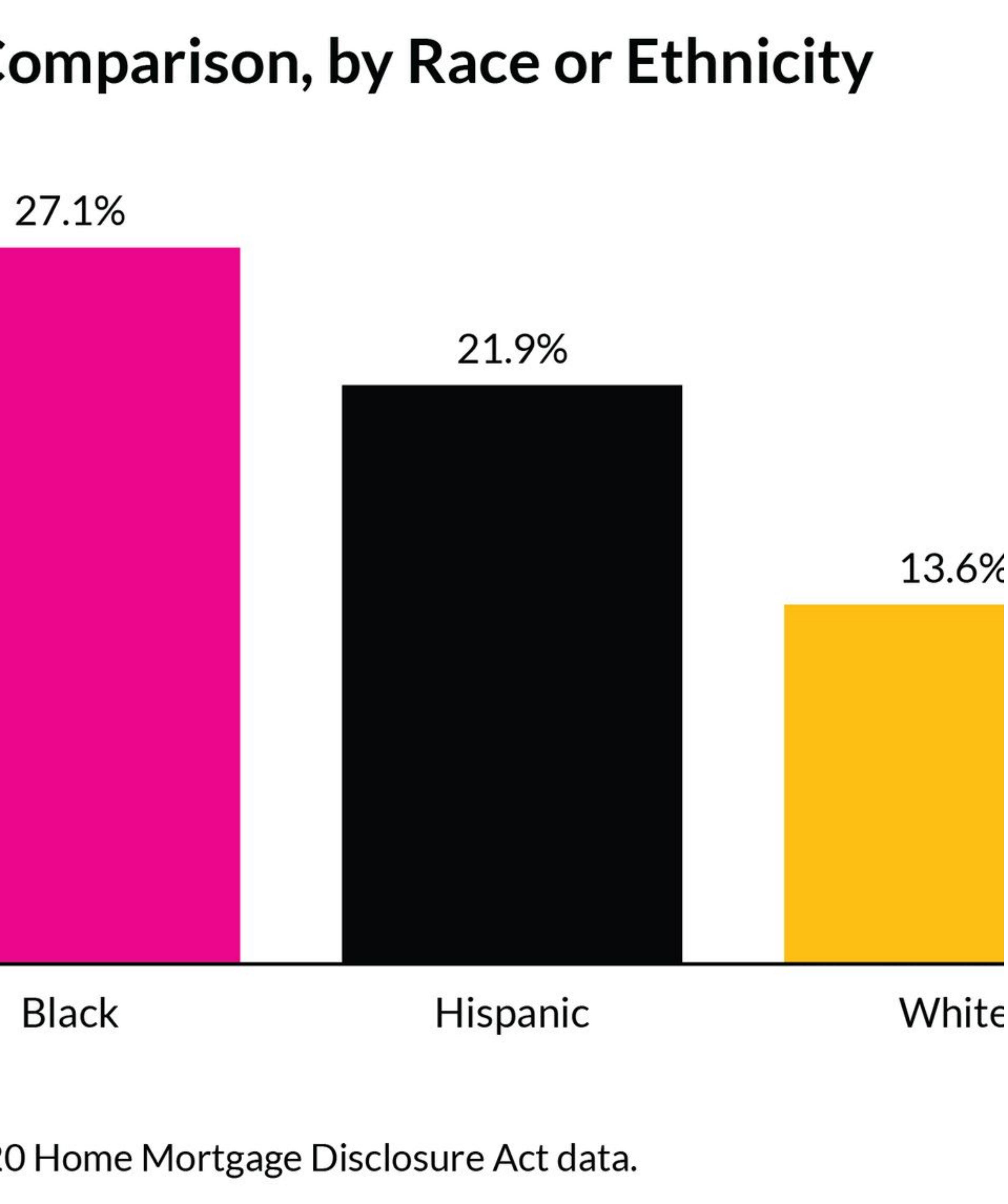
Mortgage Approval Disparities: Region, Race, and Reliability

