it can also be seen that K medoids perform better than k means because the average silhouette width measures how similar a data point is to its own cluster compared to the other clusters. (Please refer to the appendix for the corresponding visualization and results.)

Db-scan Analysis

The earlier clustering methods such as K-means and K-Medoids are sensitive to outliers as they use some form of euclidean distance metric and we recommend using DBSCAN to identify outliers and then proceed to perform Hierarchical Clustering to identify the different segments within the data. We use KNN displot to identify the elbow and select appropriate epsilon and we opted for minpts to be 50 after trial and error as we felt the separation was prominent at minpts=50.We use the Gower Distance Dissimilarity matrix as the input for the algorithm and table the classification against (of numerical only) mahalanobis outliers

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

DBSCAN clustering for 2000 objects.
Parameters: eps = 0.11, minPts = 50
The clustering contains 1 cluster(s) and 75 noise points.

0 (outlier Clusters)	1 (Non Outlier Clusters)
75	1925

We apply Mahalanobis method to identify outliers and we tabulate the comparisons

	Mahalanobis Outliers		
DBSCAN Outliers		0	1
	0	7	68
	1	78	1847

Outliers on average have higher applicant income than the non outliers and in general applied for higher loan amounts. The outliers mostly had on average higher income than the residents of the same area. They also mostly applied for mortgages in low minority population areas and in general applied to fund a home in highly populated areas.

We can visually identify the clusters and noise points more prominently in DBSCAN while Mahalnobis only considers numerical features.Categorical features are essential to the context of noise points. This dataset can particularly aid in identifying if there exists the social issues of discriminatory lending practices for people of color.

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

Hierarchical Clustering

For hierarchical clustering, we used the cleaned, randomly sampled data. The outliers are detected using the DBSCAN method on the Gower distance matrix. Unlike the outlier removal in K-mean or K-medoid, implementing the mahalanobis method to detect numerical outliers, the DBSCAN method takes the advantage of including categorical variables into consideration.

Optimal Linkage Method

As for the optimal linkage method among single, average, complete, ward.D2, we calculated and compared the avg. silhouette score for each linkage method and the number of clustering from 2 to 10.

Max (optimal k)	Single	Complete	Average	Ward.D2
Avg silhouette	0.1793(k=2)	0.1883(k=2)	0.1793(k=2)	0.1832(k=4)

Since the complete method at k = 2 produces the highest level of average silhouette score, we chose the complete method as our optimal linkage, k = 2 as the initial clustering number.

Remarks

As we can observe from the by-group boxplots and frequency table (see appendix H-1), we can summarize the following:
Cluster 1 (1674 applicants out of 1925) contains the following characteristics for the loan applicants:

1. Higher in average in loan amount, applicant income, and relative income compared with MSA/MD (local) median income
2. Located at a more owner-occupied, more 1-4 family unit housing neighborhood
3. Wider spread of all numerical variables
4. Most of the VA-guaranteed cases (veteran related)
5. More conventional loans, more refinancing cases, but less applying for primary residence
6. More racially diverse, while cluster 2 contains most of Hispanic and Latino applicants

Meanwhile, when comparing the clustering result with the actual application approval group, we can see that applicants cluster 1 are less likely to have their loan applications approved.

	cluster(complete, k = 2)	
Group (action_taken_name)	1	2
0	0.3416965	0.2549801
1	0.6583035	0.7450199

In addition, as we looked into the key variables dividing the data into two clusters with corresponding methods, we found out that “applicant_ethnicity_name” (Hispanic and Latino or not) is the dominant variable among all important factors, which coincides with our observation of that cluster 2 contains most of Hispanic and Latino applicants.

tree12$variable.importance		
applicant_ethnicity_name	loan_type_name	applicant_sex_name
355.8157122	25.7973701	24.4063492

Combining this with the fact that applicants in cluster 1 are less likely to have their loan application approved, an unverified conclusion could be made that Latino or Hispanic applicants are more favored in terms of loan application in Nevada.
However, based on many other features represented in cluster 1 such as loan type (VA-guaranteed is more favored than conventional loan), applicant's race, as well as higher income/loan amount, which may refute the conclusion, we decided to increase the cluster number to 3 or even more to explore any potential results from hierarchical analysis.

Hierarchical Analysis Extension

When we increased the cluster number to three, we found out that the cluster 1 in the previous clustering is split into two groups (684 and 990 applicants respectively from cluster 1 and cluster 3), while cluster 3 is more likely to be approved than cluster 1.

	cluster(complete, k = 3)	
Group (action_taken_name)	1	3
0	0.3684211	0.3232323
1	0.6315789	0.6767677

Extended Remarks

Similarly, we can summarize the following by observing from the by-group boxplots and frequency table (see appendix H-2):
Cluster 3 (990 applicants out of 1674) contains the following characteristics for the loan applicants, when compared with cluster 1 (684 applicants out of 1674):

1. Higher in average in loan amount, income and relative income
2. Applications located at a more-populated, more owner-occupied and more 1-4 family unit housing area
3. Contains most of the VA-guaranteed cases
4. More likely to be a male
5. More likely to have a co-applicant

In addition, when we look at the most important variables dividing cluster 1 from the previous step, there are several key factors (co applicant, loan type, applicant's gender, population, income, loan amount) which coincide with our observations from the graphs.

tree23$variable.importance			
co_applicant	loan_type_name	applicant_sex_name	population
369.84433	115.23462	83.18347	52.67716
applicant_income	loan_amount	loan_purpose_name	number_of_1_to_4_family
52.64236	39.95751	38.62509	37.98820

Labeling clusters

Since having a co-applicant means more capability to repay the loan, as there is more income backing the mortgage, we may consider having a co-applicant as more income for the application. Then, combining with the decision trees generated from the two clusterings (see appendix H-3), we can describe cluster 2 as Hispanic or Latino applicant with relatively lower income, cluster 1 as non-Hispanic or Latino applicant with lower income and applying to less populated areas, and cluster 3 as non-Hispanic or Latino applicant with higher income and applying to more populated areas. The following table shows the application approval rate for each cluster.

	cluster		
Group (action_taken_name)	1	2	3
0	0.3684211	0.2549801	0.3232323
1	0.6315789	0.7450199	0.6767677

Comparison and Conclusion

Moreover, we extended the number of clustering further to 6, to check if there are any other key factors when splitting the clusters. The key factors for splitting three clusters into four are loan type and loan purpose, the key factor for splitting four clusters into five is co applicant, and the key factors for splitting five clusters into six are gender and race.
Unlike the partition clustering methods, the hierarchical clusters are dominantly formed by categorical variables (despite having a co-applicant can be considered as more income for the application, and can be quantified if there is information about the co-applicant's income). It may make more sense, as the selection criteria for different types of loan may vary a lot. In addition,

the clustering result is more likely to be linked with income level (having a co-applicant), rather than the partition clustering.

