**Equity in Mortgage Lending**

**Research Question**Title VIII of The Civil Rights Act of 1968, also known as the Fair Housing Act, prohibits “discrimination concerning the sale, rental, and financing of housing based on race, religion, national origin, and since 1974, sex.” (“Civil Rights Act of 1968”, 2020). In an effort to gather relevant data to assess the equity of mortgage lending practices in the communities that are served by lenders, the Home Mortgage Disclosure Act (HMDA) was enacted in 1975 and requires mortgage lending and other financial institutions to maintain and report loan-level information about mortgage applications. The dataset includes lending institution information and type of loan sought, as well as borrower demographics and information about the property to be purchased. The data is anonymized by the Consumer Financial Protection Bureau and then made available for public analysis to assess the lending practices of local financial institutions for bias (“Background and Purpose of HMDA”, 2018). Various entities, from news organizations to the Bureau itself perform analyses on the data and report their findings. Summary analyses are completed and published by the Bureau, but deeper follow-up analyses need to be conducted to continually assess the equity in lending practices and report the findings to the public. The data that is collected includes information about the borrower race, ethnicity and gender, but these factors should not influence the mortgage application process.

Ho: Race, ethnicity or gender do not influence the mortgage acceptance or denial decisions by lending institutions.

Ha: Race, ethnicity or gender have a significant influence in mortgage acceptance or denial decisions by lending institutions.

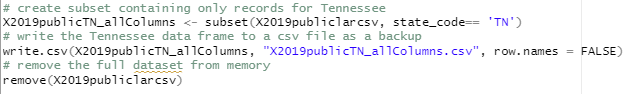
**Data Collection**The HMDA data is provided by year beginning in 1998, through 2019. The dataset for 2019 includes nationwide mortgage application data consisting of 98 independent variables as a mix of qualitative and quantitative variables, and one dependent variable. The full dataset includes 17,545,457 records and is available for download here:  <https://s3.amazonaws.com/cfpb-hmda-public/prod/snapshot-data/2019/2019_public_lar_csv.zip>. The data is anonymized by the CFPB and published for public analysis. The challenge for independent analysts is that datasets this large (6.6GB) can be a challenge to process on a personal computer. This can be mitigated by subsetting the data into smaller datasets by region or lender, for example, for individual processing and analysis.

**Data Extraction and Preparation**Describe your data-extraction and -preparation process and provide screenshots to illustrate each step. Explain the tools and techniques you used for data extraction and data preparation, including how these tools and techniques were used on the data. Justify why you used these particular tools and techniques, including any advantages or disadvantages of these when used with your data-extraction and -preparation methods.

The initial import of the full dataset requires referencing the HMDA data documentation to identify the expected datatypes for each column for import. This process reveals that some columns that one would expect contain only numeric data, such as the property\_value and income columns, to have some records that have character values; the import script was modified to accommodate these unexpected values to be examined in the context of the additional data and handled appropriately after import.

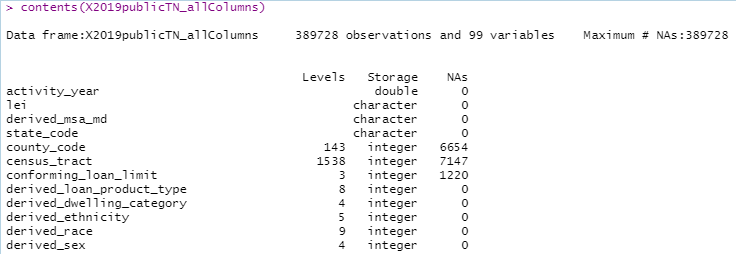


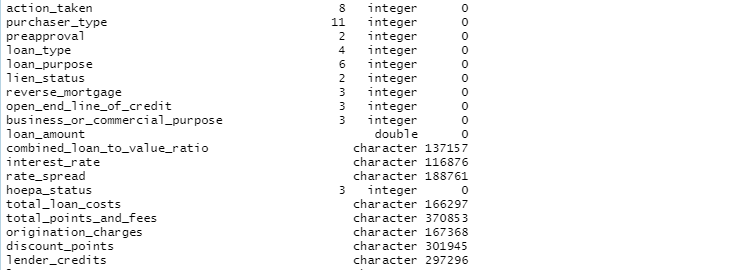
For this study we created a subset of the full dataset to reduce the analysis to include only the state of Tennessee.

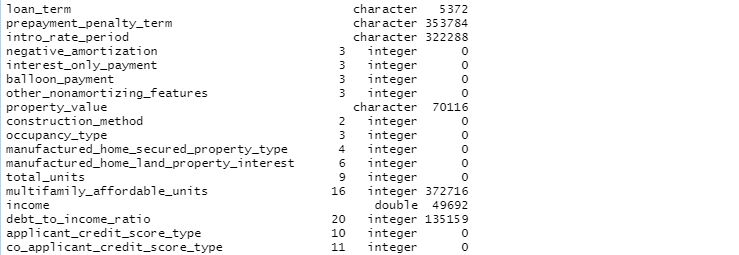


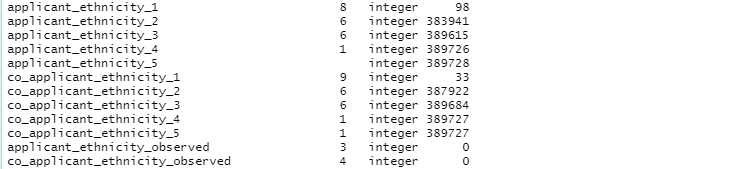
This subset of data reduced the full nationwide dataset to 389,728 rows to include only records for the state of Tennessee.

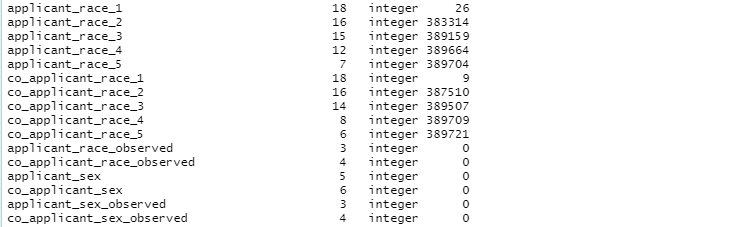
To begin, running the **contents()** method from the Hmisc library provides a quick way to see the type of data in each column and the number of NA rows to begin removing columns and rows that we cannot use for analysis.

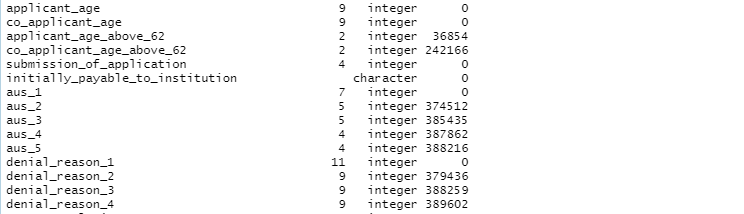


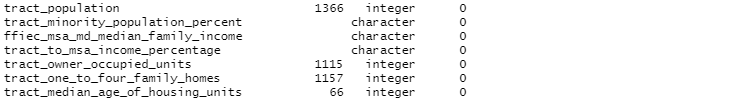




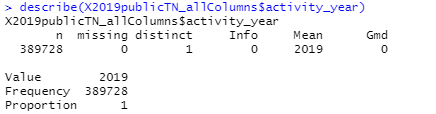


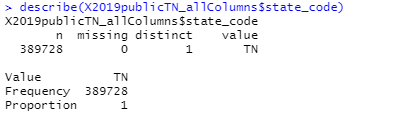






Running the **describe()** method from the Hmisc library shows the number of distinct values in each column.