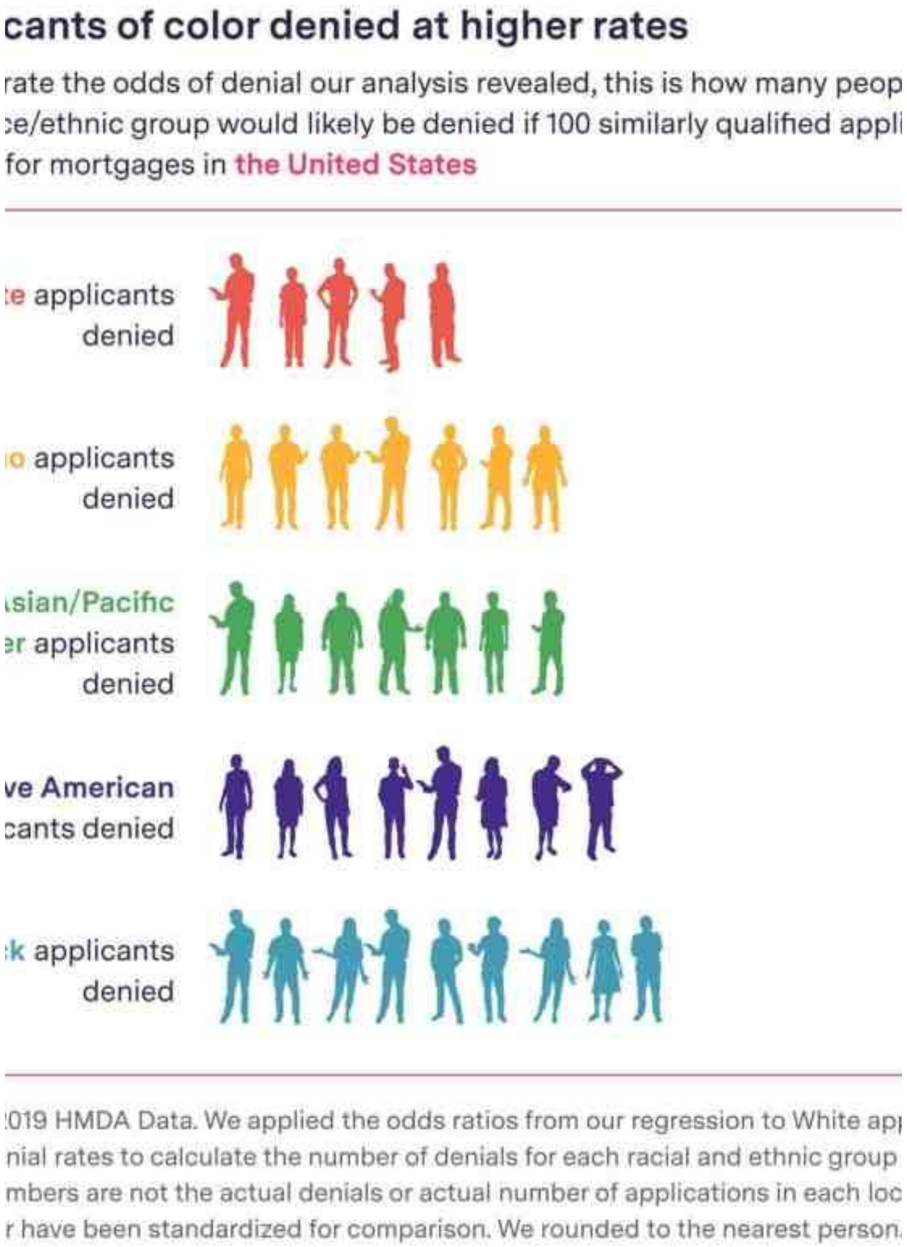
By Joseph Rodriguez


Comparison, by Race or Ethnicity

20 Home Mortgage Disclosure Act data.

Do Race or Region Affect Mortgage Approval Odds?