Conclusion & Discussion

Cluster Interpretation

According to our partitional clustering, the clusters are mainly formed by numerical variables. Due to the limitations of K means and K medoids methods we implemented, no significant differences could be seen from the categorical variables. Therefore, the results and interpretations are focused on numeric variables such as the population of the area, the number of owner occupied units, and the number of families of 1 to 4 units.
We also explored hierarchical clustering in order to mitigate this issue and to see if there is any difference in hierarchical clustering methods. Four linkage methods are compared to produce the best performance output. According to the characteristics of each cluster in hierarchical clustering, the clusters are dominantly formed by categorical variables such as racial diversity, loan types, the likelihood of a case secured by a lien, income/loan amount variation, mortgage type, co-applicant status, and gender.
Top three hierarchical clusters can be labeled as:

1. cluster 1: non-Hispanic or Latino applicant with lower income and applying to less populated areas
2. cluster 2: Hispanic or Latino applicant with relatively lower income
3. cluster 3: non-Hispanic or Latino applicant with higher income and applying to more populated areas

Implication

For the mortgage specialists, they can 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, age.. 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.

Improvement and Future research

In terms of data quality improvement, we can have the income-loan amount ratio as one of the variables to consider, as the ratio may reflect the repayment ability rather than the absolute income level or loan amount. In addition, we can quantify the co-applicant impact of the mortgage by adding the co-applicants' income to the total applicant’s income.
In terms of future research, dimension reduction techniques can be applied for the clustering analysis to allow for more variables to be used. We used 2,000 random samples that could be biased in the results and interpretations. Further studies may use larger sample sizes to reduce possible biases in the data, and the work flow approach will be altered by implementing

DBSCAN first to detect outliers then proceed to perform hierarchical clustering to identify clusters in clusters and subsequently apply K-Means or K-Medoids to separate within cluster in segments based on continuous features to see the probability of success in receiving a mortgage within group.

Acknowledgement

This research was conducted by using data provided by the Consumer Financial Protection Bureau. This research would not have been possible without their initial research and data. We also thank all our colleagues who provided insight that greatly assisted the research, although they may not agree with all of the interpretations of the report.

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

Appendix

Variables 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'

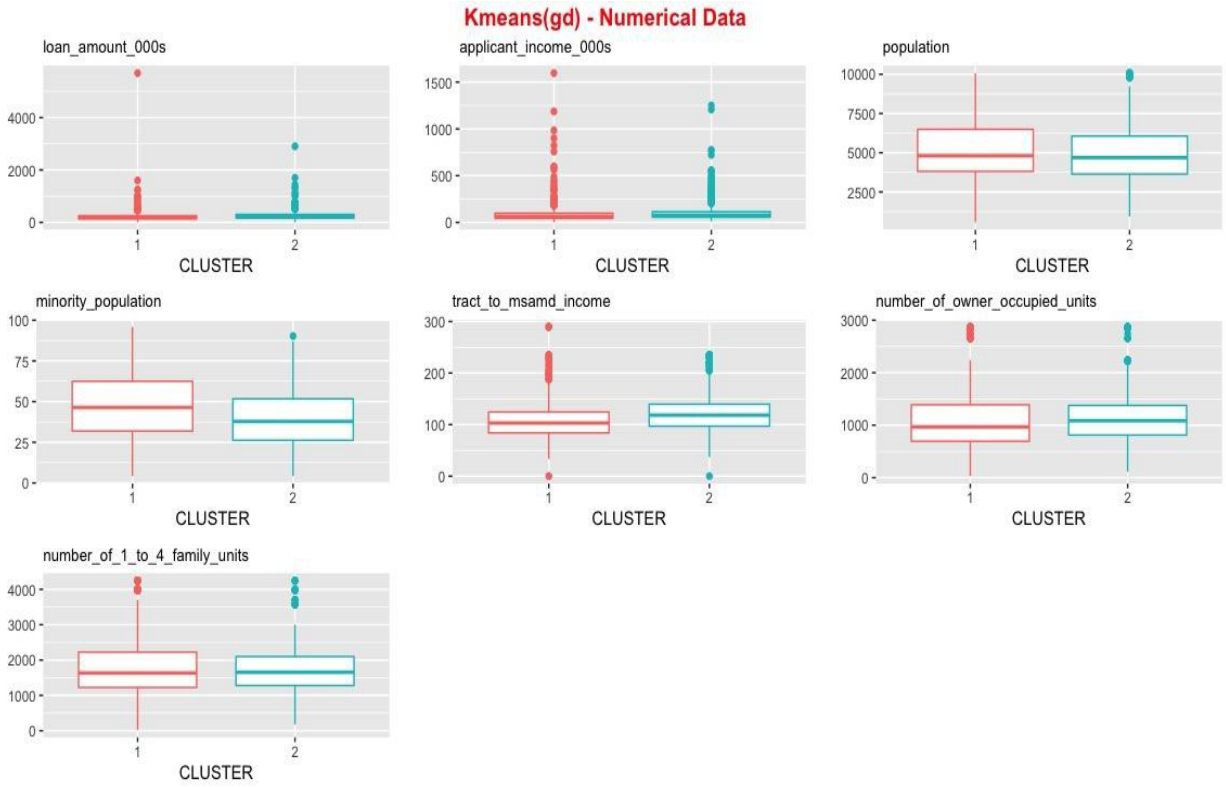
Variable Description Table

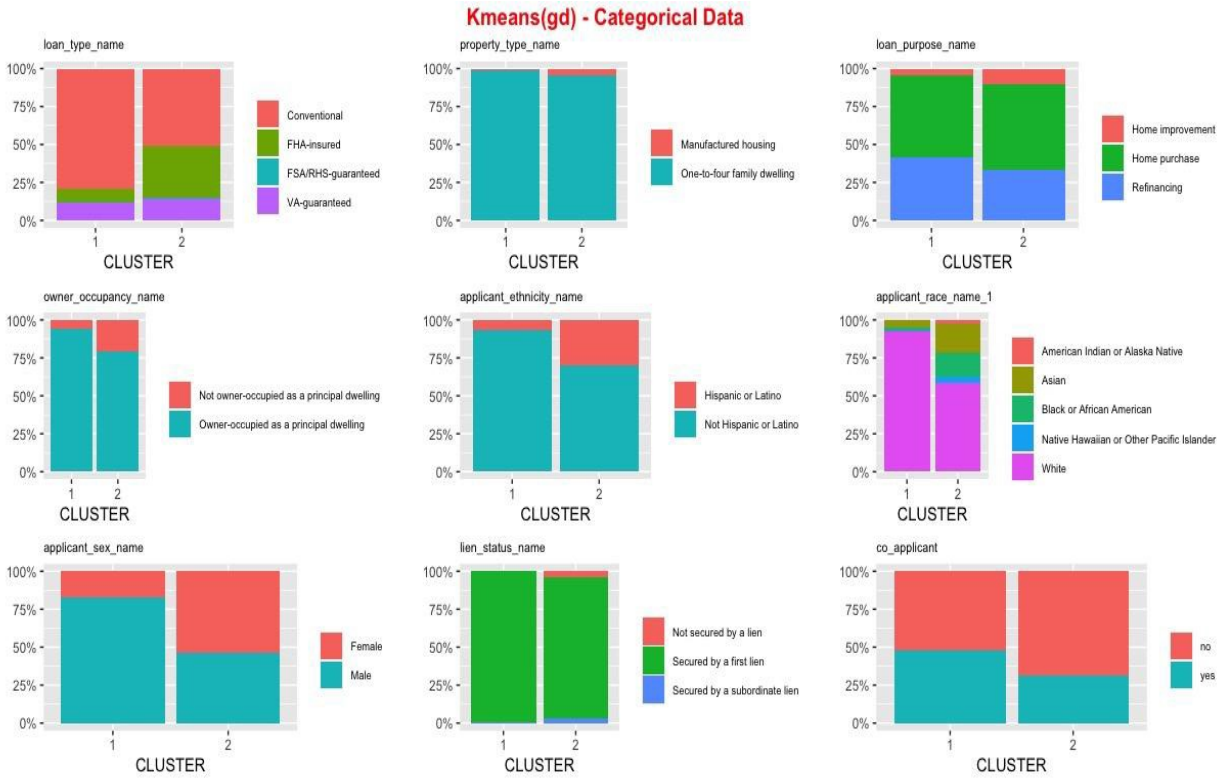
Variable	**Description**
minority_population	percentage of minority population to total population in the applicant's area
tract_to_msamd_income	percentage of tract median family income compared to MSA/MD median family income in the applicant's area
number _of_owner_occupied_units	Number of dwellings, including individual condominiums, that are lived in by the owner in the applicant's area
number_of_1_to_4_family_units	Dwellings that are built to house fewer than 5 families in the applicant's area

