Housing Discrimination in Boston

Edward Chang, Hermish Mehta, Ishaan Srivastava
University of California, Berkeley

December 18, 2020

1 Introduction

1.1 Context & Research Question

Housing discrimination has had a lasting legacy in American history, drastically shaping the United States' (US) development from the abolition of slavery to this day. Many discriminatory practices in America trace their roots to the Southern Jim Crow laws introduced at the end of the Civil War. [3] That being said, policies like exclusionary zoning led to housing discrimination and residential segregation not just in the South, but across the country. During this era, the Federal Housing Administration played a critical role in institutionalizing redlining and other explicitly discriminatory policies.

It took until the Civil Rights Act passed in the mid-twentieth century before the United States government finally deemed housing discrimination unconstitutional. Specifically, the Fair Housing Act in 1968 explicitly prohibited discrimination based on race, making racial steering, blockbusting, and redlining, all once standard practices, illegal. While this represented significant progress toward just housing policies, it did little to reverse decades of systemic discrimination: a wealth of literature suggests that housing discrimination remains a significant problem in America, leading to segregated neighborhoods and considerable wealth and educational disparities.

In this paper, we explore how the legacy of racial discrimination persists in modern housing markets. In particular, we attempt to understand whether and how racial composition affects free-market property values, holding other potentially relevant factors constant. In so doing, we challenge the common thesis that the free markets function as a race-blind instrument that automatically counters racial segregation. [1]

1.2 Data Description

Since housing markets vary drastically between cities due to different local and state regulations, we concentrate on a single cosmopolitan city: Boston, MA. We restrict ourselves to examining how race plays a role in determining house values in the city and its suburbs. To that end, we consider the Boston Housing Dataset, which contains information collected by the US Census Service in 1970. The dataset contains 506 observations of 14 variables; each observation corresponds to a neighborhood, with information about statistics such as per-capita crime rate or pupil-teacher ratio.

A detailed specification of each variable is given in Table 1, along with complete descriptions of each feature. Here, we treat owner-occupied homes' median value as the response variable, while the remaining variables act as potential explanatory variables.

Table 1. A detailed summary of each variable in the Boston Housing Dataset. Note that the quartile variable was not originally included in the data, but instead was computed using the segregation variable; this procedure is explained further in Section 2. The radial highway accessibility index variable is an ordinal variable that takes integral values from 1 through 24.

Variable	Code	Type	Description
Crime Rate	`crim`	Continuous	Per-capita crime rate by town.
Zoning	`zn`	Continuous	Proportion of residential land zoned for lots over 25,000 square feet.
Industry	`indus`	Continuous	Proportion of non-retail business acres per town.
Charles River	`chas`	Categorical	Indicator if tract borders Charles River.
Oxides	`nox`	Continuous	Nitrogen oxides concentration (parts per 10 million).
Rooms	`rm`	Continuous	Average number of rooms per dwelling.
Age	`age`	Continuous	Proportion of owner-occupied units built prior to 1940.
Distance	`dis`	Continuous	Weighted mean of distances to five Boston employment centers.
Highways	`rad`	Ordinal	Index of accessibility to radial highways.
Tax	`tax`	Continuous	Full-value property-tax rate per $10,000.
Pupil Ratio	`ptratio`	Continuous	Pupil-teacher ratio by town.
Quartile	`quart`	Ordinal	African-American proportion as a quartile; 4 corresponds to towns with the greatest proportions.
Homogeneity	`black`	Continuous	$1000(B - 0.63)^2$ where B is the proportion of African-Americans by town.
Status	`lstat`	Continuous	Lower status of the population (percent).
Median Value	`medv`	Continuous	Median value of owner-occupied homes in $1000s.

1.3 Overview

We start by understanding both variables which carry information about each neighborhood's racial composition through exploratory data analysis. We then construct two separate models, with and without race information, to predict median owner-occupied home values. Comparing these models, we attempt to estimate the effect of race on property values after controlling for other relevant variables. To avoid relying on potentially dubious distributional assumptions, we then pinpoint this estimate's uncertainty using the non-parametric bootstrap; this allows us to directly test the hypothesis that race plays no effect in determining median home values. Finally, we comment on the assumptions made in our analysis and discuss potential avenues to draw more robust causal inference type conclusions.

2 Race Data

2.1 Feature Construction

As noted in the data description, the dataset did not directly include the proportion of African-Americans in each neighborhood. Instead, it included the variable `black`, which is defined as

$$1000(B - 0.63)^2, \tag{1}$$

where B is the true recorded proportion of African Americans. Unfortunately, B cannot be perfectly recovered from the values of this variable; for example, `black` would be equal to 10 when either $B = 0.73$

or $B = 0.53$. In general, since the function is quadratic, B cannot be uniquely determined for any value of `black` below 136.9. We can nevertheless still recover some useful information about these proportions. In particular, we can still sort each neighborhood into quantiles based on its proportion of African-Americans

When `black` is below 136.9, we may not know the exact proportion of African-Americans in the neighborhood, but we can at least lower bound this at 0.26 since the absolute value of its difference from 0.63 is less than 0.37. Fortunately, this is only the case for 8% of the neighborhoods in Boston; the remaining 92% of neighborhoods have `black` greater than 136.9 and thus B can be recovered exactly. Rather than throw away the points with the highest proportions of African-Americans, we instead opt to bin these proportions. In particular, by choosing to use quartiles, all the neighborhoods with unknown proportions will comfortably fall in the fourth quartile, and therefore, there is no ambiguity.

