Enhancing the Census Income Prediction Dataset

Social Justice in Machine Learning Pedagogy

Alex West and Anusha Praturu
UC Berkeley MIDS - Capstone W210
December 2020

Abstract

The racial wealth gap is well documented in the United States and has been widening over time. While the problem cannot be solved over a 14-week semester (or even a presidential term), there are components of wealth inequality that are relevant to the field of data science that can be addressed in this time frame. One such element is the concept of equity in machine learning pedagogy and teaching tools. We made it our primary goal for this capstone term to rectify a small part of this issue with one particular machine learning teaching dataset: the Adult Census Income Dataset from 1994. This outdated dataset persists in the machine learning instructional toolbox and leads beginners to the misguided conclusion that the racial wealth gap might be solved merely by more education. In updating, supplementing, and publishing this dataset with a suite of supplemental resources, we hope to spotlight the broader issue of equity within machine learning pedagogy, initiate a conversation around fairness in the census, and encourage machine learning students and instructors to think beyond the model-based conclusions of machine learning to the people behind the data.

Introduction

Addressing the racial wealth gap in the United States is no small feat, however, breaking it down into its many components provides an avenue to more tangible, piecemeal solutions. One component of this is income, earned primarily in the labor market and tracked through the US Census. While a multitude of policy papers propose and address solutions to this problem, the field of data science has yet to tackle it in a meaningful way.

A possible avenue to confront this absence is to replace a ubiquitous dataset used in teaching machine learning: the 1996 US Adult Census Income dataset from UC Irvine[1]. Comprising 14 attributes including age, race, gender, and education, students worldwide have used this data to train and test algorithms, ultimately predicting whether an individual earns more or less than $50,000. However, the income prediction in question is problematic for several reasons. First, the target variable (income) is not a continuous variable. Rather, it is a binary variable indicating whether or not the individual as defined by the record makes less than or greater than $50,000

[1] The UCI dataset was published in 1996, but the Census data used is from 1994.

annually. Second, this dataset continues to be used in instructional blog posts and classrooms, even though it is over 25 years out of date (as of December 2020) and no longer serves as a representative sample of US households. Finally, the scope of the dataset lacks the environmental variables necessary to equitably predict income for all US families, making it a poor source of record for machine learning models.

In this paper, we outline our endeavor to rectify these issues and produce a dataset that offers a more sophisticated, nuanced, and socially informed representation of US household demographics and income in 2020. In broad terms, we do this by 1) updating upstream data sources to 2019 Census data, 2) incorporating supplemental data that captures the environment around the individuals represented in the source dataset, and 3) keeping the target variable (income) as a continuous variable, which maintains optionality of binary labels, but also opens up the possibility of more sophisticated predictions.

Finally, we aim to publish our completed, updated, and enhanced dataset to a public domain for free usage, along with supplemental literature and information for prospective machine learning students and instructors.

Literature Review

There are several important dimensions of a literature review in creating a new open-source income prediction dataset. First, to estimate the impact and reach of the original dataset with its citation in academic papers as well as instructional blogs and other online content. Second, to understand the body of work that exists on documenting the causes of the racial wealth gap, to delineate the potential for additional variables. Third, to grasp the process of creating a dataset in general, and finally, to identify recent work using Census data to potentially anticipate how this dataset might be used.

The 1996 “Adult Data Set”

The original “Adult Data Set” from the UCI Machine Learning Repository has been cited in numerous academic papers, dating from 1996 to 2005. These papers discuss different machine learning algorithms and strategies for optimizing and tuning. In general, the papers are not focused on the outcome of the machine learning model (the prediction of income) other than as a means to an end: training the model (see Appendix B for the full list). In addition to academic papers, countless blog posts, forum discussions, and textbooks address this data, discussing how it can be used to learn how different algorithms work, how to tune those algorithms, and which perform the best. Recent examples of these were published on the “Towards Data Science” and “Machine Learning Mastery” blogs in 2020.

The existing literature makes clear that this 25-year-old dataset is widely used and extremely functional in training, testing, and tuning machine learning models. It is especially useful in illustrating differences between algorithms.

The racial wealth gap

The racial wealth gap presents a dire component of US socioeconomic conditions, but also a rich environment for scholarly research. Our literature review focused on the underlying causes of the gap, income as a component of building wealth (is it enough?), and potential added variables to include in the new dataset.

