**Equity in Mortgage Lending**

**Research Question**Summarize the original real-data research question you identified in task 1. Your summary should include justification for the research question you identified in task 1, a description of the context in which the research question exists, and a discussion of your hypothesis.

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, as demonstrated by recent (as of this writing) news articles

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**Report on your data-collection process by describing the relevant data you collected, discussing any advantages and disadvantages of the data-gathering methodology you used, and discussing how you overcame any challenges you encountered during the process of collecting your data.

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, state, or lender 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.

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. The activity\_year and state\_code columns contain the same value for all rows and can be removed.

# remove columns that have info we don't need  
TN2019 <- subset(X2019publicTN\_allColumns, select=-activity\_year)  
TN2019 <- subset(TN2019, select=-state\_code)

The main demographic information that is of interest for this analyis is race, ethnicity, genger, and age of the borrower. Viewing the list of columns in the dataset shows that some age data that appears in other columns and will be unnecessary for this analysis.

# remove columns with duplicated age information

# use applicant\_age and co\_applicant\_age columns

TN2019 <- subset(TN2019, select=-applicant\_age\_above\_62)

TN2019 <- subset(TN2019, select=-co\_applicant\_age\_above\_62)

The full dataset includes aggregated data columns that have derived fields (“Derived Fields Categorization”, 2019) for race, ethnicity, sex, loan product and the action that the lender took to either approve or deny the mortgage application; this analysis retains the ‘derived’ columns for some of these instead of the original data to reduce the dimensionality of the dataset. Columns with aggregated that were removed from the dataset are as follows:

|  |  |  |
| --- | --- | --- |
| **Derived column name** | **Aggregated data columns** |  |
| derived\_loan\_product\_type | loan\_type lien\_status |  |
| derived\_dwelling\_category | construction\_method  total\_units |  |
|  |  |  |

# remove columns that are covered in 'derived' columns

# derived\_loan\_product\_type

TN2019 <- subset(TN2019, select=-loan\_type)

TN2019 <- subset(TN2019, select=-lien\_status)

# derived\_dwelling\_category

TN2019 <- subset(TN2019, select=-construction\_method)

TN2019 <- subset(TN2019, select=-total\_units)

Create subsets for each of the demographic columns (race, ethnicity, and sex) to make it easier to view and assess the summary statistics for just these columns.

TN2019\_gender <- subset(TN2019, select =c('derived\_sex', 'applicant\_sex', 'co\_applicant\_sex', 'applicant\_sex\_observed', 'co\_applicant\_sex\_observed'))

TN2019\_race <- subset(TN2019, select = c('derived\_race', 'applicant\_race\_1', 'applicant\_race\_2', 'applicant\_race\_3', 'applicant\_race\_4', 'applicant\_race\_5', 'applicant\_race\_observed', 'co\_applicant\_race\_1', 'co\_applicant\_race\_2', 'co\_applicant\_race\_3', 'co\_applicant\_race\_4', 'co\_applicant\_race\_5', 'co\_applicant\_race\_observed'))

TN2019\_ethnicity<- subset(TN2019, select =c('derived\_ethnicity', 'applicant\_ethnicity\_1', 'applicant\_ethnicity\_2', 'applicant\_ethnicity\_3', 'applicant\_ethnicity\_4', 'applicant\_ethnicity\_5', 'applicant\_ethnicity\_observed', 'co\_applicant\_ethnicity\_1', 'co\_applicant\_ethnicity\_2', 'co\_applicant\_ethnicity\_3', 'co\_applicant\_ethnicity\_4', 'co\_applicant\_ethnicity\_5', 'co\_applicant\_ethnicity\_observed'))

Running the describe() method in the Hmisc library for each of the above subsets gives summary statistics for each column which includes the number of missing values for each column.