Note the same reasoning would hold had we chosen to use quintiles or even deciles. However, looking at the data, roughly 23% have a reported proportion of 0%. To avoid having highly asymmetric quantiles or randomly assigning observations to bins, we settle on quartiles, so all neighborhoods with the same proportion lie in the same bin.

2.2 Exploratory Data Analysis

In this paper, we want to understand the effect of race, controlling for other variables. Naturally, it is crucial to understand whether the race variables included indeed provide any information not already captured by the other variables, at least in the regression context. To that end, we first attempt to regress `black` against all other variables, excluding the response variable and `quart`. Here, we find the resulting model has $R^2 = 0.258$, which suggests that while many variables are correlated with `black`, together they still only explain a small part of its variance.

We might hope to run a similar sanity check for the `quart` variable. A natural approach is to include `quart` as a categorical variable with three dummy variables. We therefore attempt to predict each of these dummies with the remaining variables, excluding median house price and `black`. More specifically, we take indicators for `quart` equal to 2, 3 and 4 as our response variables. The resulting R^2 values are 0.072, 0.080 and 0.236 respectively.[1]

Since these variables will be a large focus of our research question, we start by examining the univariate distributions of these variables and their relationship to the median owner-occupied home value. These displays can be found in Figure 1 through Figure 4. Below, we summarize some observations about these plots.

- Figure 1 shows a histogram of the homogeneity variable. The values are highly concentrated near 400, and the distribution has a strong left skew. Note that the maximum value of 396.9 corresponds to an African-American proportion of 0%; this means a large number of the neighborhoods simply had no African-Americans, at least at the granularity measured. This distribution is not surprising, given we previously discovered nearly 92% of all neighborhoods had $B \leq 0.26$.
- In Figure 2, we plot the histogram of the non-zero recovered proportions. While the variable included in our model is the quartile of each value, we plot the log-odds to visualize the distribution over a large range of values: the proportions themselves lie between 0 and 1. Note we are missing the upper tail of the distribution since the proportions above 0.26 simply could not be recovered.

Finally, the last set of four figures examine the relationship between our response variable `medv` and our two explanatory variables of interest. Notice in Figure 4, `medv` increases from quartiles one through three before dropping at quartile four. This is consistent with the weak positive relationship between `black` and `medv`; property values tend to be higher in neighborhoods with relatively high or low proportions of African-Americans. This relationship will be a focus of study throughout the remainder of the paper.

[1] Running the same regressions using the transformed variables in the final model, we also obtain similar results. In order, the R^2 values for `black`, `quart2`, `quart3` and `quart4` are 0.256, 0.075, 0.089 and 0.233.

Figure 1. Distribution of homogeneity.

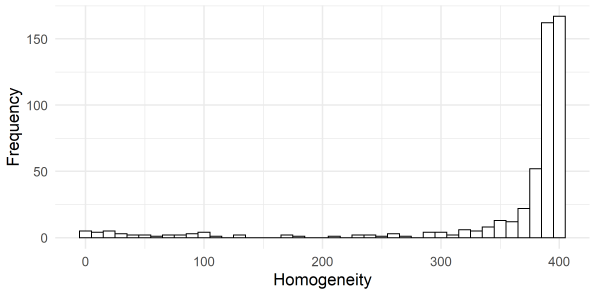


Figure 2. Distribution of proportions.

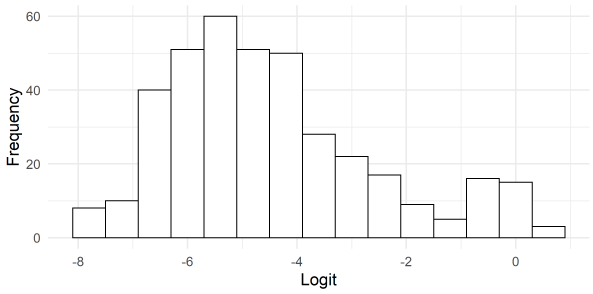


Figure 3. `medv` versus `black`.

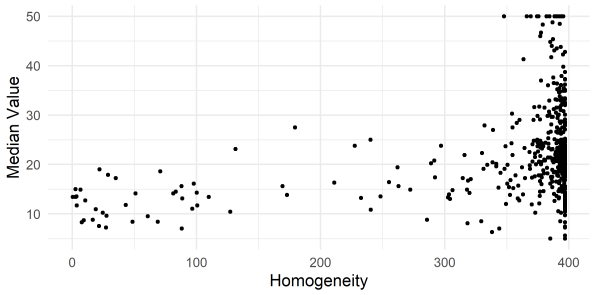


Figure 4. `medv` per `quart` level.

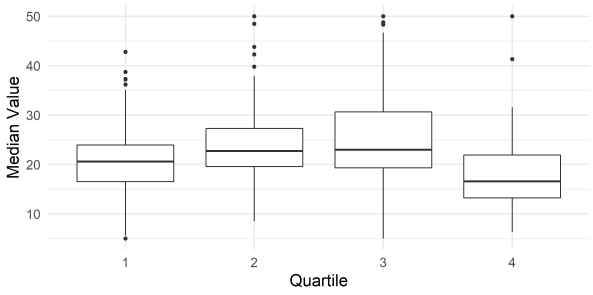


Figure 5. log(`medv`) versus `black`.

