Seven Perspectives in U.S. Income Distribution: An Invitation to Economic Inequality for Non-Experts

Anonymous

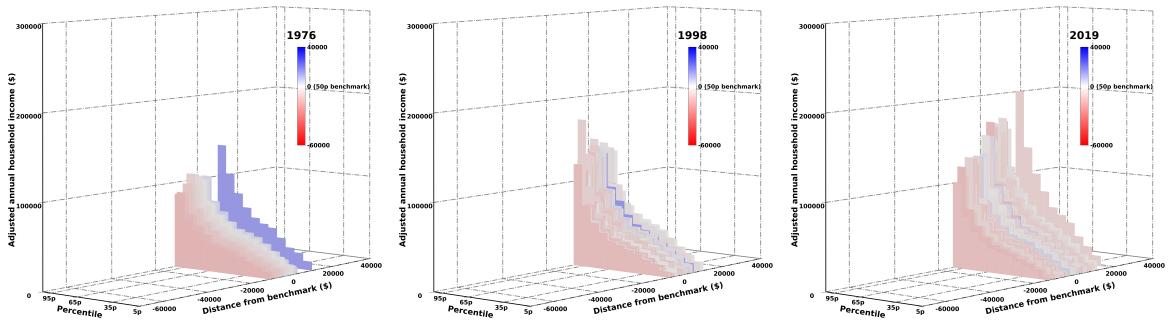


Figure 1: Visualization of adjusted household income from 1976 to 2019. The graph includes 51 slides representing 50 states and DC. The adjusted household income is shown on the vertical axis (namely "Adjusted annual household income"). The i^{th} percentile on the short horizontal axis (namely "Percentile") represents the i^{th} richest household in each slide (i.e. each state). The distance from household income benchmark is displayed on the long horizontal axis (namely "Distance from benchmark"). We utilize the red or blue color to illustrate the state below or above the benchmark in 1976, respectively. This visualization is the first perspective in the U.S. income distribution in this analysis.

Abstract

The distribution of household income is the central concern of modern economic policy due to its strong influence on life quality. Yet, this information is poorly communicated to non-expert audiences by using complex statistics, such as the Gini coefficient. To address this issue, we visualize income distribution over time using the U.S. household income microdata. The results are seven striking dynamic animations of income distribution over time, drawing public attention, and further investigation of economic inequality in the U.S.

Index Terms: Human-centered computing—Visualization—Visualization application domains—Information visualization;

1 Introduction

Income inequality is the central concern of modern economic policy because it has a strong influence on quality of life. For example, a broader income gap correlates to a shorter life expectancy on average [3, 7, 11, 33]. Income inequality also effects education opportunity. For example, higher income families can afford better education, which usually leads to higher income after graduation, which contributes to a wider income gap [8, 14]. In comparison to other developed countries (e.g. Japan and U.K.), the U.S. has not only wider income gap but also faster inequality growth rate [19, 26].

Evidences show that typical citizens in the U.S. acknowledges the effect of income inequality on their life, but they are not fully aware of the magnitude of inequality [16]. Many people believe that the wealthiest 20 percent of Americans owns just 59 percent of all wealth [25]. However, this group in fact owns 84 percent of total wealth. When surveyed about income differences by race and ethnicity, the evidence shows that the misperception in income distribution is substantial. Indeed, the respondents estimated that for every $100 earned by a White family, a Black family makes $90, when, in reality, the value is only $10. The misperception about the Latin-White wealth gap is just as significant as Black-White [22]. Not having an accurate understanding of the degree of income inequality makes it significantly harder to tighten the economic gap in America [15]. Moreover, evidences show that this problem is getting worse due to the lack of appropriate public attention [20, 30].

The most common way to communicate the quantity of economic inequality is via the Gini coefficient (denoted as G; it is also called Gini index or Gini ratio) [10, 18, 26]. For a discrete income distribution with n entities (e.g. households or individuals), the Gini coefficient is computed as [9]:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\sum_{i=1}^{n}\sum_{j=1}^{n}x_j} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\bar{x}} \quad (1)$$

where x_i is the income of entity i and $\bar{x}$ is the average income. The Gini coefficient can also be computed for a continuous income distribution similarly. Since G depends on the area between the Lorenz curve and a diagonal line of equality (Figure 2), its value lies between 0 and 1 [9]. $G = 0$ indicates a perfectly egalitarian distribution where everyone has the same income. $G = 1$ represents a distribution where all resources belongs to a single entity. While these two extremes are relatively easy for a non-expert to understand, the values of G in between the two extremes are more difficult to interpret due to the nonlinear scale of G. For example, $G = 0.5$ does not indicate an "average" degree of inequality but rather a high one. In addition, a 0.1 unit decrease from $G = 0.3$ is not the same as a comparable decrease from $G = 0.6$. Not only does G require some technical knowledge to interpret, but it also has a few undesirable properties. For example, various distributions of income give the same value of G because the entire distribution collapses into a single number. We believe that the complexity of Gini coefficient is an obstacle for a non-expert audience to recognize the magnitude of the U.S. income gap.

To address this challenge, we propose a visualization framework to communicate the U.S. income distribution with public audiences who might not have technical backgrounds. Our interactive visual-

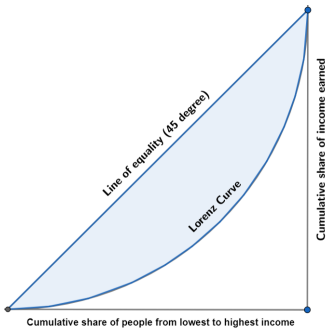


Figure 2: The Lorrenz curve [1].

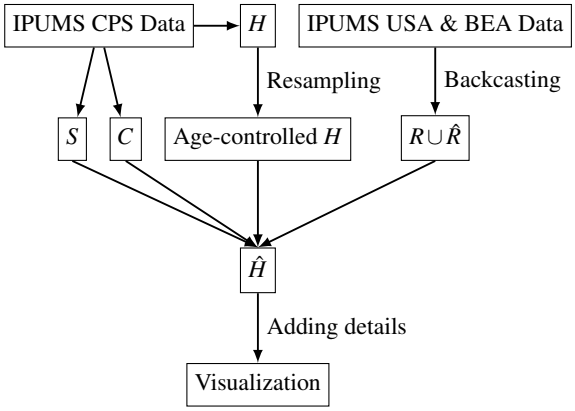


Figure 3: Data processing framework for U.S. household income. $H, \hat{H}, C, R$, and S represent household income, adjusted household income, consumer price index, regional price parity, and effective household size, respectively. This data processing framework has two steps: adjusting for household income and incorporating additional details to render a 3D visualization.

izations provide instant and accurate snapshots of the current state of income distribution. Our dynamic graphs help non-expert audiences understand the progression of economic inequality in the U.S. over the last 44 years. Unlike the Gini coefficient, these figures present the income inequality clearly and informatively without requiring advanced statistical knowledge.

