ABSTRACT

We introduce impact charts and apply them to problems of systematic bias encoded in three different data sets. Impact charts are highly visual, making the effects they find easy to understand by both domain experts and non-experts. Impact charts are nonlinear and non-parametric, so they are able to identify structural biases whose functional forms are not a priori well understood.

Impact charts are based on SHAP, an interpretability method initially designed to interpret predictions made by Machine Learning (ML) models, which is in turn based on Shapley values, an approach to assigning responsibility for economic outcomes to different factors. Although impact charts use techniques from the ML community, they are intended for use in general settings, whether ML is present or not.

Impact charts provide valuable insights even when generated from aggregate data sets. Aggregate data sets typically provide the individuals whose data they are derived from an additional level of privacy as compared to the original unaggregated data. In this work, we relied predominantly on aggregate data from the U.S. Census Bureau, which is known to have a robust privacy program.

We introduce and evaluate impact charts using three examples of their use. Our first example uses impact charts to identify racial and ethnic bias in eviction rates. Our second example uses U.S. Census data to identify racial and ethnic bias in housing prices. Our third example assesses the impact of several factors on local access to supermarkets.

All three examples not only correct for the effects of income, but also clearly demonstrate the relative impact of income as compared to racial and ethnic features. For example, we demonstrate that in some areas like DeKalb County, GA, the fraction of the population that is Black impacts eviction rates three times more than income does.

In addition to the impact charts specifically discussed herein, we have produced thousands of geographically localized impact charts for the data sets mentioned above.

There is wide variation in the shape and structure of impact plots built using data from different local areas. We hypothesize that in the future work we will be able to categorize these and identify local policy decisions, whether de jure or de facto that cause the differences from one area to the next.

KEYWORDS

Impact Charts, Systematic Bias, Machine Learning

1 INTRODUCTION

In 1881, Frederick Douglass wrote an article entitled, “The Color Line” [21]. The color line was his name for the “moral disorder,” of racial prejudice. He saw it as neither natural nor inevitable. Nineteen years after Douglass’ work, at the 1900 Paris Exposition, W.E.B. Du Bois presented the color line in visual form. He exhibited charts and graphs that presented a wide variety of data on the lives and circumstances of Black Americans [4].

We take inspiration from Du Bois’ work seeking to visually expose the color line, but using a new method inspired by and applied to the age of Machine Learning (ML). We need new tools because since the color line was first identified, it has been objectivized [40], systematized [46, 53] and digitized [6, 12, 23, 43]. Our contribution is the idea of impact charts, a new tool that builds on SHAP [37] to create a powerful means of interrogating data sets and exposing the biases they contain.

Impact charts as a bias detection mechanism are also inspired by prior work on helping users make less biased decisions [5, 59, 60], and interactive systems that help them to do so [8, 14, 33, 41, 62]. In this work, we will concentrate our attention on data sets that were initially collected to measure the status quo of social, political, and economic systems. Census data sets are good examples of this. When impact charts are built on top of this kind of data sets, they can expose the underlying systematic biases the data encode more directly and more flexibly than traditional tools such as scatter plots [15] and regression analysis [20]. Furthermore, as we shall see in examples below, impact charts can provide insights from aggregated data sets without the need for or risk of personally identifiable information. We deliberately used data sets whose raw inputs are predominantly from the U.S. Census Bureau, which has been a leader in deploying privacy preserving techniques in the generation of aggregate data [2, 55].

Impact charts can also serve a diagnostic role in ML [1] when built on top of data sets that are being used or might be used to train ML models that replicate their biases. Impact charts can be
generated as part of the testing phase of data set development [29] or as part of a data set evaluation label [28, 31].

The remainder of this work is organized as follows: In Section 2 we define the general problem of impact assessment, and define impact charts as a solution. Because impact charts are visualizations, we introduce them and work through how they are constructed and interpreted and what insights they can help us with in an example-driven way over the course of Sections 3-5. In Section 3, we introduce the visual vocabulary of impact charts and demonstrate how they show the persistence of the color line when applied to data on eviction rates, race, and ethnicity. In Section 4, we switch to data on home values, which we use not only to expose the color line again, but also to compare impact charts to alternatives. In Section 4.2 we compare impact charts to scatter plots [15]. In Section 4.3, we demonstrate the flexibility impact charts because of their lack of assumptions and flexibility of scoring metric vs. regression analysis. In Section 5, we show how a collection of impact charts, one for each of several features, can help us quickly refine alternate hypotheses and identify the feature or features with the largest impact. Finally, Section 6 offers some conclusions and suggested next steps for the further research, development, and application of impact charts.