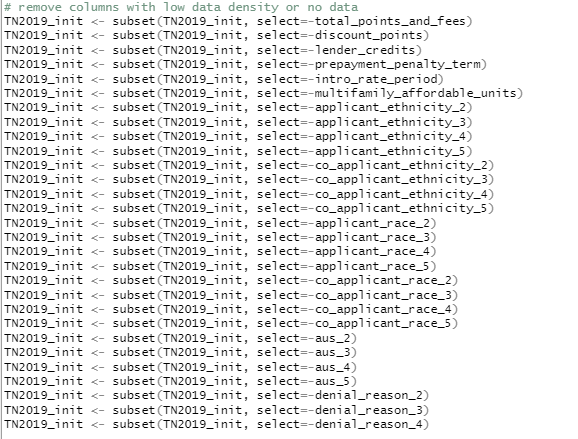




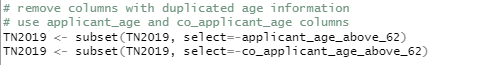
Columns that have the same values for each row can be removed.



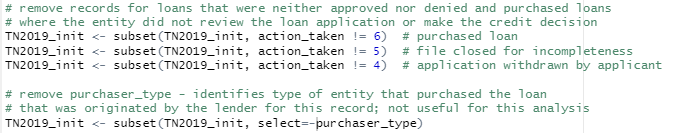
Columns in which more than half of the observations have NAs, or identified as additional columns for one discriminating column but that do not contain enough information to further the understanding of the data are removed.

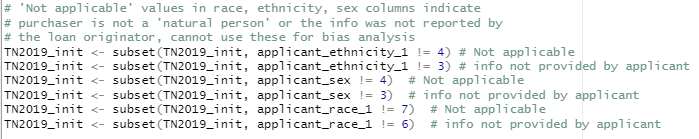


Remove columns applicant\_age\_above\_62 and co\_applicant\_age\_above\_62 as these columns contain information reported in the applicant\_age and co\_applicant\_age columns.

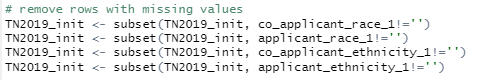


Referring to the HMDA Documentation and Guide, values of ‘Not applicable’ are submitted for race, ethnicity and sex if the data reported is for a mortgage loan that was purchased by an entity – i.e. Fannie Mae or Freddie Mac, etc. - or the applicant was “not a natural person” – i.e. a company or non-profit entity - or if the lending institution is not required to report that data. This analysis will not include information for purchased loans or those loans where the applicant was “not a natural person”. These rows are removed from the dataset, as well as rows where the applicant did not provide the information during the application process, or the application was withdrawn or closed and the loan was neither approved nor denied.

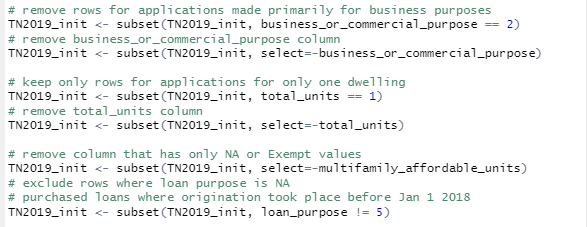




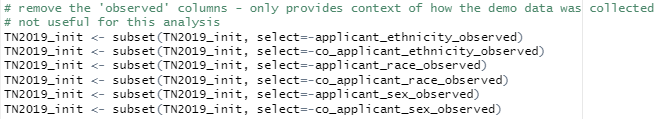
Remove rows with missing data in the race and ethnicity columns.



This analysis will only consider mortgage applications for single-family homes that are not for business purposes.



Columns that have information about how the demographic data was collected is not useful for this analysis and is removed.



Aggregated data is in the dataset as ‘derived’ fields (“Derived Fields Categorization”, 2019) and condenses the information contained in multiple columns into one column that may be easier to interpret while also reducing the dimensionality of the data to be analyzed, but might also create levels in the categorical variable column that will hava very low number of observations which can negatively affect the analysis. Each ‘derived’ column will need to be evaluated individually to determine whether or not to retain it and remove the associated columns for analysis.

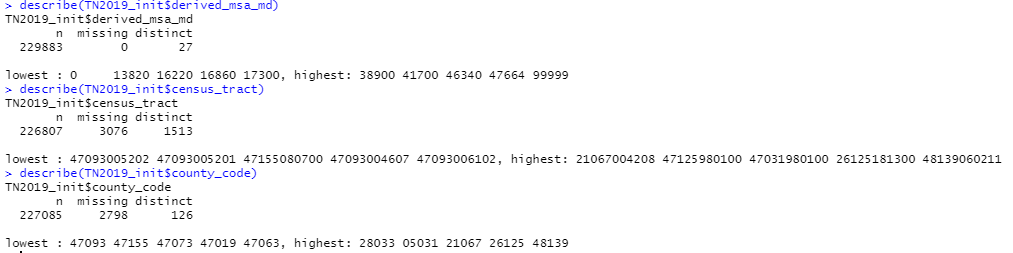
|  |  |
| --- | --- |
| **“Derived” column name** | **Condenses information from** |
| derived\_loan\_product\_type | loan\_type lien\_status |
| derived\_dwelling\_category | construction\_method total\_units |
| derived\_ethnicity | applicant\_ethnicity\_1  applicant\_ethnicity\_2 applicant\_ethnicity\_3 applicant\_ethnicity\_4 applicant\_ethnicity\_5 co\_applicant\_ethnicity\_1 co\_applicant\_ethnicity\_2 co\_applicant\_ethnicity\_3 co\_applicant\_ethnicity\_4 co\_applicant\_ethnicity\_5 |
| derived\_race | applicant\_race\_1  applicant\_race\_2  applicant\_race\_3  applicant\_race\_4  applicant\_race\_5  co\_applicant\_race\_1 co\_applicant\_race\_2 co\_applicant\_race\_3  co\_applicant\_race\_4 co\_applicant\_race\_5 |
| derived\_sex | applicant\_sex co\_applicant\_sex |
| derived\_msa\_md | county\_code census\_tract |

The ‘derived’ data in the columns above will be viewed with their corresponding columns to evaluate the quality of the derived information. Since the total\_units column has already been removed, the construction\_method column is retained for inclusion in the analysis.



To evaluate whether the ‘derived’ columns for ethnicity, race, sex, and loan product type give comparable and accurate information, the columns are retained and evaluated separately.

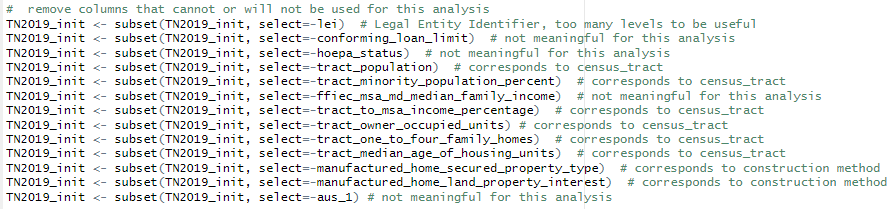
Using the **describe()** method on the derived\_msa\_md, census\_tract and county\_code columns gives a quick comparison.

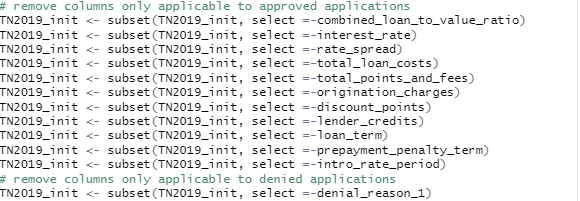


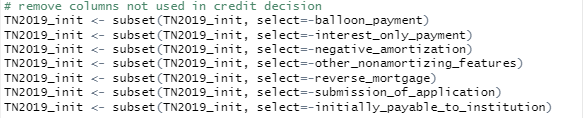
The derived\_msa\_md column will be retained as the number of distinct values for observations is much lower than the census\_tract and county\_code columns and will reduce the dimensionality of the data that is analyzed.