**> describe(TN2019\_gender)**

TN2019\_gender

5 Variables 389728 Observations

-----------------------------------------------------------------------

derived\_sex

n missing distinct

389728 0 4

Value Female Joint Male Sex Not Available

Frequency 76127 129086 121216 63299

Proportion 0.195 0.331 0.311 0.162

-----------------------------------------------------------------------

applicant\_sex

n missing distinct

389728 0 5

lowest : 2 1 3 4 6, highest: 2 1 3 4 6

Value 2 1 3 4 6

Frequency 105245 221034 22697 40602 150

Proportion 0.270 0.567 0.058 0.104 0.000

-----------------------------------------------------------------------

co\_applicant\_sex

n missing distinct

389728 0 6

lowest : 5 1 2 3 4, highest: 1 2 3 4 6

Value 5 1 2 3 4 6

Frequency 208774 32611 103299 10343 34616 85

Proportion 0.536 0.084 0.265 0.027 0.089 0.000

-----------------------------------------------------------------------

applicant\_sex\_observed

n missing distinct

389728 0 3

Value 3 2 1

Frequency 66790 304974 17964

Proportion 0.171 0.783 0.046

-----------------------------------------------------------------------

co\_applicant\_sex\_observed

n missing distinct

389728 0 4

Value 4 2 3 1

Frequency 208774 127924 45351 7679

Proportion 0.536 0.328 0.116 0.020

-----------------------------------------------------------------------

**> describe(TN2019\_race)**

TN2019\_race

13 Variables 389728 Observations

-----------------------------------------------------------------------

derived\_race

n missing distinct

389728 0 9

lowest : Black or African American White American Indian or Alaska Native Joint Asian

highest: Asian Race Not Available Native Hawaiian or Other Pacific Islander 2 or more minority races Free Form Text Only

Black or African American (28994, 0.074), White (266504, 0.684), American Indian or Alaska Native (1182, 0.003), Joint (4367, 0.011), Asian (7291, 0.019), Race

Not Available (80531, 0.207), Native Hawaiian or Other Pacific Islander (486, 0.001), 2 or more minority races (347, 0.001), Free Form Text Only (26, 0.000)

-----------------------------------------------------------------------

applicant\_race\_1

n missing distinct

389702 26 18

lowest : 3 5 1 2 6 , highest: 44 23 42 43 41

Value 3 5 1 2 6 21 27 7 4 22 24 26 25 44 23 42 43 41

Frequency 30013 268962 1785 6492 39867 688 306 40664 366 146 16 114 45 124 88 11 4 11

Proportion 0.077 0.690 0.005 0.017 0.102 0.002 0.001 0.104 0.001 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

-----------------------------------------------------------------------

applicant\_race\_2

n missing distinct

6414 383314 16

lowest : 22 21 5 26 27, highest: 4 42 44 1 41

Value 22 21 5 26 27 43 3 24 25 23 2 4 42 44 1 41

Frequency 644 1339 1878 482 610 26 203 137 363 398 69 46 23 132 9 55

Proportion 0.100 0.209 0.293 0.075 0.095 0.004 0.032 0.021 0.057 0.062 0.011 0.007 0.004 0.021 0.001 0.009

-----------------------------------------------------------------------

applicant\_race\_3

n missing distinct

569 389159 15

lowest : 27 5 26 25 24, highest: 4 3 42 43 2

Value 27 5 26 25 24 23 22 21 44 41 4 3 42 43 2

Frequency 76 110 61 60 55 52 35 37 41 17 8 8 6 2 1

Proportion 0.134 0.193 0.107 0.105 0.097 0.091 0.062 0.065 0.072 0.030 0.014 0.014 0.011 0.004 0.002

-----------------------------------------------------------------------

applicant\_race\_4

n missing distinct

64 389664 12

lowest : 24 2 22 27 23, highest: 25 21 44 4 41

Value 24 2 22 27 23 26 5 25 21 44 4 41

Frequency 9 1 9 9 5 2 7 6 2 9 4 1

Proportion 0.141 0.016 0.141 0.141 0.078 0.031 0.109 0.094 0.031 0.141 0.062 0.016

-----------------------------------------------------------------------

applicant\_race\_5

n missing distinct

24 389704 7

lowest : 41 4 27 26 5 , highest: 27 26 5 23 43

Value 41 4 27 26 5 23 43

Frequency 12 1 3 1 5 1 1

Proportion 0.500 0.042 0.125 0.042 0.208 0.042 0.042

-----------------------------------------------------------------------

applicant\_race\_observed

n missing distinct

389728 0 3

Value 3 2 1

Frequency 66459 305648 17621

Proportion 0.171 0.784 0.045

-----------------------------------------------------------------------

co\_applicant\_race\_1

n missing distinct

389719 9 18

lowest : 8 5 3 1 2 , highest: 23 42 24 44 41

Value 8 5 3 1 2 21 6 7 27 22 26 4 25 23 42 24 44 41

Frequency 208798 117867 7103 527 2593 198 17534 34595 100 52 44 156 29 56 3 12 47 5

Proportion 0.536 0.302 0.018 0.001 0.007 0.001 0.045 0.089 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000