2 DEFINITIONS

2.1 General Problem Statement

Before we dig into impact charts as a solution, we begin with a general problem, stated as follows: given \( n \) observations of the values of \( m + 1 \) variables, what is the relationship between \( m \) of them and the other one? Is there a relationship? How is it structured? What quantitative, qualitative, and visual characteristics does the relationship have? Within this relationship what is the contribution to or impact of each of the \( m \) variables on the final one.

In ML settings, whose conventions we will use herein, the \( m \) variables are called features and denoted \( x_0, x_1, ..., x_{m-1} \) while the final variable is called the target and is denoted \( y \) [9]. We will work with a set of \( n \) observations of the feature values \((x_0^{(j)}, x_1^{(j)}, ..., x_{m-1}^{(j)})\) and the target value \( y^{(j)} \) for \( 0 \leq j < n \).

2.2 Impact Charts

Impact charts are a set of \( m \) plots, one for each feature \( x_i \), that attempt to show the impact of different values of \( x_i \) that appear in the data on the corresponding values of \( y \).

Impact charts are based on SHAP [36, 37], a method designed to interpret predictions made by ML models. SHAP is in turn based on Shapley values [48], an approach to assigning responsibility for economic outcomes to different factors. SHAP has also been used to identify racial bias in credit decisions [58]. Although impact charts use SHAP and other techniques from the ML community, they are intended for use in general settings, whether ML is present or not.

Impact charts and the Shapley values that underly them are intended for use in general settings, whether ML is present or not. They are use SHAP and other techniques from the ML community, they are intended for use in general settings, whether ML is present or not. They are structured such that the impacts of each of the features on the target for a given data point add up to exactly the difference between the target value of the point and the mean value of the target across the entire data set [36, 37, 48]. Even if feature values are correlated, we tease apart relative impact across the features such that each impact

is independent of the impacts of other features. Mathematically,

\[
\hat{y}(x_0, x_1, ...x_{m-1}) = \mathbb{E}_{0 \leq j < n} [y^{(j)}] + \sum_{0 \leq i < m} I(x_i)
\]

where \( I(x_i) \) is the impact of \( x_i \). That is to say that the prediction \( \hat{y} \) made by some ML model is simply the mean of the target \( y \) values it was trained on plus the sum of the impacts of each of the features \( x_i \).

This property is important because it lets us look at a plot of \( I(x_i^{(j)}) \) vs. \( x_i^{(j)} \) for all \( 0 \leq j < n \) showing the impact of a single feature for all \( n \) feature values seen in a data set. We don’t have to worry that what we are seeing is actually the effect of some other feature, because that appears in an entirely separate impact chart for that other feature. In this respect it is fundamentally different and more powerful than an \((x_i^{(j)}, y^{(j)})\) scatter plot [15]. We will see this in Section 4.2.

Two of the examples we will discuss below come from residential real estate, which has been widely shown to harbor significant racial biases [27, 44, 46, 47, 53]. In Section 3, we consider a data set with \( x_i \) describing the median household income and racial and ethnic distribution of renters in different neighborhoods, and \( y \) indicating the eviction rate in each of those neighborhoods. We might expect an impact chart for the impact of renter income on eviction rate to be that low-income neighborhoods have higher rates of eviction and high-income neighborhoods have lower rates of eviction. And in an unbiased world, we would expect there to be no impact from the racial and ethnic features. Impact charts enable us to interrogate alternate hypotheses like these and, if they are invalid, quickly come up with alternate hypotheses.

The additive nature of impacts can be easily seen in the other real estate example that we discuss in detail in Section 4. There, we use impact charts to say things like housing prices in neighborhoods that are 80% white are impacted by $50,000. By this we mean that they are $50,000 different than prices in otherwise comparable neighborhoods where the level of whiteness varies. Similarly, in a high income neighborhood, income may have a $100,000 impact on housing prices. Even if this neighborhood is more white than average, and whiteness is correlated with higher incomes in the data, this impact is completely additive to the impact of the racial demographics of a neighborhood. So a neighborhood might have housing prices $50,000 above the mean because of it’s whiteness and an additional $100,000 because of its high income for a total of $150,000 above the mean. A similarly high-income neighborhood that is majority minority instead of majority white would see the income impact but not the racial impact, giving it housing prices $100,000 above the mean.