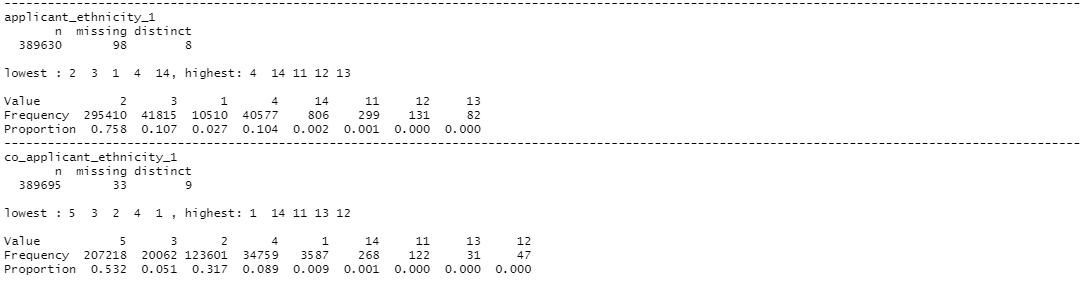
Referring again to the HMDA Documentation for each column in the dataset, many columns contain information that cannot or will not be used for this analysis.

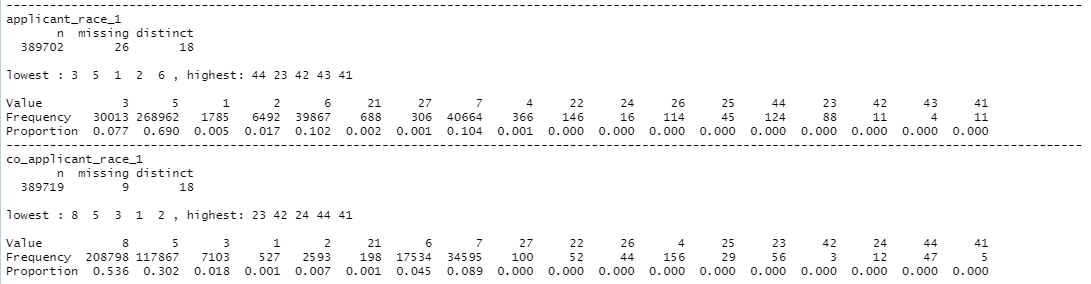






Running the **describe()** and **summary()** methods on the data provides summary statistics for the columns remaining in the dataset.

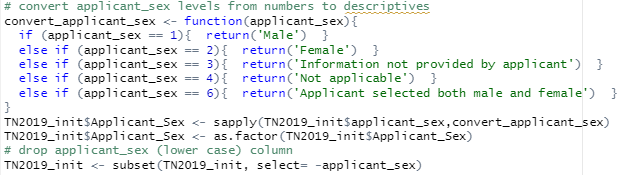


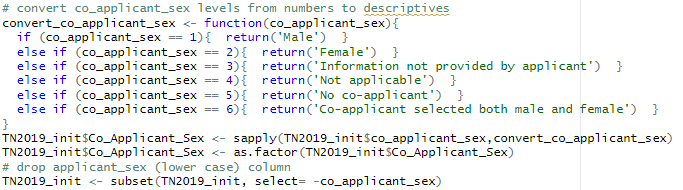


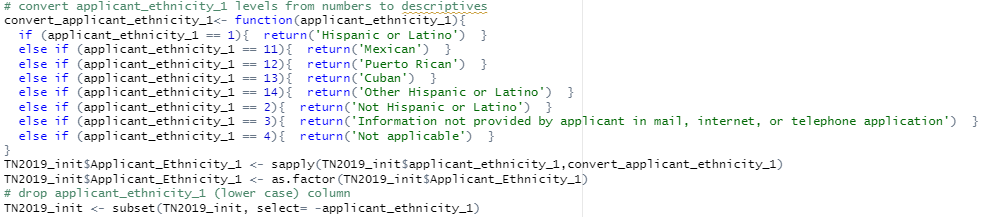


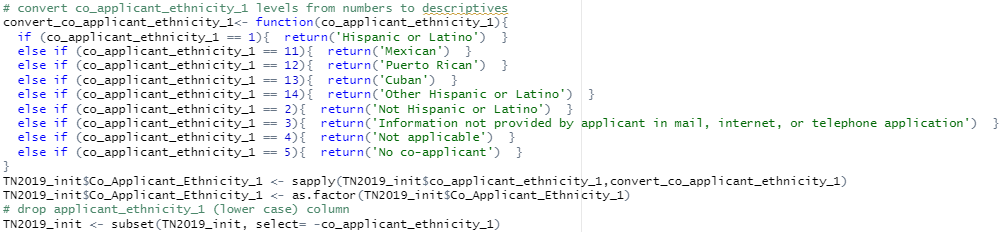
Many of the columns that contain categorical variables have numbers representing the levels for the variable, but these numbers are not easy to interpret.

The numeric levels are converted to text to facilitate visual interpretation of the information, beginning with the race, ethnicity and sex columns.





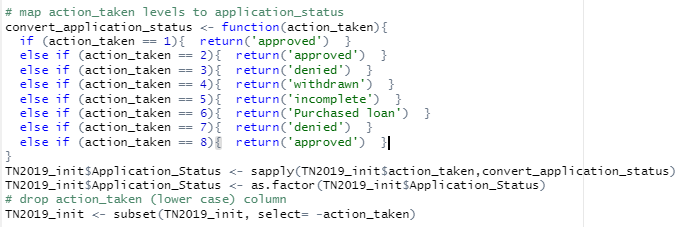




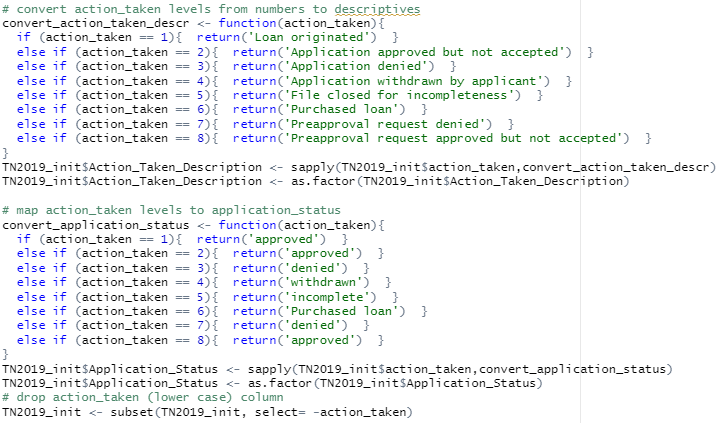




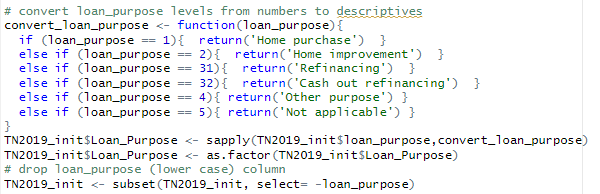
The independent variable of action\_taken is transformed from an 8-level factor to a binary variable application\_status.

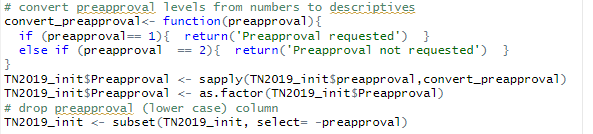


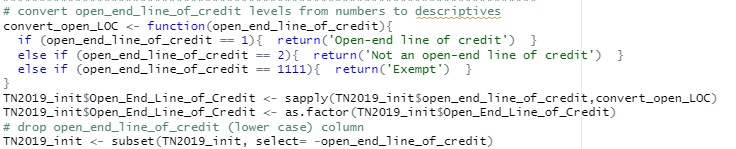
The numeric levels 4,5, and 6 were removed from the dataset in previous steps, which leaves the remaining levels of ‘approved’ and ‘denied’ and thus providing a single binary response variable for regression analysis.

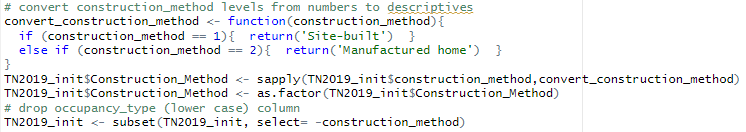


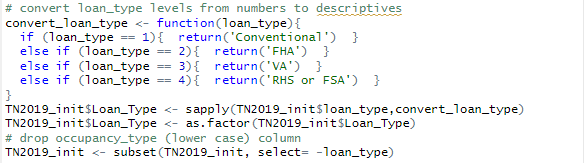
Numeric factor levels are converted to descriptive text for the remaining factor columns.