-----------------------------------------------------------------------

co\_applicant\_race\_2

n missing distinct

2218 387510 16

lowest : 21 3 22 5 27, highest: 44 2 41 43 1

Value 21 3 22 5 27 23 26 25 4 24 42 44 2 41 43 1

Frequency 434 53 254 544 198 238 177 143 17 51 13 45 26 18 3 4

Proportion 0.196 0.024 0.115 0.245 0.089 0.107 0.080 0.064 0.008 0.023 0.006 0.020 0.012 0.008 0.001 0.002

-----------------------------------------------------------------------

co\_applicant\_race\_3

n missing distinct

221 389507 14

lowest : 5 44 25 21 23, highest: 41 42 4 2 3

Value 5 44 25 21 23 27 26 22 24 41 42 4 2 3

Frequency 36 17 28 20 31 24 19 15 16 5 4 1 3 2

Proportion 0.163 0.077 0.127 0.090 0.140 0.109 0.086 0.068 0.072 0.023 0.018 0.005 0.014 0.009

-----------------------------------------------------------------------

co\_applicant\_race\_4

n missing distinct

19 389709 8

lowest : 21 44 24 27 41, highest: 27 41 23 5 4

Value 21 44 24 27 41 23 5 4

Frequency 1 1 4 2 4 4 2 1

Proportion 0.053 0.053 0.211 0.105 0.211 0.211 0.105 0.053

-----------------------------------------------------------------------

co\_applicant\_race\_5

n missing distinct

7 389721 6

lowest : 25 41 24 44 5 , highest: 41 24 44 5 23

Value 25 41 24 44 5 23

Frequency 1 2 1 1 1 1

Proportion 0.143 0.286 0.143 0.143 0.143 0.143

-----------------------------------------------------------------------

co\_applicant\_race\_observed

n missing distinct

389728 0 4

Value 4 2 3 1

Frequency 208798 128099 45311 7520

Proportion 0.536 0.329 0.116 0.019

-----------------------------------------------------------------------

**> describe(TN2019\_ethnicity)**

TN2019\_ethnicity

13 Variables 389728 Observations

-----------------------------------------------------------------------

derived\_ethnicity

n missing distinct

389728 0 5

lowest : Not Hispanic or Latino Ethnicity Not Available Hispanic or Latino Joint Free Form Text Only

highest: Not Hispanic or Latino Ethnicity Not Available Hispanic or Latino Joint Free Form Text Only

Value Not Hispanic or Latino Ethnicity Not Available Hispanic or Latino Joint Free Form Text Only