Because of the way they break impact down into additive components, impact charts can be used as a tool for identifying and visualizing social biases embedded in a variety of data sets and the social, political, and economic systems that produced them. For example, in a hypothetical world where there is no systematic racism and the data were perfectly representative, the impact charts for the housing price data set would show that the whiteness of a neighborhood had zero impact on housing prices. The same would be true of the Blackness, Asianess, Hispacness, or any other racial or demographic -ness represented in the data.
The extent to which these impacts are non-zero in particular circumstances can help guide our understanding of how and where systematic racism [24] and other systematic biases exist and what policy changes might be necessary to mitigate them.

2.2.1 Machine Learning and Prediction vs. Explanation. Impact charts are built on top of ML models. The ones presented herein use a technique called gradient boosted trees [25] as implemented in the XGBoost package [17], but our software also supports using other ML techniques to generate them. We chose gradient boosted trees because they are generally regarded to be effective for modeling tabular data [39]. Tree-based Shapley values are also less computationally intensive than those computed on other types of ML models [36].

In general ML problems are posed as prediction problems, not explanation problems [49]. Given a set of \( n \) training data points \((x_0, y_0), (x_1, y_1), \ldots, (x_{m-1}, y_{m-1})\) through \((x_0, y_0), (x_1, y_1), \ldots, (x_{n-1}, y_{n-1})\), the goal is to learn a function \( \hat{y}(x_0, x_1, \ldots, x_{m-1}) \) that minimizes some function of the difference, or error, between \( \hat{y}(x_0, x_1, \ldots, x_{m-1}) \) and \( y(j) \) across all \( 0 \leq j < n \). ML systems then use this function to make predictions \( \hat{y}(x_0, x_1, \ldots, x_{m-1}) \) for \( j \geq n \) of the value of the target for points in feature space that were not seen at training time.

While ML models can be good at prediction problems according to commonly-used metrics (with the caveat that errors are often concentrated in ways that disproportionately impact historically marginalized groups [13, 43]) they tend not to be particularly explainable, which might make them appear not to be applicable to our problem. The good news on this front is that interpretability techniques have been developed [36, 37, 45, 61] and among them SHAP [36, 37] in particular has the additive impact property we need for impact charts.

The SHAP approach has the additive characteristics we want impact charts to have, but it focuses on explaining the predictions made by a specific model trained on a specific data set, not on explaining the impact of features in a data set, much less the impact of the systematic social, political, or economic processes whose behavior is encoded in the data set.

To solve this problem, impact charts apply SHAP not to a single ML model, but to an ensemble of \( k \) ML models, each trained on a random subset of the initial data. This is a form of bagging [11] but we explain by aggregating Shapley values rather than predictions. A related approach is to use bagging with carefully designed weak but inherently explainable models [34, 35] rather than SHAP on a more general set of modeling techniques.

We use bagging of \( k \) models for two reasons. First, a single model can be prone to the problem of over-fitting, which means that it essentially memorizes the data it is trained on, but does not contain any structure that allows it to generalize to accurately predict \( \hat{y} \) for feature values it has never seen [18]. By using an ensemble, we ensure that none of the individual models is overfit to the data set as a whole.

Second, and more critically important, when we have \( k \) models, we can look at the impact of a given feature on the target for a given data point in each of the models independently. It is often the case that models trained on different random samples of the same data set have a similar overall error rate across the data set, but concentrate their errors in different areas. The net result of this, which we will see visually in Section 3 is that we get not just a single estimate of the impact of a feature on the target, but an empirical distribution of it. The structure, and in particular the variance of this distribution helps us reason about how to interpret the estimate of the impact. Intuitively, if the models in the ensemble agree, to within a small amount of variance, on the impact of a feature on the target, we can more reasonably believe that the models have independently latched on to a relationship inherent in the data or the process that created it, rather than that each of member of the ensemble has overfit to the subset of the data upon which it was trained.