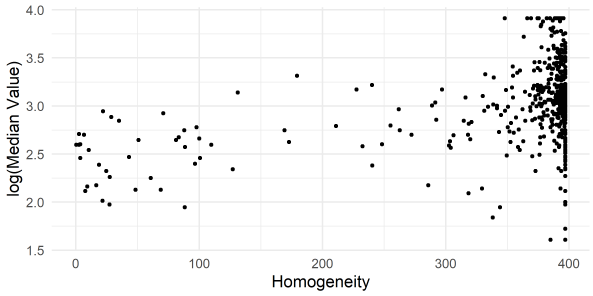
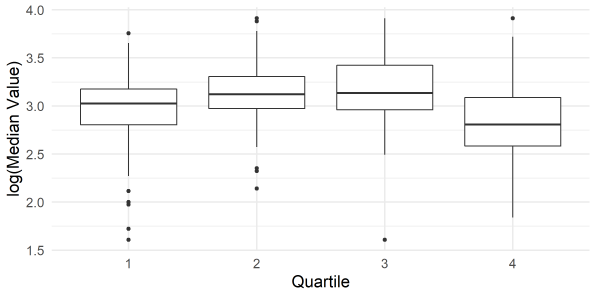


Figure 6. log(`medv`) per `quart` level.



3 Models

The previous section's exploratory data analysis suggests an association between the variables `black`, `quart` and `medv`. However, in the context of answering our research question, this relationship is relatively uninteresting: the univariate relationships between these race-based variables and the response could well be a function of confounding variables. Therefore, to draw causal conclusions, we need to understand the effect of race holding all other variables constant. This would directly answer the question of how race affects the market value of homes.

Given the observational data available, however, such analysis is impossible. Instead, we consider a reasonable approximation of understanding the effect of race on house price holding as many relevant confounds as possible constant. While this might not allow us to make strong statements about causality, we can look for a relationship between race and home values that other variables cannot easily explain. In this

setting, we are implicitly comparing a complete model that includes race information to a baseline model that omits it. To construct an appropriate null model is extremely difficult since choices of which variables to include may well influence the resulting analyses' conclusions.

Hence, we turn to the econometrics paper by Harrison and Rubinfeld [2], which first introduces this dataset. There, the authors provide a structural equation for their model specification. Choosing a simple linear power transformation for the `nox` coefficient and removing the `black` variable, we arrive at the following "race-blind" baseline model.

Model 1

$$\begin{aligned}\log(\texttt{medv}) = \beta_0 &+ \beta_1 \texttt{rm}^2 + \beta_2 \texttt{age} + \beta_3 \log(\texttt{dis}) + \beta_4 \log(\texttt{rad}) + \beta_5 \texttt{tax} + \beta_6 \texttt{ptratio} \\ &+ \beta_7 \log(\texttt{lstat}) + \beta_8 \texttt{crim} + \beta_9 \texttt{zn} + \beta_{10} \texttt{indus} + \beta_{11} \texttt{chas} + \beta_{12} \texttt{nox} + \epsilon\end{aligned} \tag{2}$$

We then derive our second model by adding in the `black` and `quart` variables. This larger model now explicitly accounts for each neighborhood's racial composition by considering its homogeneity and the proportion of African-American residents. Notice that the categorical variable `quart` is included as three dummy variables corresponding to indicators of when `quart` is 2, 3 or 4; the baseline category is the lowest quantile. This choice seems natural, given the difference in means across quartiles seen in Figure 4. The resulting alternate model is given below.

Model 2

$$\begin{aligned}\log(\texttt{medv}) = \beta_0 &+ \beta_1 \texttt{rm}^2 + \beta_2 \texttt{age} + \beta_3 \log(\texttt{dis}) + \beta_4 \log(\texttt{rad}) + \beta_5 \texttt{tax} + \beta_6 \texttt{ptratio} \\ &+ \beta_7 \log(\texttt{lstat}) + \beta_8 \texttt{crim} + \beta_9 \texttt{zn} + \beta_{10} \texttt{indus} + \beta_{11} \texttt{chas} + \beta_{12} \texttt{nox} \\ &+ \beta_{13} \texttt{black} + \beta_{14} \texttt{quart}_2 + \beta_{15} \texttt{quart}_3 + \beta_{16} \texttt{quart}_4 + \epsilon\end{aligned} \tag{3}$$

Full outputs of fitting both Model 1 and Model 2 are provided in Appendix A. We want to understand whether Model 2 fits the data better than Model 1. Equivalently, we can ask whether adding `black` and `quart` significantly improves the baseline model: do `black` and `quart` affect median house prices, holding all the variables in Equation 2 constant? Heuristically, we can try to answer this question with an F-test comparing the two models.

Table 2. Analysis of variance table comparing both models. Note the difference in degrees of freedom between the two models is exactly 4, which corresponds to the 1 `black` variable and 3 dummies added by `quart`.

Source	Res. df	RSS	df	Sum of Sq.	F	P (>F)
Model 1	493	16.832				
Model 2	489	16.029	4	0.80345	6.1279	8.114×10^{-5}

This value is extremely significant, suggesting that the relationship between race and median home prices persist even after controlling for all variables in the baseline model. However, to interpret this value as rejecting a null hypothesis that the coefficients on the added variables are zero would require verifying the assumptions of the F-test hold.

