Booth RP: Data Task

REDACTED

August 27, 2025

Abstract

This document was made in application to the open pre-doctoral position at Chicago Booth.

Table of contents

0.1 Data Cleaning

0.1.1 Shape and Column Analysis

We begin by checking the structure and columns of the dataset to ensure consistency. The data contains 47,776 rows and 12 columns, including variables such as `weight`, `year`, `age`, `education`, `race`, `asset_total`, `asset_housing`, `debt_total`, `debt_housing`, and `wealth`. For our analysis, we focus on the variables relevant to wealth and asset calculations, and note that `sex` and `income` are not used further.

Column Name	Type	Description
`weight`	float64	Survey weight
`year`	int64	Survey year
`age`	int64	Age of respondent
`sex`	object	Sex (not used in analysis)
`education`	object	Education level
`race`	object	Race/ethnicity
`asset_total`	float64	Total assets
`asset_housing`	float64	Housing assets
`debt_total`	float64	Total debts
`debt_housing`	float64	Housing debts
`income`	float64	Income (not used in analysis)
`wealth`	float64	Calculated wealth

0.1.2 Missing Values Check

A review of the dataset shows that there are no missing values in any column, so no imputation or removal of rows is necessary.

0.1.3 Data Types and Ranges

Below are the observed data types and value ranges:

Variable	Type	Min	Max
`weight`	float64	0.20	31,115.82
`year`	int64	1989	2016
`age`	int64	17	95
`sex`	object	2 unique	
`education`	object	3 unique	
`race`	object	4 unique	
`asset_total`	float64	-22,487,306.62	2,928,346,179.67
`asset_housing`	float64	0.00	182,642,128.63
`debt_total`	float64	0.00	293,486,997.64
`debt_housing`	float64	0.00	44,821,081.33
`income`	float64	0.00	351,958,858.31
`wealth`	float64	-221,985,489.24	2,929,687,834.52

0.1.4 Negative Values Check and Cleaning

We identify that `asset_total` contains 7 negative values, which is about 0.01% of the data. Since assets cannot logically be negative, we set all negative values in `asset_total` to zero. This adjustment ensures that all asset values are non-negative, as required by financial logic. After this cleaning step, `asset_total` has a minimum value of zero, and no negative values remain.

Variable	Negative Values	% of Total
`weight`	0	0.00%
`asset_total`	7 (before)	0.01%
`asset_total`	0 (after)	0.00%
`asset_housing`	0	0.00%
`debt_total`	0	0.00%
`debt_housing`	0	0.00%
`income`	0	0.00%

A table of the rows with negative `asset_total` values (before cleaning) is available in the appendix or supplementary materials.

A summary table of `asset_total` after cleaning:

Statistic	`asset_total`
Min	0
Max	2,928,346,179.67
Negative Values	0

0.1.5 Outlier Detection

We also check for outliers using the interquartile range (IQR) method. While some variables have a notable number of outliers, these are retained for analysis unless they are logically impossible (such as negative assets, which have already been addressed).

Variable	Outliers (N)	% of Total	Lower Bound	Upper Bound
weight	330	0.7%	-4,095	12,858
asset_total	8,281	17.3%	-2,215,818	3,831,215
asset_housing	5,405	11.3%	-651,383	1,085,639
debt_total	5,091	10.7%	-236,639	394,398
debt_housing	5,033	10.5%	-167,927	279,879
income	7,542	15.8%	-179,464	385,518

0.2 Categorical Distribution

Race	Count	%
white	37,044	77.5%
black	5,186	10.9%
Hispanic	3,553	7.4%
other	1,993	4.2%

Education	Count	%
college degree	19,444	40.7%
no college	17,820	37.3%
some college	10,512	22.0%

Sex	Count	%
male	37,212	77.9%
female	10,564	22.1%

0.3 Year Distribution

Year	Count
1989	3,143
1992	3,906
1995	4,299
1998	4,305
2001	4,442
2004	4,519
2007	4,417
2010	6,482
2013	6,015
2016	6,248

0.4 Wealth Variable

We define the `wealth` variable using the following formula:

$$\text{wealth} = \text{asset\_total} + \text{asset\_housing} - \text{debt\_total} - \text{debt\_housing}$$

This formula is applied after cleaning `asset_total`, ensuring that all asset values used in the calculation are non-negative and logically consistent for further analysis.

