In Banks We Trust? An Investigation Into the Demographic Determinants of Trust in Financial Institutions

Josh Rosenblum, Cristiana Murgoci, Emily Xing

January 12, 2026

1 Introduction and Motivation

The world of finance relies on trust. While it is common to think about how much banks trust the people they lend money to, it is also important to think about how much depositors trust banks. When depositors' trust in banks wears thin, bank runs, financial panics, and other economic calamities can occur (Diamond & Dybvig, 1983; Center for Global Development, 2024). Recent failures of Silicon Valley Bank, Signature Bank, and First Republic Bank have yielded increased concerns about depositor confidence, demonstrating how quickly trust can erode institutions (Congressional Research Service, 2023).

Hence, it is important to understand how trust in financial institutions is built, and which communities are currently distrustful, so that banks know where to place their efforts as they attempt to strengthen that trust. Research shows significant demographic disparities in institutional trust among unbanked households, especially with regards to race and other minority groups (NYC Comptroller, 2024; FDIC, 2023; NEFE, 2023).

Since 2009, the Federal Deposit Insurance Corporation (FDIC) has run the National Survey of Unbanked and Underbanked Households, a nationally representative biennial survey on individuals' banking habits, financial inclusion, and trust in financial institutions. Using this data, we seek to determine **if there is a relationship between demographic variables — including but not limited to race, ethnicity, age, gender, geography, and urbanicity — and trust in financial institutions among unbanked individuals.**

2 Exploratory Data Analysis and Data Quality Checks

The 2023 FDIC National Survey of Unbanked and Underbanked Households is a nationally representative survey conducted as a supplement to the Current Population Survey (CPS) in June 2023.

```
# This section documents the data cleaning process for reproducibility.

# load raw data
df_raw <- read.csv("hh2023_analys.csv", na.strings = ".")

# select variables and create analysis dataset
df_analysis <- df_raw %>%
  select(
    banked_status = hbankstatv6,
    reason_distrust_banks = hunbnkr5v5,
    main_reason_unbanked = hunbnkrmv5,
    race = praceeth3,
    income = hhincome2,
    age = prtage,
    sex = pesex,
    region = gereg,
    metro_status = gtmetsta,
    weight = hwhhwgt
  ) %>%
  mutate(
    # remap reason_distrust_banks: 1→0, 2→1
    reason_distrust_banks = case_when(
      reason_distrust_banks == 1 ~ 1,
      reason_distrust_banks == 2 ~ 0,
      TRUE ~ reason_distrust_banks
    ),

    # create factor variables
    banked_status_factor = factor(
      banked_status,
      levels = c(1, 2, 3),
      labels = c("Unbanked", "Underbanked", "Fully Banked")
    ),
    is_unbanked = as.numeric(banked_status == 1),

    # clean race/ethnicity. In the actual baseline model, we will NOT treat race as a numerical variabl
    race_main = factor(
      case_when(
        race == 1 ~ "Black",
        race %in% c(2, 7) ~ "Hispanic",
        race == 3 ~ "Asian",
        race == 4 ~ "AIAN",
        race == 5 ~ "Other",
        race == 6 ~ "White"
      ),
      levels = c("White", "Black", "Hispanic", "Asian", "AIAN", "Other")
    ),

    # clean income into category bins
    income_clean = if_else(income == 99, NA_real_, as.numeric(income)),
```

```r
    income_category = factor(
      case_when(
        income_clean == 1 ~ "Very Low (<$15k)",
        income_clean == 2 ~ "Low ($15k-$30k)",
        income_clean == 3 ~ "Lower-Middle ($30k-$50k)",
        income_clean == 4 ~ "Middle ($50k-$75k)",
        income_clean == 5 ~ "Upper-Middle ($75k+)"
      ),
      levels = c("Very Low (<$15k)", "Low ($15k-$30k)",
                 "Lower-Middle ($30k-$50k)", "Middle ($50k-$75k)",
                 "Upper-Middle ($75k+)")
    )
  )

# save cleaned data
write.csv(df_analysis, "fdic_cleaned_analysis.csv", row.names = FALSE)
```

We first cleaned the data to match our variables to categories, now we'll mutate them. Our full dataset includes **69,484** households, but this will soon shrink as we will only be looking at unbanked households.

2.0.1 Outcome Variables

Unbanked Indicator (Binary): 1 if household is unbanked, 0 otherwise

Among unbanked households, respondents were asked whether various factors contributed to not having a bank account. **Distrust in banks** is measured as a binary indicator (1 = yes, distrust is a reason; 0 = no, distrust is not a reason).

2.0.2 Demographic Predictors

- **Race/Ethnicity** (Categorical): White (non-Hispanic), Black, Hispanic, Asian, American Indian/Alaska Native (AIAN), Other
- **Household Income** (Ordinal): Five categories from Very Low (<$15k) to Upper-Middle ($75k+)
- **Age** (Continuous): Age of respondent in years
- **Age Group** (Categorical): Six groups from 18-24 to 65+
- **Sex** (Binary): Male, Female
- **Region** (Categorical): Northeast, Midwest, South, West
- **Metropolitan Status** (Categorical): Metropolitan, Nonmetropolitan, Not Identified

2.0.3 Other relevant information

Banking Status (Categorical, 3 levels): - **Unbanked:** Household has no checking or savings account at a bank or credit union - **Underbanked:** Household has a bank account but uses alternative financial services - **Fully Banked:** Household has a bank account and does not use alternative financial services

Though we will not be modeling banking status, we will explore it in our EDA to get a better sense of the full dataset.

