Modeling Mortgage Applications

Kevin Wang

Introduction

- The Consumer Financial Protection Bureau (CFPB) is a U.S. government agency responsible for protecting consumers from unfair treatment by financial institutions (banks, lenders, etc.)

- The Home Mortgage Disclosure Act (HMDA) requires many institutions to publicly disclose mortgage data to the public
  - Information is publicly available to help officials make decisions and enact policies
  - Data can reveal potential discriminatory lending patterns
  - Historical data can also be examined to detect trends over time

Data Acquisition and Exploration

Data Acquisition

- CFPB has data from 2007-2017 available for download
  - Can specify region, which kind of mortgages
- The data is available in encoded form, but plain language can also be included
  - To save memory and processing power, only examine encoded data and consult the code explanation sheet to interpret the results

```
 [1] "as_of_year"                     "respondent_id"                  "agency_code"
 [4] "loan_type"                      "property_type"                  "loan_purpose"
 [7] "owner_occupancy"                "loan_amount_000s"               "preapproval"
[10] "action_taken"                   "msamd"                          "state_code"
[13] "county_code"                    "census_tract_number"            "applicant_ethnicity"
[16] "co_applicant_ethnicity"         "applicant_race_1"               "applicant_race_2"
[19] "applicant_race_3"               "applicant_race_4"               "applicant_race_5"
[22] "co_applicant_race_1"            "co_applicant_race_2"            "co_applicant_race_3"
[25] "co_applicant_race_4"            "co_applicant_race_5"            "applicant_sex"
[28] "co_applicant_sex"               "applicant_income_000s"          "purchaser_type"
[31] "denial_reason_1"                "denial_reason_2"                "denial_reason_3"
[34] "rate_spread"                    "hoepa_status"                   "lien_status"
[37] "edit_status"                    "sequence_number"                "population"
[40] "minority_population"            "hud_median_family_income"       "tract_to_msamd_income"
[43] "number_of_owner_occupied_units" "number_of_1_to_4_family_units"  "application_date_indicator"
```

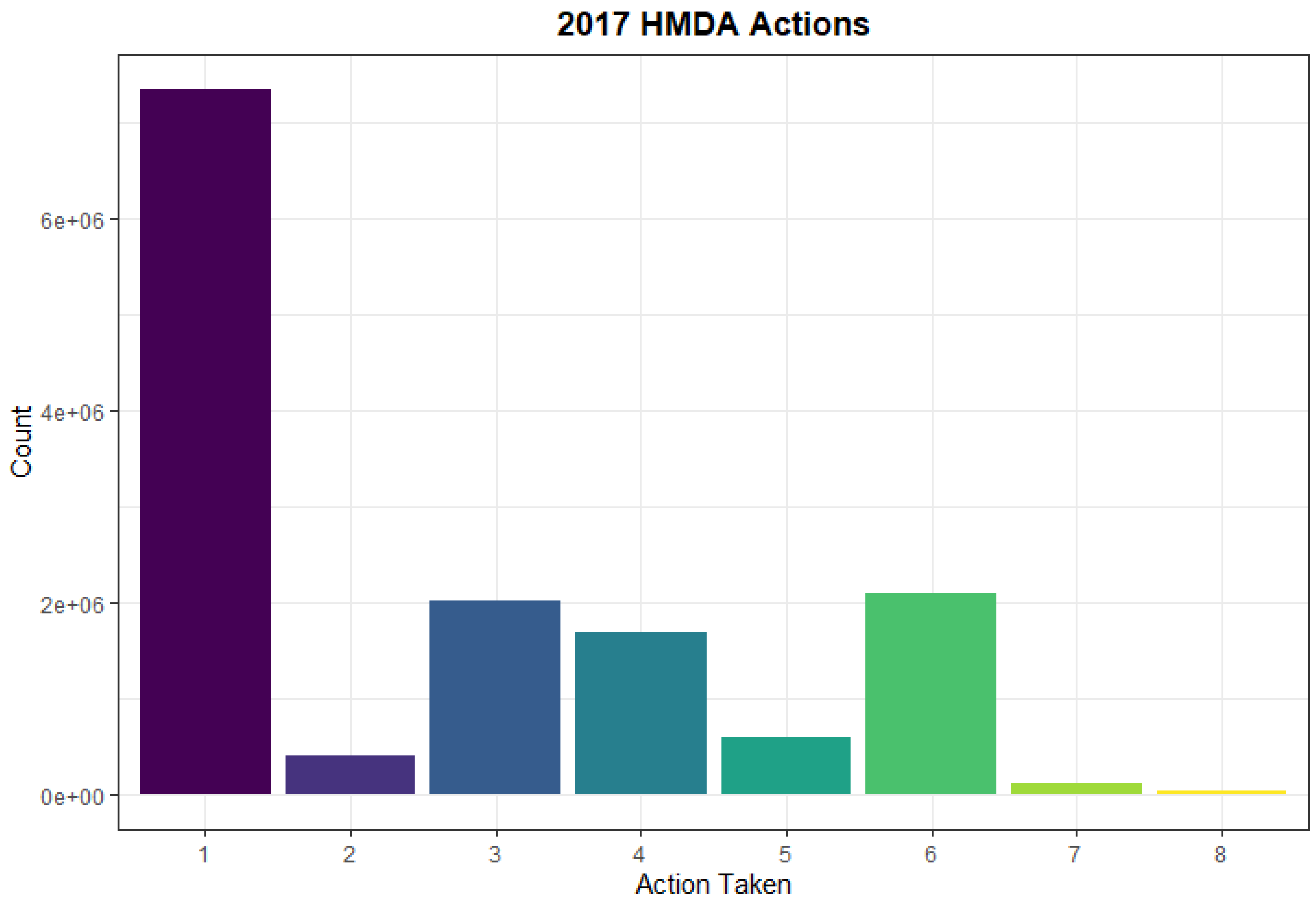
- For this project, start with data from 2017
  - 45 variables
  - 14,285,496 observations

Data Cleaning

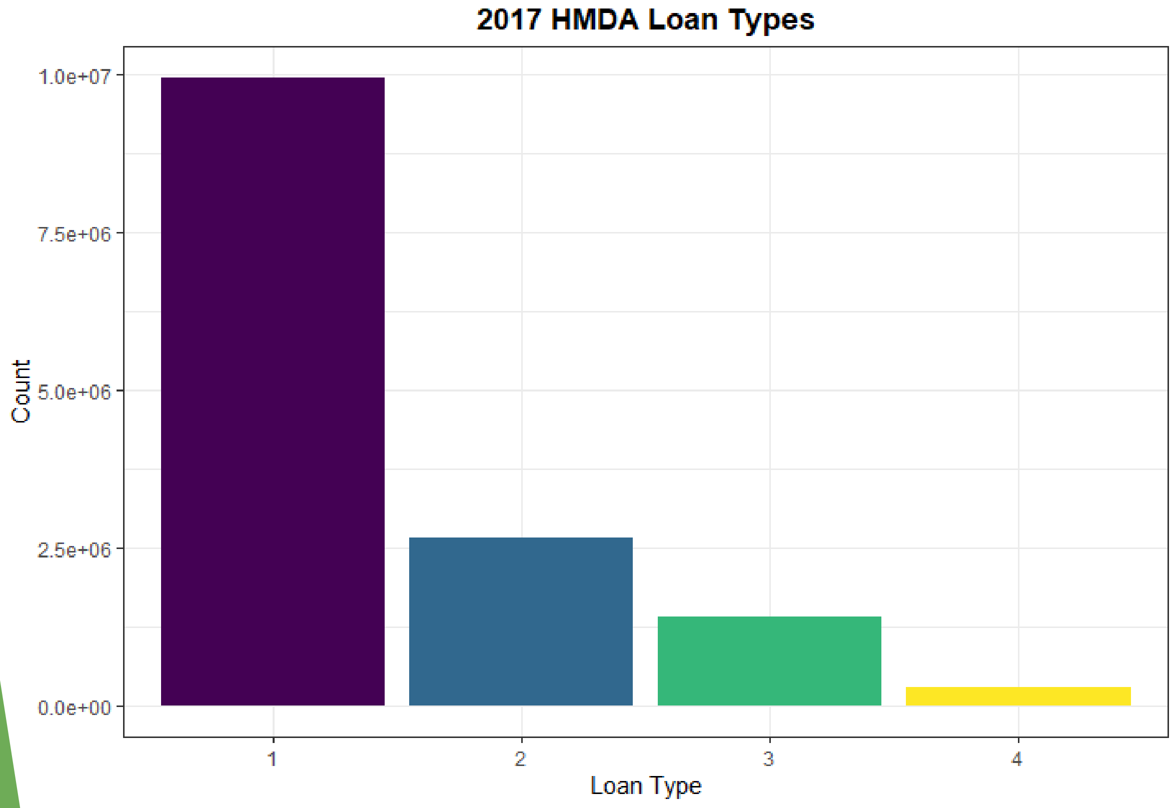
- Since there was so much data, discard fields when appropriate while cleaning
  - edit_status, sequence_number, application_date_indicator: NA only
  - as_of_year: all the same (2017); respondent_id
- Each transaction has an applicant and potentially a co-applicant
  - About 50% of rows had no co-applicant
- Each applicant/co-applicant can specify up to 5 races
  - Only 0.6% of applicants and 0.2% of co-applicants specified 2 or more races
  - Remove all secondary race features, create new Boolean feature for whether applicant was multiracial
- Action Taken: where the future dependent variable will be extracted from

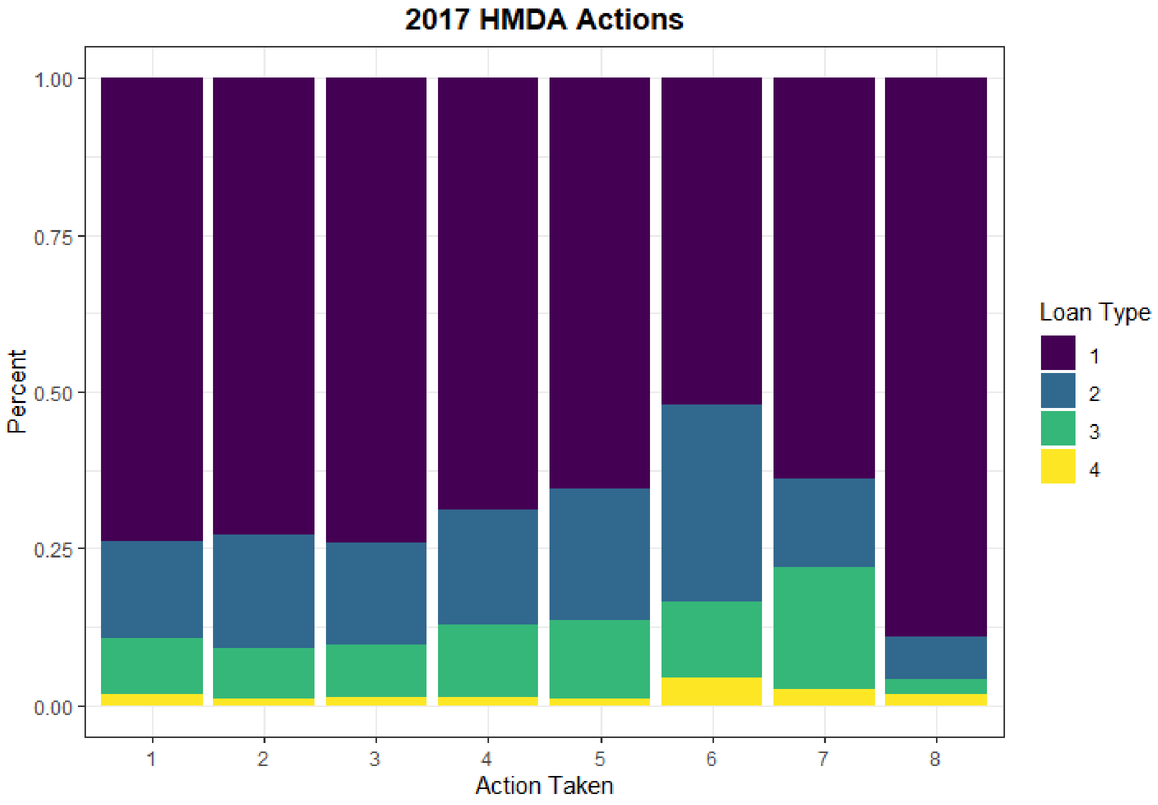
Action Taken:
1 -- Loan originated
2 -- Application approved but not accepted
3 -- Application denied by financial institution
4 -- Application withdrawn by applicant
5 -- File closed for incompleteness
6 -- Loan purchased by the institution
7 -- Preapproval request denied by financial institution
8 -- Preapproval request approved but not accepted (optional reporting)