One approach might be to look at the regression diagnostics, given in Figure 7 and Figure 8. From Figure 7, it appears that there is no clear non-linear relationship across the residuals. There are only a few high-leverage points, and these do not have considerably large Cook's distances. Unfortunately, however, the normality assumption appears questionable: the Q-Q plot suggests the distribution has heavier tails than expected. Furthermore, the trend in the scale-location plot hinting at heteroskedasticity might also pose a problem. Figure 8 is extremely similar to Figure 7 and so the diagnostics for Model 1 look very similar to that of Model 2.

Figure 7. Regression diagnostics for Model 1.

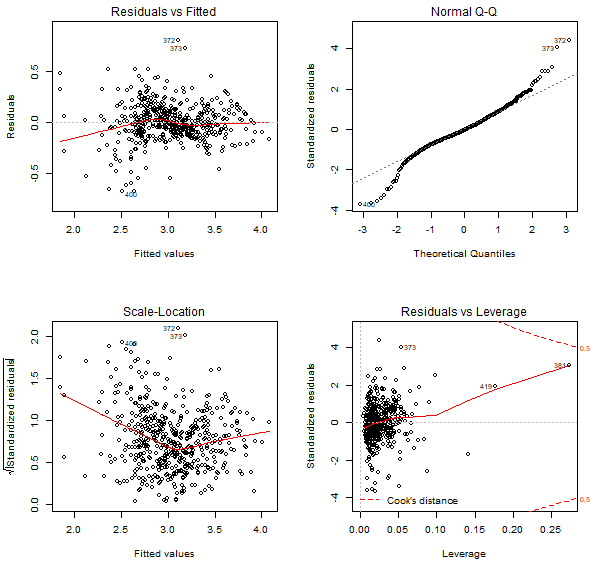
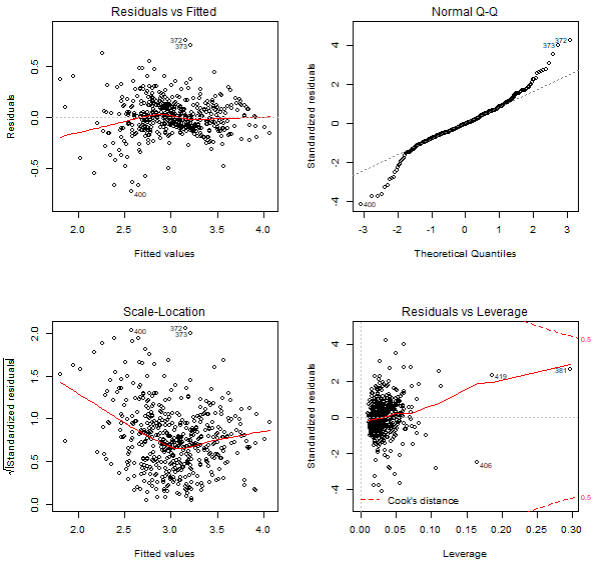


Figure 8. Regression diagnostics for Model 2.



4 Inference & Analysis

To summarize, the results in the previous section suggest the fuller model with race information (Model 2) better predicts median house prices over the baseline model (Model 1). However, rejecting the null hypothesis that the coefficients on race variables `black` and `quart` are zero with the F-test is dubious since this makes strong assumptions about the underlying conditional distribution of the response. As alluded to by Figure 8, there is no reason to believe this distribution is normal. House prices ultimately arise from a complex interaction of many factors, and either model likely does not capture this relationship exactly up to independent normal errors.

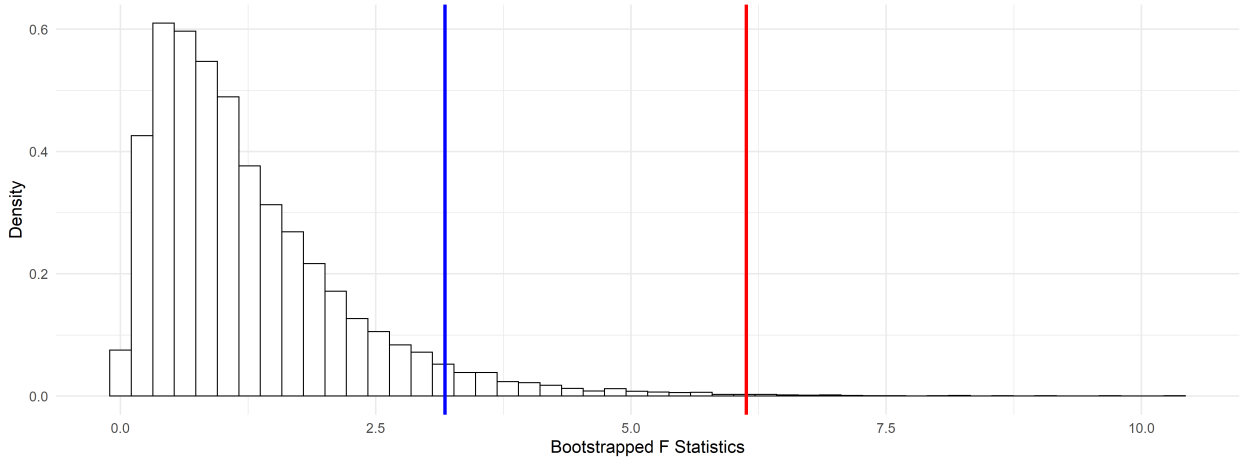
Nonetheless, we still ideally want to determine whether the difference in performance between the two models is significant. We, therefore, turn to the non-parametric bootstrap to do inference. First, we perform a bootstrap F-test; formally, we define the following test statistic for the larger model, Model 2:

$$F = \frac{\frac{1}{k}\left\{\left(L\hat{\beta} - 0\right)' \left[L\left(X'X\right)^{-1}L\right]^{-1}\left(L\hat{\beta} - 0\right)\right\}}{\frac{1}{n-p-1}\,\mathrm{ErrSS}_{\mathrm{full}}}. \tag{4}$$

