**Exploring Issues of Fair Lending Practices Using Home Mortgage Disclosure Act (HMDA) data**

**Author: Nancy Hudspeth**

**Department of Political Science & Public Administration**

**California State University, Stanislaus**

**Turlock, CA 95382**

**Email:**nhudspeth@csustan.edu

*Last modified June 7, 2019*

## **Goals of Exercise**

## **Students will use HMDA data and SPSS to answer the research questions:**

## **I**n 2017, were there differences in mortgage approval rates for different demographic groups (race, ethnicity, gender) controlling for income (and other characteristics)?

**Note to the Instructor:**

The data used in these exercises is available to the public and may be downloaded from the Consumer Financial Protection Bureau website.[[1]](#footnote-1) There are filters provided on the CFPB website, but when I recently downloaded the data included in this exercise, I found that in order to get data on denials, I needed to download all records for a particular place and year instead of using the more specific filters.

I’ve provided the data set used in this exercise, HMDA 2017 State of California.sav which is a subset of the 2017 HMDA dataset available from the Consumer Financial Protection Bureau. It contains all records for the State of California and was downloaded March 29, 2019. Some of the variables in the dataset have been recoded and some new variables have been created (which are explained in this exercise).  These exercises use SPSS to analyze the data.

I’ve also provided a smaller data set, HMDA 2017 CA 1st lien 1-4 fam dwellings.sav which is a subset of the above data set. The main difference between the two data sets is that the first contains all types of properties and all types of mortgage applications. The second excludes mortgage applications for multifamily properties, manufactured properties, and properties that would not be used as the borrower’s primary residence. It also excludes refinancing and home improvement mortgage applications.

More detail about the data is provided below.

You have permission to use this exercise and to revise it to fit your needs.

**References for using SPSS include:**

*IBM SPSS Statistics for Windows Version 23: A Basic Tutorial* by Linda Fiddler, John Korey, Edward Nelson (Editor), and Elizabeth Nelson.  The online version of the book is at [the Social Science Research and Instructional Center's Website](http://ssric.org/node/582).

*IBM SPSS for Introductory Statistics: Use and Interpretation* by George A. Morgan, Nancy L. Leech, Gene W. Gloeckner, and Karen C. Barrett (5th edition; 2013). Routledge.

*SPSS Survival Manual: A Step By Step Guide to Data Analysis Using IBM SPSS* by Julie Pallant (6th edition; 2016). McGraw Hill Education.

*Statistics for People Who (Think They) Hate Statistics* by Neil J. Salkind (6th edition; 2017). SAGE Publications, Inc. This is a good book for beginners to help them get started.

### ****Background: Fair Lending and HMDA****

### **The Home Mortgage Disclosure Act (HMDA) is a federal law that requires mortgage lenders to report information about the loans they have approved (and denied) in order to show that they are serving their communities fairly. Congress passed the HMDA in 1975 to provide citizens with information that could be used to identify and address discriminatory lending practices, such as “redlining.”**

### **The practice that became known as redlining began in the 1930s when a federal program called the Home Owners’ Loan Corporation (HOLC) created color-coded maps of more than 200 U.S. cities. These maps rated neighborhoods in terms of their desirability for investment. The “best” areas were coded blue or green; the “worst” areas were coded red. The HOLC maps were then used by the Federal Housing Administration (FHA), banks, and insurance companies when deciding whether, and at what interest rate, to approve mortgage applications for property located in various neighborhoods. In general, the redlined neighborhoods were areas inhabited by low-income people, recent immigrants, and racial, ethnic, or religious minorities, while the green and blue neighborhoods were inhabited by middle-class and wealthy white homeowners. Thus, the HOLC, a federal program that had originally been intended to boost the housing market during the Great Depression, contributed to the decline of central cities, the growth of suburbia, and hindered the efforts of low-income minorities to build wealth through homeownership. In 1968, the Fair Housing Act banned discriminatory lending practices, including redlining, but the effects are still apparent today.**

### **To learn more about the history of redlining and to view 150+ interactive maps, see Robert K. Nelson, LaDale Winling, Richard Marciano, Nathan Connolly, et al., “Mapping Inequality: Redlining in New Deal America.”**[[2]](#footnote-2) **For reviews of the literature on fair lending see, for example, Bartlett et al. (2018) and Turner and Skidmore (1999).**[[3]](#footnote-3)