2 Related Works

There is little prior visualization research in communicating U.S. income inequality to a non-expert audience. Comparative analyses of international economic inequality using visualization show that while the gap between the rich and poor decreased over time in some countries (such as France), it has increased in other countries (such as the U.K. and China) [17, 26, 31, 32]. Although these studies show a global perspective in economic inequality, they do not provide the insights about different regions within a country.

Although income inequality in the U.S. has been of interest to academics [28, 34], non-experts are typically not the target audience. A more esoteric treatment of this important issue limits the accessibility for many individuals in society, potentially impacting the awareness of the overall scale and impact of inequality. For example, while many people may be aware of income inequality along racial lines, the magnitude of the inequality is consistently underestimated [4, 16, 22]. Our visualization approach has been specifically developed not only to convey the magnitude of inequality in the U.S., but also to allow anyone (non-experts and experts alike) to engage with the issue interactively.

Variables	Description	μ	σ	min	max
Household income (H)	Total nominal income of all household members during previous year	52343.81	65094.54	-37040	3299997
Consumer price index (C)	Estimation of inflation base on 1999 price level	1.18	0.55	0.65	2.92
Regional price parity (R)	Differences in price levels across states for a given year. Expressed as a percentage of the overall national price level	97.53	8.54	84.8	119.2
Gross rent (r)	Gross monthly rental cost of the housing unit, including contract rent plus additional costs (e.g. utilities)	841.98	201.19	512.0	1600.0
Effective household size (S)	Squared root of number of household member	1.58	0.45	1	5.10
Household weight (w)	Household-level sampling weight	1,538.89	920.93	0.00	17957.53

Table 1: Summary statistics in the household levels. Each variable has a total of 2.82 million observations for the 1976-2019 period.

3 Data Processing

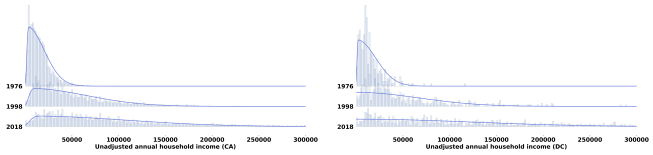
Figure 3 describes our data processing framework for the U.S. household income data. In this two-step process, the data is first adjusted with three normalizers and controlled for the age distribution for better income comparison. After that, we convert the data to the final visualization by incorporating additional visualization details. We note that this approach is generalizable. Indeed, we can utilize the same framework for datasets from other countries or other income variables (e.g. personal income or cumulative wealth) as well. More details about each step of the data processing framework are elaborated below. Our framework is implemented in a Python package and is publicly available on PyPI. More implementation details can be found at [omitted for double-blind review]. We also prepare a step-by-step Python tutorial in a Google Colab notebook for users who are interested in some technical aspects of this analysis. The tutorial is available at [omitted for double-blind review].

3.1 The Data

Our data is collected from Integrated Public Use Microdata Series - Current Population Survey (IPUMS-CPS), Integrated Public Use Microdata Series - United States of America (IPUMS-USA), and the Bureau of Economic Analysis (BEA), from 1977 to 2020 [6, 12, 29]. Since the population survey data reported during the ongoing year (t) is for the previous year ($t-1$), our analysis represents the 1976-2019 period. We only extend the analysis back to 1976 because the geographic information is not reliable before this time [12]. Although DC is not an official state, for the rest of this paper we refer to a state as one of "50 states and the DC." Table 1 presents the summary statistics of the data. We show two examples of household income distribution for CA and DC in Figure 4. Similar to other income-related variables, our household income distribution data is right-skewed. Household income can be negative because it is collected from various sources, such as business or rent. Hence, we exclude the bottom five percentiles to avoid the negative income. Similarly, removing the top five percentiles is necessary to avoid invalid high-income measurement due to the disclosure avoidance measure of IPUMS-CPS survey [12].

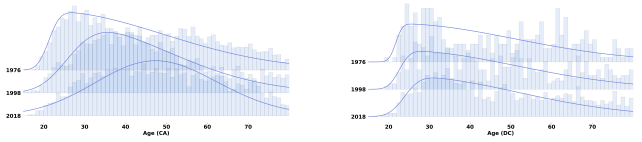
3.2 Household Income Adjustment

To improve the accuracy of income comparison across different geographic regions and time periods, we employ three normalizers: regional price parity (denoted as R), consumer price index (denoted as C), and effective household size (denoted as S). In theory, there are several other normalizers (e.g. taxes and transfers) that are accountable for income discrepancies. Unfortunately, we do not have



(a) Distribution of H in CA over time (b) Distribution of H in DC over time

Figure 4: The income distribution of both CA and DC has been increasingly right-skewed.



(a) Distribution of age in CA over time (b) Distribution of age in DC over time

Figure 5: The age distribution of CA and DC over time. While CA's population ages relatively quickly, the age of DC's population is more stable over time.

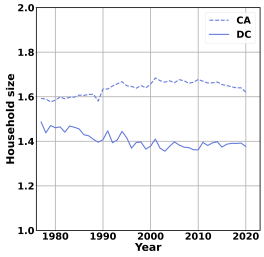


Figure 6: The effective household size of CA and DC.

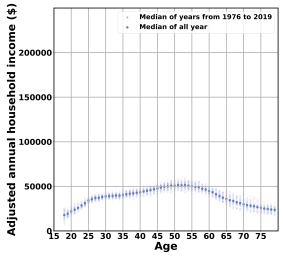


Figure 7: The median adjusted household income for each age.

access to these normalizers for the U.S. household income dataset. For this study, we first employ bootstrap resampling on household income to standardize the age distribution, then we adjust the result with three above normalizers for better comparison of income across states over time as described in Figure 3.

Adjustment for age. Since age correlates strongly to income, it is not appropriate to compare income between states that have different age distribution [2,13,24]. According to conventional wisdom, older people tends to earn higher income since they have accumulated experience and a strong network (Figure 7). The age distribution in the U.S. is significantly different across state and time period. For example, the mean age in CA grows faster than in DC (Figure 5). In 1976, both CA and DC age peaks approximately around 25 years old. As time progresses in Figure 5a, the CA age peak moves towards 50 while DC's peak stays the same position (Figure 5b). Without standardizing for age distribution, comparing income of DC and CA is not appropriate. To avoid this issue, we proposed a resampling technique that standardized the age distribution. By resampling with replacement from the original sample so that age distribution of each state is uniformly distributed, we hope to have a better comparison of income across states and time.

Adjustment for cost of living. BEA regional price parity R [6] is used to adjust the geographic differences in prices. For example, people in CA generally earn higher wages than national average but also pay higher living expenses than citizens in AL. By adjusting for cost of living using BEA data, our visualization allows meaningful comparisons of income across different geographic regions. Since the regional price parity database is only available before 2008, we perform backcasting to retrieve the unknown data with the following model:

$$R = \alpha + \beta_1 r + \beta_2 F(r) + \beta_3 F(R) + \beta_{4,i} FE_i + \varepsilon \qquad (2)$$

where $F(\cdot)$ is the forward-shift operator (i.e. lead operator), FE_i is the binary fixed effect for state i, and r is the gross rent. We choose gross rent (r) as the primary independent variable to predict R because it is the main component of the living expense within the U.S. Since data about gross rent is only available every 10 years before 2000, we perform interpolation to obtain annual dataset.