Here, $n = 504$ and $p = 14$ represent the number of observations and variables respectively; $k = 4$ is the number of coefficients we are interested in. Similarly, L is the $k \times (p+1)$ matrix which extracts the coordinates of the $\hat{\beta}$ corresponding to the added variables `black` and `quart`. In Equation 3, these correspond to estimates of β_{13} through β_{16} respectively. Qualitatively, we should expect this value to be large under an alternative hypothesis where any of these coefficients is non-zero. From the analysis of variance table above, we read off the value of this statistic as 6.2114.

To test for the significance of this value, we compare it to the approximate sampling distribution of the statistic under the null hypothesis given by the bootstrap. Figure 9 shows the observed statistic superimposed on its estimated sampling distribution. The bootstrap p-value is 0.003, which is significant at the $\alpha = 0.05$ level. This allows us to strongly reject the null hypothesis that race does not have an association with median home values ($\beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0$), controlling for the other variables given in Equation 3.

Figure 9. Approximate distribution of the F-statistic under the null hypothesis generated by the non-parametric bootstrap with $10,000$ replicates. Vertical lines are drawn at the 0.95 quantile of the sampling distribution (blue) and the actual value observed in the data (red).



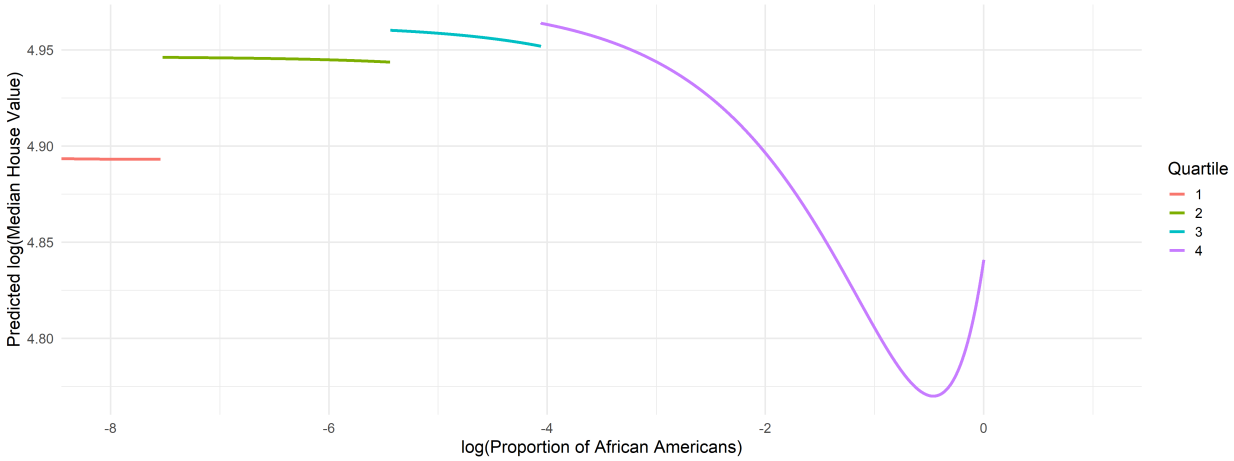
The larger model contains exactly 4 additional variables: one which corresponds to the homogeneity of each neighborhood's racial composition and three additional dummy variables characterizing which quartile it lies in depending on its proportion of African-American residents. By including all these variables, we were able to conclude that race appears to affect house prices even after controlling for various potentially relevant confounds. To estimate the coefficient on each of these individual variables, we can generate 95% confidence intervals based on the non-parametric bootstrap.

The results, shown in Table 3 allow us to understand the regression surface defined by these race-based variables and gauge whether each of these individual variables might be significant. In particular, it appears the coefficient on homogeneity is significant, while those corresponding to `quart` are not, at least at the $\alpha = 0.05$ significance level. This suggests that of variables added to the "race-blind" baseline model, `black` most improved the model's fit. In other words, median home prices are lower in mixed-race neighborhoods; this matches with the univariate pattern initially observed in the data.

Table 3. Bootstrap-based 95% confidence intervals for each coefficient generated using $10,000$ replicates. Each interval is given along with the corresponding point estimate.

Coefficient	Lower	Point Estimate	Upper
`black`	1.299×10^{-4}	5.165×10^{-4}	8.850×10^{-4}
`quart2`	-1.677×10^{-2}	5.302×10^{-2}	1.195×10^{-1}
`quart3`	-1.368×10^{-3}	6.958×10^{-2}	1.372×10^{-1}
`quart4`	-8.696×10^{-3}	8.152×10^{-2}	1.675×10^{-1}

Figure 10. Relative prediction of log(`medv`) based on African-American proportion, holding all other variables constant. Note the parabolic relationship arises from the `black` variable, which is defined as quadratic function of the proportion of African-Americans.



