Housing Affordability

Angelene Arito
Jeremy Dai
Helen Ramchandani
Ning Yan

I. Introduction

The data used in the following analyses was taken from the Housing Affordability Data System (HADS), which is a public dataset comprised of records from different public sources, primarily including the American Housing Survey (AHS) and the Census Bureau's official thresholds.[1] There are 99 variables and 64,535 records, each of which represent one sampled housing unit; a key of the variables and their definitions can be referenced in Chart 1 in the appendix. Overall, the dataset measures the affordability of housing units relative to area median incomes.

II. Exploratory Analysis

Diving into the dataset beyond the documentation, there were some immediate standouts. There were many instances of negative values for different variables across the board, though there was a pattern to them. Some variables had "-5" populated for a number of records in its column; other variables had the same pattern but with a "-6", "-9", or "'.'". The documentation on the dataset did not address these entries, but it was group consensus to mark them as missing values. Chart 2 in the appendix summarizes all missing value instances by variable as well as a percentage of the overall entries.

One other pattern that can be seen from Chart 2 is that there are many cases where a variable is missing 4438 values; looking at the data confirms that these missing values occur in the same records for each of the variables across the board (so a record missing a value for BURDEN was also missing a value for IPOV, PER, ZADEQ, etc.). There are also variables with high numbers of missing values as well, such as ASSISTED, VALUE, and VACANCY. Given this spread of missing values, the group analyzed what impact removing records with missing values would have on dataset size. All records with missing values were removed and variables with high percentages of missing values (60% or greater) were removed from analysis completely, due to their added shrinkage to dataset size. After that was performed, 35,852 records and 95 variables remained, more than sufficient still for analysis purposes.

The other primary discovery in exploratory analysis in regards to parsing down the dataset was the duplication of variables. Each of the variables with an FMT prefix had a companion variable in a different code. For example, FMTSTRUCTURETYPE and STRUCTURETYPE both capture the same exact information, but the former takes the form of "'1 Single Family'", "'2 2-4 units'" and the latter the form of "1", "2"; essentially, all FMT prefix variables were coded more cleanly in their non-prefixed counterparts. For this reason, those variables were removed; Chart 3 in the appendix summarizes which variables were kept for analysis moving forward. The final variable count after elimination was 71.

Another aspect of exploratory analysis conducted was to check for multicollinearity where many predictors proved to be highly correlated as seen in correlation matrix (in appendix). In addition, KMO (Kaiser-Meyer-Olkin), a tool to measure sample adequacy was calculated as well.

III. Model Selection

Due to the number and diversity of variable types in the dataset (even after the parsing down during exploratory analysis), several different models and iterations thereof were performed.

Principal Component Analysis (PCA)

PCA was deemed appropriate for this dataset given that 47 of the remaining variables kept in for analysis are numeric and given that one of the goals of analysis of this dataset is to discover underlying factors that may affect housing affordability. Measures of sampling adequacy were also run to ensure the validity of the PCA method on the data. The result of the KMO test was 0.5, which is considered sufficient, and Bartlett's Test of

[1] https://www.huduser.gov/portal/datasets/hads/hads.html

Sphericity was also significant at a value of 0 (reference Output 7 in the appendix). After confirming its validity on this dataset, PCA was then carried out.

For this analysis, only numerical variables in the dataset were considered. Chart 4 in the appendix identifies the variables that were used for PCA. Of these numeric variables, there are varying scales. Many of the variables represent monetary values, whether it be hundreds or thousands of dollars, such as VALUE and LMED; other variables represent a count, such as PER or ROOMS and there is also an age variable, AGE1; there are variables that represent percentages as well, such as COST06RELAMIPCT and COSTMedRELAMIPCT. Given the differences in scales, scaling the data for PCA was deemed necessary for the most interpretable results. The dataset was also split into training and test sets using k-fold cross-validation, in which half of the records were placed into a test set and the other half were placed into a training set.

In the first run of PCA, all 47 variables were used and scaled to accommodate their various units. Results from this showed that the first 8 components accounted for approximately 90% of the variance (full output can be referenced in Output 1 in the appendix), with a scree plot showing cut-off at 8 components, with an arguable "knee" prior to that. To account for most of the variance, the first 8 components were kept in for further analysis. Loadings for these 8 components were calculated and yielded some interesting results (referenced in Output 2 in the appendix), the most noteworthy being that none of the percentage-based variables were accounted for within the first 8 components. Beyond that, however, it is clear from the spread in this output that factor rotation is necessary to get a clearer picture about which components contain which variables. With that goal in mind, promax was applied to the components to make variable contributions easier to interpret. Output in the appendix shows the results of the rotation on the 8 components. This output provides much clearer of a picture of which variables contribute to each component. After running PCA on this set of data and becoming more familiar with how the variables interact, it was determined that the percentage variables may not be valuable in the context of PCA and determining underlying factors, especially when seeing that almost all of them belong to the same component (PC1). WEIGHT may also not be appropriate to include, as it does not add any unique information to the set.

With this in mind, a second round of PCA was performed with the percentage variables and WEIGHT variable omitted. The results in this iteration showed that the first 6 components account for approximately 90% of the variance (full output can be referenced in Output 4 in the appendix) and the scree plot shows that 5 or 6 could be included. To keep with consistency, the 90% cut-off was used as the determinant for number of components to carry forward for analysis, so 6 were kept. Loadings for these 6 components were then calculated (referenced in Output 5 in the appendix) and it was once again clear that rotation was needed for better interpretation of variable contribution to each of the components. Promax rotation was applied again, which yielded far clearer results. From these loadings (which can be referenced in Output 6 in the appendix), we can categorize each component and determine what they imply in regard to housing affordability and whether all the components are still necessary. Below are the component interpretations-

PC1: Income level – each of the variables weighted in this component summarize some level of income, either actual income or income adjusted against another factor. Overall, however, this component deals almost entirely with whether someone falls into the extremely low income, very low income, low income, or medium income category.
PC2: Housing costs – each of the variables weighted in this component deal with either overall housing cost at various interest rates or monthly housing costs.
PC3: Income status & number of persons – each variable reflects income level with the addition of the number of persons in the household. This may be considered too similar to the variables captured in PC1.
PC4: Number of rooms – the two primary variables in this component deal with the number of rooms in the household or structure, either total or specific breakdown (number of bedrooms).
PC5: Income amount – these variables reflect dollar amounts of income.

PC6: Costs related to number of units – the two variables here are number of units in the structure and other costs including insurance, condo, land rent, other mobile home fees, which appears more related to status of ownership being renter and perhaps deals less with single family homes and more with apartment buildings and condos.

Multiple Regression

Applying multiple regression to this dataset allowed visibility into which variables contributed to predicting the increase or decrease of our dependent variable. The exploration and modeling in a multiple regression analysis allows us to find the best explanatory variables to use in other statistical techniques.

Right from the start, there were several instances of multicollinearity. In some cases, some correlations between independent variables were a 100% positive correlation such as Area Median Income (LMED) and Growth-Adjusted Median Income (GLMED). I started with a purely numeric set as a training dataset for easier coding purposes and decreased the number of variables in the dataset from 99 total variables to 14 numeric only variables. These variables were chosen based on correlation plots and matrices showing highly correlated independent variables as well as using and comparing variables in Forward, Backward, and Stepwise selection processes. For my dependent variable, I switched between two variables which we as a group thought were significant and stood out. These were Fair Market Rent (FMR) and Value (VALUE), both of which are attributed to housing prices. My focus in this analysis was on what causes fluctuations in housing prices. I had to be careful because switching out the dependent variable caused my adjusted R squared to either go up or down, and I had to organize the independent variables that went with either dependent variable as well as keeping an eye on multicollinearity.

Looking at the results of the first linear model with FMR as the response variable (in appendix), we can see that the coefficient for FMR is negative and correlates with ROOMS and ZINC2 as they are also negative coefficient estimates. It seems that more digging into the data will be necessary because you would think that as BEDRMS go up, FMR would increase as well, but according to this, it is the opposite when FMR is the dependent variable. However, all the variables independently all have a “statistically significant” relationship with the likelihood of increasing housing prices.

I also procured a set of data for my final analysis, which included 17 numeric variables and 8 categorical variables. From this, I had to remove 3 variables (1 numeric, 2 categorical) because of singularities in the linear model which are basically variables that are too highly similar to another variable or a combination of other variables. Once those are gone, I can run VIF on the linear model which showed 3 more numeric variables related to income with factors over 10 which I also took out of the analysis. VIF is important here because the correlations did not show any highly correlated relationships with any of these 3 variables.

With this new, cleaned up dataset, the results come out more in line with what you’d expect. For example, as more beds exist in a household, the value of the house increases as well. COSTMED which is housing cost at median interest also has a positive correlation with the intercept along with BURDEN which is also cost related. The adjusted R squared is 1 in this case with most of the numeric variables being statistically significant and most of the categorical variables holding p-values of 0.1 higher, therefore being not statistically significant in this case.

Common Factor Analysis (CFA)

Once data exploration was completed, the “cleaned” data containing (~ 35K records and ~70 variables) was used for further analysis in form of factor analysis. CFA was implemented to find set of patterns in data and factors are formed for observation. Implementing CFA is a data reduction tool to help in remove redundancy from set of correlated variables. It also aids in selecting a variable to represent a group of predictors. The factors produced can be used in regression to create a more precise model. One important feature of factor analysis relates to factors calculated to maximize between group variance while minimizing in group variance.

Multicollinearity was a major issue with this dataset. The nature of CFA requires numeric data and the number of categorical variables present in dataset was a challenge to differentiate amongst the relationship between types of data. Using correlation matrix (Output 18A , Appendix), 28 significant predictors were extracted for analysis. Testing for sample adequacy (KMO) and running bartlett test (check for sample variance) proved to be statistically significant. In addition, testing for multicollinearity was priority. In order to obtain VIF values, the technique used is linear regression to extract a dependent variable from the predictor list. First, PCA was run on the correlation matrix using varimax rotation and scaling the variables followed by linear regression to obtain VIF values and extracted significant predictors. Next, a scree plot was created to view the number (2 or 3) of components to use for CFA. CFA was performed to determine latent factors caused by observed variables.

First, CFA was run (unrotated) and resulted in 2 factor implementation with cumulative variance of 55.4% and 3 factor implementation with cumulative variance of 64.5%. View factor loading, 18 significant predictors extracted. Next, the factors were rotated to see any significant changes using varimax (orthogonal) and promax (oblique) rotations. Check loadings for 2 factors (varimax) showed cumulative variance of 94% with more overlap, 3 factors cumulative variance of 91.7%. However, due to evidence of high correlation within the factors, it would be interesting to view results using oblique rotation. Surprisingly, the cumulated variance of 3 factors used in varimax was much higher value at 91.7 % (~92%) vs. oblique rotation, value at 87.4%. (Output 18, Appendix). Observing parsimony rule, findings resulted in 10 significant predictors formed by 2 factor using promax rotation. In addition, running Cronbach's alpha (measure to check factor reliability) proved to be statistically significant. The graph shows the separation of 2 factors and it's respective variables (Output 19, Appendix).

A summary of findings:

- Implementing varimax rotation captured 93.9 % (94%) with extensive overlap of variance while promax rotation valued at (94.1%) explained data with 2 factors and 10 predictors. Loadings of each rotation, varimax and promax can be viewed (Output 20, Appendix).
- Factor 1 renamed to new variable called, "income levels". Factor 1 variable consists of following: ABL30 (extremely low income level) ABL50(very low income level), ABL80(low income level), ABLMED (median income level).
- Factor 2 renamed to a new variable called, "housing cost". Factor 2 variable consists of following: COST06, COST06RELAMIPCT, COST06RELPOVPCT, COST06RELFMRPCT, COST08RELAMIPCT, COST08RELPOVPCT, COST08RELFMRPCT.
- Housing Affordability = income levels + housing cost. (Output 20A, Appendix)

Viewing Output 20A in the appendix, "housing cost06" indicates housing cost at 6% interest and "housingcost08" is housing cost at 8% interest and combined forms variable "housing cost". Both variables (housing cost and income levels) are positively related and contributes to the premise of housing affordability.

Multiple Correspondence Analysis (MCA)

In order to analyze the pattern of relationships among all the categorical variables, multiple correspondence analysis was conducted. The dataset used for MCA is the "cleaned" dataset after data exploration. In this dataset, there are 70 variables in total. Out of the 70 variables, there are 20 categorical variables. Each categorical variable has 3 to 7 levels/categories and in total, there are 94 categories(levels).

R is used to perform the MCA analysis. There are several packages in R that can perform MCA analysis and we choose "FactoMineR" package which is most commonly used to analyze multivariate dataset as in our example. The specific function used is MCA().

In order to apply all categorical values to the function, all character values are transformed into factor values. Before the analysis, we plotted all categorical variables to see if there are any variable categories with very low frequency. From the bar plots, there are several variable categories with very low frequency. However, in order to gain an overall understanding of all categorical variables, we decided to keep all categories at first. After applying

the function to all categorical variables, we get outputs of eigenvalues/variance explained and a scree plot. From these outputs, we could see that the first three dimensions can only explain 29.6% of the variance, which is a small percentage. To further understand the categorical variables, we made a biplot of our correspondence analysis. From the CA plot, all region categories are associated with cost related variables, all cost related variables are similar to each other. One interesting result is that housing type - Boat or recreation vehicle is very close to 100% to 150% Poverty category. Next, we made a correlation plot to see correlation between variables and principal dimensions. From the plot, most cost related variables are most related to dimension 2, all income related variables are most related to dimension1 and all region, housing type and metro area variables are related to dimension1 as well. In order to see quality of the representation, we used squared cosine to show degree of association between variable categories and a particular axis. The logic of this measurement is that if a variable category is well presented by two dimensions, the sum should be close to 1. In the output, variables with low square cosine values are colored in white, variables with medium square cosine values are colored in blue and variables with high square cosine values are colored in red. From the output, region categories, housing adequacy, income related variable categories and two of the housing types are not well presented the two dimensions. This may indicate that it needs more dimensions to represent these variable categories. LTE 30% AMI category is best represented by these two dimensions. To better understand the dimension, we plot contributions to the two dimensions. COST06RELFMRCAT category LTE 50% FMR, COSTMedRELAMICAT category LTE 30% AMI, COSTMedRELFMRCAT category LTE 50% FMR, COST06RELAMICAT category LTE 30% AMI and COST08RELFMRCAT category LTE 50% FMR contribute the most to dimension 1. COST08RELFMRCAT category 50.01-100%FMR, COST06RELFMRCAT category 50.01-100% FMR, COSTMedRELAMICAT category 30.01-50%AMI, COSTMedRELFMRCAT category 50.01-100%FMR and COST06RELAMICAT category30.01-50%AMI contribute the most to dimension 2.

In summary, from correspondence analysis, cost related variable and variable categories are most significant in our housing affordability analysis.

Linear Discriminant Analysis (LDA)

In an effort to observe a different angle of the data in relation to housing affordability, LDA was deemed an appropriate approach to observe the REGION variable more closely. Perhaps developing a model that is capable of predicting REGION for a new observation can shed some light on housing affordability respective to each of the four regions. The results from this analysis could possibly be looked at in parallel to some of the other methods to give a more complete interpretation.

For this analysis, only numerical variables plus region were considered to observe how well the numerical data contributes to predicting which region the observation belongs to. Because of what was discovered in PCA, the percentage variables were left out of this analysis as well. From there, the correlation matrix was consulted to determine if the variables should be pared down in any way due to high correlation values (which is related to multicollinearity and result interpretation). From this matrix (which can be referenced in Output 8 in the appendix), we can see that there are instances of variables with high correlations to a number of other variables – in some cases .9 and above. In order to avoid this, some variables were removed where this occurred – GLMED, PER, L30, L50, L80, GL30, GL50, GL80, ABL30, ABL50, ABL80, and ABLMED. The final preparatory step before running LDA was setting aside observations for cross-validation; k-fold cross-validation was chosen for this set, splitting up the dataset in two different iterations – one iteration split the data 50/50 between training and test sets and the other split it up 80/20.

LDA was then run on the 50/50 training set (Output 9 in the appendix) and its accuracy was measured against both the training and test set (Output 10 in the appendix). Prediction performance for region was relatively good in both the training and test sets, with over 60% classified correctly across all regions. Overall, there is consistency in accuracy between both training and test sets. Similar results can be seen for the 80/20 training and test sets, which are summarized in Output 23 in the appendix.

Finally, LDA was tested in one more iteration using the components uncovered during PCA. The highest loading was taken from each of the first six components (except for PC2, in which 3 were taken as they were equals and related to one another). The following variables were then tested in an LDA model: LMED, PER, COST06, COST08, COST12, BEDRMS, TOTSAL, NUNITs. Results and accuracy of results for the 50/50 training and test sets are summarized in Output 21 in the appendix. Proportions of trace were far higher for this LDA model, and the accuracy was quite a bit lower than the first model, implying that the components from PCA were not a good fit for predicting REGION, so the first model should be carried forward. Similar results can be seen for the 80/20 training and test sets as well, which are summarized in Output 24 in the appendix.

Cluster Analysis

Cluster analysis is data reduction tool which minimizes in cluster variance and aids in unsupervised classification. In the "Housing Data 2013 dataset, want to observe the regions in relationship to housing affordability, chose cluster analysis for further analysis using raw dataset as well as results from factor analysis. Since "Region" is categorical variable and categorized into 4 sections. The sections are as follows: Region '1'- Northwest; Region '2'- Midwest; Region '3'- South; Region '4'- West.

Implementation of cluster analysis on raw data findings:

K-means clustering with 4 clusters of sizes 308, 22025, 9599, 3920

Within cluster sum of squares by cluster:
[1] 46002.37 268284.94 215016.91 179169.99
(between_SS / total_SS = 50.6 %)

	'1'	'2'	'3'	'4'
'1'	119	2464	5053	1160
'2'	26	7874	1744	719
'3'	50	8816	1082	923
'4'	113	2871	1720	1118

Region '1' (Northwest) – 308 cluster size in region 1.
Region '2' (Midwest) - 22,025 cluster size in region 2.
Region '3' (South) – 9,599 cluster size in region 3.
Region '4' (West) = 3,920 cluster size in region 4.

Taking raw data into consideration, above table shows the clusters (1,2,3,4) and rows ('1', '2', '3', '4') as regions. Observing the diagonal, cluster 2 and region '2', a value of 7,874 indicates Region '2' (Midwest) is the choice for home affordability.
Note: Rest of numbers above and below diagonal are overlapping points between regions.

CONDUCTING CLUSTER ANALYSIS AFTER FACTOR ANALYSIS

Furthermore, CFA resulted in 2 factors. I chose the significant predictors amongst the factors, ran cluster analysis to view the relationship between Region and significant predictors. Findings listed below:

K-means clustering with 4 clusters of sizes 3571, 21896, 10073, 312

TABLE OF CLUSTER VS. REGIONS

	'1'	'2'	'3'	'4'
'1'	1022	2196	5458	120
'2'	646	8070	1620	27
'3'	867	8860	1093	51
'4'	1036	770	1902	114

Region ‘1’ (Northwest) – 3,571 cluster size in region 1.
Region ‘2’ (Midwest) - 21,896 cluster size in region 2.
Region ‘3’ (South) – 10,073 cluster size in region 3.
Region ‘4’ (West) - 312 cluster size in region 4.

Extracting significant variables from Factor (1,2), above table shows the clusters (1,2,3,4) and rows (‘1’, ‘2’, ‘3’, ‘4’) as regions. Observing diagonal values, cluster 2 and region ‘2’, a value of 8070 and surprisingly, **Region ‘2’ (Midwest) is the choice for home affordability.** Rest of values up and below diagonal are overlapping values from one cluster to another.

IV. Analysis of Results & Discussion

From all the analyses, housing cost is a common factor that is extracted. This factor/ component is related to original housing cost and expenses, mortgage interest rates and interest rates adjusted by FMR. From both PCA and CFA, income level is a common factor/ component, this factor relates to actual income/income levels and income/income levels adjusted by factors such as area median income, number of persons in household and FMR. Both LDA and cluster analysis evaluate relationship between region and housing affordability. This variable is not a significant output from all our analysis. The result from LDA is that it keeps the model against components extracted from PCA for prediction of region. Although this result can’t explain any of the outputs from other analysis, it can be an interesting factor from another angle to explore variables that are left out by PCA/CAF/multiple regression and MCA.

V. Appendix

Chart 1 – Dataset variable key

Variable	Description
ABL30	Extremely Low Income Adjusted for # of Bedrooms
ABL50	Very Low Income Adjusted for # of Bedrooms
ABL80	Low Income Adjusted for # of Bedrooms
ABLMED	Median Income Adjusted for # of Bedrooms
AGE1	Age of householder (1997-2009 name)
APLMED	Median income adjusted for number of persons
BEDRMS	Number of bedrooms in unit
BUILT	Year unit was built
BURDEN	Housing cost as a fraction of income
CONTROL	AHS control number
COST06	Housing cost at 6% interest
COST06RELAMICAT	Cost06 Relative to Median Income (Category)
COST06RELAMIPCT	Cost06 Relative to Median Income (Percent)
COST06RELFMRCAT	Cost06 Relative to FMR (Category)
COST06RELFMRPCT	Cost06 Relative to FMR (Percent)
COST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)
COST06RELPOVPCT	Cost06 Relative to Poverty Income (Percent)
COST08	Housing cost at 8% interest
COST08RELAMICAT	Cost08 Relative to Median Income (Category)
COST08RELAMIPCT	Cost08 Relative to Median Income (Percent)
COST08RELFMRCAT	Cost08 Relative to FMR (Category)
COST08RELFMRPCT	Cost08 Relative to FMR (Percent)
COST08RELPOVCAT	Cost08 Relative to Poverty Income (Category)
COST08RELPOVPCT	Cost08 Relative to Poverty Income (Percent)
COST12	Housing cost at 12% interest
COST12RELAMICAT	Cost12 Relative to Median Income (Category)
COST12RELAMIPCT	Cost12 Relative to Median Income (Percent)
COST12RELFMRCAT	Cost12 Relative to FMR (Category)
COST12RELFMRPCT	Cost12 Relative to FMR (Percent)
COST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)
COST12RELPOVPCT	Cost12 Relative to Poverty Income (Percent)
COSTMED	Housing cost at median interest
COSTMedRELAMICAT	CostMed Relative to Median Income (Category)
COSTMedRELAMIPCT	CostMed Relative to Median Income (Percent)
COSTMedRELFMRCAT	CostMed Relative to FMR (Category)
COSTMedRELFMRPCT	CostMed Relative to FMR (Percent)
COSTMedRELPOVCAT	CostMed Relative to Poverty Income (Category)
COSTMedRELPOVPCT	CostMed Relative to Poverty Income (Percent)
FMR	Fair market rent
GL30	Growth-adjusted extremely low income
GL50	Growth-adjusted very low income
GL80	Growth-adjusted low income
GLMED	Growth-adjusted median income
INCRELAMICAT	HH Income Relative to FMR (Category)
INCRELAMIPCT	HH Income relative to AMI (percent)
INCRELFMRCAT	HH Income relative to AMI (category)
INCRELFMRPCT	HH Income Relative to FMR (Percent)
INCRELPOVCAT	HH Income Relative to Poverty Income (Category)
INCRELPOVPCT	HH Income Relative to Poverty Income (Percent)

IPOV	Poverty income
L30	Extremely low income (AMI variable)
L50	Very low income (AMI variable)
L80	Low income (AMI variable)
LMED	Median income
METRO3	Metro status (1997-2009 name)
NUNITS	Number of units in building
OTHERCOST	Insurance, condo, land rent, other mobile home fees
OWNRENT	Tenure (adjusted)
PER	Number of persons in household
REGION	Four census regions
ROOMS	Number of rooms in unit
STRUCTURETYPE	Recoded structure type
TENURE	Owner/renter status of unit
TOTSAL	Total wage income
TYPE	Structure type
UTILITY	Monthly utility cost
VALUE	Current market value of unit
WEIGHT	Final weight
ZADEQ	Recoded adequacy of housing
ZINC2	Household income
ZSMHC	Monthly housing costs
ASSISTED	Assisted housing
FMTASSISTED	Assisted housing
FMTBEDRMS	Number of bedrooms in unit
FMTBUILT	Year unit was built
FMTBURDEN	Cost burden
FMTCOST06RELAMICAT	Cost06 Relative to Median Income (Category)
FMTCOST06RELFMRCAT	Cost06 Relative to FMR (Category)
FMTCOST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)
FMTCOST08RELAMICAT	Cost08 Relative to Median Income (Category)
FMTCOST08RELFMRCAT	Cost08 Relative to FMR (Category)
FMTCOST08RELPOVCAT	Cost 08 Relative to Poverty Income (Category)
FMTCOST12RELAMICAT	Cost12 Relative to Median Income (Category)
FMTCOST12RELFMRCAT	Cost12 Relative to FMR (Category)
FMTCOST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)
FMTCOSTMEDRELAMICAT	CostMed Relative to Median Income (Category)
FMTCOSTMEDRELFMRCAT	CostMed Relative to FMR (Category)
FMTCOSTMEDRELPOVCAT	CostMed Relative to Poverty Income (Category)
FMTINCRELAMICAT	HH Income Relative to Median Income (Category)
FMTINCRELFMRCAT	HH Income Relative to FMR (Category)
FMTINCRELPOVCAT	HH Income Relative to Poverty Income (Category)
FMTMETRO3	CENTRAL CITY / SUBURBAN STATUS
FMTOWNRENT	Owner/Renter Status (adjusted)
FMTREGION	Census region
FMTSTATUS	Occupancy status
FMTSTRUCTURETYPE	Structure type
FMTZADEQ	Adequacy of unit
STATUS	Interview status
VACANCY	Vacancy status

Chart 2 -Summary of missing values in the original dataset

Variable	Missing Value Count	Percentage
AGE1	4438	7%
APLMED	4438	7%
ASSISTED	40290	62%
BURDEN	4438	7%
COST06RELPOVCAT	4438	7%
COST06RELPOVPCT	4438	7%
COST08RELPOVCAT	4438	7%
COST08RELPOVPCT	4438	7%
COST12RELPOVCAT	4438	7%
COST12RELPOVPCT	4438	7%
COSTMedRELPOVPCT	4438	7%
FMTASSISTED	40290	62%
FMTBUILT	10058	16%
FMTBURDEN	4438	7%
FMTCOST06RELPOVCAT	4438	7%
FMTCOST08RELPOVCAT	4438	7%
FMTCOST12RELPOVCAT	4438	7%
FMTCOSTMEDRELPOVCAT	4438	7%
FMTINCRELAMICAT	4438	7%
FMTINCRELFMRCAT	4438	7%
FMTINCRELPOVCAT	4438	7%
FMTMETRO3	43042	67%
FMTREGION	53179	82%
FMTSTATUS	64535	100%
FMTSTRUCTURETYPE	2	0.003%
FMTZADEQ	4438	7%
INCRELAMICAT	4438	7%
INCRELAMIPCT	4438	7%
INCRELFMRCAT	4438	7%
INCRELFMRPCT	4438	7%
INCRELPOVCAT	4438	7%
INCRELPOVPCT	4438	7%
IPOV	4438	7%
NUNITS	2	0.003%
PER	4438	7%
STRUCTURETYPE	2	0.003%
TENURE	4438	7%
TOTSAL	4438	7%
VACANCY	60097	93%
VALUE	27389	42%
WEIGHT	32	0.05%
ZADEQ	4438	7%
ZINC2	4438	7%
ZSMHC	4438	7%