The prevailing research contends that the racial wealth gap is implanted in socioeconomic and political structure barriers, rather than a failing of financial literacy or contempt for education on the part of Black Americans. (Hamilton and Darity, 2017; Darity et al., 2018). The original "Adult" dataset leads machine learning beginners to conclude that education and race are two of the largest predictors of income. These conclusions are accurate, but incomplete without context. Reviewing the literature, we can include the frequently promoted "causes" of the racial wealth gap in the new dataset, such as educational attainment, homeownership, and other parameters. But, however important these features are to the model, they do not necessarily translate to policy conclusions. For example, encouraging Black Americans to attend college does little to address the underlying obstacles to obtaining a degree, such as childcare, cost of living, and access to banking services (to name only a few). The new dataset and its accompanying elements (this paper, a data dictionary, and further reading on the racial wealth gap) will provide this frame of reference where none existed before.

Wealth is composed of numerous factors including income, but according to recent studies, income is a large enough factor in the wealth gap "to explain the persistent difference in wealth accumulation" (Aliprantis and Carroll, 2019). Indeed, it is large enough to warrant focused policy attention. Ashman and Neumuller note that "if the income gap in our model were eliminated, racial disparities in wealth would eventually disappear. Thus, policies aimed at reducing income differences are likely to be the most direct and potent means for reducing the racial wealth gap" (2020, 237).

Our review of the literature on the racial wealth gap justified the use of income as a proxy for wealth, as well as motivated the plan to include context-setting with the data.

Creating open-source datasets

Dataset creation is a broad and extensive area of study, as it encompasses both pedagogy (*how* to create datasets that lead to teachable outcomes) and hard research (*what* results from models built on new data). Our interest in this area was twofold: to understand the accepted format for writing papers based on dataset creation, and to summarize the general mode of publishing datasets for public consumption. Meseguer-Brocal, et al. (2020) came the closest to what we are trying to accomplish with "Creating DALI, a Large Dataset of Synchronized Audio, Lyrics, and Notes." This effort focuses on an entirely different domain (audio and lyrics) yet the goal is similar (to create an open-source dataset to use in teaching). In general, the breadth of material on this subject was not surprising, but the lack of standardization in terms of dataset creation processes was unanticipated. Naturally there will be variation in terms of subject and data type, but there are innumerable methods by which these datasets are stored and made available to researchers. In some ways this feels like an area of opportunity for large research

universities to create repositories for their researchers to deposit generated datasets, and to regulate the production of these datasets to ensure a high standard.

Census Public Use Microdata Sample (PUMS) in the Wild
The final stage of the literature review concerned the use of Census PUMS data in research more generally. This allowed us to see the potential demand for a new Census Income dataset beyond machine learning pedagogy, and also how the PUMS data is typically regarded by the wider research community. Census PUMS data is used in a wide range of publications and research endeavors, both formally (in journal articles), and informally (in blog posts and other online publications). Given the breadth of data available in the census responses, the research domains range from transportation patterns (Fabusuyi et al., 2019) to Medicare beneficiaries (Procter et al., 2018), to discussions of data accuracy and error in the PUMS data itself (Kinney et al., 2017). For practitioners of machine learning, there are a few blogs illustrating the breadth of the dataset and applications for the field (Lynn, 2016) but mostly these are aimed at garnering users for specific online tools such as Emsi, MySidewalk, Carto, and others. It is possible that the few extra steps required to access PUMS data are enough to deter most bloggers and teachers, who prefer to use easily accessible data that can be re-downloaded by students.

This final piece of the literature review showcased the possible use-cases of an easily accessible PUMS dataset.

Dataset Description

Source

The dataset incorporates data from multiple sources. First and foremost, it contains the 2019 Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) files. The 2019 ACS PUMS files are a set of records from individual people or housing units, with disclosure protection enabled so that individuals or housing units cannot be identified. This protection includes geographic obfuscation using PUMAs (public use microdata areas), a Census-specific geographic designation each containing at least 100,000 people and nested within existing states or other geographic boundaries. Each row of the dataset represents an individual Census response. Before exploring the data in detail, it is important to acknowledge the history and controversy inherent in the Census itself.

The Census and Controversy
The United States Census, also known as the Decennial Census, is a procedure conducted every ten years by the U.S Census Bureau to collect information about the population demographics in the country (as mandated by Article 1, Section 2 of the Constitution).
As the population of the United States has grown and changed, traditionally accepted questions on the Census have been thrown into a debate—namely sex and race.