Adjustment for inflation. To adjust for inflation, we employ the consumer price index (C_t^0) that allows accurate comparison of income over time [5]. At period t, C_t^0 is the ratio of the market basket price m_t to the base period m_0: $C_t^0 = \frac{m_t}{m_0}$. Assuming that the market basket structure is stable over time, we use the adjusted consumer price index from 1976 to 2019 as an approximate control for the price variation over time. This assignment allows us to adjust household income H for inflation. Although the index might suffer from the biases (e.g. new good, quality, outlet, and substitution) that cause the overstate inflation, we know no other alternatives to control the R changes.

$$C_t^{2019} = \frac{m_t}{m_{2019}} = \frac{m_t/m_{1999}}{m_{2019}/m_{1999}} = \frac{C_t^{1999}}{C_{2019}^{1999}} = \frac{C_t^{1999}}{0.652} \qquad (3)$$

Adjustment for household size. The effective household size S should also be taken into account when comparing household income. For example, with $100,000 income, a household consisted of two people can have a more comfortable life than a six-person family with the same income. Therefore, we further adjust household income using the effective household size to improve the quality of income comparison. We utilize the squared root scale as suggested by [21] (which is computed as squared root of total number of household members) instead of a linear scale (i.e. simply count the number of members) because some expenses do not scale linearly. Indeed, a 6-person household does not often have 6 cars and the cost of a 2-bedroom apartment is often not doubling that of a 1-bedroom one.

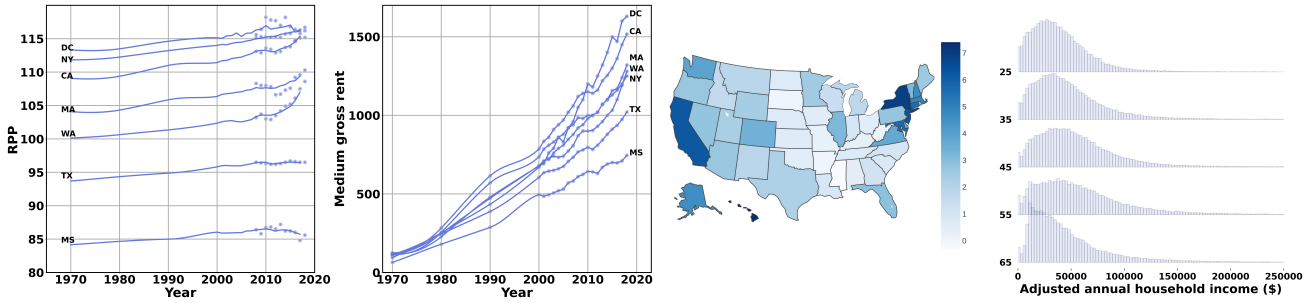
3.3 Additional Visualization Details

Having the adjusted household income, we decide to display the data on a 3D graph. 3D chart enables the visualization of income distribution on 2 axes (namely "Adjusted annual household income" and "Percentile" in Figure 1) while also allows a comparative analysis across different states among the other axis (namely "Distance to benchmark" in Figure 1). We strongly believe that further reduce the dimension of the visualization will be harmful for assisting non-expert audiences understanding the complexity of income inequality. Indeed, compressing income distribution to 1 dimensional statistic (e.g. the Gini coefficient) is shown to be difficult for non-expert audiences to understand since it requires technical knowledge. To better display the data on a 3D graph, we incorporate the following details in our visualization:

1. Segmenting the income into percentile buckets
2. Designing a ranking system that reflecting each state's economic growth over time
3. Introducing color spectrum to improve visualization
4. Incorporating information about state's population

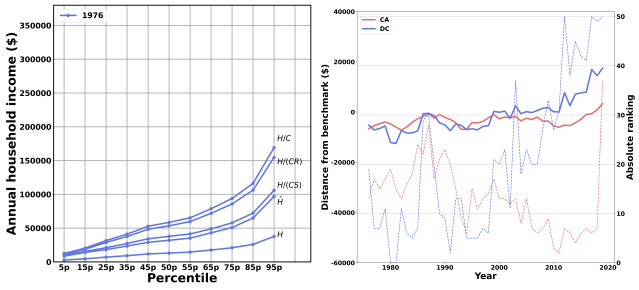
We elaborate these visualization details below.

Income segmentation. In this step, we segment adjusted household income $\hat{H}$ into several buckets and use a summary statistic of buckets for further visualization. To do so, we first calculate the



(a) Regional price parity (R and $\hat{R}$) (b) Gross rent (r and $\hat{r}$) (c) State fixed effect of regional price parity (d) Average effective household size S

Figure 8: Estimation of regional price parity for each state from 1976 to 2019 with model 2. Gross rent has a linear trend in the majority of states. The fixed effect coefficients in the backcasting model show that DC and CA are among the states with the highest living expense.



(a) Various normalizers for CA (b) Income ranking of CA and DC

Figure 9: The age distributions differ across states. DC has a relatively younger population than CA. The age distribution of each state may not be stable over time. Age and income imply a strong relationship in the U.S.

cumulative household weight (w) and the household percentile (P). For household h in-state s with total household t at year y, household percentile P is expressed as:

$$P_{h,s,y} = \frac{\sum_{i=0}^{h} w_{i,s,y}}{\sum_{i=0}^{t} w_{i,s,y}} \tag{4}$$

Then, the adjusted household income data is sorted in ascending order of percentile. Since we want to avoid over-display data, we prefer to use decide scale to percentile scale for segmentation:

$$\begin{aligned} \text{decile scale: } k &\in \{05\%, 15\%, ..., 45\%, 50\%, 55\%, ..., 95\%\} \\ \text{percentile scale: } k &\in \{05\%, 06\%, ..., 49\%, 50\%, 51\%, ..., 95\%\} \end{aligned} \tag{5}$$

Given state s and year y, a data bucket $b_{k,s,y}$ is defined as:

$$b_{k,s,y} = \{H_{h,s,y} | k-1 \leq P_{h,s,y} \leq k\} \tag{6}$$

We choose the maximum of $b_{k,s,y}$ as its summary statistic for display on our 3D graph.

Ranking. To illustrate each state's income rank and their movement over time, we experiment with two ranking methods:

1. For a given year, we sort the 50^{th} percentile of all 51 states and use the order of a state in the sorted list as its ranking.

2. For a given year, we compute the distance between each state median income to a predetermine benchmark, such as national median income in 2019.

We choose to use median instead of mean to be more robust to the skewed distribution of household income. The information about ranking is displayed on the longer horizontal axis (namely "Distance from benchmark" in Figure 1, noting that this figure is using the second ranking approach). Regardless the ranking method, from the left to right, the income ranking increase (i.e. states on the far right are richer).

Color. We assign a color from the red-blue spectrum to a state based on its ranking in the initial year (i.e. 1976). For example, if DC was a poor region in 1976, its slide will be red in all visualization from 1976 to 2019, even though DC might not be poor in 2019 anymore. Introducing this color spectrum, we want to provide the user a historical context of the economic development of each state.