3 EVICTION, RACE, ETHNICITY, AND INCOME

3.1 A First Impact Chart: Eviction and Blackness
Our first set of impact charts were generated from a data set with a point for each census tract in DeKalb County, Georgia in each of ten years from 2009 to 2018 inclusive. The data set has ten features. Seven of the features indicate the fraction of renters in the tract who identify as belonging to each of the following racial groups: White Alone; Black Alone; American Indian or Alaskan Native Alone; Asian Alone; Native Hawaiian or Pacific Islander Alone; Some Other Race Alone; and Two or More Races. The eighth feature indicates the fraction of renters who identify as being White Alone not Hispanic or Latino. The ninth feature represents the fraction of renters who identify as being Hispanic or Latino of any Race. The tenth feature is median household income for renters in the tract normalized to constant 2018 dollars. These features were derived from U.S. Census American Community Survey (ACS) 5-year data [56]. The target of the impact charts is eviction rate, measured as the number of evictions in a year per 100 rental households in the tract. This data comes from the Eviction Lab at Princeton University [26]. Although we discuss only DeKalb County here, we have prepared similar eviction impact charts for 489 counties across the country and published them at http://evl.datapinions.com/.

As in all the examples herein, we used aggregate data rather than individual data, thus lessening both our risk of failing to preserve privacy [42] and our risk of introducing noise by imputing unknown racial features in data where they are missing [27].

We will start with Figure 1, an impact chart that looks at the impact of the percentage of renters who identify as Black on the rate of eviction filings in DeKalb County, Georgia between 2009 and 2018.

Let’s look at this first impact chart and see what it tells us. We will start with the leftmost fifth of the chart, representing tracts that that are less than 20% Black. They are to the left of the light gray vertical grid line labeled 20% at the bottom. All have impacts...
Figure 1: The impact of the percentage of renters who identify as Black on the rate of eviction filing in DeKalb County, Georgia between 2009 and 2018, measured at the census tract-year level. The horizontal axis is the percent of renters who identify as Black alone. It ranges from 0% to 100%. Some census tracts in the county are almost entirely non-Black and others are almost entirely Black. The vertical axis is the impact of the percent of renters who identify as Black on the eviction filing rate, as measured by number of eviction cases filed per 100 renters per year. Each green dot represents one of the census tracts in the county during a single year.

(green dots) below -10. Most are between -10 and -15. A handful are below -15. What this means is that the eviction rate for these tracts are 10 to 15 evictions per 100 renters lower than otherwise comparable neighborhoods (similar income and mix of other racial and ethnic groups). The impact chart makes this plainly visible.

An even more extreme impact can be seen at the right side of the chart, to the right of the vertical grid line labeled 80% at the bottom. These are tracts where more than 80% of renters identify as Black. If we look at the green dots in this region, they almost all have an impact greater than +5 on the vertical axis. Many, including almost all that are above 90% Black, have an impact greater than +10. This means that neighborhoods that are overwhelmingly Black have higher rates of eviction even though the underlying models can also assign impact to income and would do so if income had an impact across different racial demographics.

3.2 A Second Impact Chart: Eviction and Whiteness

The point of an impact chart is to look at the impact of one single variable, regardless of the statistical relationship it may have to other variables in the model. We can look at other variables as well; each gets its own impact chart. For example, Figure 2 is the impact chart for the percentage of renters who are white.

In this impact chart, the green dots for tracts where renters are less than 7% white are all in the positive impact range. This means that neighborhoods that are mostly non-white have higher rates of eviction than others. Again, this is corrected for income. What this impact chart is telling us is that regardless of what groups of non-white renters live in a tract, the simple fact that few of the renters are white is, on its own, sufficient to drive eviction filing rates up.

On the other hand, when neighborhoods are 50% or more white, there is an impact of -2 to -5 points on the eviction rate. Neighborhoods with a lot of white renters have fewer evictions than neighborhoods where renters are predominately not white.

When we looked at the impact chart for Black renters, the trend looked like it might reasonably be explained with a straight line. But the white impact chart has a more complicated behavior. This ability to identify effects with shapes that are not just straight lines or other fixed parametric shapes is one of the key things that distinguishes impact charts from regression analysis. Because we used ML instead of linear or other parametric regression, we are able to identify these nonlinear impacts, even when we don’t know anything about the shapes we expect to see before we start the analysis.