The Census asks “What is Person 1’s sex?” and includes two possible answers: male, and female. Some have criticized this question recently for not including additional sexes (e.g., intersex) and gender identities that do not match with a person’s biological sex. In a 2016 study conducted by the UCLA School of Law, approximately 0.5% of respondents aged 18 or older identified as transgender or gender-nonconforming. As the Census only allows for two possible responses, many of these individuals feel excluded and not accounted for in Census data.

Since the first Census was conducted in 1790, the options for race have changed dramatically over time. In the 1790 Census, the options for race were “free white males.” “free white females,” “all other free persons,” and “slaves,” and were determined by the U.S. marshal or enumerator, not the individual. Consequently, you could not choose your own racial identity. It was not until the 1960 Census that choosing your own race became possible, and only beginning in 2000 were people given the option to identify as more than one race. The 2020 Census has included further options: people who identify as White or Black are asked to provide more details about their background (including country or other ethnicities). Despite these growing improvements, there is still some controversy regarding racial options on the Census. For example, Middle Eastern individuals are required to be counted as White by the Census, even though many of these individuals do not identify as White. In another layer of complexity, Hispanic/Latinx is not considered a race by the Census, so individuals are forced to choose between options with which they may not identify.

Although progress has been made since its inception, the Census is still a source of controversy. We recognize these pitfalls, while at the same time understanding there is very little other data available at this scale.

Supplemental Data Sources

Supplemental data also comes from the Census, but not at the individual level. Based on the literature review of potential factors influencing the racial wealth gap we included PUMA-level data on educational attainment, health insurance coverage, unemployment, median rent, and median home value. Finally, we included data from the Bureau of Justice Statistics on rates of incarceration at the state level. These supplemental data points represent new columns of data, adding information to each individual response on the state of their geographical area with respect to each additional metric.

Dataset Creation

The process of the final dataset creation began with the original 1994 dataset. One of the primary issues with the dataset to address was its obsolescence, which made the first step in the dataset creation to update the outdated 1994 data with the more recent 2019 PUMS data. The Census changes every 10 years, and so even the fields that remained from 1994 to 2019 underwent changes in the collection process. In addition to the updated collection methodology (resulting in an increased range and value type for the resulting dataset) the 2019 data also included additional columns that were not included in the original 1994 data. For example, citizenship status is a new field in the 2019 data.

After updating the data, the task remained to supplement the 2019 data to include additional columns that may provide further context in predicted individual incomes. However, barring additional data on the individual level that was joinable to the PUMS data on hand, we were relegated to using geographical data to join to the PUMA or state of each respondent in our master dataset. Per our prior research and literature review, we decided to pursue a few particular areas for supplemental data: namely, home value and gross rent, unemployment rates, health insurance coverage rates, educational attainment trends, and incarceration rates. Home value and rent, unemployment, health insurance, and educational attainment trends all originated from Census data, which we were able to join on the PUMA level, while the incarceration data were available from the Bureau of Justice which we joined on the state level.

We joined each of these datasets individually to the 2019 PUMS income dataset, and then measured the improvement in our income predictability. For each of the data supplements that displayed a measurable improvement upon our income predictability, we included it in our final dataset. The final step was cleaning up the data, removing extraneous columns of metadata, and creating the Data Dictionary for potential users to reference.

Columns

The original Adult Income dataset included fourteen attributes from the 1994 Census PUMS data. In the 2020 update, we include those columns, plus a few more.

Columns from Original Dataset (PUMS 2019 Update)

- Age
- Race
- Sex
- Marital Status
- Education
- Class of Worker
- Occupation
- Working Hours per Week
- Supplemental Income (Capital Gains/Losses)
- Native Country/Place of Birth
- Income (as a binary variable)

Additional Columns (PUMS 2019 Update)

- State of Residence
- Citizenship Status
- Field of Degree
- STEM Degree indication
- Income (as a continuous variable)

