Home Purchase Prediction

Technical Summary Report

Jun Tae Son | Omer Cheema | Yuchen Wu | Masoud Mansoury | Travis S Donoghue
Spring 2017
CSC 424 Advanced Data Analysis
DePaul University

---

1. Introduction

Every year, millions of homes are sold across the United States adding up to several trillions of dollars in transactions. Thousands of companies have built their business on taking a piece of this pie. These same companies are continuously looking for ways to beat out their competition by spending millions of dollars each year on marketing and buying leads. If there was a predictive model created to better understand who is in the market to purchase a home, that predictive model would be worth millions of dollars.

With that in mind, the purpose of this study is to examine the relationship of consumer demographics along with the consumer's home data in order to predict if they are in the market to purchase a home.

2. Data Description

The data analyzed consisted of three main groups of data with (71) variables, and (170,047) rows consisting of numeric, binary, ordinal, and categorical data (Figure 2).

Group 1. Housing Data *(i.e. home price, sq ft, beds, baths, taxes, property type, year built)*
Group 2. Demographic Data *(i.e. age, sex, marital status, children, purchase score, income)*
Group 3. Lifestyle Data *(i.e. investor, SUV owner, sports interest, outdoor interest, pets)*

3. Exploratory Data Analysis and Pre-processing:

A. Data Cleaning:

We decided to exclude 27 variables that had more than 60% missing rows. In addition, we took 3.4% of samples with complete cases from 170,047 instances to represent the original dataset. To check if the sample data can represent the original, we used dependent variable (PURCHASE), one numeric predictor (PROP_LOANTOVAL), and one categorical predictor (Wealth.Score) to compare the distributions of original and sample data. We selected these three variables based on 1. how statistically significant they were in the analysis, and 2. how many missing cases they had in the original dataset (Figure 3.A1).

As you can see from the Figure 3.A2, the count plot of dependent variable for the original data is similar to the one for the sample data. The boxplots for PROP_LOANTOVAL and the barplot for Wealth.Score also showed the similar pattern. Since the data is missing completely at random, we may conclude that the sample dataset with 5,962 rows could represent our original dataset.

B. Data Transformation:

All seven numeric variables were positively skewed, which caused the violation of normality assumption. Moreover, the unit difference between numeric variables was significantly large. For example, the third quartile for PROP_TAXAMT is 482,183 in dollar amount while the third quartile for

PROP_LOANTOVAL is 92 in ratio (Figure 3.B1). Thus, we decided to scale the numeric variables by applying natural log.

C. Data Reduction:

We had 43 independent variables in the sample data, and the number of predictors in a simple logistic regression would be more than hundreds after generating dummies for categorical variables. Such a model with too many predictors would not give us an interesting result at the end. Thus, we used PCA to reduce the dimension of data in the analysis 1.

D. Correlation Matrix and Multicollinearity:

We usually determine multicollinearity problem by looking at correlation coefficient value (>0.9) and VIF score (>10). According to the correlation Matrix (Figure 3.D1), we failed to find the highly correlated variables that have higher than 0.9 correlation coefficient value among numeric and ordinal predictors. But VIF scores in the simple logistic regression model (Figure 3.D2) shows that PROP_IND, PROP_SALESDEEDCD, and PROP_SALESTRANSCD have higher VIF scores above 10. By rotating the dimension of data using the Principal Component Analysis, we are going to solve this multicollinearity in the first analysis.

4. Analysis 1: Logistic Regression with PCA and CFA

Since there were 44 variables in the dataset, dimension reduction was required before running logistic regression. Originally, the team used Heterogenous Correlation (hetcor) for the numeric and ordinal variables and Tetrachoric Correlation (tetrachoric) to find correlation matrices for nominal data to run Principal Component Analysis. Hetcor function computes a heterogenous correlation matrix, consisting of Pearson product-moment correlations between numeric variables, polyserial correlations between numeric and ordinal variables, and polychoric correlations between ordinal variables. Tetrachor computes tetrachoric correlation matrix which uses inferred Pearson Correlation from a two x two table with the assumption of bivariate normality. Using the scores of principal components, we ran logistic regression with cross validation and found out that there were 23 significant factors in our regression model. However, the professor suggested that we should use less number of factors in our regression analysis to improve our interpretation even if we sacrifice some variance. Additionally, he suggested that we should use only hetcor to find components instead of using both hetcor and tetracor.

Following professor's recommendations we restarted our analysis. To make the data ready for hetcor, we used numeric variables as they were; converted the ordinal data into factors; and created dummies for non binary nominal variables and turned the dummies and binary nominal variables into factors. Finally we ran hetcor on the modified data but for some correlations which involved dummy variables we had NAs in the correlation matrix. We can't run PCA on a correlation matrix with even one NA value in it. To fix the issue we had to remove the dummy variables which were causing the NAs and dummy variables related to them. For instance, PROP_SALESDEEDCDX is one of the dummy variable which was creating the NAs in the correlation and its the original nominal variable was PROP_SALESDEEDCD. We had to remove all dummy variables related to the original variables. After removing the dummy variables we re-ran hetcor and used the correlation matrix to run PCA. The PCA produced 23 components with 90 percent variance (Figure 4.1), however, there were some components that had only one variable in their loadings (Figure 4.2). Therefore, to deal with such components and improve the interpretation of the loadings, we kept on decreasing the number of factors in Component Factor Analysis until we only had few factors with one surrogate variable. After several tries, we decided on using a total of 13 factors for our analysis with cumulative variance of 72% (Figure 4.3).

We ran the logistic regression on the 13 components, however, all the components turned out to be insignificant as you can see in Figure 4.4. This shows that in hope of getting better interpretation we lost some of the important variance in the data which predicted the value of the dependent variable. After manually changing the number of factors in CFA and running logistic regression on the factors we found that we need a minimum of 18 factors to build a logistic regression with significant variables (Figure 4.5). We also performed backward selection on the model (Figure 4.6) but it was a redundant step because all the components were already significant. Finally, after running cross validation on our model, we can conclude that the model can predict with 60% accuracy whether an individual/family will buy a house or not (see figure 4.7).

To conclude our first analysis we will interpret some of the significant factors to show what were our best predictors in the model(Figure 4.8). Below is the explanation about the factors:

Factor	Contributors	Explanation
RC1 (Value of House)	PROP_ASSED_VAL; PROP_MRKTVAL; PROP_TAXAMT; and PROP_SALEAMT	This factor tells us about the value of the house.
RC2 (Lifestyle)	Credit.Card.User; Donates.to.Charity. or.Causes; Parenting. and.Children.s.interest.Bundle; Book.Readers; Hi.Tech.Enthusiasts; Gaming.and.Gambling.Enthusiast; Avid.Investor	This factor tells us the lifestyles of the individual
RC8 (Size of House)	PROP_LIVINGSQFT; PROP_BEDRMS; PROP_FULLBATHS	This factor tells us about size and characteristic of the house.
RC3 (Marital Status)	Single.Parent; Marital.StatusA; Marital.StatusS	Single.Parent and Marital.StatusS (Single) are negative contributors and Marital.StatusA (Inferred Married) is a positive contributor. This factor tells us even if you are assumed as being married you will influence the data set differently from single individuals.
RC5 (Gender)	Appended.Gender.F; Appended.Gender.M	Both of them have equal contribution but different signs. This shows gender doesn't affect the outcome

6. Analysis 2: Correspondence Analysis

In order to explore the relationships between the categorical variables in the dataset, the technique of correspondence analysis was used. There were two different datasets used to complete the correspondence analysis. 1. The complete dataset used for the model that included consumers who purchased a home (1) along with consumers who did not purchase a home (0) and, 2. A dataset that only included the consumers that purchased a home (1). Although the extra step doubled the amount of analysis, the knowledge gained from identifying the correspondence between consumers that purchased a home versus the complete dataset of consumers who purchased a home with consumers who have not purchased a home was very telling. Each correspondence analysis told a unique story about the data.