Frequency 293105 82392 10165 3968 98

Proportion 0.752 0.211 0.026 0.010 0.000

-----------------------------------------------------------------------

applicant\_ethnicity\_1

n missing distinct

389630 98 8

lowest : 2 3 1 4 14, highest: 4 14 11 12 13

Value 2 3 1 4 14 11 12 13

Frequency 295410 41815 10510 40577 806 299 131 82

Proportion 0.758 0.107 0.027 0.104 0.002 0.001 0.000 0.000

-----------------------------------------------------------------------

applicant\_ethnicity\_2

n missing distinct

5787 383941 6

lowest : 12 13 11 14 2 , highest: 13 11 14 2 1

Value 12 13 11 14 2 1

Frequency 744 353 2523 1884 240 43

Proportion 0.129 0.061 0.436 0.326 0.041 0.007

-----------------------------------------------------------------------

applicant\_ethnicity\_3

n missing distinct

113 389615 6

lowest : 12 13 11 14 1 , highest: 13 11 14 1 2

Value 12 13 11 14 1 2

Frequency 23 19 11 57 2 1

Proportion 0.204 0.168 0.097 0.504 0.018 0.009

-----------------------------------------------------------------------

applicant\_ethnicity\_4

n missing distinct value

2 389726 1 2

Value 2

Frequency 2

Proportion 1

-----------------------------------------------------------------------

applicant\_ethnicity\_observed

n missing distinct

389728 0 3

Value 3 2 1

Frequency 66810 305330 17588

Proportion 0.171 0.783 0.045

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_1

n missing distinct

389695 33 9

lowest : 5 3 2 4 1 , highest: 1 14 11 13 12

Value 5 3 2 4 1 14 11 13 12

Frequency 207218 20062 123601 34759 3587 268 122 31 47

Proportion 0.532 0.051 0.317 0.089 0.009 0.001 0.000 0.000 0.000

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_2

n missing distinct

1806 387922 6

lowest : 13 14 11 12 2 , highest: 14 11 12 2 1

Value 13 14 11 12 2 1

Frequency 123 500 808 252 106 17

Proportion 0.068 0.277 0.447 0.140 0.059 0.009

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_3

n missing distinct

44 389684 6

lowest : 14 11 12 13 2 , highest: 11 12 13 2 1

Value 14 11 12 13 2 1

Frequency 21 4 7 10 1 1

Proportion 0.477 0.091 0.159 0.227 0.023 0.023

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_4

n missing distinct value

1 389727 1 13

Value 13

Frequency 1

Proportion 1

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_5

n missing distinct value

1 389727 1 14

Value 14

Frequency 1

Proportion 1

-----------------------------------------------------------------------

co\_applicant\_ethnicity\_observed

n missing distinct

389728 0 4

Value 4 2 3 1

Frequency 207218 127941 47035 7534

Proportion 0.532 0.328 0.121 0.019

-----------------------------------------------------------------------

Variables with all observations missing:

[1] applicant\_ethnicity\_5

Columns with all missing data or low data density are removed from the dataset.

# remove columns for demographics with low data density or no data

TN2019 <- subset(TN2019, select=-co\_applicant\_race\_2)

TN2019 <- subset(TN2019, select=-co\_applicant\_race\_3)

TN2019 <- subset(TN2019, select=-co\_applicant\_race\_4)

TN2019 <- subset(TN2019, select=-co\_applicant\_race\_5)

TN2019 <- subset(TN2019, select=-applicant\_race\_2)

TN2019 <- subset(TN2019, select=-applicant\_race\_3)

TN2019 <- subset(TN2019, select=-applicant\_race\_4)

TN2019 <- subset(TN2019, select=-applicant\_race\_5)

TN2019 <- subset(TN2019, select=-co\_applicant\_ethnicity\_2)

TN2019 <- subset(TN2019, select=-co\_applicant\_ethnicity\_3)

TN2019 <- subset(TN2019, select=-co\_applicant\_ethnicity\_4)

TN2019 <- subset(TN2019, select=-co\_applicant\_ethnicity\_5)

TN2019 <- subset(TN2019, select=-applicant\_ethnicity\_2)

TN2019 <- subset(TN2019, select=-applicant\_ethnicity\_3)

TN2019 <- subset(TN2019, select=-applicant\_ethnicity\_4)

TN2019 <- subset(TN2019, select=-applicant\_ethnicity\_5)

# remove rows for demographics with missing values

TN2019 <- subset(TN2019, co\_applicant\_race\_1!='')

TN2019 <- subset(TN2019, applicant\_race\_1!='')

TN2019 <- subset(TN2019, co\_applicant\_ethnicity\_1!='')

TN2019 <- subset(TN2019, applicant\_ethnicity\_1!='')

Running the describe() method on the dataset gives additional information to identify other very low-density data columns and rows with missing data that can be removed from the dataset:

# remove other columns with low data density

TN2019 <- subset(TN2019, select=-aus\_2)

TN2019 <- subset(TN2019, select=-aus\_3)

TN2019 <- subset(TN2019, select=-aus\_4)

TN2019 <- subset(TN2019, select=-aus\_5)

TN2019 <- subset(TN2019, select=-denial\_reason\_2)

TN2019 <- subset(TN2019, select=-denial\_reason\_3)

TN2019 <- subset(TN2019, select=-denial\_reason\_4)

The response variable for this analysis is the action\_taken column. Referencing the data dictionary for insight into the numeric values represented, we can remove rows with action\_taken in (4,5,6) as those levels are equivalent to NA for this project.

# remove rows with action\_taken in (4,5,6)

# action\_taken = 4 (application withdrawn)

# action\_taken = 5 (application closed as incomplete)

# action\_taken = 6 (purchased loan, meaning an entity purchased the loan)

TN2019 <- subset(TN2019, action\_taken!=4)

TN2019 <- subset(TN2019, action\_taken!=5)

TN2019 <- subset(TN2019, action\_taken!=6)

To facilitate interpetation of the data within the action\_taken column, the data is converted from the numeric value to the description of the values:

# map action\_taken levels to application\_status

convert\_application\_status <- function(action\_taken){

if (action\_taken == 1){ return('approved') }

else if (action\_taken == 2){ return('approved') }

else if (action\_taken == 3){ return('denied') }

else if (action\_taken == 4){ return('withdrawn') }

else if (action\_taken == 5){ return('incomplete') }

else if (action\_taken == 6){ return('Purchased loan') }

else if (action\_taken == 7){ return('denied') }

else if (action\_taken == 8){ return('approved') }

}

TN2019$Application\_Status <- sapply(TN2019$action\_taken,convert\_application\_status)

TN2019$Application\_Status <- as.factor(TN2019$Application\_Status)

# drop action\_taken (lower case) column

TN2019 <- subset(TN2019, select= -action\_taken)

Consulting the data dictionary and the data submission guide, the values in the race, ethnicity, and sex columns that map to a ‘Not applicable’ value indicates that the loan was purchased and the financial institution chose not to report the data, or that the loan applicant was not a ‘natural person’ (i.e. a business or other entity) (p.33) for which this demographic data does not apply. Rows with these values are removed from the dataset, as well as rows where the information was not provided by the applicant and is therefore not available for analysis.