Population size representation. The information about population size of each state is represented using the thickness of its slide. For state s in year y with h households, the thickness of its slide is proportional to its normalized population $\bar{p}$, which computes as:

$$\bar{p}_s = \frac{\sum_{i=0}^{h} w_i}{min(p_s)} \tag{7}$$

where w_i represents the weight of household i in the sample.

4 Result and Discussion

The real and predicted regional price parity inferred from model 2 are shown in Figure 8a. DC and CA are among the states with highest cost of living. In addition, these states have not only the highest gross rent in the country but also a high rate of gross rent growth (Figure 8b). Even after controlling for gross rent, DC and CA are still among the most expensive region in the U.S. as indicated by their size of fixed effect coefficient (Figure 8c). With an exceptionally high cost of living as well as its growth rate in these areas in comparison to other less expensive states, such as MS, we believe that adjust household income with regional price parity is crucial for a meaningful income comparison across states over time. Resampling to standardize age distribution has almost no effect on the income distribution, which can be explained by the fact that income distribution of household with various ages are approximately the same in our data (Figure 8d).

We show the effect of different normalizers for adjusting household income of CA in 1976 in Figure 9a. Because of inflation, the purchasing power of one dollar is larger in 1976 than in 2019. Therefore H/C curve is above the H curve due to the fact that we are using consumer price index based on 2019 value. The effective household size S tends to have larger effect than regional price parity R since the household size of CA tends to be large (See Figure 6 for more details). The effects of S and R partially neutralize the effect of C, and the final adjusted household income $\hat{H} = H/(CRS)$ is more suitable to be compared across states over time.

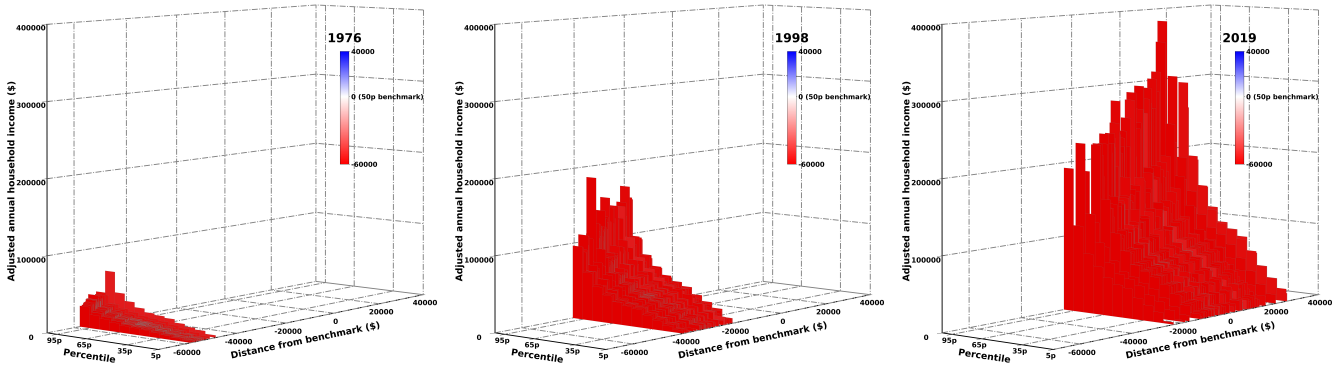


Figure 10: The second perspective demonstrates the U.S. household income without adjustment from 1976 to 2019. The tremendous movement through years illustrates a growth illusion.

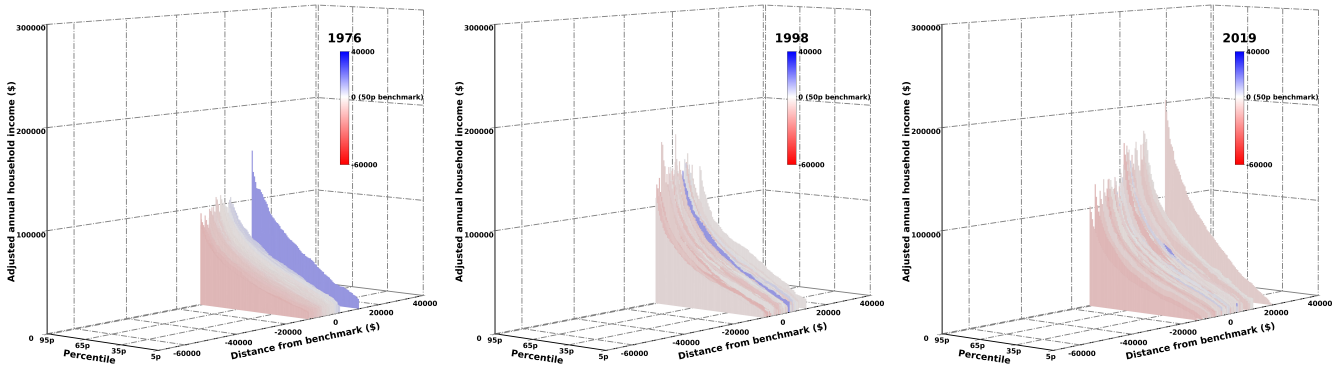


Figure 11: The third perspective takes the advantage of the percentile scale to increase the resolution of $\hat{H}$. The higher resolution of percentile enables the users to clearly observe the shape of income distribution.

In addition to household income adjustment, we experiment with two approaches for ranking region based on income as introduced in Section 3.3. We find that sorting regions by median household income introduces an artificial amount of variance in the ranking of each state. For example, using this ranking system (dashed line in Figure 9b), CA ranks 8^{th} in 2018, indicating that it is one of the richest states in the country. Nonetheless, CA ranks 38^{th} in the next year, pushing it closer to the poorest states. This fluctuation is unlikely explainable by economic force but rather statistical noise. The other ranking approach, which is the distance of state's medium income to a benchmark, is more stable (bold line in Figure 9b). We believe that the second approach provide a more realistic picture of real economic development.

Having all necessary data, we demonstrate the utility of our visualization by showing seven perspectives in U.S. income inequality. We show that even though economic inequality is a complex issue, our visualization can help users have a better understanding of this problem by observing it from various angles. Although we only include three sample graphs (1976, 1998, and 2019) for each perspective in this paper due to the space limitation, an animated set of 44 graphs (corresponding to 44 years from 1976 to 2019) for each perspective is publicly available at [omitted for double-blind review]. The third perspective implement percentile scale (explained in Equation 5) for a higher resolution view of income distribution. All other perspectives implement the decile scale.

First perspective: The first perspective is our main view of the U.S. income inequality in this paper. This set of graphs visualizes the adjusted household income $\hat{H}$ from 1976 to 2019 (Figure 1). As time progresses, the poor (e.g. 5^{th} percentile) stays nearly the same, while the rich (e.g. 95^{th} percentile) gets richer at an increasing rate. Indeed, the height of the tallest block increases from $150,000 in 1976 to almost $200,000 in 2019, while the shortest block does not change much during this period. The income distribution for all states is more linear in 1976 than that in 2019, suggesting that the gap between the lowest and highest adjusted household income enlarges overtime.