The goal of my project was to investigate whether race or region affect mortgage approval odds in the United States, even when controlling for financial factors. Through a combination of data analysis, logistic regression modeling, and a custom Monte Carlo simulation, I planed to uncover potential disparities.



cants of color denied at higher rates

rate the odds of denial our analysis revealed, this is how many peop
ce/ethnic group would likely be denied if 100 similarly qualified appli
for mortgages in the United States

019 HMDA Data. We applied the odds ratios from our regression to White ap
nial rates to calculate the number of denials for each racial and ethnic group
mbers are not the actual denials or actual number of applications in each loc
r have been standardized for comparison. We rounded to the nearest person.

Introduction – Why This Matters

Research question

If two people have the same income and loan request, but are of different races or live in different states — do they have the same odds of getting approved for a mortgage?

Goal

Investigate fairness in mortgage approval using data

Focus regions

California (West), New York (East), and Georgia (South)

Data Wrangling

- **Source: 2023 HMDA Mortgage Data**

  Obtained over 3 million rows of mortgage application data from the 2023 Home Mortgage Disclosure Act (HMDA) dataset.

- **Filtered to 3 States**

  Filtered the data to focus on mortgage applications from California (West), New York (East), and Georgia (South).

- **Selected Key Variables**

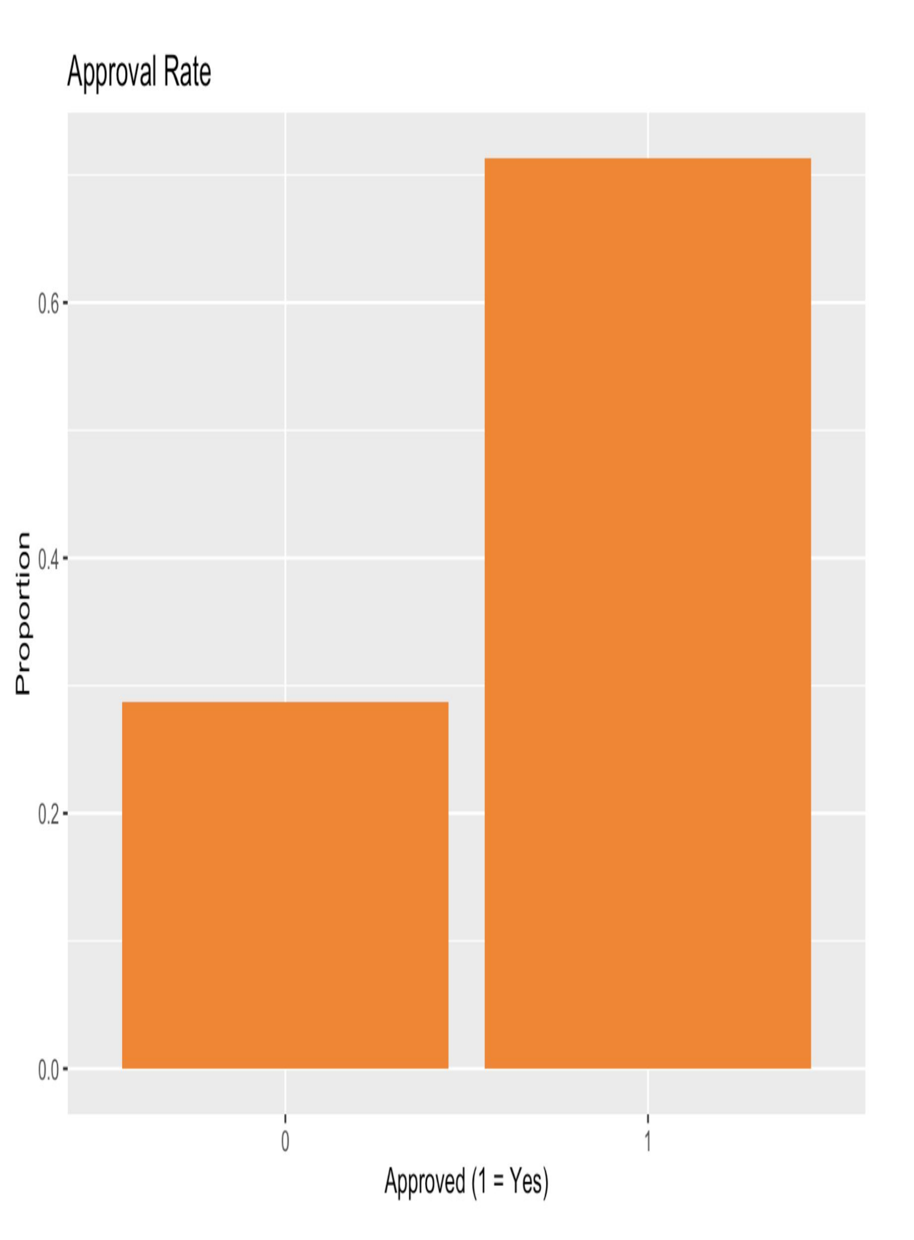
  Included variables such as loan amount, income, race, ethnicity, sex, region, and neighborhood demographics (e.g., percent minority population).