0.5 Weight Variable

We also make use of the weight variable in the following way:

1. NaN values are removed from both the values and weights arrays.
2. The values and weights are sorted by value.
3. The cumulative sum of the sorted weights is computed.
4. The total weight is divided by 2 to find the "median weight."
5. The function finds the first position (index) where the cumulative weight meets or exceeds the median weight.
6. The value at this position is the weighted median entry. If the cumulative weight at that index exactly equals the median weight, the weighted median is the average of the value at that index and the next one. Otherwise, it is simply the value at the found index.

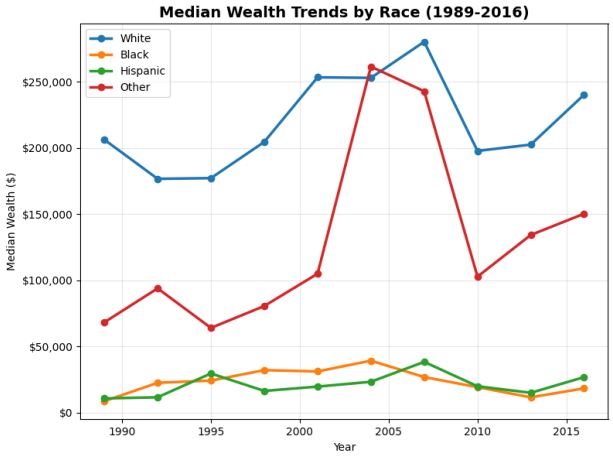
In other words, the weighted median is the value in the sorted data where the cumulative sum of weights first reaches at least half the total weight. This ensures that the weighted median reflects the distribution of the variable in the population, accounting for the importance (weight) of each observation.

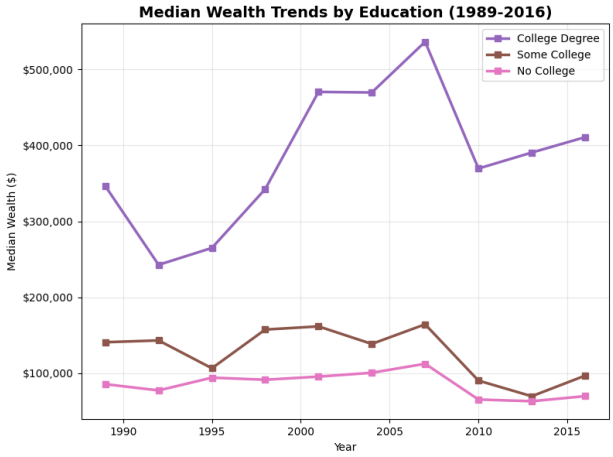
Mathematically, the weighted median m of a set of values $x_1, x_2, \ldots, x_n$ with corresponding non-negative weights $w_1, w_2, \ldots, w_n$ is defined as the value m such that:

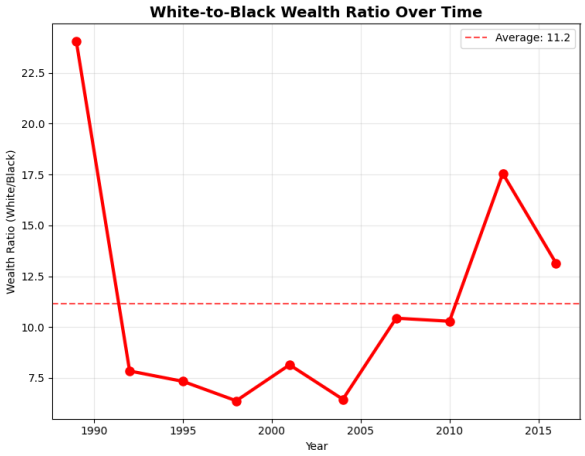
$$\sum_{i:x_i<m} w_i \leq \frac{1}{2}\sum_{i=1}^{n} w_i \quad \text{and} \quad \sum_{i:x_i>m} w_i \leq \frac{1}{2}\sum_{i=1}^{n} w_i$$