The first interesting story was told by the two variables Purchase Power and Wealth Score using the dataset containing only consumers who purchased a home in the last six months. To add color, Purchase Power is similar to a FICO score where it ranges from 0-850 with 0 representing a high risk consumer and 850 represents low risk consumer of paying back a loan. Wealth score amounts to the total net worth of the

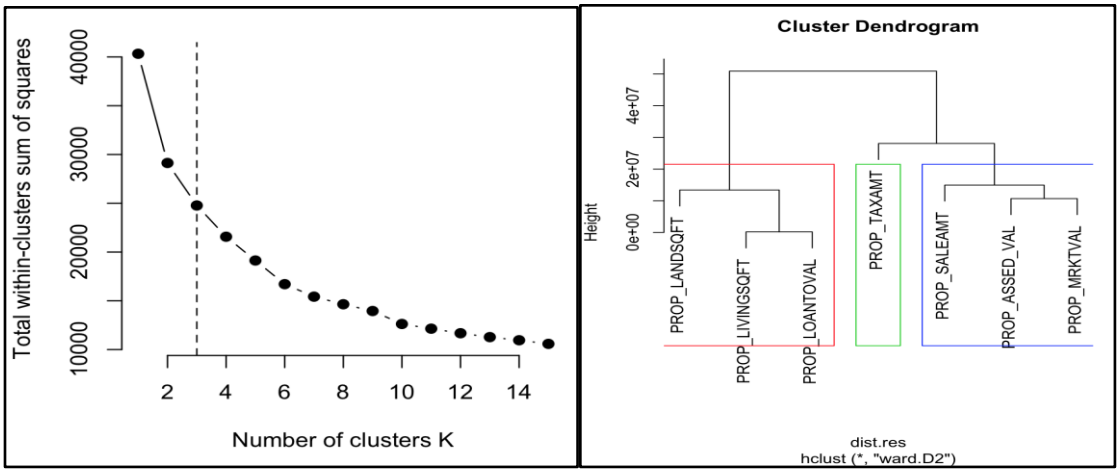
consumer. Reviewing the biplot in Figure 6.1, the consumers that were low risk had high correspondence with consumers that had a low net worth. The story the data is telling involves the 2008 Financial Crisis where many banks went bankrupt due to bad mortgage loans. In response, the banks were strict on which consumers they would provide mortgages for, only providing mortgages to consumers who had high credit and high net worth. Just recently, the banks began to offer more lenient loans for consumers with high credit and low net worth. The biplot in Figure 6.1 and the mosaic plot in Figure 6.2 reiterate this story by illustrating consumers with low net worth and high purchase power have purchased the most homes.

Another interesting story is observed when the two variables Income and Marital Status are analyzed. Within the marital status categorical variable, there is are deterministic values which represent Married (M) or Single (S), and probabilistic data for Inferred Married (A) or Inferred Single (B). From the biplot in Figure 6.3 and the mosaic plot in Figure 6.4, a clear conclusion is observed. Consumers that are single or inferred single fall in the lowest income bracket of <$20,000 (A), and $20,000-$29,999 (B). It is an interesting story and one that spouses would appreciate. Two hypotheses immediately come to mind. The first is that the data includes both spouses' incomes. The second assumption, one that Freud would respect, is the consumer works harder to increase their income to provide for their family.

7. Analysis 3: Cluster Analysis

We used cluster analysis to group a set of variables and compared the meaning of each variable to get a better insight of our data distribution. We tried several new techniques for our analysis. For creating distance matrix, we used ***grow*** distance function. ***Elbow*** and ***Sihouette*** methods helped us to determine the best number of clusters. For clustering the data, we applied ***k-means*** on numeric variables and ***PAM*** on whole of the data.

In the first cluster analysis, we only took into account of the 7 numeric variables. We scaled numeric variables (Prop_assed_val,Prop_mrktval, Prop_landsqft,Prop_livingsqft, Prop_taxamt, Prop_saleamt, and Prop_loantoval). Besides, we used Elbow method to determine the optimal number of clusters for k-means clustering. The factoextra package can be easily computed using the function fviz_nbclust. The result shows three clusters are suggested in our data set. Furthermore, we computed pairwise distance matrices for finding hierarchical clustering result by using euclidean distance and wardD2. The cluster dendrogram shows the seven variables are the leaves. Prop_Taxamt is the simplicifolious and it means the distribution of the variable is substantially different from remaining chunks. The arrangement of the clades interpret which leaves are most similar to each other. Prop_landsqft, Prop_livingsqft and Prop_loantoval variables are belongs to housing attributes, and it is the crucial determinants for customers to purchase a house. Prop_taxamt is the external factor which determined by multiple factors, such as state and federal aid, budgets for the city, etc. Prop_saleamt, Prop_assed_val, and Prop_mrktval are the variables related with real estate market.



In the second cluster analysis we took into account of all the variables; however, we faced two major issues:

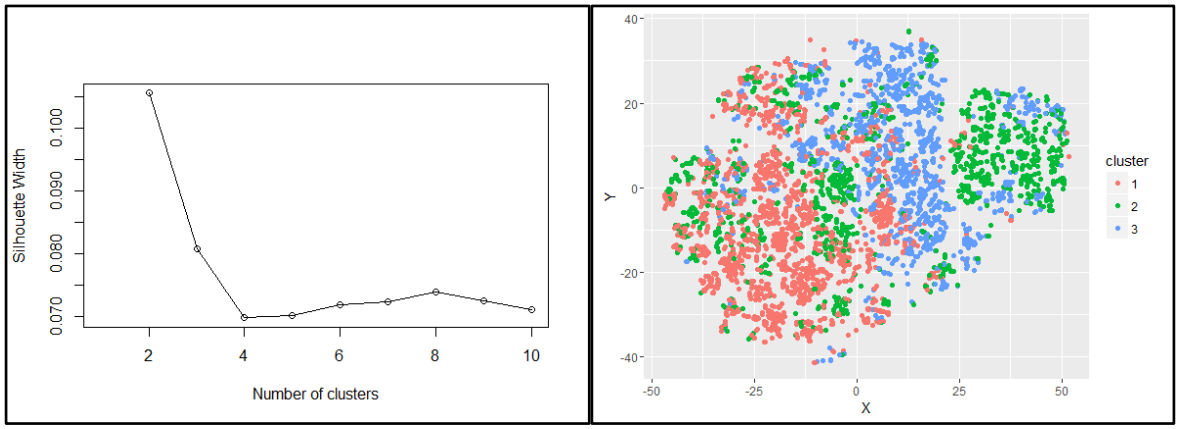
1) Dealing with categorical and numerical variables
2) Visualizing large data to be able to interpret them

For the first one, we need to find a distance function that works on multiple data types and a clustering technique that works well on this distance matrix. We used *Gower* distance function which properly works on multiple data types and also for clustering technique we used *Partitioning Around Medoids (PAM)*. For second issue, interpretation of results, we used descriptive statistics to find the meaning of each cluster. Also, we used scatter plot to visualize the clusters.

Since there are many variables in the dataset, first we need to select some meaningful variables for cluster analysis by using domain knowledge. We used below variables for cluster analysis.

Age	***PROP_MRKTVAL***	***PROP_LIVINGSQFT***	***PROP_BEDRMS***
Estimated.Income	***Wealth.Score***	***Purchasing.Power.Score***	***Home.Market.Value***

We used *Sihouette* to determine the best number of clusters. Below is the result of sihouette which shows that for the specified variables, the best number of clusters is 3. After calculating the distance function and determining the number of clusters, we use them as input for PAM clustering technique. Summary of clusters are shown in Figure 7.1. Also, below is the scatter plot which shows the distribution of clusters:



Interpretation of each cluster:

- ***Cluster 1:*** *Wealthiest people with higher income, but their purchasing power is lower than people in other clusters. Also, people in this cluster tend to buy homes with higher market value.*
- ***Cluster 2:*** *People with lowest wealth score and lowest income, but their purchasing power is higher than people in other clusters. Also, people in this cluster do not care for market value of the home.*
- ***Cluster 3:*** *People in this cluster are almost between people in clusters 1 and 2 in terms of wealth score, income, and purchasing power. People in this cluster tend to buy home with more bedrooms.*

8. Conclusion:

Data science and predictive model building can be interpreted with different accuracies depending on the steps taken through the process. The main steps taken in building this model involved data cleaning, principal component analysis, correspondence analysis, cluster analysis, and logistic regression. The final regression model obtained an accuracy of 60%. Interpreting the factors showed that the value of individual's/family's current house, size of their current house, their lifestyle, and their marital status are one of the most important predictors of whether they going to buy a house or not.