The movement of color allows us to observe the change in state's ranking over time in Figure 1. For example, in 1976, AK was the only state colored blue because it had the medium household income above the 2019 national benchmark. As time progresses, AK no longer stands out from the rest since its ranking decrements gradually [1]. In 2019, we can hardly see AK's blue slide anymore because it blends into the surrounding slides. Although there are a few states (e.g. AK) starting with high ranking in 1976 then slowly fall down and vice versa[2], the ranking of most states change very little in the studied period, reflecting a stable economic condition over time. Because there is not a lot of movement along the "Distance from benchmark" axis, we do not observe a significant economic growth in the U.S. in this period. However, we observe significant change in the income of the 95^{th} percentile, which confirm our previous conclusion about the unequal opportunity for economic development in different household percentile.

[1]Note that, by design, AK's color (in this case blue) does not change since 1976

[2]This phenomenon is known as the convergence effect (i.e. catch-up effect) in economic literature [23].

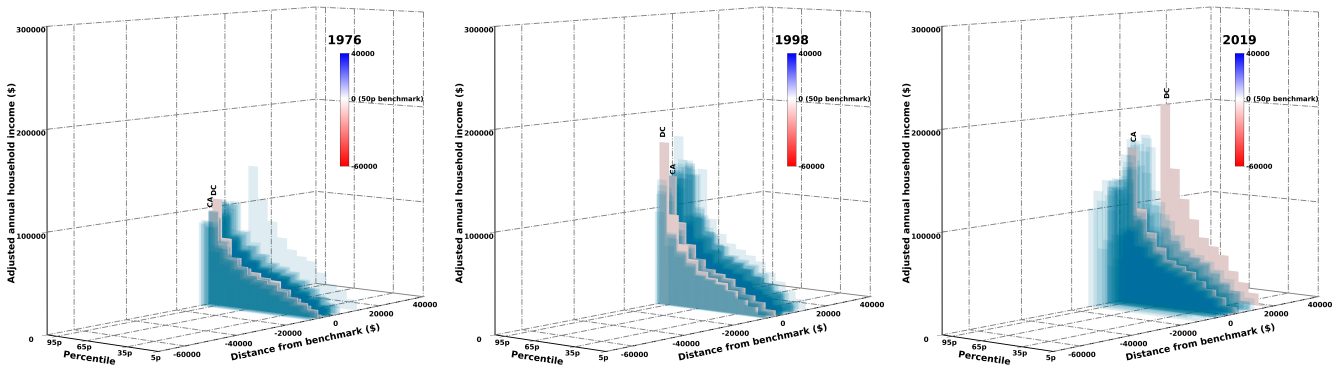


Figure 12: The fourth perspective highlights $\hat{H}$ of CA and DC from 1976 to 2019.

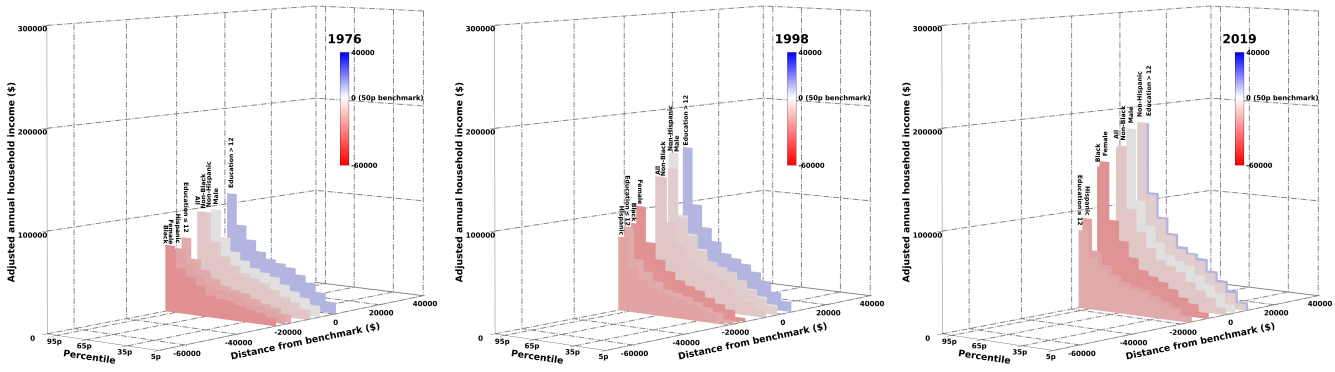


Figure 13: The fifth perspective is the adjusted household income, including all CA and its subcategories: Black, Non-Black, Hispanic, Non-Hispanic, Education with or without high school diploma (Education ≤ 12), Education with higher degrees after high school (Education > 12), Male, and Female.

Second perspective: We present the unadjusted household income H from 1976 to 2019 in our second perspective (Figure 10). Similar to the first perspective, the second one shows that the gap between the rich and poor enlarges over time. Specifically, the height of the tallest block increases from $50,000 to $400,000 from 1976 to 2019, while the shortest block are still capped at $20,000. Not surprisingly, the shape of income distribution was more linear in 1976 but become more curvature towards 2019. Unlike the first perspective, it is harder to distinguish the ranks among the states in the second perspective because all states started below the 2019 benchmark in 1976, leading to similar red colors.

Through the second perspective, we observe a significant movement of all states from the left to the right on the long horizontal axis from 1976 to 2019, which we did not see in the first one. Indeed, in Figure 10, in 1976, the position of every state varied in the range of (-55, 000 to -40,000). In 2019, its movement shifts drastically to the right (-20,000 to 25,000). Although this great movement of household income seems to suggest an overall economic advancement in the U.S., inflation, cost of living, and household size also change significantly during the same period. The first and second perspective remind the users to be critical and mindful when working with the complexity of income data. Carefully adjusting income, we avoid falling into the illusion of economic growth and having a better understanding of economic condition in the U.S.

Third perspective: The third perspective in this analysis is a higher resolution version of the first one (Figure 11) by introducing a finer scale in the percentile axis (the shorter horizontal axis). Compared to the two previous perspectives, the slides in Figure 11 are smoother, enabling the user to see the shape of the income distribution of each

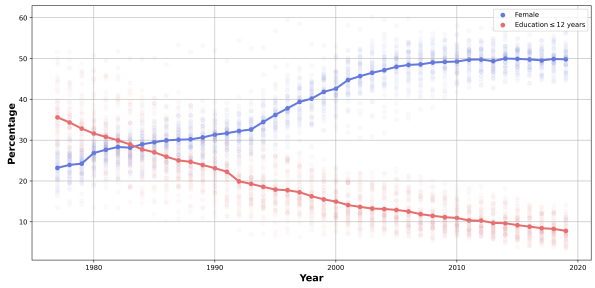


Figure 14: The percentage of household with female head or head with ≤ 12 years of education. As time progresses, we see that the percentile of female household approaches 50%, and more households have higher than high school education.