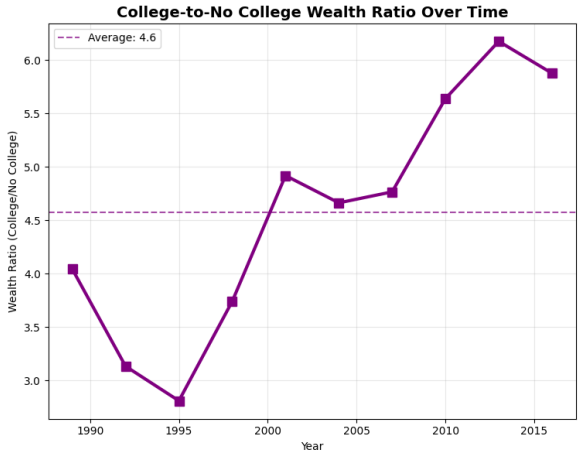
That is, m is the smallest value for which the cumulative sum of the weights of all values less than m is at most half the total weight, and the cumulative sum of the weights of all values greater than m is also at most half the total weight.

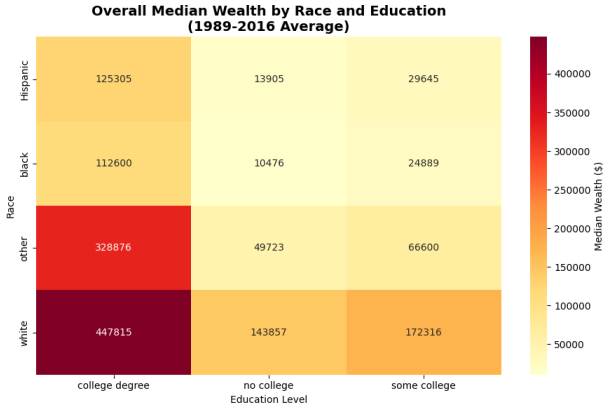
1 Please summarize key trends in median total wealth over the last 30 years by race and education using plots and in writing.

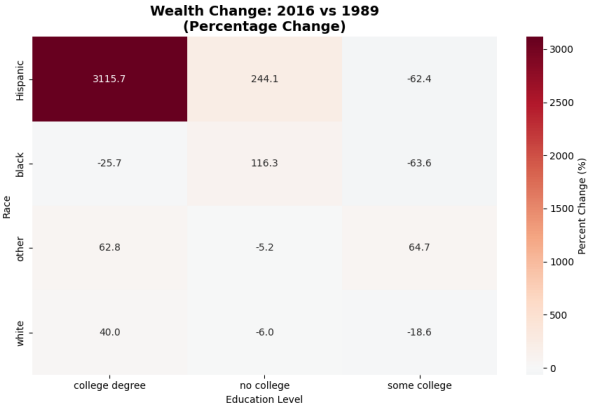












1.1 Race

- **White households** consistently maintain the highest median wealth
- **Black households** have the lowest median wealth throughout most years
- **Hispanic households** show similar patterns to Black households, with relatively

higher wealth in recent years
- **Other race households** show high volatility, with a dramatic spike in 2004–2007
- The **White-to-Black wealth ratio** averages around 11x over the entire period
- The gap was extreme in 1989 but narrowed considerably by the mid-1990s
- The ratio has fluctuated throughout the sample period.

1.2 Education

- **College degree holders** consistently have the highest median wealth.
- **Some college** group falls in the middle.
- **No college** group has the lowest wealth.
- The **College-to-No College ratio** averages around 4.6 throughout the sample period.
- This gap has **widened significantly** over time.
- The education premium peaked around 2010–2013.
- The education gap shows a trend toward expansion.