3.3 What About the Impact of Income?

The impacts in Figures 1 and 2 demonstrate that the racial demographics of a neighborhood systematically impact eviction rates. But what about income? That is the one feature that we would hypothesize would have an impact, independent of the color line. Figure 3 looks at the impact of the income feature.
Figure 2: The impact of the percentage of renters who identify as non-Hispanic or Latino white on the rate of eviction filing in DeKalb County, Georgia between 2009 and 2018, measured at the census tract-year level. Compare to Figure 1, which is for the feature representing the percent of renters who identify as Black in those same tracts over the same time period.

Figure 3: The impact of the median income for renters in a census tract of eviction filing in DeKalb County, Georgia between 2009 and 2018. Compare to Figures 1 and 2, which are for the features representing the percent of renters who identify as Black and non-Hispanic or Latino white in those same tracts over the same time period.
4 HOUSING PRICES, INCOME, RACE, AND ETHNICITY

Our next data set moves us from rental markets to owner-occupied housing. The underlying data set in this case is derived entirely from U.S. Census ACS 5-year data [56] from 2021. Like our first data set, it has features for race and ethnicity, but for those of homeowners rather than renters, and a feature for the median income of homeowners. The data is at the block group level, which is finer granularity than the census tract level and captures data about small neighborhood-sized areas. The target is median home value for owner-occupied homes in the block group. We will look at impact charts for data from the Los Angeles, California area (a CBSA in Census terminology [54]). As was the case with our eviction data, we generated hundreds of additional impact charts for other CBSAs. They can be found at http://rih.datapinions.com/impact.html.

Our goal in generating impact charts with this data is to get a more nuanced understanding of the impact of the racial and ethnic makeup of neighborhoods on home values than previous studies were able to do using regression analysis [44, 47]. We were also able to do this using only aggregate data, thus reducing the possibility of exposing private information about individuals. This does not mean that researchers using impact charts needn’t be concerned with privacy and related ethical issues, but it does mean that they can often do their work with data that makes it easier to preserve privacy.

4.1 The Impact of Whiteness

Consider Figure 4, an impact chart which looks at the impact of whiteness on median home values at the block group level in the Los Angeles area CBSA.

The green dots are denser in Figure 4 than we saw in the impact charts in Figures 1, 2, and 3, mainly because there are a lot more of them (7,019 vs. 1,340). There are also a lot more grey dots, (7,019 × 50 = 350,950) but for most of the graph, the band of grey dots is between $50,000 and $75,000 wide from top to bottom. More importantly, for block groups that are below about 24% white, all of the grey dots are below zero.

We built k = 50 different models with random 80% subsets of our data, but not a single one of them ever indicated there was a single block group with white population under 24% where the impact of this low level of whiteness was positive. So homes in low-white population neighborhoods have lower value, even when income is included in the models and regardless of what races and ethnicities are present. Non-whiteness is a color line that systematically devalues a neighborhood.

At the other end of the scale, neighborhoods that are overwhelmingly white saw home values an average (green dots) of $75,000 or more higher than comparable neighborhoods. With the exception of a few outliers, the error bars implied by the grey dots measured this impact as being at least $50,000 in all cases.

We hypothesized that low-income tracts would have higher eviction rates. But the impact chart tells us something more nuanced. There is a spike between $20,000 and $40,000, but it drops away on either side. For tracts on the right side of the spike, this fits the hypothesis that higher income areas have lower rates of eviction and above a certain income level ($75,000 here) income has essentially no effect on eviction rate. But on the left side, the opposite happens. Lower income tracts have stronger negative impact, including the lowest seen anywhere on the chart, below -5 for some tracts where median renter household income is around $20,000. It is possible that this is the case because low-income renters qualify for programs or housing with less stringent eviction practices than the open market. It is certainly something that points us in that direction for further inquiry.

Taken together, these three impact charts tell two compelling stories. The first is that the color line of systematic racism is alive and well in eviction filings in DeKalb County, Georgia. The impact of fraction of renters in a census tract who are Black (a range of ±15) is three times the size of the impact of the median income of renters in the tract (a range of ±5). The second, less complete but potentially more positive story of the existence of some kind of eviction safety net for very low income households.

3.4 What About the Gray Dots?

The green dots we have been looking at represent the average impact from 50 different machine learning models. Each of these models uses the same code, but is trained on a different random sample of 80% of the data. The grey dots are the impact of the variable in each of the 50 different models. So for each green dot, there are 50 gray dots behind it, in a vertical line.