**Data used**

According to HMDA’s website, “The Home Mortgage Disclosure Act (HMDA) requires many financial institutions to maintain, report, and publicly disclose loan-level information about mortgages. These data help show whether lenders are serving the housing needs of their communities; they give public officials information that helps them make decisions and policies; and they shed light on lending patterns that could be discriminatory. The public data are modified to protect applicant and borrower privacy.”[[4]](#footnote-4)

HMDA data is available from the Consumer Financial Protection Bureau.[[5]](#footnote-5) The data can be filtered by year (2007-2017), geographic area (state, county, metropolitan statistical area), and loan type and purpose (e.g., for first-lien, owner-occupied, 1-4 family homes) and type of financing. However, as I noted above, in order to get data that includes denials I needed to download all records for a place and year. The more specific filters will only generate data on loans that were approved.

For this exercise, I have attached an SPSS dataset for the State of California, 2017, that contains all data on applications for residential mortgages, including the applicant and co-applicant (race, ethnicity, gender, income), loan purpose, loan amount, and decision on the loan which includes several categories for reasons why a loan is denied. If desired, the dataset can be sub-divided by county – or other datasets may be created and downloaded by the instructor. The name of the SPSS file is “CA 2017 hmda\_lar 1st Lien 1-4 family dwellings.sav”

I’ve also provided a smaller data set, HMDA 2017 CA 1st lien 1-4 fam dwellings.sav which is a subset of the above data set. The main difference between the two data sets is that the first contains all types of properties and all types of mortgage applications. The second excludes mortgage applications for multifamily properties, manufactured properties, and properties that would not be used as the borrower’s primary residence. It also excludes refinancing and home improvement mortgage applications.

This second data set was created from the first, using the “select cases” function and the following filters:

* Property type = one to four family dwelling (other than manufactured housing)
	+ *Manufactured housing and multi-family housing were excluded*
* Loan purpose = home purchase
	+ *Refinancing and home improvement were excluded*
* Owner occupancy = Owner-occupied as a principal dwelling
	+ *Not owner-occupied as a principal dwelling was excluded*

**Either dataset will work for the exercises; the output described is for the complete dataset (all cases).**

The data set includes 55 variables, which are listed along the top row (columns or “fields”). In the complete State of California dataset, in 2017 there are 1,048,575 rows. Each row indicates a mortgage application.

The variables are collected using the HMDA Reporting Form (example below) which is completed by the lending institutions.[[6]](#footnote-6) Detailed descriptions of the variables (“field reference”) are listed on the CPFB website here <https://cfpb.github.io/api/hmda/fields.html>

### Example of HDMA Reporting Form

### ****Steps in the Analysis****

### ****Frequency distributions****

Open the Home Mortgage Disclosure Act (HMDA) data set provided with this exercise.

Run frequencies in SPSS for the variables listed below. The name in parentheses is the variable name.

* ***SPSS Commands*:** Analyze 🡪 Descriptive Statistics 🡪 Frequencies; Move variables into box; OK).
* Loan originated or denied (action\_taken\_name),
* Loan purpose: home purchase, refinance, home improvement (loan\_purpose\_name),
* Loan type – either guaranteed by government program such as FHA, VA, or RHS, or conventional (loan\_type\_name)
* Denial reasons – 3 variables (denial\_reason\_1, denial\_reason\_2, denial\_reason\_3)
* Applicant’s ethnicity (applicant\_ethnicity\_name)
* Applicant’s race – 3 variables – applicants may select up to 5 races to identify themselves (applicant\_race\_name\_1, applicant\_race\_name\_2, applicant\_race\_name\_3)
* Applicant’s gender (applicant\_sex\_name)
* Owner occupancy status of the property – second homes, vacation homes, and rental properties are classified as “not owner occupied” (owner\_occupancy\_name)
* Property type (property\_type\_name)
* ***SPSS Shortcut:*** To get summary statistics of a particular variable (including frequencies), you may simply right-click on the variable in Data View.

***SPSS Output*:**

Frequency tables include four columns of numbers.

The “frequency” column tells you the number of loan applications that correspond to each category in the variable.