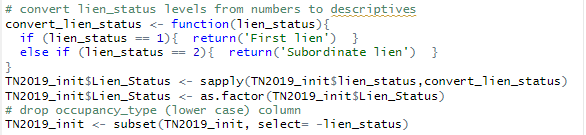


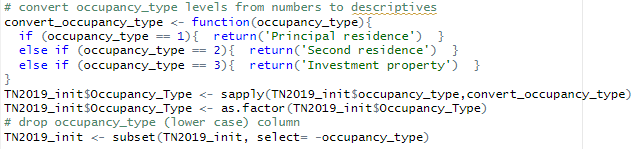


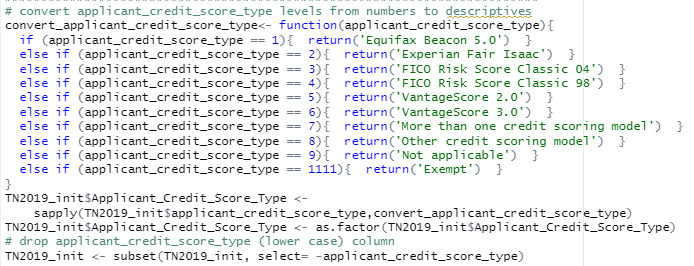


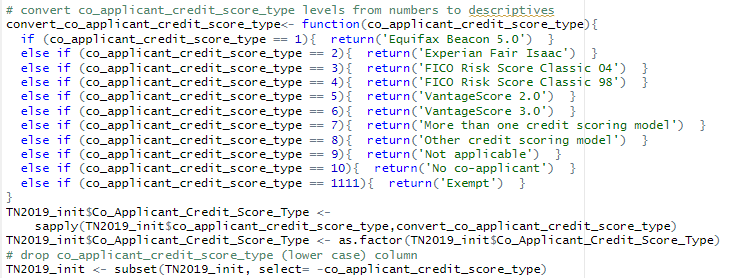


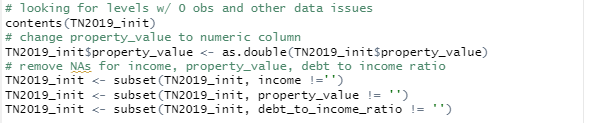


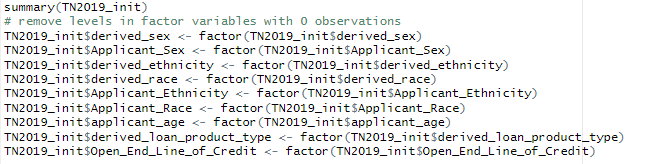












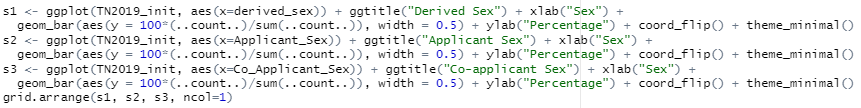
**Analysis**Report on your data-analysis process by describing the analysis technique(s) you used to appropriately analyze the data and by justifying the tools used in your data analysis. Include the calculations you performed and their outputs. Justify how you selected the analysis technique(s) you used, including any advantages or disadvantages of these technique(s).

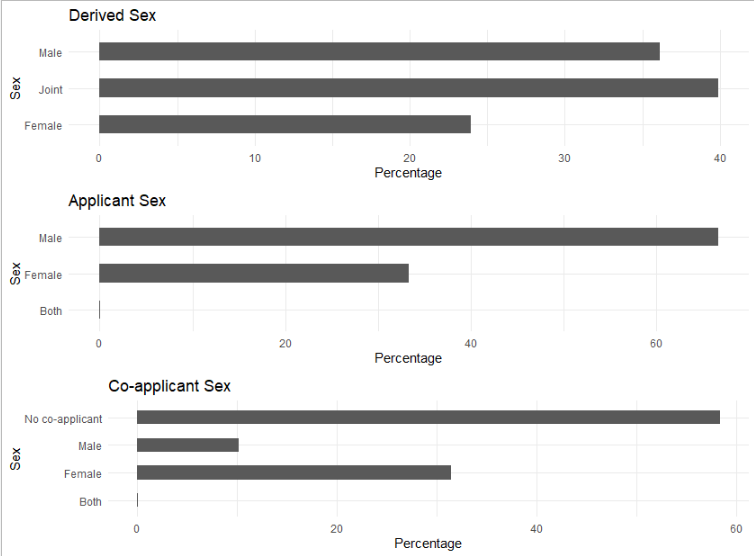
We will use descriptive analysis on the independent categorical and binary variables to determine which variables should be used for the final analysis to reduce the dimensions of the analysis (Tuffery, 2011). Because our independent variables consist of continuous and nominal variables, we will have to employ a factorial analysis of mixed (FAMD) data method to give insight into which variables in the data may be exceptional or which variables may be linked to each other (Tuffery, 2011). Logistic regression and decision tree analysis will be run after the factor analysis and removing any variables that do not contribute significantly to the outcome.

For this project used R to extract, clean and analyse the data. R is an open-source tool that was developed for statistical analysis and graphing (What is R?, 2020) that has a wide selection of packages to enhance statistical analysis that are freely available and continuously being updated with improvements and bug fixes.

########## UNIVARIATE   
library(ggplot2)  
library(gridExtra)