Data Exploration



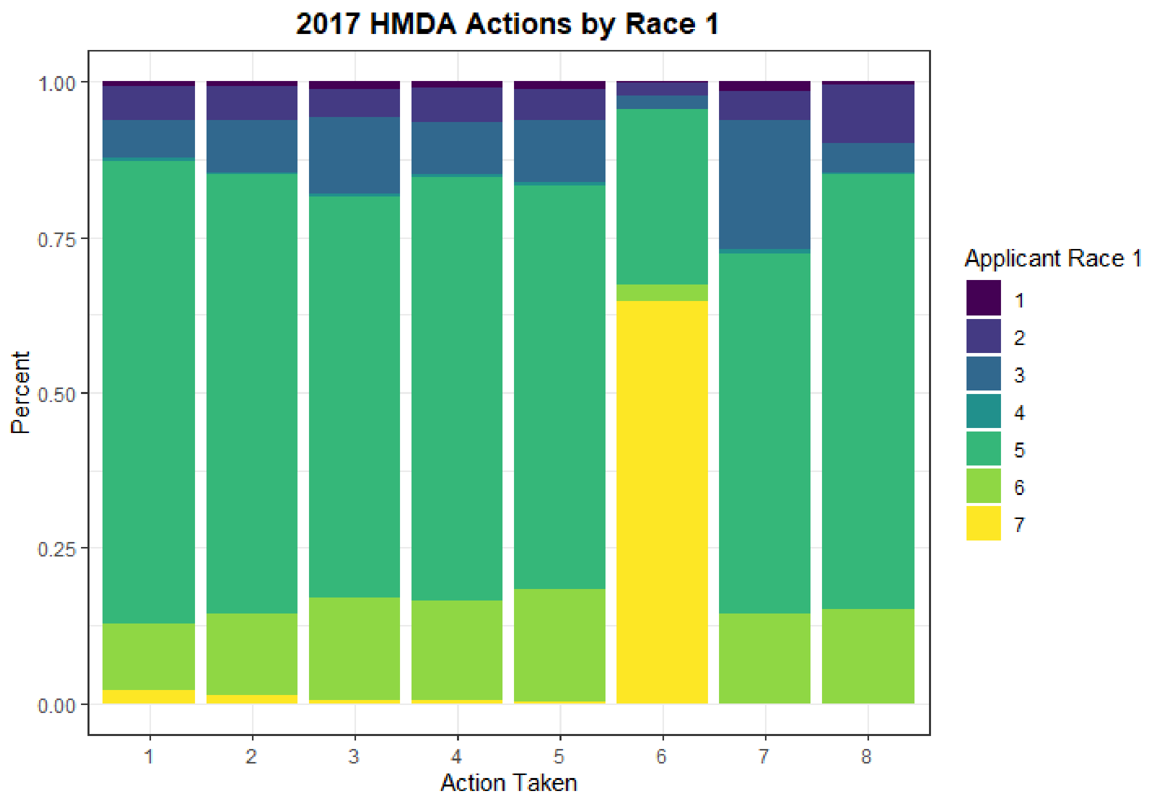
Data Exploration

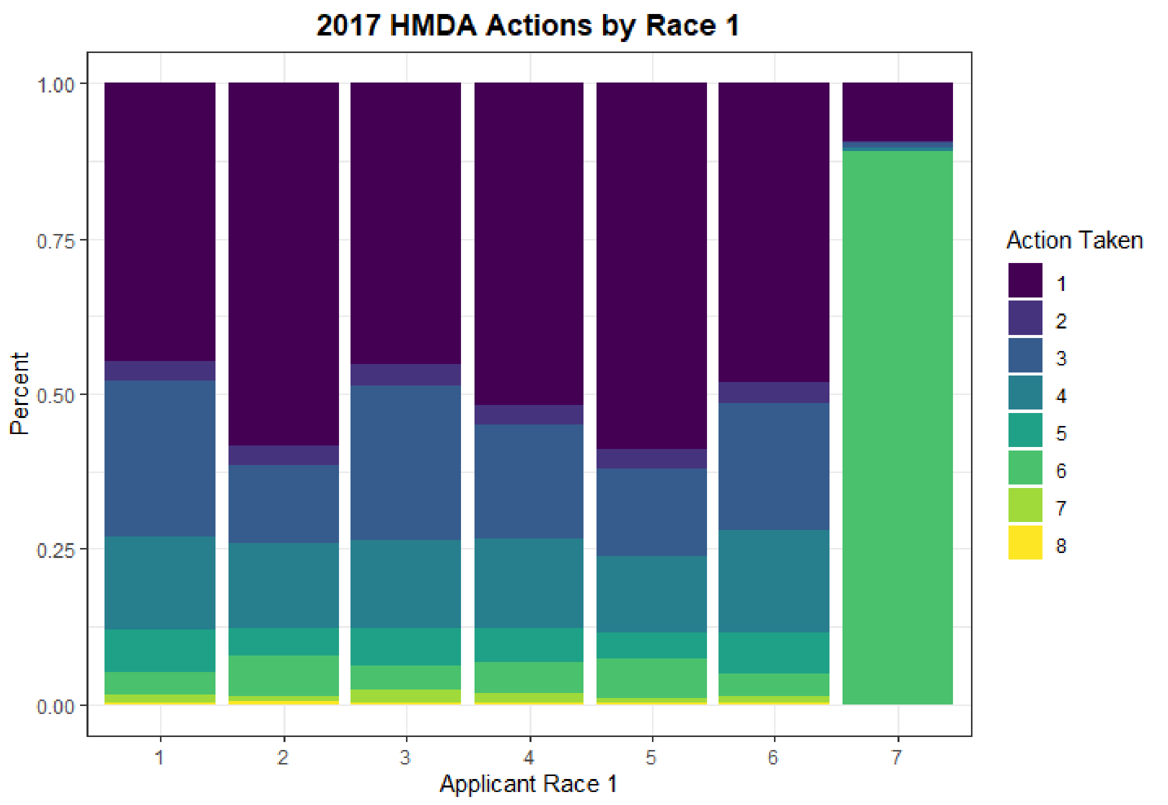




Loan Type:
1 -- Conventional (any loan other than FHA, VA, FSA, or RHS loans)
2 -- FHA-insured (Federal Housing Administration)
3 -- VA-guaranteed (Veterans Administration)
4 -- FSA/RHS (Farm Service Agency or Rural Housing Service)

Data Exploration





Race:

1 -- American Indian or Alaska Native

2 -- Asian

3 -- Black or African American

4 -- Native Hawaiian or Other Pacific Islander

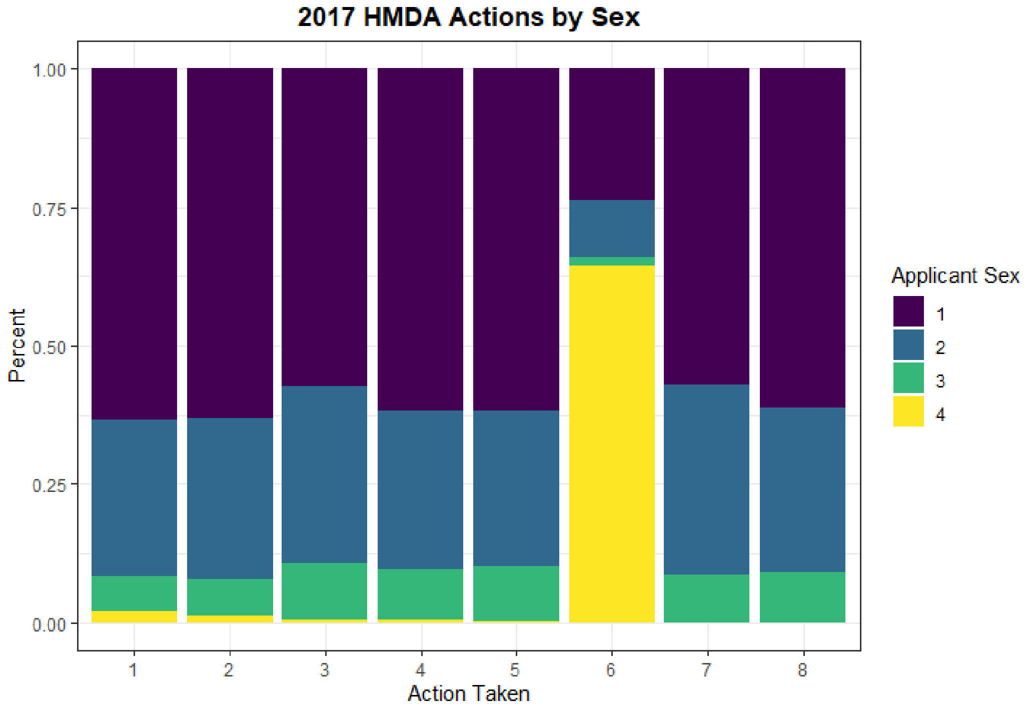
5 -- White

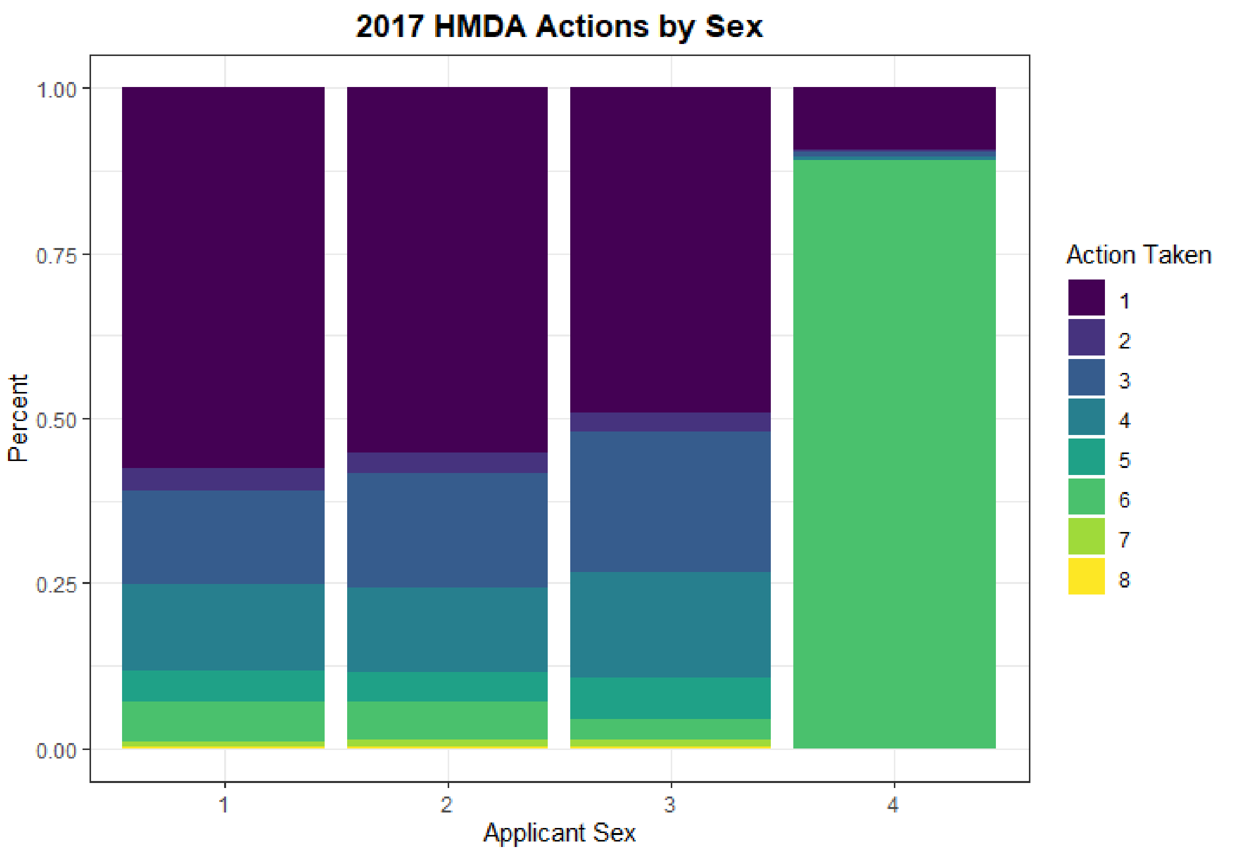
6 -- Information not provided by applicant in mail, Internet, or telephone application

7 -- Not applicable

8 -- No co-applicant

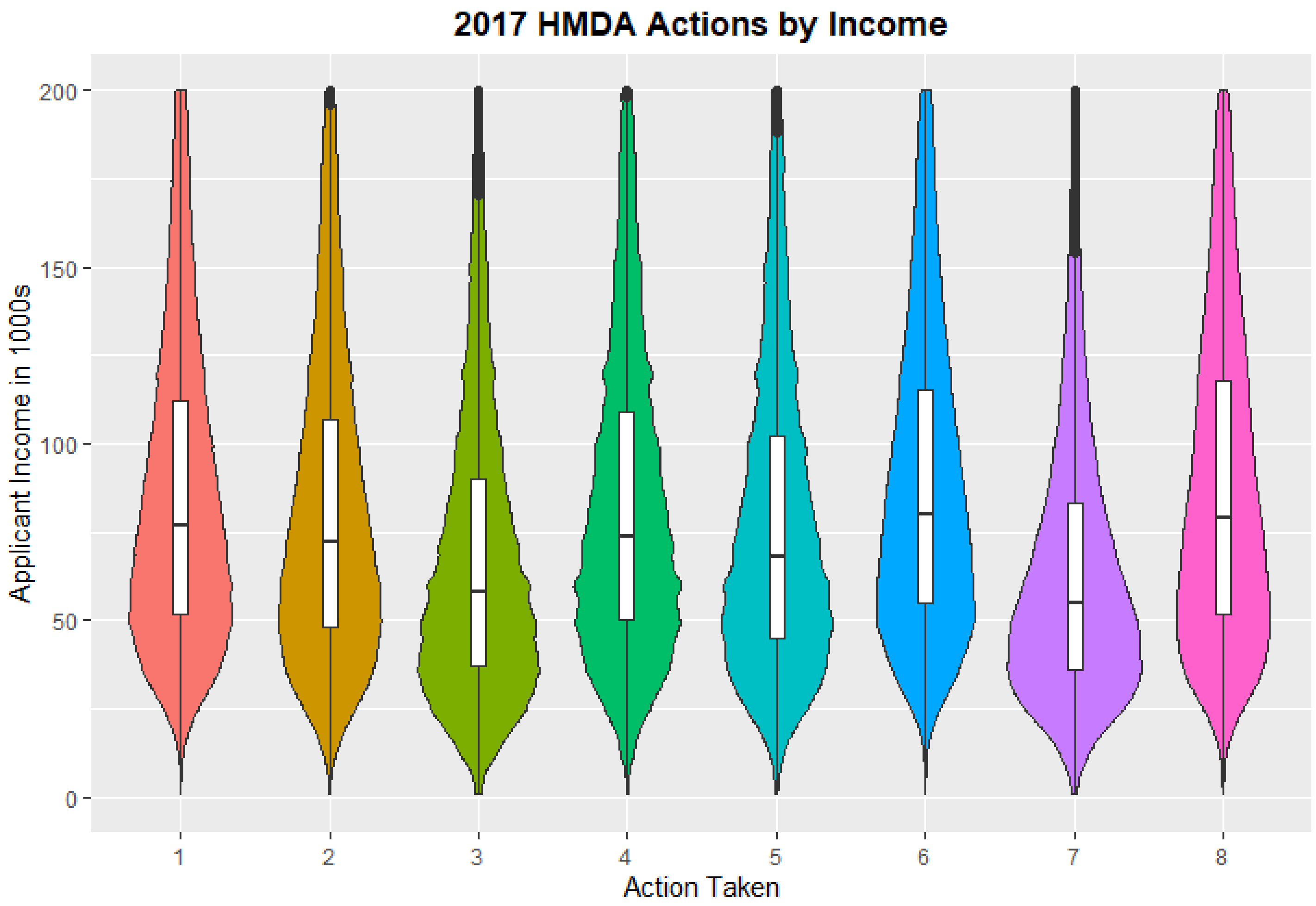
Data Exploration





Sex:
1 -- Male
2 -- Female
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 -- Not applicable
5 -- No co-applicant