co_applicant	Whether applicant has a co-applicant

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

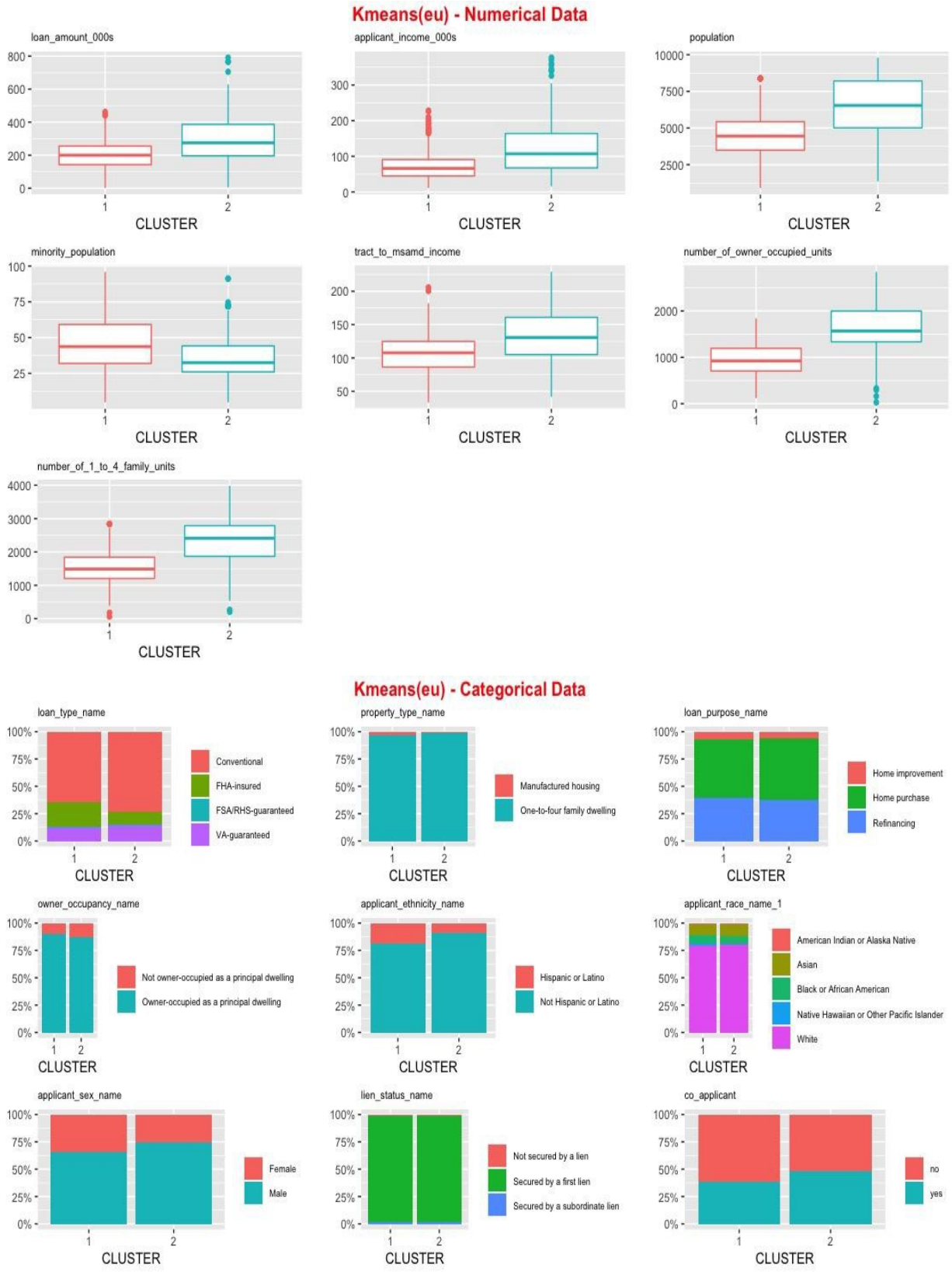
K-means (Gower) results





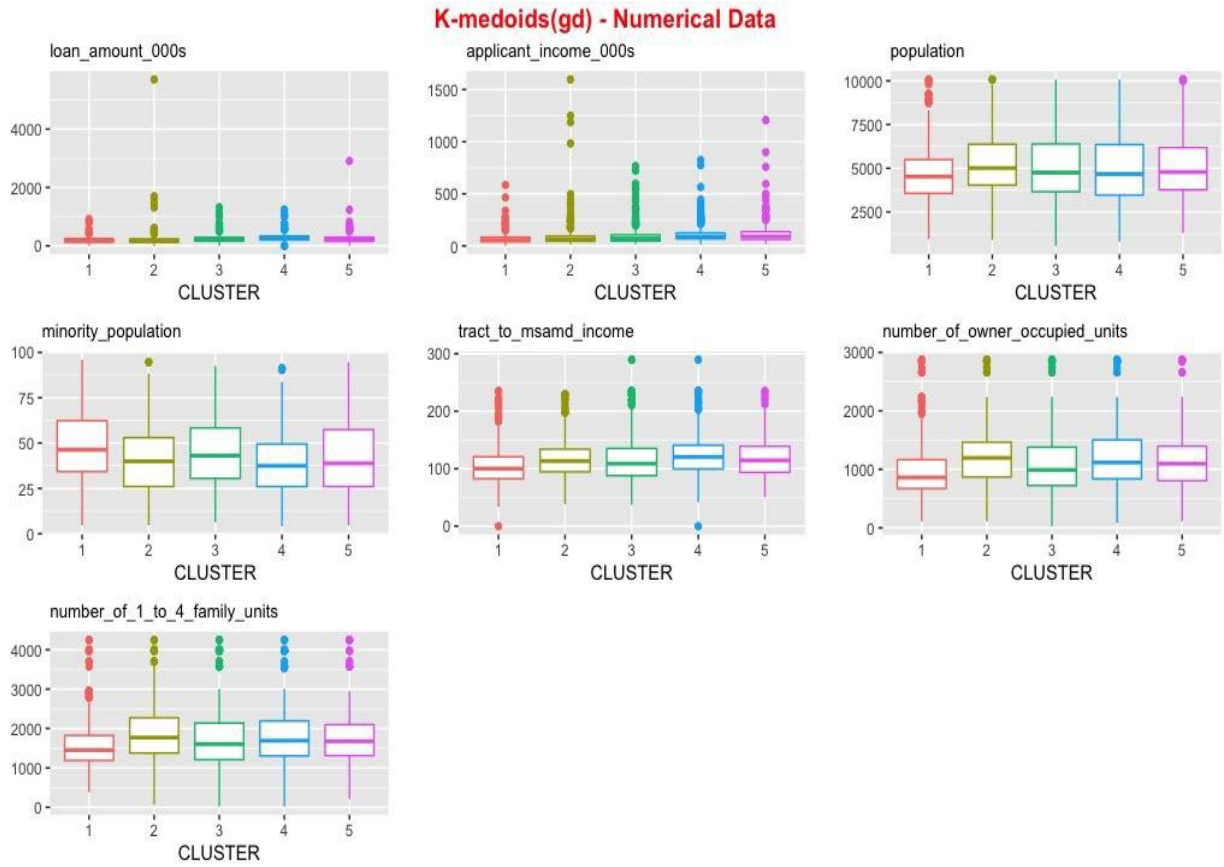
Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