# 'Not applicable' values in race, ethnicity, sex columns indicate

# purchaser is not a 'natural person' or the info was not reported by

# the loan originator, cannot use these for bias analysis

TN2019 <- subset(TN2019, applicant\_ethnicity\_1 != 4)

TN2019 <- subset(TN2019, applicant\_sex != 4)

TN2019 <- subset(TN2019, applicant\_race\_1 != 7)

TN2019 <- subset(TN2019, applicant\_age!='8888')

TN2019 <- subset(TN2019, co\_applicant\_age!='8888')

# remove rows in race, ethnicity, sex info columns

# with codes that indicate information was not provided by the applicant

TN2019 <- subset(TN2019, applicant\_ethnicity\_1 != 3)

TN2019 <- subset(TN2019, applicant\_sex != 3)

TN2019 <- subset(TN2019, applicant\_race\_1 != 6)

TN2019 <- subset(TN2019, co\_applicant\_ethnicity\_1 != 3)

TN2019 <- subset(TN2019, co\_applicant\_sex != 3)

TN2019 <- subset(TN2019, co\_applicant\_race\_1 != 6)

To make visual identification of the coded values in the main columns of interest in this analysis, the numeric codes are converted into the descriptive terms provided in the data dictionary for the dataset as follows:

# convert applicant\_sex levels from numbers to descriptives

convert\_applicant\_sex <- function(applicant\_sex){

if (applicant\_sex == 1){ return('Male') }

else if (applicant\_sex == 2){ return('Female') }

else if (applicant\_sex == 3){ return('Information not provided by applicant') }

else if (applicant\_sex == 4){ return('Not applicable') }

else if (applicant\_sex == 6){ return('Applicant selected both male and female') }

}

TN2019$Applicant\_Sex <- sapply(TN2019$applicant\_sex,convert\_applicant\_sex)

TN2019$Applicant\_Sex <- as.factor(TN2019$Applicant\_Sex)

# drop applicant\_sex (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_sex)

# convert co\_applicant\_sex levels from numbers to descriptives

convert\_co\_applicant\_sex <- function(co\_applicant\_sex){

if (co\_applicant\_sex == 1){ return('Male') }

else if (co\_applicant\_sex == 2){ return('Female') }

else if (co\_applicant\_sex == 3){ return('Information not provided by applicant') }

else if (co\_applicant\_sex == 4){ return('Not applicable') }

else if (co\_applicant\_sex == 5){ return('No co-applicant') }

else if (co\_applicant\_sex == 6){ return('Applicant selected both male and female') }

}

TN2019$Co\_Applicant\_Sex <- sapply(TN2019$co\_applicant\_sex,convert\_co\_applicant\_sex)

TN2019$Co\_Applicant\_Sex <- as.factor(TN2019$Co\_Applicant\_Sex)

# drop co\_applicant\_sex (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_sex)

# convert applicant\_sex\_observed levels from numbers to descriptives

convert\_applicant\_sex\_observed <- function(applicant\_sex\_observed){

if (applicant\_sex\_observed == 1){ return('Collected on the basis of visual observation or surname') }

else if (applicant\_sex\_observed == 2){ return('Not collected on the basis of visual observation or surname') }

else if (applicant\_sex\_observed == 3){ return('Not applicable') }

}

TN2019$Applicant\_Sex\_Observed <- sapply(TN2019$applicant\_sex\_observed,convert\_applicant\_sex\_observed)

TN2019$Applicant\_Sex\_Observed <- as.factor(TN2019$Applicant\_Sex\_Observed)

# drop applicant\_sex\_observed (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_sex\_observed)

# convert co\_applicant\_sex\_observed levels from numbers to descriptives

convert\_co\_applicant\_sex\_observed <- function(co\_applicant\_sex\_observed){

if (co\_applicant\_sex\_observed == 1){ return('Collected on the basis of visual observation or surname')}

else if (co\_applicant\_sex\_observed == 2){ return('Not collected on the basis of visual observation or surname')}

else if (co\_applicant\_sex\_observed == 3){ return('Not applicable')}

else if (co\_applicant\_sex\_observed == 4){ return('No co-applicant')}

}

TN2019$Co\_Applicant\_Sex\_Observed <- sapply(TN2019$co\_applicant\_sex\_observed,convert\_co\_applicant\_sex\_observed)