For the HMDA 2017 State of California data (all records):

For the variable action\_taken\_name, there are 3 categories: “application approved but not accepted,” “application denied by financial institution,” and “loan originated.” The SPSS output shows that of a total of 1,048,575 loan applications, 53,535 were approved but not accepted, 117,287 were denied, and 877,753 were originated.

The “percent” column converts the frequencies to percentages using all cases (1,048,575) as the denominator. For the variable action\_taken\_name, 5.1% of the loans were approved but not accepted, 11.2% were denied, and 83.7% were originated.

The “valid percent” column accounts for missing cases. For action\_taken\_name all the percentages are valid because none of the information is missing. For some of the variables we will examine later (e.g., applicant income) some of the information was unavailable (perhaps the applicant didn’t know or refused to answer the question) and the “valid percent” will reflect the difference by using only those cases with valid (complete) information. The denominator will be reduced for these variables.

The “cumulative percent” column cumulates the valid percentages into a running total.  Look at the table for action\_taken\_name.  The first entry in the cumulative percent column is 5.1 because 5.1% of the cases were approved but not accepted.  The second entry is 16.3 because 16.3% were either approved but not accepted or denied by the financial institution.  The third entry is 100.0 because all 100% of the cases were either approved but not accepted, denied, or resulted in a loan originated by the financial institution. This feature is more useful with categories that are continuous or ordinal such as “strongly disagree,” “disagree,” neither agree nor disagree,” “agree,” “strongly agree.”

Note – for the denial variables (denial\_reason\_name\_1, 2, and 3), most cases were approved, and were not denied. So, in these cases the values for denial\_reason\_name\_1, 2, and 3 are blank. For example, looking at the frequency output for denial\_reason\_name\_1 the category is blank for 969,736 loan applications.

* 1. **Rank-order, crosstabs**

Social scientists look for variations. Variations in data may occur over time, in different places, or among different groups of people. If there is no variation in data it suggests that everyone or everything is the same with respect to those measures.

Let’s say we want to look at loan applications, approvals and denials by county. We would do a crosstabulation which creates a table showing the frequencies, or number of cases, for each combination of variables.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Crosstabs; Put county\_name in the Row(s) box and action\_taken\_name in the Column(s) box; 🡪 OK.

***SPSS Output*:**

You will see a table with each county name listed as a row and the three categories of action\_taken\_name (application approved but not accepted; application denied by financial institution; loan originated) and total in columns across the top. Each cell indicates the number of cases that correspond to that combination of rows and columns. For example, using the 2017 State of California data (all records), in Humboldt County there were 2,394 loan applications in total, of which 382 were denied and 1,906 resulted in loans originated. (Note that there are 608 missing cases - applications that didn’t indicate a county.)

***Export SPSS output into Excel and explore:***

You can export the output table into an Excel spreadsheet by right-clicking on the SPSS output table and following the prompts.

After exporting into Excel, we can compute the rejection rate of loan applications by county. (Divide the number of applications that were denied by the total number of applications).

Using the sort function in Excel, rank the counties from highest rejection rate to lowest. Colusa County had the highest rejection rate (26.4%) and San Luis Obispo County had the lowest (2.8%). Refer to the county map in the Appendix (page 24). Does there appear to be a geographical pattern in the loan rejection rate by county? Compare the pattern you find in your data to the map on page 25, which shows Median Household Income for Counties 2013-2017.

* 1. **Recoding variables: Income brackets**

Let’s say one preliminary hypothesis is that rejection of mortgage applications is related to applicants having low income. There are two variables in the HMDA data that are measures of income. These include:

* The percentage of the median family income for the census tract where the property is located compared to the median family income for the Metropolitan Statistical Area /Metropolitan Division that the property is located in.[[7]](#footnote-7) (tract\_to\_msamd\_income)
* Applicant’s income in thousands (applicant\_income\_000s)

Generating a frequency table of either of these variables will create a *very long* output table that is not very useful.

For example, a frequency table of applicant\_income\_000s displays the number of applicants with a particular income, and each row may represent a $1 increment, and may contain only 1 person. For this reason, in order to use this variable, we must create brackets. In the map on page 25, for example, there are 5 brackets of median family income.