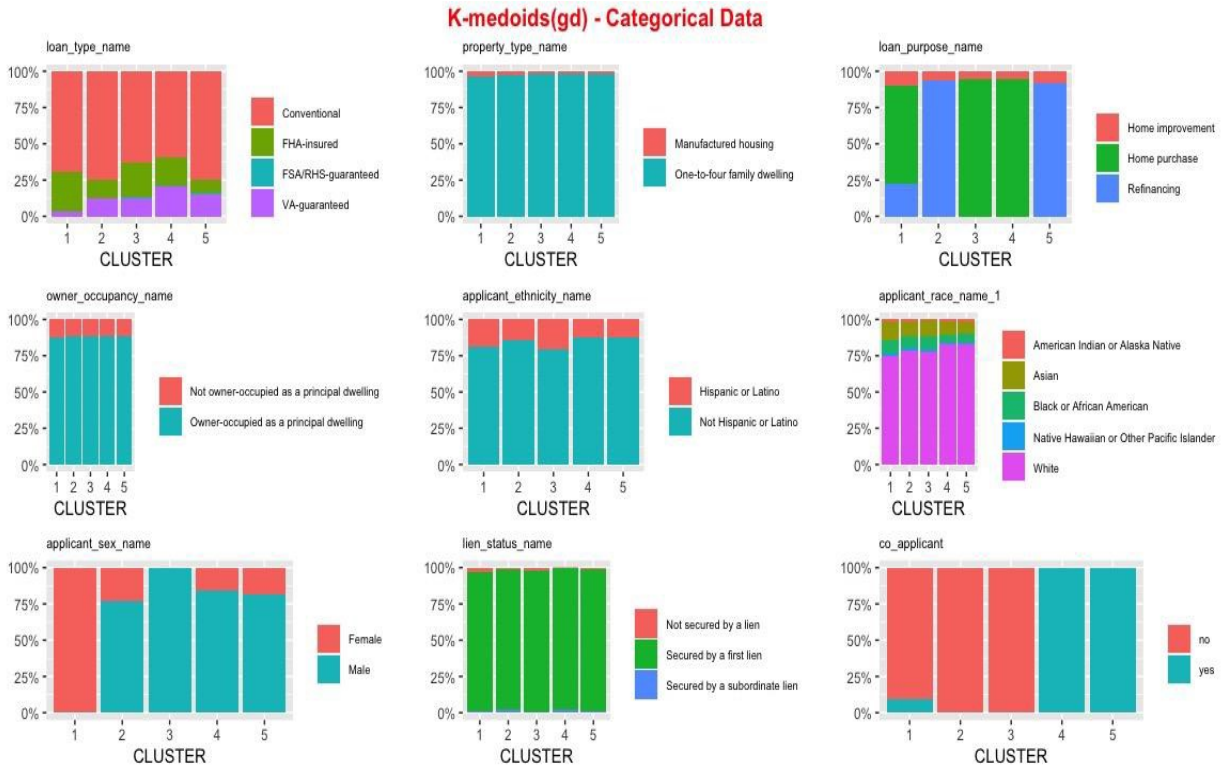
K-means (euclidean) results



Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

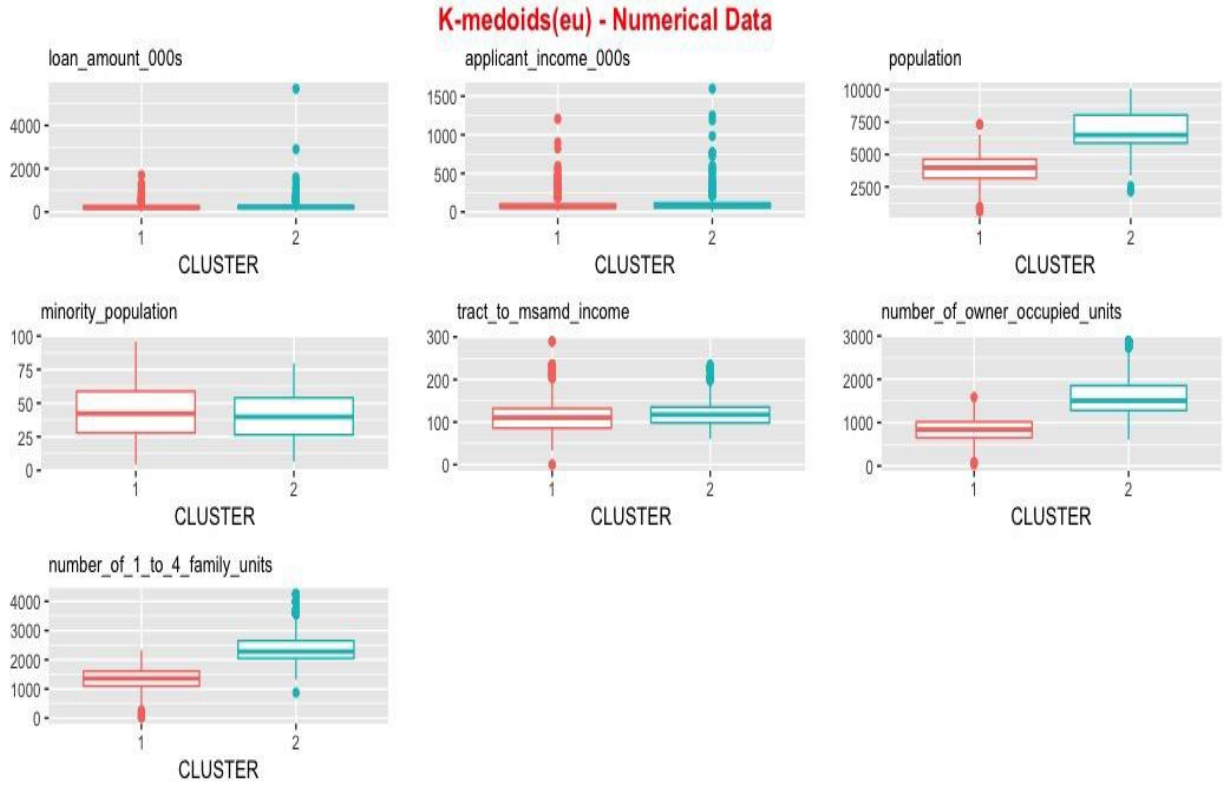
K-medoid (Gower's) results