TN2019$Co\_Applicant\_Sex\_Observed <- as.factor(TN2019$Co\_Applicant\_Sex\_Observed)

# drop co\_applicant\_sex\_observed (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_sex\_observed)

# convert applicant\_ethnicity\_1 levels from numbers to descriptives

convert\_applicant\_ethnicity\_1<- function(applicant\_ethnicity\_1){

if (applicant\_ethnicity\_1 == 1){ return('Hispanic or Latino') }

else if (applicant\_ethnicity\_1 == 11){ return('Mexican') }

else if (applicant\_ethnicity\_1 == 12){ return('Puerto Rican') }

else if (applicant\_ethnicity\_1 == 13){ return('Cuban' }

else if (applicant\_ethnicity\_1 == 14){ return('Other Hispanic or Latino') }

else if (applicant\_ethnicity\_1 == 2){ return('Not Hispanic or Latino') }

else if (applicant\_ethnicity\_1 == 3){ return('Information not provided by applicant in mail, internet, or telephone application') }

else if (applicant\_ethnicity\_1 == 4){ return('Not applicable') }

}

TN2019$Applicant\_Ethnicity\_1 <- sapply(TN2019$applicant\_ethnicity\_1,convert\_applicant\_ethnicity\_1)

TN2019$Applicant\_Ethnicity\_1 <- as.factor(TN2019$Applicant\_Ethnicity\_1)

# drop applicant\_ethnicity\_1 (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_ethnicity\_1)

# convert co\_applicant\_ethnicity\_1 levels from numbers to descriptives

convert\_co\_applicant\_ethnicity\_1<- function(co\_applicant\_ethnicity\_1){

if (co\_applicant\_ethnicity\_1 == 1) return('Hispanic or Latino' }

else if (co\_applicant\_ethnicity\_1 == 11){ return('Mexican') }

else if (co\_applicant\_ethnicity\_1 == 12){ return('Puerto Rican') }

else if (co\_applicant\_ethnicity\_1 == 13){ return('Cuban') }

else if (co\_applicant\_ethnicity\_1 == 14){ return('Other Hispanic or Latino') }

else if (co\_applicant\_ethnicity\_1 == 2){ return('Not Hispanic or Latino') }

else if (co\_applicant\_ethnicity\_1 == 3){ return('Information not provided by applicant in mail, internet, or telephone application') }

else if (co\_applicant\_ethnicity\_1 == 4){ return('Not applicable') }

else if (co\_applicant\_ethnicity\_1 == 5){ return('No co-applicant') }

}

TN2019$Co\_Applicant\_Ethnicity\_1 <- sapply(TN2019$co\_applicant\_ethnicity\_1,convert\_co\_applicant\_ethnicity\_1)

TN2019$Co\_Applicant\_Ethnicity\_1 <- as.factor(TN2019$Co\_Applicant\_Ethnicity\_1)

# drop co\_applicant\_ethnicity\_1 (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_ethnicity\_1)

# convert applicant\_ethnicity\_observed levels from numbers to descriptives

convert\_applicant\_ethnicity\_observed <- function(applicant\_ethnicity\_observed){

if (applicant\_ethnicity\_observed == 1){ return('Collected on the basis of visual observation or surname') }

else if (applicant\_ethnicity\_observed == 2){ return('Not collected on the basis of visual observation or surname') }

else if (applicant\_ethnicity\_observed == 3){ return('Not applicable') }

}

TN2019$Applicant\_Ethnicity\_Observed <- sapply(TN2019$applicant\_ethnicity\_observed,convert\_applicant\_ethnicity\_observed)

TN2019$Applicant\_Ethnicity\_Observed <- as.factor(TN2019$Applicant\_Ethnicity\_Observed)

# drop applicant\_ethnicity\_observed (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_ethnicity\_observed)

# convert co\_applicant\_ethnicity\_observed levels from numbers to descriptives

convert\_co\_applicant\_ethnicity\_observed <- function(co\_applicant\_ethnicity\_observed){

if (co\_applicant\_ethnicity\_observed == 1){ return('Collected on the basis of visual observation or surname') }

else if (co\_applicant\_ethnicity\_observed == 2){ return('Not collected on the basis of visual observation or surname') }

else if (co\_applicant\_ethnicity\_observed == 3){ return('Not applicable') }

else if (co\_applicant\_ethnicity\_observed == 4){ return('No co-applicant') }

}

TN2019$Co\_Applicant\_Ethnicity\_Observed <- sapply(TN2019$co\_applicant\_ethnicity\_observed,convert\_co\_applicant\_ethnicity\_observed)