In our analysis, we will create 10 brackets:

1. 0 – $49,999
2. $50,000 – $74,999
3. $75,000 – $99,999
4. $100,000 – $149,999
5. $150,000 – $199,999
6. $200,000 – $249,999
7. $250,000 – $299,999
8. $300,000 – $349,999
9. $350,000 – $499,999
10. $500,000 and up

To create brackets of income, we will use the “recode” function in SPSS.

* ***SPSS Commands:*** Transform 🡪 Recode into *different* variables; Move applicant\_income\_000s into the Numeric Variables box. Under “Output Variable” create a new name for your recoded variable (app\_inc\_recode\_3) Then 🡪 Old and New Values…
* ***Note 1:*** you could recode your variable into the *same* variable without creating a new one, but this will wipe out the original data and will prevent you from recoding it into different brackets in the future, so it’s not recommended.
* ***Note 2:*** The data set already contains this recoded variable, app\_inc\_recode\_2.

Now you will put in the ranges of income in the box as shown:

1. Click “Range, LOWEST through value:” and enter 39.999 in the box. (Remember, the variable is income in 1,000s). Then click New Value and enter 1 in the box; 🡪 Add.
2. Click “Range” and enter 40.0 and 49.999 in the boxes. Then click New Value and enter 2 in the box; 🡪 Add.
3. Continue for all 7 categories of income.



1. Now, when you examine your data, you should see the new variable.
2. Go to variable view (click on tab at the bottom) and on the row for app\_inc\_recode\_3 click on “Values,” and enter the value labels for the brackets (i.e., 1.00 = “less than or equal to 49,999”….)



* 1. **Histograms**

Histograms are graphical representations of frequencies. A histogram can tell us if our data is normally distributed, which is important to further statistical analysis. In this example, let’s create a histogram of app\_inc\_recode\_2.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Frequencies; Put app\_inc\_recode\_2 in the Variable(s) box; 🡪 Charts 🡪 Histograms 🡪 Continue 🡪 OK.

You will see the output below. Double-clicking this output chart will bring up an editor window that you can use to change the axes titles, axes increments, fonts, colors, and other aspects of the graph.

What the histogram shows us is that the largest number of applicants, about 239,000, had incomes in the range of $100,000 to $149,999. This is consistent with the summary descriptive statistics below.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Frequencies; Put applicant\_income\_000s in the Variable(s) box; 🡪 Statistics 🡪 Click Mean and Median 🡪 OK.

|  |
| --- |
| **Statistics** |
| applicant\_income\_000s  |
| N | Valid | 991654 |
| Missing | 56921 |
| Mean | 158.83 |
| Median | 105.00 |
| Minimum | 1 |
| Maximum | 610715 |



Later, we might want to revise or “fine tune” our brackets so that we have a bracket for the applicants who are between $105,000 (the median) and $159,000 (the mean); this way we can easily show the number who are above/below the mean and median.

Note that the histogram is skewed; the mean is higher than the median, due to the applicants with very high incomes.

* 1. **Creating a new variable: Combine race and ethnicity**

In order to determine if there is discrimination in mortgage lending, HMDA mandates that lenders collect and report data on the race and ethnicity of loan applicants.

Applicants and co-applicants may choose up to 5 races to identify themselves and there are 5 categories that they may choose from: White; Black or African American; Asian; Native Hawaiian or Other Pacific Islander; American Indian or Alaska Native.

Applicants and co-applicants may choose one of two ethnicities: Hispanic or Latino, or Not Hispanic or Latino.

In the HMDA data, the following variables represent race/ethnicity:

* applicant race\_name\_1 thru applicant\_race\_name\_5
* co\_applicant\_race\_name\_1 thru co\_applicant race\_name\_5
* applicant\_ethnicity\_name
* co\_applicant\_ethnicity\_name

Let’s examine the data on applicant race and ethnicity.

In the frequency table for applicant\_race\_name\_1 below, we see that in addition to the 5 categories of race, there are 2 categories indicating that information was not collected from some applicants.

|  |
| --- |
| **applicant\_race\_name\_1** |
|  | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | American Indian or Alaska Native | 9714 | .9 | .9 | .9 |
| Asian | 147245 | 14.0 | 14.0 | 15.0 |
| Black or African American | 43180 | 4.1 | 4.1 | 19.1 |
| Information not provided by applicant in mail, Internet, or telephone application | 169071 | 16.1 | 16.1 | 35.2 |
| Native Hawaiian or Other Pacific Islander | 10595 | 1.0 | 1.0 | 36.2 |
| Not applicable | 14712 | 1.4 | 1.4 | 37.6 |
| White | 654058 | 62.4 | 62.4 | 100.0 |
| Total | 1048575 | 100.0 | 100.0 |  |

In the frequency table for applicant\_ethnicity\_name below, we see that in addition to the 2 categories of ethnicity, there are 2 categories indicating that information was not collected from some applicants.