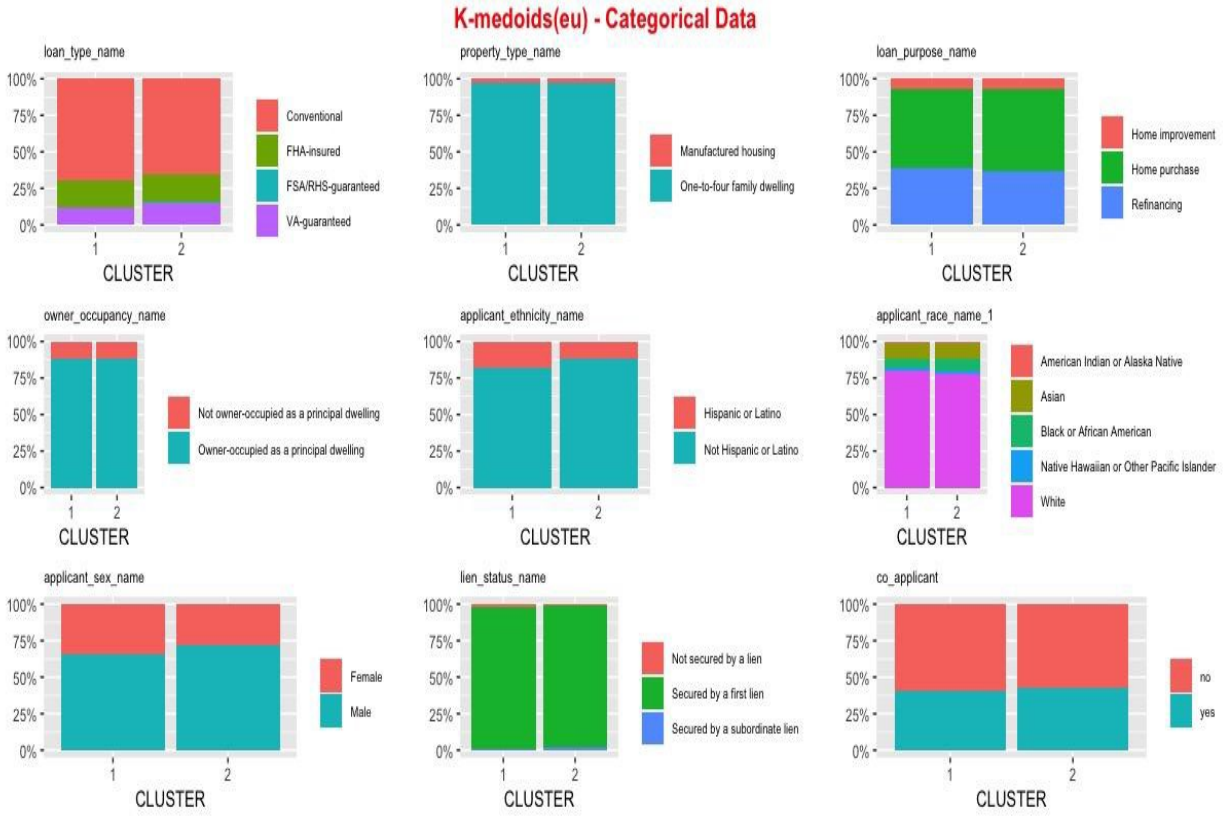




Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

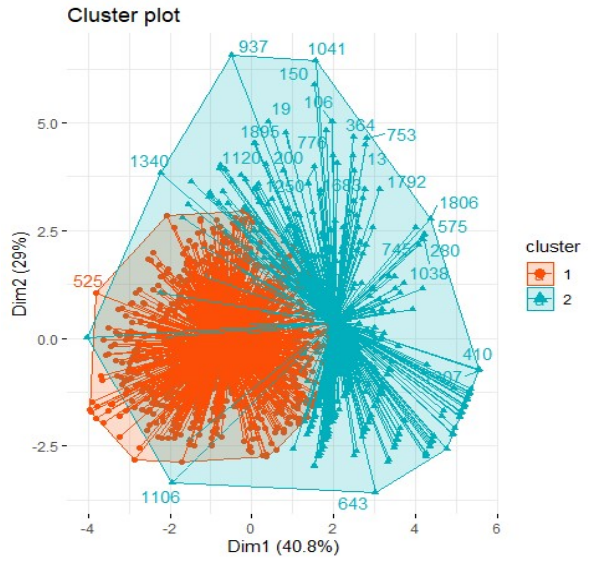
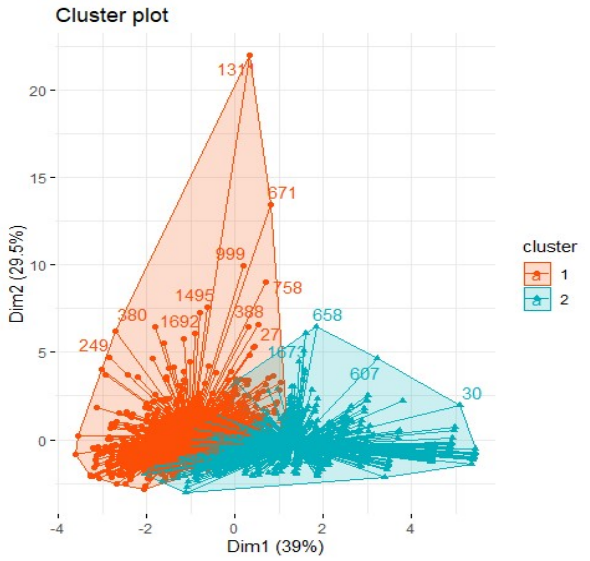
K-medoid (Euclidean) results