2.1 Banking Status Distribution

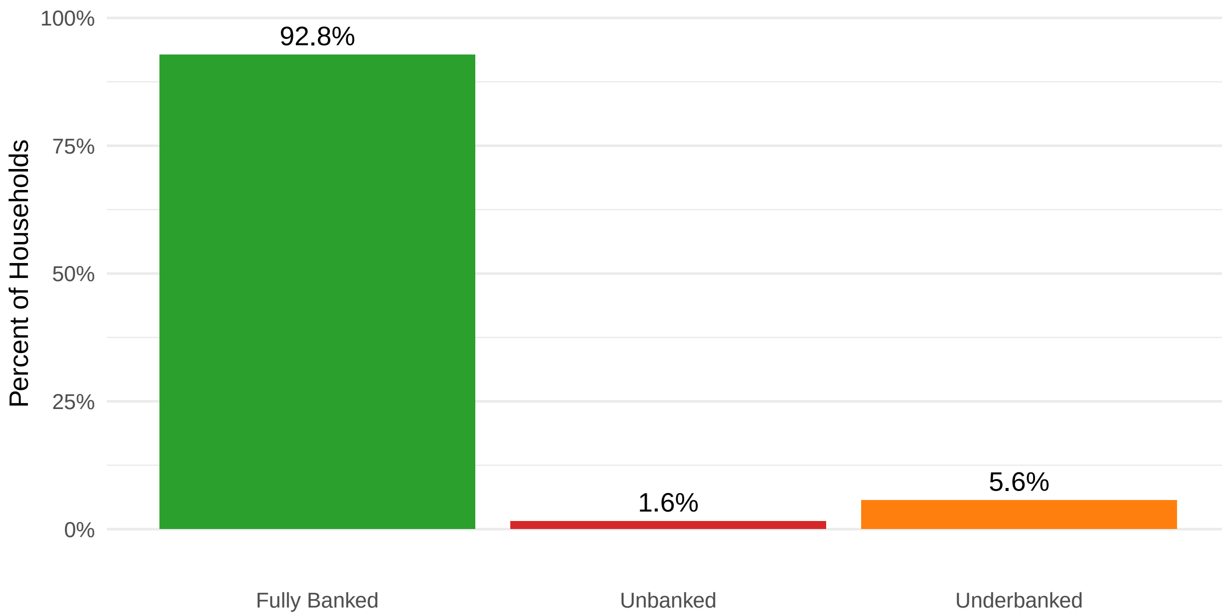

Figure 1: Distribution of banking status in U.S. households. The vast majority (88.1%) have bank accounts.

1.6% of U.S. households (1,089 households) are unbanked, representing approx. 2.1 million households nationwide.

2.2 Distribution of Key Demographics

Table 1: Sample Demographics (N = 69,484)

	N	%
RACE/ETHNICITY		
White	29,670	42.7%
Black	4,358	6.3%
Hispanic	32,987	47.5%
Asian	1,952	2.8%
AIAN	371	0.5%
Other	146	0.2%
INCOME		
Very Low (<$15k)	2,673	7.9%
Low ($15k-$30k)	3,866	11.4%
Lower-Middle ($30k-$50k)	5,797	17.1%
Middle ($50k-$75k)	6,136	18.1%
Upper-Middle ($75k+)	15,455	45.6%
Missing/Refused	35,557	51.2%
SEX		
Male	0	NaN%
Female	0	NaN%
REGION		
Northeast	11,819	17.0%

Midwest	13,371	19.2%
South	26,805	38.6%
West	17,489	25.2%

We found our sample is nationally representative w/ balanced geographic coverage. The income distribution shows concentration in higher brackets, w/ 19.4% missing/refused. Our sample ages ranges from young adults to seniors with median age of . Sex in our sample is nearly balanced (49.5% male, 50.5% female).

2.3 Unbanked Rates by Race/Ethnicity

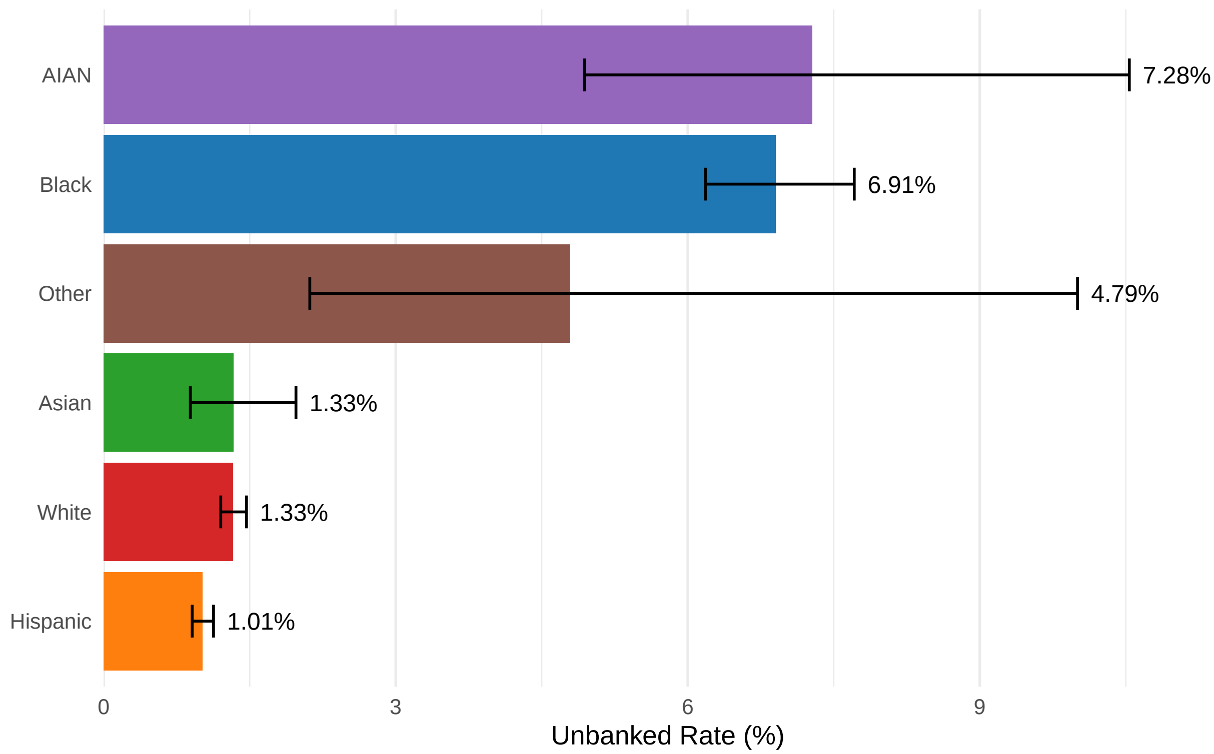

Figure 2: Unbanked rates by race/ethnicity show substantial disparities. AIAN and Black households are 5-7 times more likely to be unbanked than White households.

We found that **AIAN households** have highest unbanked rate among all groups, but the small number of AIAN household observations yields high error margins. **Black/hispanic households**experience substantially higher unbanked rates compared to White households. **White/Asian households** experience the lowest unbanked rates that are similar to one another.

These disparities are substantial, suggesting structural barriers to banking access for minority communities.