Chart 3 – Summary of variables to keep and omit for analysis

Variable	Description	Keep/Omit
ABL30	Extremely Low Income Adjusted for # of Bedrooms	KEEP
ABL50	Very Low Income Adjusted for # of Bedrooms	KEEP
ABL80	Low Income Adjusted for # of Bedrooms	KEEP
ABLMED	Median Income Adjusted for # of Bedrooms	KEEP
AGE1	Age of householder (1997-2009 name)	KEEP
APLMED	Median income adjusted for number of persons	KEEP
BEDRMS	Number of bedrooms in unit	KEEP
BUILT	Year unit was built	KEEP
BURDEN	Housing cost as a fraction of income	KEEP
CONTROL	AHS control number	KEEP
COST06	Housing cost at 6% interest	KEEP
COST06RELAMICAT	Cost06 Relative to Median Income (Category)	KEEP
COST06RELAMIPCT	Cost06 Relative to Median Income (Percent)	KEEP
COST06RELFMRCAT	Cost06 Relative to FMR (Category)	KEEP
COST06RELFMRPCT	Cost06 Relative to FMR (Percent)	KEEP
COST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)	KEEP
COST06RELPOVPCT	Cost06 Relative to Poverty Income (Percent)	KEEP
COST08	Housing cost at 8% interest	KEEP
COST08RELAMICAT	Cost08 Relative to Median Income (Category)	KEEP
COST08RELAMIPCT	Cost08 Relative to Median Income (Percent)	KEEP
COST08RELFMRCAT	Cost08 Relative to FMR (Category)	KEEP
COST08RELFMRPCT	Cost08 Relative to FMR (Percent)	KEEP
COST08RELPOVCAT	Cost08 Relative to Poverty Income (Category)	KEEP
COST08RELPOVPCT	Cost08 Relative to Poverty Income (Percent)	KEEP
COST12	Housing cost at 12% interest	KEEP
COST12RELAMICAT	Cost12 Relative to Median Income (Category)	KEEP
COST12RELAMIPCT	Cost12 Relative to Median Income (Percent)	KEEP
COST12RELFMRCAT	Cost12 Relative to FMR (Category)	KEEP
COST12RELFMRPCT	Cost12 Relative to FMR (Percent)	KEEP
COST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)	KEEP
COST12RELPOVPCT	Cost12 Relative to Poverty Income (Percent)	KEEP
COSTMED	Housing cost at median interest	KEEP
COSTMedRELAMICAT	CostMed Relative to Median Income (Category)	KEEP
COSTMedRELAMIPCT	CostMed Relative to Median Income (Percent)	KEEP
COSTMedRELFMRCAT	CostMed Relative to FMR (Category)	KEEP
COSTMedRELFMRPCT	CostMed Relative to FMR (Percent)	KEEP
COSTMedRELPOVCAT	CostMed Relative to Poverty Income (Category)	KEEP
COSTMedRELPOVPCT	CostMed Relative to Poverty Income (Percent)	KEEP
FMR	Fair market rent	KEEP
GL30	Growth-adjusted extremely low income	KEEP
GL50	Growth-adjusted very low income	KEEP
GL80	Growth-adjusted low income	KEEP
GLMED	Growth-adjusted median income	KEEP
INCRELAMICAT	HH Income Relative to FMR (Category)	KEEP
INCRELAMIPCT	HH Income relative to AMI (percent)	KEEP
INCRELFMRCAT	HH Income relative to AMI (category)	KEEP
INCRELFMRPCT	HH Income Relative to FMR (Percent)	KEEP
INCRELPOVCAT	HH Income Relative to Poverty Income (Category)	KEEP
INCRELPOVPCT	HH Income Relative to Poverty Income (Percent)	KEEP

IPOV	Poverty income	KEEP
L30	Extremely low income (AMI variable)	KEEP
L50	Very low income (AMI variable)	KEEP
L80	Low income (AMI variable)	KEEP
LMED	Median income	KEEP
METRO3	Metro status (1997-2009 name)	KEEP
NUNITS	Number of units in building	KEEP
OTHERCOST	Insurance, condo, land rent, other mobile home fees	KEEP
OWNRENT	Tenure (adjusted)	KEEP
PER	Number of persons in household	KEEP
REGION	Four census regions	KEEP
ROOMS	Number of rooms in unit	KEEP
STRUCTURETYPE	Recoded structure type	KEEP
TENURE	Owner/renter status of unit	KEEP
TOTSAL	Total wage income	KEEP
TYPE	Structure type	KEEP
UTILITY	Monthly utility cost	KEEP
VALUE	Current market value of unit	KEEP
WEIGHT	Final weight	KEEP
ZADEQ	Recoded adequacy of housing	KEEP
ZINC2	Household income	KEEP
ZSMHC	Monthly housing costs	KEEP
ASSISTED	Assisted housing	OMIT
FMTASSISTED	Assisted housing	OMIT
FMTBEDRMS	Number of bedrooms in unit	OMIT
FMTBUILT	Year unit was built	OMIT
FMTBURDEN	Cost burden	OMIT
FMTCOST06RELAMICAT	Cost06 Relative to Median Income (Category)	OMIT
FMTCOST06RELFMRCAT	Cost06 Relative to FMR (Category)	OMIT
FMTCOST06RELPOVCAT	Cost06 Relative to Poverty Income (Category)	OMIT
FMTCOST08RELAMICAT	Cost08 Relative to Median Income (Category)	OMIT
FMTCOST08RELFMRCAT	Cost08 Relative to FMR (Category)	OMIT
FMTCOST08RELPOVCAT	Cost 08 Relative to Poverty Income (Category)	OMIT
FMTCOST12RELAMICAT	Cost12 Relative to Median Income (Category)	OMIT
FMTCOST12RELFMRCAT	Cost12 Relative to FMR (Category)	OMIT
FMTCOST12RELPOVCAT	Cost12 Relative to Poverty Income (Category)	OMIT
FMTCOSTMEDRELAMICAT	CostMed Relative to Median Income (Category)	OMIT
FMTCOSTMEDRELFMRCAT	CostMed Relative to FMR (Category)	OMIT
FMTCOSTMEDRELPOVCAT	CostMed Relative to Poverty Income (Category)	OMIT
FMTINCRELAMICAT	HH Income Relative to Median Income (Category)	OMIT
FMTINCRELFMRCAT	HH Income Relative to FMR (Category)	OMIT
FMTINCRELPOVCAT	HH Income Relative to Poverty Income (Category)	OMIT
FMTMETRO3	CENTRAL CITY / SUBURBAN STATUS	OMIT
FMTOWNRENT	Owner/Renter Status (adjusted)	OMIT
FMTREGION	Census region	OMIT
FMTSTATUS	Occupancy status	OMIT
FMTSTRUCTURETYPE	Structure type	OMIT
FMTZADEQ	Adequacy of unit	OMIT
STATUS	Interview status	OMIT
VACANCY	Vacancy status	OMIT

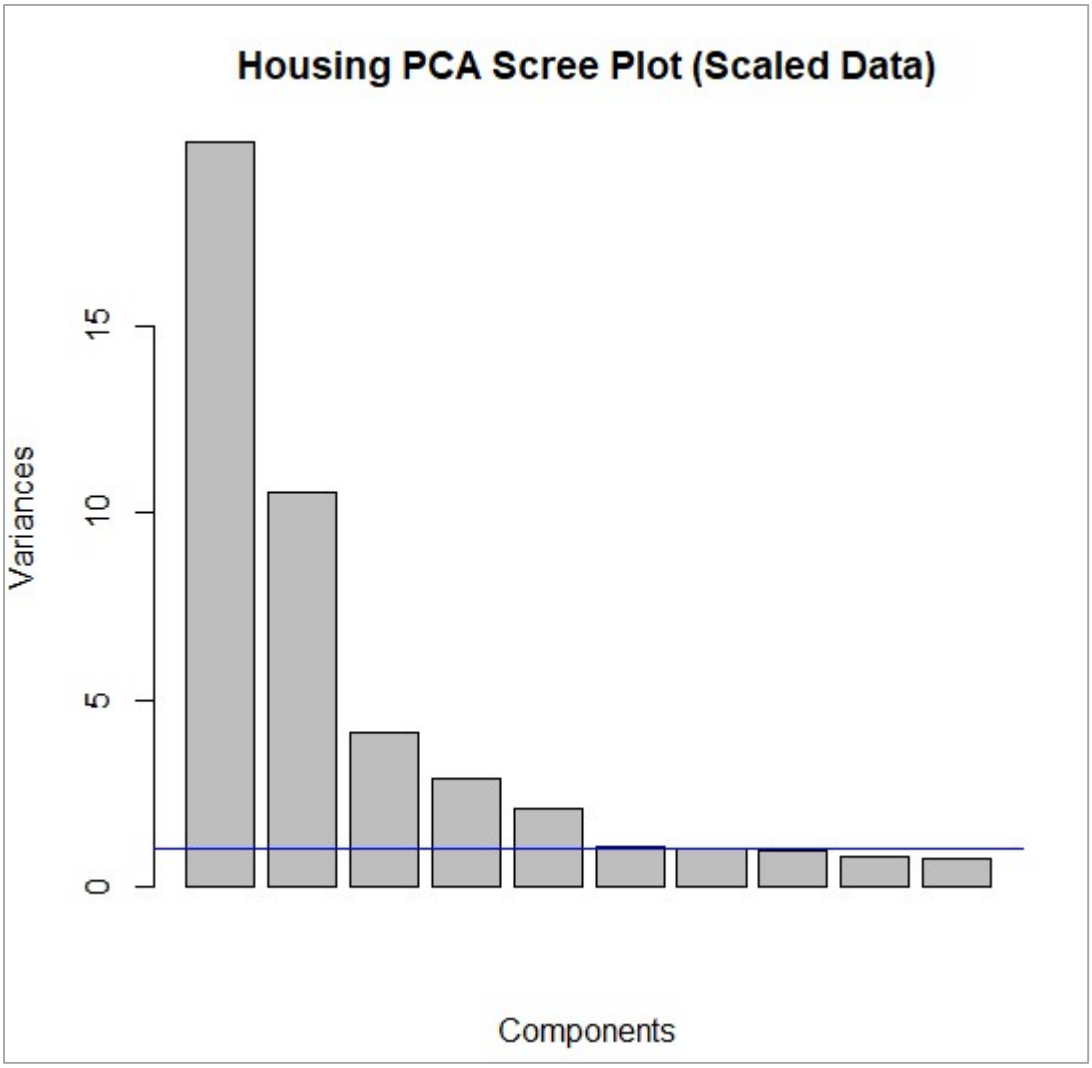
Chart 4 – Numeric variables from the dataset

Variable	Description
AGE1	Age of householder (1997-2009 name)
LMED	Median income
FMR	Fair market rent
L30	Extremely low income (AMI variable)
L50	Very low income (AMI variable)
L80	Low income (AMI variable)
IPOV	Poverty income
BEDRMS	Number of bedrooms in unit
VALUE	Current market value of unit
NUNITS	Number of units in building
ROOMS	Number of rooms in unit
WEIGHT	Final w eight
PER	Number of persons in household
ZINC2	Household income
ZSMHC	Monthly housing costs
UTILITY	Monthly utility cost
OTHERCOST	Insurance, condo, land rent, other mobile home fees
COST06	Housing cost at 6% interest
COST12	Housing cost at 12% interest
COST08	Housing cost at 8% interest
COSTMED	Housing cost at median interest
TOTSAL	Total w age income
GLMED	Grow th-adjusted median income
GL30	Grow th-adjusted extremely low income
GL50	Grow th-adjusted very low income
GL80	Grow th-adjusted low income
APLMED	Median income adjusted for number of persons
ABL30	Extremely Low Income Adjusted for # of Bedrooms
ABL50	Very Low Income Adjusted for # of Bedrooms
ABL80	Low Income Adjusted for # of Bedrooms
ABLMED	Median Income Adjusted for # of Bedrooms
BURDEN	Housing cost as a fraction of income
INCRELAMIPCT	HH Income relative to AMI (percent)
INCRELPOVPCT	HH Income Relative to Poverty Income (Percent)
INCRELFMRPCT	HH Income Relative to FMR (Percent)
COST06RELAMIPCT	Cost06 Relative to Median Income (Percent)
COST06RELPOVPCT	Cost06 Relative to Poverty Income (Percent)
COST06RELFMRPCT	Cost06 Relative to FMR (Percent)
COST08RELAMIPCT	Cost08 Relative to Median Income (Percent)
COST08RELPOVPCT	Cost08 Relative to Poverty Income (Percent)
COST08RELFMRPCT	Cost08 Relative to FMR (Percent)
COST12RELAMIPCT	Cost12 Relative to Median Income (Percent)
COST12RELPOVPCT	Cost12 Relative to Poverty Income (Percent)
COST12RELFMRPCT	Cost12 Relative to FMR (Percent)
COSTMedRELAMIPCT	CostMed Relative to Median Income (Percent)
COSTMedRELPOVPCT	CostMed Relative to Poverty Income (Percent)
COSTMedRELFMRPCT	CostMed Relative to FMR (Percent)

Output 1 – PCA results with all 47 numeric variables, scaled

```
Importance of components%s:
                            PC1    PC2     PC3     PC4    PC5     PC6    PC7
Standard deviation     4.4621 3.2453 2.02767 1.69089 1.4445 1.03167 1.0028
Proportion of Variance 0.4236 0.2241 0.08748 0.06083 0.0444 0.02265 0.0214
Cumulative Proportion  0.4236 0.6477 0.73519 0.79603 0.8404 0.86307 0.8845
                           PC8    PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.98560 0.9043 0.86406 0.82925 0.75432 0.63861 0.53509
Proportion of Variance 0.02067 0.0174 0.01589 0.01463 0.01211 0.00868 0.00609
Cumulative Proportion  0.90513 0.9225 0.93841 0.95305 0.96515 0.97383 0.97992
                          PC15    PC16    PC17    PC18    PC19    PC20    PC21
Standard deviation     0.5037 0.47716 0.36225 0.32649 0.23950 0.22468 0.21416
Proportion of Variance 0.0054 0.00484 0.00279 0.00227 0.00122 0.00107 0.00098
Cumulative Proportion  0.9853 0.99016 0.99296 0.99522 0.99644 0.99752 0.99849
                          PC22    PC23    PC24    PC25    PC26    PC27    PC28
Standard deviation     0.14761 0.14184 0.11560 0.08586 0.07808 0.03554 0.01844
Proportion of Variance 0.00046 0.00043 0.00028 0.00016 0.00013 0.00003 0.00001
Cumulative Proportion  0.99896 0.99939 0.99967 0.99983 0.99996 0.99998 0.99999
                          PC29     PC30    PC31     PC32      PC33      PC34
Standard deviation     0.01717 0.009399 0.00683 0.004575 0.0003488 2.537e-10
Proportion of Variance 0.00001 0.000000 0.00000 0.000000 0.0000000 0.000e+00
Cumulative Proportion  1.00000 1.000000 1.00000 1.000000 1.0000000 1.000e+00
                            PC35      PC36      PC37      PC38      PC39
Standard deviation     2.431e-10 2.371e-10 2.298e-10 2.275e-10 1.691e-10
Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
                            PC40      PC41      PC42      PC43      PC44
Standard deviation     1.637e-10 1.421e-10 1.265e-10 7.788e-11 5.851e-15
Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.000e+00 1.000e+00
                            PC45      PC46      PC47
Standard deviation     3.131e-15 8.206e-16 2.618e-16
Proportion of Variance 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00
```

Graph 1 – Scree plot for PCA results with all 47 numeric variables, scaled



Output 2 - PCA loadings results with all 47 numeric variables, scaled

```
Loadings:
                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
AGE1
LMED                            -0.34   0.32  -0.45  -0.14
FMR
L30                             -0.11  -0.12                 0.23
L50                             -0.19  -0.20         -0.10   0.39
L80                     -0.10   -0.25  -0.37         -0.28  -0.25
IPOV                                   -0.36   0.11   0.36  -0.13
BEDRMS
VALUE           -0.99   0.16
NUNITS
ROOMS
WEIGHT                                                        0.16
PER
ZINC2           -0.13  -0.72   0.66  -0.15
ZSMHC
UTILITY
OTHERCOST
COST06
COST12
COST08
COSTMED
TOTSAL                  -0.67  -0.69   0.27
GLMED                           -0.34   0.32  -0.45  -0.14
GL30                            -0.11  -0.12                 0.23
GL50                            -0.19  -0.20         -0.10   0.39
GL80                    -0.10   -0.25  -0.37         -0.28  -0.25
APLMED                  -0.14   -0.39  -0.30  -0.20   0.48
ABL30                           -0.12          0.16          0.28
ABL50                           -0.20   0.14   0.27  -0.13   0.46
ABL80                           -0.26   0.15   0.54  -0.37  -0.32
ABLMED                  -0.11   -0.42   0.37   0.38   0.50
BURDEN
INCRELAMIPCT
INCRELPOVPCT
INCRELFMRPCT
COST06RELAMIPCT
COST06RELPOVPCT
COST06RELFMRPCT
COST08RELAMIPCT
COST08RELPOVPCT
COST08RELFMRPCT
COST12RELAMIPCT
COST12RELPOVPCT
COST12RELFMRPCT
COSTMedRELAMIPCT
COSTMedRELPOVPCT
COSTMedRELFMRPCT
```

Output 3 – PCA rotated loadings results with all 47 numeric variables, scaled

Loadings:

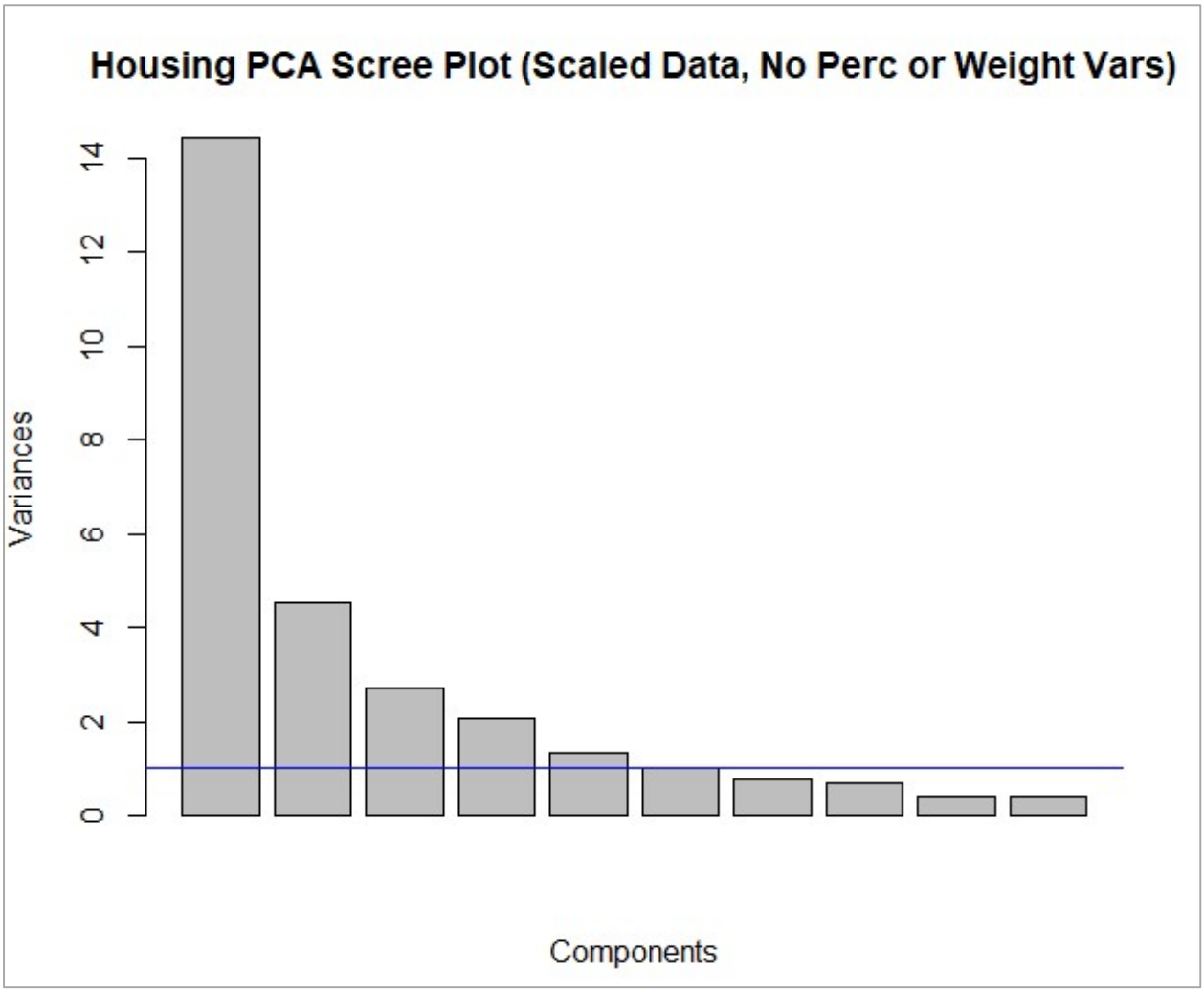
	RC1	RC2	RC4	RC3	RC5	RC6	RC8	RC7
VALUE	0.941							
COST06	0.942							
COST12	0.943							
COST08	0.942							
COSTMED	0.941							
COST06RELAMIPCT	0.974							
COST06RELPOVPCT	0.934							
COST06RELFMRPCT	0.955							
COST08RELAMIPCT	0.975							
COST08RELPOVPCT	0.935							
COST08RELFMRPCT	0.956							
COST12RELAMIPCT	0.975							
COST12RELPOVPCT	0.936							
COST12RELFMRPCT	0.957							
COSTMedRELAMIPCT	0.973							
COSTMedRELPOVPCT	0.934							
COSTMedRELFMRPCT	0.953							
LMED		0.940						
FMR		0.811						
L30		0.802	0.575					
L50		0.802	0.575					
L80		0.727	0.653					
GLMED		0.940						
GL30		0.802	0.575					
GL50		0.802	0.575					
GL80		0.727	0.653					
APLMED		0.805	0.544					
ABL30		0.919						
ABL50		0.919						
ABL80		0.862						
ABLMED		0.909						
IPOV			0.938					
PER			0.937					
ZINC2				0.921				
TOTSAL				0.792				
INCRELAMIPCT				0.940				
INCRELPOVPCT				0.914				
INCRELFMRPCT				0.944				
BEDRMS					0.901			
ROOMS					0.815			
NUNITS						0.810		
OTHERCOST						0.694		
WEIGHT							-0.831	
BURDEN								0.997
AGE1			-0.439				0.478	
ZSMHC	0.488							
UTILITY					0.408			

	RC1	RC2	RC4	RC3	RC5	RC6	RC8	RC7
SS loadings	16.363	10.870	4.945	4.558	2.477	1.266	1.061	1.001
Proportion Var	0.348	0.231	0.105	0.097	0.053	0.027	0.023	0.021
Cumulative Var	0.348	0.579	0.685	0.782	0.834	0.861	0.884	0.905
Proportion Explained	0.385	0.256	0.116	0.107	0.058	0.030	0.025	0.024
Cumulative Proportion	0.385	0.640	0.756	0.864	0.922	0.952	0.976	1.000

Output 4 – PCA results with percentage variables removed, scaled

```
Importance of components%s:
                            PC1    PC2     PC3     PC4     PC5     PC6    PC7
Standard deviation     3.7999 2.1319 1.64424 1.44128 1.15215 1.00965 0.8700
Proportion of Variance 0.4979 0.1567 0.09323 0.07163 0.04577 0.03515 0.0261
Cumulative Proportion  0.4979 0.6546 0.74785 0.81948 0.86525 0.90041 0.9265
                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
Standard deviation     0.83738 0.65006 0.63604 0.46822 0.42222 0.34540 0.21939
Proportion of Variance 0.02418 0.01457 0.01395 0.00756 0.00615 0.00411 0.00166
Cumulative Proportion  0.95069 0.96526 0.97921 0.98677 0.99292 0.99703 0.99869
                          PC15    PC16    PC17    PC18    PC19    PC20
Standard deviation     0.14723 0.09066 0.08126 0.03404 0.01710 0.00461
Proportion of Variance 0.00075 0.00028 0.00023 0.00004 0.00001 0.00000
Cumulative Proportion  0.99944 0.99972 0.99995 0.99999 1.00000 1.00000
                             PC21       PC22       PC23       PC24       PC25
Standard deviation     0.0003485 2.475e-10 1.423e-10 1.285e-10 7.925e-11
Proportion of Variance 0.0000000 0.000e+00 0.000e+00 0.000e+00 0.000e+00
Cumulative Proportion  1.0000000 1.000e+00 1.000e+00 1.000e+00 1.000e+00
                             PC26       PC27       PC28      PC29
Standard deviation     8.688e-14 3.066e-15 6.326e-16 2.87e-16
Proportion of Variance 0.000e+00 0.000e+00 0.000e+00 0.00e+00
Cumulative Proportion  1.000e+00 1.000e+00 1.000e+00 1.00e+00
```

Graph 2 - Scree plot for PCA results with percentage variables removed, scaled



Output 5 - PCA loadings results with percentage & WEIGHT variables removed, scaled

```
Loadings:
          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
LMED                     0.10  -0.33   0.32   0.45   0.14
FMR
L30                            -0.11  -0.12
L50                            -0.19  -0.20
L80                      0.10  -0.25  -0.37          0.28
IPOV                                  -0.36  -0.12  -0.36
BEDRMS
VALUE     -0.99  -0.16
NUNITS
ROOMS
PER
ZINC2     -0.13   0.72  -0.66  -0.16
ZSMHC
UTILITY
OTHERCOST
COST06
COST12
COST08
COSTMED
TOTSAL             0.67   0.68   0.28
GLMED                    0.10  -0.33   0.32   0.45   0.14
GL30                           -0.11  -0.12
GL50                           -0.19  -0.20
GL80                     0.10  -0.25  -0.37          0.28
APLMED                   0.14  -0.39  -0.30   0.20  -0.48
ABL30                          -0.12         -0.16
ABL50                          -0.20   0.14  -0.27   0.13
ABL80                          -0.26   0.15  -0.54   0.38
ABLMED                   0.12  -0.42   0.38  -0.38  -0.49
```

Output 6 – PCA rotated loadings results with percentage & WEIGHT variables removed, scaled

```
Loadings:
          RC1    RC3    RC2    RC4    RC5    RC6
LMED       1.114
FMR        0.758
GLMED      1.114
ABL30      0.940
ABL50      0.940
ABL80      0.846
ABLMED     0.978
L30        0.507  0.634
L50        0.507  0.633
L80               0.739
IPOV      -0.415  1.140
PER       -0.421  1.147
GL30       0.507  0.634
GL50       0.507  0.633
GL80               0.739
APLMED     0.559  0.578
VALUE                1.042
COST06               1.032
COST12               1.036
COST08               1.034
COSTMED              1.029
BEDRMS                      0.964
ROOMS                       0.850
ZINC2                              0.940
TOTSAL                             0.985
NUNITS                                    0.876
OTHERCOST                                 0.670
ZSMHC                0.424
UTILITY

                 RC1   RC3   RC2   RC4   RC5   RC6
SS loadings    8.477 5.839 5.745 2.395 1.965 1.287
Proportion Var 0.292 0.201 0.198 0.083 0.068 0.044
Cumulative Var 0.292 0.494 0.692 0.774 0.842 0.886
```