Demographers often combine race and ethnicity when analyzing population data. A commonly-used list of demographic categories is:

* 1. White, Non-Hispanic
	2. African American, Non-Hispanic
	3. Asian, Non-Hispanic
	4. Native Hawaiian or Other Pacific Islander, Non-Hispanic
	5. American Indian or Alaska Native, Non-Hispanic
	6. Two or More Races, Non-Hispanic
	7. Hispanic, All Races

Sometimes, if the numbers of people in categories 4 through 6 are relatively small, these will be aggregated and reported as “Other.”

To create the above categories, we need to combine the race and ethnicity variables. This means we will use the recode function with if/then conditions for inclusion into the new categories.

***Note:*** *I have created this combined variable, App\_Race\_Eth, and it’s included in the dataset provided with this exercise. For those who want to practice creating a new variable, or are using a different dataset, the steps are described below. Or, you may wish to repeat these steps to create the same variable for the co-applicants. Otherwise, you may use the variable provided (App\_Race\_Eth) and skip to Exercise 6. Crosstabs, below.*

First, we create new variables called “app\_race\_code\_1 thru 5” which will take on a value of 1 through 7 based on the values of the variable (1 thru 5 as listed above; 6 = not applicable and 7 = information not provided…).

* ***SPSS Commands:*** Transform 🡪 Recode into Different Variables; input Old and New values into the box. SPSS will save these, so doing the variables 2 through 5 is much quicker.

|  |  |
| --- | --- |
| **Old Values** | **New Values** |
| White  | 1 |
| Black or African American | 2 |
| Asian | 3 |
| Native Hawaiian or Other Pacific Islander | 4 |
| American Indian or Alaska Native | 5 |
| Not applicable | 6 |
| Information not provided by applicant in mail, Internet, or telephone application | 7 |

* Using “recode into same variable” assign values of 0 to the “system- or user-missing” blank cells in the app\_race\_code variables. This is useful for app\_race\_code\_2 and beyond, because those applicants who identified themselves as only 1 race have left these cells blank. If race\_code\_2 has a value, it means the person identifies as mixed-race.
* Go into Variable View and define the values of the app\_race\_code series so you don’t forget what they mean.

Repeat the above process to create “app\_eth\_code” from “applicant\_ethnicity\_name.”

* ***SPSS Commands:*** Transform 🡪 Recode into Different Variables; input Old and New values into the box.

|  |  |
| --- | --- |
| **Old Values** | **New Values** |
| Hispanic | 1 |
| Non-Hispanic | 2 |
| Not applicable | 3 |
| Information not provided by applicant in mail, Internet, or telephone application | 4 |

Now, we will create a new variable called App\_Race\_Eth by recoding app\_race\_code\_1 into a different variable called App\_Race\_Eth. To create the new variable we will choose cases based on the criteria below using “if/then” statements in SPSS. We can only do 1 if/then statement at a time so we will recode into App\_Race\_Eth a total of 7 times.

The categories of App\_Race\_Eth are:

1. White, Non-Hispanic

app\_race\_code\_1 = 1 (White) AND app\_eth\_code = 2 (Not Hispanic or Latino) AND app\_race\_code\_2 = 0 (if app\_race\_code\_2 = 0 it means that the applicant has identified him/herself as being of only 1 race).