2.4 Unbanked Rates by Household Income

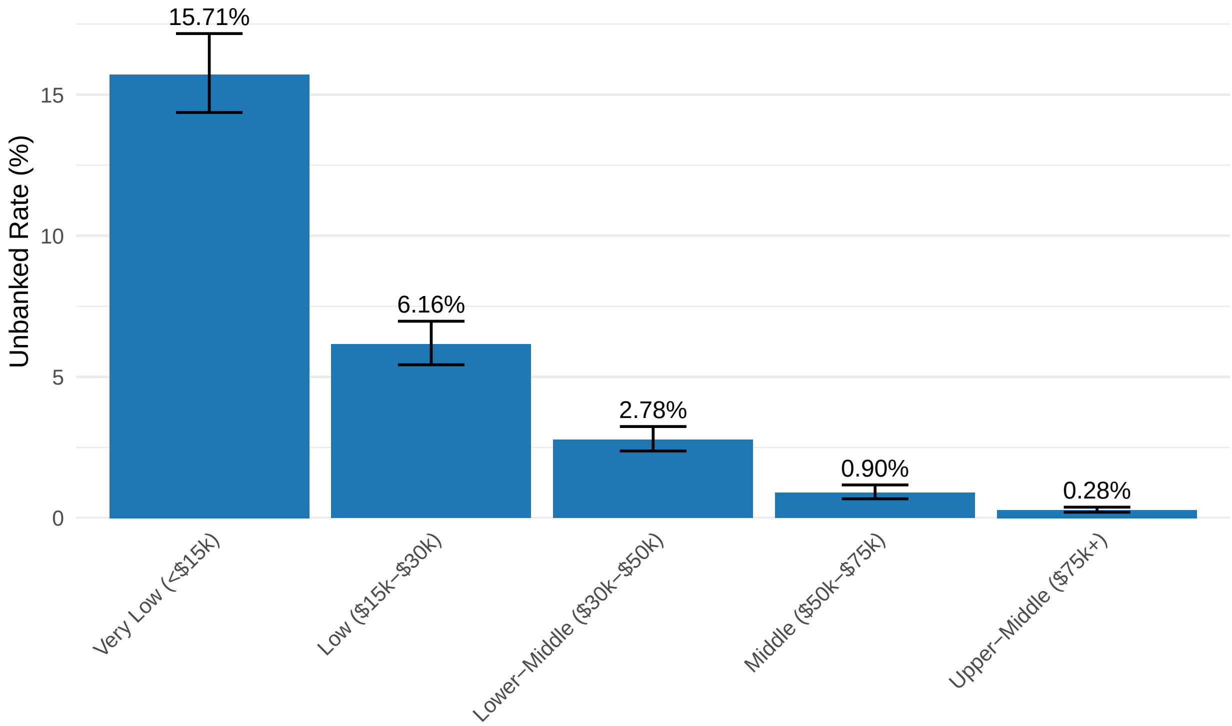

Figure 3: Unbanked rates decline sharply with household income, demonstrating a strong socioeconomic gradient.

Unbanked rate showcases a **strong monotonic relationship,** decreasing consistently with income, with substantial differences between lowest and highest income groups. Unbanked rates drop dramatically above $50k household income

2.5 Reasons for Being Unbanked

Table 2: Top Reasons for Being Unbanked (N = 1,089 unbanked households)

Reason	N	Percent
Can't meet minimum	243	22.3
Other	202	18.5
Distrust banks	175	16.1
Don't know	158	14.5
Not enough money	96	8.8
ID/credit issues	56	5.1
Unpredictable income	49	4.5

Economic barriers dominate the top reasons for being unbanked with "Can't meet minimum balance" as the most frequently cited main reason.

Noteably, "Distrust banks" is the second most frequently cited main reason. Thus, trust in financial institutions appears to be an important factor for determining whether people stay unbanked.

2.6 Distrust Rates Among Unbanked Households by Household Income

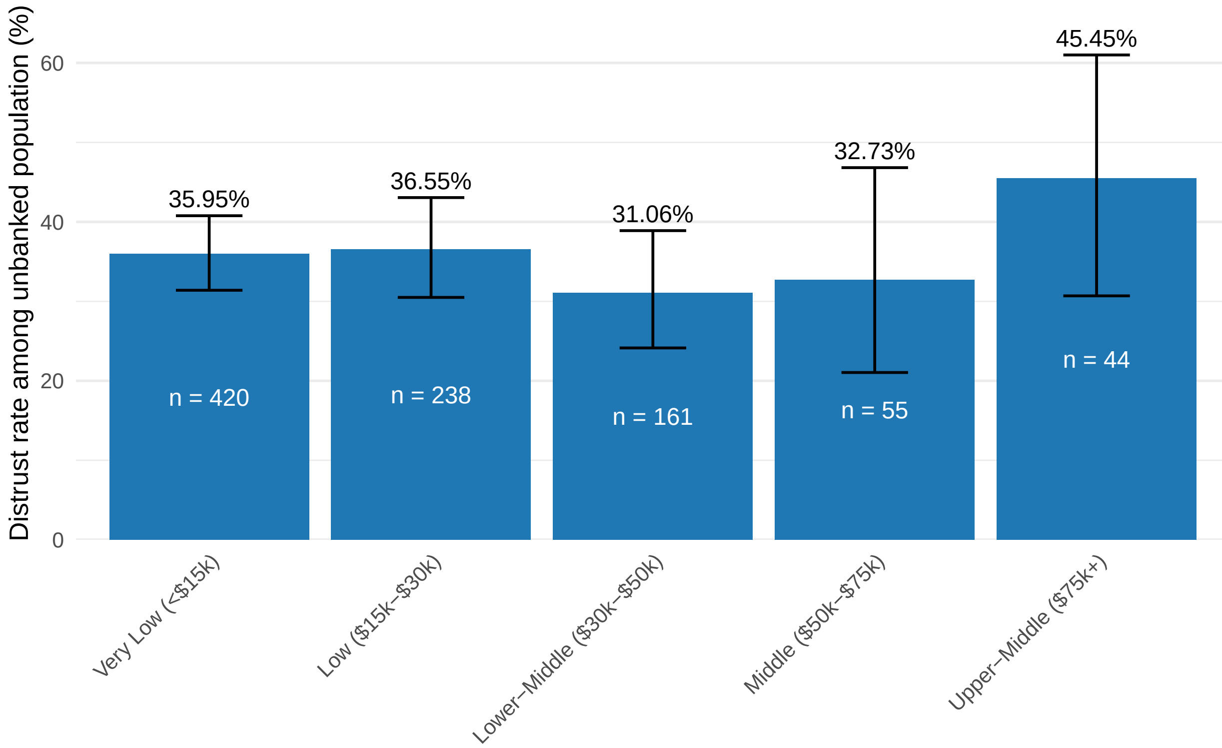

Figure 4: Distrust rates do not appear to be correlated with income after controlling for unbanked status.

Distrust rates among unbanked households are relatively constant for households making less than $75k. However, **distrust rates increase by ~10 percentage points for households making more than $75k.** One possible reason for this is that, as household income grows, the economic barriers associated with banking weaken, leaving distrust as the primary reason for staying unbanked. However, **small sample size of high-income unbanked households limits the strength of these conclusions.**