Output 7 – MSA tests: KMO & Bartlett's Test of Sphericity for PCA

```
Overall MSA =  0.5
MSA for each item =
            AGE1             LMED              FMR              L30
             0.5              0.5              0.5              0.5
             L50              L80             IPOV          BEDRMS
             0.5              0.5              0.5              0.5
           VALUE           NUNITS            ROOMS          WEIGHT
             0.5              0.5              0.5              0.5
             PER            ZINC2            ZSMHC         UTILITY
             0.5              0.5              0.5              0.5
       OTHERCOST           COST06           COST12           COST08
             0.5              0.5              0.5              0.5
         COSTMED           TOTSAL            GLMED             GL30
             0.5              0.5              0.5              0.5
            GL50             GL80           APLMED            ABL30
             0.5              0.5              0.5              0.5
           ABL50            ABL80           ABLMED           BURDEN
             0.5              0.5              0.5              0.5
    INCRELAMIPCT     INCRELPOVPCT     INCRELFMRPCT  COST06RELAMIPCT
             0.5              0.5              0.5              0.5
 COST06RELPOVPCT  COST06RELFMRPCT  COST08RELAMIPCT  COST08RELPOVPCT
             0.5              0.5              0.5              0.5
 COST08RELFMRPCT  COST12RELAMIPCT  COST12RELPOVPCT  COST12RELFMRPCT
             0.5              0.5              0.5              0.5
COSTMedRELAMIPCT COSTMedRELPOVPCT COSTMedRELFMRPCT
             0.5              0.5              0.5
```

```
> cortest.bartlett(housing.cor, n = 35852,diag=TRUE)
$chisq
[1] Inf

$p.value
[1] 0

$df
[1] 1081
```

Output 8 – Correlation matrix for all numeric variables except percentages

	LMED	FMR	L30	L50	L80
LMED	1.00000000	0.65569662	0.72020562	0.72031050	0.624196475
FMR	0.65569662	1.00000000	0.69671179	0.69669047	0.656319172
L30	0.72020562	0.69671179	1.00000000	0.99999341	0.978922214
L50	0.72031050	0.69669047	0.99999341	1.00000000	0.978924936
L80	0.62419647	0.65631917	0.97892221	0.97892494	1.000000000
IPOV	0.06362384	0.22317380	0.67535363	0.67513846	0.737911496
BEDRMS	0.07337018	0.48504695	0.26212115	0.26197522	0.278154979
VALUE	0.29234292	0.46447249	0.32337374	0.32336209	0.309235925
NUNITS	0.03406063	-0.02341883	0.01017022	0.01024791	0.007644454
ROOMS	0.10938605	0.37492896	0.24185606	0.24173760	0.250774704
WEIGHT	-0.29739697	-0.15634111	-0.21912996	-0.21961918	-0.218415712
PER	0.06190327	0.22665271	0.67901174	0.67879507	0.742027743
ZINC2	0.17692928	0.26316355	0.29714868	0.29714187	0.299269280
ZSMHC	0.32790053	0.44602909	0.44326680	0.44321941	0.435314235
UTILITY	0.18068680	0.32599847	0.33381311	0.33400163	0.337642115
OTHERCOST	0.09311435	0.10220817	0.07218353	0.07221924	0.066520908
COST06	0.29877505	0.47300577	0.33611077	0.33611470	0.322345675
COST12	0.29682934	0.47051465	0.33195713	0.33195550	0.318042716
COST08	0.29801643	0.47204879	0.33444357	0.33444520	0.320614456
COSTMED	0.29941493	0.47379393	0.33758040	0.33758646	0.323877022
TOTSAL	0.16837218	0.23902153	0.33168247	0.33165749	0.338904885
GLMED	1.00000000	0.65569662	0.72020562	0.72031050	0.624196475
GL30	0.72020562	0.69671179	1.00000000	0.99999341	0.978922214
GL50	0.72031050	0.69669047	0.99999341	1.00000000	0.978924936
GL80	0.62419647	0.65631917	0.97892221	0.97892494	1.000000000
APLMED	0.77010627	0.62736782	0.96574832	0.96568454	0.932312173
ABL30	0.82055958	0.88589181	0.77173388	0.77181065	0.710363164
ABL50	0.82069630	0.88578429	0.77206414	0.77215600	0.710737999
ABL80	0.73585982	0.87766174	0.73709606	0.73719126	0.718908392
ABLMED	0.86379852	0.79428903	0.73014964	0.73016538	0.657189995

	IPOV	BEDRMS	VALUE	NUNITS	ROOMS
LMED	0.06362384	0.07337018	0.29234292	0.034060627	0.10938605
FMR	0.22317380	0.48504695	0.46447249	-0.023418826	0.37492896
L30	0.67535363	0.26212115	0.32337374	0.010170224	0.24185606
L50	0.67513846	0.26197522	0.32336209	0.010247910	0.24173760
L80	0.73791150	0.27815498	0.30923593	0.007644454	0.25077470
IPOV	1.00000000	0.34815143	0.08343373	-0.065343528	0.28009022
BEDRMS	0.34815143	1.00000000	0.27496561	-0.167773247	0.77601822
VALUE	0.08343373	0.27496561	1.00000000	0.033348672	0.33872707
NUNITS	-0.06534353	-0.16777325	0.03334867	1.000000000	-0.14448181
ROOMS	0.28009022	0.77601822	0.33872707	-0.144481806	1.00000000
WEIGHT	-0.01990160	-0.02442890	-0.07100953	-0.043688586	-0.04536781
PER	0.98960751	0.35495776	0.08936476	-0.067546447	0.29091090
ZINC2	0.24088270	0.26898381	0.40588023	0.006555490	0.34926497
ZSMHC	0.27484782	0.31320176	0.58833863	0.016993840	0.36655293
UTILITY	0.29840870	0.37391558	0.27904737	-0.137311021	0.40077729
OTHERCOST	-0.03493426	-0.05341976	0.25368842	0.249247514	-0.01032280
COST06	0.09520373	0.28108434	0.99488077	0.045715114	0.34751858
COST12	0.09111198	0.27922975	0.99787996	0.041352092	0.34478756
COST08	0.09352407	0.28036066	0.99628029	0.043915276	0.34644240
COSTMED	0.09673251	0.28169553	0.99339448	0.047364409	0.34844141
TOTSAL	0.31861243	0.24500159	0.29200059	-0.001439194	0.28937549
GLMED	0.06362384	0.07337018	0.29234292	0.034060627	0.10938605
GL30	0.67535363	0.26212115	0.32337374	0.010170224	0.24185606
GL50	0.67513846	0.26197522	0.32336209	0.010247910	0.24173760
GL80	0.73791150	0.27815498	0.30923593	0.007644454	0.25077470
APLMED	0.66700746	0.28167572	0.27415054	-0.022100351	0.26771580
ABL30	0.23046084	0.55058709	0.44559980	-0.033447525	0.45718672
ABL50	0.23071219	0.55019119	0.44551237	-0.033333193	0.45688785
ABL80	0.25549489	0.61540687	0.45125870	-0.041968920	0.50299496
ABLMED	0.22537019	0.55638085	0.38293669	-0.064762716	0.47356053

	WEIGHT	PER	ZINC2	ZSMHC	UTILITY
LMED	-0.29739697	0.06190327	0.17692928	0.32790053	0.18068680
FMR	-0.15634111	0.22665271	0.26316355	0.44602909	0.32599847
L30	-0.21912996	0.67901174	0.29714868	0.44326680	0.33381311
L50	-0.21961918	0.67879507	0.29714187	0.44321941	0.33400163
L80	-0.21841571	0.74202774	0.29926928	0.43531423	0.33764211
IPOV	-0.01990160	0.98960751	0.24088270	0.27484782	0.29840870
BEDRMS	-0.02442890	0.35495776	0.26898381	0.31320176	0.37391558
VALUE	-0.07100953	0.08936476	0.40588023	0.58833863	0.27904737
NUNITS	-0.04368859	-0.06754645	0.00655549	0.01699384	-0.13731102
ROOMS	-0.04536781	0.29091090	0.34926497	0.36655293	0.40077729
WEIGHT	1.00000000	-0.02099777	-0.05511017	-0.11793506	-0.08129991
PER	-0.02099777	1.00000000	0.24177815	0.26765797	0.30642289
ZINC2	-0.05511017	0.24177815	1.00000000	0.46382790	0.25562975
ZSMHC	-0.11793506	0.26765797	0.46382790	1.00000000	0.39331919
UTILITY	-0.08129991	0.30642289	0.25562975	0.39331919	1.00000000
OTHERCOST	-0.04058178	-0.03466755	0.13089724	0.29850223	-0.02117469
COST06	-0.07654070	0.10139163	0.41525589	0.61317575	0.32717120
COST12	-0.07465525	0.09721541	0.41239594	0.60502648	0.31036823
COST08	-0.07577190	0.09967796	0.41413702	0.60989746	0.32026352
COSTMED	-0.07723393	0.10295051	0.41620464	0.61607501	0.33347109
TOTSAL	-0.04981412	0.30465051	0.80492047	0.44971751	0.20941998
GLMED	-0.29739697	0.06190327	0.17692928	0.32790053	0.18068680
GL30	-0.21912996	0.67901174	0.29714868	0.44326680	0.33381311
GL50	-0.21961918	0.67879507	0.29714187	0.44321941	0.33400163
GL80	-0.21841571	0.74202774	0.29926928	0.43531423	0.33764211
APLMED	-0.23143832	0.67284914	0.29108529	0.41802427	0.33103247
ABL30	-0.24677546	0.23165137	0.29049441	0.46081246	0.34954106
ABL50	-0.24732768	0.23191046	0.29053012	0.46079184	0.34979414
ABL80	-0.25497270	0.25689696	0.29890547	0.46336196	0.36274851
ABLMED	-0.25790821	0.22773088	0.28070358	0.42791697	0.34018207

	OTHERCOST	COST06	COST12	COST08	COSTMED
LMED	0.09311435	0.29877505	0.29682934	0.29801643	0.29941493
FMR	0.10220817	0.47300577	0.47051465	0.47204879	0.47379393
L30	0.07218353	0.33611077	0.33195713	0.33444357	0.33758040
L50	0.07221924	0.33611470	0.33195550	0.33444520	0.33758646
L80	0.06652091	0.32234568	0.31804272	0.32061446	0.32387702
IPOV	-0.03493426	0.09520373	0.09111198	0.09352407	0.09673251
BEDRMS	-0.05341976	0.28108434	0.27922975	0.28036066	0.28169553
VALUE	0.25368842	0.99488077	0.99787996	0.99628029	0.99339448
NUNITS	0.24924751	0.04571511	0.04135209	0.04391528	0.04736441
ROOMS	-0.01032280	0.34751858	0.34478756	0.34644240	0.34844141
WEIGHT	-0.04058178	-0.07654070	-0.07465525	-0.07577190	-0.07723393
PER	-0.03466755	0.10139163	0.09721541	0.09967796	0.10295051
ZINC2	0.13089724	0.41525589	0.41239594	0.41413702	0.41620464
ZSMHC	0.29850223	0.61317575	0.60502648	0.60989746	0.61607501
UTILITY	-0.02117469	0.32717120	0.31036823	0.32026352	0.33347109
OTHERCOST	1.00000000	0.33108007	0.30382836	0.31984537	0.34136603
COST06	0.33108007	1.00000000	0.99934843	0.99988826	0.99990514
COST12	0.30382836	0.99934843	1.00000000	0.99977632	0.99875649
COST08	0.31984537	0.99988826	0.99977632	1.00000000	0.99958750
COSTMED	0.34136603	0.99990514	0.99875649	0.99958750	1.00000000
TOTSAL	0.08813055	0.29980113	0.29736789	0.29884074	0.30062671
GLMED	0.09311435	0.29877505	0.29682934	0.29801643	0.29941493
GL30	0.07218353	0.33611077	0.33195713	0.33444357	0.33758040
GL50	0.07221924	0.33611470	0.33195550	0.33444520	0.33758646
GL80	0.06652091	0.32234568	0.31804272	0.32061446	0.32387702
APLMED	0.04305585	0.28632474	0.28231353	0.28470866	0.28775713
ABL30	0.07459346	0.45397656	0.45151895	0.45303046	0.45475848
ABL50	0.07463127	0.45391241	0.45144641	0.45296282	0.45469755
ABL80	0.06995657	0.45978789	0.45728268	0.45882299	0.46058598
ABLMED	0.04271604	0.39070157	0.38838803	0.38980513	0.39145024

	TOTSAL	GLMED	GL30	GL50	GL80
LMED	0.168372176	1.00000000	0.72020562	0.72031050	0.624196475
FMR	0.239021526	0.65569662	0.69671179	0.69669047	0.656319172
L30	0.331682469	0.72020562	1.00000000	0.99999341	0.978922214
L50	0.331657489	0.72031050	0.99999341	1.00000000	0.978924936
L80	0.338904885	0.62419647	0.97892221	0.97892494	1.000000000
IPOV	0.318612430	0.06362384	0.67535363	0.67513846	0.737911496
BEDRMS	0.245001585	0.07337018	0.26212115	0.26197522	0.278154979
VALUE	0.292000594	0.29234292	0.32337374	0.32336209	0.309235925
NUNITS	-0.001439194	0.03406063	0.01017022	0.01024791	0.007644454
ROOMS	0.289375487	0.10938605	0.24185606	0.24173760	0.250774704
WEIGHT	-0.049814117	-0.29739697	-0.21912996	-0.21961918	-0.218415712
PER	0.304650514	0.06190327	0.67901174	0.67879507	0.742027743
ZINC2	0.804920469	0.17692928	0.29714868	0.29714187	0.299269280
ZSMHC	0.449717506	0.32790053	0.44326680	0.44321941	0.435314235
UTILITY	0.209419982	0.18068680	0.33381311	0.33400163	0.337642115
OTHERCOST	0.088130550	0.09311435	0.07218353	0.07221924	0.066520908
COST06	0.299801129	0.29877505	0.33611077	0.33611470	0.322345675
COST12	0.297367888	0.29682934	0.33195713	0.33195550	0.318042716
COST08	0.298840740	0.29801643	0.33444357	0.33444520	0.320614456
COSTMED	0.300626706	0.29941493	0.33758040	0.33758646	0.323877022
TOTSAL	1.000000000	0.16837218	0.33168247	0.33165749	0.338904885
GLMED	0.168372176	1.00000000	0.72020562	0.72031050	0.624196475
GL30	0.331682469	0.72020562	1.00000000	0.99999341	0.978922214
GL50	0.331657489	0.72031050	0.99999341	1.00000000	0.978924936
GL80	0.338904885	0.62419647	0.97892221	0.97892494	1.000000000
APLMED	0.326310736	0.77010627	0.96574832	0.96568454	0.932312173
ABL30	0.268781078	0.82055958	0.77173388	0.77181065	0.710363164
ABL50	0.268849447	0.82069630	0.77206414	0.77215600	0.710737999
ABL80	0.276483445	0.73585982	0.73709606	0.73719126	0.718908392
ABLMED	0.261605030	0.86379852	0.73014964	0.73016538	0.657189995

	APLMED	ABL30	ABL50	ABL80	ABLMED
LMED	0.77010627	0.82055958	0.82069630	0.73585982	0.86379852
FMR	0.62736782	0.88589181	0.88578429	0.87766174	0.79428903
L30	0.96574832	0.77173388	0.77206414	0.73709606	0.73014964
L50	0.96568454	0.77181065	0.77215600	0.73719126	0.73016538
L80	0.93231217	0.71036316	0.71073800	0.71890839	0.65718999
IPOV	0.66700746	0.23046084	0.23071219	0.25549489	0.22537019
BEDRMS	0.28167572	0.55058709	0.55019119	0.61540687	0.55638085
VALUE	0.27415054	0.44559980	0.44551237	0.45125870	0.38293669
NUNITS	-0.02210035	-0.03344753	-0.03333319	-0.04196892	-0.06476272
ROOMS	0.26771580	0.45718672	0.45688785	0.50299496	0.47356053
WEIGHT	-0.23143832	-0.24677546	-0.24732768	-0.25497270	-0.25790821
PER	0.67284914	0.23165137	0.23191046	0.25689696	0.22773088
ZINC2	0.29108529	0.29049441	0.29053012	0.29890547	0.28070358
ZSMHC	0.41802427	0.46081246	0.46079184	0.46336196	0.42791697
UTILITY	0.33103247	0.34954106	0.34979414	0.36274851	0.34018207
OTHERCOST	0.04305585	0.07459346	0.07463127	0.06995657	0.04271604
COST06	0.28632474	0.45397656	0.45391241	0.45978789	0.39070157
COST12	0.28231353	0.45151895	0.45144641	0.45728268	0.38838803
COST08	0.28470866	0.45303046	0.45296282	0.45882299	0.38980513
COSTMED	0.28775713	0.45475848	0.45469755	0.46058598	0.39145024
TOTSAL	0.32631074	0.26878108	0.26884945	0.27648345	0.26160503
GLMED	0.77010627	0.82055958	0.82069630	0.73585982	0.86379852
GL30	0.96574832	0.77173388	0.77206414	0.73709606	0.73014964
GL50	0.96568454	0.77181065	0.77215600	0.73719126	0.73016538
GL80	0.93231217	0.71036316	0.71073800	0.71890839	0.65718999
APLMED	1.00000000	0.75131523	0.75157069	0.70390893	0.78123778
ABL30	0.75131523	1.00000000	0.99999093	0.97378677	0.96116711
ABL50	0.75157069	0.99999093	1.00000000	0.97378533	0.96108547
ABL80	0.70390893	0.97378677	0.97378533	1.00000000	0.92108124
ABLMED	0.78123778	0.96116711	0.96108547	0.92108124	1.00000000

Output 9 – LDA output for pared down set

```
Call:
lda(REGION ~ LMED + FMR + IPOV + BEDRMS + VALUE + NUNITS + ROOMS +
    WEIGHT + ZINC2 + ZSMHC + UTILITY + OTHERCOST + TOTSAL + APLMED,
    data = training)

Prior probabilities of groups:
      '1'       '2'       '3'       '4'
0.2485217 0.2902488 0.3007364 0.1604931

Group means:
        LMED      FMR     IPOV   BEDRMS    VALUE   NUNITS    ROOMS
'1' 79829.85 1488.916 17791.15 3.140741 318653.2 5.473850 6.615264
'2' 65646.05 1038.757 17422.41 3.132616 175214.3 3.655391 6.536421
'3' 60681.67 1132.164 17182.21 3.147468 194667.0 3.075682 6.495641
'4' 68798.17 1549.448 17912.48 3.229753 369603.8 2.894334 6.605839
      WEIGHT    ZINC2    ZSMHC  UTILITY OTHERCOST   TOTSAL   APLMED
'1' 1539.715 96301.56 1647.958 284.3851 123.94010 70980.48 68832.50
'2' 1737.072 79384.73 1147.584 216.9104  97.31527 57011.37 56058.10
'3' 2634.971 75102.73 1093.087 231.4352 111.76839 54120.30 51484.44
'4' 2635.735 93832.55 1559.335 235.3505 126.99366 62938.52 59496.09

Coefficients of linear discriminants:
                    LD1           LD2           LD3
LMED       -5.179318e-05  8.695368e-05  1.720052e-05
FMR         4.681767e-03  1.570396e-04  3.770067e-04
IPOV        3.499035e-05  7.751850e-06  1.101024e-04
BEDRMS     -9.043355e-01 -1.802648e-01  1.012114e-01
VALUE       3.521044e-07  1.280425e-07  1.296134e-06
NUNITS     -1.648624e-03  1.188715e-03 -4.532729e-03
ROOMS       4.247929e-02 -3.914970e-02  1.378391e-03
WEIGHT      4.224993e-04 -2.395174e-04 -3.431555e-04
ZINC2       5.750569e-07  3.180146e-08  4.911930e-06
ZSMHC      -4.853864e-05  6.866666e-06  2.471138e-04
UTILITY    -2.455751e-04  1.885820e-03 -7.888664e-03
OTHERCOST -1.467220e-05 -4.838443e-04 -1.536632e-03
TOTSAL     -9.570873e-07 -2.078525e-07 -6.256985e-06
APLMED     -2.493489e-05 -5.872264e-06 -3.932378e-05

Proportion of trace:
   LD1    LD2    LD3
0.5290 0.4498 0.0212
```

Output 10 – Accuracy of LDA model on 50/50 training and test data

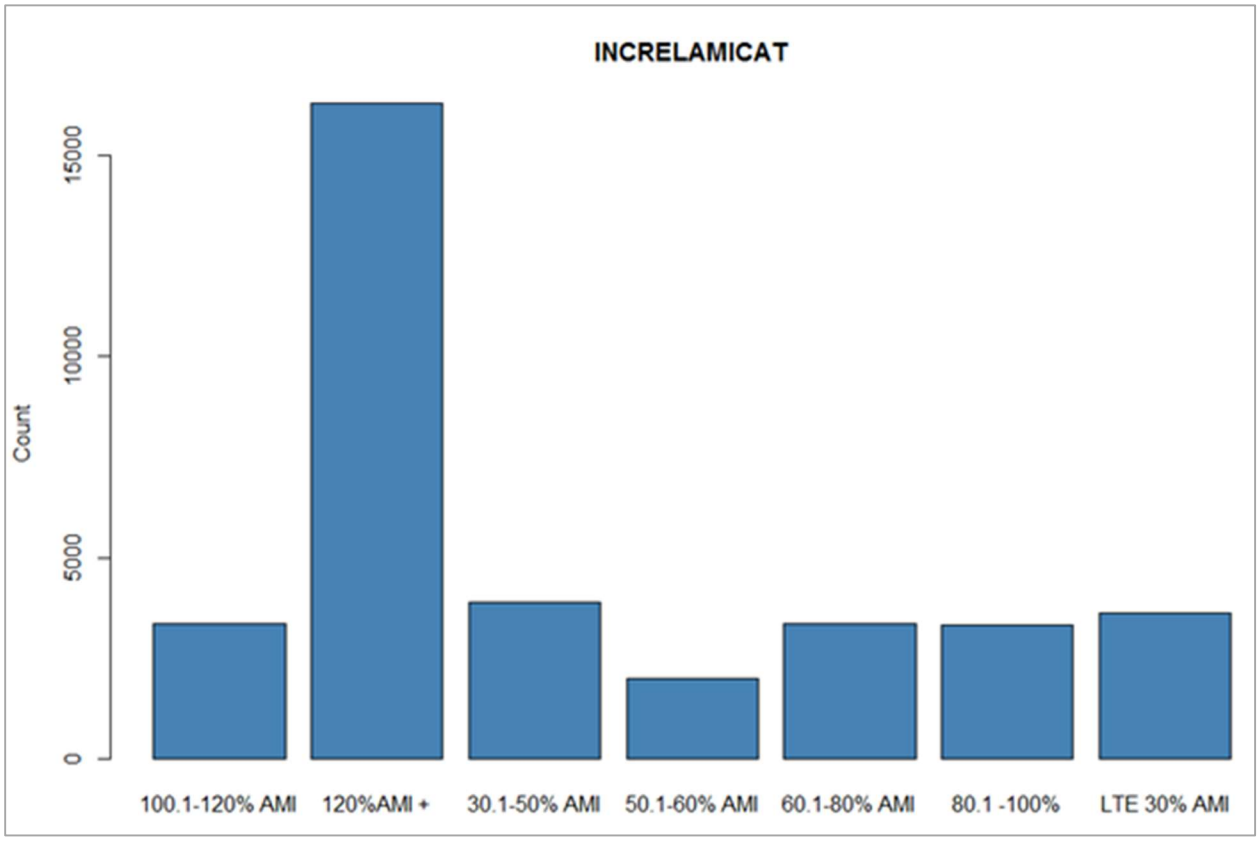
```
> table(housingtrainclass, training[,15])

housingtrainclass  '1'  '2'  '3'  '4'
              '1' 2835  217  433  324
              '2'  652 4200  678  212
              '3'  619  782 4019 1052
              '4'  349    4  261 1289
> table(housingtestclass, test[,15])

housingtestclass  '1'  '2'  '3'  '4'
             '1' 2754  221  420  318
             '2'  629 4123  720  226
             '3'  584  816 4071 1086
             '4'  374    0  269 1315
```

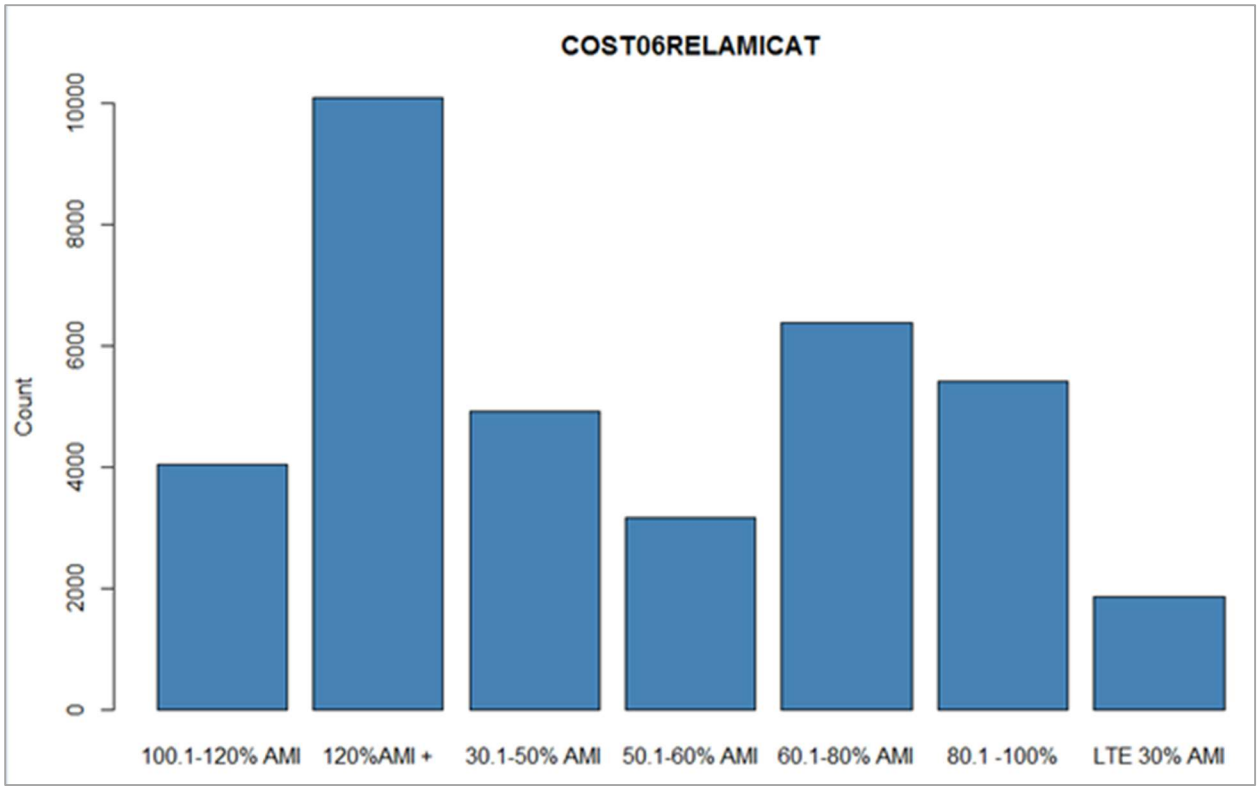
housingtrainclass	'1'	'2'	'3'	'4'	% correct
'1'	2835	217	433	324	74%
'2'	652	4200	678	212	73%
'3'	619	782	4019	1052	62%
'4'	349	4	261	1289	68%