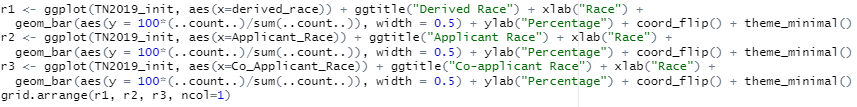
**SEX**

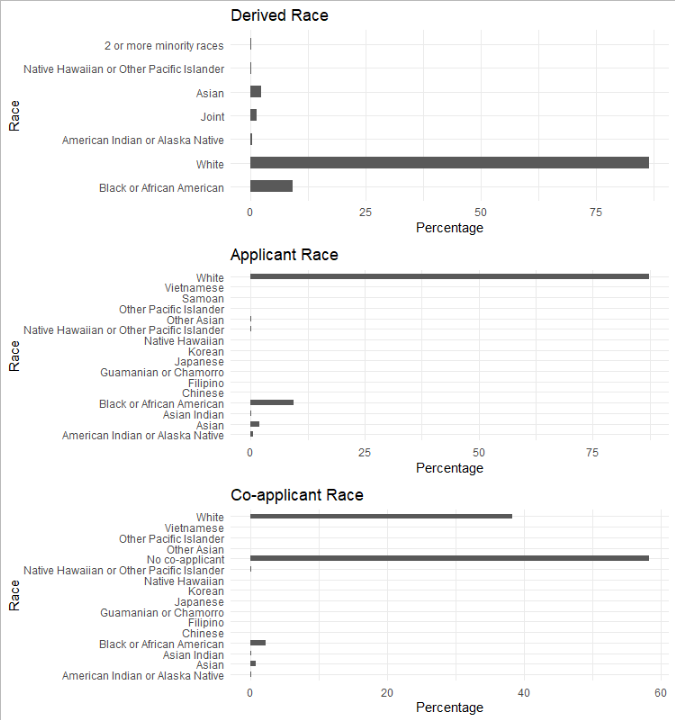






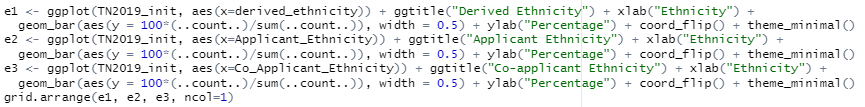
**RACE**

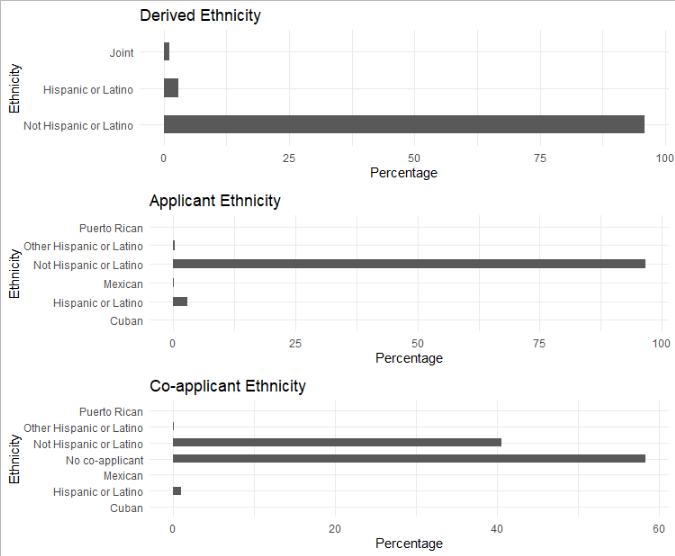






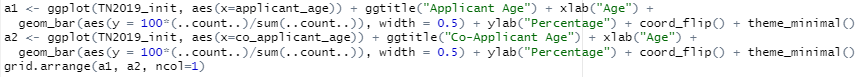
**ETHNICITY**

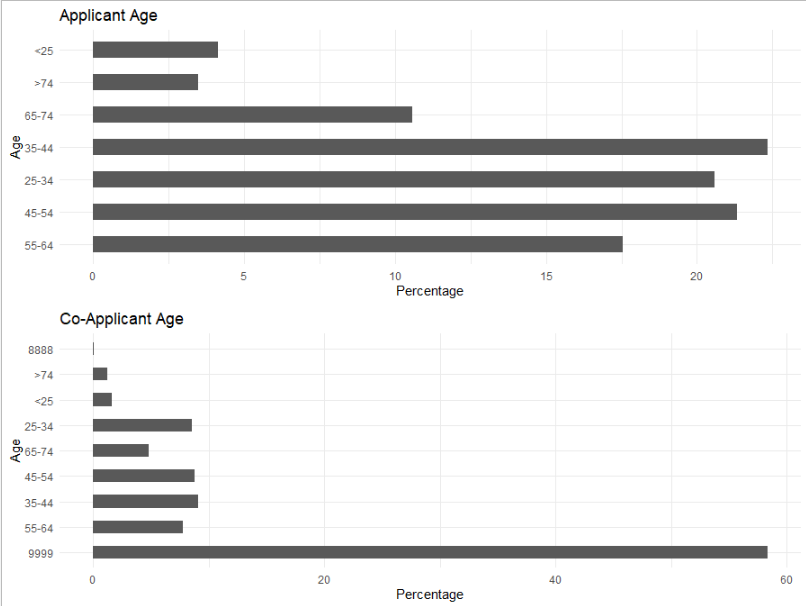






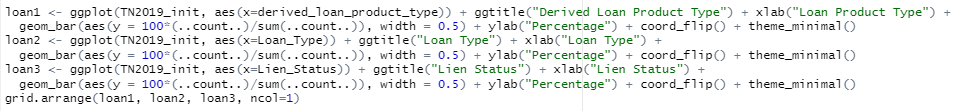
**AGE**

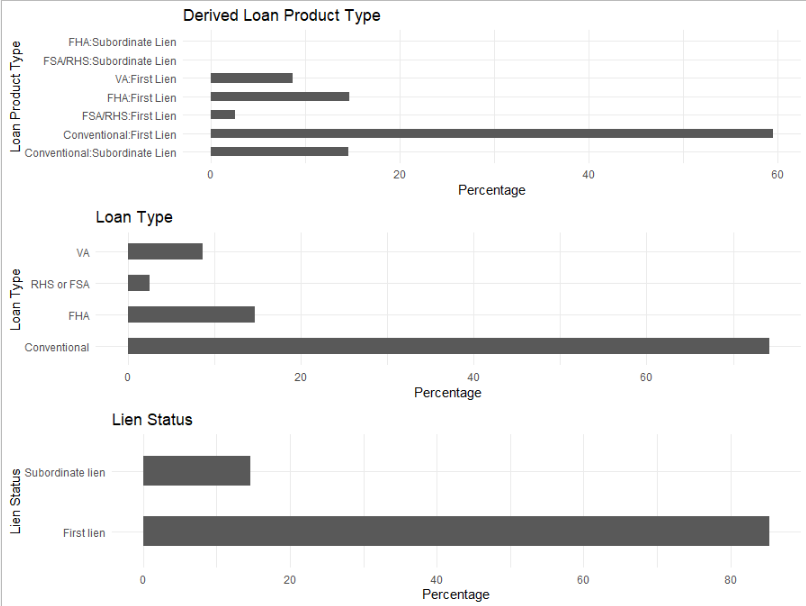






**LOAN TYPE**

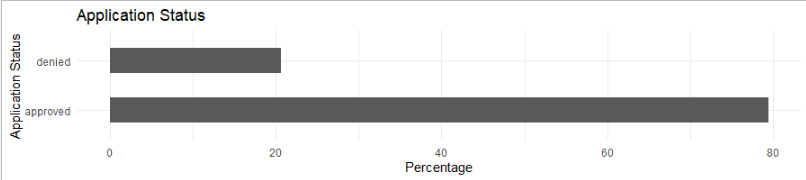




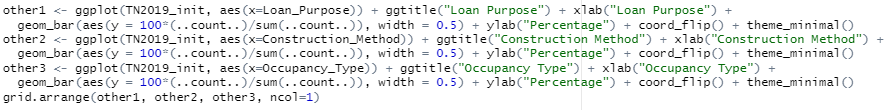


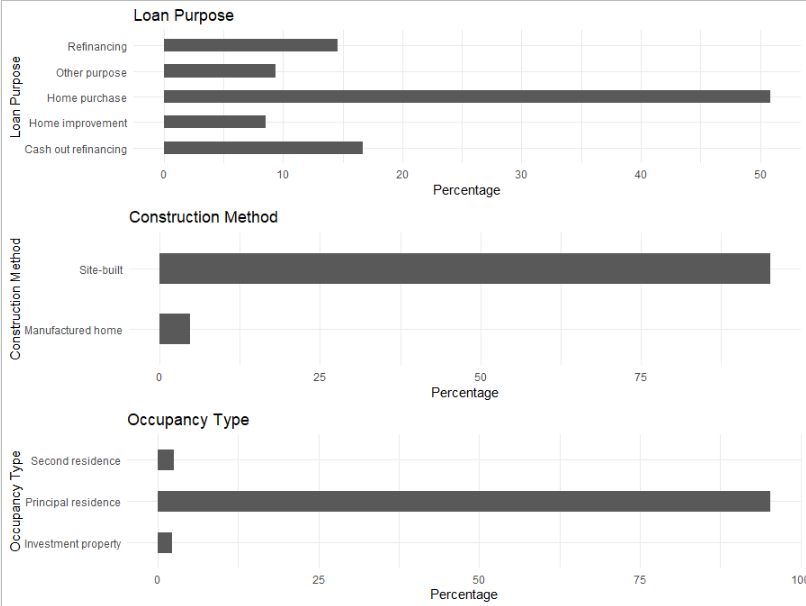
**ACTION TAKEN**

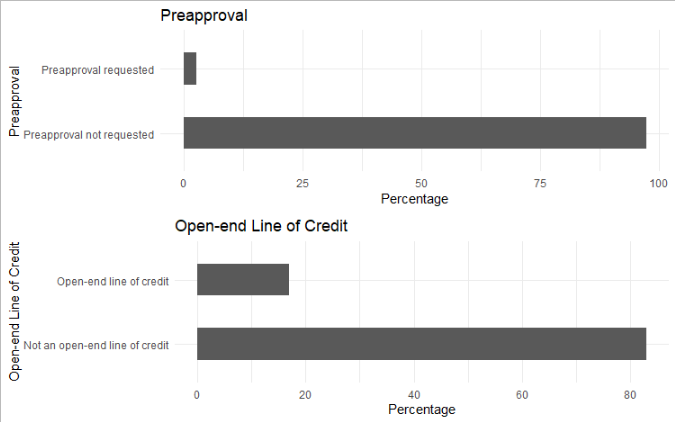
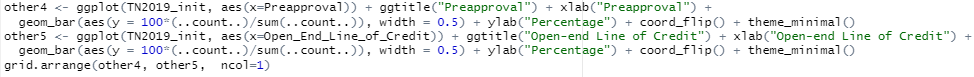