People with higher wealth score mainly consider the market value of property as an important indicator for buying home. While people with lower wealth score do not consider market value of property when buying a home. Also, another interesting pattern in clusters is that wealthier people have lower purchasing power. Besides, the key factors for people to purchase a house are the housing features, tax rates, and real estate market. Even though when a customer likes the housing feature, but if the real estate market tends to be at a high peak, the customer will not choose to purchase a house.

Future work on the model will be conducted in a couple ways. First, the data will be dissected determining a) which variables cost the most to collect, and b) which variables have the most NULL values with least effect on the model. The parameters found in b) will be removed in order to put as many records as possible through the model. Continuous monitoring of the model and testing will be conducted insuring new leads fall within the 60% accuracy range.

Appendix:

Plots:

Figure 2.1

< Group 1. Housing Data >

REAL ESTATE AND PROPERTY

Assessor Parcel Number	**Property Type**	**Property Construction**	**Property Style**
Year Home Built	Single Family	Wood	Colonial
Home Value	Commercial	Brick	Bungalow
Appraised Value	Condominium	Stucco	Art Deco
Home Square Footage	Apartment	Metal Frame	Coach
Property Square Footage	**Property Condition**	Concrete	A Frame
Property Acreage	Excellent	Masonry	Victorian
Owner-occupied	Very Good	Glass	Tudor
Number of Rooms	Good	**Roof Type**	Spanish
Number of Bedrooms	Fair	Clay Tile	Mediterranean
Number of Stories	Poor	Aluminum	**Pool**
Number of Baths	Unsound	Shake	**Fireplace**
Subdivision	Under Construction	**Garage**	
Mobile Home	**Fuel Type**	**Basement**	
Heating Type	**Water Type**	**Air Conditioning**	

Figure 2.2

< Group 2. Demographic Data >

CONSUMER DEMOGRAPHIC DATA

Age	**Income**	**Gender**	**Homeowner**
Year Home Built	**Length of Residence**	**Home Value**	**Renter**
Presence of Children	**Number of Children**	**Marital Status**	**Education**
Ethnicity	**Dwelling Type**	**Employment Type**	**Credit Card User**
Wealth Score	**Children Age**	**Charitable Donor**	**Car Ownership**

Figure 2.3

< Group 3. Lifestyle Data >

LIFESTYLE ATTRIBUTES

Women's Apparel	**Hobbies, Home & Garden**	**Travel Enthusiast**	**Price Club Participant**
Plus Size Apparel	Sewing and Knitting	General Travel	**Parenting Interest**
Men's Apparel	Woodworking	Domestic	**Automotive DIY**
Big and Tall	Photography	International	**Sports Spectator**
Pet Lovers	**Home and Garden**	Cruise	General Sports
Cat Owner	Home Improvement	**Physical Fitness**	Football
Dog Owner	Cooking and Wine Enthusiast	Health Exercise	Baseball
Equestrian	Gourmet Food and Wine	Running	Golf
Art Interest	Cooking and Wine Enthusiast	Walking	Tennis
Arts	Natural Foods	Aerobics	Auto Racing
Music	**Avid Investor**	**Hi-Tech Enthusiast**	**Outdoor Enthusiast**
Antiques	**Gaming and Gambling**	**Self-Improvement**	General Outdoors
Performing Arts	**Motorcycle Enthusiast**	Health & Medical	Snow Sports
Collectibles Interest	**Boat Enthusiast**	Dieting Weight Loss	Water Sports
Antique Collectibles	**Book Buyer**	Gen. Self-Improvement	Hunting and Fishing
Sports Collectibles	**Book Reader**		

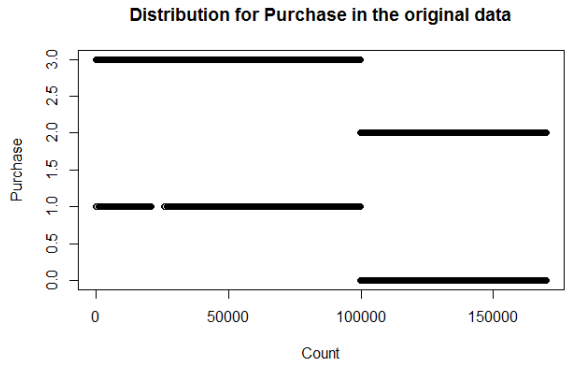
Figure 3.A1

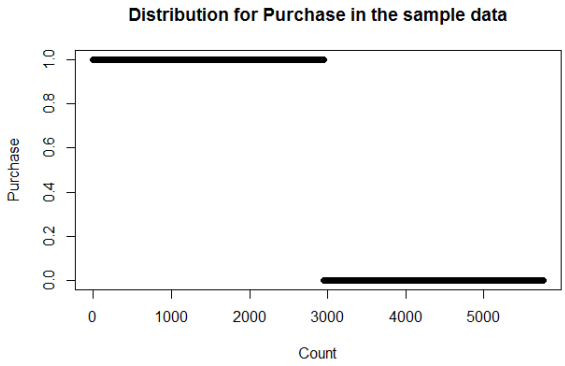
< A set of variables with missing cases >

```
> sapply(missing, function(x) round(sum(is.na(x))/nrow(missing),2))
        PROP_LOANTOVAL             PROP_BEDRMS           PROP_FULLBATHS     Length.of.Residence
                  0.29                    0.20                    0.04                    0.33
          PROP_MTGTERM       Home.Market.Value       Estimated.Income           Wealth.Score
                  0.23                    0.36                    0.33                    0.43
Purchasing.Power.Score               Education      Number.of.Children          PROP_MTGLOANCD
                  0.33                    0.52                    0.56                    0.22
  Appended.Gender....1          Marital.Status
                  0.38                    0.39
```

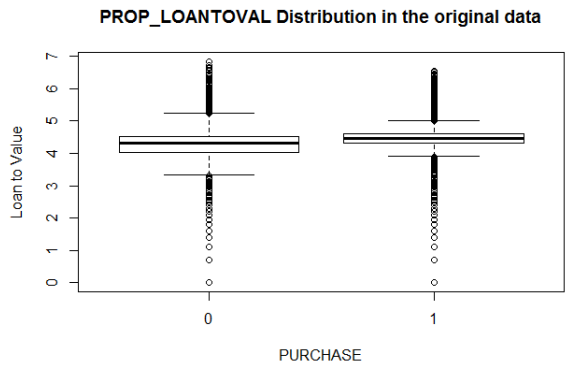
Figure 3.A2

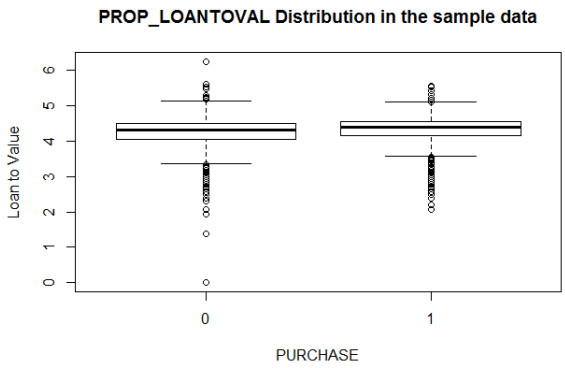
< Comparison of distributions for original and sample data >

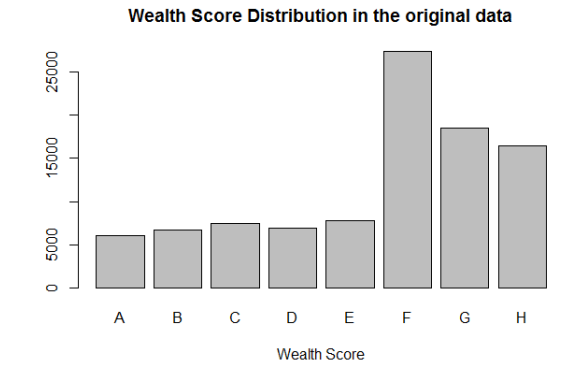




We assigned a value of "3" if a person bought a house (PURCHASE = 1) and some information are missing, and a value of "2" if a person didn't purchase a house (PURCHASE = 2) and some information are missing.