Supplemental Columns

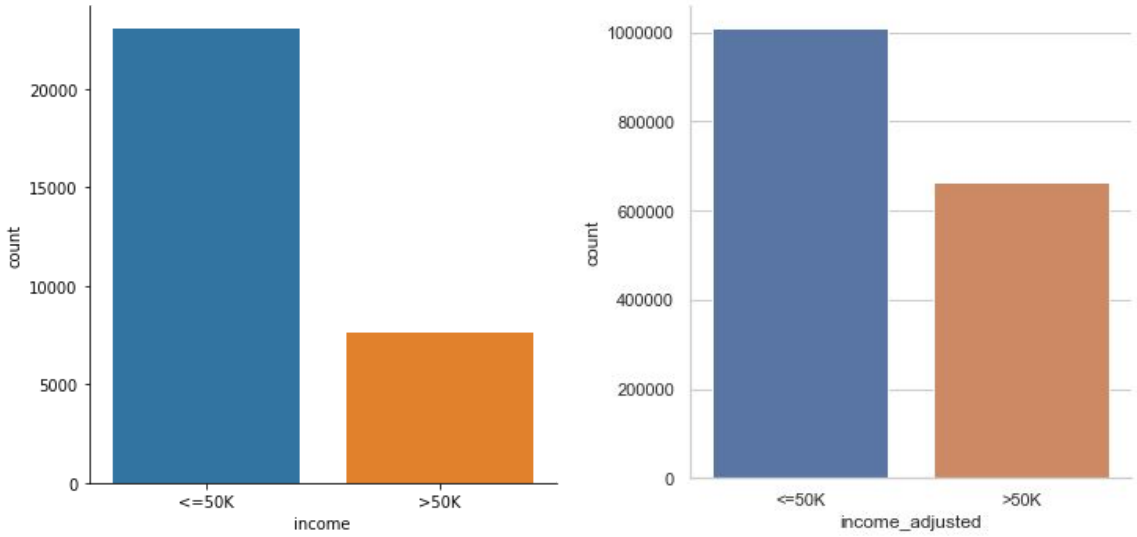
- Median Home Value in PUMA (Census ACS)
- Median Gross Rent in PUMA (Census ACS)

- Unemployment Rate in PUMA (Census ACS)
- Health Insurance Coverage Rate in PUMA (Census ACS)
- Educational Attainment Levels in PUMA (multiple columns, Census ACS)
- Incarceration Rate per 100,000 in state (Bureau of Justice Statistics)

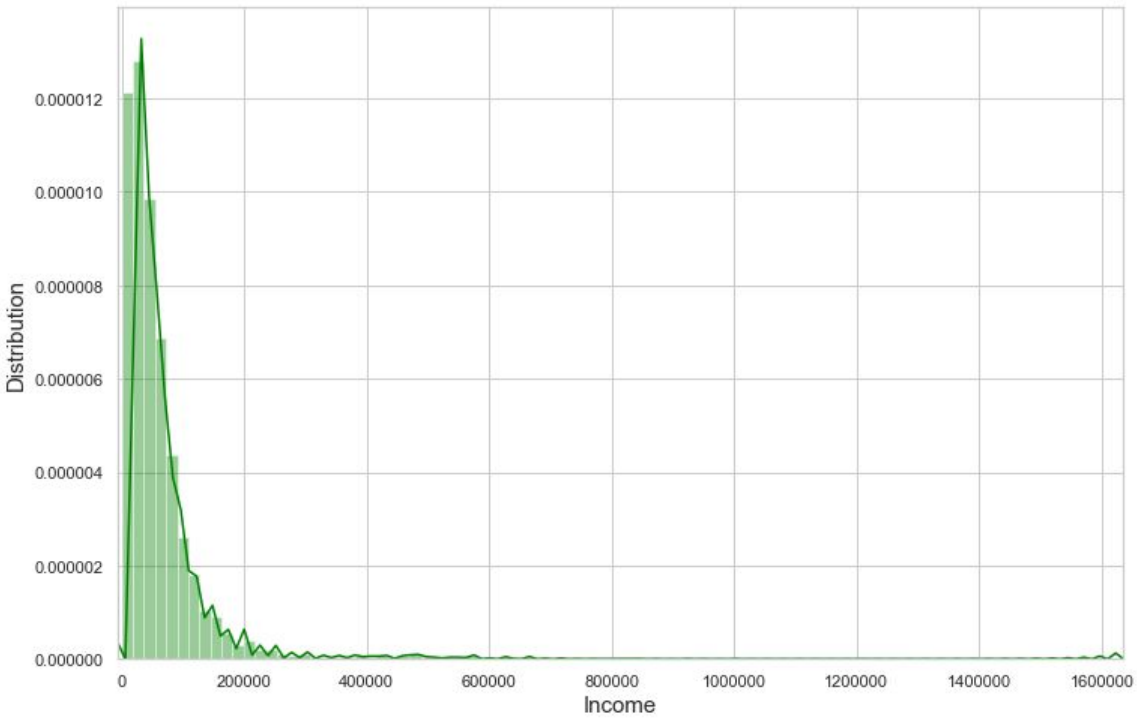
Dataset Analysis

The dataset analysis summarized below provides a deeper look into the final dataset that we have published, with the intent of allowing users to become more familiar with the data before using it.