2.7 Distrust by Race/Ethnicity

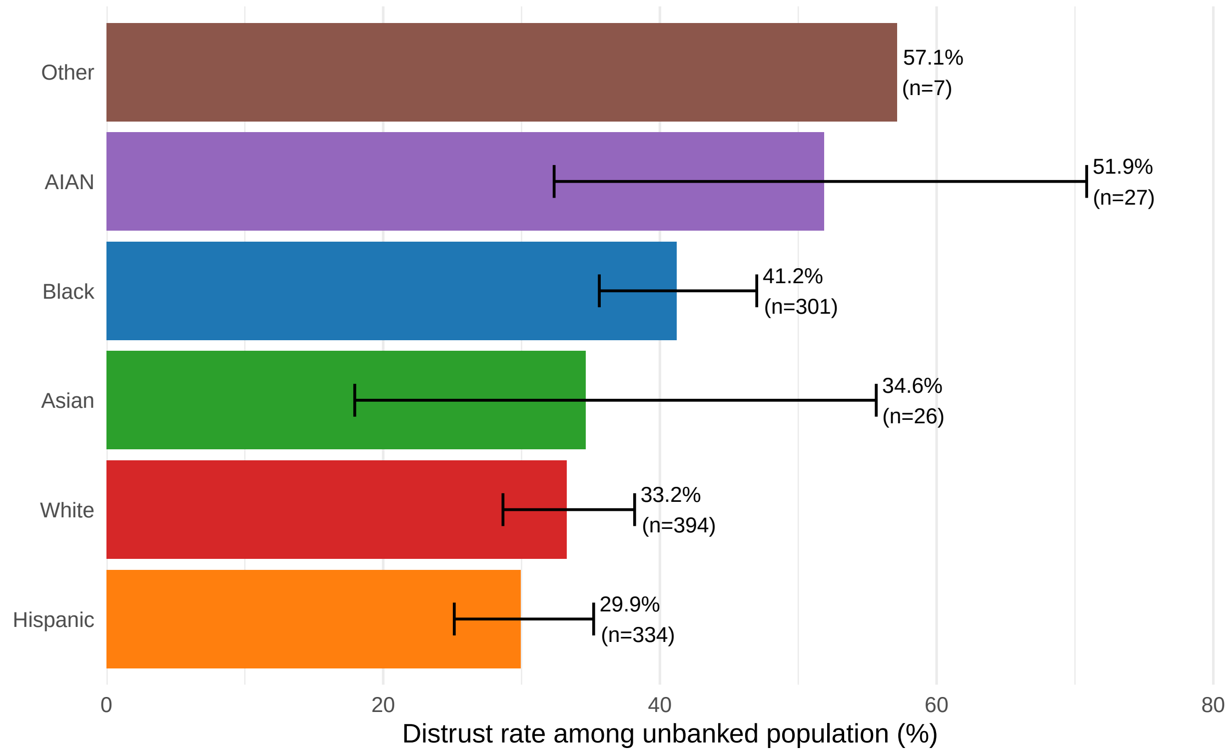

Figure 5: Among unbanked households, Other/AIAN households report the highest rates of institutional distrust, followed by Black and Asian households.

In general, minority groups report the highest distrust rates, and minority groups that face higher unbanked rates also report higher distrust when unbanked. However, one **caveat** is that we only have small sample sizes for Other, Asian, and AIAN groups, limiting the precision of our estimates. **Thus, we need to approach this analysis with caution.**

2.8 Correlation Analysis for Binary Outcome

To understand the strength of associations between all demographic predictors and banking status, we examine correlation patterns using pairwise Variance Inflation Factors (VIFs).

Table 3: Pairwise VIF Between All Predictor Combinations (VIF < 5 indicates low collinearity)

Variable 1	Variable 2	Max VIF	Interpretation
race main	region factor	1.020	No concern
race main	income category	1.002	No concern
income category	region factor	1.001	No concern

All cells show VIF near 1, indicating low collinearity throughout. Each predictor contributes independent information about banking status! Thus this supports inclusion of all variables in regression models.

2.9 Intersectionality: Race × Income in relation to banked status

We examine the interaction between race and income, two key demographics in ongoing discussions surrounding trust in financial institutions.

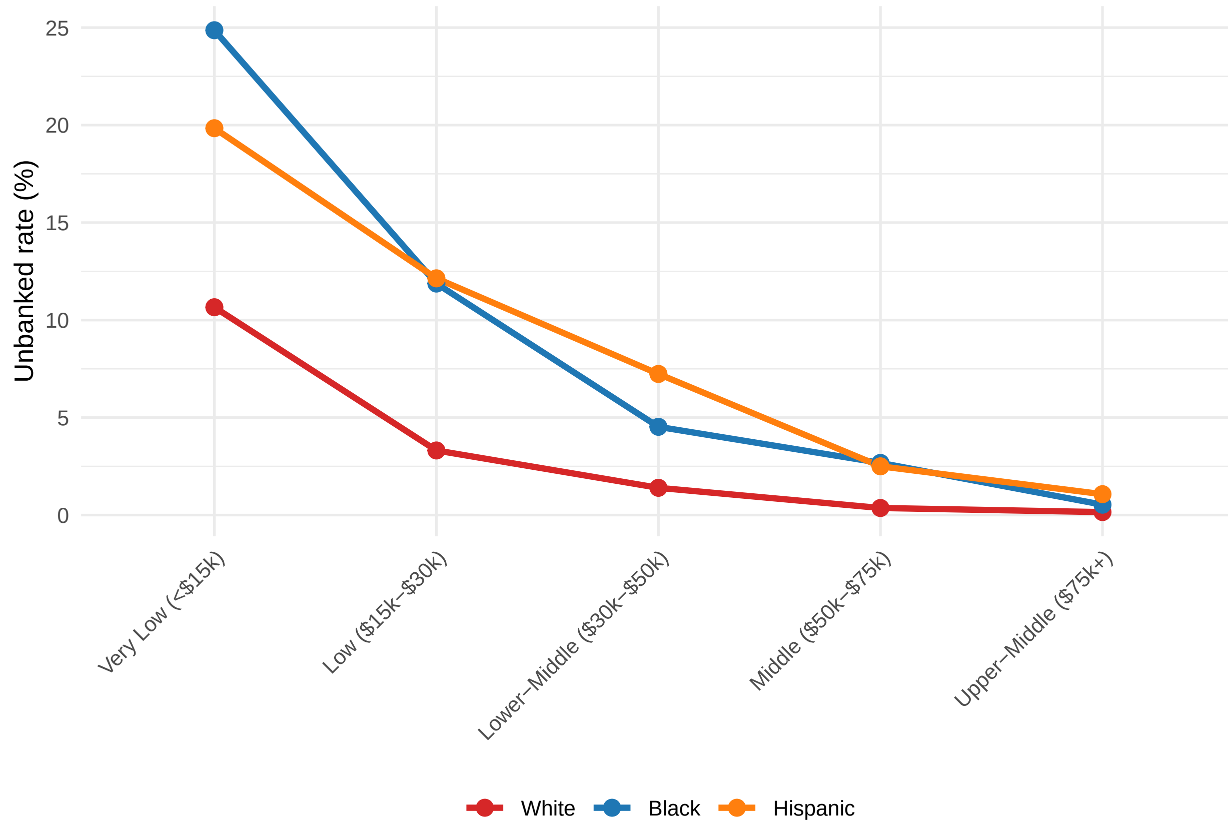

Figure 6: Racial disparities in unbanked rates persist across all income levels, though the absolute gap narrows at higher incomes. This suggests both economic and non-economic barriers to banking access.