1.3 Comprehensive Wealth Analysis Summary (1989–2016)

Metric	Group	1989 Value	2016 Value	% Change (1989–2016)	CAGR	Notes
Median Wealth	White	$206,364	$240,350	+16.5%	+0.57%	Highest absolute wealth
	Black	$8,583	$18,300	+113.2%	+2.84%	Fastest growth rate
	Hispanic	$10,710	$26,800	+150.2%	+3.46%	Largest % increase
	Other	$68,234	$150,350	+120.3%	+2.97%	High volatility group
	College Degree	$346,490	$410,800	+18.6%	+0.63%	Highest absolute wealth
	Some College	$140,928	$96,905	-31.2%	-1.38%	Significant decline
	No College	$85,699	$69,921	-18.4%	-0.75%	Moderate decline
Wealth Gap Ratios	White-to-Black	24.0	13.1	-45.4%	—	Gap narrowed significantly

Metric	Group	1989 Value	2016 Value	% Change (1989–2016)	CAGR	Notes
	College-to-No College	4.0	5.9	+45.3%	—	Gap widened substantially
Wealth Volatility (CV)	White	—	—	—	—	16.1% (lowest)
	Black	—	—	—	—	40.4% (high)
	Hispanic	—	—	—	—	40.8% (high)
	Other	—	—	—	—	53.5% (highest)
	College Degree	—	—	—	—	23.9% (moderate)
	Some College	—	—	—	—	26.3% (moderate)
	No College	—	—	—	—	19.0% (low)
Financial Crisis Impact (2007-2010)	White	—	—	-29.4%	—	Moderate decline
	Black	—	—	-28.4%	—	Similar to White
	Hispanic	—	—	-48.1%	—	Severe impact
	Other	—	—	-57.7%	—	Most severe impact
	College Degree	—	—	-31.1%	—	Moderate decline
	Some College	—	—	-45.0%	—	Large decline
	No College	—	—	-41.8%	—	Significant decline