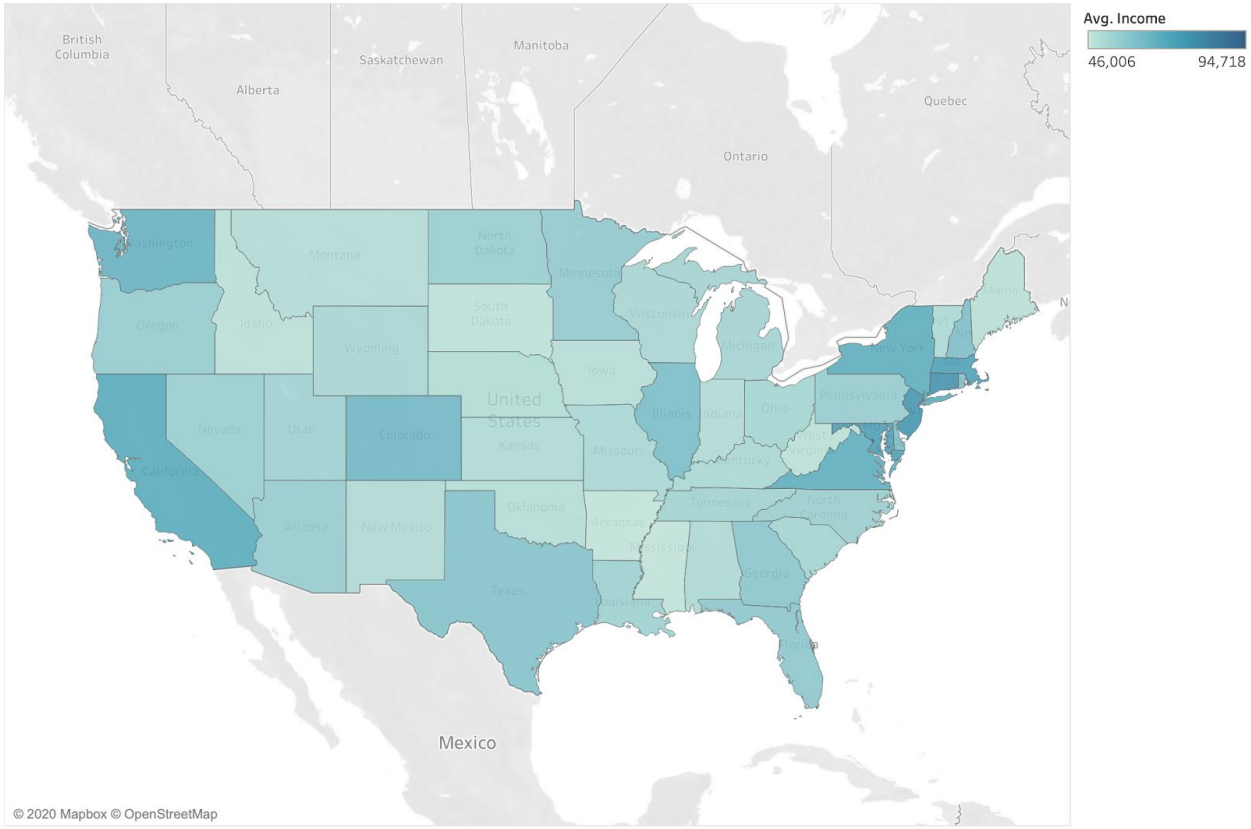
The final published dataset has a total of 28 columns and 1.7M rows. The primary variable is income, which carries a range of $104 to $1.6M. Income as a continuous variable is an improvement over the 1994 dataset, which only coded income as a binary variable, being greater than or less than $50,000 annually. The following figures compare income as a binary variable and demonstrate the shift in gross income between 1994 (left) and 2019 (right).



While in both samples, the data is skewed toward those earning less than $50,000, we can see that by 2019 the bias is less extreme, with the proportion of the population earning more than $50,000 increasing from 24% in 1994 to 40% in 2019. Still, we can clearly see the long tail of income when looking at the full distribution of the continuous income variable (below). We can see quite easily that while the range of income is large, the vast majority of people in the sample report annual incomes of less than $200,000.