Finally, given the scope of our research question, we comment on our analysis from a causal inference paradigm. If we appropriately accounted for all relevant confounding variables in Model 2, we could claim that changing a neighborhood's racial composition, ceteris paribus, would affect house prices. Based on the specific results above, such a statement would directly imply interventions that increase neighborhood diversity would be expected to cause property values to decrease. However, as we discussed earlier, this is a tough sell given the available.

It is impossible to be sure that all relevant confounding variables have been appropriately controlled for. For example, the association between `black` and median home prices observed in the data could simply be a product of an additional variable entirely unrelated to race. Even worse, such omitted variable bias cannot even be investigated using only the data itself. Our willingness to believe a causal interpretation of our results therefore depends on our belief about how well Equation 2 captures all possible relevant confounds. Unfortunately, this is par for the course given the nature of the observational data.

5 Conclusion

This report began with a discussion of the pervasiveness of institutional racism in the American housing market. Through statistically sound analyses, paying attention to underlying model assumptions, we found that the proportion of African-Americans living in a neighborhood had a statistically significant impact on its median house price, controlling for potentially relevant confounding variables. Rather than a simple monotonic relation, we discovered that both lower and higher proportions of African-Americans were associated with higher median neighborhood prices. In other words, mixed-race neighborhoods had lower property values. These findings challenge the notion that the free market ignores racial bases.

That being said, our results ultimately fall short of making any causal claim, given the data available. Possible follow-up work to address this question could involve studying the effect of interventions designed to desegregate neighborhoods on housing prices. One open question, which is particularly relevant for policymakers, is whether these patterns even still exist today.

References

[1] Friedman, M. (1975). *Capitalism and freedom.* Chicago: Univ. of Chicago Press.

[2] Harrison, David & Rubinfeld, Daniel. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management.* 5. 81-102. 10.1016/0095-0696(78)90006-2.

[3] Zonta, M. (2019, July 15). Racial Disparities in Home Appreciation. Retrieved December 18, 2020, from https://www.americanprogress.org/issues/economy/reports/2019/07/15/469838/racial-disparities-home-appreciation/

6 Appendices

6.1 Appendix A: Model 1 and Model 2 Output

Table 4

	Dependent variable:	
	log(medv)	
	Model 1 Estimates (SE)	Model 2 Estimates (SE)
cri	-0.013^{***} (0.001)	-0.011^{***} (0.001)
zn	-0.0001 (0.001)	0.00003 (0.001)
indus	-0.00002 (0.002)	-0.0002 (0.002)
chas	0.093^{***} (0.034)	0.080^{**} (0.033)
nox	-0.893^{***} (0.154)	-0.845^{***} (0.151)
age	0.0003 (0.001)	0.0001 (0.001)
tax	-0.0004^{***} (0.0001)	-0.0004^{***} (0.0001)
ptratio	-0.030^{***} (0.005)	-0.028^{***} (0.005)
black		0.001^{***} (0.0001)
quart2		0.053^{**} (0.023)
quart3		0.070^{***} (0.024)
quart4		0.082^{***} (0.030)
rm^2	0.006^{***} (0.001)	0.006^{***} (0.001)
log(dis)	-0.205^{***} (0.034)	-0.211^{***} (0.034)
log(rad)	0.095^{***} (0.019)	0.098^{***} (0.019)
log(lstat)	-0.385^{***} (0.025)	-0.368^{***} (0.025)
Constant	5.020^{***} (0.168)	4.688^{***} (0.183)
Observations	506	506
R^2	0.801	0.810
Adjusted R^2	0.796	0.804
Residual Std. Error	0.185 (df = 493)	0.181 (df = 489)
F Statistic	164.862^{***} (df = 12; 493)	130.323^{***} (df = 16; 489)
Note:		$^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