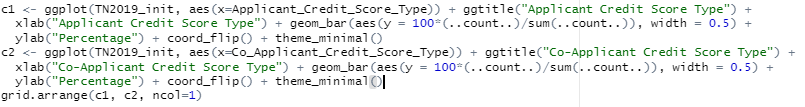


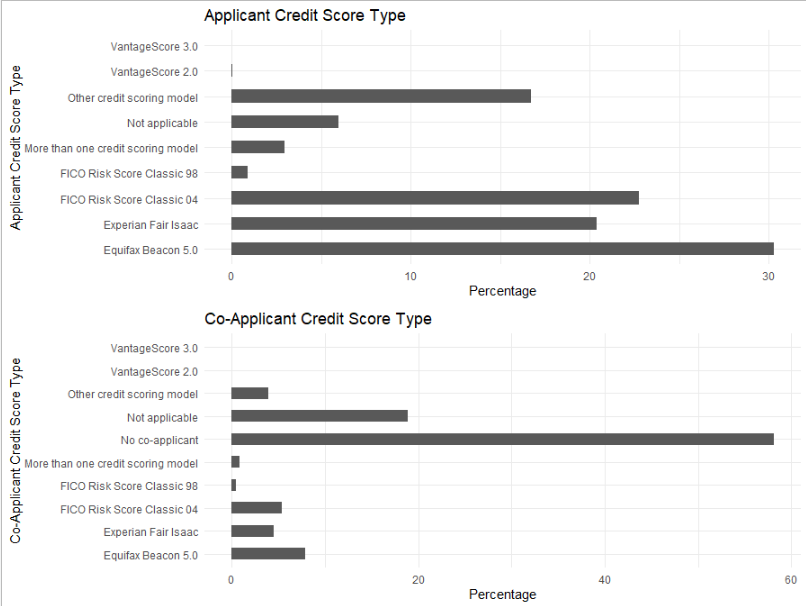
**OTHER 1**



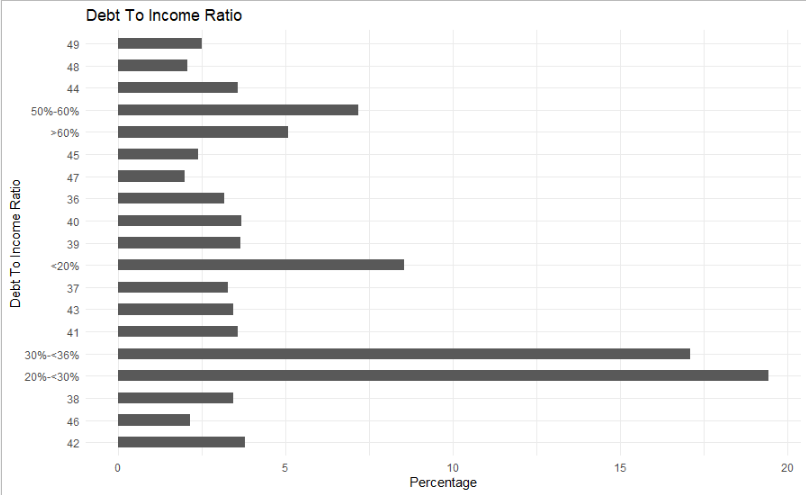




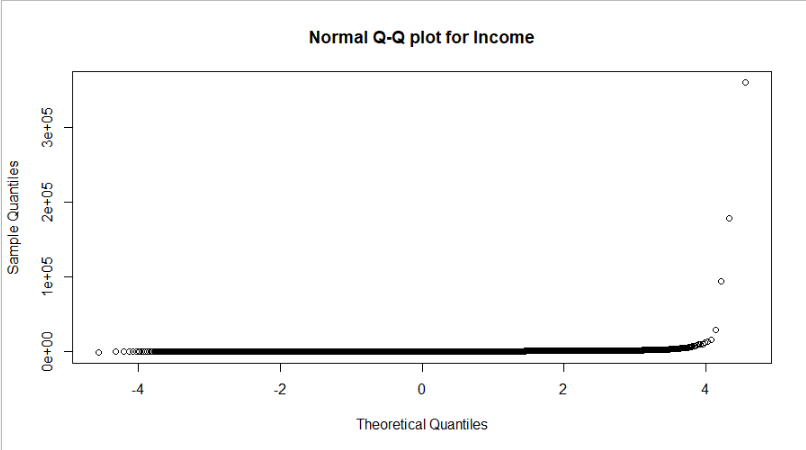


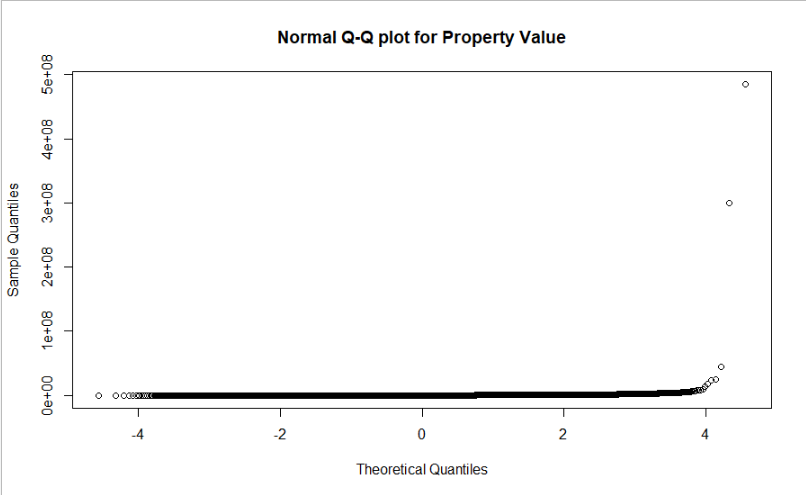


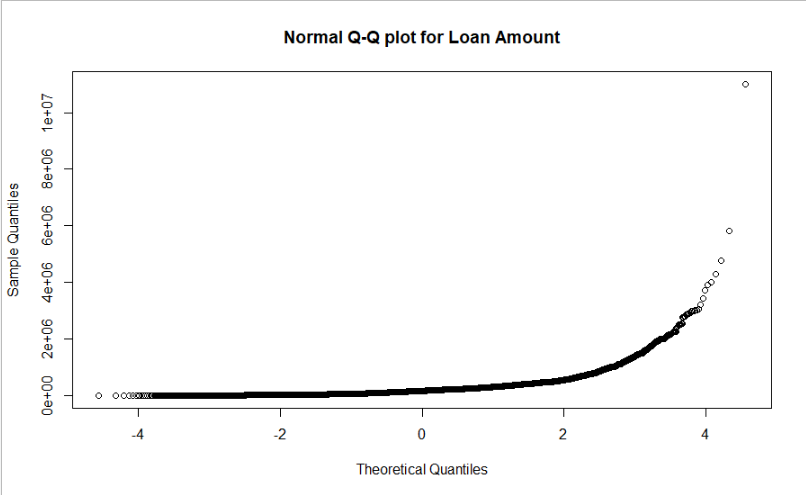








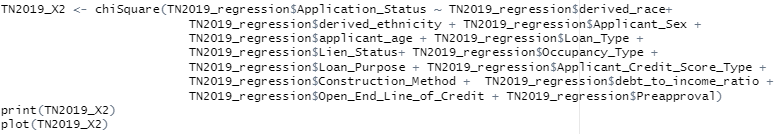


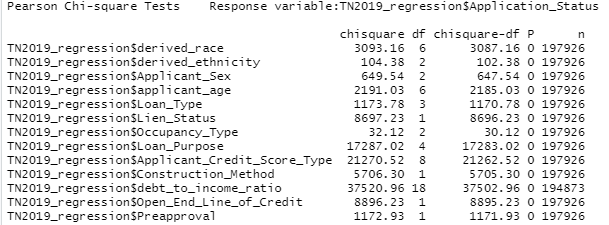




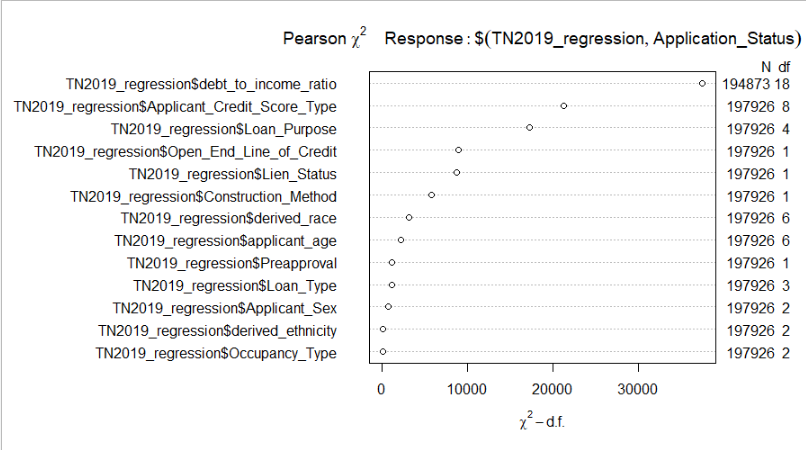
**CHI SQUARE ANALYSIS**

Calculating chi-square statistics for the categorical variables against the response gives values that indicate the strength of the relationship between each of the variables and the response.

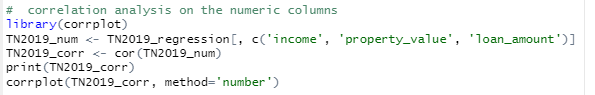


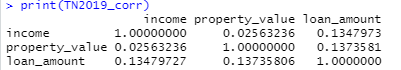


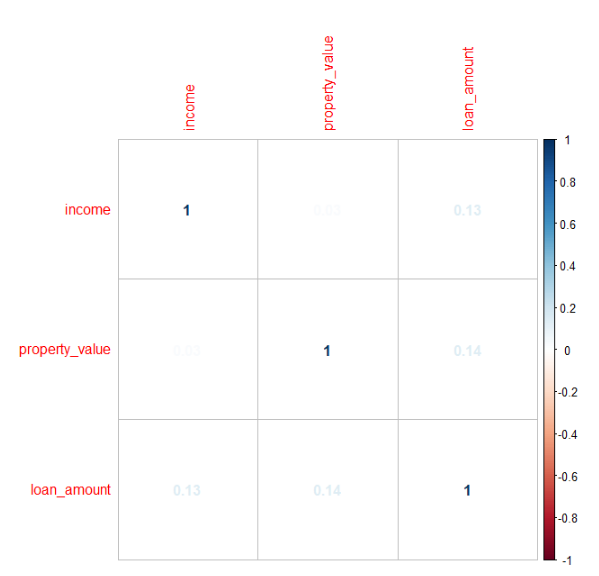
Plotting the chi-square statistics for the categorical variables against the response gives a picture of the relationship between each of the variables and the response that may be easier to interpret.