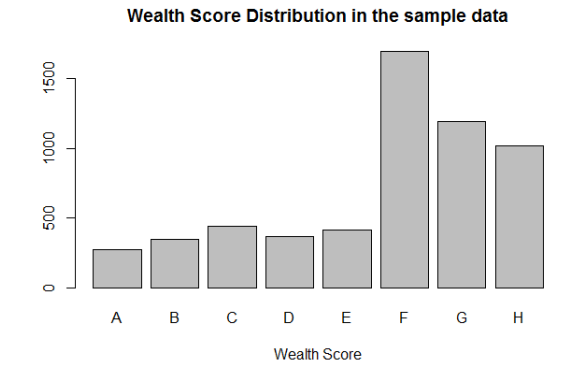


Figure 3.B1

< Summary of seven numeric variables >

```
> summary(numeric)
 PROP_ASSED_VAL     PROP_MRKTVAL      PROP_LANDSQFT     PROP_LIVINGSQFT  PROP_TAXAMT
 Min.   :   1920   Min.   :  11075   Min.   :      1   Min.   :  480   Min.   :    252
 1st Qu.:  56887   1st Qu.: 139800   1st Qu.:   7314   1st Qu.: 1497   1st Qu.: 188728
 Median : 134796   Median : 215000   Median :  10542   Median : 1994   Median : 301208
 Mean   : 200636   Mean   : 272060   Mean   :  34951   Mean   : 2201   Mean   : 395841
 3rd Qu.: 258775   3rd Qu.: 327060   3rd Qu.:  19490   3rd Qu.: 2687   3rd Qu.: 482813
 Max.   :6340000   Max.   :6340000   Max.   :5227200   Max.   :10435   Max.   :8079963
  PROP_SALEAMT      PROP_LOANTOVAL
 Min.   :    100   Min.   :  1.00
 1st Qu.: 130373   1st Qu.: 61.00
 Median : 202806   Median : 78.00
 Mean   : 259314   Mean   : 75.69
 3rd Qu.: 315000   3rd Qu.: 92.00
 Max.   :4050000   Max.   :522.00
```

Figure 3.B2

< Distribution of each numeric variables before and after log transformation >

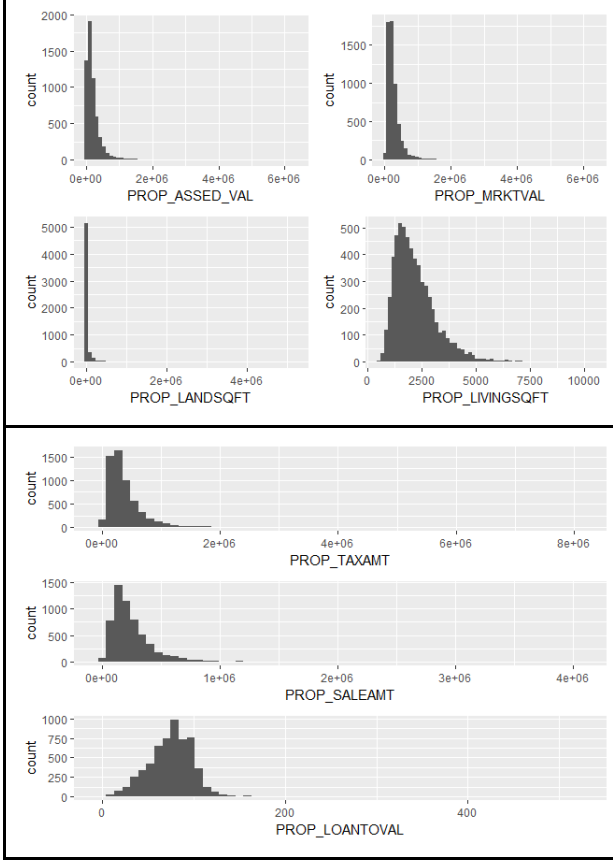
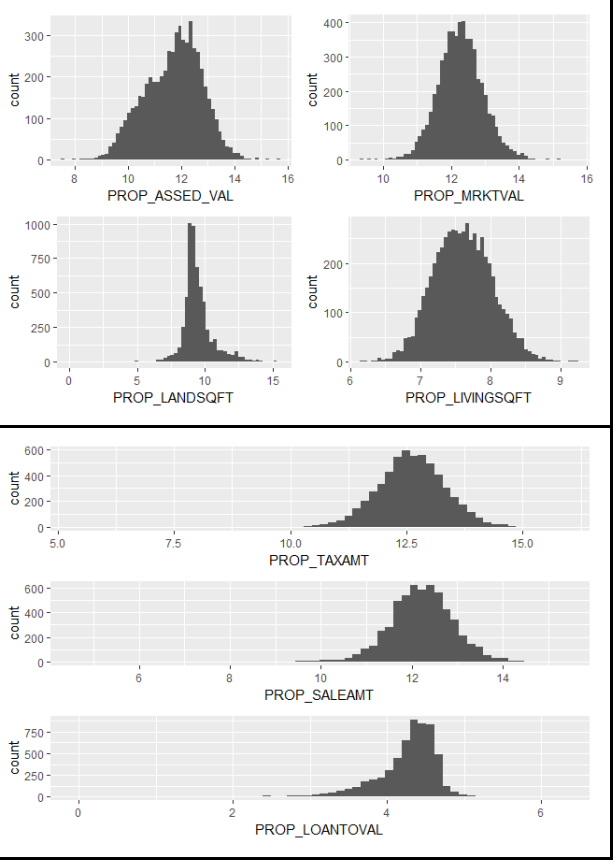
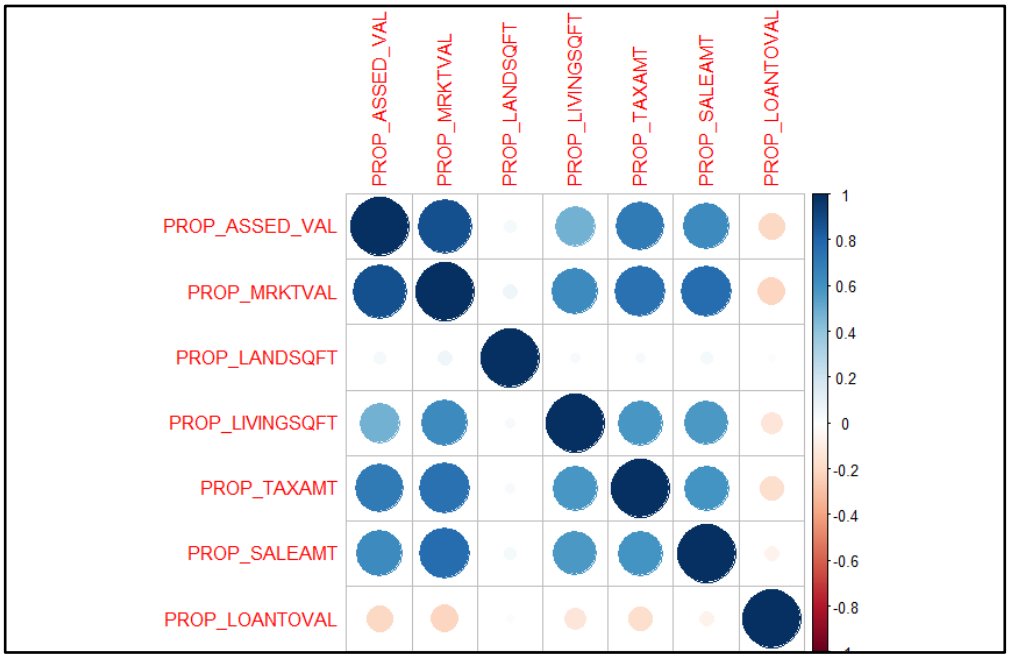
The distribution of numeric variables before log transformation.	The distribution of numeric variables after log transformation
	