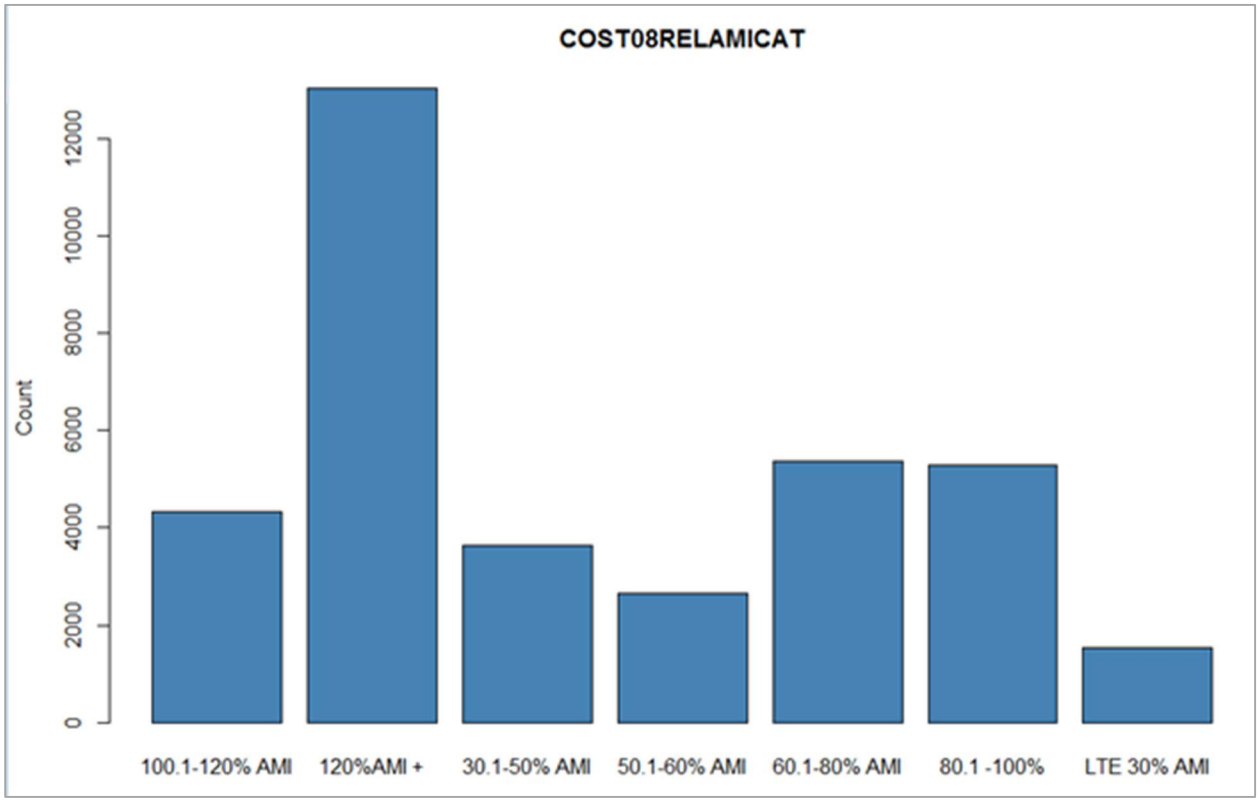
housingtestclass	'1'	'2'	'3'	'4'	% correct
'1'	2754	221	420	318	74%
'2'	629	4123	720	226	72%
'3'	584	816	4071	1086	62%
'4'	374	0	269	1315	67%

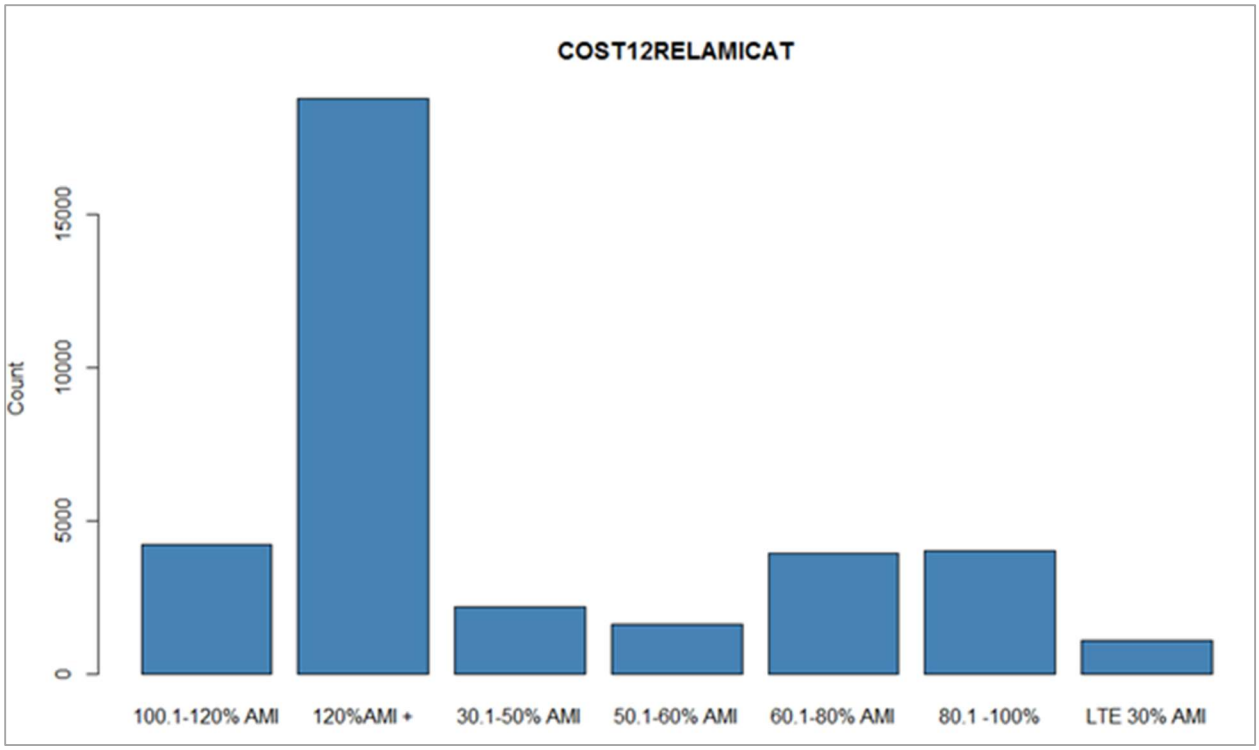
Output 11 MCA - Categorical variable summary and variable bar plots

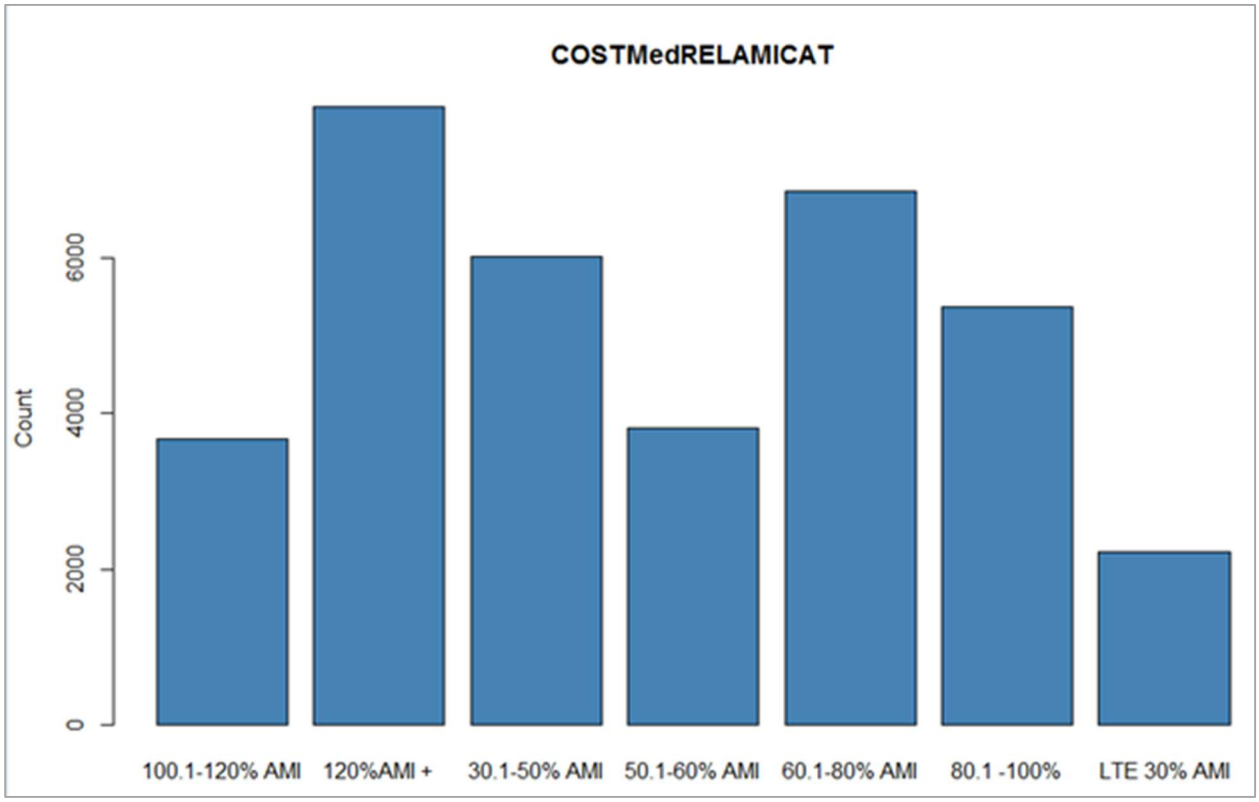
METRO3	REGION	TYPE	ZADEQ	STRUCTURETYPE
5	4	6	3	6
INCRELAMICAT	INCRELPOVCAT	INCRELFMRCAT	COST06RELAMICAT	COST06RELPOVCAT
7	4	3	7	4
COST08RELAMICAT	COST08RELPOVCAT	COST08RELFMRCAT	COST06RELFMRCAT	COST12RELAMICAT
7	4	3	3	7
COST12RELPOVCAT	COST12RELFMRCAT	COSTMedRELAMICAT	COSTMedRELPOVCAT	COSTMedRELFMRCAT
4	3	7	4	3

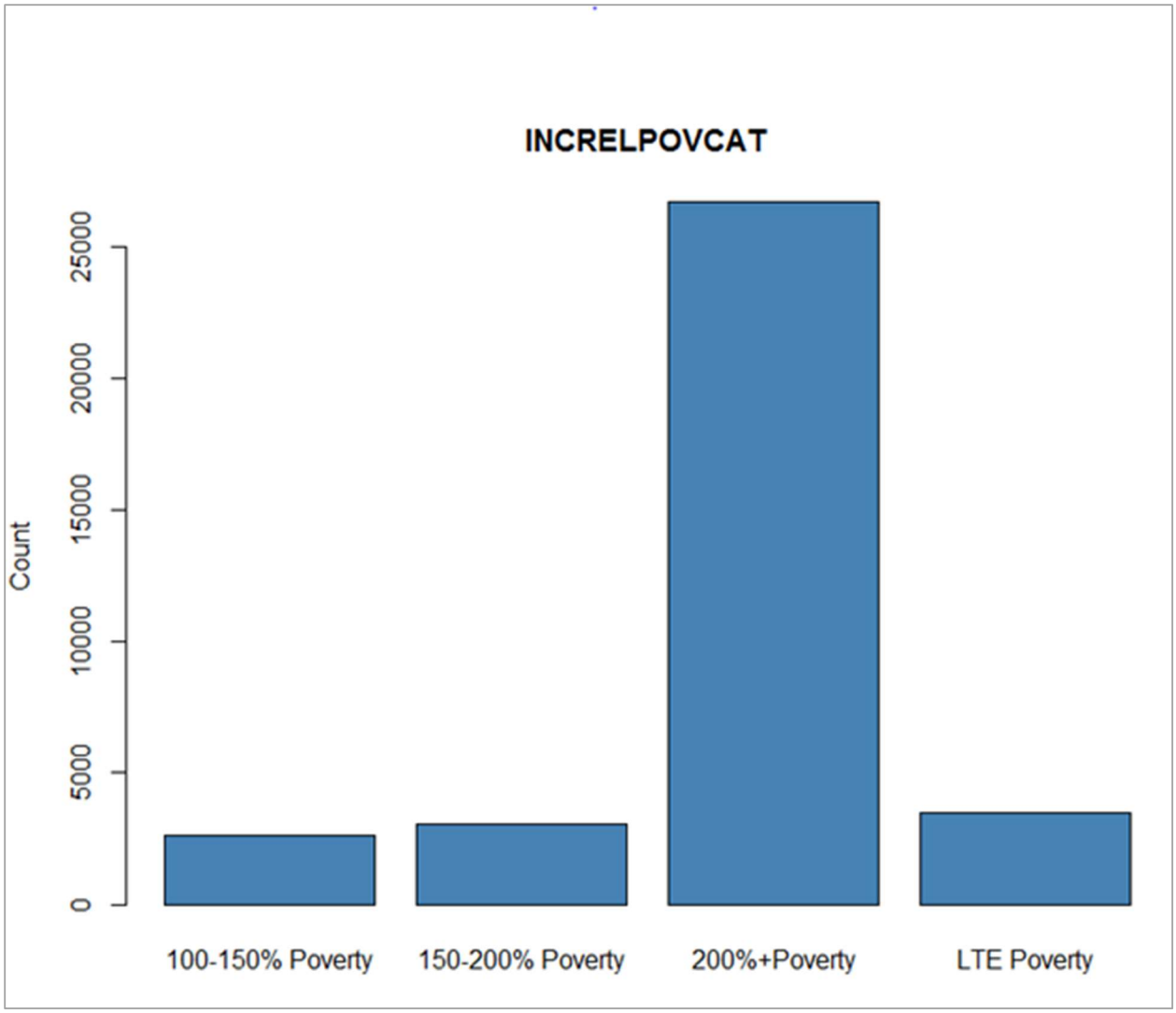


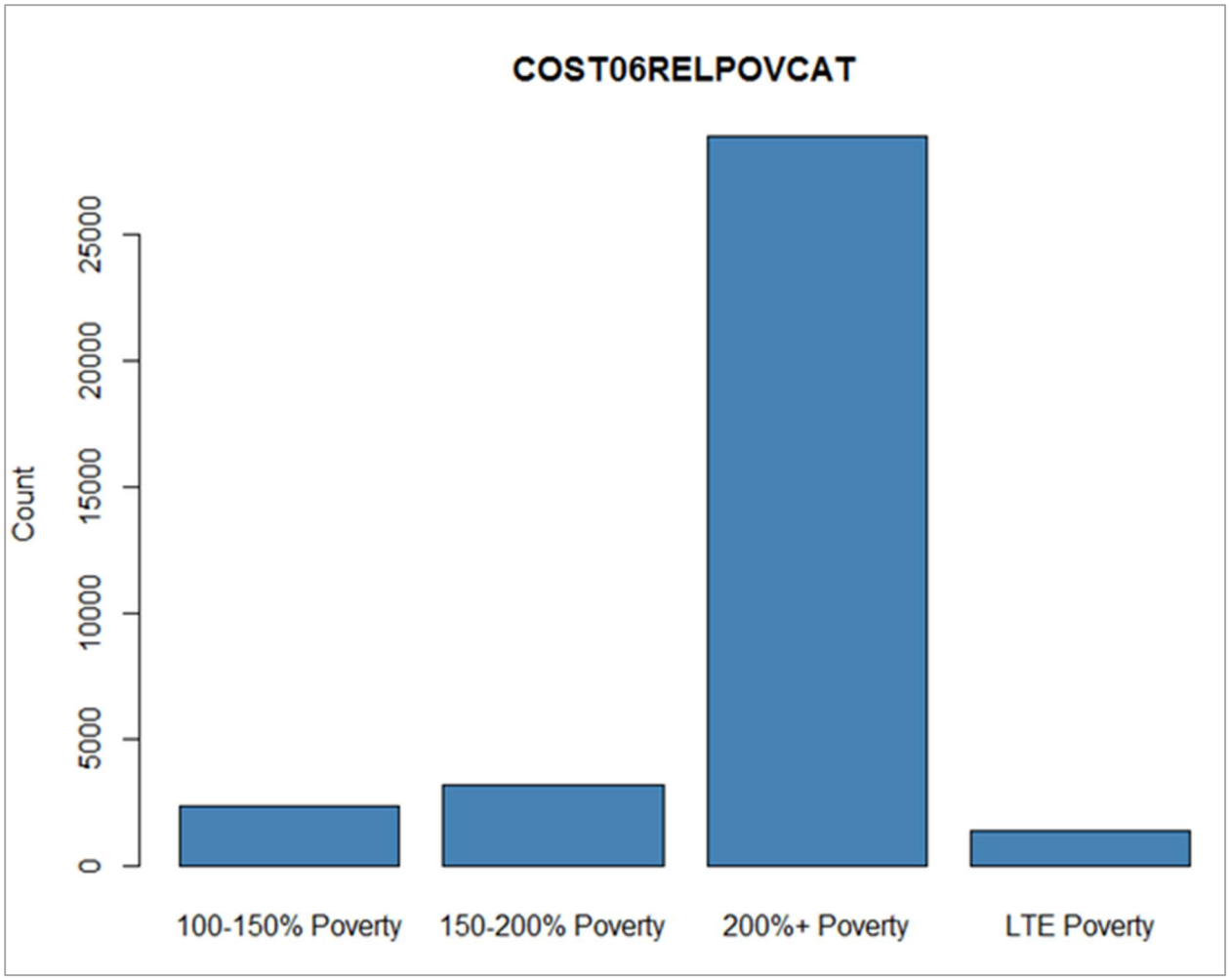

COST06RELAMICAT
Count
0
2000
4000
6000
8000
10000
100.1-120% AMI
120%AMI +
30.1-50% AMI
50.1-60% AMI
60.1-80% AMI
80.1 -100%
LTE 30% AMI

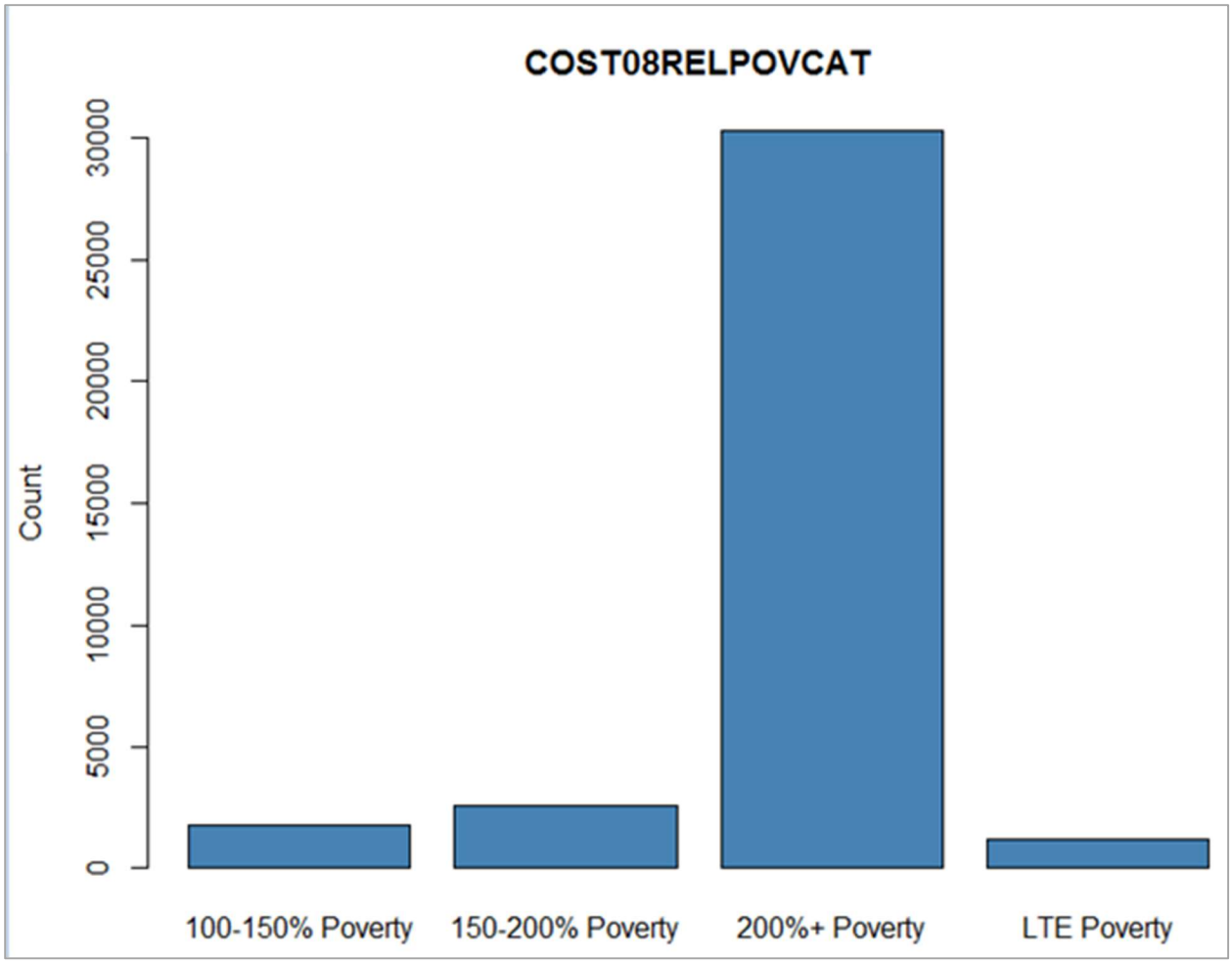

COST08RELAMICAT
Count
0
2000
4000
6000
8000
10000
12000
100.1-120% AMI
120%AMI +
30.1-50% AMI
50.1-60% AMI
60.1-80% AMI
80.1 -100%
LTE 30% AMI

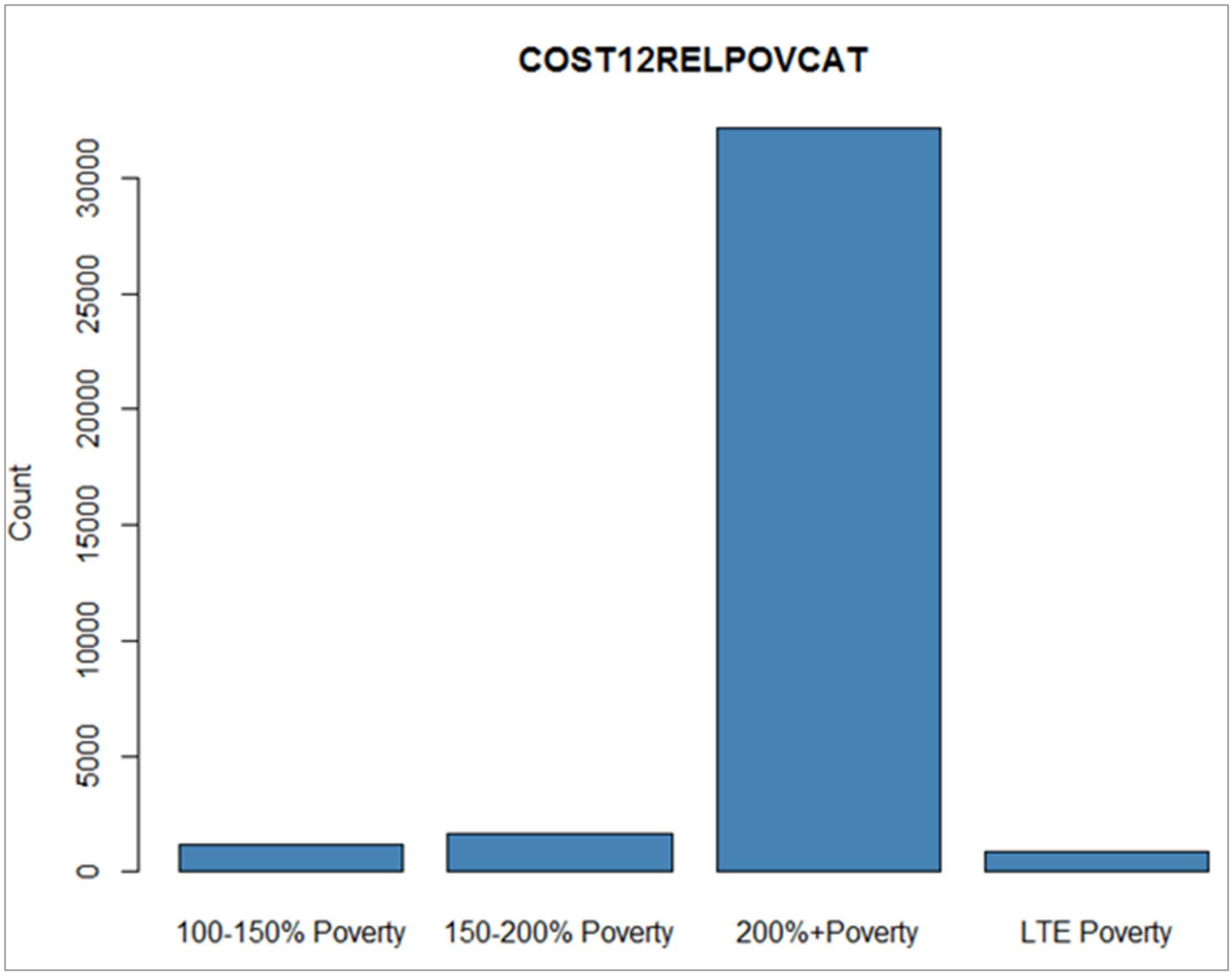

COST12RELAMICAT
Count
0
5000
10000
15000
100.1-120% AMI
120%AMI +
30.1-50% AMI
50.1-60% AMI
60.1-80% AMI
80.1 -100%
LTE 30% AMI

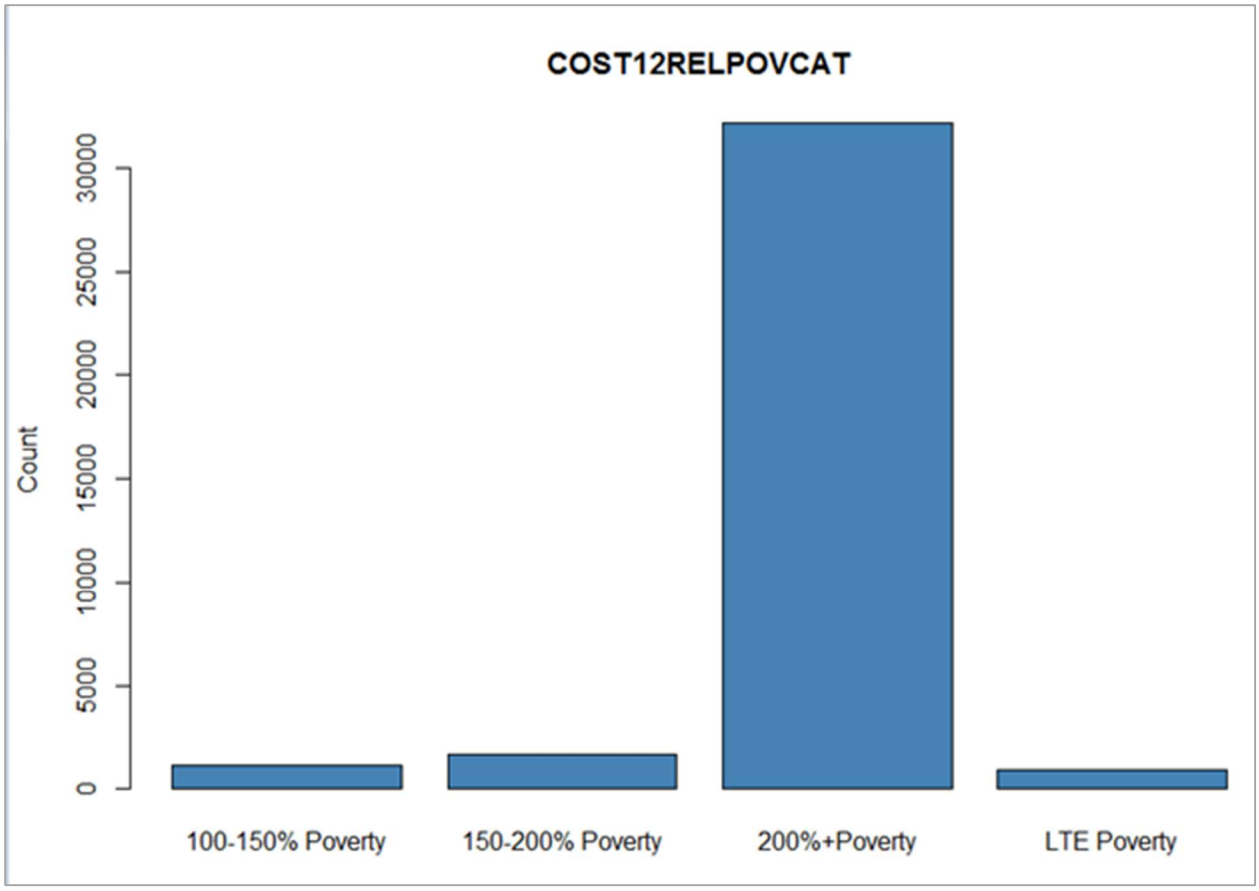

COSTMedRELAMICAT
Count
0
2000
4000
6000
100.1-120% AMI
120%AMI +
30.1-50% AMI
50.1-60% AMI
60.1-80% AMI
80.1 -100%
LTE 30% AMI