Old values = 1; New values = 1

1. African American, Non-Hispanic

app\_race\_code\_1 = 2 (Black or African American) AND app\_eth\_code = 2 AND app\_race\_code\_2 = 0

Old values = 2; New values = 2

1. Asian, Non-Hispanic

app\_race\_code\_1 = 3 (Asian) AND app\_eth\_code = 2 AND app\_race\_code\_2 = 0

Old values = 3; New values = 3

1. Native Hawaiian or Other Pacific Islander, Non-Hispanic

app\_race\_code\_1 = 4 (Native Hawaiian or Other Pacific Islander) AND app\_eth\_code = 2 AND app\_race\_code\_2 = 0

Old values = 4; New values = 4

1. American Indian or Alaska Native, Non-Hispanic

app\_race\_code\_1 = 5 (American Indian or Alaska Native) AND app\_eth\_code = 2 AND app\_race\_code\_2 = 0

Old values = 5; New values = 5

1. Mixed Race, Non-Hispanic

app\_race\_code\_2 > 0 AND app\_eth\_code = 2

Old values = 1, 2, 3, 4, 5; New values = 6

1. Hispanic, All Races

app\_eth\_code = 1

Old values = 1, 2, 3, 4, 5; New values = 7

* ***SPSS Commands:*** Transform 🡪 Recode into Different Variables; Put app\_race\_code\_1 and App\_Race\_Eth” into the Numeric Variable and Output Variables boxes. Click “Old and New Values…” and input the values in 1 thru 7 above, depending on which category you’re doing. (For example, if you are creating the category for Asian Non-Hispanic, enter Old values 3, New values 3.) Then, click “IF” and enter the If/then statement in 1 thru 7 above, depending on which category you’re doing. See below.



* 1. **Crosstabs, Chi-Square: Race & ethnicity x action taken**

Our inquiry is whether different demographic groups are treated the same or differently in the mortgage application process. We hypothesize that there will be significant differences; due to the history of redlining and discrimination in housing we hypothesize that White, Non-Hispanic mortgage applicants will be approved at higher rates than other groups.

Now that we have our demographic variable for race and ethnicity App\_Race\_Eth, we can crosstabulate this with action\_taken\_name to see if different groups of applicants are approved and rejected at the same rate. The crosstabulation command will create an output table of frequencies that indicate the number of cases that fall into the particular categories.

Let’s see whether the different groups were approved or rejected at the same rate.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Crosstabs; Put App\_Race\_Eth in the Rows box and action\_taken\_name in the Columns box 🡪 OK.

***SPSS Output:***

The output table is shown below. Each demographic group is listed as a row and the three categories of action\_taken\_name (application approved but not accepted; application denied by financial institution; loan originated) and total are in the columns across the top. Each cell indicates the number of cases that correspond to that combination of rows and columns. For example there were 12,905 Asian, Non-Hispanic applicants who were rejected.

Using this data, does it appear that different groups were rejected at different rates, or were all treated the same?

* Which group had the lowest rejection rate?
* Which group had the highest?
* What was the overall rate of rejection?



Next, let’s look at whether the frequencies we’ve obtained above are significantly different than what we’d expect to obtain due to random chance.

For example, there were 851,586 total applicants, of which 40,209 (4.7%) were Black, Non-Hispanic. There were a total of 88,986 applications that were rejected. Since 4.7% of the applicant population is Black, we’d expect 4,202 of the total who were rejected to be Black applicants. Instead, we see that 6,368 Black applicants were rejected.

On the flip side, 54.6% of the total applicants were White, Non-Hispanic. Were 54.6% of the *rejected* applicants White, Non-Hispanic?

We want to know whether these differences are significant. We can answer this question by using the Chi-Square test.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Crosstabs; Put App\_Race\_Eth in the Rows box and action\_taken\_name in the Columns box. Statistics 🡪 Chi-Square. Cells 🡪 Counts (select Observed); Percentages (select Row) 🡪 OK.
* ***Note:*** In this analysis we are investigating whether an applicant’s race/ethnicity has an effect on how they are treated in the mortgage application process. So the independent variable is App\_Race\_Eth and the dependent variable is action\_taken. In the crosstabs computation it doesn’t matter whether the independent or dependent variable is placed in the rows or the columns, but the percentages should be computed for the independent variable (in this case, the rows).

***SPSS Output:***

The SPSS output is shown below. First, the crosstabulation table clearly shows us the different rejection rates for each demographic group. For example, 9.7% of White, Non-Hispanic applicants were rejected compared to 18.2% of American Indian and 15.8% of Black, Non-Hispanic applicants.

The result is statistically significant with a large value of Chi-Square. So our research hypothesis is supported.