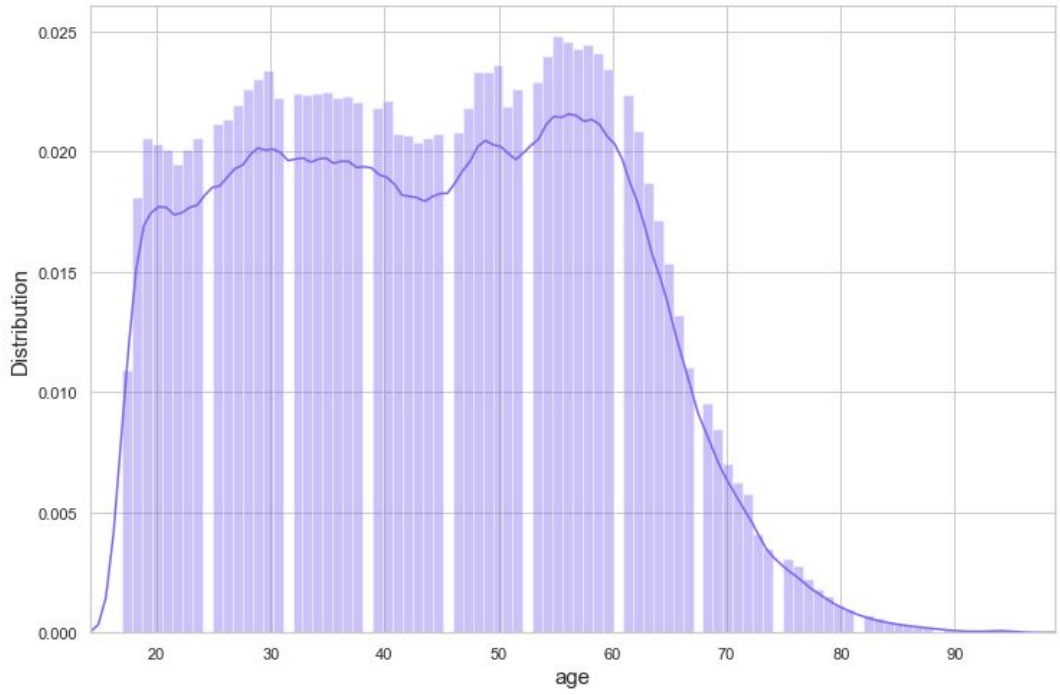


We can also look at this data geographically, such as in the map below which shows average income by state.

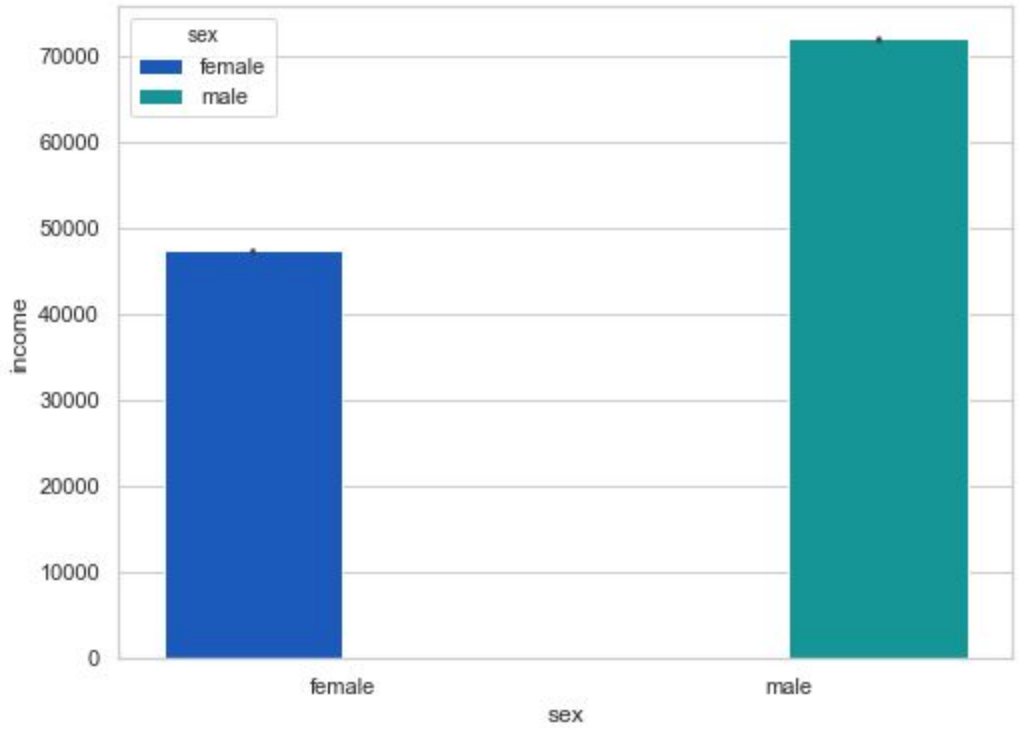


The figure above shows the high level of variance for incomes geographically within the United States, with coastal states such as California, Washington, Connecticut, New Jersey, and New York earning high incomes close to $80,000, and southern states like Arkansas and Mississippi having lower incomes closer to $50,000.