2 Repeat your analysis for just median housing wealth for black and white households

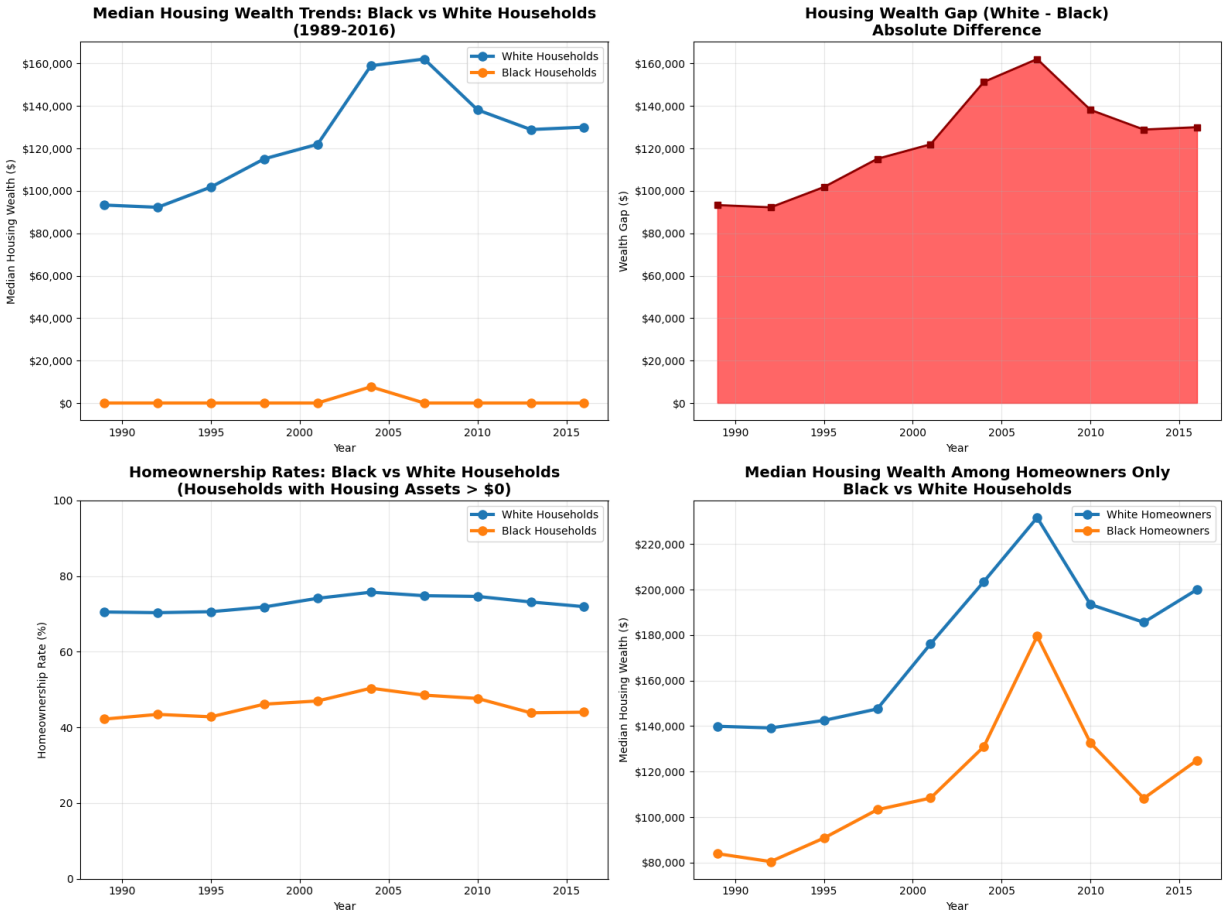


Figure 1: Median Housing Wealth Black vs White

Metric	Group	1989	2016	% Change	CAGR	Notes
Housing Wealth (All)	White	$93,293	$130,000	+39.4%	+1.24%	Substantial growth
	Black	$0	$0	0%	—	Positive in 1/10 years
Home ownership Rate	White	70.5%	71.9%	+1.9%	—	Avg: 72.8%
	Black	42.2%	44.0%	+4.3%	—	Avg: 45.6%

Metric	Group	1989	2016	% Change	CAGR	Notes
Ownership Gap	White-Black	28.3 pp	27.9 pp	-1.4%	—	Avg: 27.2 pp
Housing Wealth (Owners)	White	$139,940	$200,000	+42.9%	—	Among homeowners
	Black	$83,964	$125,000	+48.9%	—	Among homeowners
Wealth Ratio (Owners)	White/Black	1.7	1.6	-5.9%	—	Gap narrowed
Volatility (CV)	White	—	—	—	—	19.8%
	Black	—	—	—	—	25.7%

2.1 Findings

2.1.1 Stark housing wealth divide

- In 9 out of 10 survey years, more than half of Black households had zero median housing wealth.
- White households maintained substantial median housing wealth ($93K–$162K) throughout.

2.1.2 Persistent homeownership gap

- White households: average homeownership rate 72.8% (range: 70.5%–75.7%)
- Black households: average homeownership rate 45.6% (range: 42.2%–50.4%)
- The homeownership gap (27.2 percentage points) has remained stable for nearly three decades.
- This gap is a major barrier to wealth accumulation for Black families.

2.1.3 Housing wealth among homeowners

- Among homeowners, the racial gap narrows but persists.
- In 1989, White homeowners had 1.7× the housing wealth of Black homeowners; in 2016, this ratio was 1.6×.
- Both groups saw strong growth: +43% for White homeowners, +49% for Black homeowners.

2.1.4 Growth patterns

- White households experienced steady housing wealth growth (1.24% annually), peaking during the 2004–2007 housing boom.
- Black households showed a volatile pattern, with most years at zero median housing wealth.
- The 2008 financial crisis affected both groups, but White households recovered more fully.

2.1.5 Economic cycle impact

- During the 2001–2007 boom, both groups saw housing wealth increases.
- The 2007–2010 crisis brought sharp declines for both groups.