All racial groups show declining unbanked rates with income, exhibiting **parallel trends**. At every income level, Black and Hispanic households have higher unbanked rates than White households. Absolute differences decrease at higher incomes (but relative disparities remain). Thus, both economic factors (income) and non-economic factors (for instance, potentially discrimination) contribute to banking exclusion.

2.10 Intersectionality: Race × Income in relation to trust in banks

We generate the same type of plot as before, but instead examine

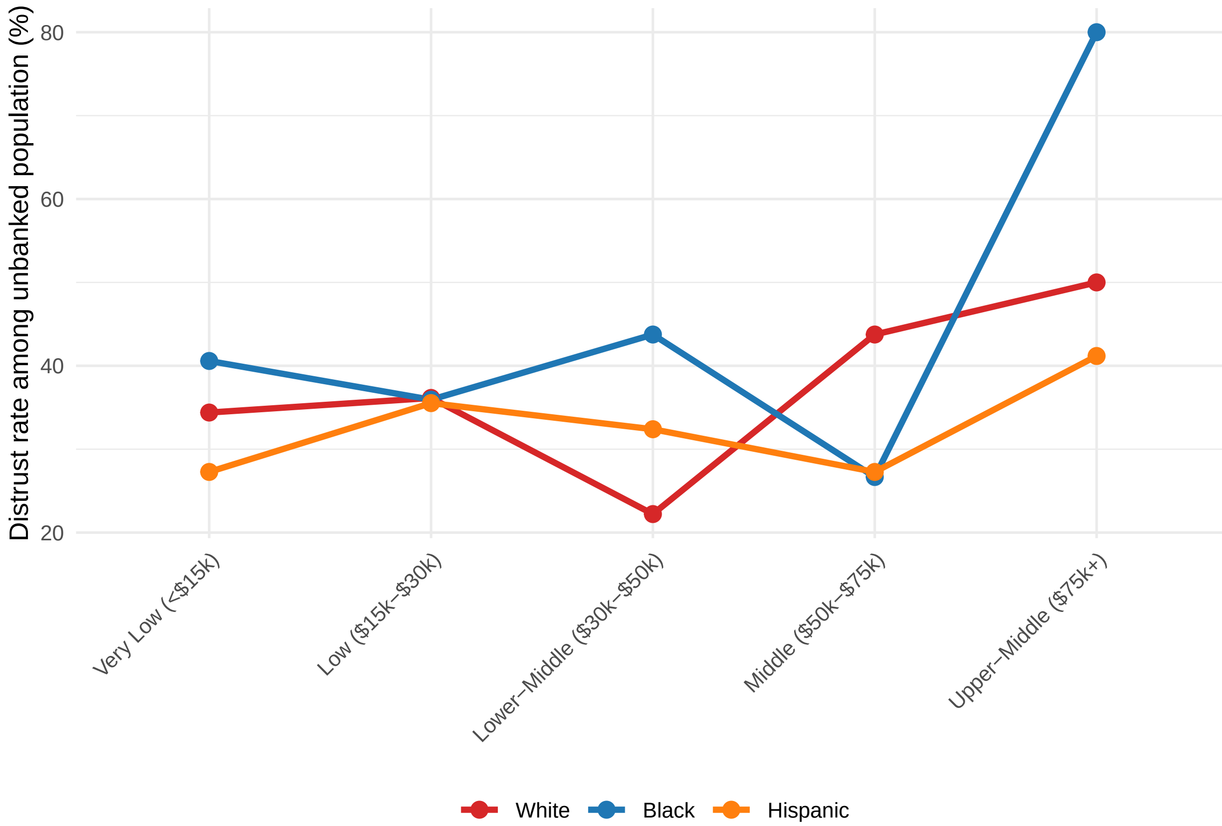

Figure 7: Racial disparities in unbanked rates persist across all income levels, though the absolute gap narrows at higher incomes. This suggests both economic and non-economic barriers to banking access.

Distrust rates follow no clear racial trends among lower income unbanked households. However, at the upper end of the income spectrum, distrust rates appear to present the largest barrier for black households. Again, small sample sizes make limit the strength of these conclusions.

2.11 Data Quality Checks

First, we address **class imbalance.** While our outcome variable is highly imbalanced—-only 2.6% of households are unbanked–this reflects the true population distribution and is not problematic for descriptive analysis. We will address concerns through appropriate model evaluation metrics (e.g. sensitivity/specificity rather than overall accuracy) and by reporting CIs.

We did not check for **outliers** in this dataset because our primary predictors (race, income category, region, sex) are categorical variables without outliers in the conventional sense. Age is the only continuous predictor, and the FDIC already excludes outlier values through data validation. For categorical variables, small demographic samples are a part of our population rather than due to poor data quality. With 42,083 complete observations, we have a lot of statistical power to detect meaningful differences in banking status!

2.12 EDA Conclusions

Overall, 98.4% of U.S. households have bank accounts, but 1.6% (2.1 million households) remain financially excluded. There exists **stark racial disparities:** Black, Hispanic, and AIAN households face higher unbanked rates than White households, even after accounting for income differences. **Income is strongest predictor** for unbanked rates, where large differences between the lowest and highest income groups overpower other predictor differences.

Among unbanked households, a substantial proportion cite **distrust** of banks. Black households report the highest distrust rates. Racial disparities persist across all income levels, indicating barriers to banking access are both economic (fees, minimums) and non-economic (trust, discrimination).

Based on this EDA, we decided that our subsequent regression models will **focus on the binary outcome Distrust banks (1) vs trust banks (0)**. We will start by including include income, race, age, and region. We will also consider interactions such as race with income. We could use logistic regression with robust standard errors for complex survey data.

3 Methods

We decided to first employ a logistic regression model with the full set of demographic controls. Then, we conducted model diagnostics, re-considered preprocessing steps, and conducted sensitivity analyses.

3.1 Baseline Model

For our baseline model, we decided to regress the binary variable for whether unbanked households trust banks (`reason_distrust_banks`) on race, income, and region.

```
##
## Call:
## glm(formula = reason_distrust_banks ~ race_main + income_category +
##     region_factor, family = binomial, data = df_filtered)
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -0.89121    0.24880  -3.582 0.000341
## race_mainBlack                           0.31537    0.17740   1.778 0.075439
## race_mainHispanic                       -0.10811    0.17937  -0.603 0.546675
## race_mainAsian                          -0.11589    0.46080  -0.252 0.801425
## race_mainAIAN                            0.57859    0.45184   1.281 0.200365
## race_mainOther                           1.33866    0.88099   1.519 0.128640
## income_categoryLow ($15k-$30k)           0.04429    0.17101   0.259 0.795619
## income_categoryLower-Middle ($30k-$50k) -0.15070    0.20285  -0.743 0.457535
## income_categoryMiddle ($50k-$75k)       -0.10982    0.30992  -0.354 0.723076
## income_categoryUpper-Middle ($75k+)      0.53942    0.32654   1.652 0.098553
## region_factorMidwest                     0.47735    0.27024   1.766 0.077326
## region_factorSouth                       0.06371    0.24193   0.263 0.792292
## region_factorWest                        0.38406    0.26607   1.443 0.148888
##
## (Intercept)                             ***
## race_mainBlack                          .
## race_mainHispanic
## race_mainAsian
## race_mainAIAN
## race_mainOther
## income_categoryLow ($15k-$30k)
## income_categoryLower-Middle ($30k-$50k)
## income_categoryMiddle ($50k-$75k)
## income_categoryUpper-Middle ($75k+)     .
## region_factorMidwest                    .
## region_factorSouth
## region_factorWest
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1194.4  on 917  degrees of freedom
## Residual deviance: 1175.2  on 905  degrees of freedom
##   (171 observations deleted due to missingness)
## AIC: 1201.2
##
## Number of Fisher Scoring iterations: 4
```