The reason we plot the grey dots is that ML models can be fickle, giving very different results when trained on data that to humans looks relatively the same [50, 52]. By training many models, we can see if they agree on the impact of each input or if they are all over the place, indicating that our final estimate of the impact (the green dot) isn’t that accurate.

The length of the vertical distribution of gray dots gives us a visual idea of how confident we can be on what the corresponding green dot tells us. They are like error bars. If all 50 models closely agree, the grey dots don’t extend very far above or below the green dot. But in some cases they do. You can see a few examples in each of the impact charts above. Often, the grey dots for these tracts look somewhat out of place relative to the trend of the green dots immediately around them. This means that the models are having trouble agreeing on what is going on in the tract, perhaps because it is heavily influenced by some other feature that we did not include in our model.

We also sometimes see cases, like for incomes between $40,000 and $80,000, where there are what look like two different impact lines of green dots, each surrounded by tight groupings of grey dots. In cases like these the difference between the two groups of green dots may lie in a missing feature that could explain the impact but was not given to the model.

3 For more on the variables that racial and ethnic features were based on, see the ACS variables B25003A_002E, B25003B_002E, B25003C_002E, B25003A_006E as described at https://api.census.gov/data/2021/acs/acs5/variables/B25003A_002E.html and similar. For income data, see the variable B25119_002E as described at https://api.census.gov/data/2021/acs/acs5/variables/B25119_002E.html.
4.2 Scatter Plots as an Alternative Visualization

The impact of whiteness on home values is made abundantly clear in the impact chart in Figure 4. It is so clear, in fact, that one might wonder why this kind of impact is more normally publicized through individual stories of specific individuals who were impacted. Rather than systematic studies, stories are regularly published that describe how Black families replaced family photos and other personal effects in their homes with artifacts chosen to imply that the home was owned by a white family. The result was that professional appraisers increased their assessment of the value of the home, sometimes by hundreds of thousands of dollars [22, 30, 31].

Impact charts let us go from individual anecdotes to compelling systematic conclusions using readily available privacy-preserving data. Recall that everything in Figure 4 was derived from publicly available aggregate U.S. Census data.

For the sake of comparison, the most widely used visualization that is designed to help us visualize the impact of a feature on a target is the scatter plot [15]. Figure 5 is a scatter plot of the data set behind the impact chart in Figure 4.

The fundamental problem with Figure 5 is that we are showing the home values $y_{ij}$ on the vertical axis, not the impact $I(x_{ij})$ of the single whiteness variable as we did in Figure 4. The scattering across $2,000,000$ on the vertical axis obscures the effect of the single variable on the horizontal axis with the effects of all the other variables. We could add a regression line, or a regression curve if we fit a non-linear model, but even if we did so, the impact would not be explicitly shown. At best we could say something about impact based on the slope $m$ of the line, like, “for every 10 points the percentage of homeowners who are white goes up, home values tend to go up by $0.1m$.”

Alternatively, we could partition the block groups into those with high white (or Black, etc...) population, say 95% or more, and low white (or Black, etc...), say 50% or less. We could then use classical statistical techniques to evaluate the difference between the distribution of median home values in the different groups, correcting for median income [44]. We think that impact charts produce a richer and more compelling story than this approach.

Some existing approaches to understand bias using scatter plots [33] focus on helping users understand the relationship and impact of interacting features, which is complementary to the insights on the impact of individual features that impact charts offer.

4.3 Regression Analysis and Measures of Error

Another consideration, which is often useful in constructing impact charts, and which was critical in the construction of the impact charts for the housing value data set, is the choice of error function. All ML models try to minimize error on the data they are trained on, but how we decide to measure error can significantly influence their behavior. In regression models, Mean Squared Error (MSE) is widely used both because it often gives good results and because it is computationally efficient in many cases [20].

Among the assumptions that regression analysis makes is that of homoscedasticity [20], which means that the variance in the observations of the target, median housing price in our example, is the same regardless of the value of the variable’s value.

But it’s hard to imagine this condition holds in our case. Suppose, for example, that for homes that are actually worth $200,000, our input data captures them to within an accuracy of ±$20,000, or ±10%. What about homes worth $1,000,000? We don’t expect to be able to also measure their value to within ±$20,000, which would be ±2%. 
Figure 5: A scatter plot of median home value (vertical axis) vs. the fraction of homeowners in a block group who identify as white (horizontal axis) at the block group level in the Los Angeles-Long Beach-Anaheim, CA CBSA.