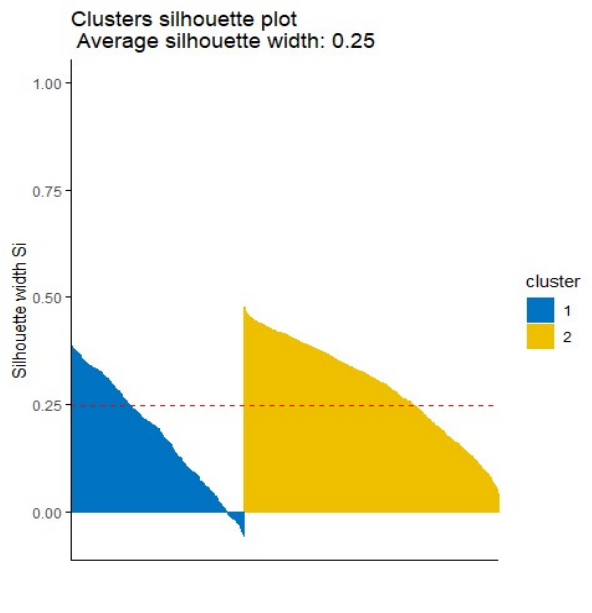
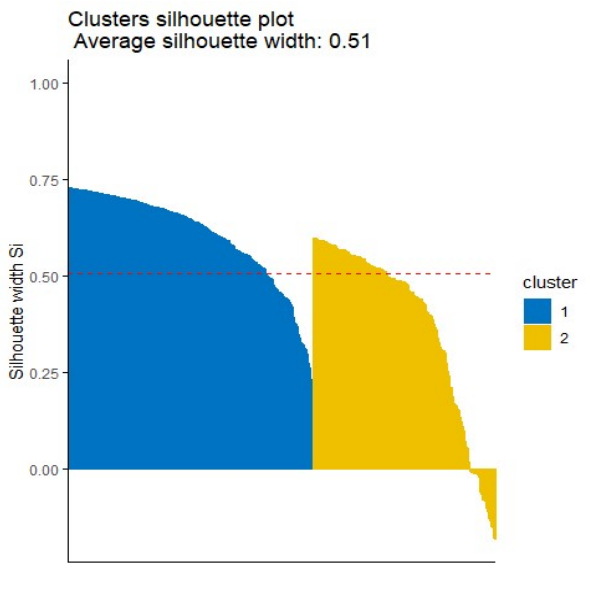


Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

K-means vs K-medoids - Cluster Plot (Numerical Variables / First 2 Principal Components)

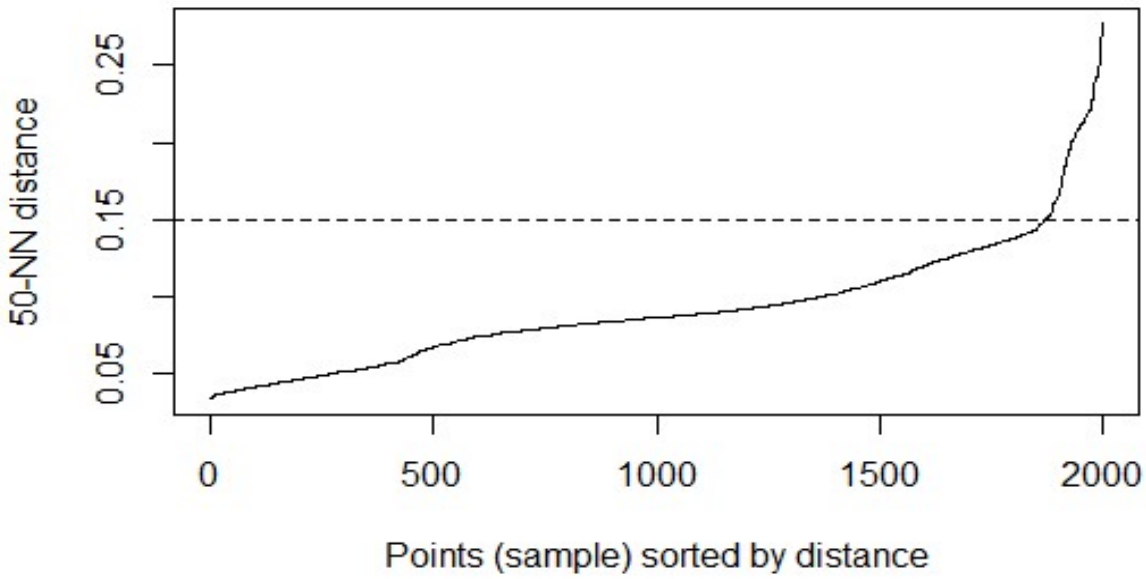
K-means	K-medoid
	

K-means vs K-medoids - Average Silhouette Width (All variables)

K-means	K-medoid
	
cluster size ave.sil.width 1 1 735 0.18 2 2 1080 0.29	cluster size ave.sil.width 1 1 1148 0.60 2 2 852 0.37

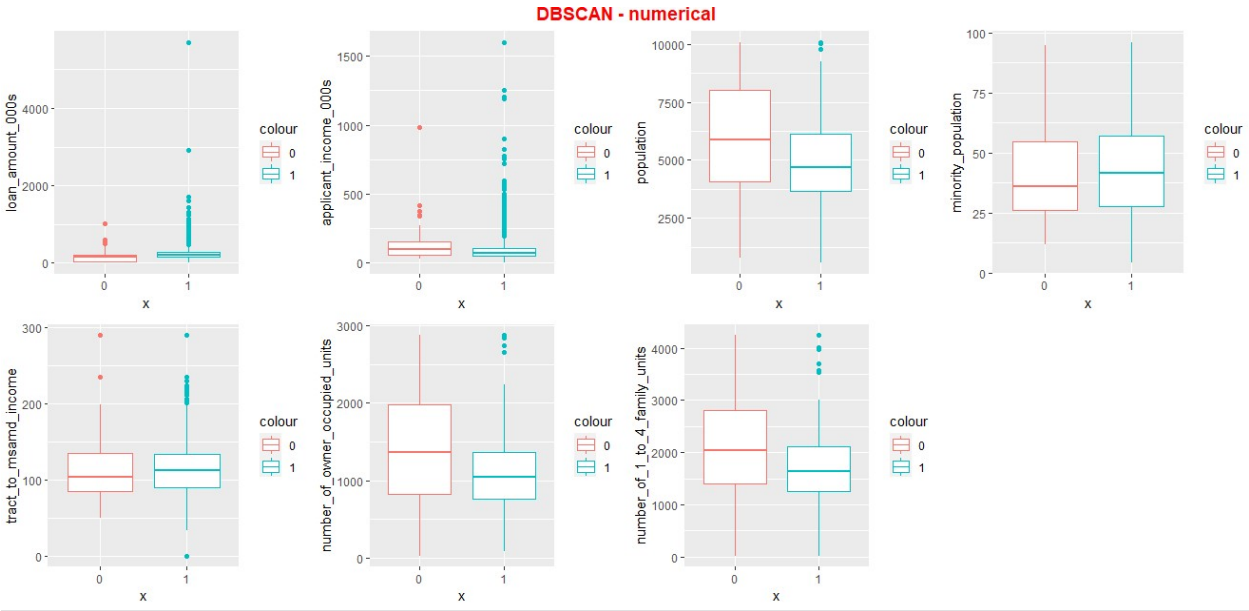
Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

