**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.

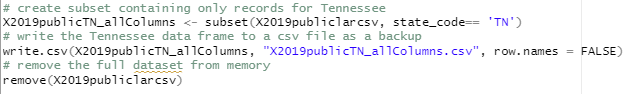
**Data Collection**The HMDA data is provided by year beginning in 1998, through 2019. The dataset for 2019 includes nationwide mortgage application data consisting of 94 independent variables as a mix of qualitative and quantitative variables, and five dependent variables. The full dataset includes 17,545,457 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.

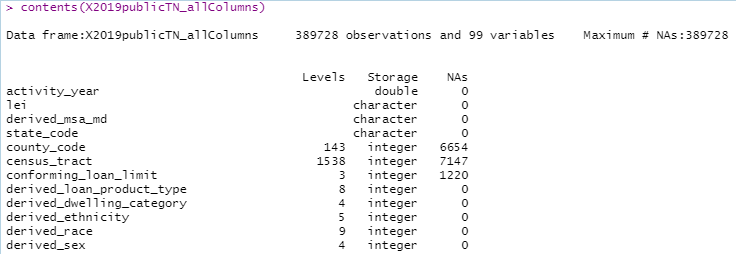
The initial import of the full dataset requires referencing the HMDA data documentation to identify the expected datatypes for each column for import. This process reveals that some columns that one would expect contain only numeric data, such as the property\_value and income columns, to have some records 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.

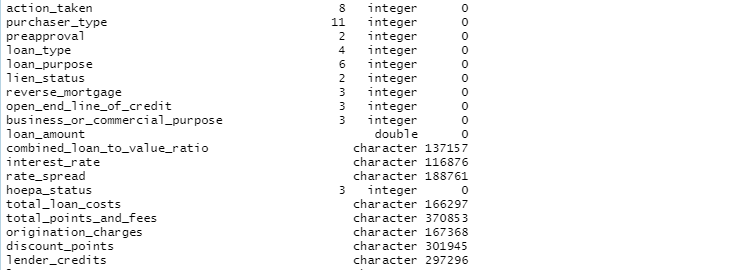


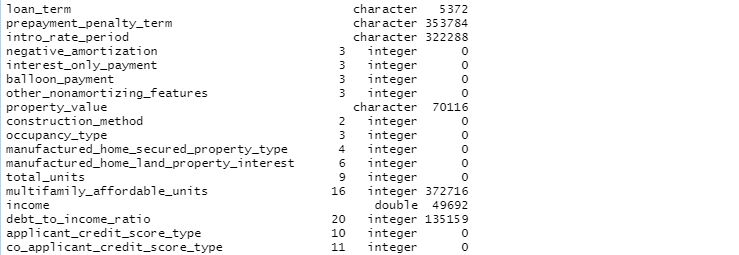
For this study we created a subset of the full dataset to reduce the analysis to include only the state of Tennessee.

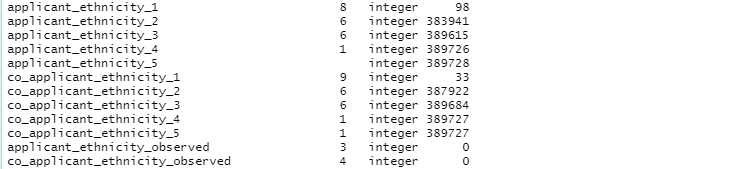


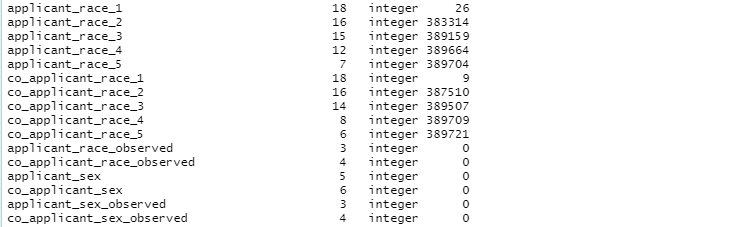
This subset of data reduced the full nationwide dataset to 389,728 rows. Running the **contents()** method from the Hmisc library provides a quick way to see the number of levels for factor variables and the number of NA rows in each column.

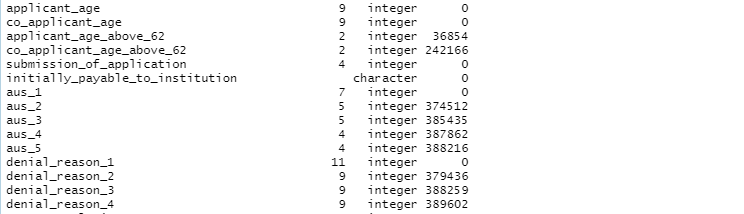


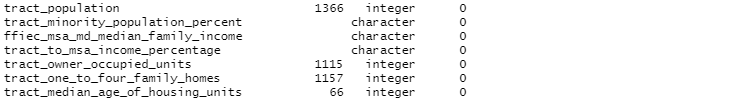












First, the columns that have the same values for each row can be removed.



Columns in which more than half of the observations have NAs are identified as additional columns for one discriminating column but that don’t contain enough information to further the understanding will be removed.

The data required cleaning to convert the 4 factor response variables to a single binary response variable, remove columns that contain NA for all rows in the Tennessee subset, remove columns such as applicant\_age\_above\_62 and co\_applicant\_age\_above\_62 that contain information duplicated in other columns, and account for values of ‘Exempt’ found in some of the continuous data columns. Aggregated data is in the original dataset as ‘derived’ fields (“Derived Fields Categorization”, 2019) for race and ethnicity information; we will keep the aggregated data fields for race and ethnicity from the original dataset and remove all of the applicant\_race\_1, applicant\_race\_2, applicant\_race\_3, etc. columns with duplicate data for race and ethnicity.  We will remove the additional columns for automated underwriting system data that have a very high number of missing values, keeping the primary aus\_1 column of data for analysis. The data density after removing duplicate, aggregated and low data density columns is 75%.

**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 we will use 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

**Data Summary and Implications**Summarize the implications of your data analysis by discussing the results of your data analysis in the context of the research question, including any limitations of your analysis. Within the context of your research question, recommend a course of action based on your results. Then propose two directions or approaches for future study of the data set.

**Sources**

“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.  Accessed August 8, 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>.