- **Transformed Variables**

  Applied log transformation to loan amount and income to address their highly skewed distributions.

- **Dropped Missing Data**

  Removed rows with missing data to ensure a cleaner dataset for modeling.

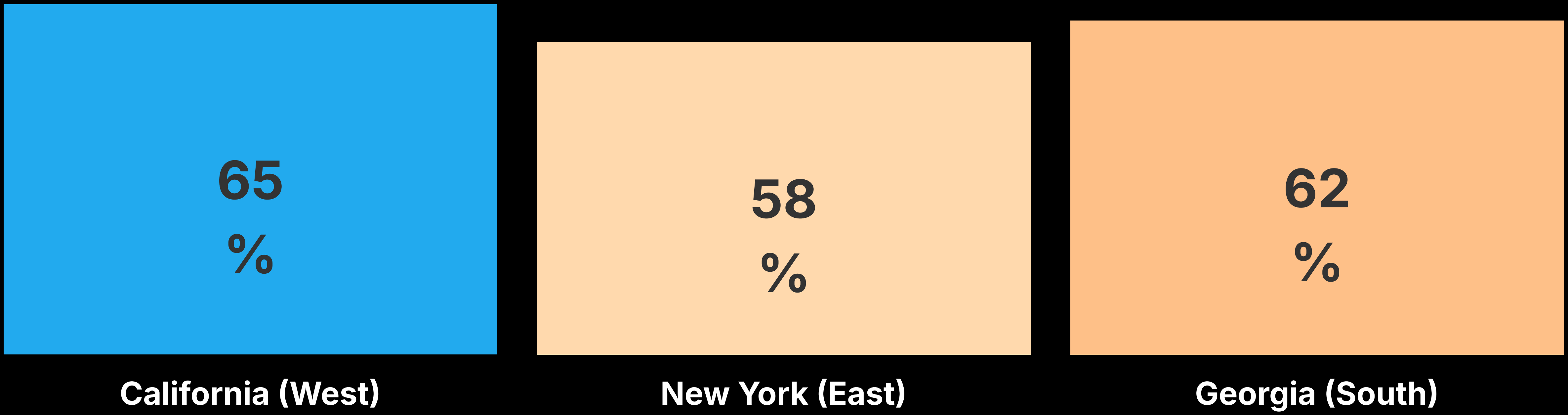


How Was the Data Distributed?

The exploratory data analysis (EDA) revealed the distributions of key variables in the mortgage dataset. The log-transformed loan amount and income both showed long-tailed distributions, with some applicants having extremely high values. Additionally, the overall mortgage approval rate was calculated to be around 60%, providing important baseline context before diving into the modeling.

Do Race and Region Affect Approval?

Approval rates (%) in each of the three regions



Modeling Approval Odds: Who Gets In?

Logistic Regression Model

Built a logistic regression model to predict mortgage approval (yes/no) as the outcome variable.