* 1. **Crosstabs, Chi-Square: Race & ethnicity x action taken x applicant’s income**

The work that we did in Exercises 2 and 3 suggested that mortgage rejection could be related to the applicants’ incomes, and we recoded the applicants’ income variable into a new variable with 10 brackets (app\_inc\_recode\_2).

Let’s see whether income is related to the two previous variables we examined: race/ethnicity and action taken. What do you expect the relationships between the three variables will be?

Run crosstabs:

app\_inc\_recode\_2 x App\_Race\_Eth

app\_inc\_recode\_2 x action\_taken

Describe your findings. With income as the independent variable (rows), read down the “application denied” column.

* 1. **Controlling for applicant’s income**

Our research question is whether different demographic groups experience differential treatment in the mortgage application process—whether racial and ethnic discrimination occurs. However, the results in Exercise 7 showed that applicant’s income is related to both Race/Ethnicity and mortgage approval. Perhaps racial and ethnic minorities are rejected at higher rates than Whites because their incomes are lower. For this reason we want to control for applicants’ income.

We do this by telling SPSS to divide our population into sub-groups based on income, and crosstabulate App\_Race\_Eth x action\_taken for each sub-group. We control for income by analyzing applicants who are all at the same income level. This way we can see whether there are still differences in rejection/approval rates for the different race/ethnic groups.

* ***SPSS Commands:*** Analyze 🡪 Descriptive Statistics 🡪 Crosstabs; Put App\_Race\_Eth in the Rows box and action\_taken\_name in the Columns box and app\_inc\_recode\_2 in the Layer 1 of 1 box. Statistics 🡪 Chi-Square. Cells 🡪 Counts (select Observed); Percentages (select Row) 🡪 OK.

C

The part of the output tables are shown on the following pages. If the relationship between an applicant’s race/ethnicity and mortgage approval was due to income, then the results for each of the income brackets wouldn’t be significant. Instead, we see that there is still a strong relationship.

|  |
| --- |
| **Chi-Square Tests** |
| app\_inc\_recode\_2 | Value | df | Asymptotic Significance (2-sided) |
| less than or equal to 49,999 | Pearson Chi-Square | 463.681b | 12 | .000 |
| Likelihood Ratio | 435.866 | 12 | .000 |
| N of Valid Cases | 89289 |  |  |
| 50,000 - 74,999 | Pearson Chi-Square | 428.093c | 12 | .000 |
| Likelihood Ratio | 397.121 | 12 | .000 |
| N of Valid Cases | 146208 |  |  |
| 75,000 - 99,999 | Pearson Chi-Square | 433.869d | 12 | .000 |
| Likelihood Ratio | 400.227 | 12 | .000 |
| N of Valid Cases | 147083 |  |  |
| 100,000 - 149,999 | Pearson Chi-Square | 557.060e | 12 | .000 |
| Likelihood Ratio | 506.838 | 12 | .000 |
| N of Valid Cases | 196618 |  |  |
| 150,000 - 199,999 | Pearson Chi-Square | 249.966f | 12 | .000 |
| Likelihood Ratio | 224.093 | 12 | .000 |
| N of Valid Cases | 101422 |  |  |
| 200,000 - 249,999 | Pearson Chi-Square | 86.619g | 12 | .000 |
| Likelihood Ratio | 76.724 | 12 | .000 |
| N of Valid Cases | 51141 |  |  |
| 250,000 - 299,999 | Pearson Chi-Square | 87.998h | 12 | .000 |
| Likelihood Ratio | 78.963 | 12 | .000 |
| N of Valid Cases | 28424 |  |  |
| 300,000 - 349,999 | Pearson Chi-Square | 51.243i | 12 | .000 |
| Likelihood Ratio | 46.155 | 12 | .000 |
| N of Valid Cases | 17252 |  |  |
| 350,000 - 499,999 | Pearson Chi-Square | 89.096j | 12 | .000 |
| Likelihood Ratio | 73.361 | 12 | .000 |
| N of Valid Cases | 22281 |  |  |
| 500,000 and up | Pearson Chi-Square | 115.495k | 12 | .000 |
| Likelihood Ratio | 97.170 | 12 | .000 |
| N of Valid Cases | 20008 |  |  |
| Total | Pearson Chi-Square | 3038.139a | 12 | .000 |
| Likelihood Ratio | 2808.915 | 12 | .000 |
| N of Valid Cases | 819726 |  |  |
| a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 132.26. |
| b. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.32. |
| c. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 28.76. |
| d. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 24.28. |
| e. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 28.84. |
| f. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 12.11. |
| g. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 7.13. |
| h. 1 cells (4.8%) have expected count less than 5. The minimum expected count is 2.90. |
| i. 4 cells (19.0%) have expected count less than 5. The minimum expected count is .98. |
| j. 4 cells (19.0%) have expected count less than 5. The minimum expected count is 1.90. |
| k. 3 cells (14.3%) have expected count less than 5. The minimum expected count is 2.94. |