state more clearly. In 1976, even the richest state (AK) has a relatively linear shape of income distribution. In 2019, almost every states adopt a more curved distribution, suggesting an higher level of inequality. Although this information is displayed in the first perspective, the higher resolution helps the users observe the shape of the distribution more clearly. Interestingly, in 2019, although DC is the richest region (most far to the right), its shape of income distribution is not as curved as that of the poorest states (most far to the left), suggesting that DC has a lower degree of inequality. Although we believe that a policy research focusing on this observation is very interesting, it is beyond the scope of this paper. Nonetheless, we

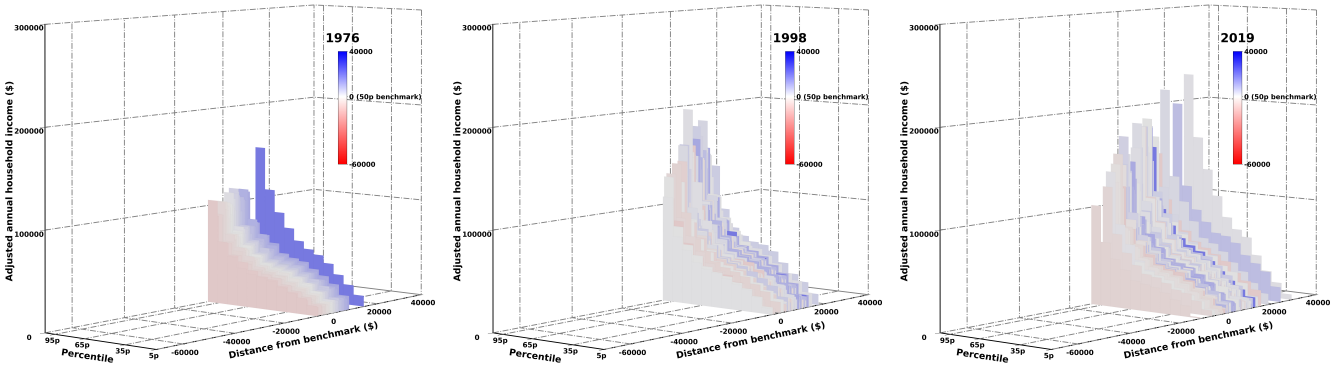


Figure 15: The sixth perspective shows $\hat{H}$ distribution of males household only from 1976 to 2019.

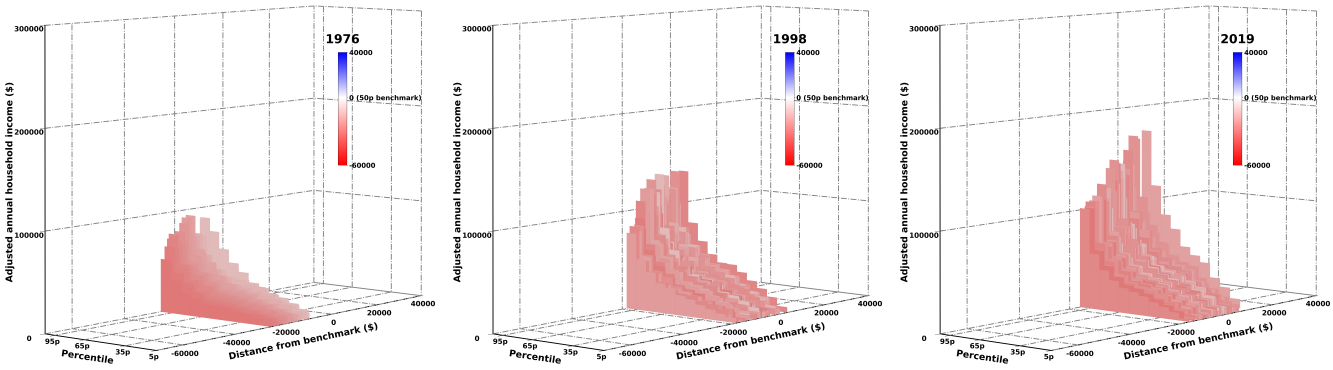


Figure 16: The seventh perspective shows $\hat{H}$ distribution of females household only from 1976 to 2019.

hope our visualization can draw the attention from non-expert users for further research of underlying conditions in these states.

Fourth perspective: The fourth perspective highlights the movement of adjusted household income for CA and DC (Figure 12). Before the 2000s, DC and CA slide were close to each other. After that, ranking of the two states deviate apart. Eventually, DC attains the highest rank in 2019, while CA remains at almost the same position. Although the inequality level in both region has increased over time, DC manages to have a more even economic growth, hence its distribution is less curved and its ranking increases over time. CA has significant growth in the upper percentile, but since its median does not change much, its ranking stays at about the same place. More information about ranking of DC and CA is previously shown in Figure 9b. We again believe that it is fascinating to investigate why some regions have more evenly distributed growth opportunity than the others, but this question is beyond the scope of our paper. As the users take a deep look into our visualization, we hope to spark their interest for further investigation in this topic.

Fifth perspective: In the fifth perspective (Figure 13) we present the adjusted income for various sub-populations in CA, including genders (male/female), races (Black/Non-Black), ethnicity (Hispanic/Non-Hispanic), and education ($\leq$ 12 years of education/ $>$12 years of education). We choose CA for demonstration because all of its sub-populations have sufficiently large sample size, helping us avoid statistical problem associated with small sample. Male, highly educated ($>$12 years of education), non-Black, and non-Hispanic are the group with highest adjusted household income. In fact, in 1976, Female, Black, or Hispanic make less than people who have below high school level of education. These groups have lower income than all CA on average before 2019. As time progresses, their positions change. In 2019, Female, Black, and Hispanic move closer to all CA. Over the 1976-2019 period, we found education to be the most important factors that determines income ranking. Highly educated group consistently has the highest ranking among all. From 1976 to 2019, due to the advancement in education and the demand of the labor market, the number of household head who only have high school diploma drops from around 40% to below 10% (Figure 14). The higher education standard makes it harder for household with less than 12 years of education to earn high income. Not surprisingly, household without higher education is the poorest group in CA in 2019.

Sixth and Seventh perspective: Our last two perspectives shows adjusted income for households whose head are only male (Figure 15) or only female (Figure 16). Unlike in the fifth perspective, we can work with all states here because all states have sufficiently large sample size of male and female household [3]. If we only look at data for male household in Figure 15, most states start above the 2019 benchmark in 1976 (hence having very light red or blue colors, indicating a rich group) and they even go far further to the right as time progresses. In contrast, female household in all states start below the 2019 benchmark in 1976 (hence having red color). As time progresses, we observe a faster movement of male cluster to the right than female one, suggesting that male household started at a better position and have been growing faster than female household. Although there is more and more household with female head as indicated in Figure 14, the underlying cause that slows down growth

[3]Note that this is not true for other subpopulations. For example, in 1976, there are 27 states (e.g. AK, AZ, CO, and CT) with $\leq$ 50 Black households, making it unpractical to construct percentile statistics for these samples. Hence we are unable to construct another perspective to understand the progression of Black income distribution for all states.