Data Exploration



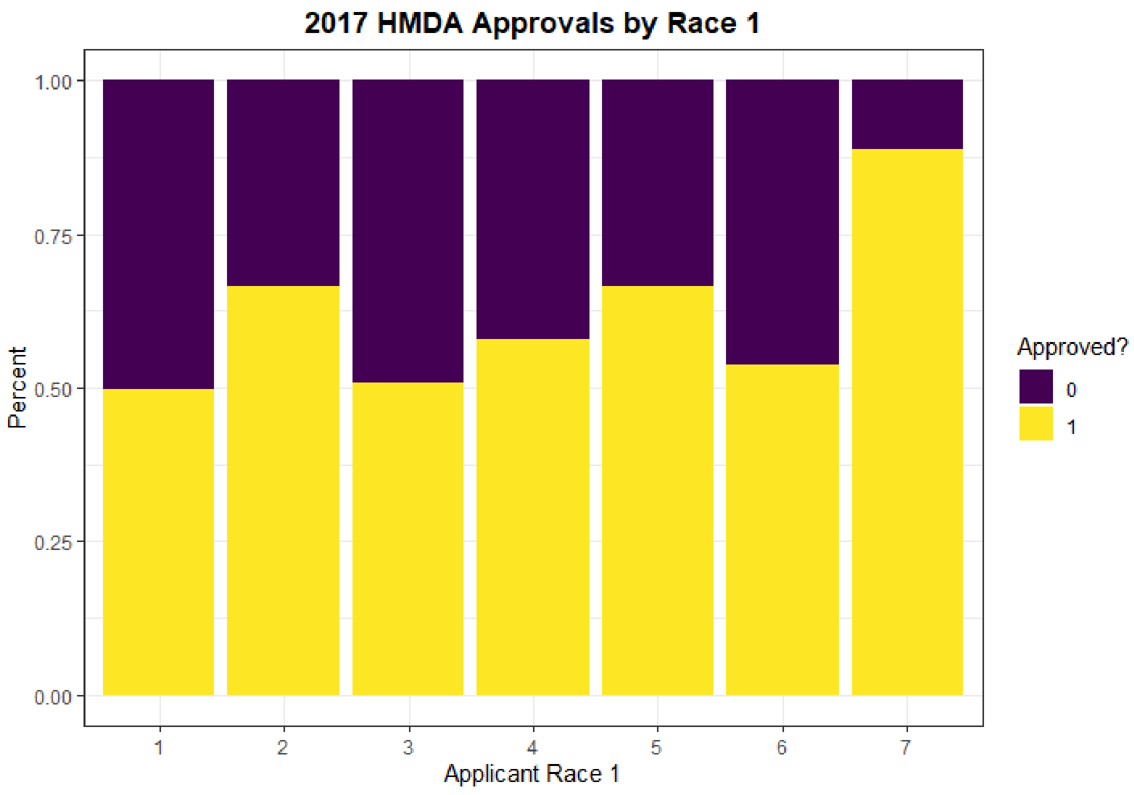
Action Taken:
1 -- Loan originated
2 -- Application approved but not accepted
3 -- Application denied by financial institution
4 -- Application withdrawn by applicant
5 -- File closed for incompleteness
6 -- Loan purchased by the institution
7 -- Preapproval request denied by financial institution
8 -- Preapproval request approved but not accepted (optional reporting)

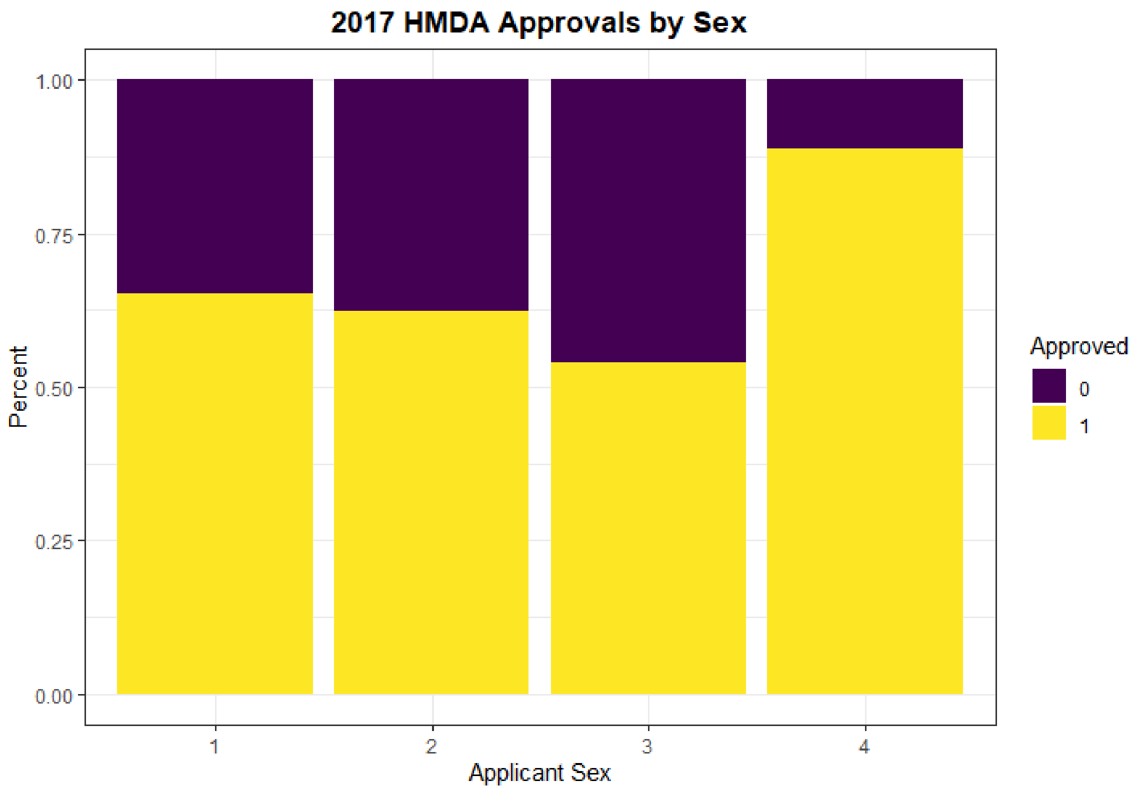
Data Cleaning

- Action 6: “Loan purchased by the institution”
  - Transactions involving a loan being sold from one institution to another
  - Remove these rows
- Action 4: “Application withdrawn by applicant”
  - The applicant took themselves out of consideration
  - Remove these rows
- Create a new feature to act as a dependent variable in future modeling
  - Group actions 1, 2, 8
    - Loan originated, application/preapproval was approved but not accepted
  - Remainder actions 3, 5, 7
    - Application/preapproval denied, closed for incompleteness
  - Now dependent variable is binary for easier use and ability to use logistic regression