6.2 Appendix B: Code

```
# for general data cleaning ease
library(tidyverse)
# for bootstrapping
library(boot)

## DATA CLEANING

# save data as boston
boston <- MASS::Boston

# attempt to recover black proportion from data
inter <- (sqrt(boston$black/1000) - 0.63) *-1
# categorize data into quartiles of recovered black proportion
boston <- mutate(boston,
                 quart = as.factor(ntile(inter, n = 4)))

## 2. RACE DATA

# make a data frame copy with quart modified to be numerical
boston.copy = mutate(boston, quart = as.integer(quart))

# attempt to regress black on all vars except for
# quart and medv
model = lm(black ~ . -quart -medv, data = boston.copy)
summary(model)

# attempt to regress quart2 on all vars except for
# black, quart, and medv
boston.copy$quart2 = boston.copy$quart == 2
model2 = lm(quart2 ~ . -black - quart - medv, boston.copy)
summary(model2)

# attempt to regress quart3 on all vars except for
# black, quart, quart2, and medv
boston.copy$quart3 = boston.copy$quart == 3
model3 = lm(quart3 ~ . -black - quart - medv -quart2, boston.copy)
summary(model3)

# attempt to regress quart3 on all vars except for
# black, quart, quart2, quart3, and medv
boston.copy$quart4 = boston.copy$quart == 4
model4 = lm(quart4 ~ . -black -quart -medv -quart2 -quart3, boston.copy)
summary(model4)

# drop quart columns and "reset" data
boston.copy = select(boston.copy, !c(quart2, quart3, quart4))

# Using transformed data in accordance with Harrison and Rubinfeld (1976) model.
# More mentioned elsewhere in report
```

```
boston.copy <- boston %>%
    mutate(sq.rm = rm^2,
           log.dis = log(dis),
           log.rad = log(rad),
           log.lstat = log(lstat),
           log.medv = log(medv)) %>%
    select(!c(rm, dis, rad, lstat, medv))

# attempt to regress black on all vars except for
# quart and log.medv
model1 = lm(black ~ .-log.medv - quart, boston.copy)
summary(model1)

# attempt to regress quart2 on all vars except for
# black, quart, and medv
boston.copy$quart2 = boston.copy$quart == 2
model2 = lm(quart2 ~ . -black - quart - log.medv, boston.copy)
summary(model2)

# attempt to regress quart3 on all vars except for
# black, quart, quart2, and medv
boston.copy$quart3 = boston.copy$quart == 3
model3 = lm(quart3 ~ . -black - quart - log.medv -quart2, boston.copy)
summary(model3)

# attempt to regress quart3 on all vars except for
# black, quart, quart2, quart3, and medv
boston.copy$quart4 = boston.copy$quart == 4
model4 = lm(quart4 ~ . -black -quart -log.medv -quart2 -quart3, boston.copy)
summary(model4)

## EDA FIGURES

# ggplot object for boston data set w/theming
boston.plot <- ggplot(boston) + theme_minimal(base_size = 16)

# Figure 1. Histogram of Homogeneity
eda1 <- boston.plot +
    geom_histogram(aes(x = black),
                   binwidth = 10, color="black", fill="white") +
    labs(x = 'Homogeneity', y = 'Frequency')
ggsave('eda1.png', plot = eda1, width = 8, height = 4)

# Figure 2. Histogram of logit(proportion)
eda2 <- boston.plot +
    geom_histogram(aes(x = log(inter/(1-inter))),
                   binwidth = 0.6, color="black", fill="white") +
    labs(x = 'Logit', y = 'Frequency')
ggsave("eda2.png", plot = eda2, width = 8, height = 4)

# Figure 3. Scatterplot of Median Value vs. Homogeneity
```

```
eda3 <- boston.plot +
    geom_point(aes(x = black, y = medv)) +
    labs(x = 'Homogeneity', y = 'Median Value')
ggsave("eda3.png", plot = eda3, width = 8, height = 4)

# Figure 4. Boxplot of Median Values per quart
eda4 <- boston.plot +
    geom_boxplot(aes(x = quart, y = medv)) +
    labs(x = 'Quartile', y = 'Median Value')
ggsave("eda4.png", plot = eda4, width = 8, height = 4)

# Figure 5. Scatterplot of log(Median Value) vs. Homogeneity
eda5 <- boston.plot +
    geom_point(aes(x = black, y = log(medv))) +
    labs(x = 'Homogeneity', y = 'log(Median Value)')
ggsave("eda5.png", plot = eda5, width = 8, height = 4)

# Figure 6. Boxplot of log(Median Values) per quart
eda6 <- boston.plot +
    geom_boxplot(aes(x = quart, y = log(medv))) +
    labs(x = 'Quartile', y = 'log(Median Value)')
ggsave("eda6.png", plot = eda6, width = 8, height = 4)

## 3. MODELS

# transform based on Harrison and Rubinfeld (1976) model
boston <- boston %>%
    mutate(sq.rm = rm^2,
           log.dis = log(dis),
           log.rad = log(rad),
           log.lstat = log(lstat),
           log.medv = log(medv)) %>%
    select(!c(rm, dis, rad, lstat, medv))

# regress log(medv) on all vars except for racial data (MODEL 2)
noblack.mod <- lm(log.medv ~ . - black - quart, boston)

# regress log(medv) on all vars (MODEL 3)
full.mod <- lm(log.medv ~ ., boston)

# look at model output
summary(noblack.mod)
summary(full.mod)

anova(lm(log.medv ~ . - quart, boston), full.mod)

# compare models
anova(noblack.mod, full.mod)

# Figure 7. plot diagnostics for model 2
png(filename='diag1.png', width = 600, height = 600)
```

```
par(mfrow = c(2, 2))
plot(noblack.mod)
dev.off()

# Figure 8. plot diagnostics for model 3
png(filename='diag2.png', width = 600, height = 600)
par(mfrow = c(2, 2))
plot(full.mod)
dev.off()

## 4. INFERENCE

# Bootstrap F-Test
# Our variables of interest
special <- c("black", "quart2", "quart3", "quart4")
# Get our estimates from the full model (model 3)
beta.hat.full <- full.mod$coefficients

l.matrix <- matrix(0, nrow = length(special), ncol = length(beta.hat.full))
for (row in 1:4) {
    index <- match(special[row], names(beta.hat.full))
    l.matrix[row, index] <- 1
}

glh.statistic <- function(formula, hypothesis, output.dim, data, indices) {
    data.sample <- data[indices, ]
    result <- lm(formula, data.sample)
    data.matrix.sample <- model.matrix(formula, data.sample)
    beta.sample <- result$coefficients

    denominator <- summary(result)$sigma ^ 2
    numerator <- t(l.matrix %*% beta.sample - hypothesis) %*%
        solve(l.matrix %*% solve(t(data.matrix.sample) %*%
                                     data.matrix.sample) %*% t(l.matrix)) %*%
        (l.matrix %*% beta.sample - hypothesis)
    numerator <- numerator / output.dim
    numerator / denominator
}

# get our estimated F-Statistics
full.f.stat <- glh.statistic(log.medv ~ ., rep(0, length(special)),
                             length(special), boston, 1:nrow(boston))

# set seed for reproducibility
set.seed(123)

# get our bootstrapped F-Statistics
results <- boot(
    data=boston, statistic=glh.statistic, R=10000,
    formula=log.medv ~ ., hypothesis=l.matrix %*% beta.hat.full,
    output.dim=length(special)
```

```
    )

# get our p-value for this test
p.value <- 1 - ecdf(results$t)(full.f.stat)
p.value

f.plot <- ggplot() +
    geom_histogram(aes(x = results$t, y = ..density..),
                   fill = 'white', color = 'black', bins = 50) +
    geom_vline(xintercept = full.f.stat,
               color = 'red', size = 1.5) +
    geom_vline(xintercept = quantile(results$t, .95),
               color = 'blue', size = 1.5) +
    labs(x = 'Bootstrapped F Statistics', y = 'Density') +
    theme_minimal(base_size = 16)
ggsave('fplot.png', plot = f.plot, width = 16, height = 6)

# Bootstrapped CI's for estimate of our variables
# set seed for reproducibility
set.seed(123)
# 10000 bootstraps
n <- 10000
# 0.05 significance level -> 95% confidence intervals
alpha <- 0.05
# get our t-stat estimates
t.hat <- summary(full.mod)$coefficients[special, 3]

# get our bootstrapped t-statistics
bootstraps <- replicate(n, (function() {
    bootstrap.i <- sample(1:nrow(boston), nrow(boston), repl = T)
    bootstrap.frame <- boston[bootstrap.i,]
    model <- lm(log.medv ~ ., bootstrap.frame)
    model.sum <- summary(model)
    model.t <- model.sum$coefficients[special, 3]
    bootstrap.vals <- model.t - t.hat
    bootstrap.vals
})())

# get our ci's
cis <- sapply(1:4, function(x) {
    row <- bootstraps[1,]
    ci <- t.hat[x] - quantile(row, c(1 - alpha/2, alpha/2))
})

t.hat.se <- summary(full.mod)$coefficients[special, 2]

# clean format a bit
colnames(cis) <- special
cis <- rbind(cis, t.hat)
rownames(cis) <- c("lower", "upper", "estimate")
# multiply by SE's to get estimate CI
```

```
cis <- t(cis) * t.hat.se
cis


# plot for predicting log(medv) with all other variables held
# constant except for black and quarter
# theoretical black proportion
B_0 <- seq(0, 1, 0.00001)
# get black var from B_0
black_0 <- 1000 * (B_0 - .63) ^ 2

# get our quarter var to predict
# quantile split from our data set
quants <- quantile(inter, seq(0, 1, .25))
quants[1] <- -Inf
quants[5] <- Inf
inter_0 <- (sqrt(black_0/1000) - 0.63) *-1
# split according to data set
binned_inter_0 <- cut(inter_0, quants, labels = 1:4)

# predict our log medv's
pred_log_medv <- predict(full.mod,
                         data.frame(black = black_0, quart = binned_inter_0,
                                    crim = 0, zn = 0, indus = 0, chas = 0,
                                    nox = 0, age = 0, tax = 0, ptratio = 0,
                                    sq.rm = 0, log.dis = 0, log.rad = 0,
                                    log.lstat = 0))

theo_pred <- ggplot() +
    geom_line(aes(x = log(B_0), y = pred_log_medv, color = binned_inter_0),
              size = 1.25) +
    xlim(-8, 1) + labs(x = 'log(Proportion of African Americans)',
                       y = 'Predicted log(Median House Value)',
                       color = 'Quartile') +
    theme_minimal(base_size = 16)

ggsave('theo_b.png', theo_pred, width = 16, height = 6)
```