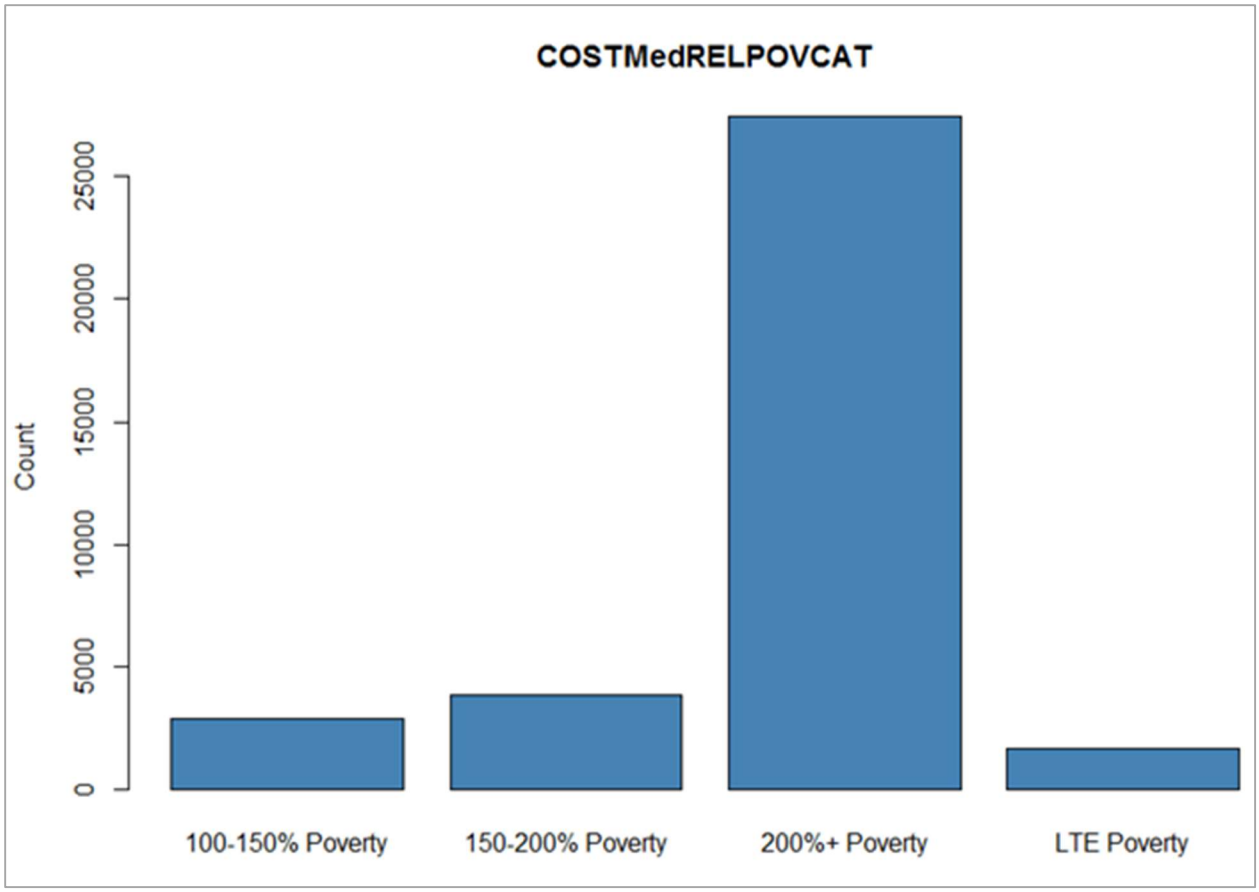

INCRELPOVCAT
Count
0
5000
10000
15000
20000
25000
100-150% Poverty
150-200% Poverty
200%+Poverty
LTE Poverty

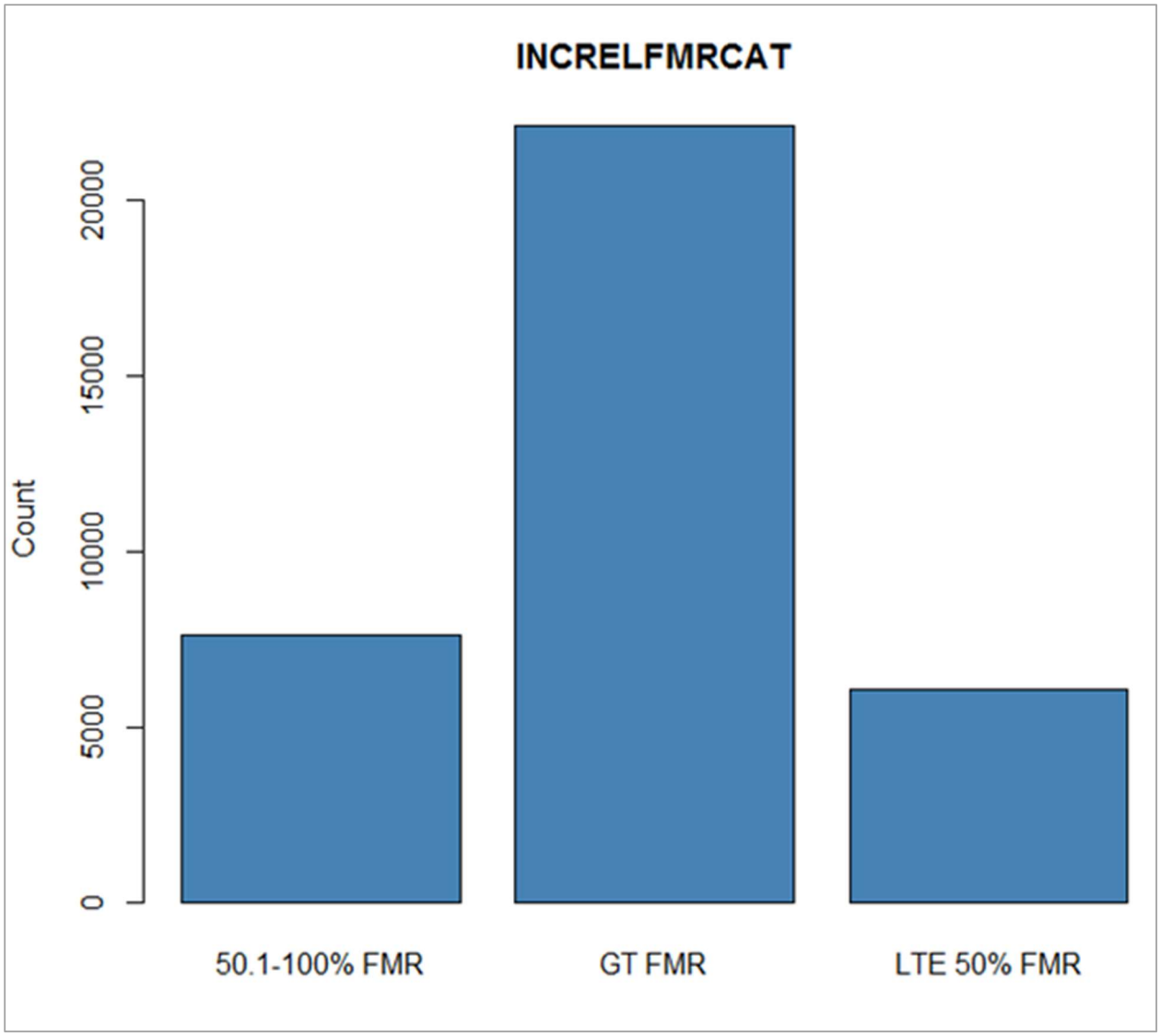

COST06RELPOVCAT
Count
0
5000
10000
15000
20000
25000
100-150% Poverty
150-200% Poverty
200%+ Poverty
LTE Poverty

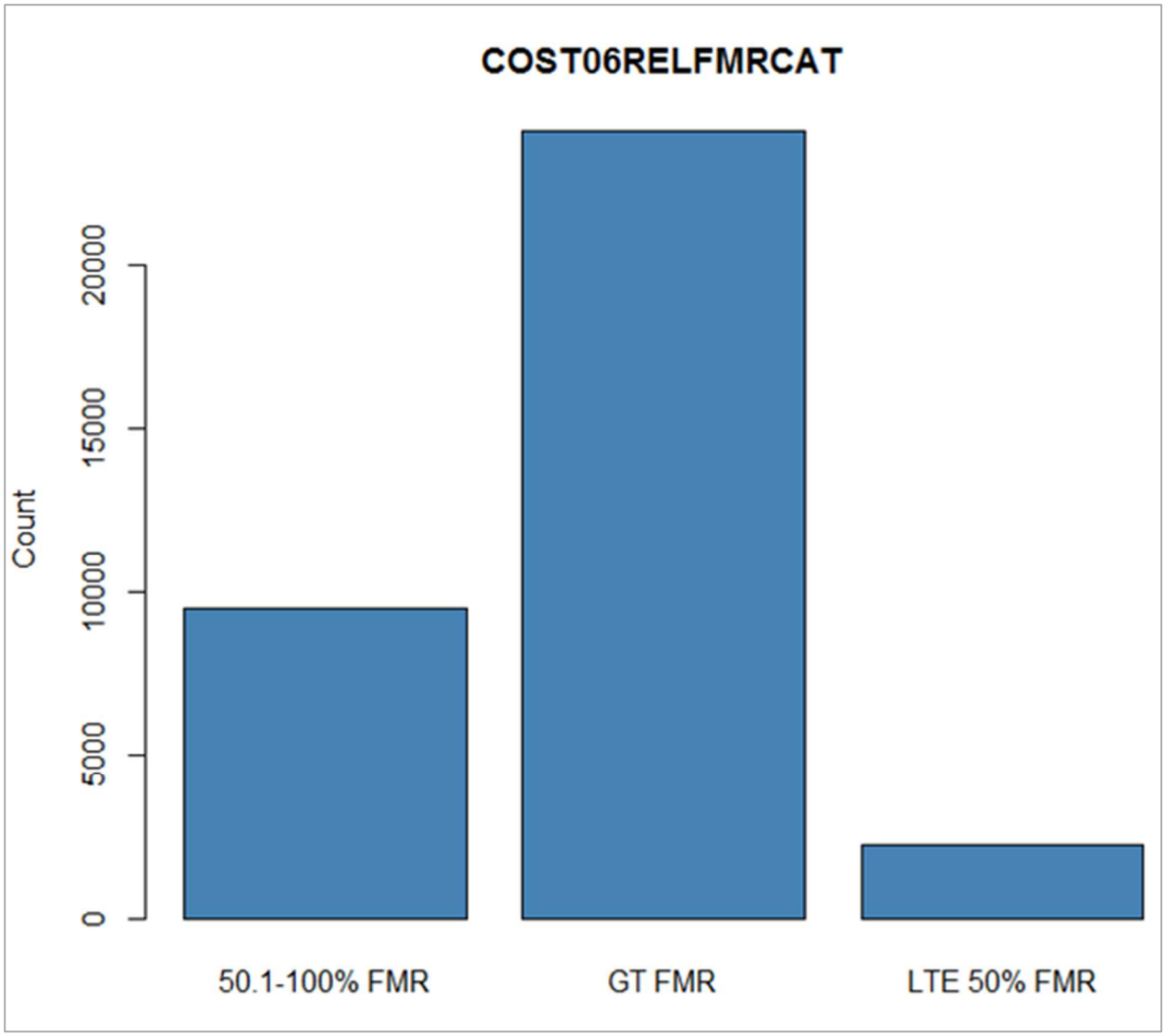

COST08RELPOVCAT
Count
0
5000
10000
15000
20000
25000
30000
100-150% Poverty
150-200% Poverty
200%+ Poverty
LTE Poverty

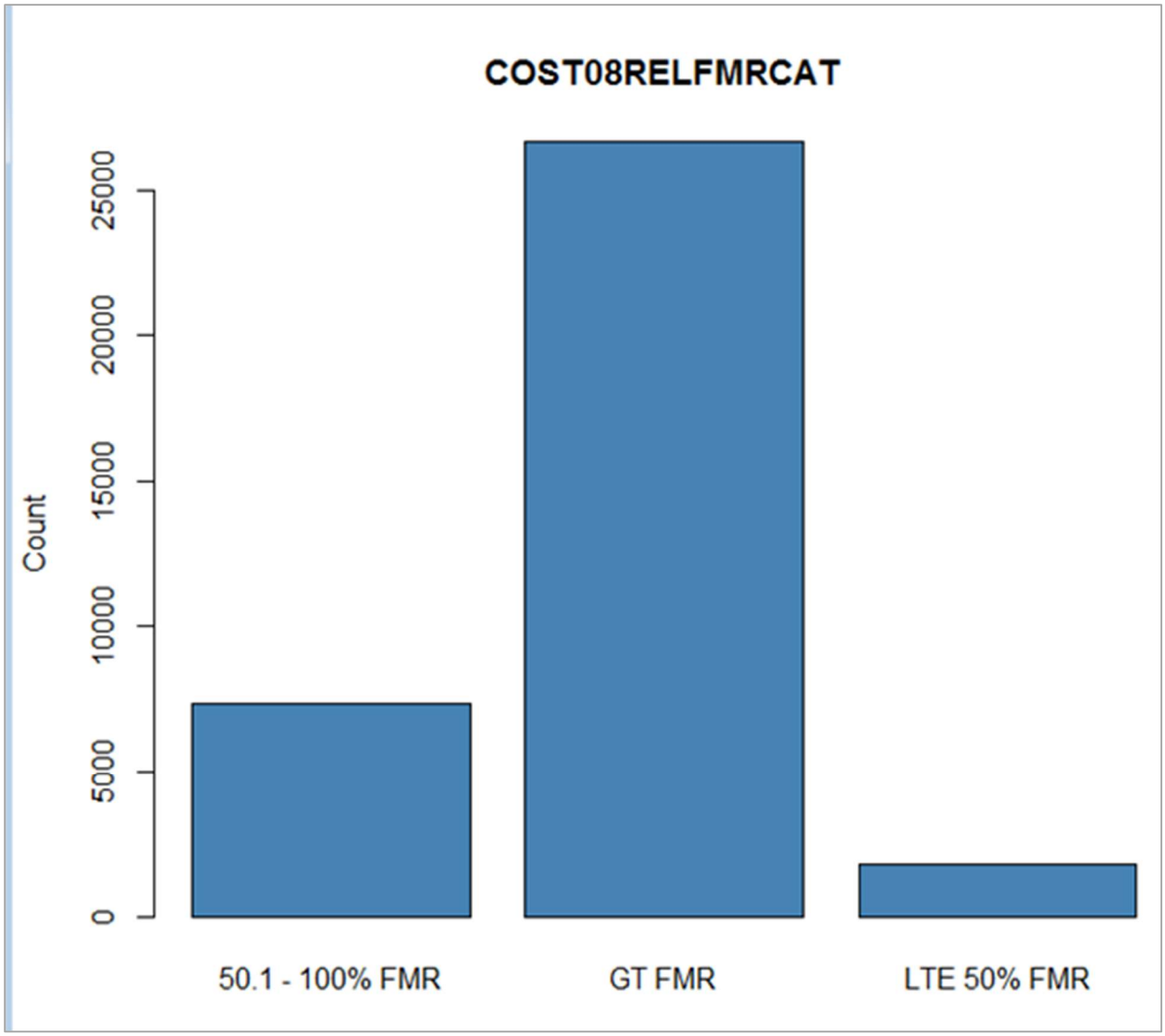

COST12RELPOVCAT
Count
0
5000
10000
15000
20000
25000
30000
100-150% Poverty
150-200% Poverty
200%+Poverty
LTE Poverty

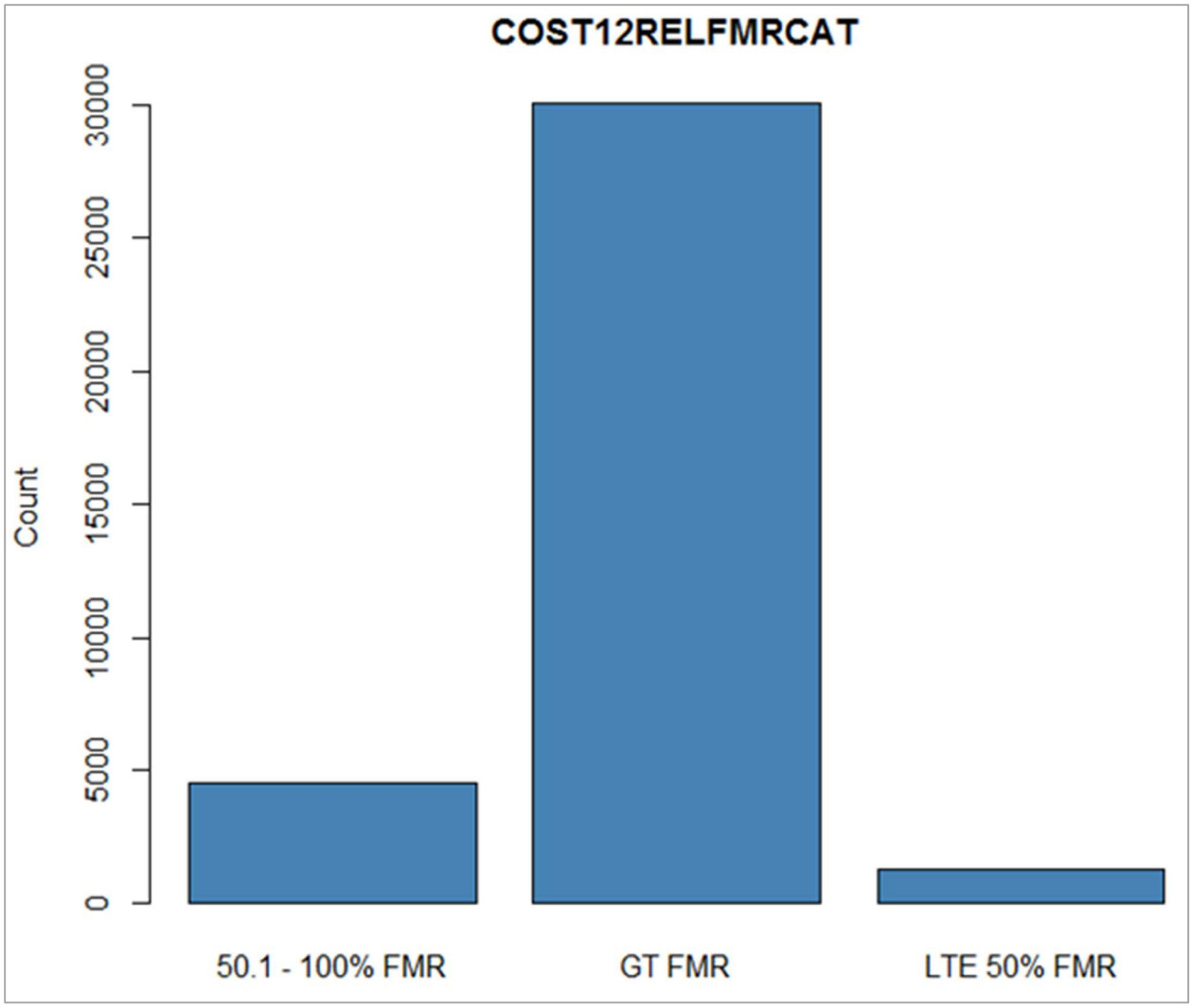

COST12RELPOVCAT
Count
0
5000
10000
15000
20000
25000
30000
100-150% Poverty
150-200% Poverty
200%+Poverty
LTE Poverty

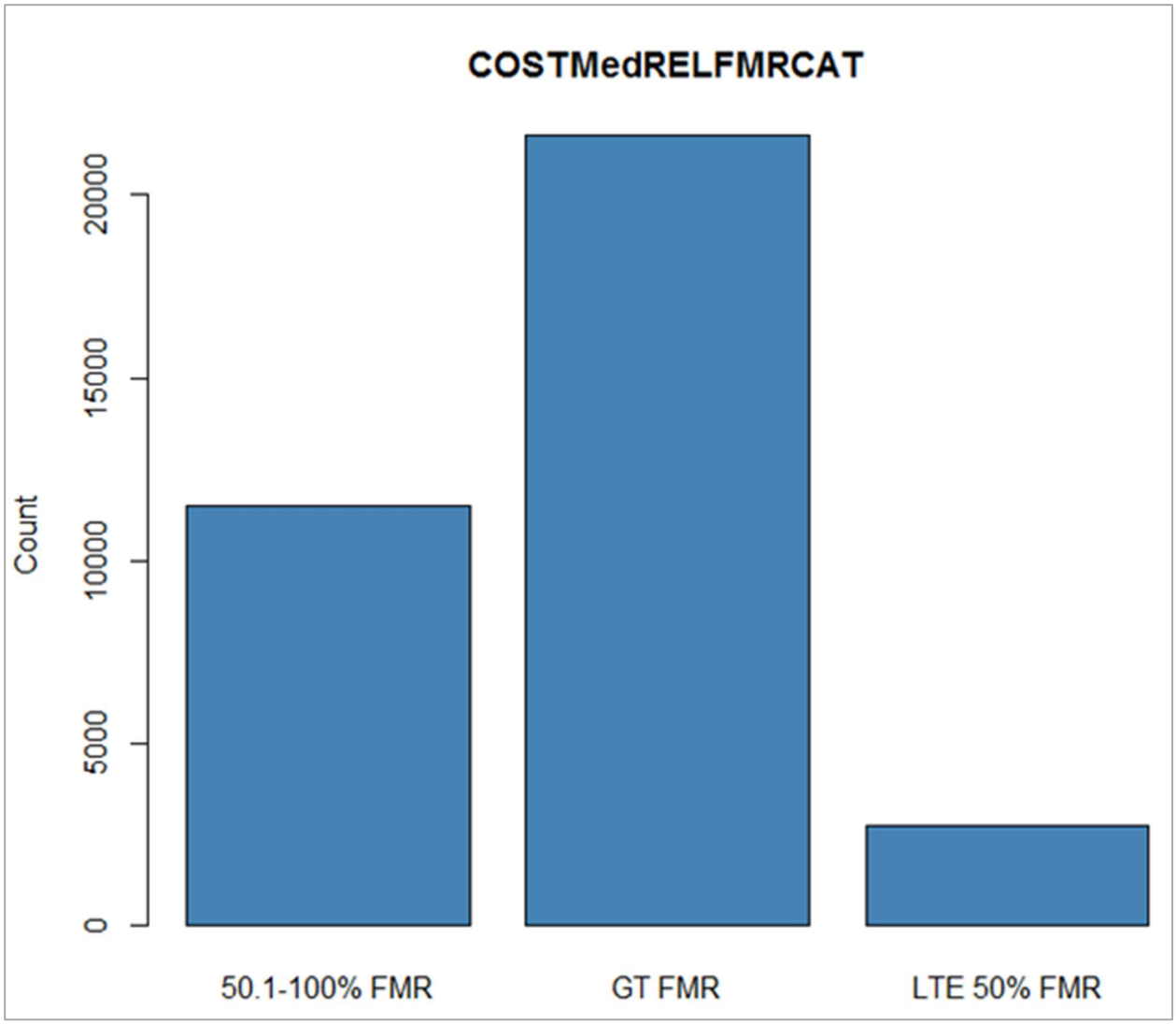

COSTMedRELPOVCAT
Count
0
5000
10000
15000
20000
25000
100-150% Poverty
150-200% Poverty
200%+ Poverty
LTE Poverty