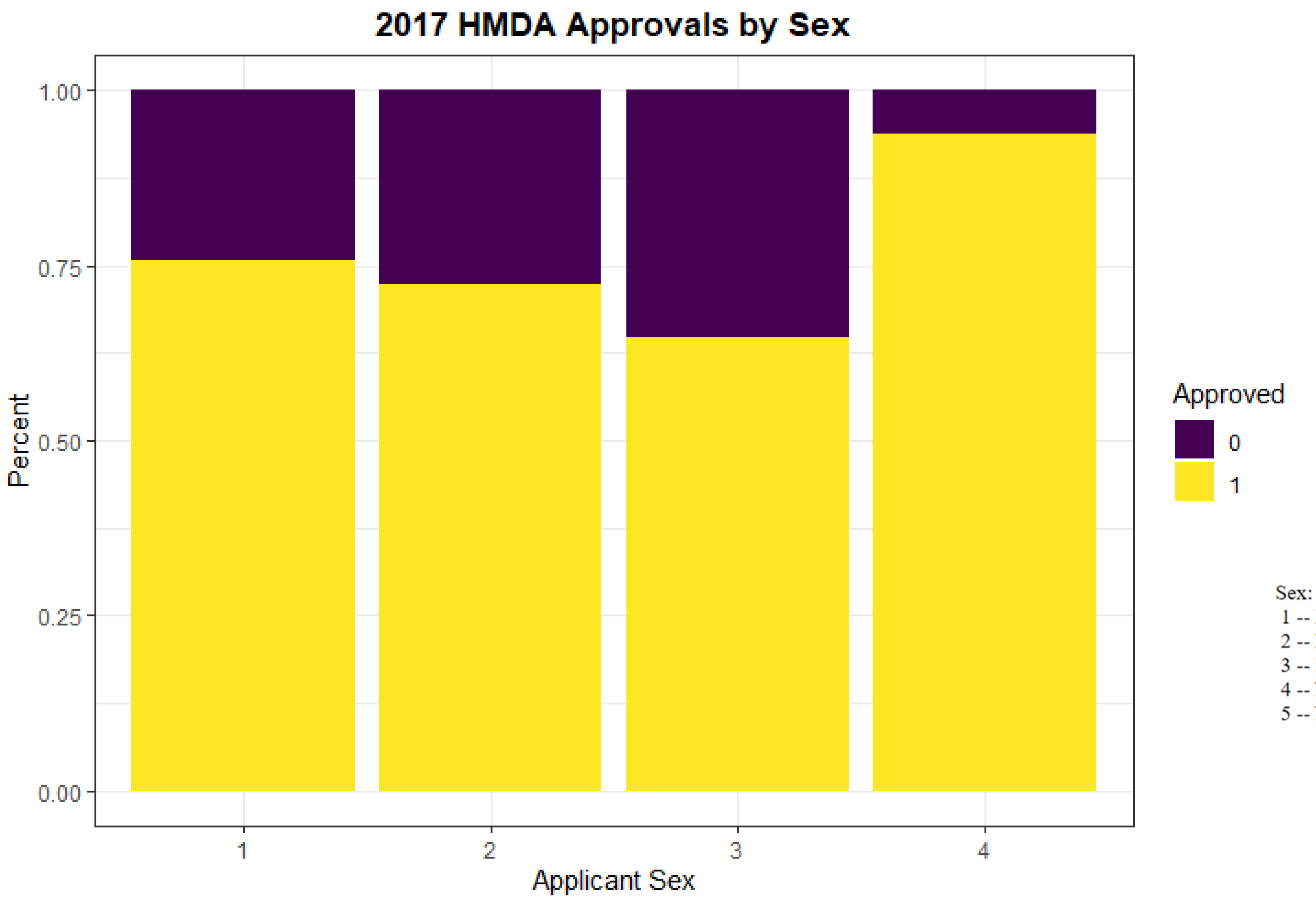
Data Exploration

- Example plots with only action 6 removed





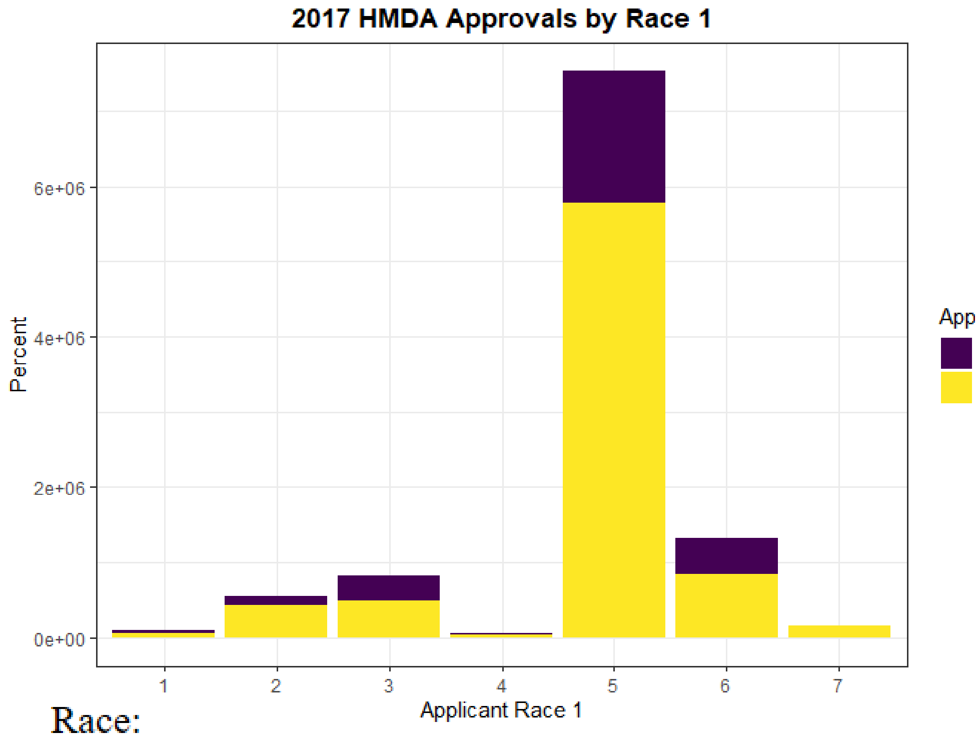
Data Exploration

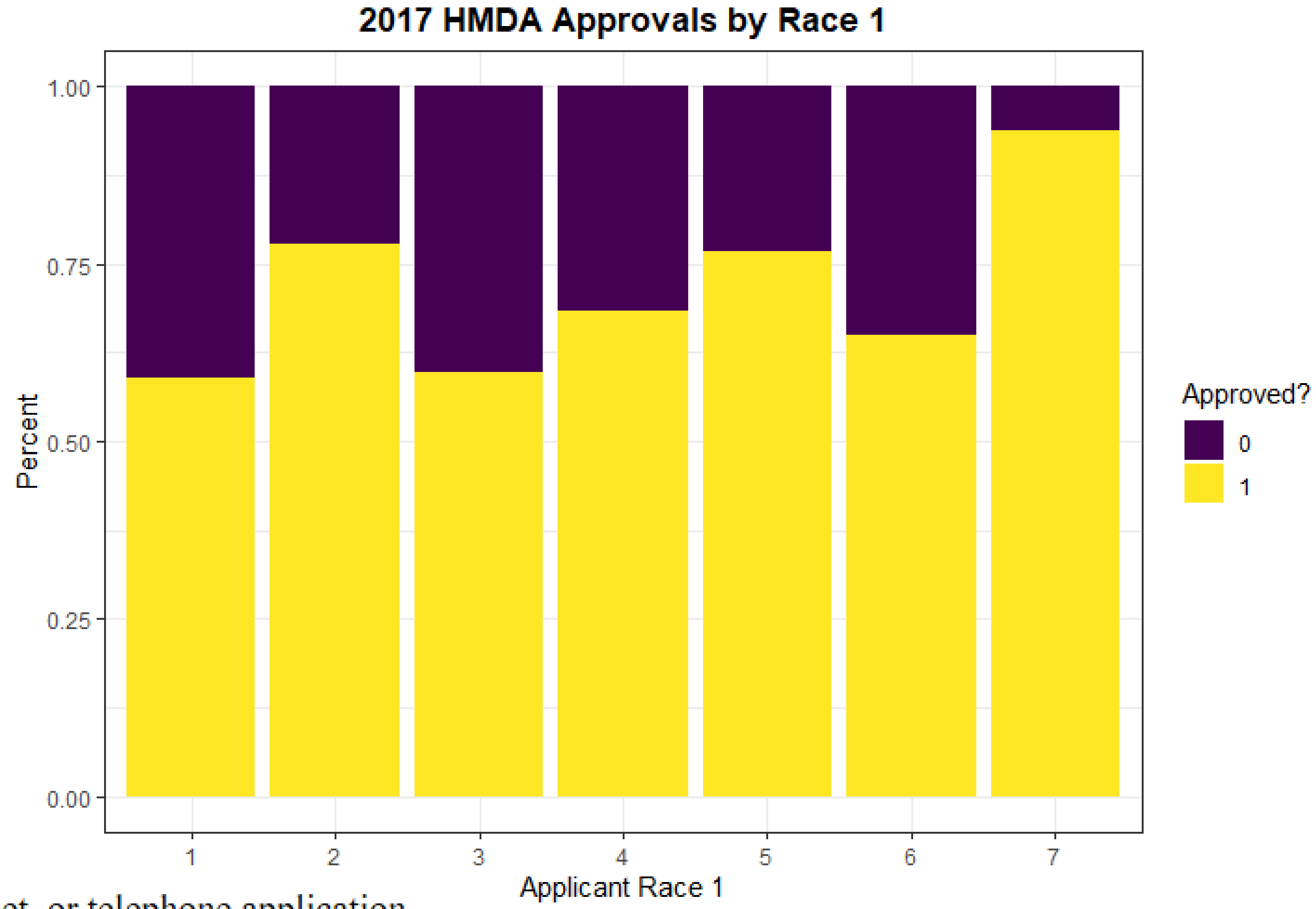


- Future plots in these slides only show results of removing both action 4 and 6

Sex:
1 -- Male
2 -- Female
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 -- Not applicable
5 -- No co-applicant

Data Exploration

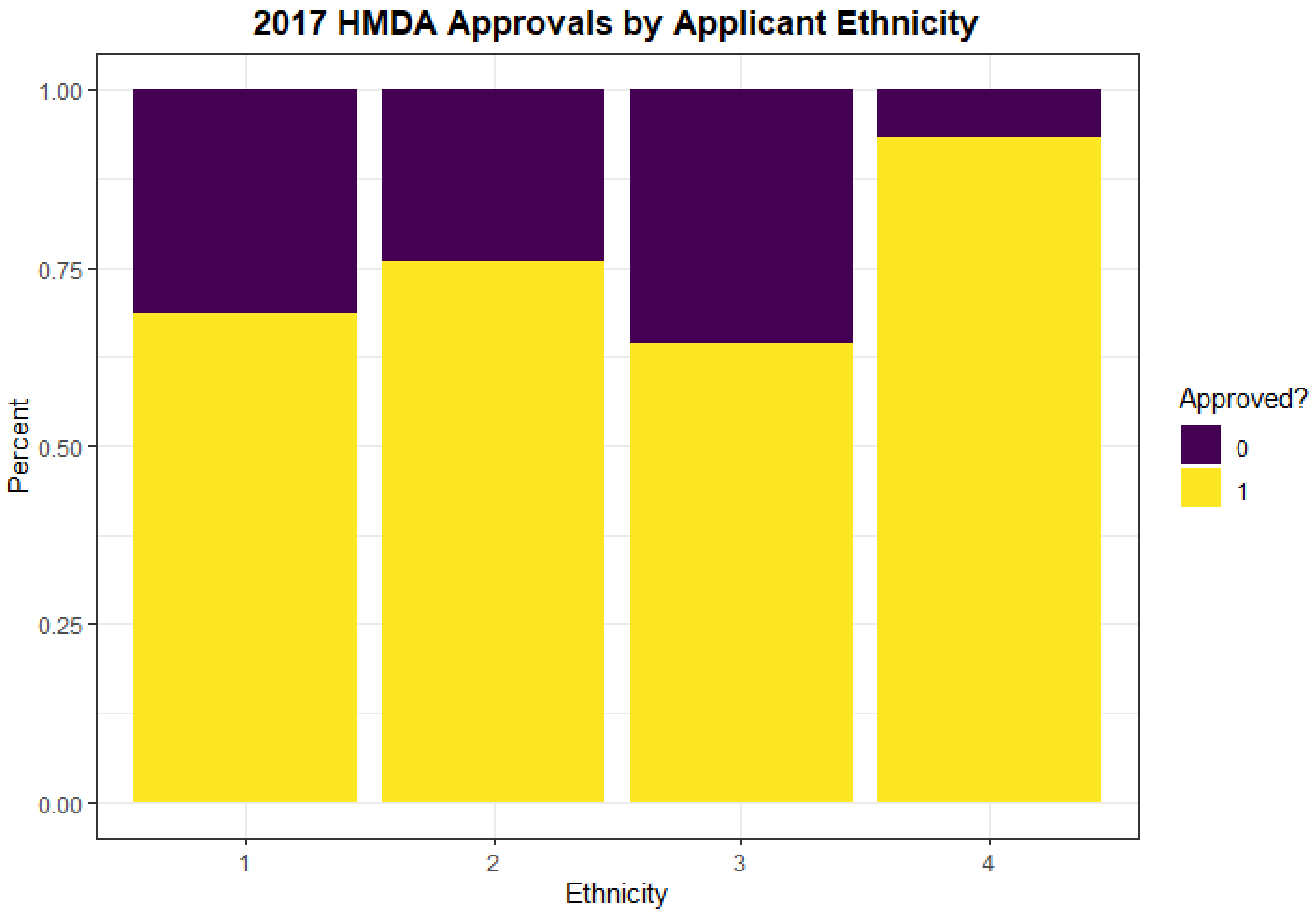




Race:

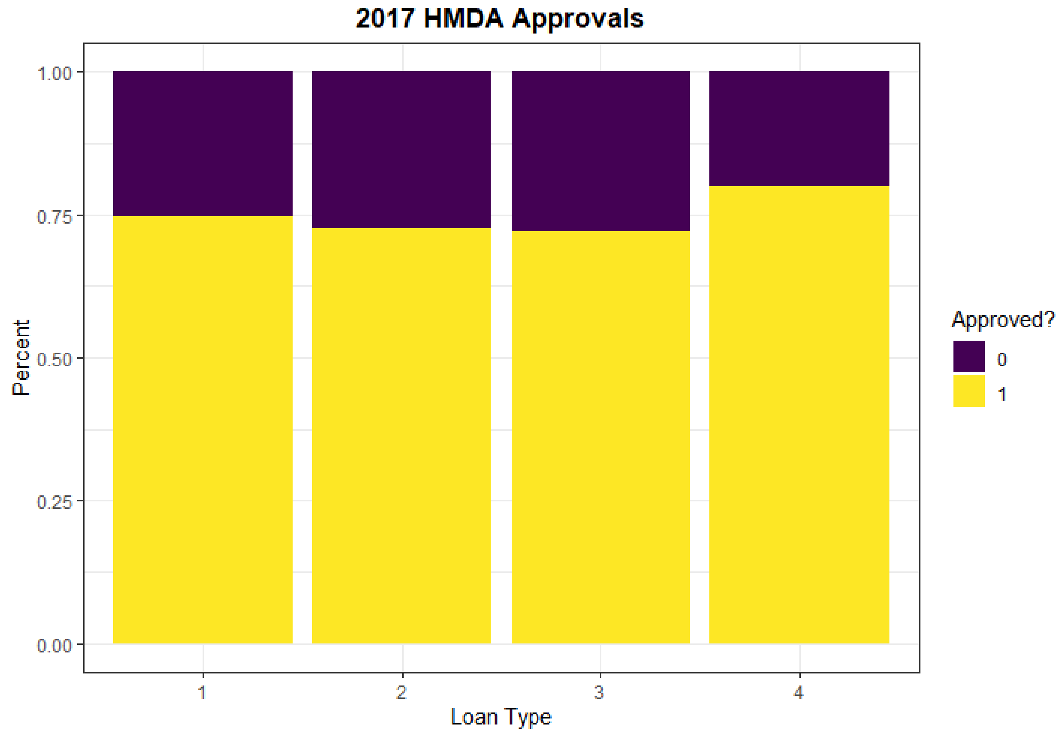
1 -- American Indian or Alaska Native
2 -- Asian
3 -- Black or African American
4 -- Native Hawaiian or Other Pacific Islander
5 -- White
6 -- Information not provided by applicant in mail, Internet, or telephone application
7 -- Not applicable
8 -- No co-applicant

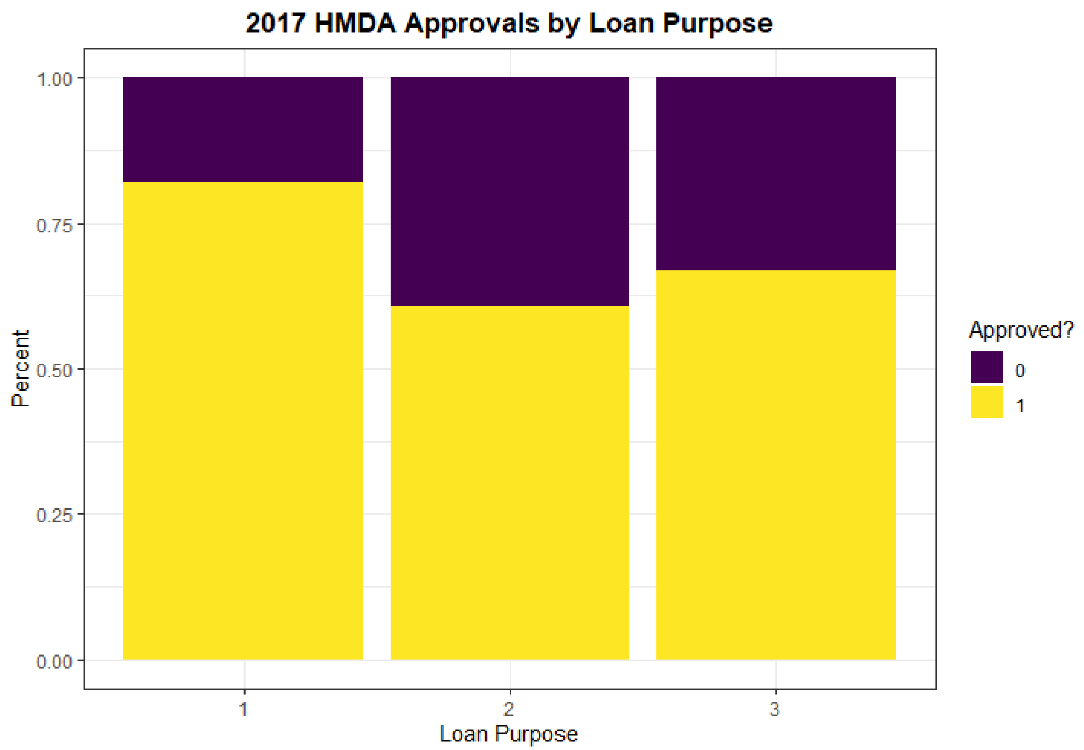
Data Exploration



Ethnicity:
1 -- Hispanic or Latino
2 -- Not Hispanic or Latino
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 -- Not applicable
5 -- No co-applicant