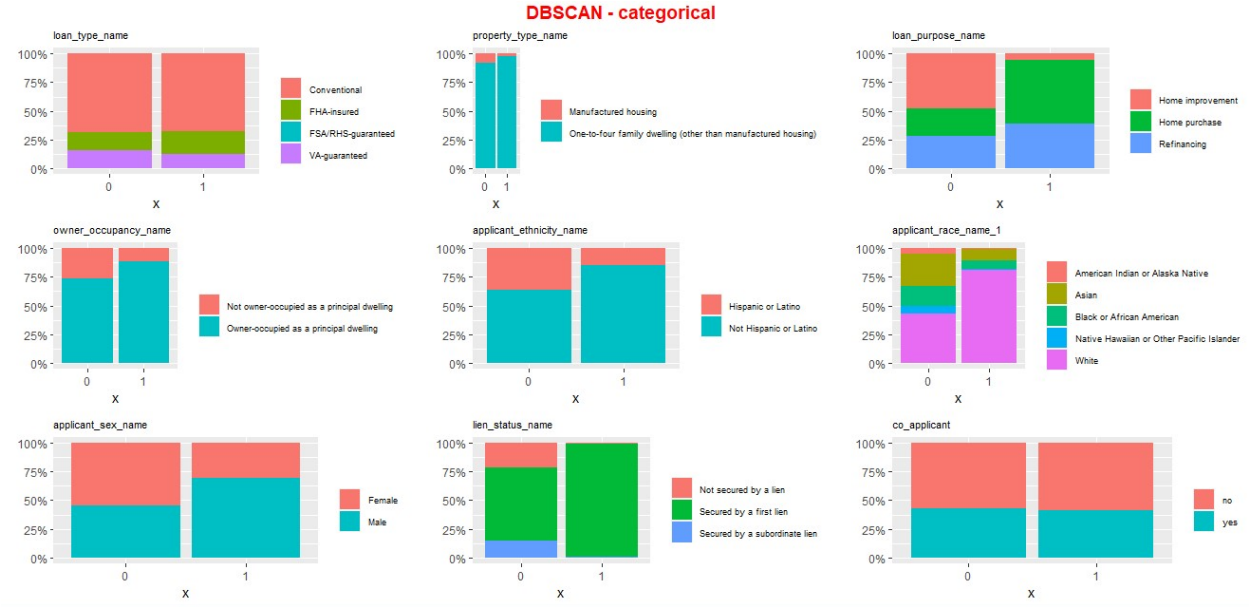
KNN -Displot for Optimal EP



Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

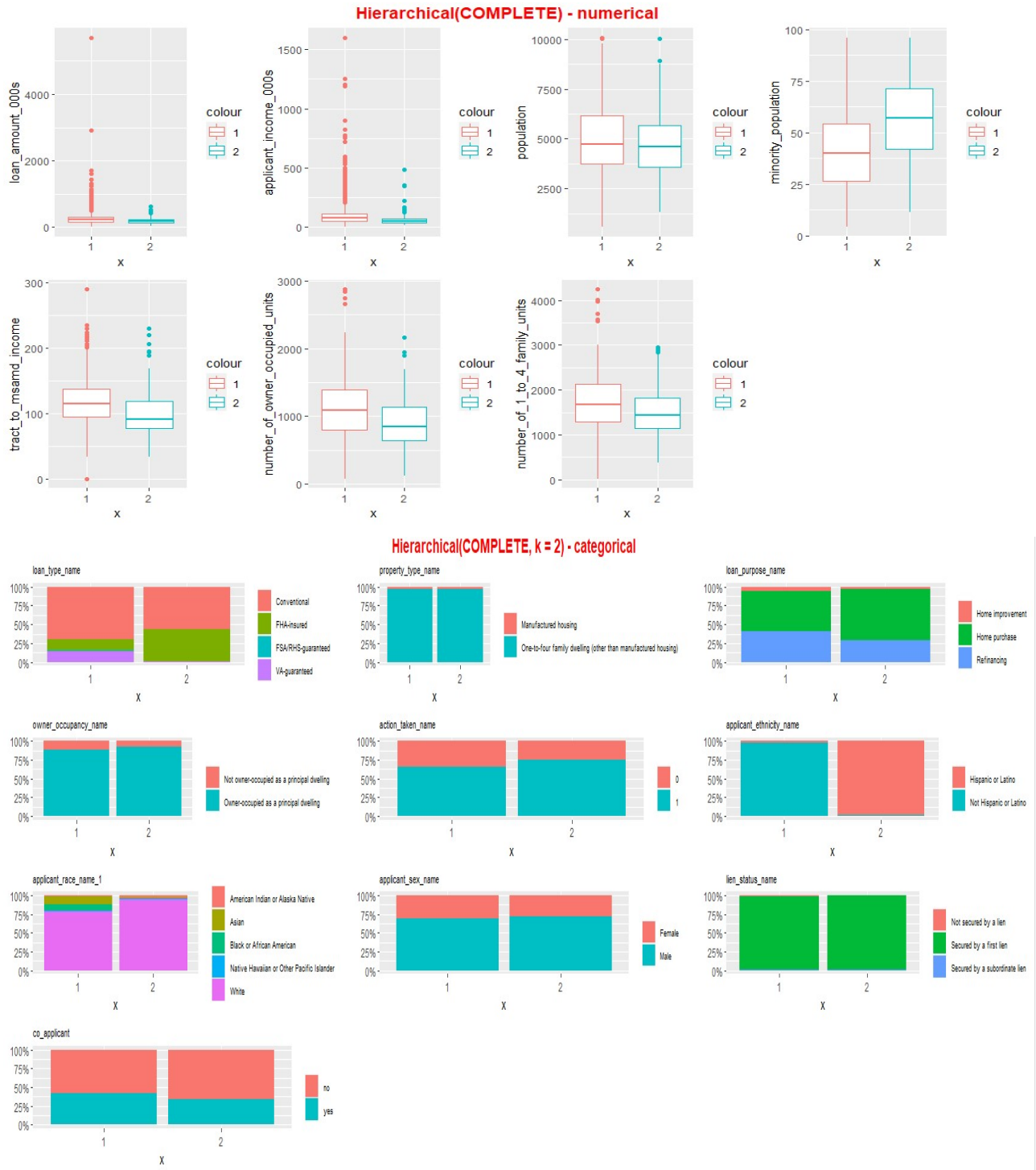
DBSCAN Outlier Characteristics





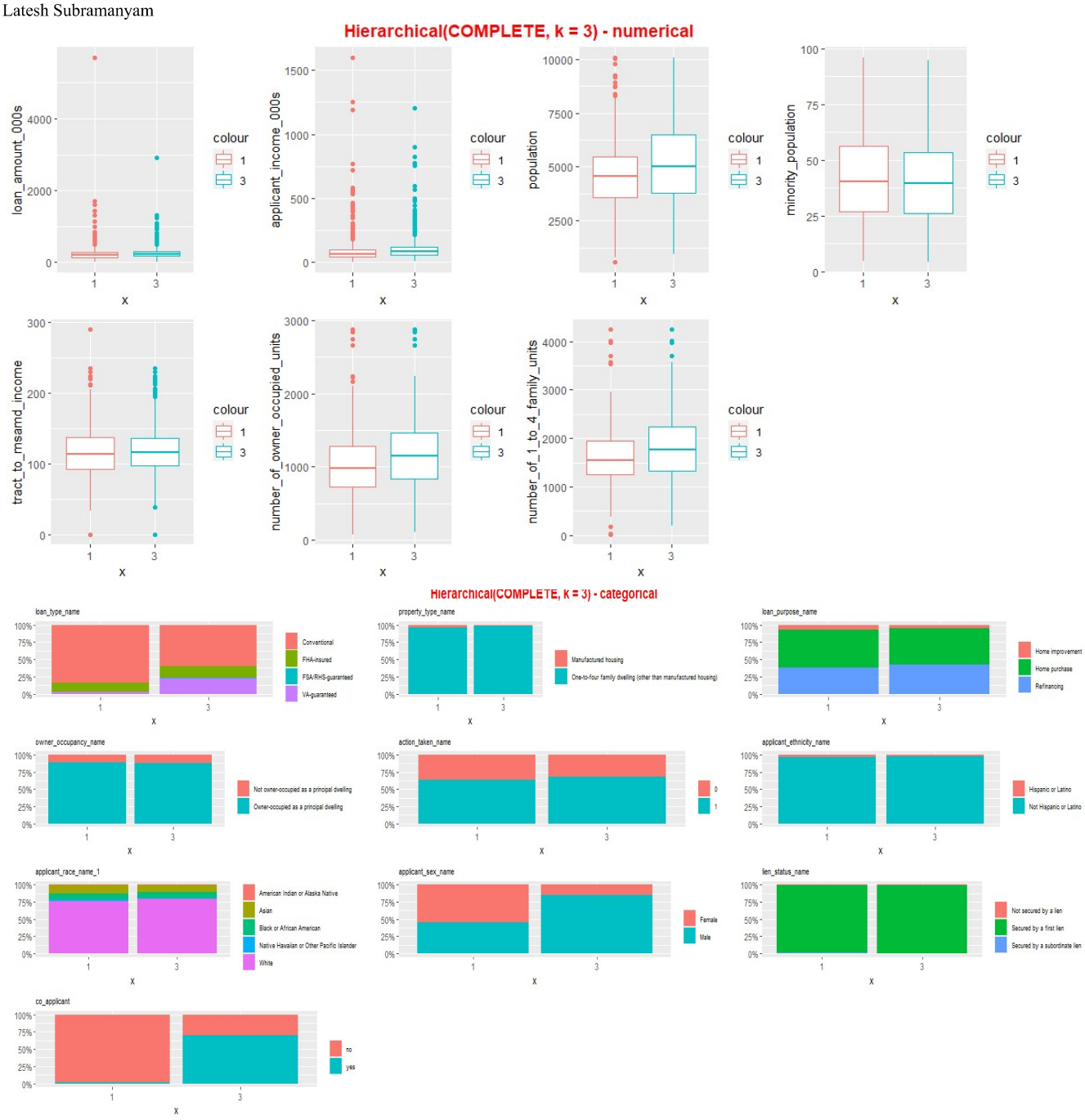
Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

Hierarchical Analysis result:



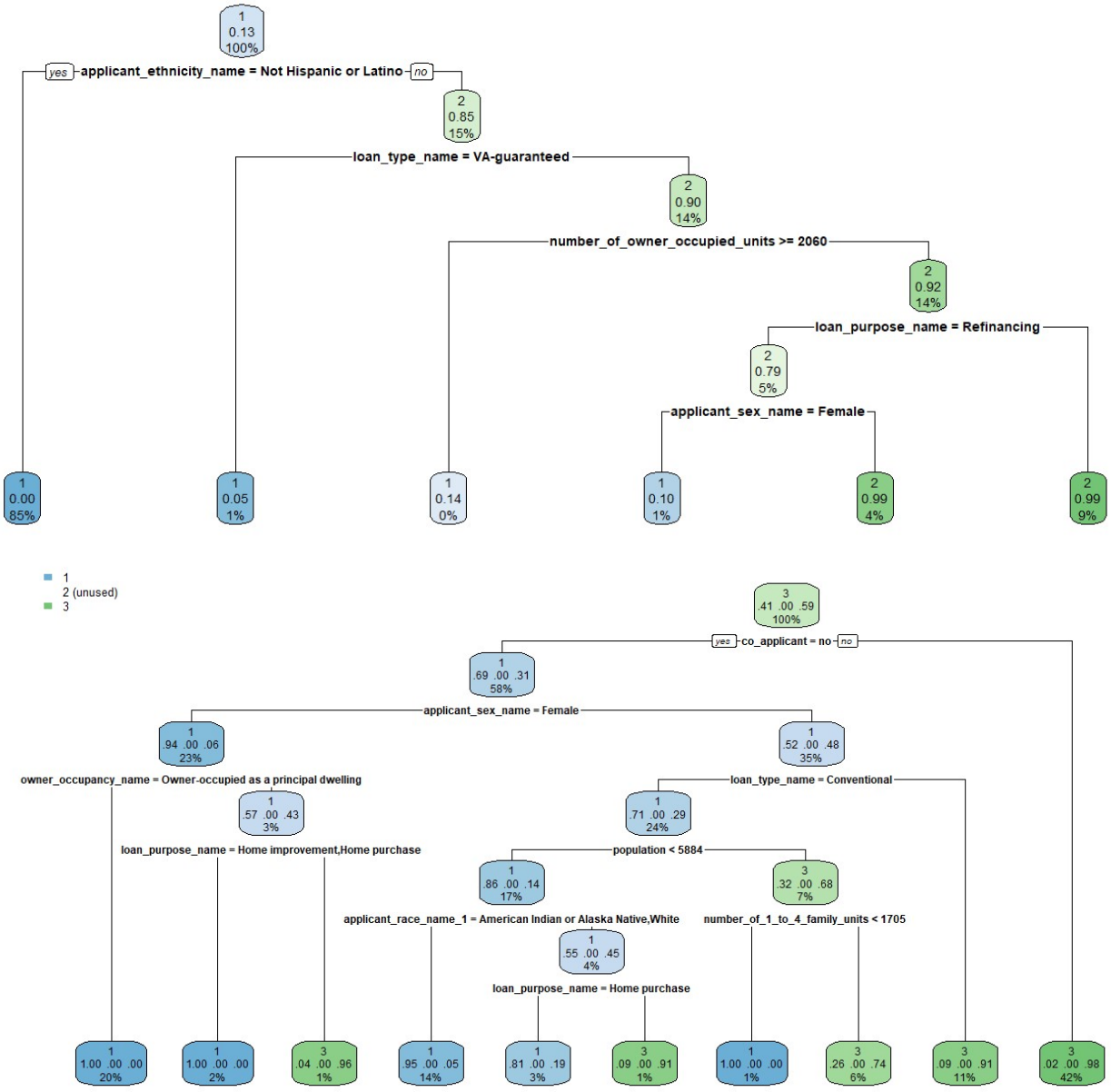
(Graph H-1)

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam



(Graph H-2)

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam



(Graph H-3)

Leo Liang
Qirui Cao
Young Lee
Latesh Subramanyam

Reference

Historic HMDA Data. Consumer Financial Protection Bureau. (n.d.). Retrieved April 14, 2022, from https://www.consumerfinance.gov/data-research/hmda/historic-data/?geo=wa&records=all-records&field_descriptions=labels

Top 10 worst states for affordable housing. Gov1. (2016, August 12). Retrieved April 14, 2022, from https://www.gov1.com/housing/articles/top-10-worst-states-for-affordable-housing-U1kGxUIB5UowVmOe/