**CORRELATION ANALYSIS**

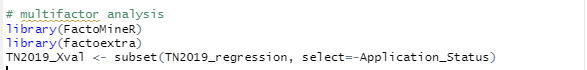




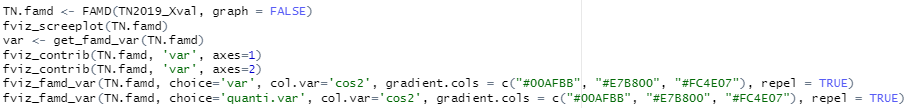


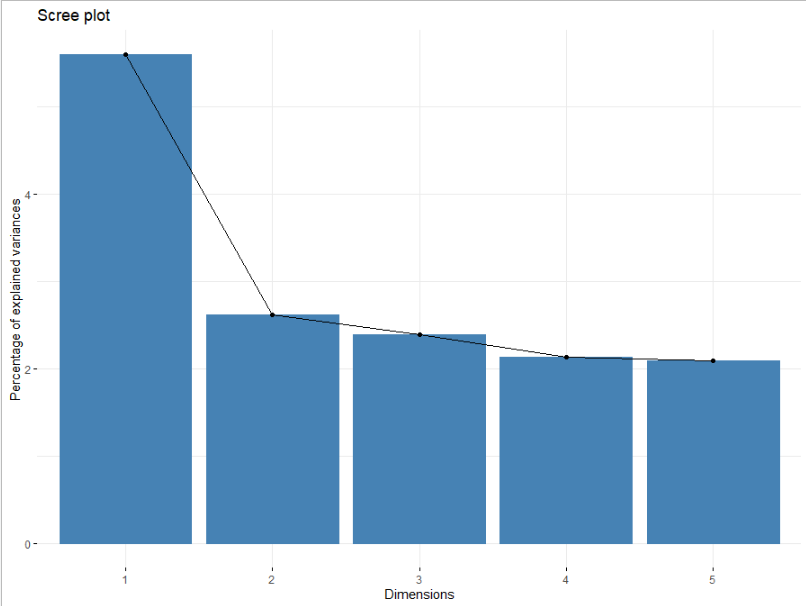
**MULTIFACTOR ANALYSIS**

Since the **stats** library **cor()** method requires numeric data for analysis, we must use another library that can can perform correlation analysis on numeric and factor data. The **FactoMineR** library provides tools to analyze this mixed data using **FAMD()** method:

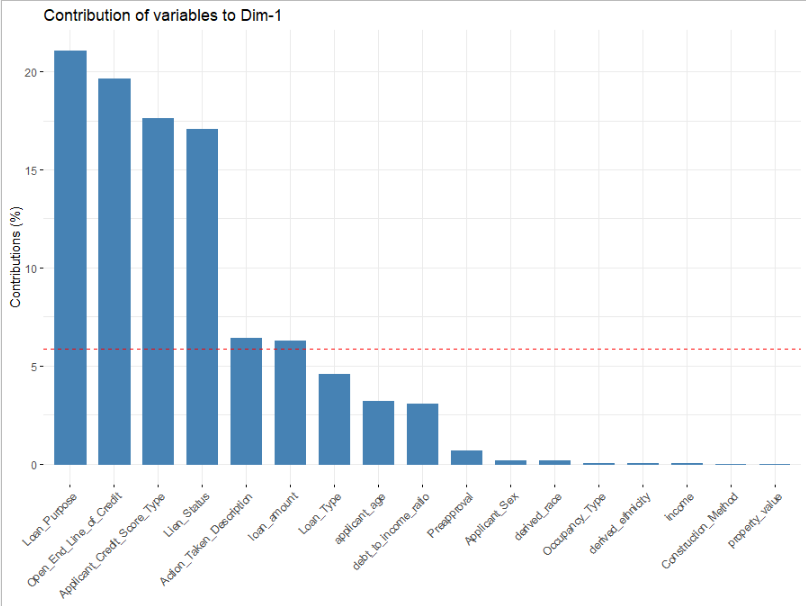


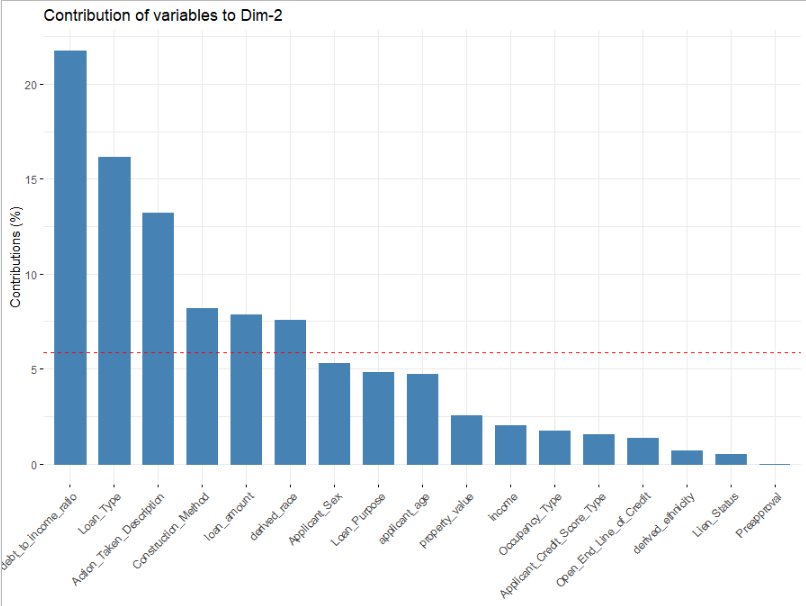
Next, we can view the scree plot to show us the percentages of inertia explained by each FAMD dimensions. This is analogous to PCA for continuous data. The **factoextra** library has methods that graph the information generated by the FAMD() method.