Data Exploration

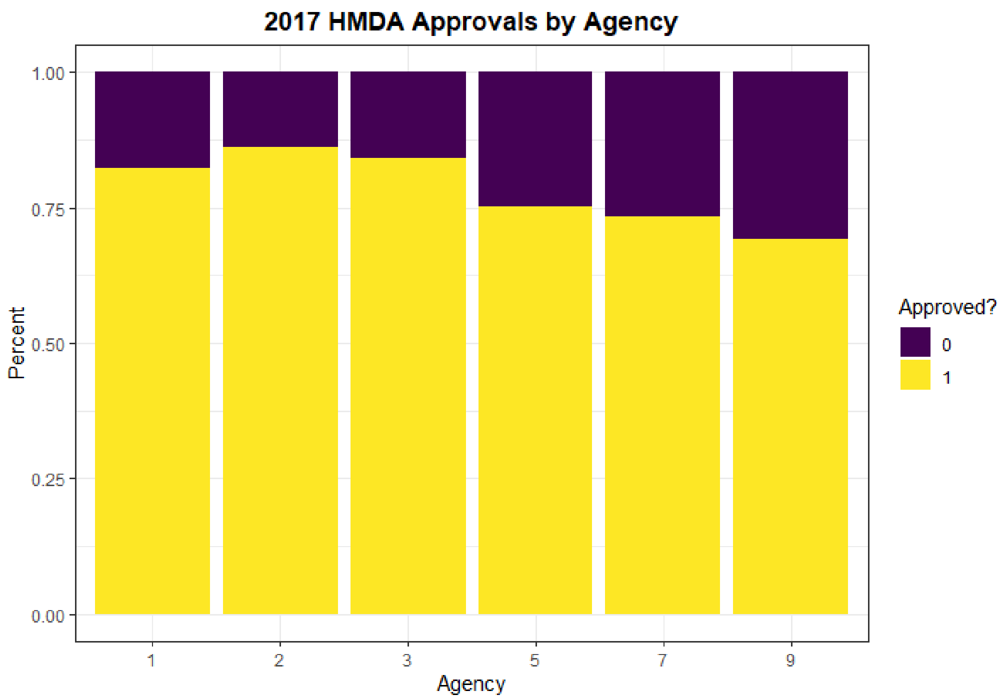




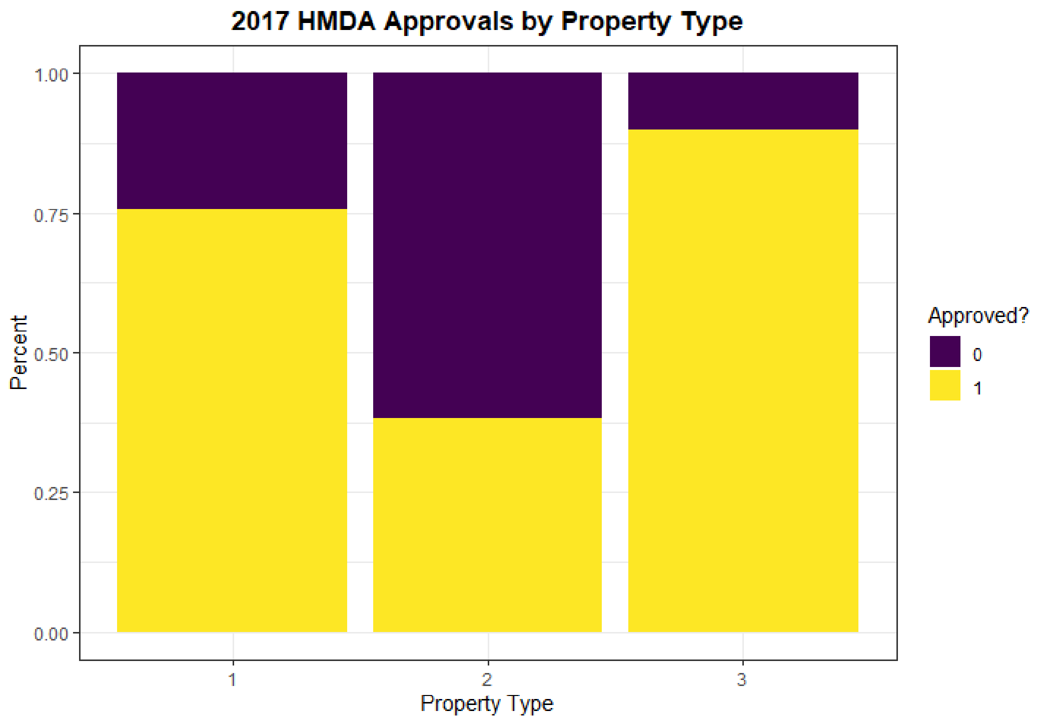
Loan Type:
1 -- Conventional (any loan other than FHA, VA, FSA, or RHS loans)
2 -- FHA-insured (Federal Housing Administration)
3 -- VA-guaranteed (Veterans Administration)
4 -- FSA/RHS (Farm Service Agency or Rural Housing Service)

Loan Purpose:
1 -- Home purchase
2 -- Home improvement
3 -- Refinancing

Data Exploration

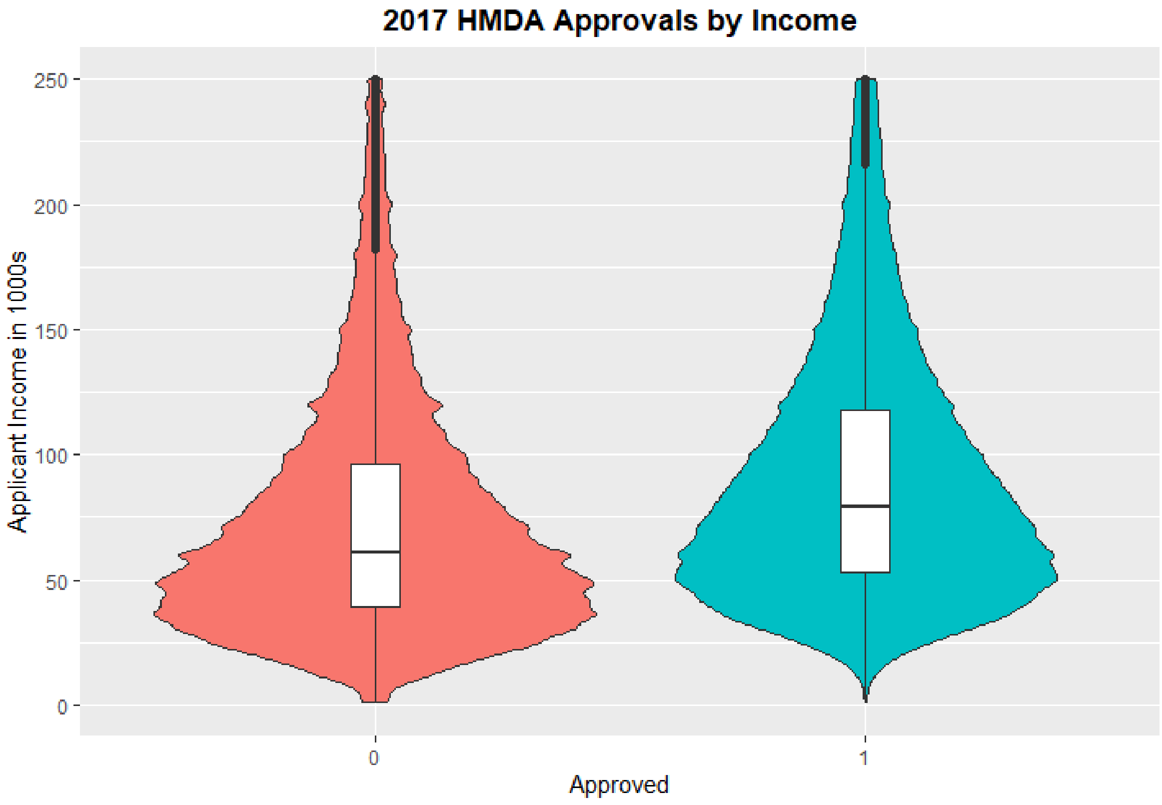


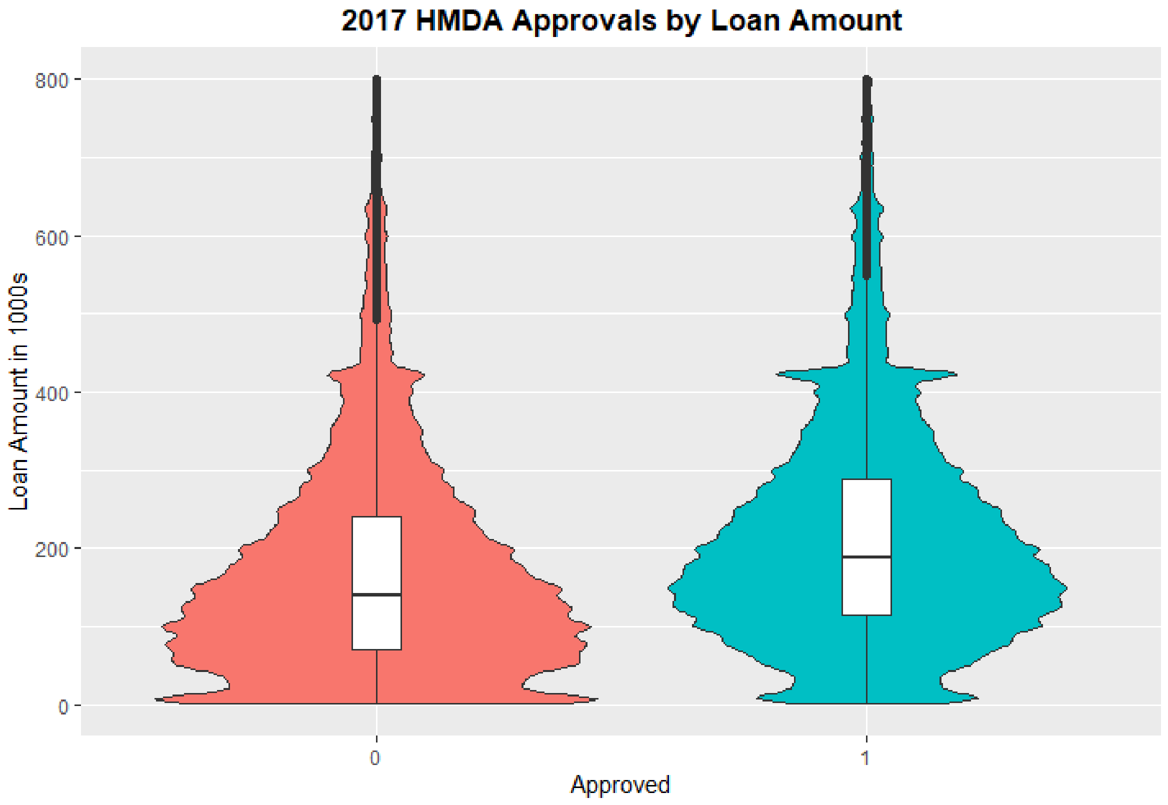
Agency:
1 -- Office of the Comptroller of the Currency (OCC)
2 -- Federal Reserve System (FRS)
3 -- Federal Deposit Insurance Corporation (FDIC)
5 -- National Credit Union Administration (NCUA)
7 -- Department of Housing and Urban Development (HUD)
9 -- Consumer Financial Protection Bureau (CFPB)



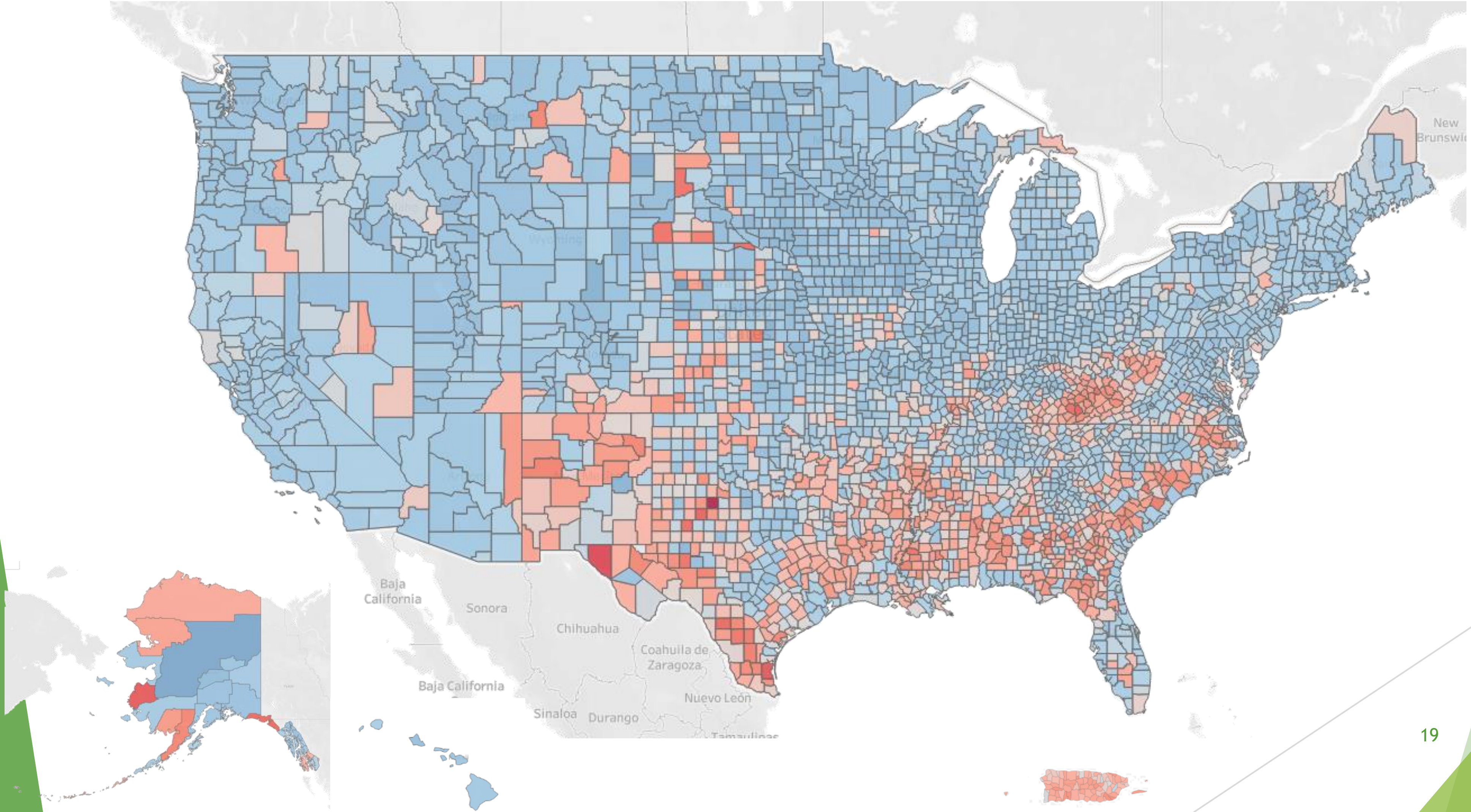
Property Type:
1 -- One to four-family (other than manufactured housing)
2 -- Manufactured housing
3 -- Multifamily

Data Exploration





Approval Rates per County/FIPS



Modeling / Machine Learning

Preparation for Modeling

- Many possible features to evaluate in machine learning models
- Can set categorical variables as factors
  - Already encoded as numbers in original CSV
- Denial Reason(s) removed
  - Cannot have a reason unless application was rejected
  - Even then, most rejections had no denial reason listed
- Some features had excessive amounts of NA values; remove them
- Some were almost always the same value; remove them

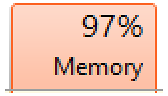
```
 rate_spread
Min.    :  2
1st Qu.:  2
Median :  2
Mean   :  2
3rd Qu.:  2
Max.   :100
NA's   :10009258
```

```
table(hmda_reduced$hoepa_status)

    1        2
 3538 10498993
```