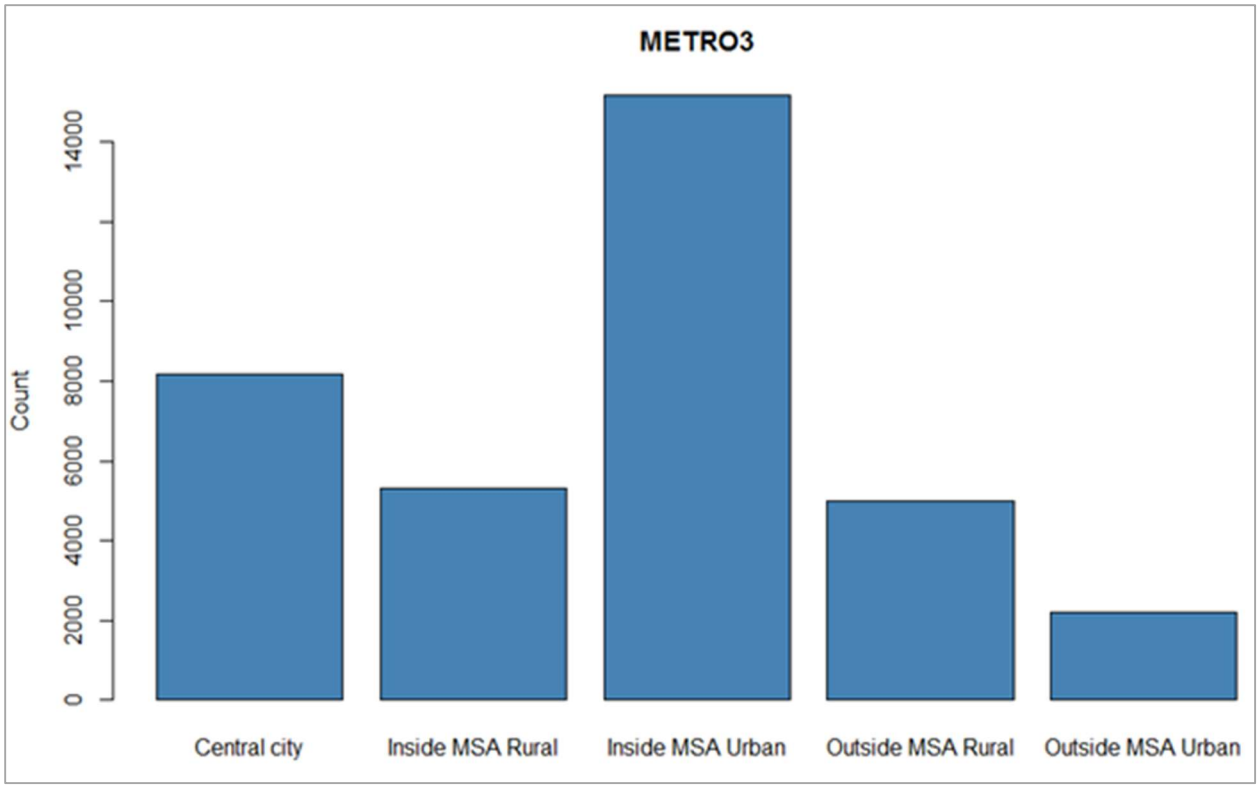

INCRELFMRCAT
Count
0
5000
10000
15000
20000
50.1-100% FMR
GT FMR
LTE 50% FMR

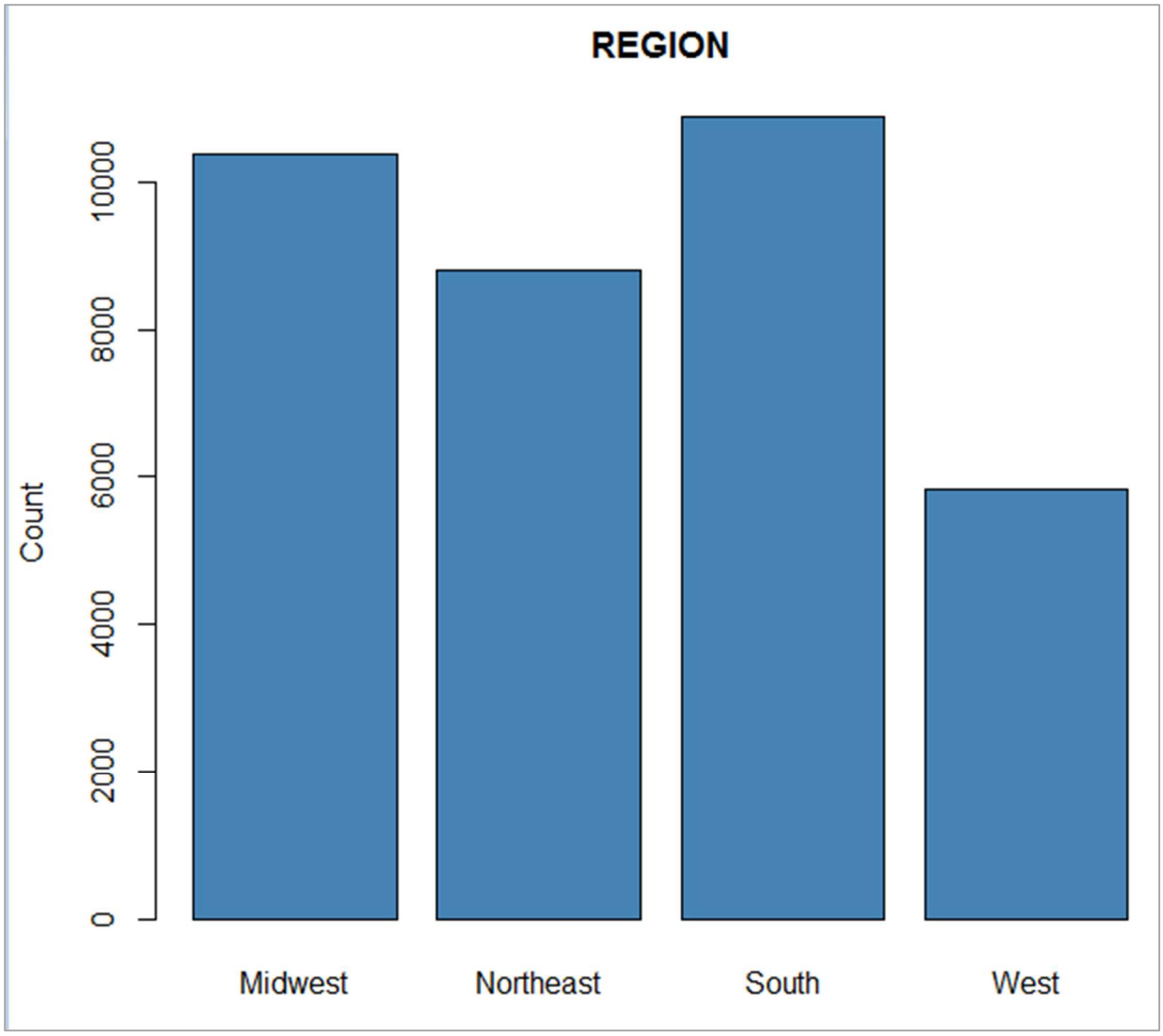

COST06RELFMRCAT
Count
0
5000
10000
15000
20000
50.1-100% FMR
GT FMR
LTE 50% FMR

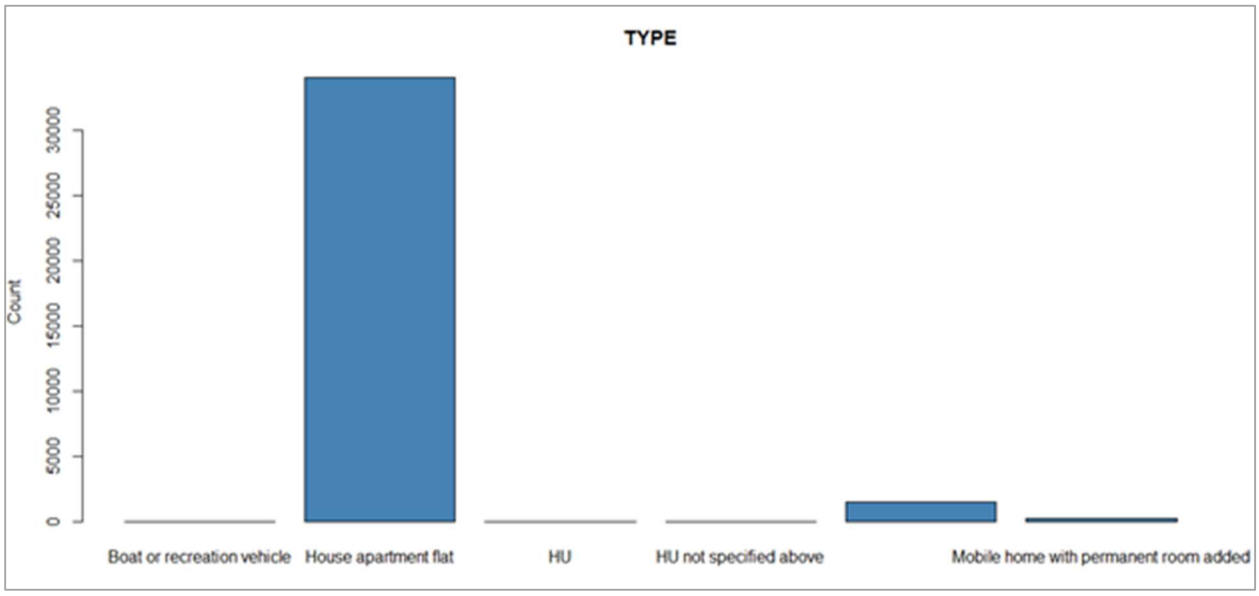

COST08RELFMRCAT
Count
0
5000
10000
15000
20000
25000
50.1 - 100% FMR
GT FMR
LTE 50% FMR

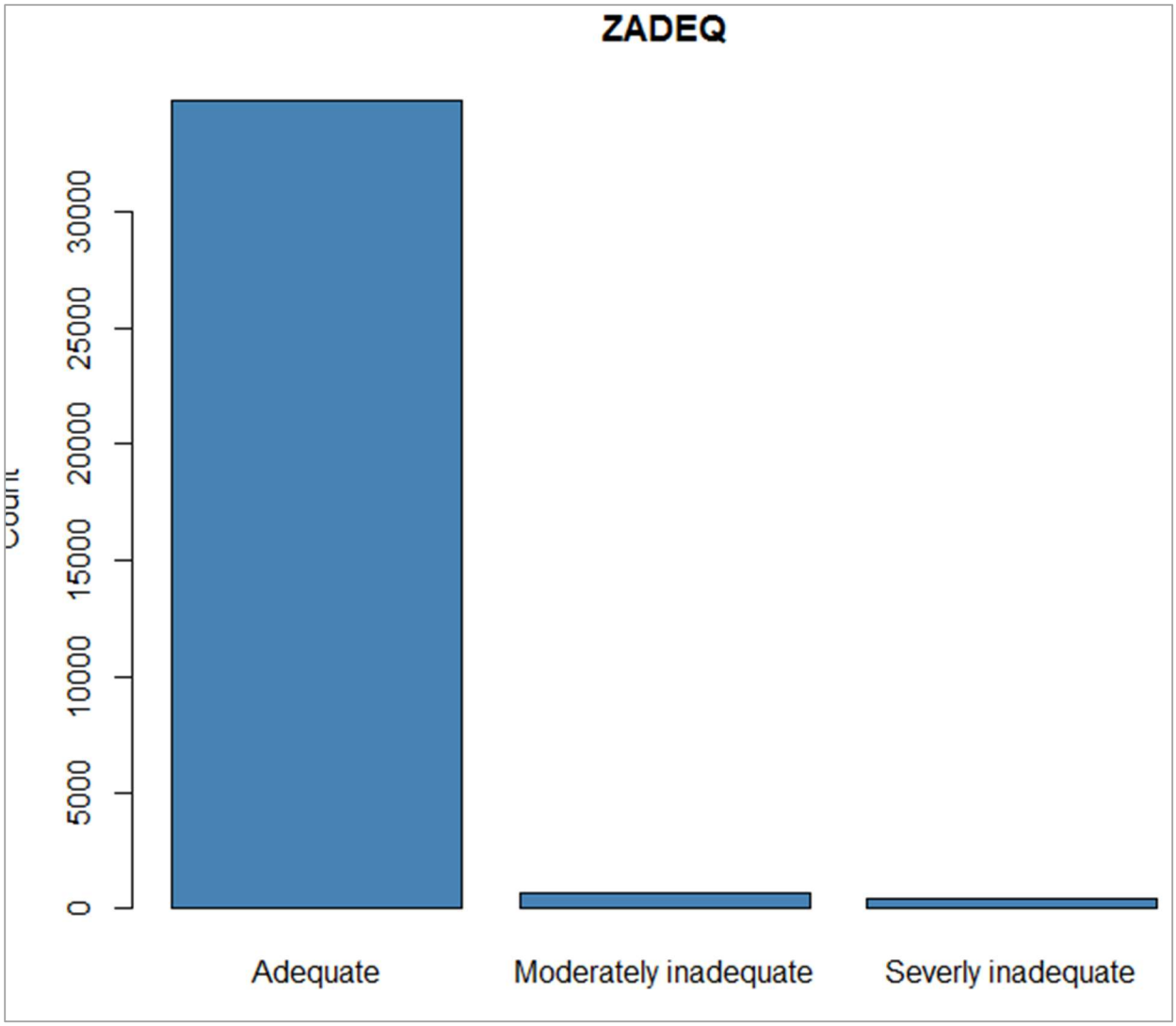

COST12RELFMRCAT
Count
0
5000
10000
15000
20000
25000
30000
50.1 - 100% FMR
GT FMR
LTE 50% FMR

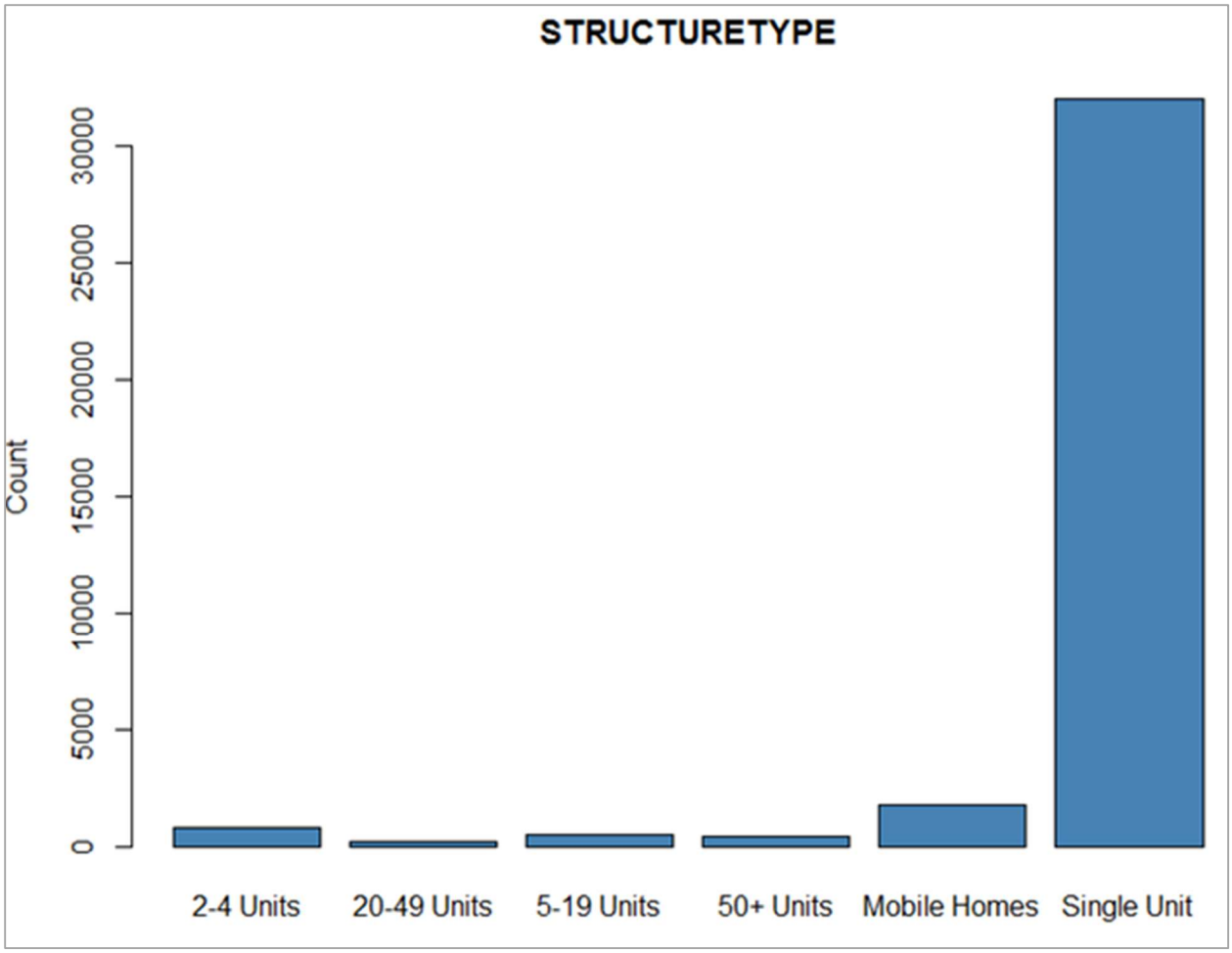

COSTMedRELFMRCAT
Count
0
5000
10000
15000
20000
50.1-100% FMR
GT FMR
LTE 50% FMR


METRO3
Count
0
2000
4000
6000
8000
10000
14000
Central city
Inside MSA Rural
Inside MSA Urban
Outside MSA Rural
Outside MSA Urban


REGION
Count
0
2000
4000
6000
8000
10000
Midwest
Northeast
South
West


TYPE
Count
0
5000
10000
15000
20000
25000
30000
Boat or recreation vehicle
House apartment flat
HU
HU not specified above
Mobile home with permanent room added


ZADEQ
Count
0
5000
10000
15000
20000
25000
30000
Adequate
Moderately inadequate
Severly inadequate


STRUCTURETYPE
Count
0
5000
10000
15000
20000
25000
30000
2-4 Units
20-49 Units
5-19 Units
50+ Units
Mobile Homes
Single Unit

Output 12 MCA - Eigenvalues

```
> mca1$eig
 [1] 0.5388214107 0.3592439961 0.1952034312 0.1474755875 0.1351367128 0.1238198560
 [7] 0.1165542403 0.1006170041 0.0971970776 0.0914060450 0.0806982810 0.0689251805
[13] 0.0668302519 0.0622766616 0.0605565210 0.0578709419 0.0547857882 0.0546104414
[19] 0.0534553278 0.0515454185 0.0513655049 0.0506938696 0.0504827633 0.0502662641
[25] 0.0497500690 0.0494526814 0.0493427752 0.0487083038 0.0486908270 0.0477680172
[31] 0.0468614056 0.0456770531 0.0430973974 0.0426799367 0.0418160001 0.0396950047
[37] 0.0376649159 0.0368037934 0.0342276213 0.0321894552 0.0224144130 0.0218532101
[43] 0.0207684395 0.0188443152 0.0183029076 0.0157510466 0.0141907160 0.0138322051
[49] 0.0133226297 0.0122858576 0.0112164905 0.0101434521 0.0089313121 0.0086770714
[55] 0.0083309183 0.0076931671 0.0066173926 0.0060700260 0.0057557808 0.0055349639
[61] 0.0052178149 0.0049655973 0.0041516626 0.0032952096 0.0030653022 0.0027510034
[67] 0.0026297019 0.0024469022 0.0019996676 0.0015865466 0.0012414342 0.0010547195
[73] 0.0007942884
```

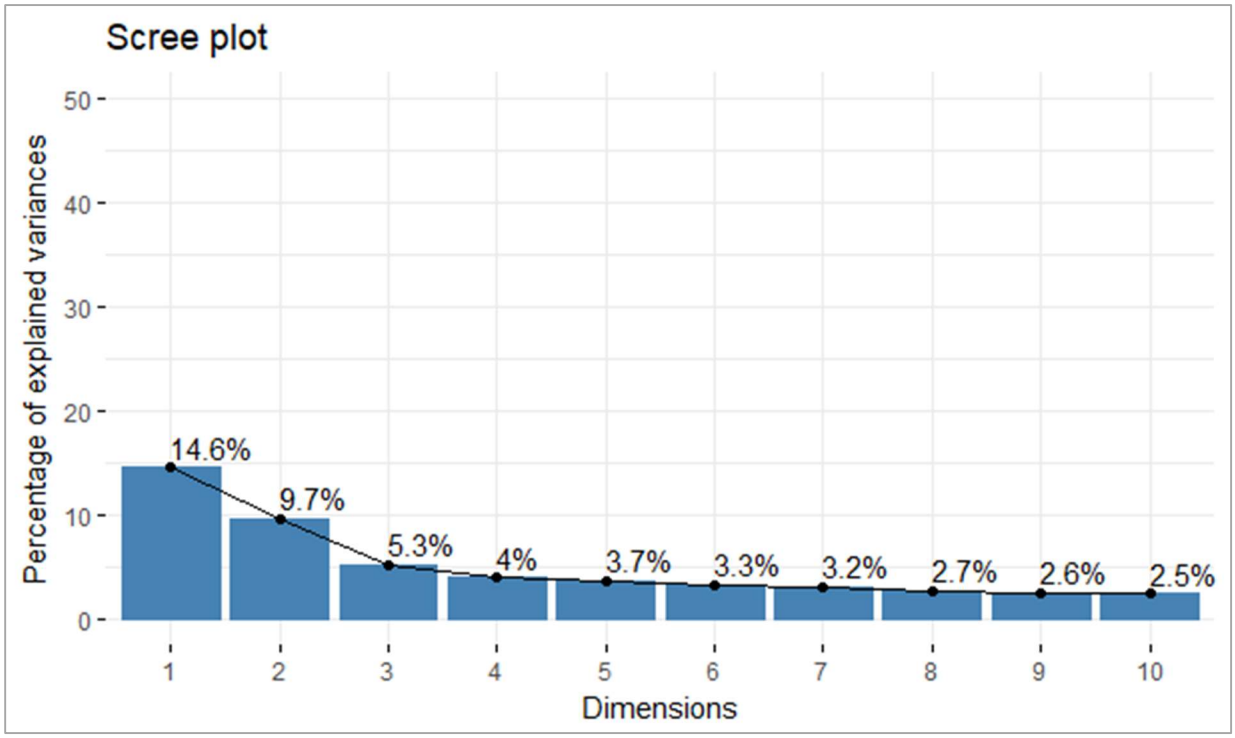
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	0.5388214107	14.56274083	14.56274
Dim.2	0.3592439961	9.70929719	24.27204
Dim.3	0.1952034312	5.27576841	29.54781
Dim.4	0.1474755875	3.98582669	33.53363
Dim.5	0.1351367128	3.65234359	37.18598
Dim.6	0.1238198560	3.34648259	40.53246
Dim.7	0.1165542403	3.15011460	43.68257
Dim.8	0.1006170041	2.71937849	46.40195
Dim.9	0.0971970776	2.62694804	49.02890
Dim.10	0.0914060450	2.47043365	51.49933
Dim.11	0.0806982810	2.18103462	53.68037
Dim.12	0.0689251805	1.86284272	55.54321
Dim.13	0.0668302519	1.80622302	57.34943
Dim.14	0.0622766616	1.68315302	59.03259
Dim.15	0.0605565210	1.63666273	60.66925
Dim.16	0.0578709419	1.56407951	62.23333
Dim.17	0.0547857882	1.48069698	63.71403
Dim.18	0.0546104414	1.47595788	65.18998
Dim.19	0.0534553278	1.44473859	66.63472
Dim.20	0.0515454185	1.39311942	68.02784
Dim.21	0.0513655049	1.38825689	69.41610
Dim.22	0.0506938696	1.37010458	70.78620

Dim.23	0.0504827633	1.36439901	72.15060
Dim.24	0.0502662641	1.35854768	73.50915
Dim.25	0.0497500690	1.34459646	74.85375
Dim.26	0.0494526814	1.33655896	76.19031
Dim.27	0.0493427752	1.33358852	77.52389
Dim.28	0.0487083038	1.31644064	78.84034
Dim.29	0.0486908270	1.31596830	80.15630
Dim.30	0.0477680172	1.29102749	81.44733
Dim.31	0.0468614056	1.26652447	82.71386
Dim.32	0.0456770531	1.23451495	83.94837
Dim.33	0.0430973974	1.16479453	85.11317
Dim.34	0.0426799367	1.15351180	86.26668
Dim.35	0.0418160001	1.13016217	87.39684
Dim.36	0.0396950047	1.07283797	88.46968
Dim.37	0.0376649159	1.01797070	89.48765
Dim.38	0.0368037934	0.99469712	90.48234
Dim.39	0.0342276213	0.92507084	91.40742
Dim.40	0.0321894552	0.86998528	92.27740
Dim.41	0.0224144130	0.60579495	92.88320
Dim.42	0.0218532101	0.59062730	93.47382
Dim.43	0.0207684395	0.56130917	94.03513
Dim.44	0.0188443152	0.50930582	94.54444
Dim.45	0.0183029076	0.49467318	95.03911

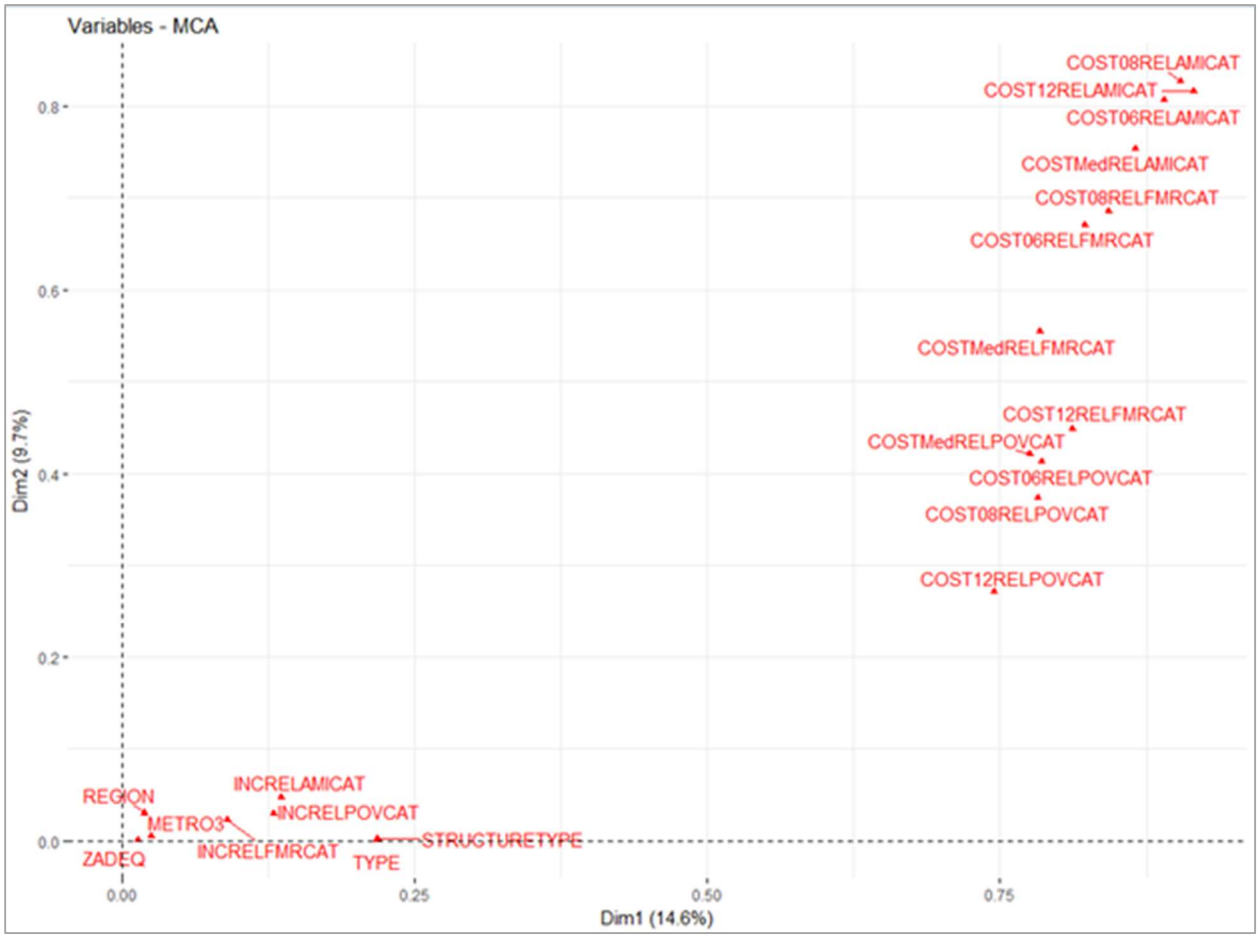
Output 13 MCA - Correlation Ratio

	RS1	RS2	RS3	RS4	RS5
METRO3	0.02536151	0.0054844752	0.0032686657	0.0013807840	0.0007267033
REGION	0.01950515	0.0305301110	0.0326369025	0.0090588841	0.0072037487
TYPE	0.21802612	0.0007654315	0.0005400545	0.0006620685	0.0050484862
ZADEQ	0.01395689	0.0007766034	0.0002757687	0.0001241178	0.0035529492
STRUCTURETYPE	0.21862866	0.0022492897	0.0025783767	0.0033967138	0.0050032681
INCRELAMICAT	0.13620845	0.0478643472	0.0301055220	0.0048957373	0.2518044710
INCRELPOVCAT	0.12970564	0.0297542554	0.0081230222	0.0008480369	0.2196452013
INCRELFMRCAT	0.09052787	0.0225130133	0.0116529041	0.0002253058	0.2499097627
COST06RELAMICAT	0.89033803	0.8069292849	0.7962257551	0.5095356535	0.2177586863
COST06RELPOVCAT	0.78553993	0.4138087676	0.0909661646	0.1430397420	0.2806399669
COST08RELAMICAT	0.90410522	0.8271186232	0.8516025467	0.6752610920	0.2669430592
COST08RELPOVCAT	0.78258944	0.3735448113	0.0912510851	0.1327210613	0.2195108713
COST08RELFMRCAT	0.84246588	0.6854423016	0.0177848484	0.0187442858	0.0220789574
COST06RELFMRCAT	0.82267166	0.6711118472	0.0183320281	0.0503634149	0.0088452623
COST12RELAMICAT	0.91531048	0.8155727890	0.7882597682	0.5327831931	0.1725358016
COST12RELPOVCAT	0.74488264	0.2721485434	0.1039134201	0.1384772386	0.2229961108
COST12RELFMRCAT	0.81183183	0.4485174923	0.1241294608	0.0222037696	0.0092005869
COSTMedRELAMICAT	0.86566656	0.7537988884	0.7337850487	0.5602757508	0.2786417752
COSTMedRELPOVCAT	0.77534406	0.4219929604	0.0748650024	0.1091195776	0.2557019495
COSTMedRELFMRCAT	0.78376219	0.5549560867	0.1237722805	0.0363953224	0.0049866376