Predictors

The model included the following predictors: log(income), log(loan amount), region (West, East, South), race (White, Black, Other), and percent minority population in the neighborhood.

Goal

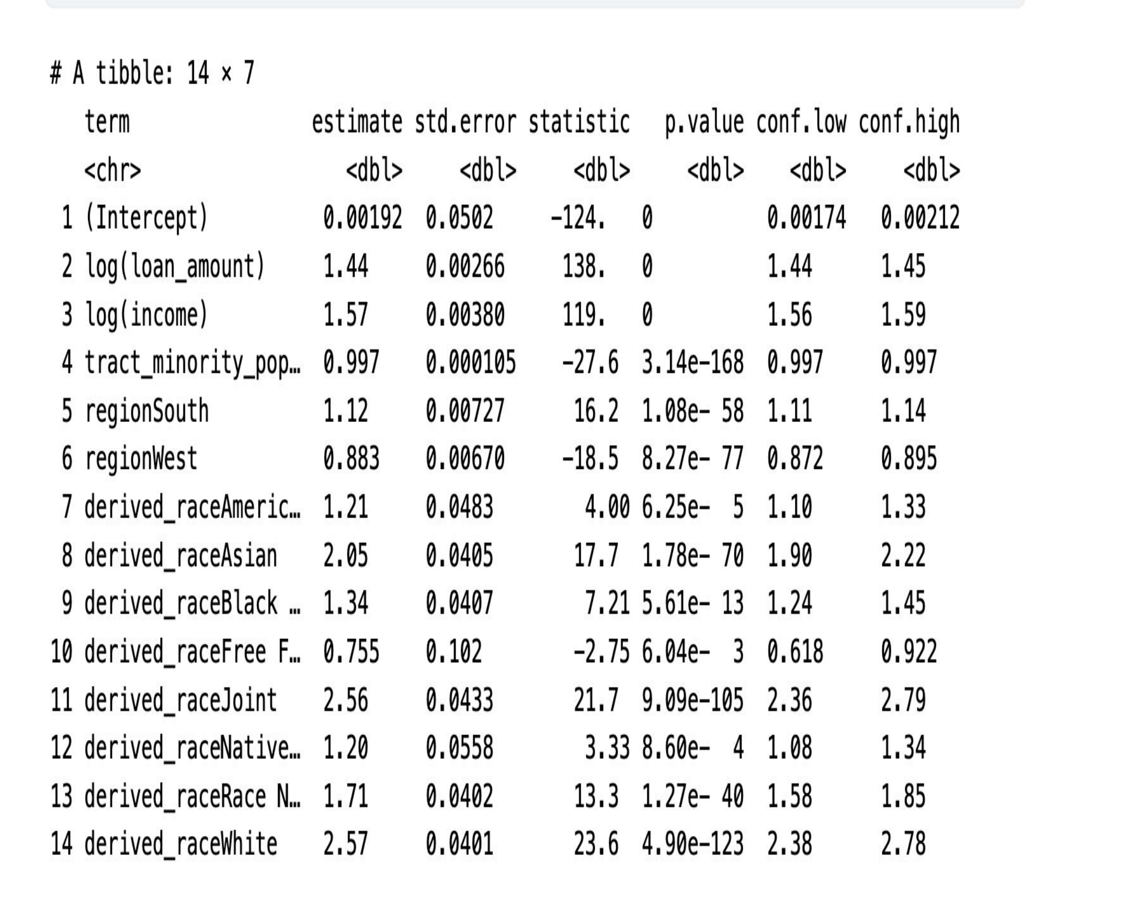
The goal was to see which variables were significant predictors of mortgage approval, even after controlling for financial factors like income and loan amount.

Justification

Logistic regression was chosen because the outcome variable (approved/not approved) is binary in nature.