Preparation for Modeling

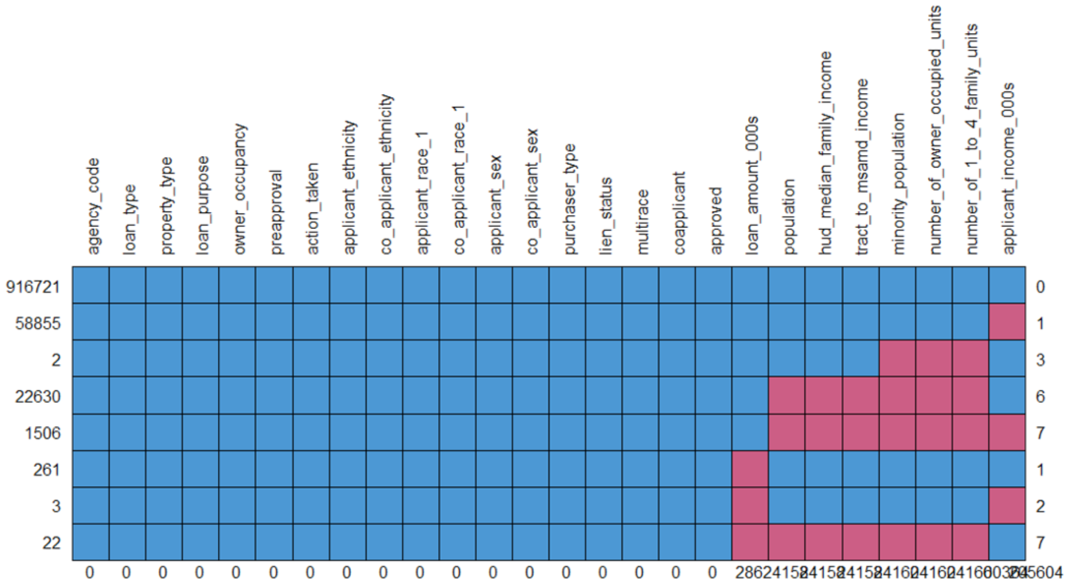
- After removing action 4+6, still 10,502,531 observations remain
- To reduce complexity, do not consider categorical features with a very large number of factors
  - Example: State/County (FIPS code)
  - Drop and consider numerical features based on location of applications only
    - Minority population %, Median family income, etc.

- Still, any attempt to run models on 10 million+ observations with available hardware was unreasonable
  - A single model could take many hours, even days to complete!



Preparation for Modeling

- Take only a subset of data for initial consideration, further reduce amount of observations when necessary due to hardware limitations
  - Start with ~10% of data (1 million observations)
- There are still multiple missing values
  - Use mice library to run imputation to fill in missing values
  - Set max iterations to 1 otherwise computation time would take too long



Preliminary Modeling

- A binary classification problem
  - True/False for whether an application for a mortgage was accepted
- Initial modeling in R
  - GLM Logistic regression on all features to get a sense of their significance
- For race, set “white” (5) as reference instead since it comprised most of the observations
- Expect some singularities since co-applicant features have value of “No co-applicant”
  - Vs. Feature of whether there is a co-applicant

```
Coefficients: (3 not defined because of singularities)
                             Estimate Std. Error z value Pr(>|z|)    
(Intercept)                 3.329e-01  3.821e-02   8.713  < 2e-16 ***
agency_code2               -2.447e-02  2.472e-02  -0.990 0.322246    
agency_code3                2.329e-02  1.853e-02   1.257 0.208791    
agency_code5                1.797e-01  1.778e-02  10.104  < 2e-16 ***
agency_code7               -1.506e+00  1.675e-02 -89.879  < 2e-16 ***
agency_code9               -6.729e-01  1.638e-02 -41.092  < 2e-16 ***
loan_type2                 -5.593e-01  1.189e-02 -47.030  < 2e-16 ***
loan_type3                 -7.040e-01  1.479e-02 -47.582  < 2e-16 ***
loan_type4                 -1.039e+00  3.710e-02 -28.001  < 2e-16 ***
property_type2             -7.449e-01  1.406e-02 -52.972  < 2e-16 ***
property_type3              2.673e-01  8.100e-02   3.300 0.000967 ***
loan_purpose2              -6.852e-01  1.408e-02 -48.660  < 2e-16 ***
loan_purpose3              -7.235e-01  8.614e-03 -83.994  < 2e-16 ***
owner_occupancy2            7.605e-02  1.113e-02   6.836 8.13e-12 ***
owner_occupancy3           -1.599e-01  7.048e-02  -2.268 0.023314 *  
loan_amount_000s           -2.359e-06  3.913e-06  -0.603 0.546604    
preapproval2                1.577e-01  1.940e-02   8.131 4.27e-16 ***
preapproval3                7.182e-02  1.728e-02   4.156 3.24e-05 ***
applicant_ethnicity2        2.354e-01  1.315e-02  17.900  < 2e-16 ***
applicant_ethnicity3        1.632e-01  2.234e-02   7.304 2.80e-13 ***
applicant_ethnicity4        1.755e-01  1.771e-01   0.991 0.321723    
co_applicant_ethnicity2     1.082e-01  1.990e-02   5.437 5.40e-08 ***
co_applicant_ethnicity3    -1.385e-02  3.293e-02  -0.421 0.674112    
co_applicant_ethnicity4     1.308e-01  2.801e-01   0.467 0.640479    
co_applicant_ethnicity5    -1.323e-01  2.157e-02  -6.134 8.60e-10 ***
applicant_race_11          -2.977e-01  3.662e-02  -8.130 4.30e-16 ***
applicant_race_12           4.590e-02  1.752e-02   2.620 0.008788 ** 
applicant_race_13          -4.449e-01  1.420e-02 -31.327  < 2e-16 ***
applicant_race_14          -2.453e-01  5.069e-02  -4.839 1.31e-06 ***
applicant_race_16          -1.794e-01  1.982e-02  -9.053  < 2e-16 ***
applicant_race_17           4.945e-01  2.501e-01   1.978 0.047965 *  
co_applicant_race_11       -2.786e-01  5.947e-02  -4.684 2.81e-06 ***
co_applicant_race_12       -1.504e-01  2.532e-02  -5.940 2.85e-09 ***
co_applicant_race_13       -1.675e-01  2.664e-02  -6.288 3.21e-10 ***
co_applicant_race_14       -2.119e-01  7.118e-02  -2.977 0.002908 ** 
co_applicant_race_16        1.174e-02  2.975e-02   0.395 0.693121    
co_applicant_race_17        3.462e-01  3.799e-01   0.911 0.362086    
co_applicant_race_18               NA         NA      NA       NA    
```

```
applicant_sex2                -8.191e-02  8.541e-03  -9.590  < 2e-16 ***
applicant_sex3                -2.694e-01  2.226e-02 -12.104  < 2e-16 ***
applicant_sex4                 1.584e+00  2.027e-01   7.812 5.61e-15 ***
co_applicant_sex2              9.975e-02  1.417e-02   7.040 1.92e-12 ***
co_applicant_sex3              1.831e-01  3.348e-02   5.469 4.53e-08 ***
co_applicant_sex4             -7.257e-02  2.914e-01  -0.249 0.803314    
co_applicant_sex5                     NA         NA      NA       NA    
applicant_income_000s          8.023e-05  1.092e-05   7.345 2.06e-13 ***
purchaser_type1                2.092e+01  4.639e+01   0.451 0.651957    
purchaser_type2                2.186e+01  5.609e+01   0.390 0.696684    
purchaser_type3                2.087e+01  5.761e+01   0.362 0.717218    
purchaser_type4                2.100e+01  1.544e+03   0.014 0.989150    
purchaser_type5                2.126e+01  2.520e+02   0.084 0.932756    
purchaser_type6                2.094e+01  6.295e+01   0.333 0.739345    
purchaser_type7                2.123e+01  5.933e+01   0.358 0.720430    
purchaser_type8                2.076e+01  1.923e+02   0.108 0.914031    
purchaser_type9                2.106e+01  8.548e+01   0.246 0.805402    
lien_status2                   2.495e-01  1.554e-02  16.056  < 2e-16 ***
lien_status3                   1.441e-01  1.670e-02   8.627  < 2e-16 ***
population                    -6.515e-05  2.469e-06 -26.382  < 2e-16 ***
minority_population           -2.038e-03  1.701e-04 -11.987  < 2e-16 ***
hud_median_family_income       6.743e-06  2.214e-07  30.452  < 2e-16 ***
tract_to_msamd_income          3.952e-03  8.813e-05  44.845  < 2e-16 ***
number_of_owner_occupied_units 7.430e-05  1.053e-05   7.053 1.75e-12 ***
number_of_1_to_4_family_units  4.153e-05  6.753e-06   6.150 7.75e-10 ***
multirace1                    -1.588e-01  4.229e-02  -3.756 0.000173 ***
coapplicant1                          NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1143851  on 999999  degrees of freedom
Residual deviance:  561813  on 999938  degrees of freedom
AIC: 561937

Number of Fisher Scoring iterations: 19
```

Preliminary Modeling Observations

- Conventional loans had higher acceptance than other types of loans
- Loan amount had a negative coefficient but was not significant
- Higher applicant income was significantly positive
- Ethnicity: not Hispanic/Latino was significantly positive
- Minority races were all significantly negative compared to White
  - Except Asian for the primary applicant
- Listing more than one race was significantly negative

Feature Selection

- Try other models with the lesser features from the initial model removed
  - AIC typically increased
  - Example: removing loan amount, purchaser type, preapproval, and all co-applicant fields (while retaining coapplicant? feature) actually almost doubled the AIC
- In a perfect situation, could employ stepAIC to assist in obtaining a model with an ideal number of retained features
  - Unfortunately, hardware limitations left stepAIC running for a very long time with no iterations being finished

Caret Modeling

- Model data using caret library
- Perform 5-fold cross-validation to save computation time
- Run train with logistic regression first
- Run “rpart” tree model and compare results
  - Runtime already starting to get very lengthy
- Decent accuracy, but poor Kappa

```
Accuracy
          Min.  1st Qu.   Median     Mean   3rd Qu.      Max. NA's
log  0.7593800 0.759625 0.759835 0.759849 0.7599212 0.7604838    0
tree 0.7588012 0.760070 0.762400 0.761362 0.7626888 0.7628500    0

Kappa
          Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log  0.1937740 0.1954350 0.1958226 0.1961131 0.1971580 0.1983759    0
tree 0.1701033 0.1748366 0.2080637 0.1946368 0.2100327 0.2101475    0
```

KNN Classifier

- Attempting to model with the 1 million dataset was not possible within a reasonable amount of time
- Further take a random sample of 100,000 rows in order to run some additional models in order to get an idea of how a potential final model would perform
  - Still perform 5-fold cross-validation
  - Also compare with glm
- K=9 in selected model
- For further refinement, scale the features before attempting improvements

```
Accuracy
         Min.   1st Qu.    Median    Mean   3rd Qu.      Max. NA's
log 0.7567378 0.7570000 0.7606120 0.75929 0.7607380 0.7613619    0
knn 0.7234500 0.7236638 0.7238862 0.72475 0.7260137 0.7267363    0

Kappa
         Min.    1st Qu.    Median       Mean    3rd Qu.       Max. NA's
log 0.1867703 0.18717408 0.1937972 0.19407420 0.20079777 0.20183168    0
knn 0.0642095 0.06762325 0.0700631 0.06962949 0.07230683 0.07394474    0
```

Random Forest

- Keep using 100k dataset
  - Still perform 5-fold cross-validation; compare with glm
- Already took ~1 hour to run on just 100k rows
- mtry = 20

```
Accuracy
         Min.   1st Qu.    Median    Mean   3rd Qu.      Max. NA's
log 0.7567378 0.7570000 0.7606120 0.75929 0.7607380 0.7613619    0
rf  0.7720000 0.7720386 0.7731113 0.77340 0.7747613 0.7750888    0

Kappa
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log 0.1867703 0.1871741 0.1937972 0.1940742 0.2007978 0.2018317    0
rf  0.2888405 0.2904534 0.2939125 0.2952974 0.3011968 0.3020837    0
```

```
> confusionMatrix.train(log_fit)
Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  5.0  3.2
         1 20.9 70.9

 Accuracy (average) : 0.7593
```

```
> confusionMatrix.train(rf_fit)
Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  8.1  4.9
         1 17.8 69.2

 Accuracy (average) : 0.7734
```

XGBoost

- Keep using 100k dataset
  - Still perform 5-fold cross-validation; compare with glm
- Much faster compared to random forest with comparable results

```
Accuracy
         Min.  1st Qu.    Median    Mean   3rd Qu.      Max. NA's
log 0.7567378 0.75700 0.7606120 0.75929 0.7607380 0.7613619    0
xgb 0.7761888 0.77645 0.7767612 0.77721 0.7777389 0.7789111    0

Kappa
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log 0.1867703 0.1871741 0.1937972 0.1940742 0.2007978 0.2018317    0
xgb 0.2906202 0.2956944 0.2993302 0.2984985 0.2996688 0.3071789    0
```

```
> confusionMatrix.train(xgb_fit)
Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  7.9  4.3
         1 18.0 69.8

 Accuracy (average) : 0.7772
```