* 1. **Further analysis – your turn**

Now it’s your turn – here are some ideas for further analysis.

1. If you did the analysis for the complete data set, do you get the same results using the second data set (home purchase, 1-4 family dwellings, owner-occupied)?
2. Income is just one of many factors that could explain why a mortgage application is rejected. Others include credit history, debt-to-income ratio, and insufficient collateral. The HMDA dataset includes denial reasons (denial\_reason\_name\_1 thru 3). Could denial reasons explain the differences in rejection rates among different demographic groups? Are there differences in denial reasons among different demographic groups? Re-create the analysis above using denial reasons instead of income.

*Hint: Most applicants were denied based on one reason, but some were denied based on 3 reasons. Using the recode function, you can create a new variable that accounts for applicants who have 2 or 3 reasons that they were denied similar to the way we created a new race/ethnicity variable.*

1. Are there differences between genders in mortgage denial rates? Re-create the analysis above to examine differences between male and female applicants.
2. Approval or denial of a mortgage application is just part of the story. Are there differences between demographic groups in terms of the interest rate that the applicant receives, or the amount of the loan that the applicant receives? Use the variables rate\_spread and loan\_amount\_000s to examine this question.
3. Repeat the analysis for a particular county (or group of counties) in California.
4. Compare California to another state.
5. Have there been changes over time?

**APPENDIX 1. COUNTY MAPS**



Source: geology.com <https://geology.com/state-map/california.shtml>



Source: U.S. Census American Community Survey <https://www.census.gov/content/dam/Census/library/visualizations/2018/comm/acs-5yr-mhi-all-counties.jpg>

1. <https://www.consumerfinance.gov/data-research/hmda/explore> [↑](#footnote-ref-1)
2. Interactive maps are available at Mapping Inequality: Redlining in New Deal America <https://dsl.richmond.edu/panorama/redlining/#loc=4/36.71/-96.93&opacity=0.8> ; for descriptions of the project also see Miller, Greg. 2016. “Newly Released Maps Show How Housing Discrimination Happened.” *National Geographic* <https://news.nationalgeographic.com/2016/10/housing-discrimination-redlining-maps/>

and Domonoske, Camila. 2016. “Interactive Redlining Map Zooms in on America’s History of Discrimination, *National Public Radio* (October 19, 2016).<https://www.npr.org/sections/thetwo-way/2016/10/19/498536077/interactive-redlining-map-zooms-in-on-americas-history-of-discrimination> [↑](#footnote-ref-2)
3. Bartlett, R., Morse, A., Stanton, R., Wallace, N. 2018. Consumer-Lending Discrimination in the Era of FinTech. UC Berkeley Public Law Research Paper <https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf> and Turner, M.A., Skidmore, F., Eds. 1999. Mortgage Lending Discrimination: A Review of the Existing Evidence. The Urban Institute. <https://www.urban.org/sites/default/files/publication/66151/309090-Mortgage-Lending-Discrimination.PDF> [↑](#footnote-ref-3)
4. Concumer Financial Protection Bureau. “About HMDA” <https://www.consumerfinance.gov/data-research/hmda/> [↑](#footnote-ref-4)
5. Consumer Financial Protection Bureau. <https://www.consumerfinance.gov/data-research/hmda/explore> [↑](#footnote-ref-5)
6. <https://www.ffiec.gov/hmda/forms.htm> [↑](#footnote-ref-6)
7. According to the CFPB website, “An MSA is a region with a relatively high population density at its core (usually a single large city) and close economic ties throughout. Larger MSAs are divided into MDs.” <https://cfpb.github.io/api/hmda/fields.html> [↑](#footnote-ref-7)