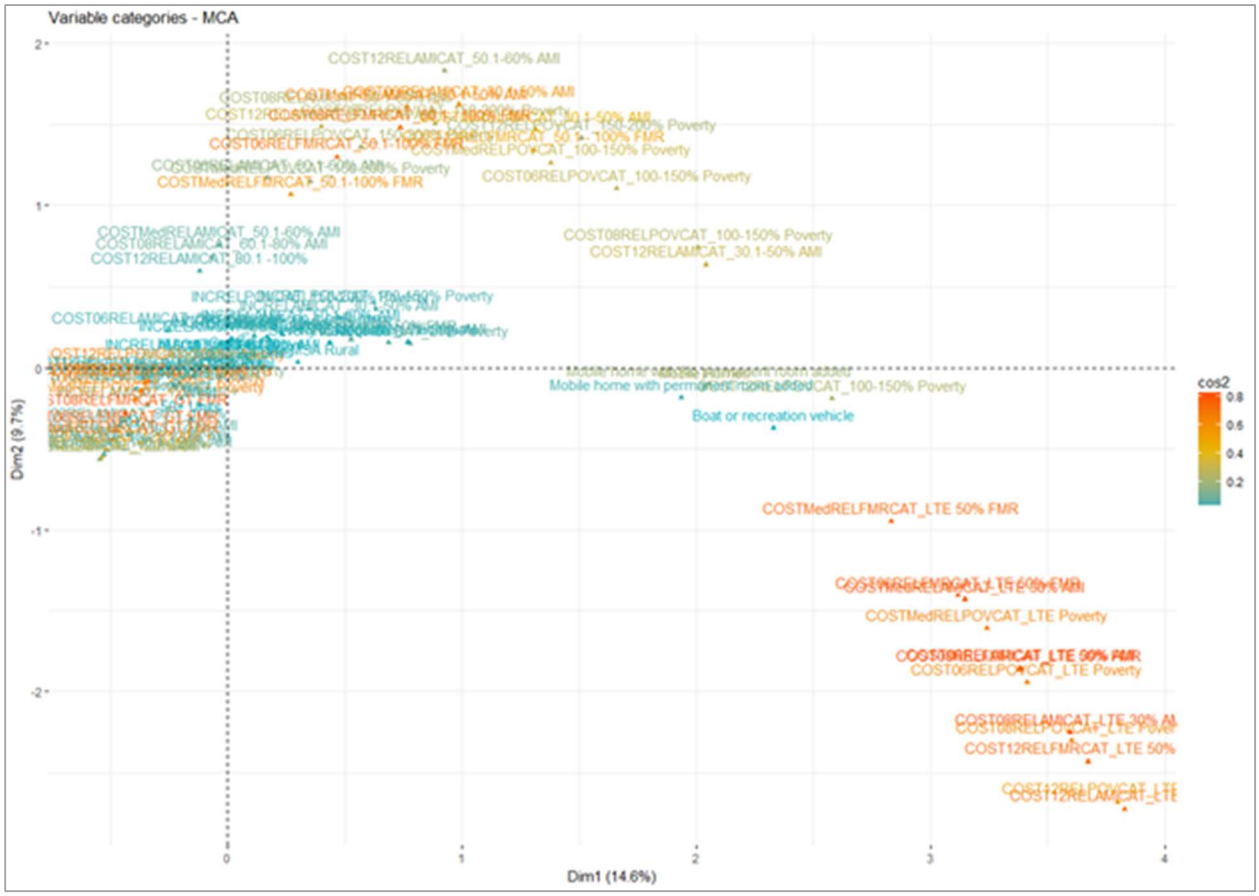
Output 14 MCA - Scree Plot

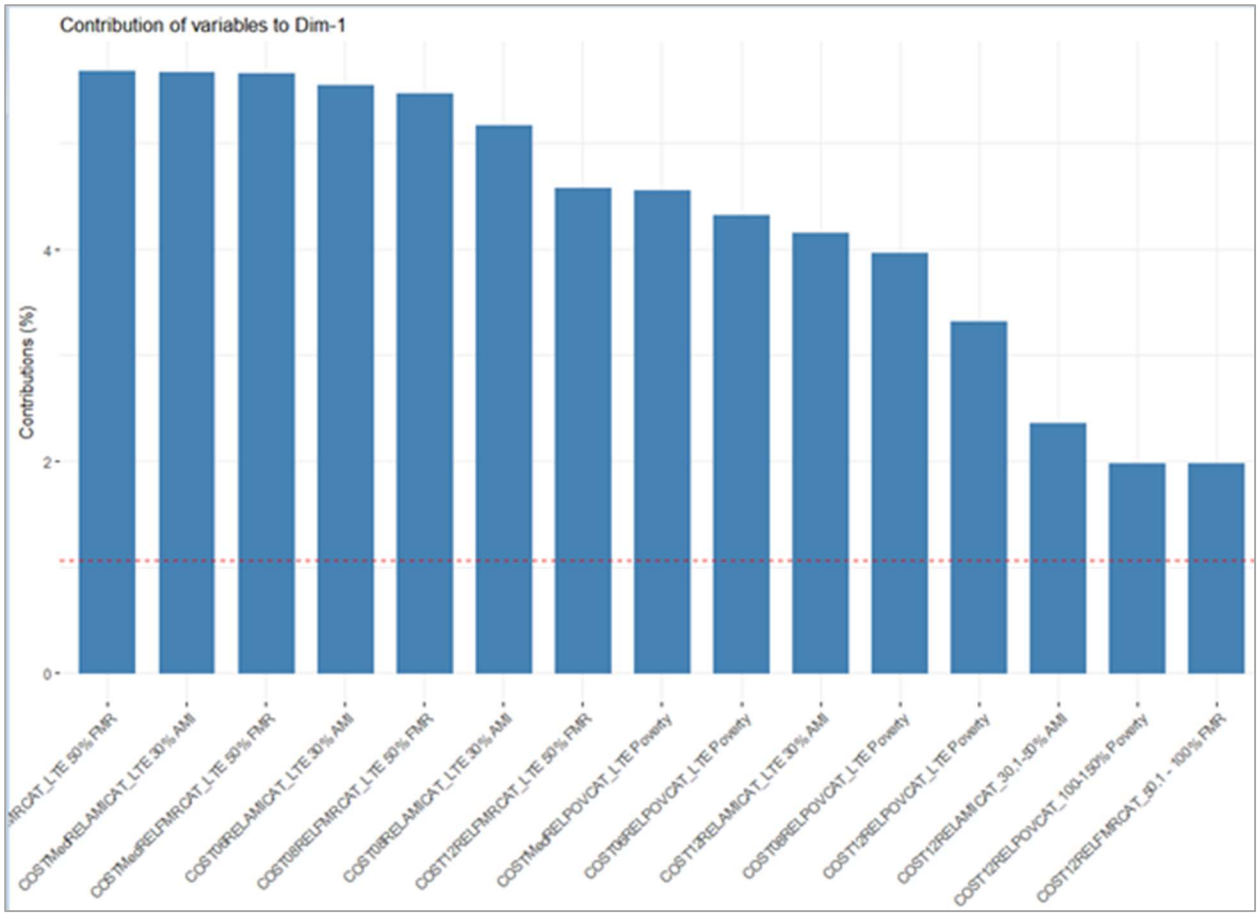


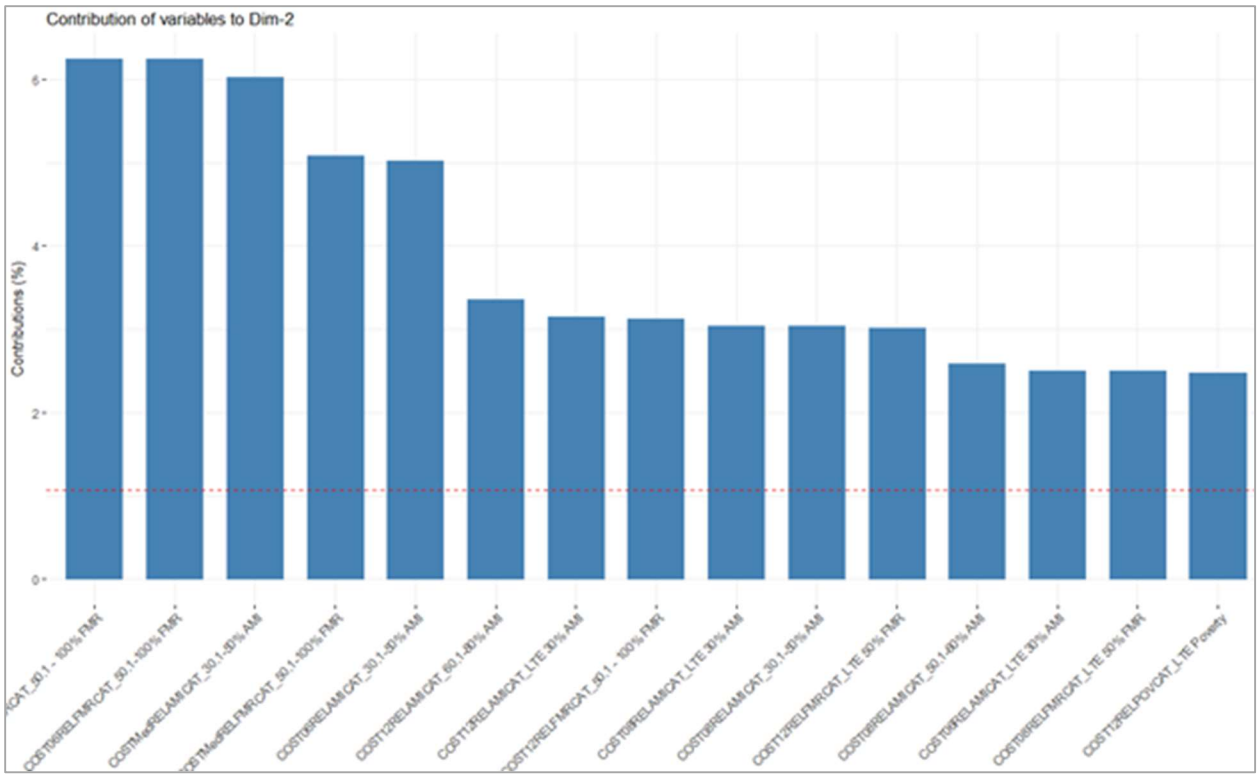
Output 15 MCA - CA plot

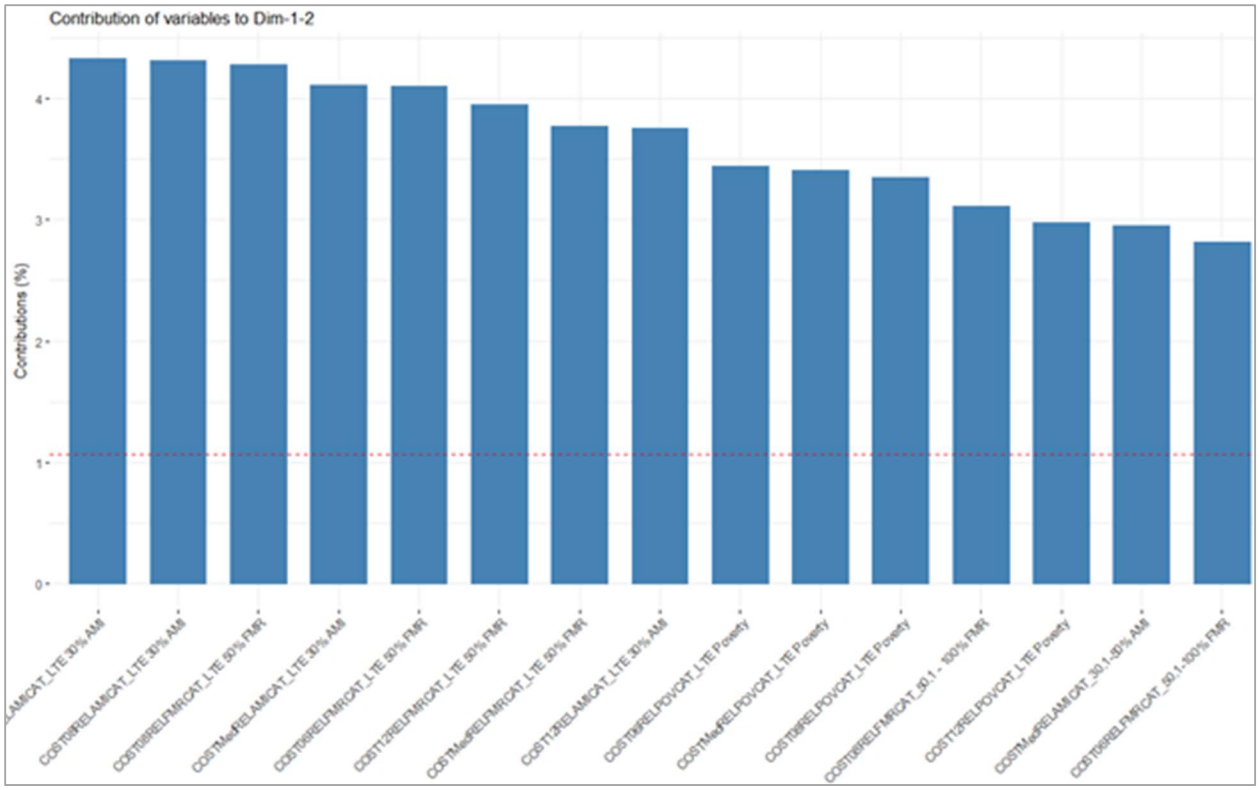


Output 16 - MCA - Quality of representation on dimensions

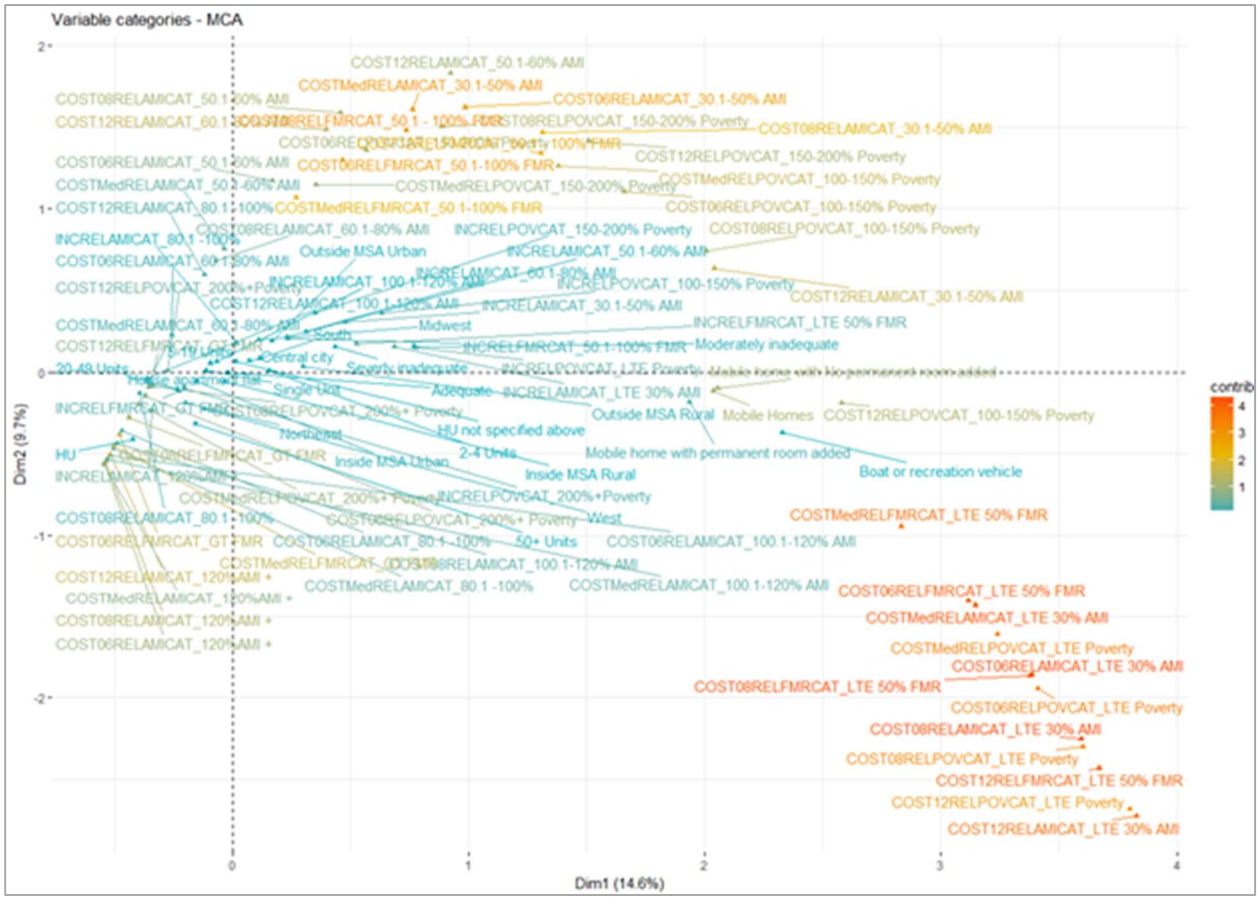



Contribution of variables to Dim-1
Contributions (%)
0
2
4
COST06RELAMICAT_LTE 50% FMR
COSTMedRELAMICAT_LTE 30% AMI
COSTMedRELFMRCAT_LTE 50% FMR
COST06RELAMICAT_LTE 30% AMI
COST06RELFMRCAT_LTE 50% FMR
COST06RELAMICAT_LTE 30% AMI
COST12RELFMRCAT_LTE 50% FMR
COSTMedRELPOVCAT_LTE Poverty
COST06RELPOVCAT_LTE Poverty
COST12RELAMICAT_LTE 30% AMI
COST06RELPOVCAT_LTE Poverty
COST12RELPOVCAT_LTE Poverty
COST12RELAMICAT_30.1-50% AMI
COST12RELPOVCAT_100-150% Poverty
COST12RELFMRCAT_50.1 - 100% FMR

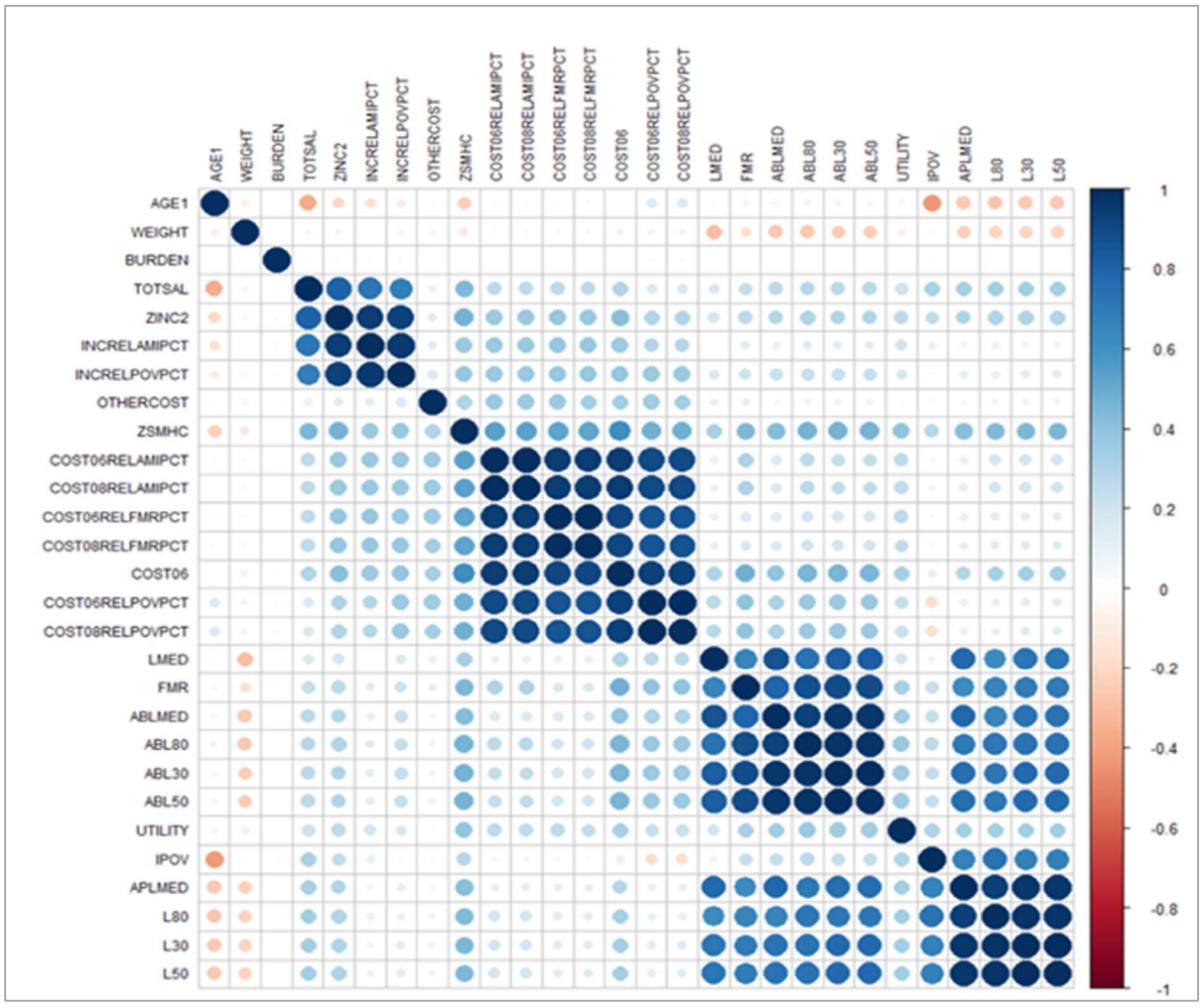

Contribution of variables to Dim-2
Contributions (%)
0
2
4
6
COST06RELFMRCAT_50.1-100% FMR
COSTMedRELAMICAT_30.1-50% AMI
COST06RELAMICAT_30.1-50% AMI
COST12RELAMICAT_60.1-80% AMI
COST12RELAMICAT_LTE 30% AMI
COST12RELFMRCAT_50.1 - 100% FMR
COST06RELAMICAT_LTE 30% AMI
COST06RELAMICAT_30.1-50% AMI
COST12RELFMRCAT_LTE 50% FMR
COST06RELAMICAT_60.1-80% AMI
COST06RELAMICAT_LTE 30% AMI
COST06RELFMRCAT_LTE 50% FMR
COST12RELPOVCAT_LTE Poverty


Contribution of variables to Dim-1-2
Contributions (%)
0
1
2
3
4
COST0REL AMICAT_LTE:30% AMI
COST0RELAMICAT_LTE:30% AMI
COST0RELFMRICAT_LTE:50% FMR
COSTMedRELAMICAT_LTE:30% AMI
COST0RELFMRICAT_LTE:50% FMR
COST12RELFMRICAT_LTE:50% FMR
COSTMedRELFMRICAT_LTE:50% FMR
COST13RELAMICAT_LTE:30% AMI
COST0RELPOVCAT_LTE Poverty
COSTMedRELPOVCAT_LTE Poverty
COST0RELPOVCAT_LTE Poverty
COST0RELFMRICAT_50.1 - 100% FMR
COST13RELPOVCAT_LTE Poverty
COSTMedRELAMICAT_30.1-50% AMI
COST0RELFMRICAT_50.1-100% FMR

Output 17 - MCA - Most important variable categories



Output 18 A - correlation matrix for significant values (18) predictors, however, viewing the graph, more predictors seem to deem insignificant.