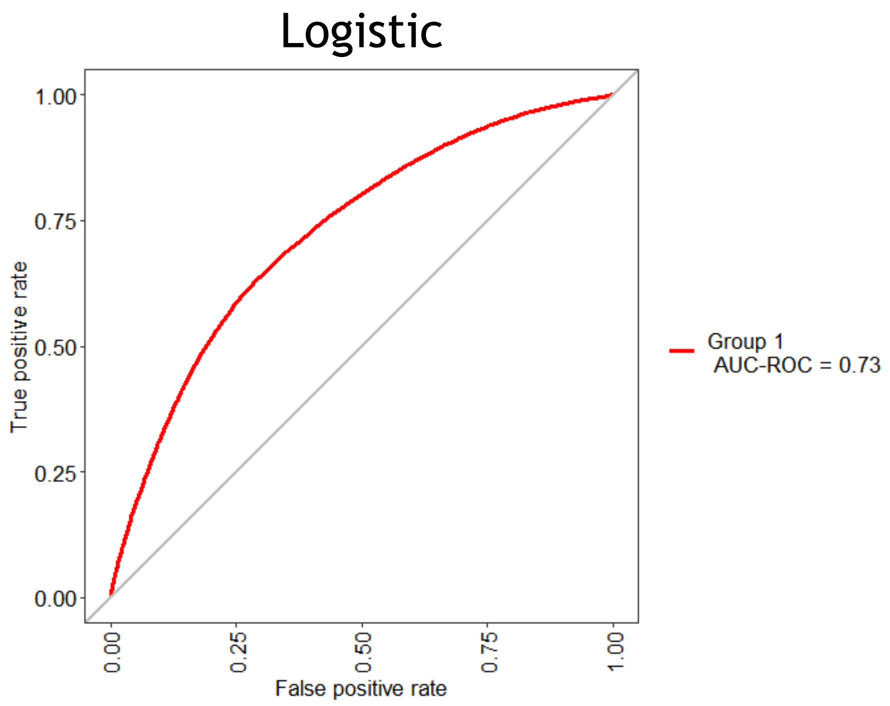
```
> confusionMatrix.train(rf_fit)
Cross-Validated (5 fold) Confusion Matrix

(entries are percentual average cell counts across resamples)

          Reference
Prediction    0    1
         0  8.1  4.9
         1 17.8 69.2

 Accuracy (average) : 0.7734
```

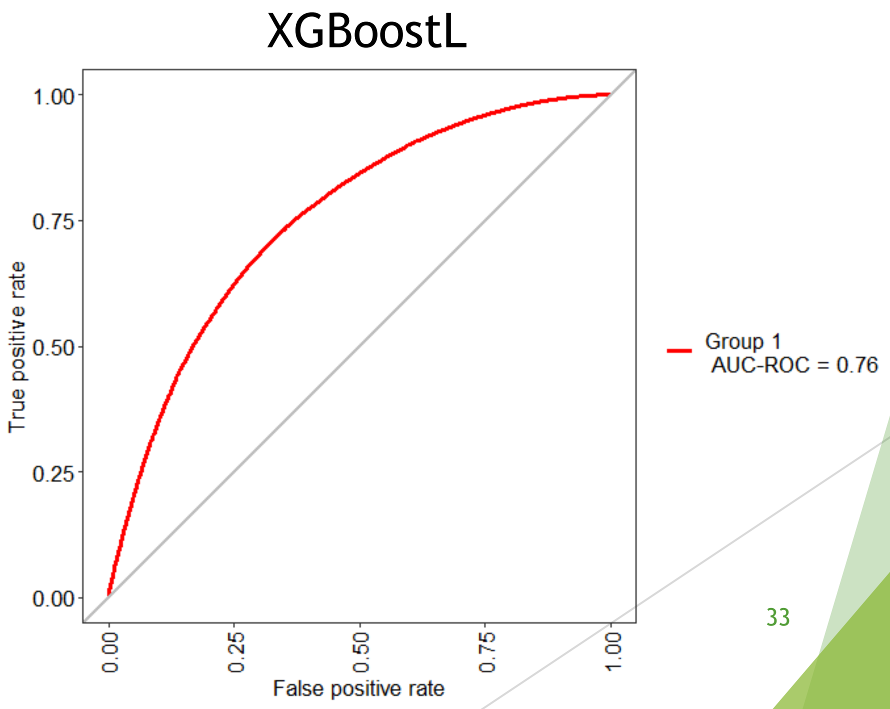
Classification Models

- All models had approximately the same accuracy
- Kappa noticeably improved for Random Forest and XGBoost Linear
  - Random Forest model hampered by much longer runtime in R
- Other models like SVM had extremely long runtimes that did not complete within a reasonable time frame
- Take GLM and XGB models as references to explore possible improvements
  - StepAIC with simplified GLM model, while able to complete even with limited iterations, did not produce any improvement in AIC

Classifier Testing

- Generate some AUC-ROC curves for GLM and XGBoost models
  - Use MLeval library
  - Use Data frames instead of Data Tables, rename levels to valid R names





Imbalanced Learning

- Data: roughly 75-25 split for Accepted/Rejected
- Consider using imbalanced learning techniques with Logistic Regression and XGBoost
- Down-sampling
  - Use fewer “Accepted” rows to match “Rejected”
- Up-sampling
  - Use more “Rejected” rows to match “Accepted”
- Improved Kappa at the cost of Accuracy

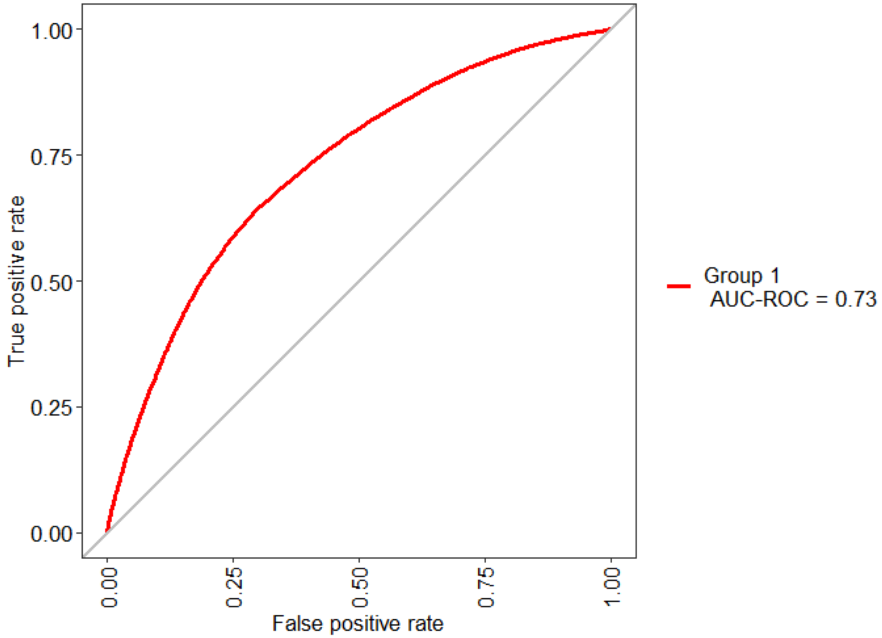
```
Accuracy
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log_d 0.6597000 0.6604330 0.6627331 0.6630999 0.6662667 0.6663667    0
log_u 0.6590500 0.6612331 0.6622831 0.6629999 0.6647668 0.6676666    0
xgb_d 0.6933653 0.6947653 0.6949000 0.6953401 0.6961848 0.6974849    0
xgb_u 0.7108500 0.7122356 0.7125644 0.7141700 0.7173359 0.7178641    0

Kappa
           Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log_d 0.2724874 0.2760515 0.2810292 0.2806956 0.2863706 0.2875394    0
log_u 0.2716611 0.2758211 0.2806246 0.2796564 0.2833849 0.2867902    0
xgb_d 0.3171845 0.3176508 0.3219302 0.3215193 0.3230611 0.3277697    0
xgb_u 0.3169378 0.3242647 0.3249225 0.3280799 0.3353678 0.3389069    0
```

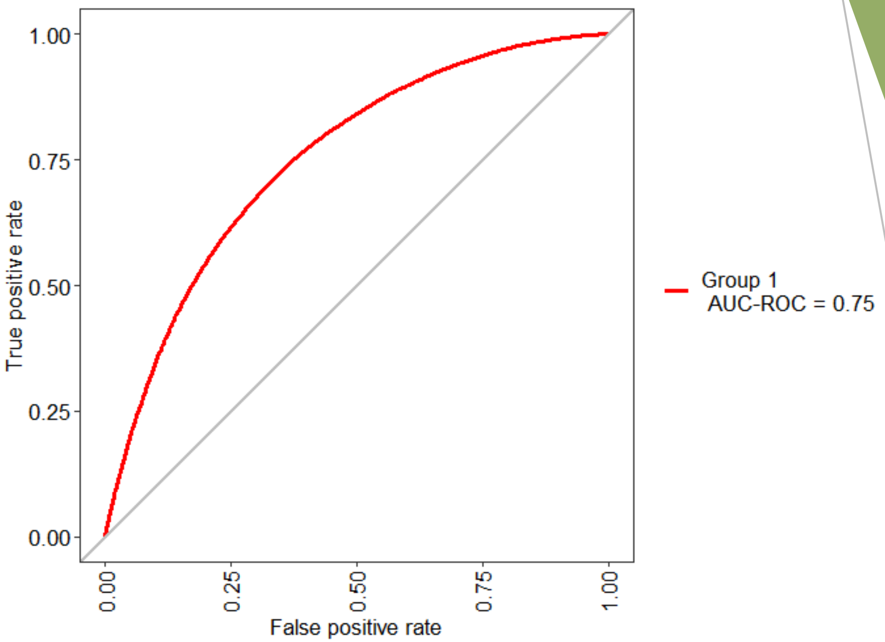
```
Accuracy
         Min. 1st Qu.    Median    Mean   3rd Qu.      Max. NA's
log 0.7567378 0.75700 0.7606120 0.75929 0.7607380 0.7613619    0
xgb 0.7761888 0.77645 0.7767612 0.77721 0.7777389 0.7789111    0

Kappa
         Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
log 0.1867703 0.1871741 0.1937972 0.1940742 0.2007978 0.2018317    0
xgb 0.2906202 0.2956944 0.2993302 0.2984985 0.2996688 0.3071789    0
```

Logistic

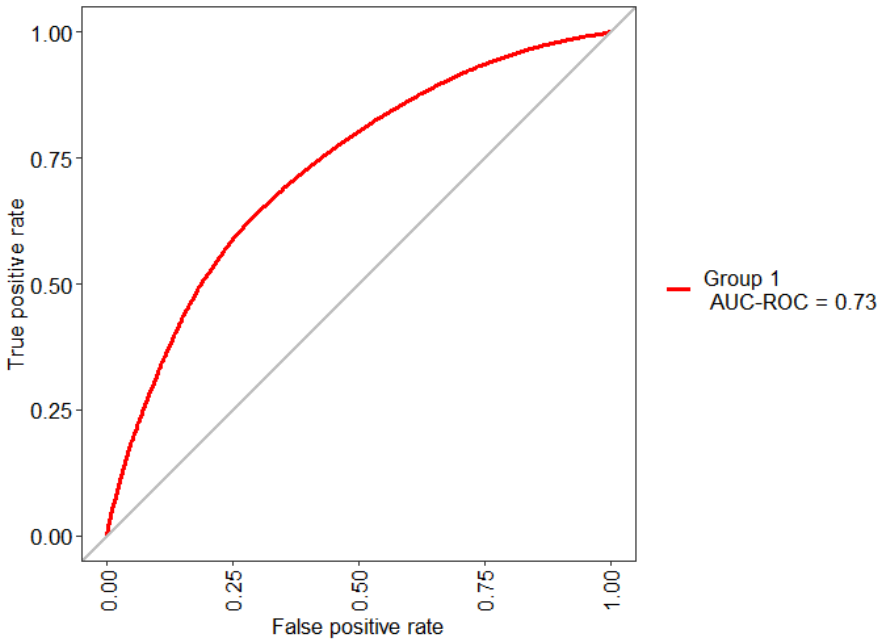
XGBoostL

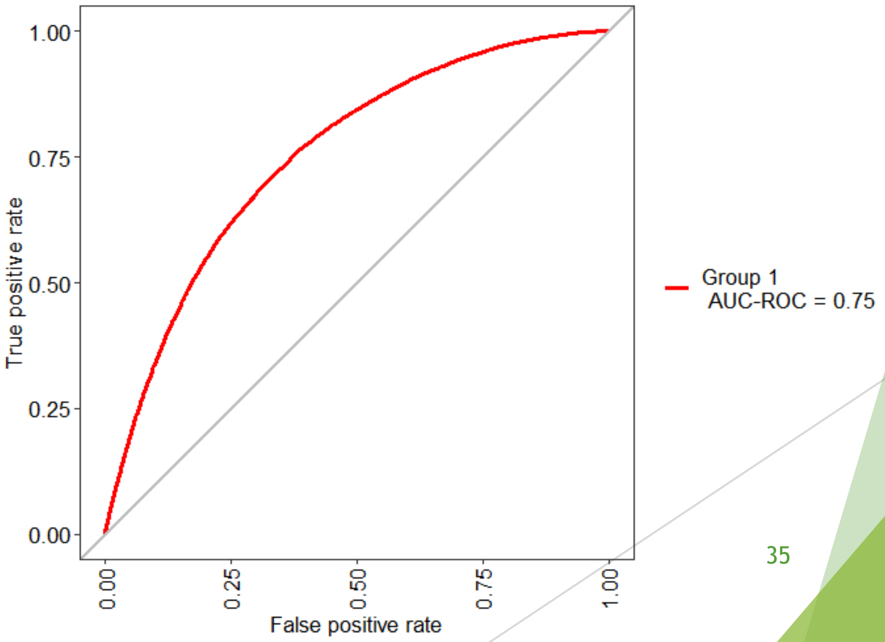
Down





Up





Spark

- Try to use Spark libraries in R to process big data
- Time how long it takes to model a caret train and compare
  - Typical 10-15 seconds for a simple logistic regression over the 100k dataset
  - Include 5-fold cross-validation
- Time how long it takes to model in Spark
  - Must use spark-specific model fitting functions
    - Can't use caret
  - Also try to include 5-fold cross validation
  - Unfortunately, modeling would run indefinitely before freezing on my machine
  - Advantage is copying the data into Spark would save on some RAM usage