Instead, we might measure them to within the same \( \pm 10\% \) as the lower priced homes, which would be \( \pm $100,000 \). This is very much a classic case of heteroscedasticity, the opposite of homoscedasticity.

Similarly, if we built a model to predict median home prices, we would not expect its mean error to be the same in dollar terms on homes worth around $200,000 as homes worth around $1,000,000. At best, we might hope for the same mean relative error. For example, the model might have a mean absolute error of 20% of the actual home value regardless of whether it was $200,000 or $1,000,000. In the former the error would be \( \pm $40,000 \) and in the latter it would be \( \pm $200,000 \). This measure of error is called Mean Absolute Percent Error (MAPE).

Note that we use the mean of the absolute value of the error percentage rather than the mean of the percentage error. Otherwise a large positive error in half the cases (say an 80% overestimate) could be counteracted by a large negative error (an 80% underestimate) in the other half. The mean would be 0%, making this poor model look like an excellent one.

Optimizing model construction to minimize MAPE tends, especially in cases of heteroscedasticity, to produce very different models than optimizing for minimum MSE. If we optimize for MSE in cases like ours, the influence of the high-priced end of the market can overwhelm the influence of the lower priced end because absolute errors are larger there. Squaring compounds this effect. So in effect, we try really hard to be good at predicting prices at the high end of the market even if that means, in MAPE terms, we end up with a pretty bad model at the low end of the market. We'd like to have a good model at both ends of the market, so we'd prefer to optimize our model for MAPE rather than MSE.

XGBoost [17], the open-source implementation of gradient boosted trees that we used behind the impact charts herein, supports MAPE. We used it for all of the housing price impact charts we generated. For the eviction rate, we use Mean Average Error (MAE), since the target was already expressed as a rate of eviction per 100 renters.

5 ACCESS TO SUPERMARKETS, INCOME, RACE, AND ETHNICITY

Our final motivating example uses 2019 data from the U.S. Department of Agriculture (USDA) Food Access Research Atlas (FARA) [57] along with 2019 U.S. Census ACS data [56] to assess the impact of race, ethnicity, income and access to a vehicle on access to local supermarkets that carry fresh healthy food. Areas without such access are commonly referred to as food deserts. The target variable, as defined by the FARA data set is called lapophalfshare. It represents the fraction of the population of a census tract that lacks access to a supermarket within a distance of 1/2 mile of their home. This is expressed on a scale from 0, meaning everyone in the tract has a supermarket within 1/2 mile of their residence, to 100, meaning nobody does. 30 would means 30% of the tract’s population does not have access to a supermarket within 1/2 mile but the other 70% does.

The data sets we constructed are at the census tract level for all census tracts within the most populated census places (cities and towns, typically) across the United States. It is sparser than the the previous two data sets, because the FARA data is itself more sparse. We will concentrate on the New York City data set herein.

As was the case for the two previous data sets, the impact charts we generated for food access in New York City indicated that racial
We introduced impact charts and applied them to the problem of visualizing the impact of the color line in two housing-related applications. In the eviction analysis in Section 3, we showed using impact charts that the impact of Blackness was 3 times that of the impact of income on eviction rates. In the home valuation analysis in Section 4, we illustrated the impact of whiteness, introducing the use of MAPE as an error metric, and demonstrated the degree to which impact charts can clearly demonstrate impact when scatter plots and regression analysis do not. We also discussed impact charts relative to regression analysis using constructed sub-populations. Finally, in Section 5, we used USDA FARA data to illustrate how by visually scanning a collection impact charts, we can refine hypotheses.

Our current work is empirical, but we are also working on a theoretical analysis to characterize the impacts that impact charts expose vs. any they may miss by looking at their behavior on synthetic data produced by parametric and causal models.

Finally, we note that the case studies here each involved a data set representing a single geographic area (a county, a CBSA, and a city). Because we chose ML techniques that produce accurate models and also enable us to compute impact charts efficiently (as opposed to more general techniques that consume resources exponential in the number of features [37]) we have been able to generate thousands of impact charts for different geographies.

### 6.1 Code and Data Availability

Having introduced impact charts and shown their use on three data sets, we believe we have only scratched the surface. In order to enable further development and make it easy for researchers to bring their own data sets and generate impact charts from them, we have made a reference implementation of impact charts available at https://github.com/impactchart/impactchart. This repository is the recommended starting point and holds the most current version of the impact chart library. We hope that in using our code, researchers will both find additional insights like those discussed herein and identify limitations of either the approach or the software implementing it so that we can make appropriate improvements.