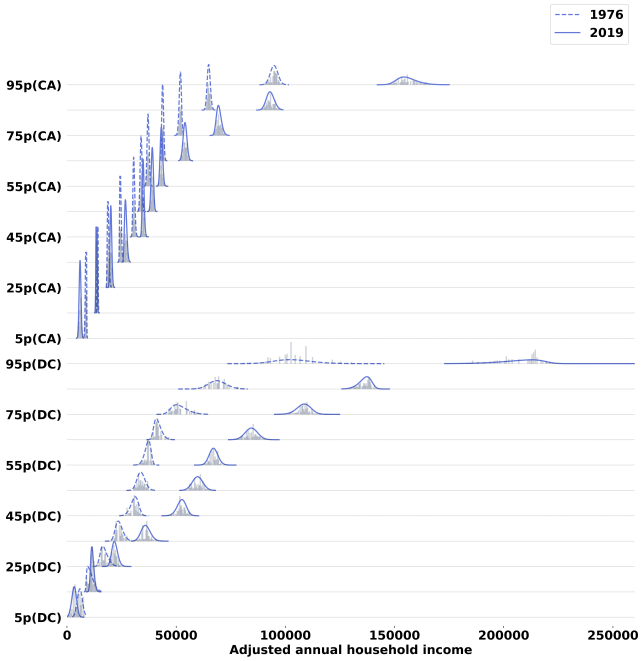


Figure 17: Bootstrap resampling for computing standard error of adjusted annual household income distribution in CA (top) and DC (bottom) in 1976 (dashed line) and 2019 (bold line).

rate of female household is another excellent question for further investigation.

Our seven perspectives in U.S. household income are intuitive and approachable for non-experts while introducing the audiences various factors that may affect the income inequality. We hope our visualization can be a point of reference for general audiences or domain experts in their discussion about economic inequality.

We note that the information about household income communicated via our visualization is best interpreted with sampling variability due to its empirical nature. For example, since the height of each block in Figure 1 is a sample estimation of $\hat{H}$ percentile for every region in the U.S, it has a standard error. We employ bootstrap resampling to estimate the standard error of the height of each column depicted in the graph of DC and CA in 1976 and 2019. These two states are chosen for demonstration of standard error because their sample size are quite different. Indeed, the sample size in CA in 1976 and 2019 is 5,412 and 5,529 observations, respectively, whereas the sample size in DC in 1976 and 2020 is 320 and 1,118 observations, respectively.

Figure 17 shows that the value of the standard error for percentile estimation rises as the sample size increases. The standard error of highly populated states (e.g. CA) is several thousand dollars for the 95^{th} percentile, but for state with smaller population (e.g. DC), the bounce is in the tens of thousands of dollars. For example, in Figure 17, the standard error in CA is generally smaller than in DC since CA has a higher population. In other words, if the IPUMS-CPS had carried out a second survey for 1976, we would see the 95^{th} percentile of real household income in DC varies by plus or minus approximately $13,500. In addition, the degree of fluctuation due to sampling variability increases as we go from lower percentile to higher one because the sample size decreases gradually as the population get richer. Indeed, a sharp-eyed viewer may notice that in Figure 1, the 95^{th} percentile in 2019 is not only taller but more jagged than it was in 1976. Most of DC's 95^{th} fluctuations are too large to be explained by actual economic forces. These ups and

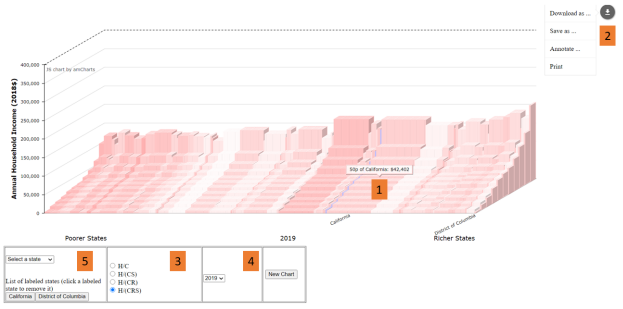


Figure 18: An example of the interactive version. We demonstrate the visualization of $\hat{H}$ (i.e. the first perspective) in 2019 in which CA and DC are noted regions while pointing our mouse to the 50^{th} percentile of CA. The users can choose different type of household income in the table below the graph.

downs reflect the effects of random sampling. CA's 95^{th} is smoother than DC's because its sample size is much more significant.

We attempt to further engage with the non-expert audiences by implementing a public web-based user-friendly visualizations at [omitted for double-blind review]. A snapshot of the website is shown in Figure 18. We hope that this visualization can be used as a resource for various general purposes, such as classroom case study or supporting material for policy research. The highlight of our web-based visualization is its abilities to provide the user an interactive and highly customized experience.

When interacting with the visualization on the website, the user can hoover to a percentile of a state to see a small window pop-up with income information of that state's percentile (box 1 in Figure 18). The users can also look up their income percentile on the graph to understand their position in their state. For example, if the adjusted household income of a user in CA is about $40,000 a year, they will be in the 50^{th} percentile as indicated in box 1 in Figure 18. Because income inequality affects different household quite differently, a household with an individualized knowledge about their relative economic position can make a more informed decision when, for example, they need to decide which party to vote for or which state to live in. In addition, this message-like box allows the user to see the actual value on the chart more precisely rather a mere estimation based on the vertical axis. We enable the user to export the visualization for a report or download underlying data for further analysis by creating a download button (box 2 in Figure 18). Various file formats are allow, such as PNG and JPG for picture or CSV and XLXS for tabular data.

The visualization on the website is highly customizable. Various type of income (e.g. real household income H/C, or adjusted household income $H/(CRS) = \hat{H}$) can be display at the convenience of the users by selecting one of the four options under the graph (box 3 in Figure 18). The user can also choose to display a certain time period (from 1976 to 2019) or mark a certain state of interest (box 4 and 5 in Figure 18, respectively). For example, the graph displayed on Figure 18 is on 2019 and CA and DC are marked. We intend to make this visualization customizable and flexible so that it can be a helpful general resource.

5 Conclusion and Further Research

We introduce a framework for visualizing income distribution in the U.S. To achieve a better income comparison across states over time, we adjust household income using regional price parity, consumer price index, and effective household size on top of resampling for age distribution standardization. Information about economic inequality

from our visualization does not require technical background to understand, hence is more accessible for non-expert audiences.

There are various direction to extend this project. For example, a case study on real users with various technical background might provide some insights for further developing this visualization. As with any other empirical long-term population study, our analysis identifies some underlying limitations. The IPUMS-CPS survey is not the same every year. Questions change, and so do the data collection methods [27]. Nonetheless, we provide an entrée to the income distribution and inequality study. A visualization for household income in the U.S. over time offers an excellent starting point because it lays bare the facts, captures attention, and stimulates many questions about causes and remedies.

6 Acknowledgement

We thanks to [omitted for double-blind review] for helping us adjust for household size and [omitted for double-blind review] for writing JavaScript code that enabled choosing year and state labels. We truly appreciate the support of [omitted for double-blind review] for their extensive support on revising this paper.