TN2019$Co\_Applicant\_Ethnicity\_Observed <- as.factor(TN2019$Co\_Applicant\_Ethnicity\_Observed)

# drop co\_applicant\_ethnicity\_observed (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_ethnicity\_observed)

# convert applicant\_race\_1 levels from numbers to descriptives

convert\_applicant\_race\_1<- function(applicant\_race\_1){

if (applicant\_race\_1 == 1){ return('American Indian or Alaska Native') }

else if (applicant\_race\_1 == 2){ return('Asian') }

else if (applicant\_race\_1 == 21){ return('Asian Indian') }

else if (applicant\_race\_1 == 22){ return('Chinese') }

else if (applicant\_race\_1 == 23){ return('Filipino') }

else if (applicant\_race\_1 == 24){ return('Japanese') }

else if (applicant\_race\_1 == 25){ return('Korean') }

else if (applicant\_race\_1 == 26){ return('Vietnamese') }

else if (applicant\_race\_1 == 27){ return('Other Asian') }

else if (applicant\_race\_1 == 3){ return('Black or African American') }

else if (applicant\_race\_1 == 4){ return('Native Hawaiian or Other Pacific Islander') }

else if (applicant\_race\_1 == 41){ return('Native Hawaiian') }

else if (applicant\_race\_1 == 42){ return('Guamanian or Chamorro') }

else if (applicant\_race\_1 == 43){ return('Samoan') }

else if (applicant\_race\_1 == 44){ return('Other Pacific Islander') }

else if (applicant\_race\_1 == 5){ return('White') }

else if (applicant\_race\_1 == 6){ return('Information not provided by applicant in mail, internet, or telephone application') }

else if (applicant\_race\_1 == 7){ return('Not applicable') }

}

TN2019$Applicant\_Race\_1 <- sapply(TN2019$applicant\_race\_1,convert\_applicant\_race\_1)

TN2019$Applicant\_Race\_1 <- as.factor(TN2019$Applicant\_Race\_1)

# drop applicant\_race\_1 (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_race\_1)

# convert co\_applicant\_race\_1 levels from numbers to descriptives

convert\_co\_applicant\_race\_1<- function(co\_applicant\_race\_1){

if (co\_applicant\_race\_1 == 1){ return('American Indian or Alaska Native') }

else if (co\_applicant\_race\_1 == 2){ return('Asian') }

else if (co\_applicant\_race\_1 == 21){ return('Asian Indian') }

else if (co\_applicant\_race\_1 == 22){ return('Chinese') }

else if (co\_applicant\_race\_1 == 23){ return('Filipino') }

else if (co\_applicant\_race\_1 == 24){ return('Japanese') }

else if (co\_applicant\_race\_1 == 25){ return('Korean') }

else if (co\_applicant\_race\_1 == 26){ return('Vietnamese') }

else if (co\_applicant\_race\_1 == 27){ return('Other Asian') }

else if (co\_applicant\_race\_1 == 3){ return('Black or African American') }

else if (co\_applicant\_race\_1 == 4){ return('Native Hawaiian or Other Pacific Islander') }

else if (co\_applicant\_race\_1 == 41){ return('Native Hawaiian') }

else if (co\_applicant\_race\_1 == 42){ return('Guamanian or Chamorro') }

else if (co\_applicant\_race\_1 == 43){ return('Samoan') }

else if (co\_applicant\_race\_1 == 44){ return('Other Pacific Islander') }

else if (co\_applicant\_race\_1 == 5){ return('White')

else if (co\_applicant\_race\_1 == 6){ return('Information not provided by applicant in mail, internet, or telephone application') }

else if (co\_applicant\_race\_1 == 7){ return('Not applicable') }

else if (co\_applicant\_race\_1 == 8){ return('No co-applicant') }

}

TN2019$Co\_Applicant\_Race\_1 <- sapply(TN2019$co\_applicant\_race\_1,convert\_co\_applicant\_race\_1)

TN2019$Co\_Applicant\_Race\_1 <- as.factor(TN2019$Co\_Applicant\_Race\_1)