Looking at other variables besides income proves interesting as well, such as the below distribution of age from the 2019 sample. We see a fairly even distribution between ages 18 and 60, followed by a sharp decline in frequency. Another interesting facet is a spike (increased frequency) of people between the ages of 50 and 60, a manifestation of the aging Baby Boomer generation.

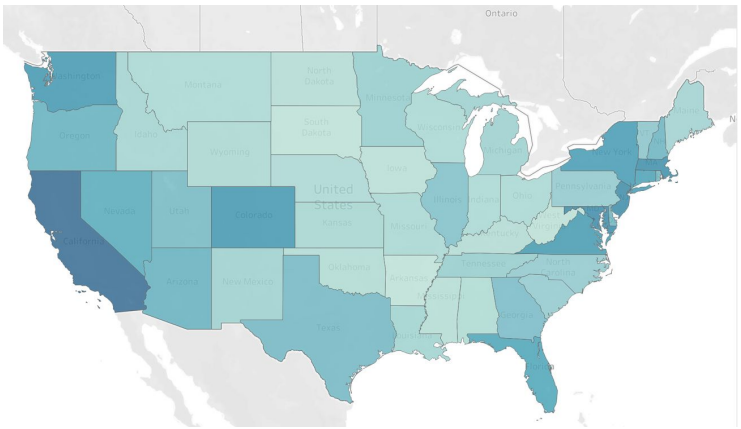


Another illuminating way of looking at the data is by combining variables. The below chart combines income with gender, and demonstrates the wide income gap between males and females in the US, with the mean female income of under $50,000 and the mean male income above $70,000.



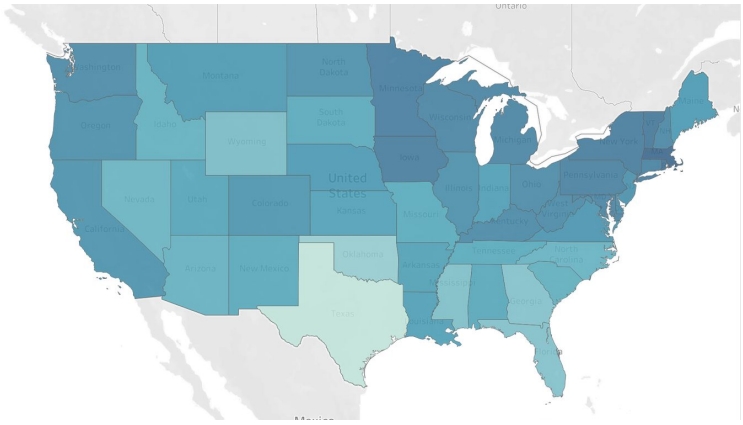
In addition to the census variables included in the 2019 sample, we also supplemented our dataset by adding data related to home values, health insurance coverage rates, unemployment rates, educational attainment rates, and incarceration rates by PUMA.

Avg. Gross Rent



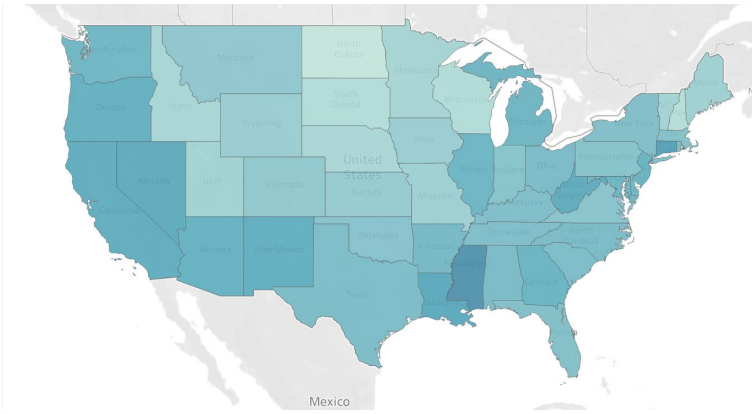
California has the highest avg. gross rent of all states, but other coastal states are not far behind. States in the midwest and south tend to have relatively lower gross rent prices.