References

[1] R. Aaberge. Axiomatic characterization of the gini coefficient and lorenz curve orderings. *Journal of Economic Theory*, 101(1):115–132, 2001. doi: 10.1006/jeth.2000.2749

[2] I. Almas and M. Mogstad. Older or wealthier? the impact of age adjustment on wealth inequality. *The Scandinavian Journal of Economics*, 114(1):24–54, 2012. doi: 10.1111/j.1467-9442.2011.01662.x

[3] O. Attanasio, E. Hurst, and L. Pistaferri. The evolution of income, consumption, and leisure inequality in the us, 1980-2010. Working Paper 17982, National Bureau of Economic Research, April 2012. doi: 10.3386/w17982

[4] S. R. Bailey, A. Saperstein, and A. M. Penner. Race, color, and income inequality across the americas. *Demographic Research*, 31:735–756, 2014.

[5] M. F. Bryan and S. G. Cecchetti. The consumer price index as a measure of inflation. Working Paper 4505, National Bureau of Economic Research, October 1993. doi: 10.3386/w4505

[6] Bureau of Economic Analysis. Regional price parities by state and metro area. `https://www.bea.gov/data/prices-inflation/regional-price-parities-state-and-metro-area`, 2020. Accessed: 2021-01-31.

[7] R. Chetty, M. Stepner, and S. Abraham. The association between income and life expectancy in the united states, 2001-2014. *The Journal of the American Medical Association*, April 2016. doi: 10.1001/jama.2016.4226

[8] D. Coady and A. Dizioli. Income inequality and education revisited: persistence, endogeneity and heterogeneity. *Applied Economics*, 50(25):2747–2761, 2018. doi: 10.1080/00036846.2017.1406659

[9] P. M. Dixon, J. Weiner, T. Mitchell-Olds, and R. Woodley. Bootstrapping the gini coefficient of inequality. *Ecology*, 68(5):1548–1551, 1987.

[10] R. Dorfman. A formula for the gini coefficient. *The Review of Economics and Statistics*, 61(1):146–149, 1979. `http://www.jstor.org/stable/1924845`.

[11] P. Dünhaupt. An empirical assessment of the contribution of financialization and corporate governance to the rise in income inequality. Technical report, Berlin School of Economics and Law, Institute for International Political Economy (IPE), 2014. `http://hdl.handle.net/10419/102709`.

[12] S. Flood, M. King, R. Rodgers, S. Ruggles, and J. R. Warren. Integrated public use microdata series, current population survey: Version 8.0 [dataset]. `https://doi.org/10.18128/D030.V8.0`, 2020. Accessed: 2021-01-31.

[13] J. P. Formby and T. G. Seaks. Paglin's gini measure of inequality: A modification. *The American Economic Review*, 70(3):479–482, 1980.

[14] P. FÃ¶ldvÃ¡ri and B. van Leeuwen. Should less inequality in education lead to a more equal income distribution? *Education Economics*, 19(5):537–554, 2011. doi: 10.1080/09645292.2010.488472

[15] L. S. Giordono, M. D. Jones, and D. W. Rothwell. Social policy perspectives on economic inequality in wealthy countries. *Policy Studies Journal*, 47(S1):S96–S118, 2019. doi: 10.1111/psj.12315

[16] E. Gudrais. What we know about wealth. `https://harvardmagazine.com/2011/11/what-we-know-about-wealth`, Mar 2014. Accessed: 2021-01-31.

[17] J. Han. The analysis of opportunity inequality, income-gap and public policy. In *2011 International Conference on Computer Science and Service System (CSSS)*, pp. 2665–2668, 2011. doi: 10.1109/CSSS.2011.5974555

[18] D. Hartmann, M. R. Guevara, C. Jara-Figueroa, M. Aristarán, and C. A. Hidalgo. Linking economic complexity, institutions, and income inequality. *World Development*, 93:75 – 93, 2017. doi: 10.1016/j.worlddev.2016.12.020

[19] J. Heathcote, F. Perri, and G. L. Violante. Unequal we stand: An empirical analysis of economic inequality in the united states, 1967-2006. *Review of Economic Dynamics*, 13(1):15 – 51, 2010. Special issue: Cross-Sectional Facts for Macroeconomists. doi: 10.1016/j.red.2009.10.010

[20] J. M. Horowitz, R. Igielnik, and R. Kochhar. Americansâ views on u.s. economic inequality, Jan 2020.

[21] D. Johnson, T. Smeeding, and B. Torrey. Economic inequality through the prisms of income and consumption. *Monthly Labor Review*, pp. 11–24, 2005. `www.bls.gov/opub/mlr/2005/04/art2full.pdf`.

[22] M. Kraus, I. Onyeador, N. Daumeyer, J. Rucker, and J. Richeson. The misperception of racial economic inequality. *Perspectives on Psychological Science*, pp. 899–921, 2019. `https://doi.org/10.1177/1745691619863049`.

[23] J. Mathews. Catch-up strategies and the latecomer effect in industrial development. *New Political Economy*, 11(3):313–335, Sept. 2006. doi: 10.1080/13563460600840142

[24] K. M. Murphy and F. Welch. Empirical age-earnings profiles. 8(2), April 1990. doi: 10.1086/298220

[25] M. Norton. Living beyond your means when youâre not rich. `https://www.nytimes.com/roomfordebate/2011/03/21/rising-wealth-inequality-should-we-care/living-beyond-your-means-when-youre-not-rich`, May 2011. Accessed: 2021-01-31.

[26] M. Roser and E. Ortiz-Ospina. Income inequality. *Our World in Data*, 2013. `https://ourworldindata.org/income-inequality`.

[27] J. Rothbaum and A. Edwards. Survey redesigns make comparisons to years before 2017 difficult. www.census.gov/library/stories/2019/09/us-median-household-income-not-significantly-different-from-2017.html, 2019. Accessed: 2021-01-31.

[28] A. Rowhani-Rahbar, D. A. Quistberg, E. R. Morgan, A. Hajat, and F. P. Rivara. Income inequality and firearm homicide in the us: a county-level cohort study. *Injury Prevention*, 25(Suppl 1):i25–i30, 2019. doi: 10.1136/injuryprev-2018-043080

[29] S. Ruggles, S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek. Ipums usa: Version 10.0 [dataset]. `https://doi.org/10.18128/D010.V10.0`, 2020. Accessed: 2021-01-31.

[30] S. Subramanian and I. Kawachi. The association between state income inequality and worse health is not confounded by race. *International Journal of Epidemiology*, 32(6):1022–1028, 12 2003. doi: 10.1093/ije/dyg245

[31] B. Sutcliffe. *100 ways of seeing an unequal world*. London: Zed Books, 2001.

[32] The CORE Team. The economy. `https://core-econ.org/the-economy/book/text/19.html`, 2020. Accessed: 2021-01-31.

[33] B. C. Truesdale and C. Jencks. The health effects of income inequality: Averages and disparities. *Annual Review of Public Health*, 37(1):413–430, 2016. PMID: 26735427. doi: 10.1146/annurev-publhealth-032315-021606

[34] D. Vilda, M. Wallace, L. Dyer, E. Harville, and K. Theall. Income inequality and racial disparities in pregnancy-related mortality in the us. *SSM - Population Health*, 9:100477, 2019. doi: 10.1016/j.ssmph.2019.100477