3 Many households are not homeowners and so your analysis for the prior...

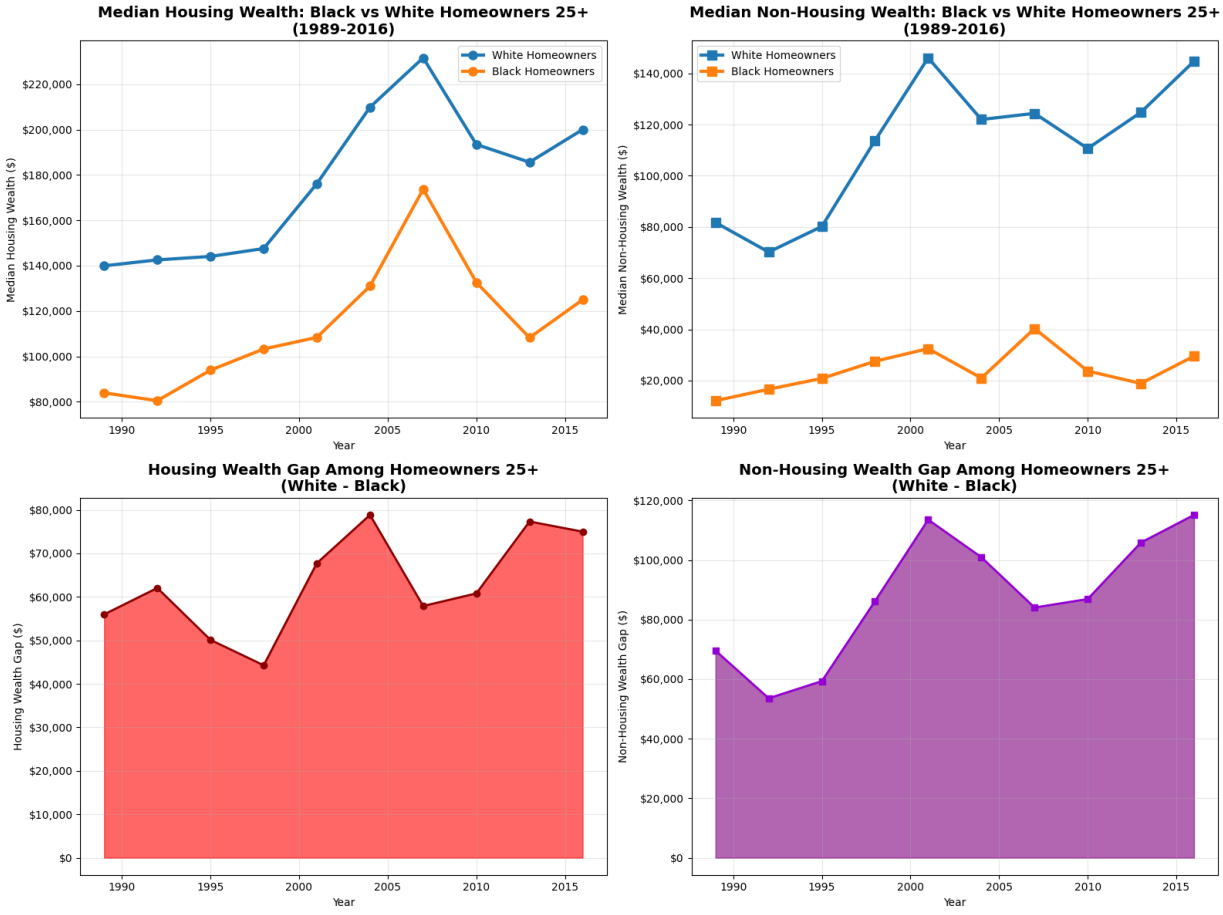


Figure 2: Financial Crisis Impact Homeowners 25+

3.1 Homeowners Aged 25+ Analysis Summary

Dataset: 33,292 homeowners aged 25+ (69.7% of total sample)
Race Distribution: White: 85.2%, Black: 6.3%, Hispanic: 4.7%, Other: 3.7%

3.1.1 Median Housing Wealth by Year (Homeowners 25+)

Year	Black	White	White/Black Ratio
1989	$83,964	$139,940	1.7
1992	$80,499	$142,551	1.8
1995	$93,960	$144,072	1.5
1998	$103,282	$147,546	1.4
2001	$108,388	$176,131	1.6
2004	$130,994	$209,844	1.6
2007	$173,702	$231,603	1.3
2010	$132,639	$193,433	1.5
2013	$108,266	$185,599	1.7
2016	$125,000	$200,000	1.6