For completeness, the original code that constructed the data sets and produced the impact charts in Section 3 has been preserved at https://github.com/datapinions/evldata and https://github.com/datapinions/evlcharts respectively. The original code for Section 4 has been preserved at https://github.com/datapinions/rihdata and https://github.com/datapinions/rihcharts. The original code for Section 5 has been preserved at https://github.com/vengroff/faradata and https://github.com/vengroff/faracharts.

### 6.2 Future Work

In future work, we believe it will be possible to characterize the impacts we see across hundreds of geographies based on political features of the geographies themselves. For example, we will be able to say that the color line creates greater disparities in cities in one state or region than in another, or in cities that have implemented a particular program vs. those that have not. This, we believe, will provide policy makers and citizens alike with powerful information on how to address color line impacts in their communities.

We would also like to look into integrating the generation of impact charts into the lifecycle of data sets used to train ML models [29] and summary reports on such data sets, such as data set nutrition labels [28, 51]. Impact charts could also be integrated into existing systems for visually assessing bias in data sets and ML systems [5, 8, 14, 41, 59, 60, 62].

In engaging practitioners, whether ML practitioners, data engineers, data scientists, social scientists, or policymakers, it will be critical that we understand how they use and interpret impact charts. In studying this we intend to build on existing work on how

---

1For reference, a full run that generates almost 5,000 eviction impact charts runs in a matter of hours on a single laptop.
Figure 6: Impact charts showing the impact of the percentage of residents identifying as members of certain racial and ethnic groups ((a)-(d)), the impact of median income (e), and the fraction of residents with access to a vehicle (f) on the fraction of residents who lack of access to a supermarket within half a mile of home.

practitioners work with anti-bias tools and how tools can be constructed to be more useful to practitioners [3, 7, 10, 16, 19, 32, 38].

Coming full circle to Douglass’ color line [21] as we introduced it at the outset, we hope that the use of impact charts can contribute in some small way to exposing the color line as it exists today and inspiring solutions that lead to its elimination.

7 ETHICS AND SOCIAL IMPACT

When working at the intersection of data and structural racism, we believe there are important ethical and social concerns that researchers must address.

7.1 Ethical Concerns and Mitigation

Whenever we deal with data including race, ethnicity, gender identity or any of many other sensitive variables or proxy variables,
we have to be concerned about privacy. We were while developing impact charts. A common approach to managing privacy involves researchers gaining access to sensitive data in a highly controlled environment, for example via https://www.researchdatagov.org/, doing their analysis, and carefully structuring the results they publish so as to preserve the privacy or the (possibly unwilling or unknowing) participants whose data they studied. We deliberately did not take this approach, and instead challenged ourselves to develop an analysis technique that could effectively find structural impacts using aggregate data. In particular, most of our data comes from the U.S. Census, which has a robust privacy program [2, 55]. We cannot claim that aggregate data can never be manipulated in ways that expose individuals [42], but we can say that it can be analyzed with substantially less risk in most circumstances.

7.2 Author Positioning and Reflection
Having been trained as a Computer Scientist in an era when ethics were not a part of most undergraduate or graduate CS curricula, the author has historically been biased towards technology-centric solutions to problems. And without question impact charts are technology-centric.

However, like all technologies, we must come to view them as tools and evaluate the extent to which their uses are concentrated in applications that are socially useful or socially damaging. We believe that they are the former and that this has been demonstrated by the example in Sections 3, 4, and 5.

That having been said, impact charts are only one small tool. Pointing out adverse impacts does not eliminate them. At best, it is diagnostic [1]. But as such it can motivate those generating the charts to become a part of a larger movement, to engage with, to listen to, and to be led by those often less privileged individuals who bear the brunt of the impacts the charts highlight.

7.3 Adverse Impact
Impact charts, no less than any other data-driven technological artifacts, can be useless or even dangerous when fed the wrong data. While at its core the technique is designed to identify and expose biased impacts in society as reflected in data, there is nothing stopping motivated actors from repurposing the vocabulary of impact charts to tell the stories they want told on top of data they have generated, manipulated, or commissioned to meet their own needs.

REFERENCES