Recommendations for Further Study and Takeaways

Future Recommendations

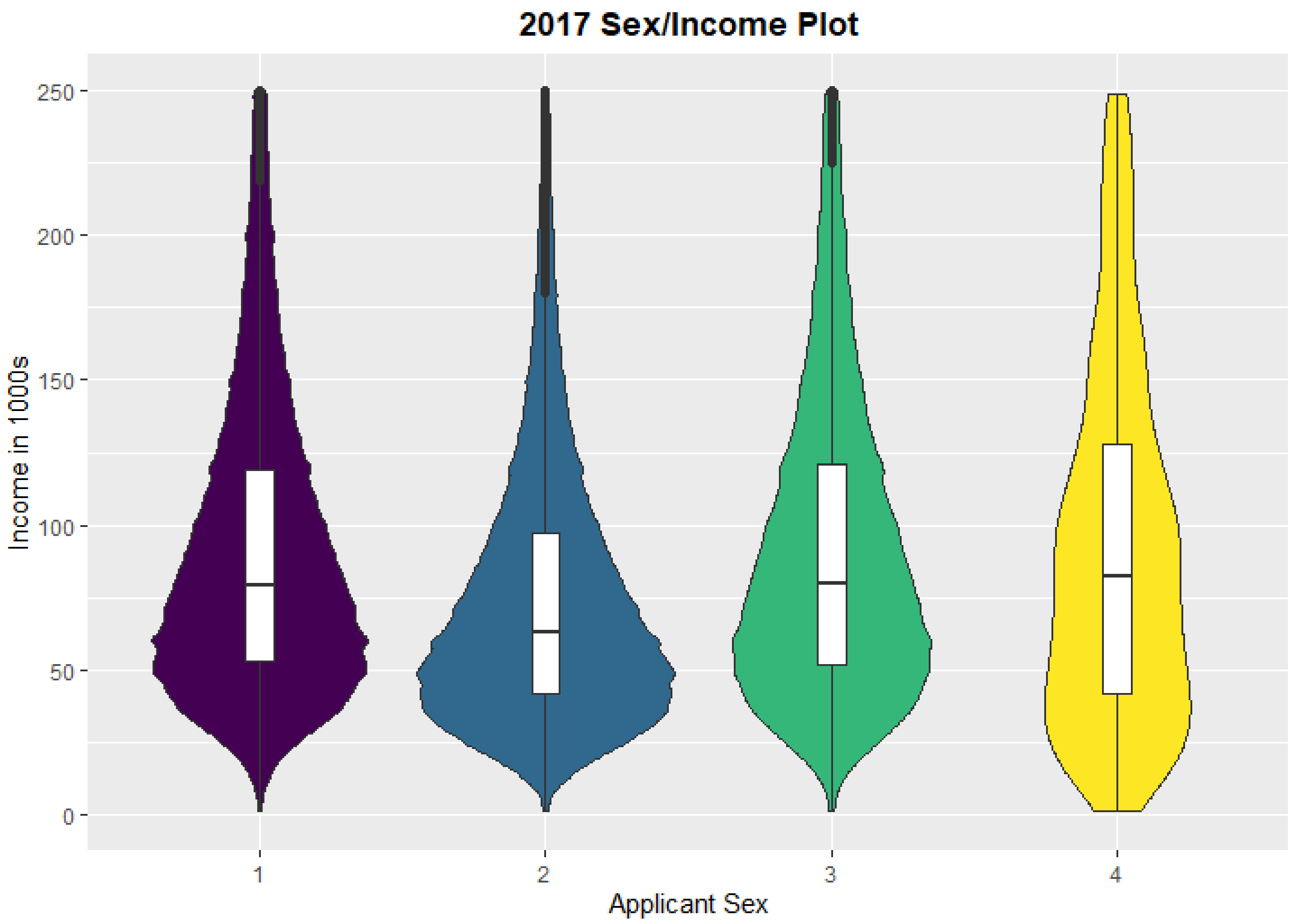
- Consider more modeling algorithms
  - Tune a wider variety of parameters
- Change dependent variable
  - Multiclass classification using Action Taken instead of Approved
- Include analysis of multiple years of data
  - Download and import multiple years of HMDA data
  - Panel Data / Predictive Analytics
- Better hardware required
  - Faster processing for certain machine learning models
  - More memory to handle extremely large datasets

Conclusions and Takeaways

- Models can predict with ~75% accuracy whether a mortgage application will be approved
- XGBoost had the best performance when taking into account processing time in addition to scoring metrics
- Is there really bias in mortgage approvals?

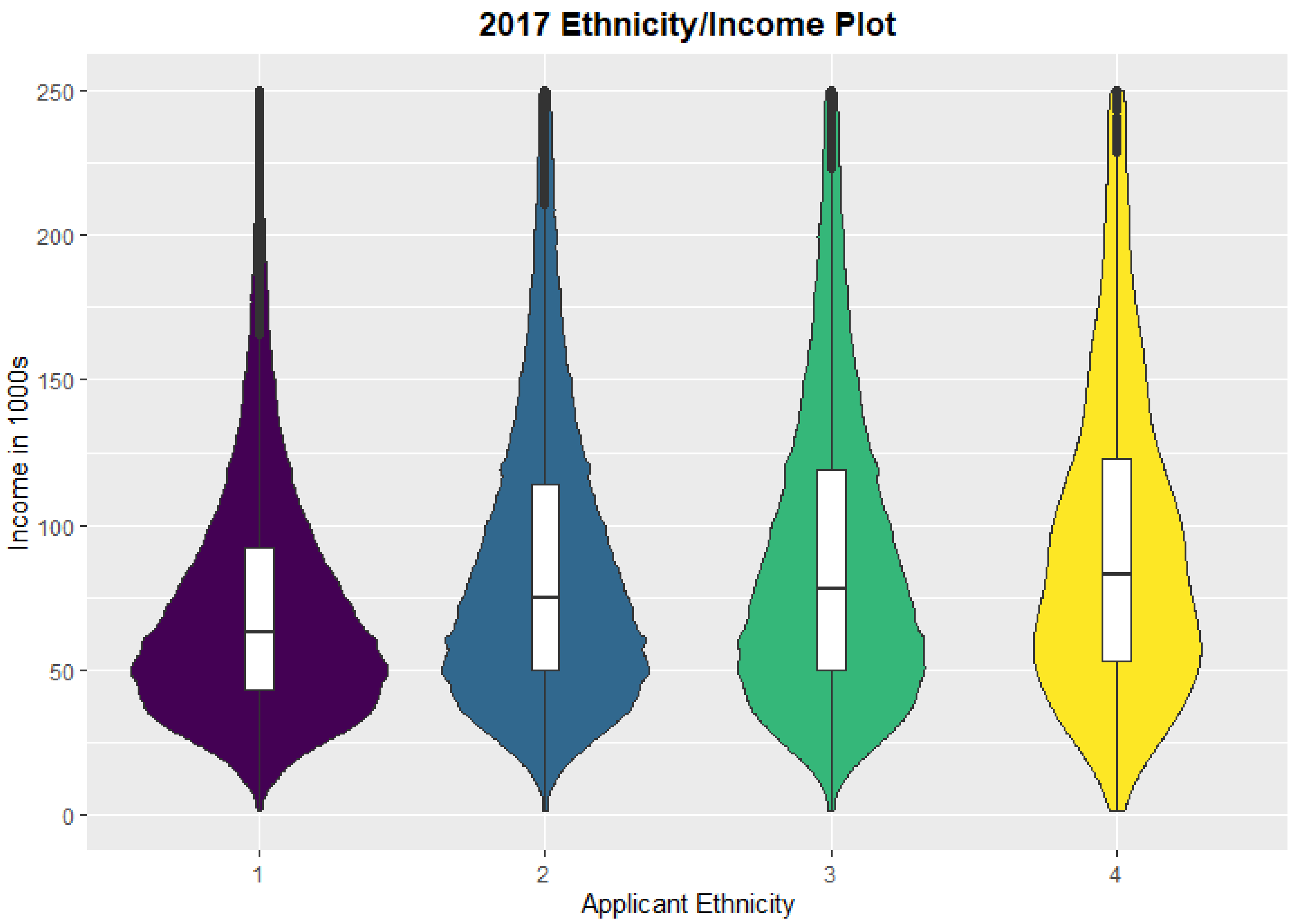
applicant_ethnicity2	2.354e-01	1.315e-02	17.900	< 2e-16	***
applicant_ethnicity3	1.632e-01	2.234e-02	7.304	2.80e-13	***
applicant_ethnicity4	1.755e-01	1.771e-01	0.991	0.321723	
co_applicant_ethnicity2	1.082e-01	1.990e-02	5.437	5.40e-08	***
co_applicant_ethnicity3	-1.385e-02	3.293e-02	-0.421	0.674112	
co_applicant_ethnicity4	1.308e-01	2.801e-01	0.467	0.640479	
co_applicant_ethnicity5	-1.323e-01	2.157e-02	-6.134	8.60e-10	***
applicant_race_11	-2.977e-01	3.662e-02	-8.130	4.30e-16	***
applicant_race_12	4.590e-02	1.752e-02	2.620	0.008788	**
applicant_race_13	-4.449e-01	1.420e-02	-31.327	< 2e-16	***
applicant_race_14	-2.453e-01	5.069e-02	-4.839	1.31e-06	***
applicant_race_16	-1.794e-01	1.982e-02	-9.053	< 2e-16	***
applicant_race_17	4.945e-01	2.501e-01	1.978	0.047965	*
co_applicant_race_11	-2.786e-01	5.947e-02	-4.684	2.81e-06	***
co_applicant_race_12	-1.504e-01	2.532e-02	-5.940	2.85e-09	***
co_applicant_race_13	-1.675e-01	2.664e-02	-6.288	3.21e-10	***
co_applicant_race_14	-2.119e-01	7.118e-02	-2.977	0.002908	**
co_applicant_race_16	1.174e-02	2.975e-02	0.395	0.693121	
co_applicant_race_17	3.462e-01	3.799e-01	0.911	0.362086	
applicant_sex2	-8.191e-02	8.541e-03	-9.590	< 2e-16	***
applicant_sex3	-2.694e-01	2.226e-02	-12.104	< 2e-16	***
applicant_sex4	1.584e+00	2.027e-01	7.812	5.61e-15	***
co_applicant_sex2	9.975e-02	1.417e-02	7.040	1.92e-12	***
co_applicant_sex3	1.831e-01	3.348e-02	5.469	4.53e-08	***
co_applicant_sex4	-7.257e-02	2.914e-01	-0.249	0.803314	

Bias?



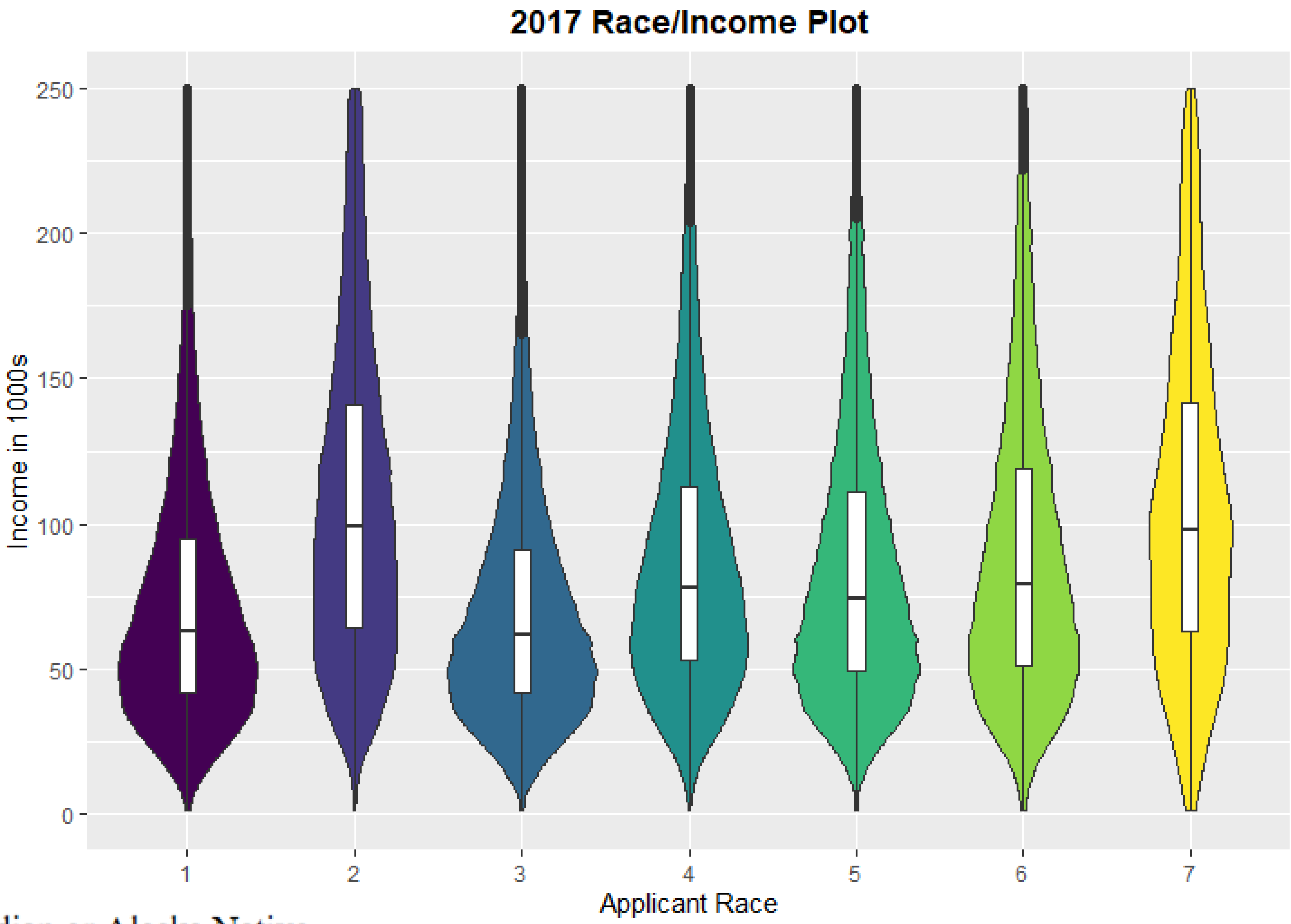
Sex:
1 -- Male
2 -- Female
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 -- Not applicable
5 -- No co-applicant

Bias?



Ethnicity:
1 -- Hispanic or Latino
2 -- Not Hispanic or Latino
3 -- Information not provided by applicant in mail, Internet, or telephone application
4 -- Not applicable

Bias?



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

Conclusion

- Ethnicity/race/sex were significant in modeling
- However, should consider socioeconomic factors before drawing any conclusions
  - As shown in previous slides, median applicant income varied between the different races, ethnicities, and sexes

- Data source: https://www.consumerfinance.gov/data-research/hmda/historic-data/