3.1.2 Median Non-Housing Wealth by Year (Homeowners 25+)

Year	Black	White	White/Black Ratio
1989	$12,128	$81,725	6.7
1992	$16,603	$70,185	4.2
1995	$20,828	$80,179	3.8
1998	$27,517	$113,758	4.1
2001	$32,462	$146,053	4.5
2004	$20,984	$121,964	5.8
2007	$40,299	$124,371	3.1
2010	$23,720	$110,643	4.7
2013	$18,869	$124,764	6.6
2016	$29,540	$144,730	4.9

3.1.3 Financial Crisis Impact and Recovery Analysis

CategoryGroup	2007 Peak	2010 Crisis	Dollar Change	% Change	2016 Recovery	Recovery Rate	Volatility (CV)
Housing Wealth White	$231,603	$193,433	-$38,170	-16.5%	$200,000	86.4% (Partial)	18.3%
Black	$173,702	$132,639	-$41,063	-23.6%	$125,000	72.0% (Partial)	24.3%

Category	Group	2007 Peak	2010 Crisis	Dollar Change	% Change	2016 Recovery	Recovery Rate	Volatility (CV)
Non-Housing Wealth	White	$124,371	$110,643	-$13,727	-11.0%	$144,730	116.4% (Full)	23.7%
	Black	$40,299	$23,720	-$16,579	-41.1%	$29,540	73.3% (Partial)	34.2%

3.1.4 Long-Term Growth and Composition Analysis (1989-2016)

Metric	Group	Housing Wealth	Non-Housing Wealth
CAGR	White	1.33%	2.14%
	Black	1.48%	3.35%
Total Growth	White	42.9%	77.1%
	Black	48.9%	143.6%
1989 Composition	White	63.1% housing	36.9% non-housing
	Black	87.4% housing	12.6% non-housing
2016 Composition	White	58.0% housing	42.0% non-housing
	Black	80.9% housing	19.1% non-housing

3.1.5 Findings

3.1.5.1 Median Housing Wealth:

Both Black and White homeowners saw growth from 1989 to 2007, peaking in 2007.

- 2007 Peak: White: $231,603 | Black: $173,702
- 2010 Crisis: White: $193,433 | Black: $132,639
- 2016 Recovery: White: $200,000 | Black: $125,000

3.1.5.2 Median Non-Housing Wealth:

Both groups saw non-housing wealth peak in 2007, decline in 2010, and partial recovery by 2016.

- 2007 Peak: White: $124,371 | Black: $40,299
- 2010 Crisis: White: $110,643 | Black: $23,720
- 2016 Recovery: White: $144,730 | Black: $29,540

3.1.5.3 Loss in Housing Wealth (2007)

Dollar Terms:

- White: $231,603 → $193,433 = –$38,170
- Black: $173,702 → $132,639 = –$41,063
- Black homeowners had the largest dollar loss ($41,063 vs. $38,170).

Proportional Terms (% Loss):

- White: –16.5%
- Black: –23.6%
- Black homeowners also had the largest proportional loss.

3.1.5.4 Summary

- Both Black and White homeowners aged 25+ experienced significant declines in housing wealth during the financial crisis (2007–2010).
- Black homeowners had the largest loss in both dollar terms and proportional terms.
- By 2016, neither group had fully recovered to 2007 levels, but White homeowners recovered a greater share of their losses.

4 Many potential channels have been identified for explaining the wealth

4.1 Hypothesis 1: Workplace Income Discrimination

One key mechanism by which discrimination widens racial wealth gaps is through **income labor discrimination**. Persistent disparities in income make it more difficult for Black households to accumulate assets or manage debt as reliably as White households.

4.1.1 Longitudinal Analysis of Income Variation

- Use longitudinal datasets such as the *Panel Study of Income Dynamics (PSID)* to measure year-to-year income and income by race.
- Estimate income outcome coefficients, controlling for education, occupation, experience, sex, and location.
- This approach isolates the contribution of discrimination to income differences, above and beyond observable characteristics.