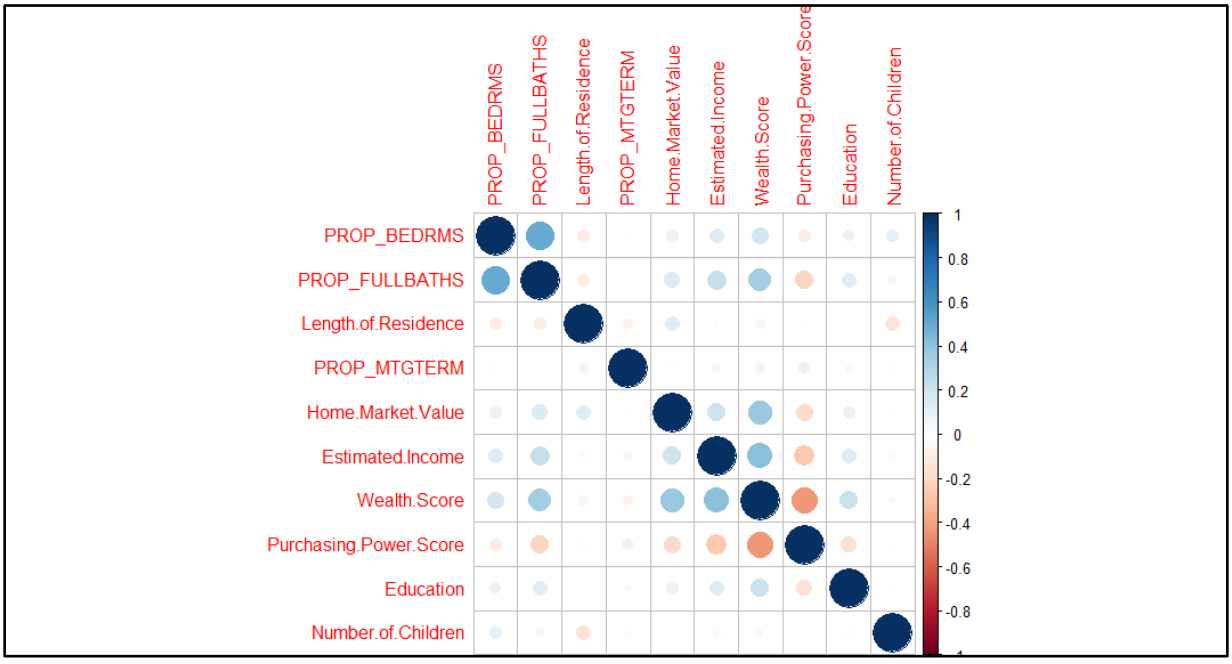
Figure 3.D1

< Correlation Plot for Numeric Variables >



	PROP_ASSED_VAL	PROP_MRKTVAL	PROP_LANDSQFT	PROP_LIVINGSQFT	PROP_TAXAMT	PROP_SALEAMT	PROP_LOANTOVAL
PROP_ASSED_VAL	1.00000000	0.87318619	0.04794541	0.47621379	0.70498663	0.63153188	-0.20783194
PROP_MRKTVAL	0.87318619	1.00000000	0.06295264	0.63153459	0.74125158	0.76984270	-0.21982043
PROP_LANDSQFT	0.04794541	0.06295264	1.00000000	0.03267441	0.03125792	0.04428101	-0.01936755
PROP_LIVINGSQFT	0.47621379	0.63153459	0.03267441	1.00000000	0.58027206	0.57208813	-0.13415156
PROP_TAXAMT	0.70498663	0.74125158	0.03125792	0.58027206	1.00000000	0.59161024	-0.17811627
PROP_SALEAMT	0.63153188	0.76984270	0.04428101	0.57208813	0.59161024	1.00000000	-0.06353069
PROP_LOANTOVAL	-0.20783194	-0.21982043	-0.01936755	-0.13415156	-0.17811627	-0.06353069	1.00000000

< Correlation Plot for Ordinal Variables >



	PROP_BEDRMS	PROP_FULLBATHS	Length.of.Residence	PROP_MTGTERM	Home.Market.Value	Estimated.Income	Wealth.Score	Purchasing.Power.Score	Education	Number.of.Children
PROP_BEDRMS	1.00000000	0.505723958	-0.100770981	-0.010349192	0.099924185	0.14771127	0.18535959	-0.104534397	0.087087499	0.107095809
PROP_FULLBATHS	0.50572396	1.000000000	-0.100828501	0.004842165	0.156286939	0.23343029	0.34441830	-0.216286949	0.130197144	0.047339753
Length.of.Residence	-0.10077098	-0.100828501	1.000000000	-0.065180958	0.135356448	0.02858849	0.05970043	-0.018199446	-0.006549868	-0.149632115
PROP_MTGTERM	-0.01034919	0.004842165	-0.065180958	1.000000000	-0.015678780	-0.04747262	-0.06748706	0.084108959	-0.037356328	0.017052354
Home.Market.Value	0.09992419	0.156286939	0.135356448	-0.015678780	1.000000000	0.20421219	0.37278159	-0.182828165	0.098482941	-0.008554159
Estimated.Income	0.14771127	0.233430292	0.028588488	-0.047472622	0.204212193	1.00000000	0.41480962	-0.267240776	0.146059375	0.035800539
Wealth.Score	0.18535959	0.344418298	0.059700427	-0.067487057	0.372781590	0.41480962	1.00000000	-0.434614755	0.212688351	-0.031859871
Purchasing.Power.Score	-0.10453440	-0.216286949	-0.018199446	0.084108959	-0.182828165	-0.26724078	-0.43461475	1.000000000	-0.155917528	0.006020992
Education	0.08708750	0.130197144	-0.006549868	-0.037356328	0.098482941	0.14605938	0.21268835	-0.155917528	1.000000000	-0.017352268
Number.of.Children	0.10709581	0.047339753	-0.149632115	0.017052354	-0.008554159	0.03580054	-0.03185987	0.006020992	-0.017352268	1.000000000

Figure 3.D2

< VIF scores for each predictor in a simple logistic regression model >

	GVIF	Df	GVIF^(1/(2*Df))
PROP_ASSED_VAL	-0.01	1	NaN
PROP_MRKTVAL	-1.42	1	NaN
PROP_LANDSQFT	0.17	1	0.42
PROP_LIVINGSQFT	-0.08	1	NaN
PROP_TAXAMT	-0.12	1	NaN
PROP_SALEAMT	1.33	1	1.15
PROP_LOANTOVAL	0.22	1	0.47
PROP_BEDRMS	-0.01	8	NaN
PROP_FULLBATHS	0.10	5	0.79
Length.of.Residence	-0.36	2	NaN
PROP_MTGTERM	0.03	2	0.41
Home.Market.Value	0.60	18	0.99
Estimated.Income	0.55	7	0.96
Wealth.Score	-0.17	7	NaN
Purchasing.Power.Score	0.97	19	1.00
Education	1.99	3	1.12
Number.of.Children	1.36	2	1.08
Mobile.Home.Indicator	0.41	1	0.64
Single.Parent	5.57	1	2.36
Fireplace.in.Home	1.87	1	1.37
Pool.Owner	1.69	1	1.30
Senior.in.HouseHold	0.96	1	0.98
Credit.Card.User	2.58	1	1.61
Donator.to.Charity.or.Causes	1.29	1	1.14
Luxury.Vehicle.Owner	1.94	1	1.39
SUV.Owner	0.83	1	0.91
Pickup.Truck.Owner	1.09	1	1.04
Price.Club.and.value.Purchasing.Indicator	0.95	1	0.97
Parenting.and.Children.s.interest.Bundle	0.64	1	0.80
Book.Readers	1.28	1	1.13
Hi.Tech.Enthusiasts	1.90	1	1.38
Gaming.and.Gambling.Enthusiast	1.19	1	1.09
Avid.Investors	1.04	1	1.02
PROP_IND10	95.17	1	9.76
PROP_IND11	45.66	1	6.76
PROP_IND21	21.13	1	4.60
PROP_SALESDEEDCDG	-1631082000000000	1	NaN