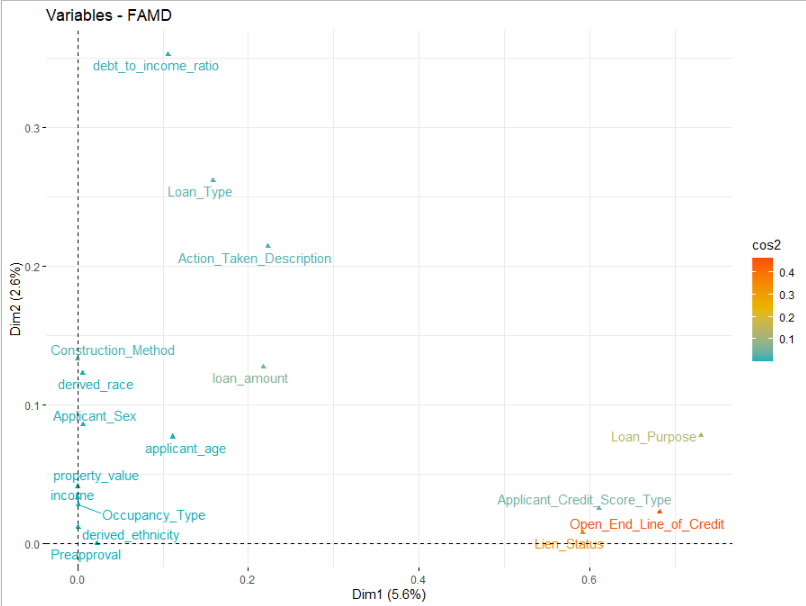
**fviz\_screeplot()** method generates the screeplot of the dimensions of the factor analysis shows the proportions of the variances, and then we can see the individual factors that contribute to each dimension with the **fviz\_contrib()** method.



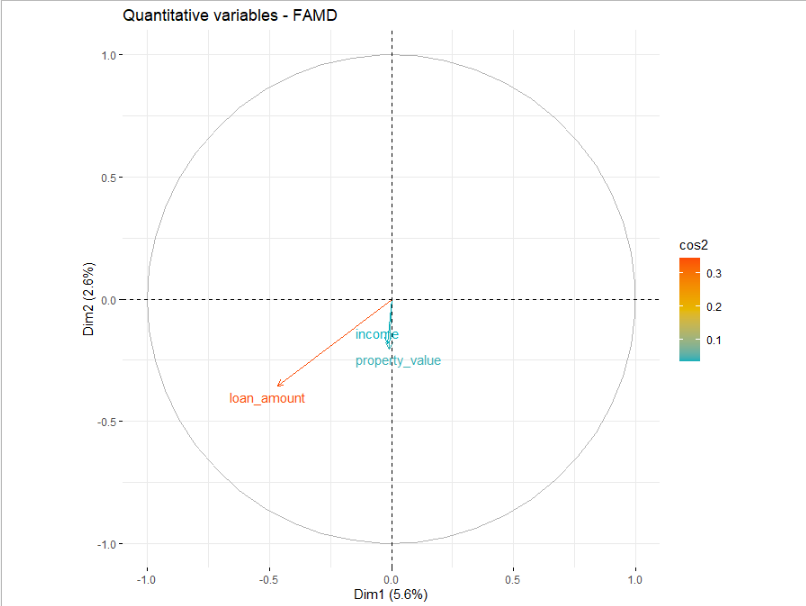


**fviz\_famd\_var()** with the ‘var’ option shows us all of the independent variables in the factor analysis results.





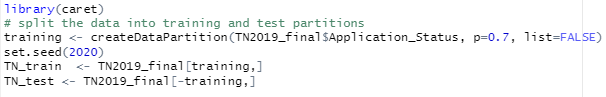




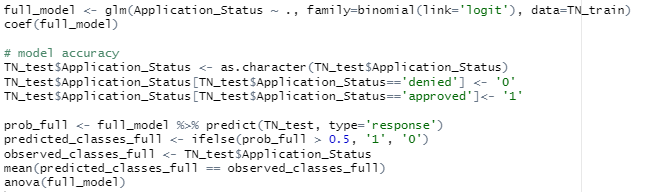
Based on the results of the factor analysis, columns income and property\_value are correlated, income is removed from the dataset.

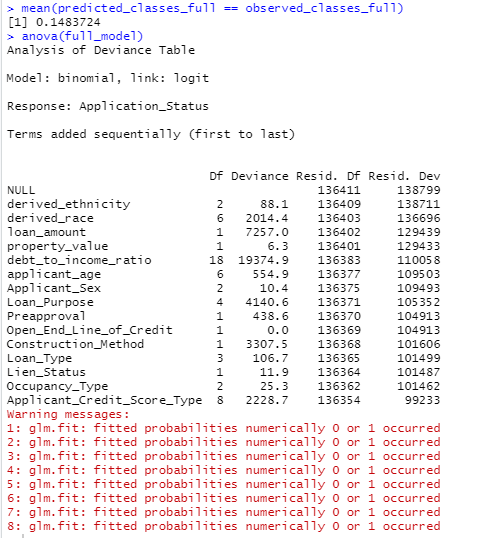


From the final cleaned and pruned data set, we partition the data set into training and test sets using the **createDataPartition()** method from the **caret** library to calculate a regression model.



The training data set is used generate the model, and then use the test data set to assess the accuracy of the model that is generated.





**Data Summary and Implications**

The results of the regression analysis demonstrate that given the data that we have, this model is not a good fit for prediction. This is because the public dataset does not contain the same data that the lending institutions have that their application scoring models use for making loan application decisions. For example, the lending institutions also report the applicant’s credit score, but this information is removed from the public dataset to protect applicant privacy.

The chi square and multi factor analyses clearly shows that there are differences between the proportions of application approvals and denials based on race, sex and ethnicity. If the models that are used by lending institutions to make credit decisions are trained using data that is biased, then the results of using that model to make decisions will also have bias. This cannot be addressed without having access to the same data the models are using.

Further analysis could include analysing lending decisions per each lending institution for bias in lending decisions, either nationwide or by region. Another investigation could compare lending decisions for bias between states and investigating historical lending practices within each state to assess trends.

**Sources**

A Guide to HMDA Reporting: Getting It Right! <https://www.ffiec.gov/hmda/pdf/2019guide.pdf>.  
Retrieved August 10, 2020.

“Background and Purpose of HMDA.” September 6, 2018. Retrieved August 8, 2020 from <https://www.ffiec.gov/hmda/history.htm>.

“Civil Rights Act of 1968.” Wikimedia Foundation. August 2, 2020.  Retrieved August 10, 2020 from  <https://en.wikipedia.org/wiki/Civil_Rights_Act_of_1968>.

“Derived Fields Categorization.” <https://github.com/cfpb/hmda-platform/wiki/Derived-Fields-Categorization>. Updated August 18, 2019.  Retrieved August 8, 2020.

“HMDA Documentation.” <https://ffiec.cfpb.gov/documentation/2019/lar-data-fields/>. Retrieved July 25, 2020.

Tuffery S. *Data Mining and Statistics for Decision Making.* Wiley; 2011. Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=cat07141a&AN=ebc.EBC792450&authtype=sso&custid=ns017578&site=eds-live&scope=site.

What is R? The R Project. The R Foundation. Retrieved July 12, 2020 from <https://www.r-project.org/about.html>.