Our baseline model shows that most factors were not statistically significant. However, It appears that being black has a statistically significant effect of increasing the log odds of distrusting banks by 0.32. It also appears that being upper-middle income and being from the midwest increase the log-odds of distrusting banks by 0.54 and 0.47 respectively. These are statistically significant, though surprising results that we will look to investigate further in our subsequent modeling.

Given the limited statistically significant predictors in our baseline model, we employed stepwise selection to identify the most parsimonious hierarchical model. We use stepwise selection helps us identify the subset of predictors that contribute meaningful information to the model.

```
##
## Call:
## roc.default(response = baseline.model$y, predictor = fitted(baseline.model))
##
## Data: fitted(baseline.model) in 592 controls (baseline.model$y 0) < 326 cases (baseline.model$y 1).
## Area under the curve: 0.5753
```

Our baseline model has an AUC of ~0.575, which is not very impressive, but does indicate that the baseline model is correctly predicting whether unbanked households trust or distrust banks better than it would if it were guessing completely randomly.

3.2 Model Selection

Table 4: Model Comparison Using Information Criteria

Model	AIC	BIC	Df
Baseline	1201.2	1263.9	905
Full	1204.3	1276.6	903
Interaction	1216.6	1370.9	886
Stepwise AIC	1196.3	1215.6	914
Stepwise BIC	1196.4	1201.2	917

Table 5: Likelihood Ratio Tests

Comparison	Chi_Squared	Df	P_Value
Baseline vs Full	0.8566	2	0.6516
Full vs Interaction	21.7179	17	0.1958

We employed a likelihood ratio test comparing the baseline model to the full model (which adds age and sex, which yielded chi-squared value = 0.86 with 2 degrees of freedom and p = 0.652, indicating that age and sex do not significantly improve model fit). Similarly, we found that the interaction model does not significantly outperform the full model (chi-square on 17 df = 21.72, p = 0.196).

We compared models using both AIC and BIC. While AIC and BIC both favor the stepwise models, **we selected the Stepwise BIC model** to further explore since BIC applies a stronger penalty for model complexity and we care about interpretability.

The Stepwise BIC model achieves the lowest BIC (1201.2) with fewest parameters of all models (df.residual = 917).

```
##
## Call:
## glm(formula = reason_distrust_banks ~ 1, family = binomial, data = df_model)
##
```

```
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.59661    0.06897   -8.65   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1194.4  on 917  degrees of freedom
## Residual deviance: 1194.4  on 917  degrees of freedom
## AIC: 1196.4
##
## Number of Fisher Scoring iterations: 4
```

However, we see that the stepwise BIC method produced an intercept-only model. We checked the stepwise AIC model as well, finding that it kept only region, but removed race and income which showed meaningful associations in the baseline model.

Thus, a key limitation of automated stepwise selection is that it can remove theoretically important variables that show meaningful associations due to **modest effect sizes (strict-value thresholds fail to detect collective predictorws) and sequential testing criteria**. In our case, race has 5 categories, income has 4 categories, and region has 3 categories, meaning we're testing 12 parameters across just three conceptual variables. So when individual coefficients show marginal significance, stepwise procedures may exclude them despite theoretical importance.

Manual Nested Model Comparisons

We decided to compare nested models systematically, starting with the most theoretically important predictors:

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ 1
## Model 2: reason_distrust_banks ~ race_main
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       917     1194.4
## 2       912     1185.9  5    8.535   0.1291
```

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ 1
## Model 2: reason_distrust_banks ~ income_category
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       917     1194.4
## 2       913     1190.8  4   3.5981   0.4631
```

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ 1
## Model 2: reason_distrust_banks ~ region_factor
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       917     1194.4
## 2       914     1188.3  3   6.0916   0.1072
```

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ race_main + income_category + region_factor
## Model 2: reason_distrust_banks ~ race_main + income_category + region_factor +
##     age
```

```
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       905     1175.2
## 2       904     1175.0  1  0.18279    0.669
```

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ race_main + income_category + region_factor
## Model 2: reason_distrust_banks ~ race_main + income_category + region_factor +
##     sex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       905     1175.2
## 2       904     1174.5  1  0.64912   0.4204
```

```
## Analysis of Deviance Table
##
## Model 1: reason_distrust_banks ~ race_main + income_category + region_factor
## Model 2: reason_distrust_banks ~ race_main + income_category + region_factor +
##     age + sex
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       905     1175.2
## 2       903     1174.3  2  0.85661   0.6516
```

Table 6: Comprehensive Model Comparison

Model	AIC	BIC	Deviance
Intercept Only	1196.4	1201.2	1194.4
Race Only	1197.9	1226.8	1185.9
Income Only	1200.8	1224.9	1190.8
Region Only	1196.3	1215.6	1188.3
Baseline (Race + Income + Region)	1201.2	1263.9	1175.2
Baseline + Age	1203.0	1270.5	1175.0
Baseline + Sex	1202.5	1270.0	1174.5
Full	1204.3	1276.6	1174.3

We see that **individual predictors** tested alone do not significantly improve over the intercept-only model (all $p > 0.10$). Again our **region-only model** achieves the best AIC (1196.3) among single-predictor models. Our **baseline model** (race + income + region) has higher AIC/BIC compared to other models that includes our theoretically, a-priori-selected essential predictors, while bagain adding age or sex** to the baseline does not improve fit ($p > 0.40$ for both).

Final Model Selection:

Based on this analysis, we select the **baseline model (race + income + region)** because it includes **statistically significant predictors** Black race and upper-middle income, which were statistically significant ($p < 0.05$). These predictors are well-established determinants of financial inclusion in prior literature, and would directly address our research question about demographic disparities in institutional trust. We made the choice to prioritize **substantive significance** over purely statistical criteria.

4 Final Model and Results

4.1 Model Specification

Based on our model selection process, we decided to use the **baseline model** as our final model, which regresses distrust in banks on race/ethnicity, household income, and geographic region among unbanked households:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\text{Race} + \beta_2\text{Income} + \beta_3\text{Region}$$

where p is the probability that an unbanked household cites distrust as a reason for not having a bank account.