PROP_SALESDEEDCDQ	-36246270000000	1	NaN
PROP_SALESDEEDCDU	1250496000000000	1	35362360
PROP_SALESDEEDCDX	960526300000000	1	30992360
PROP_SALESTRANSCD1	20.04	1	4.48
PROP_SALESTRANSCD2	3.61	1	1.90
PROP_SALESTRANSCD3	17.32	1	4.16
PROP_SALESTRANSCD6	1.98	1	1.41
PROP_SALESTRANSCD7	1.94	1	1.39
PROP_MTGLOANCDCNV	2.99	1	1.73
PROP_MTGLOANCDFHA	2.82	1	1.68
PROP_MTGLOANCDPP	1.49	1	1.22
PROP_MTGLOANCDSBA	1.02	1	1.01
Appended.Gender....1F	1.10	1	1.05
Marital.StatusA	2.44	1	1.56
Marital.StatusB	1.05	1	1.03
Marital.StatusM	4.56	1	2.13

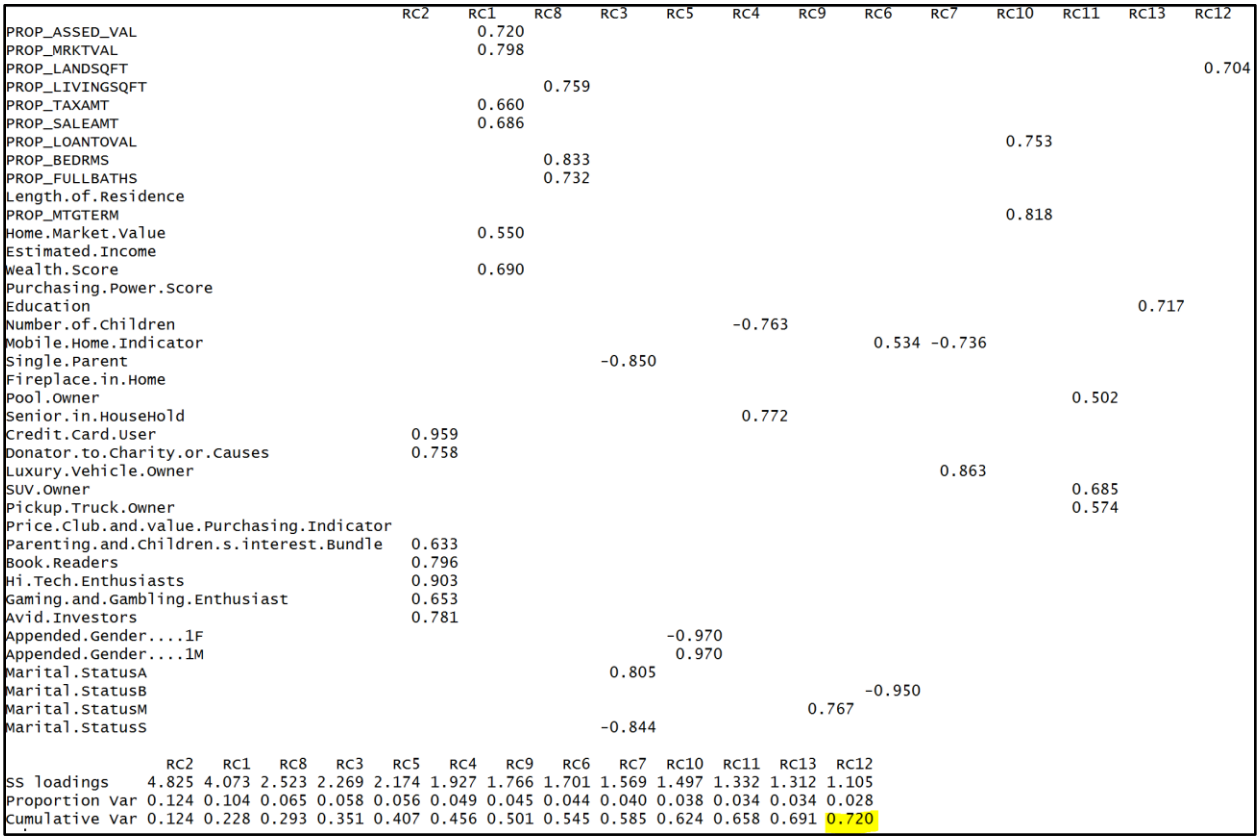
Figure 4.1

```
Importance of components:
                          Comp.1    Comp.2     Comp.3     Comp.4    Comp.5     Comp.6     Comp.7     Comp.8
Standard deviation     2.4993612 2.2439649 1.72594491 1.47739351 1.3230216 1.25344913 1.21956473 1.15063239
Proportion of Variance 0.1601745 0.1291123 0.07638169 0.05596645 0.0448817 0.04028551 0.03813688 0.03394756
Cumulative Proportion  0.1601745 0.2892868 0.36566848 0.42163493 0.4665166 0.50680213 0.54493901 0.57888657
                          Comp.9    Comp.10    Comp.11    Comp.12    Comp.13    Comp.14    Comp.15    Comp.16
Standard deviation     1.09890727 1.07855449 1.04601838 1.03420908 0.98032435 0.97070707 0.92838349 0.91921096
Proportion of Variance 0.03096403 0.02982769 0.02805524 0.02742534 0.02464194 0.02416083 0.02209989 0.02166535
Cumulative Proportion  0.60985060 0.63967829 0.66773353 0.69515887 0.71980082 0.74396165 0.76606154 0.78772689
                         Comp.17    Comp.18    Comp.19    Comp.20    Comp.21    Comp.22    Comp.23    Comp.24
Standard deviation     0.89579518 0.86320967 0.82671173 0.8167636 0.80200384 0.75898631 0.75492091 0.73765195
Proportion of Variance 0.02057562 0.01910592 0.01752442 0.0171052 0.01649257 0.01477077 0.01461296 0.01395206
Cumulative Proportion  0.80830251 0.82740843 0.84493285 0.8620380 0.87853062 0.89330139 0.90791435 0.92186641
```

Figure 4.2

	RC21	RC9	RC16	RC10	RC23	RC19	RC15	RC14	RC13	RC18
PROP_ASSED_VAL										
PROP_MRKTVAL										
PROP_LANDSQFT										
PROP_LIVINGSQFT										
PROP_TAXAMT										
PROP_SALEAMT										
PROP_LOANTOVAL					0.868					
PROP_BEDRMS										
PROP_FULLBATHS										
Length.of.Residence						0.893				
PROP_MTGTERM				0.897						
Home.Market.Value			0.948							
Estimated.Income										
Wealth.Score										
Purchasing.Power.Score	-0.910									
Education										
Number.of.Children										
Mobile.Home.Indicator										
Single.Parent										
Fireplace.in.Home									0.947	
Pool.Owner										
Senior.in.HouseHold										
Credit.Card.User										
Donator.to.Charity.or.Causes										
Luxury.Vehicle.Owner										
SUV.Owner							0.967			
Pickup.Truck.Owner										0.964
Price.Club.and.value.Purchasing.Indicator								0.903		
Parenting.and.Children.s.interest.Bundle										
Book.Readers										
Hi.Tech.Enthusiasts										
Gaming.and.Gambling.Enthusiast										
Avid.Investors										
Appended.Gender....1F										
Appended.Gender....1M										
Marital.StatusA		-0.672								
Marital.StatusB										
Marital.StatusM		0.750								
Marital.StatusS										