Output 18- 3 factor loadings using orthogonal (varimax) and oblique (promax) rotations, (18 predictors), Varimax rotation:

Loadings:

	Factor1	Factor2	Factor3
LMED	0.781	0.306	
FMR	0.827		
ABL30	0.957		
ABL50	0.957		
ABL80	0.932		
ABLMED	0.941		
COST06	0.925		
COST06REL	0.993		
COST06REL	0.905		
COST06REL	0.951		
COST08REL	0.993		
COST08REL	0.905		
COST08REL	0.951		
L30	0.603	0.793	
L50	0.603	0.792	
L80	0.531	0.823	
IPOV	0.823		
APLMED	0.604	0.757	

	Factor1	Factor2	Factor3
SS loadings	6.472	6.454	3.580
Proportion Var	0.360	0.359	0.199
Cumulative Var	0.360	0.718	0.917

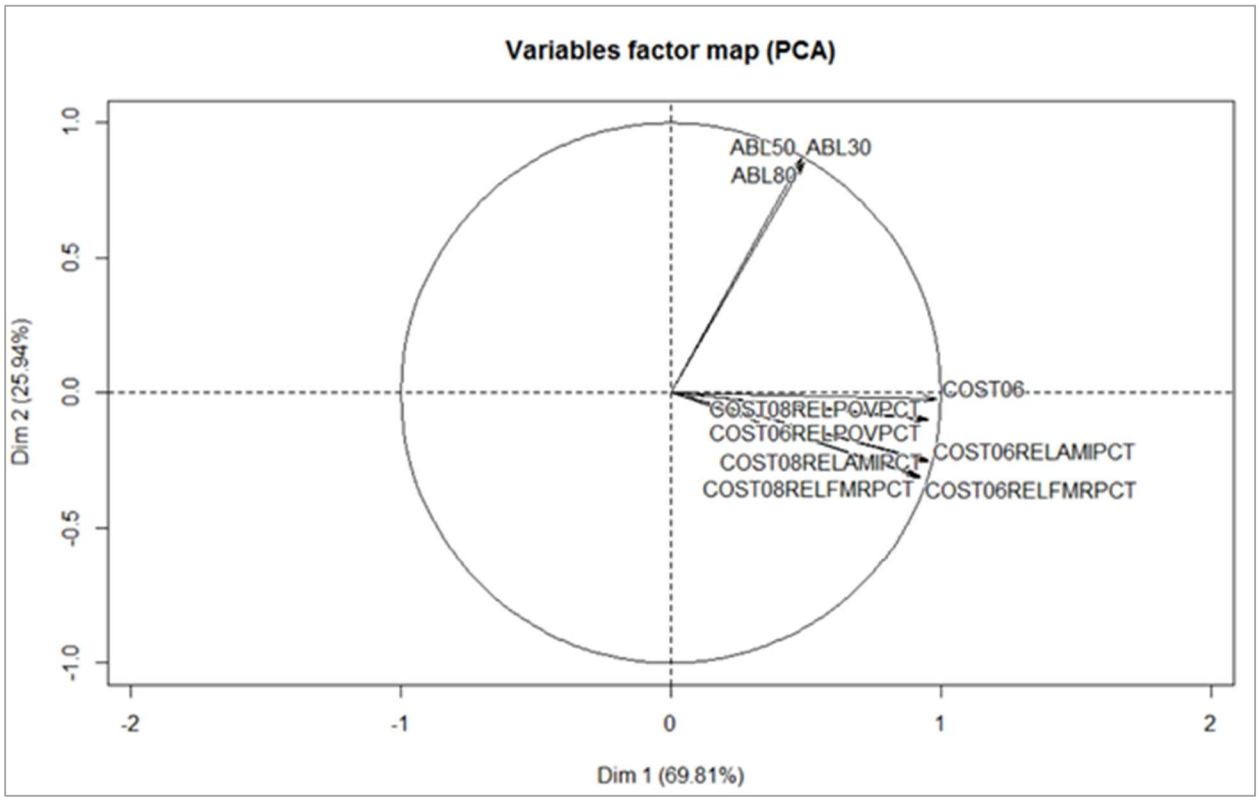
Promax (oblique rotation- 3 factors)

Loadings:

	Factor1	Factor2	Factor3
COST06	0.906		
COST06RELAMIPCT	1.021		
COST06RELPOVPCT	0.872		
COST06RELFMRPCT	0.984		
COST08RELAMIPCT	1.020		
COST08RELPOVPCT	0.873		
COST08RELFMRPCT	0.984		
LMED		0.763	
FMR		0.827	
ABL30		0.987	
ABL50		0.986	
ABL80		0.965	
ABLMED		0.992	
L30		0.320	0.781
L50		0.321	0.781
L80			0.835
IPOV		-0.324	0.960
APLMED		0.349	0.738

	Factor1	Factor2	Factor3
SS loadings	6.399	5.805	3.528
Proportion Var	0.356	0.322	0.196
Cumulative Var	0.356	0.678	0.874

Output 19- Graph of factor 1 and 2 displayed with respective variables



Output 20- 2 factor loadings using orthogonal (varimax) and oblique (promax) rotations, (10 predictors), promax rotation

Loadings:

	Factor1	Factor2
COST06	0.897	
COST06RELAMIPCT	1.017	
COST06RELPOVPCT	0.876	
COST06RELFMRPCT	0.983	
COST08RELAMIPCT	1.016	
COST08RELPOVPCT	0.877	
COST08RELFMRPCT	0.983	
ABL30		1.008
ABL50		1.008
ABL80		0.976

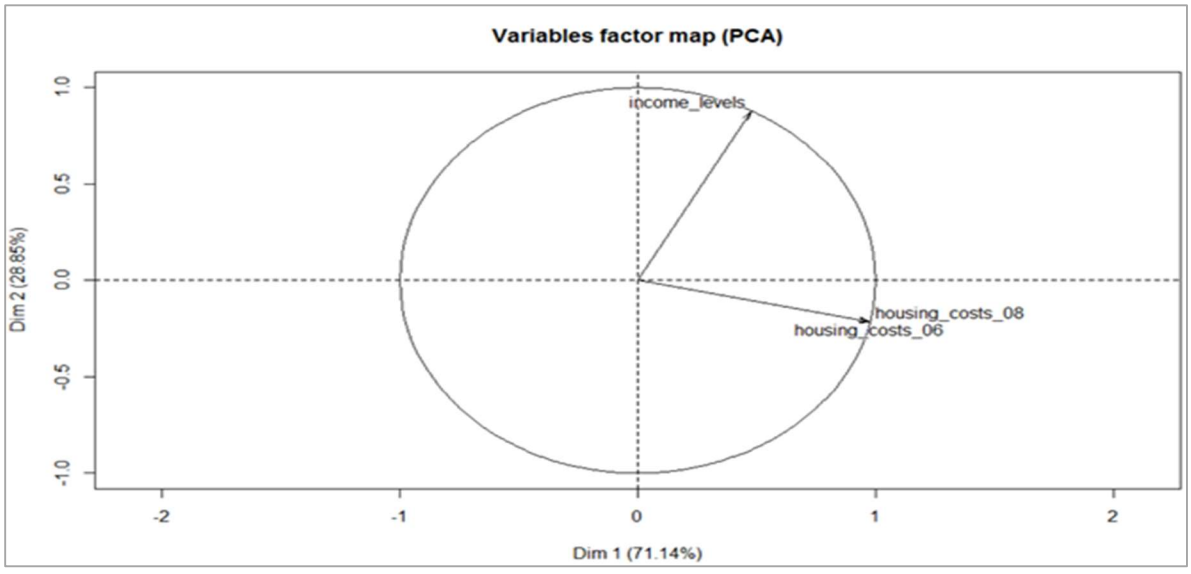
	Factor1	Factor2
SS loadings	6.341	3.072
Proportion Var	0.634	0.307
Cumulative Var	0.634	0.941

Varimax rotation Loadings:

	Factor1	Factor2
COST06	0.915	0.333
COST06RELAMIPCT	0.993	
COST06RELPOVPCT	0.882	0.248
COST06RELFMRPCT	0.952	
COST08RELAMIPCT	0.993	
COST08RELPOVPCT	0.882	0.248
COST08RELFMRPCT	0.952	
ABL30		0.989
ABL50		0.989
ABL80		0.963

	Factor1	Factor2
SS loadings	6.239	3.147
Proportion Var	0.624	0.315
Cumulative Var	0.624	0.939

OUTPUT 20A- GRAPH OF VARIABLES FORMED FROM FACTOR 1 AND FACTOR 2



Output 21 – Results & accuracy of LDA model on 50/50 training and test sets for PCA variables

```
Call:
lda(REGION ~ LMED + PER + COST06 + COST08 + COST12 + BEDRMS +
    TOTSAL + NUNITS, data = training2)

Prior probabilities of groups:
      '1'       '2'       '3'       '4'
0.2485217 0.2902488 0.3007364 0.1604931

Group means:
        LMED      PER   COST06   COST08   COST12   BEDRMS   TOTSAL   NUNITS
'1' 79829.85 2.646016 2526.080 2910.990 3756.578 3.140741 70980.48 5.473850
'2' 65646.05 2.558332 1478.692 1690.338 2155.292 3.132616 57011.37 3.655391
'3' 60681.67 2.498238 1636.952 1872.095 2388.670 3.147468 54120.30 3.075682
'4' 68798.17 2.681613 2818.714 3265.168 4245.961 3.229753 62938.52 2.894334

Coefficients of linear discriminants:
                  LD1           LD2           LD3
LMED    -9.417962e-05 -2.619160e-05  1.433369e-05
PER     -3.290506e-02  1.084208e-01  1.995232e-01
COST06  -6.988617e-04 -2.793654e-03 -1.304200e-02
COST08  -8.612966e-05 -1.514401e-04 -1.389387e-03
COST12   4.796234e-04  2.268899e-03  9.348002e-03
BEDRMS   1.493554e-01 -1.131166e-01  1.269653e-01
TOTSAL   3.772893e-07 -2.768569e-06 -4.163847e-07
NUNITS   1.652899e-04 -2.959382e-03  3.128593e-03

Proportion of trace:
   LD1    LD2    LD3
0.8978 0.0871 0.0151
```

```
housingtrainclass2  '1'  '2'  '3'  '4'
                '1' 3048  476  432  615
                '2'  725 2733  862  846
                '3'  527 1889 4013 1153
                '4'  155  105   84  263
> table(housingtestclass2, test2[,48])

housingtestclass2  '1'  '2'  '3'  '4'
               '1' 2929  452  430  606
               '2'  727 2660  902  898
               '3'  527 1928 4058 1185
               '4'  158  120   90  256
```

housingtrainclass2	'1'	'2'	'3'	'4'	% correct
'1'	3048	476	432	615	67%
'2'	725	2733	862	846	53%
'3'	527	1889	4013	1153	53%
'4'	155	105	84	263	43%

housingtestclass2	'1'	'2'	'3'	'4'	% correct
'1'	2929	452	430	606	66%
'2'	727	2660	902	898	51%
'3'	527	1928	4058	1185	53%
'4'	158	120	90	256	41%

Output 22 - Combining variables within factors to categorize into two distinct variables, income levels and housing costs.

Multiple Regression summary of linear model with FMR as dependent variable (numeric variables only):

```
Call:
lm(formula = FMR ~ VALUE + LMED + IPOV + BEDRMS + NUNITS + ROOMS +
    WEIGHT + ZINC2 + ZSMHC + UTILITY + OTHERCOST + COSTMED +
    BURDEN, data = myNumData)

Residuals:
     Min       1Q   Median       3Q      Max
-1110.46  -134.89   -45.70    76.64  1299.66

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.843e+02  9.779e+00 -59.752  < 2e-16 ***
VALUE        2.640e-04  5.634e-06  46.855  < 2e-16 ***
LMED         1.806e-02  1.076e-04 167.872  < 2e-16 ***
IPOV         2.113e-03  2.285e-04   9.247  < 2e-16 ***
BEDRMS       2.133e+02  2.235e+00  95.437  < 2e-16 ***
NUNITS       2.088e-01  4.736e-02   4.408 1.04e-05 ***
ROOMS       -3.504e+01  1.186e+00 -29.539  < 2e-16 ***
WEIGHT       1.316e-02  1.032e-03  12.746  < 2e-16 ***
ZINC2       -1.336e-04  1.697e-05  -7.868 3.71e-15 ***
ZSMHC        1.488e-02  1.615e-03   9.213  < 2e-16 ***
UTILITY      1.210e-01  1.157e-02  10.460  < 2e-16 ***
OTHERCOST    3.773e-02  7.576e-03   4.980 6.37e-07 ***
COSTMED             NA         NA      NA       NA
BURDEN      -1.385e-02  1.806e-02  -0.767    0.443
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 227.8 on 35272 degrees of freedom
Multiple R-squared:  0.6667,    Adjusted R-squared:  0.6666
F-statistic:  5881 on 12 and 35272 DF,  p-value: < 2.2e-16
```

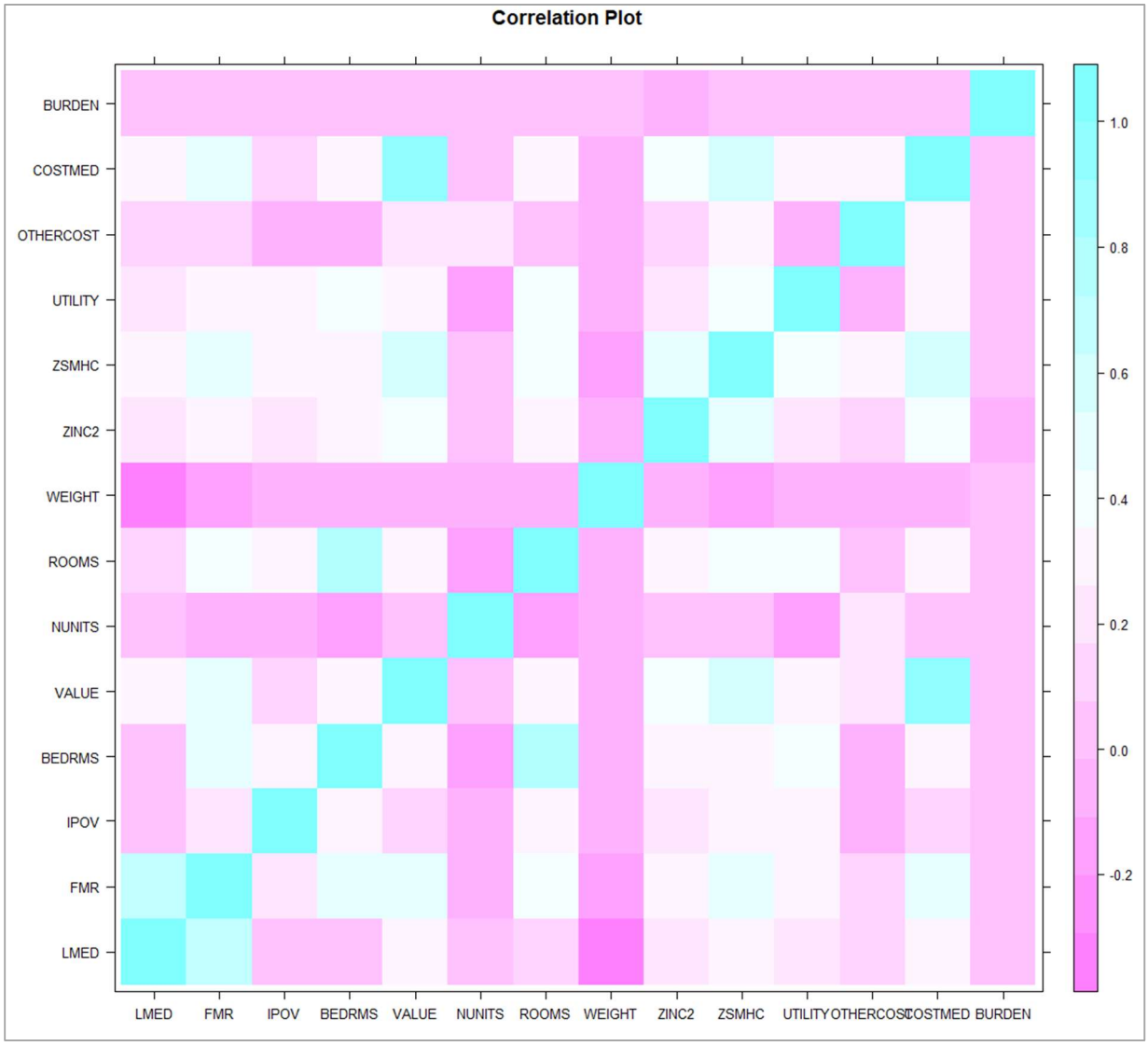
Multiple Regression summary of linear model with VALUE as dependent variable (numeric and categorical variables):

```
Call:
lm(formula = VALUE ~ ., data = myNumCatData)

Residuals:
       Min         1Q     Median         3Q        Max
-6.707e-04 -2.652e-05  4.150e-06  2.132e-05  5.858e-04

Coefficients:
                Estimate Std. Error    t value Pr(>|t|)
(Intercept)    9.950e-05  2.576e-05  3.863e+00 0.000112 ***
AGE1          -6.851e-08  2.218e-08 -3.089e+00 0.002008 **
METRO3        -3.589e-07  2.858e-07 -1.256e+00 0.209246
REGION        -5.044e-07  3.521e-07 -1.432e+00 0.152014
FMR           -1.018e-08  1.143e-09 -8.902e+00  < 2e-16 ***
BEDRMS         2.874e-06  6.309e-07  4.556e+00 5.24e-06 ***
BUILT         -4.520e-08  1.313e-08 -3.441e+00 0.000579 ***
TYPE           1.143e-06  9.189e-07  1.243e+00 0.213722
NUNITS        -2.198e-08  1.304e-08 -1.686e+00 0.091732 .
ROOMS         -1.544e-06  3.192e-07 -4.836e+00 1.33e-06 ***
WEIGHT         1.521e-11  2.912e-10  5.200e-02 0.958350
ZINC2         -2.977e-11  4.515e-12 -6.594e+00 4.36e-11 ***
ZADEQ         -1.077e-06  1.286e-06 -8.380e-01 0.402058
ZSMHC         -3.025e-09  4.403e-10 -6.871e+00 6.49e-12 ***
STRUCTURETYPE  5.712e-07  3.587e-07  1.592e+00 0.111300
UTILITY       -1.721e+02  3.080e-09 -5.587e+10  < 2e-16 ***
OTHERCOST     -1.721e+02  2.122e-09 -8.111e+10  < 2e-16 ***
COSTMED        1.721e+02  2.664e-10  6.461e+11  < 2e-16 ***
BURDEN         2.188e-08  4.768e-09  4.589e+00 4.47e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.015e-05 on 35266 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 4.297e+22 on 18 and 35266 DF,  p-value: < 2.2e-16
```

VIF model using 'car' package showing acceptable values for categorical and numeric model:

```
> vif(numCatModel)
      AGE1     METRO3     REGION        FMR    BEDRMS     BUILT      TYPE     NUNITS     ROOMS     WEIGHT
  1.183899   1.322395   1.271727   1.984574   3.063313   1.184677   1.472047   1.203404   2.873694   1.256192
     ZINC2      ZADEQ      ZSMHC STRUCTURETYPE    UTILITY  OTHERCOST    COSTMED     BURDEN
  1.422335   1.012839   2.195167   1.810016   1.403564   1.377967   2.064027   1.002269
```

Plot of correlation matrix between numeric variables shown above:



Output 23 - Accuracy of LDA model on 80/20 training and test data

housingtrainclass	'1'	'2'	'3'	'4'	% correct
'1'	4518	347	688	523	74%
'2'	1007	6620	1103	346	73%
'3'	986	1289	6457	1690	62%
'4'	583	3	420	2102	68%
housingtestclass	**'1'**	**'2'**	**'3'**	**'4'**	**% correct**
'1'	1085	103	165	111	74%
'2'	246	1678	304	96	72%
'3'	235	322	1616	432	62%
'4'	136	1	118	522	67%

Output 24 – Accuracy of LDA model on 80/20 training and test sets for PCA variables

housingtrainclass2	'1'	'2'	'3'	'4'	% correct
'1'	4837	775	705	1010	66%
'2'	1141	4168	1300	1319	53%
'3'	878	3157	6531	1926	52%
'4'	238	159	132	406	43%
housingtestclass2	**'1'**	**'2'**	**'3'**	**'4'**	**% correct**
'1'	1167	202	175	231	66%
'2'	261	1024	322	357	52%
'3'	214	822	1676	475	53%
'4'	60	56	30	98	40%

VI. Sources

1. *American Housing Survey: Housing Affordability Data System*. Retrieved from https://www.huduser.gov/portal/datasets/hads/hads.html

VII. Individual Summaries

Individual Contribution and Reflection

Ning Yan - MCA

In the final project, I participated in the group discussions on the project. Also, I did exploratory analysis on our dataset to create boxplots for the variables and also created correlation matrix for all the variables to see their correlation. For the final presentation, I help with finalizing the PowerPoint slides. Also, as the only in-class student, I presented and played presentations for all other group members of our project findings.

I used multivariate correspondence analysis method for the project to explore all categorical variables in our dataset. In total, there are 93 levels/variable categories among all the categorical variables. In order to see frequencies of each variable category, I did bar charts for each variable and because I want to make sure that I haven't left out any information that are significant to the final output, I decided to keep all categories first. The reason why I did bar charts in the first place is that low frequency variable categories have a possibility of distorting the final result. I used R built-in package FactoMineR and function MCA() from this package to do my analysis. Firstly, I transformed all character values to factors to fit the function. Then, I applied the function to all my categorical variables. Most of my time is spent on interpretation on all outputs. The nice thing about this package is that it generates eigenvalues and contributions of each dimension so that I can better understand each dimension and what they mean in regards to the whole dataset and how it can relate to other analysis on numeric variables like PCA and CFA. To visualize the final outputs, I did different plots and bar charts to show relationship among different variable/ variable categories and relationship between variable/ variable category and dimensions. I also created two bar charts to see contributions of different variable categories to the two dimensions I included in my final interpretation of the result.

From my multivariate analysis, I learned a new package in R that can contribute to analysis on categorical variables. This is very useful and important because almost every dataset collected from survey contains categorical/ordinal variables and a lot of the times, there is a general pattern among these categorical variables that would contribute to the conclusion of the dataset. By using this R package to do the analysis, I managed to discover underlying structure among all categorical values in the dataset by reading into the eigenvalues and squared cosine values of all the dimensions and variable categories. Although the variance explained by the two dimensions I included in my final interpretation, the result is accurate on some extent that it corresponds to the outputs from other analysis by the team. I also did some thinking in how to make the final result better. One approach is to get more information on all the categorical variables to see if certain low-frequency variable categories are not important/ influential and thus can be removed from the dataset. Also, as suggested by Professor Besser, another way is to see if certain low-frequency variable categories can be coded/combined together and make a new variable category with normal frequency. The other way to improve the final result is to include more dimensions in final representation. Since there are more than 15 dimensions from the analysis, including more dimensions will have more variance explained by MCA and thus the result will be much more accurate. Variable/ variable categories with low square cosine values could be better represented by the dimensions.

Finally, by accomplishing this project, I become more familiar with correspondence analysis especially in how to deal with different variable data type, how to re-code the categorical variables to fit in the analysis/ models and how to interpret all the outputs from the analysis. I also learned different ways to visualize the results. The most important thing is that I learned to how to related the findings from correspondence analysis back to the original dataset and other methods on numeric values to explain the structure in the dataset.

Angelene Arito

Contribution

My first contribution to the final project was in searching for a dataset to use; we had each agreed to do some research on our first chain of emails and then meet online to discuss which we were going to move forward with. I found two potential datasets that I linked to the group for our initial meeting. Neither of my datasets wound up being chosen, but I was part of the conversation in determining which of the datasets everyone brought was the best to use and helped to ensure it met all the requirements.

After we selected the dataset to use, we had another meeting to discuss exploratory analysis, what sort of cleaning we would have to do the set, and who would be performing which analysis. I maintained a master file for our dataset in excel in which I manually created a key for all the variables to have a clean list (since the documentation on the set was all over the place and in PDF format). I also went through and tagged all the variables with missing values and created a summary of the instances. I was part of the discussion in determining which variables should be removed from analysis as well, after which I generated a new, cleaned dataset without missing values and the variables the group agreed to omit for analysis. I then sent this out to the group so that everyone was working from the same file to prevent any errors. I also documented which variables we decided to keep and gave them either a "numeric" or "categorical" tag, so that it was easy to scan the variable list. Beyond that exploratory analysis piece, I also contributed to answering some of the questions in part 1 of homework 3 (which had been posted to a google doc to make group contributions easier to manage); I was also the one who finished touching up that part of the homework and submitting it to D2L.

In terms of actual analysis, I performed PCA on the numerical variables in the cleaned dataset. I pulled out the numeric variables and used R to run a few iterations of PCA. Most of the PCA work I did on this dataset can be seen in my homework 3 submission, but to summarize techniques at a high level – 1) I ran MSA tests (KMO & Bartlett's test) to ensure PCA was a valid approach to the data 2) I ran PCA on all numerical variables in the cleaned dataset, scaled to account for the varying units between variables 3) I applied varimax rotation after determining the loadings were too difficult to interpret 4) after analyzing the new loadings, I determined that some of the variables don't appear helpful or relevant so I removed them and ran another PCA iteration same as the first (ran scaled PCA, applied varimax rotation, analyzed the rotated loadings). After getting more satisfactory results on the second PCA iteration, I analyzed each principal component and gave a name to each category with a description of the information weighted there.

I also performed LDA. Because I was already looking at and familiar with the numerical variables from running PCA, I decided to use them again in LDA and look at them in relation to census region, which seems like it might have an interesting relationship with housing affordability. Most of the LDA work I did on this dataset can be seen in my homework 4 submission, but to summarize the techniques at a high level – 1) I computed the correlation matrix for the variables (minus the percentage variables, which I kept out due to what I concluded in PCA) to determine if there were any high correlation relationships between variables that might contribute to multicollinearity and mess with interpretation 2) I removed the highly correlated variables and then performed k-fold cross-validation to separate the dataset into training and test sets – I put half the observations in one and half in the other 3) I ran LDA on the data and observed the accuracy of the classifications in both the training and test sets and briefly analyzed the results.

After performing the individual analyses, I wrote about the processes to complete them and summarized the results in the PCA and LDA sections of both the technical summary and PowerPoint and about PCA in the non-technical summary. I recorded my portion of the PowerPoint slides for the group presentation on 11/8/17 and then made some changes to PCA and LDA based on feedback from the professor. I also wrote the introductory

portion of the technical summary and cleaned it up for delivery. I was on all the group meetings to discuss what was needed to finish up the documentation for final submission.

Reflection

The biggest learning piece that I got from performing analyses on the dataset for this final project has been tying everything in together. It can be difficult to picture how the different analyses and components of those analyses work together when you're answering specific questions on homework assignments. For example, performing PCA on a much larger dataset like the one we chose for our final project required cleaning, exploratory analysis, MSA testing, and interpretation – all of which we learned in class or on the homework assignments, but in those contexts the concepts were usually in isolation and the datasets we had were usually very small and straightforward (and clean). Especially getting to the interpretation of the loadings on PCA made the whole point of PCA come full circle in a way – because the results from my PCA analysis may have been useful to another group member's analysis. From the LDA analysis, I learned that even though a dataset may be appropriate for a certain technique and that even though I may think performing it will shed some light on the overall analysis, the results may wind up not being as applicable or meaningful as originally imagined.

Jeremy Dai Individual Contribution and Reflection

Before we all met for the first time, we had all agreed through email that we should each procure some data sets we could use in the final project for class. I decided against using data from my job solely because there weren't enough numerical variables in any of our databases. Also, there's potential for sensitive information leaving the office at that point. Anyways, I surfed through Kaggle.com and found a couple good data sets with mainly all numerical variables. We met up as a group and discussed which data set we would like to explore more. We all agreed that the housing data set would give us the most flexibility as it had a good mix of numerical and categorical variables.

We met at a later date and discussed the group assignment where we divided up the roles and decided what we were all going to do for the project - my analysis was to be done with multiple regression. Exploring the data as well as cleaning it was a massive chore not only for myself, but for the entire group as well. We met up after everyone had a chance to research the data, and everyone had the same reaction to the amount of variables and missing values. At this point, everyone was working together to figure out which variables to keep and which to flat out throw away. I used both manual and automatic techniques to decide which variables to keep in my analysis, which included correlation matrices, correlation plots, VIF validations, and manually going through the data and deleting any null or odd values such as negative numbers.

Once I was happy with my data set, I made several iterations of different sets of data such as one with all categorical variables just to check for any outliers or weird correlations after turning them into dummy values. I also made a set with just numerical values and did same, as well as my final set with numeric and categorical variables combined. My numeric set was mainly for training the data because it was easier all around to use mostly numeric variables in R. My test set was the last set I made with the numeric and categorical variables.

In the end, we each had slightly different data sets depending on the statistical technique we used, but our results all agreed with each other that cost was a factor in determining housing values. I contributed to the presentation slides as well as helping with certain aspects of the presentation itself when it came to the online members presenting. The final paper came together really well with all of us adding to it through Google Docs.

I thoroughly enjoyed working with my group and was glad they had as much patience as me when it came to exploring our data set. It was very interesting and enlightening to go over the other members' results and seeing how they applied the techniques learned in class. The data set really pushed the limits of my multiple regression model. There was a need to validate everything to make sure the numbers were right. Thus, learning VIF in class was extremely beneficial and I will definitely be taking that away with me for use in future analyses.

Helen Ramchandani

Contribution

As soon as group was formed, I created google hangout to include all members for collaboration and decide on a topic for project. Once dataset chosen, discussion of which variables to keep and discard resulted in an updated dataset used for analysis. Method selection was initiated and for my part of this project, I chose to implement factor analysis and cluster analysis. A complete exploratory analysis was performed and include the following: create correlation matrix, check VIF for multicollinearity, KMO test for sample adequacy, Bartlett test check for overall variance, Cronbach's Alpha for factor reliability. Next step, analyze factors, extract variables deem significant and create new variables to categorize the data into a generalized variable for better understanding the factors. These new variables give a better description to form new hypothesis.

Next method I chose to implement is cluster analysis using k-means. I used the same dataset (clean) and ran cluster analysis on raw data. In addition, ran cluster analysis after implementation of factor analysis where results are same in both tests. The tables are shown in the description pertaining to cluster analysis along with corresponding confusion matrix and statistical output.

Once analysis was complete, included my report, graphs and conclusions in corresponding section, "technical summary" of google docs. Power point slides created for submission as a group presentation held 11/08/17. In addition, I initiated a new document, "CSC424 Non-technical Summary" for all members to include their part of conclusions and explanation of results in layman terms.

Reflection

I enjoyed working as a group with this data set and gained knowledge about various multivariate regression methods. This was my first time using R language and I am amazed of its contents including libraries to generate statistical calculations dealing with matrices. In addition, it has many library functions to generate high quality visual graphs. To generate these types of graphs in another programming language would be tedious to say the least.

Given the size of dataset along with many categorical variables, implementing cluster analysis with k-means was interesting. I ran cluster analysis on raw dataset to view results and implement cluster analysis on results from factor analysis. To my surprise, the results were the same. It is fascinating how "kmeans" algorithm worked on raw data without providing any information about variables. One can clearly learn from data! Also, I found MCA particularly interesting by observing the relationships between categorical variables and their graphical representation in 2-dimensional space. MCA is similar to PCA, only difference, MCA works with categorical variables and quantitative variables in PCA. I look forward enhancing my knowledge in learning more methods of multivariate regression.