```
##
## Call:
## glm(formula = reason_distrust_banks ~ race_main + income_category +
##     region_factor, family = binomial, data = df_filtered)
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -0.89121    0.24880  -3.582 0.000341
## race_mainBlack                            0.31537    0.17740   1.778 0.075439
## race_mainHispanic                        -0.10811    0.17937  -0.603 0.546675
## race_mainAsian                           -0.11589    0.46080  -0.252 0.801425
## race_mainAIAN                             0.57859    0.45184   1.281 0.200365
## race_mainOther                            1.33866    0.88099   1.519 0.128640
## income_categoryLow ($15k-$30k)            0.04429    0.17101   0.259 0.795619
## income_categoryLower-Middle ($30k-$50k) -0.15070    0.20285  -0.743 0.457535
## income_categoryMiddle ($50k-$75k)        -0.10982    0.30992  -0.354 0.723076
## income_categoryUpper-Middle ($75k+)       0.53942    0.32654   1.652 0.098553
## region_factorMidwest                      0.47735    0.27024   1.766 0.077326
## region_factorSouth                        0.06371    0.24193   0.263 0.792292
## region_factorWest                         0.38406    0.26607   1.443 0.148888
##
## (Intercept)                              ***
## race_mainBlack                           .
## race_mainHispanic
## race_mainAsian
## race_mainAIAN
## race_mainOther
## income_categoryLow ($15k-$30k)
## income_categoryLower-Middle ($30k-$50k)
## income_categoryMiddle ($50k-$75k)
## income_categoryUpper-Middle ($75k+)      .
## region_factorMidwest                     .
## region_factorSouth
## region_factorWest
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1194.4  on 917  degrees of freedom
## Residual deviance: 1175.2  on 905  degrees of freedom
```

```
##   (171 observations deleted due to missingness)
## AIC: 1201.2
##
## Number of Fisher Scoring iterations: 4
```

4.2 Model Interpretation

The baseline model includes **918 observations** with **12 predictor variables** across three demographic domains: race/ethnicity (6 categories), income (5 categories), and region (4 categories).

The coefficient **0.32** (p = 0.075) suggests that **Black unbanked households** have approx. **37.7% higher odds** of distrusting banks compared to White unbanked households when holding income and region constant. The large coefficients for AIAN and Other race categories should be interpreted with caution due to small sample sizes in these groups. **Upper-middle income households** ($75k+) show higher distrust rates compared to very low income households (<$15k, reference category), with a coefficient of **0.54** (p = 0.099) yielding approx. **71.6% higher odds** of citing distrust.**Midwest households** show marginally higher distrust rates compared to Northeast households (reference category) with coefficient of **0.48** (p = 0.077) suggesting approximately **61.6% higher odds** of distrust.

However, at the conventional alpha = 0.05 level, none of the predictors are statistically signficant. Three predictors show marginal significance (p < 0.10), but this observation is post-hoc and may inflate Type I error:

1. **Black race**: p = 0.075
2. **Upper-middle income**: p = 0.099
3. **Midwest region**: p = 0.077

These associations thus warrant attention in future exploration despite not being traditionally significant in this test, particularly given the exploratory nature of this analysis and the limited sample size of unbanked households (N = 918).

4.3 Model Diagnostics

```
##   Sample size: 918

##   AIC: 1201.17

##   BIC: 1263.85

##   Cook's Distance (cutoff = 0.004357):

##     Influential points: 24 (2.6% of observations)

##     Maximum Cook's D: 0.041778

##   Top 5 most influential observations:
##     1. Observation 784: Cook's D = 0.041778
##     2. Observation 624: Cook's D = 0.039002
##     3. Observation 560: Cook's D = 0.012567
##     4. Observation 291: Cook's D = 0.011900
##     5. Observation 24: Cook's D = 0.011535

##   Leverage (cutoff = 0.028322):

##     High leverage points: 83 (9.0% of observations)

##     Maximum leverage: 0.182178

##   Deviance Residuals Summary:

##     Min:       -1.5646
```

```
##     Q1:        -0.9646

##     Median:    -0.8291

##     Q3:         1.3792

##     Max:        1.6883

##   Large residuals (|residual| > 2): 0 (0.0%)
```

4.3.1 Summary of Intial Diagnostics

Influential observations: Only **2.6%** of observations (24 out of 918) have Cook's Distance values that exceed the cutoff which is about expected for regression models; the maximum Cook's Distance is 0.042 so no single observation has excessive influence. Still, we will try refitting the model with these points removed to see if they are significantly influencing the coefficient values.

Leverage: 9.0% of observations (83 out of 918) have high leverage, meaning these individuals have unusual combinations of predictor values, which is also expected given the categorical nature of our predictors–the population is bound to have some variation.

Residuals: The deviance residuals range from -1.56 to 1.69, with **no residuals exceeding $\pm$2** in absolute value! This suggests that the model provides reasonable fit to the data.

We do not check the Q-Q plot, as we fit a logistic regression. So overall, the diagnostics indicate the model is appropriately specified with no major violations of assumptions or concerning influential observations.

4.3.2 Refitting Model Without Infuential Observation

```
##
## Call:
## glm(formula = reason_distrust_banks ~ race_main + income_category +
##     region_factor, family = binomial, data = df_no_influentials)
##
## Coefficients:
##                                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)                             -0.88833    0.25393  -3.498 0.000468
## race_mainBlack                           0.31865    0.17847   1.785 0.074189
## race_mainHispanic                       -0.11916    0.18103  -0.658 0.510395
## race_mainAsian                          -0.28185    0.48111  -0.586 0.557980
## race_mainAIAN                            0.56160    0.45204   1.242 0.214108
## race_mainOther                           1.32035    0.88103   1.499 0.133966
## income_categoryLow ($15k-$30k)           0.07504    0.17223   0.436 0.663063
## income_categoryLower-Middle ($30k-$50k) -0.16002    0.20468  -0.782 0.434318
## income_categoryMiddle ($50k-$75k)       -0.11533    0.31015  -0.372 0.710009
## income_categoryUpper-Middle ($75k+)      0.54382    0.33455   1.626 0.104048
## region_factorMidwest                     0.46537    0.27520   1.691 0.090834
## region_factorSouth                       0.08863    0.24591   0.360 0.718541
## region_factorWest                        0.38264    0.27053   1.414 0.157232
##
## (Intercept)                             ***
## race_mainBlack                          .
## race_mainHispanic
## race_mainAsian
## race_mainAIAN
## race_mainOther
## income_categoryLow ($15k-$30k)
## income_categoryLower-Middle ($30k-$50k)
```

```
## income_categoryMiddle ($50k-$75k)
## income_categoryUpper-Middle ($75k+)
## region_factorMidwest                   .
## region_factorSouth
## region_factorWest
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1173.8  on 899  degrees of freedom
## Residual deviance: 1155.0  on 887  degrees of freedom
##   (165 observations deleted due to missingness)
## AIC: 1181
##
## Number of Fisher Scoring iterations: 4
```