4.1.2 Event Study Strategy

- Compare firms, industries, or states with **inclusionary policies** (e.g., diversity initiatives, pay transparency laws) to those with **exclusionary practices** (e.g., documented discrimination cases).
- Include a **neutral group** of firms or states as a control.
- Estimate effects on income controlling for race. A triple-difference specification,

$$(\text{Inclusionary} - \text{Exclusionary} - \text{Control}),$$

provides stronger identification, while a simpler difference (Exclusionary vs. Non-exclusionary) serves as a robustness check.

4.1.3 Expected Contribution

By focusing on income β coefficient across different policy environments, this approach highlights how workplace discrimination affects income. The key test is whether inclusive versus exclusionary environments produce systematically different income patterns across racial groups.

4.2 Hypothesis 2: Disparities in the Transmission of Investment Knowledge

Another important mechanism sustaining the racial wealth gap is **unequal transmission of investment knowledge**. Even among households with similar incomes and access to financial products, Black parents may be less likely to transmit financial knowledge—such as saving habits or familiarity with financial instruments like 401(k)s—to their children. These differences in financial literacy and exposure to sound financial advice can generate divergent wealth trajectories.

4.2.1 Intergenerational Transmission of Financial Knowledge

- Construct a dataset linking parents and children (e.g., using the PSID or other intergenerational surveys).
- Use indicators of parental financial sophistication (e.g., whether parents have a 401(k), stock holdings, or savings rate).
- Test whether parental financial sophistication predicts children's financial sophistication.
- Estimate a triple-difference specification:

$$\begin{aligned}&\Big(\text{Black w/ financially active parents}\\&\quad-\text{Black w/o financially active parents}\Big)\\&-\Big(\text{White w/ financially active parents}\\&\quad-\text{White w/o financially active parents}\Big)\end{aligned}$$

- This identifies whether financial knowledge is transmitted differently across racial groups, which could be a driving force behind the wealth gap.
- Additionally, examine whether the savings rate of parents is similarly transmitted to children, controlling for race.

4.2.2 Neighborhood Transmission of Financial Knowledge

This analysis can be extended to the neighborhood level. Children may acquire financial knowledge not only from parents but also through neighborhood connections. - Measure neighborhood financial sophistication (e.g., share of households with retirement accounts, stock ownership, or savings rate). - Test whether Black individuals in financially sophisticated neighborhoods experience high levels of financial sophistication. This may be an important observation, given that white-majority, financially sophisticated neighborhoods

may not necessarily transmit financial sophistication to minority Black members of the neighborhood. - Compare whether these neighborhood effects operate equally for Black and White households, using the same triple-difference estimator as above.

4.2.3 Expected Contribution

By separately identifying family-based and neighborhood-based channels, and testing whether these operate differently by race, we can assess how much of the wealth gap is driven by **differences in transmission of financial knowledge** that affect asset holdings.

4.3 Assessing the Importance of Each Channel

To evaluate the importance of these channels, we can extract the coefficients from our income labor discrimination estimates. This would give us a dollar amount that we can use to evaluate the magnitude of our estimates and their relation to overall wealth. These coefficients would also be helpful in comparing against other estimates academics have produced.

Next, with regard to the financial sophistication transmission channel, we can use these estimates to obtain coefficients that measure the financial sophistication associated with neighbor or family transmission. To assess the importance of these, we would create an additional test examining whether individuals with greater financial sophistication tend to have higher asset holdings. This can be done using a simple regression, controlling for age, race, gender, etc., with financial sophistication as the independent variable and assets as the dependent variable. The coefficient from this regression would identify asset increases associated with higher financial sophistication. This provides a dollar estimate to assess the magnitude and importance of financial literacy, and thus the significance of the financial sophistication transmission channel. Furthermore, this is a well-studied area, so other academic papers can be referenced to benchmark the magnitude and importance of financial sophistication for asset accumulation.

Finally, we can compare our estimates to findings from other academic studies, while also checking for statistical significance and robustness.