6.3 Appendix C: R Session Info

```
R version 3.6.3 (2020-02-29)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 18363)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.1252
[2] LC_CTYPE=English_United States.1252
[3] LC_MONETARY=English_United States.1252
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] boot_1.3-24      forcats_0.5.0    stringr_1.4.0    dplyr_1.0.2
 [5] purrr_0.3.4      readr_1.4.0      tidyr_1.1.2      tibble_3.0.4
 [9] ggplot2_3.3.2    tidyverse_1.3.0

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.5          cellranger_1.1.0    pillar_1.4.7
 [4] compiler_3.6.3      dbplyr_2.0.0        tools_3.6.3
 [7] digest_0.6.25       jsonlite_1.7.2      lubridate_1.7.9.2
[10] lifecycle_0.2.0     gtable_0.3.0        pkgconfig_2.0.3
[13] rlang_0.4.8         reprex_0.3.0        cli_2.2.0
[16] DBI_1.1.0           rstudioapi_0.13     yaml_2.2.1
[19] haven_2.3.1         withr_2.3.0         xml2_1.3.2
[22] httr_1.4.2          fs_1.5.0            generics_0.1.0
[25] vctrs_0.3.5         hms_0.5.3           grid_3.6.3
[28] tidyselect_1.1.0    glue_1.4.2          R6_2.5.0
[31] fansi_0.4.1         readxl_1.3.1        farver_2.0.3
[34] modelr_0.1.8        magrittr_2.0.1      MASS_7.3-51.5
[37] backports_1.2.0     scales_1.1.1        ellipsis_0.3.1
[40] rvest_0.3.6         assertthat_0.2.1    colorspace_2.0-0
[43] labeling_0.4.2      stringi_1.4.6       munsell_0.5.0
[46] broom_0.7.2         crayon_1.3.4
```