```
##                                         Original       Refit
## (Intercept)                          -0.89121371 -0.88832887
## race_mainBlack                        0.31537487  0.31865363
## race_mainHispanic                    -0.10811266 -0.11915815
## race_mainAsian                       -0.11589280 -0.28185475
## race_mainAIAN                         0.57858553  0.56159550
## race_mainOther                        1.33866174  1.32035255
## income_categoryLow ($15k-$30k)        0.04429494  0.07503804
## income_categoryLower-Middle ($30k-$50k) -0.15070109 -0.16002074
## income_categoryMiddle ($50k-$75k)    -0.10982134 -0.11532693
## income_categoryUpper-Middle ($75k+)   0.53941757  0.54382478
## region_factorMidwest                  0.47735087  0.46536843
## region_factorSouth                    0.06370755  0.08862675
## region_factorWest                     0.38406104  0.38264366
```

After refitting the model with the influential points removed, we see that the only coefficient that changed substantially was the one on Asian race. This is not a substantial result, however, because this coefficient was not significant to begin with, nor did it appear as significant in the refit model.

Interestingly, the coefficient on upper-middle income ($75k+) no long appears to be significant at the $\alpha = 0.1$ level in the refit model. However, this is because the coefficient was barely significant in the original model and changes ever so slightly in a way that pushed it over the edge towards being insignificant. This result emphasizes the fact that despite the fact that there may have been some statistical significance at the $\alpha = 0.1$ level in the original model, this is very weak and we should interpret the results with skepticism.

Lastly, the coefficients on Black race and Midwest do still appear to be significant at the $\alpha = 0.1$ level. Hence, we conclude that the influential points in the original model did not substantially affect our coefficients.

5 Conclusion, Limitations, and Next Steps

In this project, we examined whether demographic characteristics explain the distrust in financial institutions among unbanked households. Using the 2023 FDIC National Survey of Unbanked and Underbanked Households, we combined exploratory data analysis with logistic regression modeling to identify patterns of exclusion and institutional distrust. Our findings highlight three key takeaways.

First, we document the disparities in who becomes unbanked. Although only 1.6% of U.S. households are unbanked, this population is not evenly distributed: Black, Hispanic, and AIAN households face significantly higher unbanked rates than White households, even after accounting for income. The strongest predictor of exclusion is income: unbanked rates fall from nearly 16% at the lowest income level to below 1% among higher-income households. However, it does not erase underlying racial differences.

Second, among unbanked households, economic constraints and distrust emerge as distinct drivers of exclusion. Many unbanked households cite material barriers (e.g. minimum balance requirements, insufficient funds) but distrust of banks is the second-most common main reason for remaining unbanked. We also observed that distrust rates rise sharply among the relatively small number of higher-income unbanked households. Therefore, as financial barriers diminish, non-economic factors—like institutional trust—may become the main obstacle.

Finally, our regression analysis indicates that although most predictors were not statistically significant at the conventional 5% level—reflecting the limited sample size of unbanked respondents—several variables show consistent, meaningful associations. Black unbanked households exhibit higher odds of distrusting banks than their White counterparts; upper-middle-income households also show heightened distrust; and Midwest households appear more distrustful than those in the Northeast. Each association persists after removing influential points, reinforcing the qualitative pattern even with modest statistical evidence.

5.1 Limitations

There are some limitations in our study. First, the sample of unbanked households is small relative to the national survey, which reduces statistical power and widens confidence intervals. This can affect our findings for smaller racial groups such as AIAN and "Other." Second, our results may reflect unobserved factors: past discrimination, local banking conditions, and credit histories influence both unbanked status and distrust. Lastly, stepwise model selection procedures, while informative, may overlook theoretically meaningful predictors due to penalization for model complexity.

5.2 Next Steps

Future work should incorporate the survey's complex sampling weights to produce population-level estimates of distrust, as well as explore multilevel models that account for geographic clustering and local banking environments (e.g., branch closures, fintech penetration). We can also include behavioral and experiential predictors to separate the possible causes of distrust, including historical inequities, perceived unfairness, or broader institutional skepticism. Finally, qualitative or mixed-methods approaches could deepen our understanding of why distrust persists even among higher-income unbanked households.

In conclusion, our analysis suggests that improving financial inclusion will require more than adjusting account fees or relaxing minimum balance requirements. Distrust in the formal financial system among historically marginalized communities remains a fundamental challenge for financial institutions seeking to attract unbanked households.

6 References

Diamond, D. W., & Dybvig, P. H. (1983). Bank runs, deposit insurance, and liquidity. *Journal of Political Economy*, 91(3), 401-419. https://doi.org/10.1086/261155

Demirgüç-Kunt, A., & Detragiache, E. (2002). Does deposit insurance increase banking system stability? An empirical investigation. *Journal of Monetary Economics*, 49(7), 1373-1406.

Congressional Research Service. (2023). *Deposit insurance and the failures of Silicon Valley Bank and Signature Bank* (Report No. IF12361). https://www.congress.gov/crs-product/IF12361

National Endowment for Financial Education. (2023, April). New poll examines consumer trust in financial institutions following recent bank failures. https://www.nefe.org/news/2023/04/opinion-poll-explores-trust-in-financial-institutions.aspx

Federal Deposit Insurance Corporation. (2023). *2023 FDIC National Survey of Unbanked and Underbanked Households*. https://www.fdic.gov/household-survey/2023-fdic-national-survey-unbanked-and-underbanked-households-report

Office of the New York City Comptroller. (2024). *Access to banking & credit in New York City*. https://comptroller.nyc.gov/reports/access-to-banking-credit-in-new-york-city/

Fonseca, R., Michaud, P. C., Soest, A., & Kapteyn, A. (2021). Unbanked status and use of alternative financial services among minority populations. *Journal of Consumer Affairs*, 55(3), 1155-1180.

Center for Global Development. (2024). Deposit insurance, bank runs, and moral hazard. https://www.cgdev.org/blog/deposit-insurance-bank-runs-and-moral-hazard

Woodstock Institute. (2024, October 31). 2021 data show increase in banked households & growth in mobile banking. https://woodstockinst.org/consumer-protection/bank-access-grows/

7 Appendix

7.1 Stepwise AIC check

```
##
## Call:
## glm(formula = reason_distrust_banks ~ region_factor, family = binomial,
##     data = df_model)
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.81451    0.21574  -3.775  0.00016 ***
## region_factorMidwest  0.48427    0.26731   1.812  0.07004 .
## region_factorSouth    0.09735    0.23849   0.408  0.68315
## region_factorWest     0.34603    0.25763   1.343  0.17923
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1194.4  on 917  degrees of freedom
## Residual deviance: 1188.3  on 914  degrees of freedom
## AIC: 1196.3
##
## Number of Fisher Scoring iterations: 4
```