Modeling Results

Even after controlling for income and loan size, race and region still made a difference. White applicants had about 2.5 times the odds of getting approved compared to others, which means those disparities didn’t go away just because the financials were the same.



```
# A tibble: 14 × 7
   term                estimate std.error statistic   p.value conf.low conf.high
   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>
 1 (Intercept)          0.00192  0.0502     -124.   0          0.00174  0.00212
 2 log(loan_amount)     1.44     0.00266     138.   0          1.44     1.45
 3 log(income)          1.57     0.00380     119.   0          1.56     1.59
 4 tract_minority_pop…  0.997    0.000105    -27.6  3.14e-168  0.997    0.997
 5 regionSouth          1.12     0.00727      16.2  1.08e- 58  1.11     1.14
 6 regionWest           0.883    0.00670     -18.5  8.27e- 77  0.872    0.895
 7 derived_raceAmeric…  1.21     0.0483        4.00 6.25e-  5  1.10     1.33
 8 derived_raceAsian    2.05     0.0405       17.7  1.78e- 70  1.90     2.22
 9 derived_raceBlack …  1.34     0.0407        7.21 5.61e- 13  1.24     1.45
10 derived_raceFree F…  0.755    0.102        -2.75 6.04e-  3  0.618    0.922
11 derived_raceJoint    2.56     0.0433       21.7  9.09e-105  2.36     2.79
12 derived_raceNative…  1.20     0.0558        3.33 8.60e-  4  1.08     1.34
13 derived_raceRace N…  1.71     0.0402       13.3  1.27e- 40  1.58     1.85
14 derived_raceWhite    2.57     0.0401       23.6  4.90e-123  2.38     2.78
```

Monte Carlo Simulation – Stress Testing the Model

Synthetic Data Generation

Created a synthetic dataset with known 'true' coefficients for race and region to test the logistic regression model.

Simulation Runs

Run 1,000 simulations, each time using the logistic regression model to predict approval probabilities on the synthetic data.

Evaluation Metrics

Measured the bias, confidence interval coverage, and variance of the estimated coefficients across the simulation runs to assess the model's reliability.

Stress Testing

Explored how the model's performance changed under different stress test conditions, such as when the overall approval rate was very low (e.g., 10%).

Summary

Logistic regression model found race and region effects

Even after controlling for income and loan amount, the model showed differences in approval odds based on race and region

Simulation tested model reliability

A custom Monte Carlo simulation revealed the conditions under which the logistic regression could reliably detect the disparities

Reliable under 'ideal' conditions

The model performed well in detecting the true effects when approval rates were typical and predictors had strong signals

Less reliable in 'stress test' scenarios

The model's performance degraded when approval rates were very low or predictors were noisy, leading to biased estimates and poor confidence interval coverage

The logistic regression model was able to identify the effects of race and region on mortgage approval, but its reliability depended on the specific conditions of the data and approval process.

Reflection

- **Cleaning matters more than I thought**

  I realized that extensive data cleaning and preparation was crucial for building a reliable model. Small decisions around encoding, transformations, and handling missing data had a big impact on the results.

- **EDA gives useful direction, but modeling confirms it**

  The exploratory data analysis surfaced some interesting patterns, but formally testing them with a logistic regression model provided stronger evidence and insights about the drivers of mortgage approval disparities.