Figure 4.3



	RC2	RC1	RC8	RC3	RC5	RC4	RC9	RC6	RC7	RC10	RC11	RC13	RC12
PROP_ASSED_VAL		0.720											
PROP_MRKTVAL		0.798											
PROP_LANDSQFT													0.704
PROP_LIVINGSQFT			0.759										
PROP_TAXAMT		0.660											
PROP_SALEAMT		0.686											
PROP_LOANTOVAL										0.753			
PROP_BEDRMS			0.833										
PROP_FULLBATHS			0.732										
Length.of.Residence													
PROP_MTGTERM										0.818			
Home.Market.Value		0.550											
Estimated.Income													
Wealth.Score		0.690											
Purchasing.Power.Score													
Education												0.717	
Number.of.Children						-0.763							
Mobile.Home.Indicator								0.534	-0.736				
Single.Parent				-0.850									
Fireplace.in.Home													
Pool.Owner											0.502		
Senior.in.HouseHold						0.772							
Credit.Card.User	0.959												
Donator.to.Charity.or.Causes	0.758												
Luxury.Vehicle.Owner									0.863				
SUV.Owner											0.685		
Pickup.Truck.Owner											0.574		
Price.Club.and.value.Purchasing.Indicator													
Parenting.and.Children.s.interest.Bundle	0.633												
Book.Readers	0.796												
Hi.Tech.Enthusiasts	0.903												
Gaming.and.Gambling.Enthusiast	0.653												
Avid.Investors	0.781												
Appended.Gender....1F					-0.970								
Appended.Gender....1M					0.970								
Marital.StatusA				0.805									
Marital.StatusB								-0.950					
Marital.StatusM							0.767						
Marital.StatusS				-0.844									

	RC2	RC1	RC8	RC3	RC5	RC4	RC9	RC6	RC7	RC10	RC11	RC13	RC12
SS loadings	4.825	4.073	2.523	2.269	2.174	1.927	1.766	1.701	1.569	1.497	1.332	1.312	1.105
Proportion Var	0.124	0.104	0.065	0.058	0.056	0.049	0.045	0.044	0.040	0.038	0.034	0.034	0.028
Cumulative Var	0.124	0.228	0.293	0.351	0.407	0.456	0.501	0.545	0.585	0.624	0.658	0.691	0.720

Figure 4.4

```
Call:
glm(formula = PURCHASE ~ ., family = binomial(link = "logit"),
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8968  -1.1400   0.7179   1.1114   1.9417

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.016e-02  2.721e-02   1.843   0.0653 .
RC2          1.849e+01  1.375e+01   1.344   0.1788
RC1          1.179e+02  8.762e+01   1.345   0.1785
RC8         -1.109e+02  8.237e+01  -1.347   0.1781
RC3          1.904e+02  1.417e+02   1.344   0.1791
RC5          4.818e+01  3.593e+01   1.341   0.1800
RC4         -8.650e+00  6.507e+00  -1.329   0.1838
RC9         -5.561e+02  4.138e+02  -1.344   0.1791
RC6         -1.038e+03  7.727e+02  -1.344   0.1790
RC7         -7.181e+01  5.340e+01  -1.345   0.1787
RC10        -4.078e+01  3.059e+01  -1.333   0.1826
RC11        -6.585e+01  4.919e+01  -1.339   0.1807
RC13        -5.645e+01  4.205e+01  -1.342   0.1795
RC12         7.362e-01  5.927e-01   1.242   0.2142
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7984.5  on 5761  degrees of freedom
Residual deviance: 7619.3  on 5748  degrees of freedom
AIC: 7647.3
```

Figure 4.5

```
Call:
glm(formula = PURCHASE ~ ., family = binomial(link = "logit"),
    data = data4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9097  -1.1352   0.7091   1.1084   1.9334

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.077e-02  2.728e-02   1.861   0.0628 .
RC2          1.708e+01  8.358e+00   2.044   0.0410 *
RC1          1.017e+02  4.974e+01   2.045   0.0409 *
RC8         -1.305e+02  6.376e+01  -2.046   0.0407 *
RC3          3.009e+02  1.473e+02   2.043   0.0411 *
RC5          8.203e+01  4.021e+01   2.040   0.0414 *
RC4          4.246e+01  2.082e+01   2.040   0.0414 *
RC9         -9.426e+02  4.615e+02  -2.043   0.0411 *
RC6         -1.554e+03  7.610e+02  -2.043   0.0411 *
RC10        -4.830e+01  2.381e+01  -2.029   0.0425 *
RC7         -3.268e+02  1.599e+02  -2.043   0.0411 *
RC13         3.054e+01  1.490e+01   2.050   0.0404 *
RC16         1.605e+02  7.857e+01   2.043   0.0410 *
RC14         9.307e+01  4.554e+01   2.044   0.0410 *
RC11         2.332e+02  1.141e+02   2.045   0.0409 *
RC15        -6.457e+01  3.168e+01  -2.038   0.0415 *
RC18        -2.205e+01  1.084e+01  -2.034   0.0420 *
RC12         6.114e+01  2.995e+01   2.042   0.0412 *
RC17        -2.414e+01  1.182e+01  -2.042   0.0411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7984.5  on 5761  degrees of freedom
Residual deviance: 7593.0  on 5743  degrees of freedom
AIC: 7631
```

Figure 4.6

```
> summary(model_step)

Call:
glm(formula = PURCHASE ~ RC2 + RC1 + RC8 + RC3 + RC5 + RC4 +
    RC9 + RC6 + RC10 + RC7 + RC13 + RC16 + RC14 + RC11 + RC15 +
    RC18 + RC12 + RC17, family = binomial(link = "logit"), data = data4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9097  -1.1352   0.7091   1.1084   1.9334

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.077e-02  2.728e-02   1.861   0.0628 .
RC2          1.708e+01  8.358e+00   2.044   0.0410 *
RC1          1.017e+02  4.974e+01   2.045   0.0409 *
RC8         -1.305e+02  6.376e+01  -2.046   0.0407 *
RC3          3.009e+02  1.473e+02   2.043   0.0411 *
RC5          8.203e+01  4.021e+01   2.040   0.0414 *
RC4          4.246e+01  2.082e+01   2.040   0.0414 *
RC9         -9.426e+02  4.615e+02  -2.043   0.0411 *
RC6         -1.554e+03  7.610e+02  -2.043   0.0411 *
RC10        -4.830e+01  2.381e+01  -2.029   0.0425 *
RC7         -3.268e+02  1.599e+02  -2.043   0.0411 *
RC13         3.054e+01  1.490e+01   2.050   0.0404 *
RC16         1.605e+02  7.857e+01   2.043   0.0410 *
RC14         9.307e+01  4.554e+01   2.044   0.0410 *
RC11         2.332e+02  1.141e+02   2.045   0.0409 *
RC15        -6.457e+01  3.168e+01  -2.038   0.0415 *
RC18        -2.205e+01  1.084e+01  -2.034   0.0420 *
RC12         6.114e+01  2.995e+01   2.042   0.0412 *
RC17        -2.414e+01  1.182e+01  -2.042   0.0411 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7984.5  on 5761  degrees of freedom
Residual deviance: 7593.0  on 5743  degrees of freedom
AIC: 7631
```