# drop co\_applicant\_race\_1 (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_race\_1)

# convert applicant\_race\_observed levels from numbers to descriptives

convert\_applicant\_race\_observed <- function(applicant\_race\_observed){

if (applicant\_race\_observed == 1){ return('Collected on the basis of visual observation or surname') }

else if (applicant\_race\_observed == 2){ return('Not collected on the basis of visual observation or surname') }

else if (applicant\_race\_observed == 3){ return('Not applicable') }

}

TN2019$Applicant\_Race\_Observed <- sapply(TN2019$applicant\_race\_observed,convert\_applicant\_race\_observed)

TN2019$Applicant\_Race\_Observed <- as.factor(TN2019$Applicant\_Race\_Observed)

# drop applicant\_race\_observed (lower case) column

TN2019 <- subset(TN2019, select= -applicant\_race\_observed)

# convert co\_applicant\_race\_observed levels from numbers to descriptives

convert\_co\_applicant\_race\_observed <- function(co\_applicant\_race\_observed){

if (co\_applicant\_race\_observed == 1){ return('Collected on the basis of visual observation or surname') }

else if (co\_applicant\_race\_observed == 2){ return('Not collected on the basis of visual observation or surname') }

else if (co\_applicant\_race\_observed == 3){ return('Not applicable') }

else if (co\_applicant\_race\_observed == 4){ return('No co-applicant') }

}

TN2019$Co\_Applicant\_Race\_Observed <- sapply(TN2019$co\_applicant\_race\_observed,convert\_co\_applicant\_race\_observed)

TN2019$Co\_Applicant\_Race\_Observed <- as.factor(TN2019$Co\_Applicant\_Race\_Observed)

# drop co\_applicant\_race\_observed (lower case) column

TN2019 <- subset(TN2019, select= -co\_applicant\_race\_observed)

Subsets of the data by sex, race, and ethnicity show that the ‘observed’ columns do not add to the understanding of the data, and are removed from analyis.

######################################################

# check the 'observed' columns for information

TN2019\_Gender <- subset(TN2019, select =c('Application\_Status',

'derived\_sex', 'Applicant\_Sex', 'Applicant\_Sex\_Observed',

'Co\_Applicant\_Sex','Co\_Applicant\_Sex\_Observed'))

TN2019\_Race <- subset(TN2019, select = c('Application\_Status',

'derived\_race', 'Applicant\_Race\_1', 'Applicant\_Race\_Observed',

'Co\_Applicant\_Race\_1', 'Co\_Applicant\_Race\_Observed'))

TN2019\_Ethnicity<- subset(TN2019, select =c('Application\_Status',

'derived\_ethnicity', 'Applicant\_Ethnicity\_1',

'Applicant\_Ethnicity\_Observed', 'Co\_Applicant\_Ethnicity\_1',

'Co\_Applicant\_Ethnicity\_Observed'))

# remove 'observed' columns as unneeded information

TN2019 <- subset(TN2019, select=-Applicant\_Sex\_Observed)

TN2019 <- subset(TN2019, select=-Applicant\_Race\_Observed)

TN2019 <- subset(TN2019, select=-Applicant\_Ethnicity\_Observed)

The remaining categorical variables are converted from numeric levels to descriptive levels for easier interpretation:

Histograms give us an approximate visual representation of the distribution of continuous numerical data that is easier to interpret than quantiles or other summary statistics given for the continuous variables.

Bar graphs allow us to see the data distribution for the categorical variables, which allows for quicker visual comparison of the distribution of the frequency of the levels within the categorical variables and to compare the distributions between the different variables.

check of # of home sales in TN with THDA.org for sanity check of final numbers – seem to be in line with 2018 numbers that are available

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.  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.

R was used for this project 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**

A Guide to HMDA Reporting: Getting It Right! <https://www.ffiec.gov/hmda/guide.htm>.  
<https://www.ffiec.gov/hmda/pdf/2019guide.pdf>. Last Edited March 7, 2019.

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

Chiwaya, Nigel and Ross, Janell. “The American dream while Black: ‘Locked in a vicious cycle’”. NBC News. <https://www.nbcnews.com/specials/american-dream-while-black-homeownership/index.html>. August 3, 2020.

“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.

Olick, Diana. “Lenders deny mortgages for Black homeowners at a rate 80 percent higher than whites.” NBC News. <https://www.nbcnews.com/news/nbcblk/lenders-deny-mortgages-black-homeowners-rate-80-percent-higher-whites-n1237441>. August 20, 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>.

<https://www.investopedia.com/articles/economics/08/fannie-mae-freddie-mac-credit-crisis.asp>

<https://thda.org/research-planning/home-sales-price-by-county>