- **Simulations are a powerful tool to test modeling assumptions**

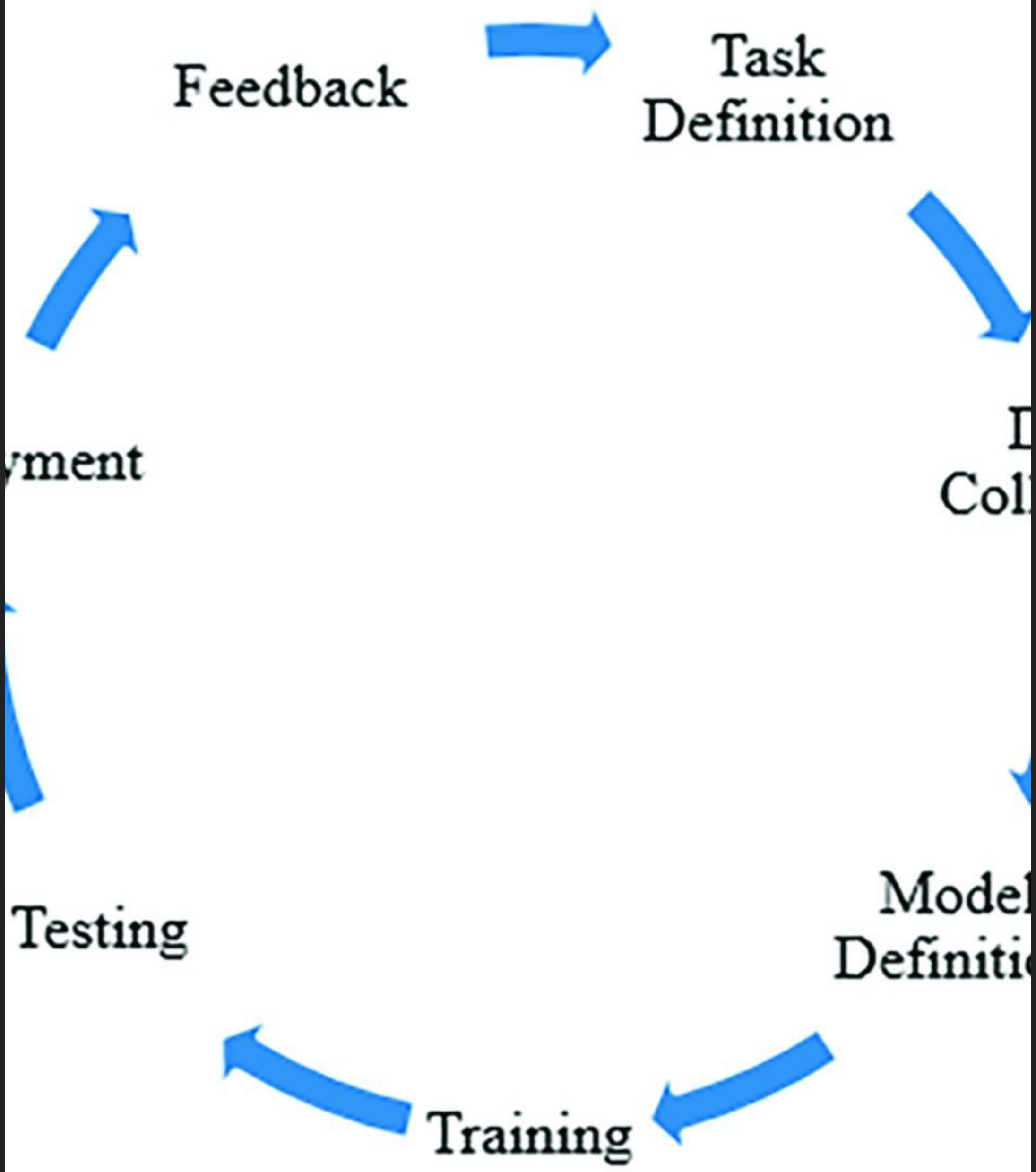
  The custom Monte Carlo simulation I designed would allow me to stress test the logistic regression model under different scenarios, revealing its strengths and limitations in detecting bias.

- **Feedback improved the simulation write-up**

  After receiving feedback, I clarified my variable encoding choices and provided a more detailed explanation of the simulation methodology, making the analysis more transparent and robust.

- **Try non-logistic models for comparison**

  If I were to do this project again, I would explore other modeling approaches like random forests or fairness-aware algorithms to see if they provide additional insights or more reliable results.



Why This Actually Matters

- Mortgage approval affects access to housing, education, and opportunity
- Bias in models can quietly reinforce real-world inequality
- Even “neutral” algorithms can produce unfair results
- It’s not just about numbers — it’s about people’s lives
- Modeling well means thinking ethically, not just statistically

This project made me realize: if we’re going to build models that influence real decisions, we have to treat fairness like a technical goal — not just a moral one