Figure 4.7

	RC2	RC1	RC8	RC3	RC5
PROP_ASSED_VAL		0.831			
PROP_MRKTVAL		0.788			
PROP_LANDSQFT					
PROP_LIVINGSQFT			0.738		
PROP_TAXAMT		0.780			
PROP_SALEAMT		0.671			
PROP_LOANTOVAL					
PROP_BEDRMS			0.856		
PROP_FULLBATHS			0.731		
Length.of.Residence					
PROP_MTGTERM					
Home.Market.Value					
Estimated.Income					
Wealth.Score					
Purchasing.Power.Score					
Education					
Number.of.Children					
Mobile.Home.Indicator					
Single.Parent				-0.852	
Fireplace.in.Home					
Pool.Owner					
Senior.in.HouseHold					
Credit.Card.User	0.937				
Donator.to.Charity.or.Causes	0.750				
Luxury.Vehicle.Owner					
SUV.Owner					
Pickup.Truck.Owner					
Price.Club.and.value.Purchasing.Indicator					
Parenting.and.Children.s.interest.Bundle	0.635				
Book.Readers	0.770				
Hi.Tech.Enthusiasts	0.931				
Gaming.and.Gambling.Enthusiast	0.677				
Avid.Investors	0.794				
Appended.Gender....1F					-0.971
Appended.Gender....1M					0.971
Marital.StatusA				0.818	
Marital.StatusB					
Marital.StatusM					
Marital.StatusS				-0.833	

	RC2	RC1	RC8	RC3	RC5	RC4	RC9
SS loadings	4.700	3.664	2.315	2.252	2.166	1.747	1.708
Proportion Var	0.121	0.094	0.059	0.058	0.056	0.045	0.044
Cumulative Var	0.121	0.214	0.274	0.332	0.387	0.432	0.476

Figure 4.8

```
> cvmodel$results
  parameter  Accuracy
1      none 0.5956305
```

Figure 6.1

< Purchase Power vs Wealth Power >

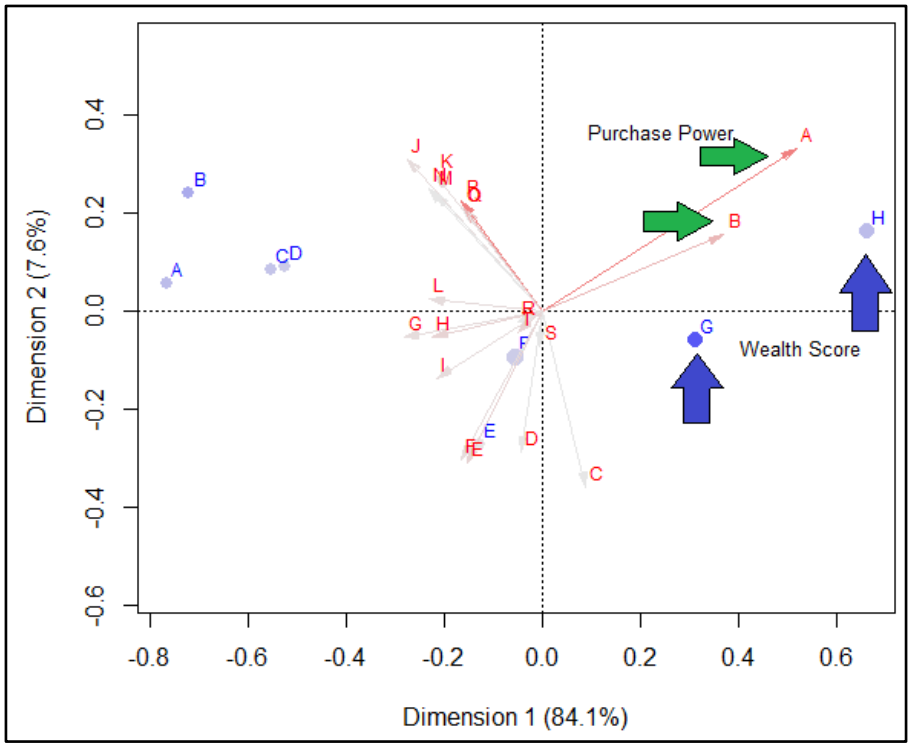


Figure 6.2

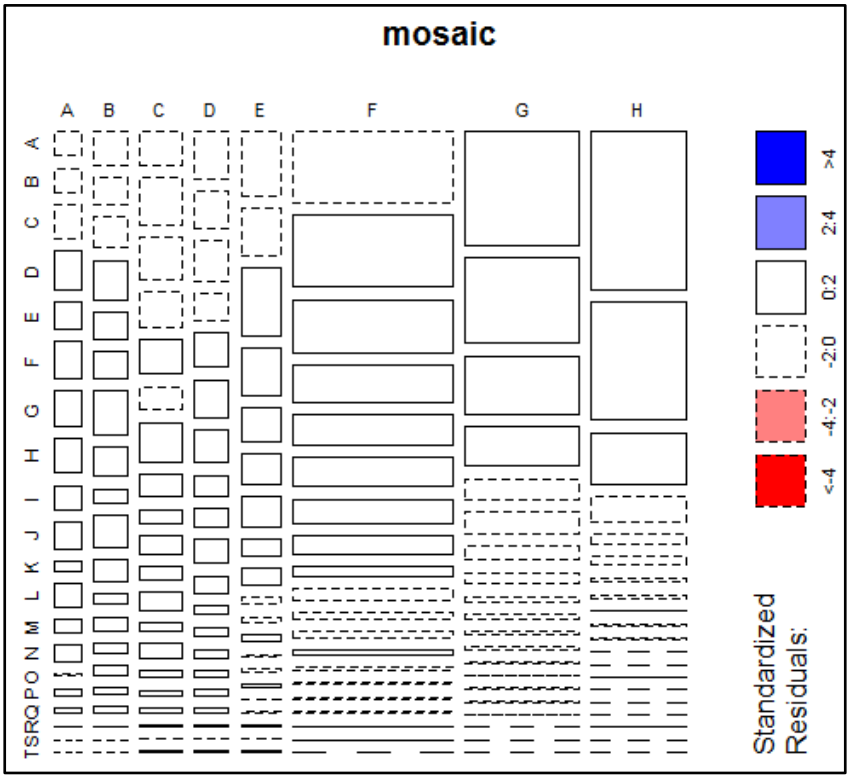


Figure 6.3

<Marital Status vs Income>

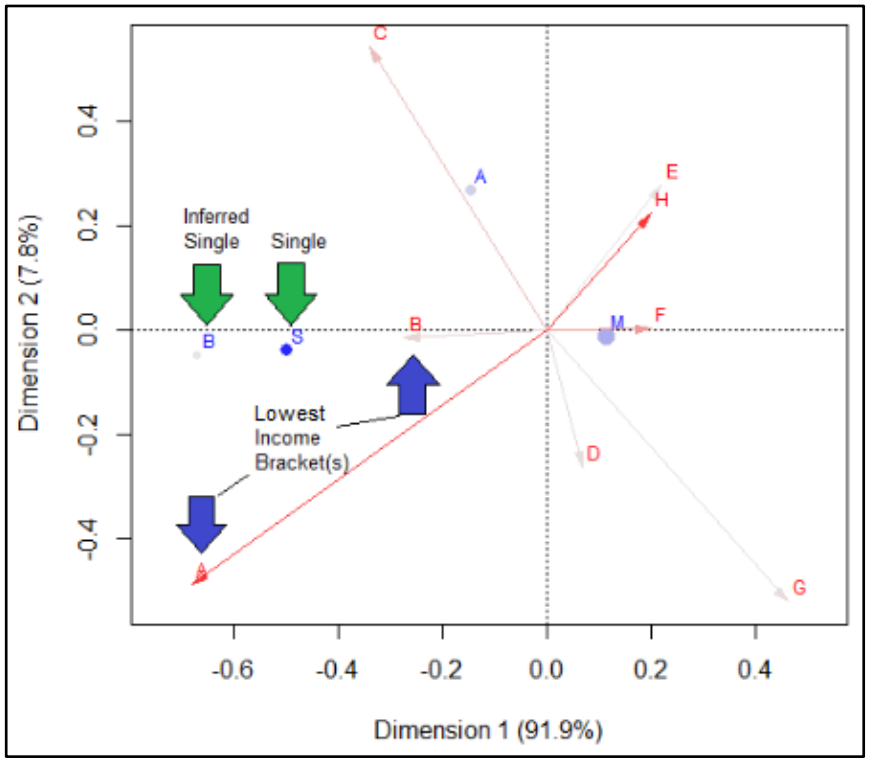


Figure 6.4