Avg. Insured Rate



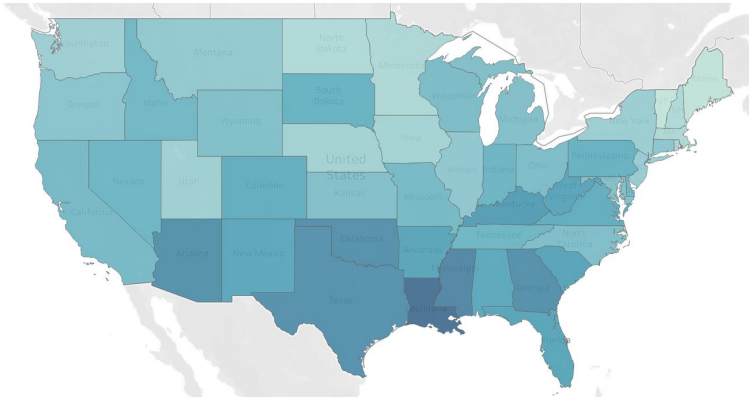
It appears that most states have very high rates of health insurance coverage, with Texas being the notable exception at 82% insured. Other states in the south also have relatively lower insurance coverage rates, but none lower than Texas.

Unemployment Rate



With regard to unemployment, many coastal states appear to have higher unemployment rates than midwestern states and states with low population density and fewer urban areas. However, the highest unemployment rate falls in the state of Mississippi in the south.

Incarceration Rate



Finally, coming to incarceration rates as provided by the US Bureau of Justice Statistics, it appears that southern states have the highest rates of incarceration in the country. With Louisiana recording as many as 1,300 incarcerated individuals per 100,000 in the population. Northern states have incarceration rates well below 1%.

Dataset Evaluation

This new dataset is intended to be used as a machine learning teaching tool, in the same way as the original “adult” dataset. In order to measure the improvements that our work had on the first income dataset, we employed an iterative modeling approach which allowed us to independently estimate the value of each individual improvement, and finally measure the value of all the data enhancements in aggregate.

This process began with establishing baseline models on the original 1994 dataset. We evaluated several models but ultimately chose logistic regression to use as our model to compare against variations. We saw a strong performance on this dataset, with the model predicting income greater than or less than $50,000 with 83% accuracy.

Our first and most significant change was updating the dataset from 1994 to 2018 (and subsequently 2019 when it was released mid-workstream), choosing the same attributes or their equivalents in the newer survey. Running the same baseline models on the updated data from 2019 yielded an interesting result: a decline in performance across key model evaluation metrics, including accuracy which dipped to 75%. However, despite this decline, other less quantifiable improvements from this update outweigh the performance loss. First and foremost, the dataset in its 2019 form is more reflective of the state of the country and income distribution today. Further, the 2019 update includes income as a continuous variable, rather than the binary variable indicating income greater than or less than $50,000. This change also has several benefits, primarily by offering flexibility on the target variable. Users of the new dataset have the option to transform this variable into any format that may meet their modeling needs. This may include reverting to the binary indicator, defining income buckets specific to their use case, predicting income decile or percentile, or even predicting approximate income as a continuous variable.

After the 2019 update, we began supplementing our data with further variables which our literature review indicated would provide improved predictability on income. We joined each of these datasets independently to our 2019 update and ran the same baseline models on the result. The below data scorecard presents the individual improvement across key model evaluation KPIs. The result is an average 1% improvement in accuracy, precision, recall for each supplemental dataset, which also provided justification for including the supplement in the final published dataset.

Finally, after concluding which supplemental datasets would remain and which provided no incremental value, we joined all the data into one final dataset ready for publication. We ran the baseline models one final time over this dataset, which yielded a final accuracy of 78%, 3% higher than the 2019 update alone, and effectively demonstrating the added value of expanding the dataset scope beyond the Census PUMS data alone.

Dataset Scorecard

Data	1994	2019	Incarceration	Home Value / Gross Rent	Unemployment	Health Insurance	Educational Attainment	Final Dataset
Accuracy	0.83	0.75	0.76	0.76	0.75	0.76	0.76	0.78
F1 Score	0.81	0.75	0.75	0.76	0.75	0.75	0.76	0.77
Precision	0.82	0.75	0.75	0.76	0.75	0.75	0.76	0.77
Recall	0.83	0.75	0.76	0.76	0.75	0.76	0.76	0.78

Discussion and Conclusions

Through the process outlined above, we created the new census income prediction dataset and made it available open-source at this link. The accompanying web page includes supplemental materials such as the data dictionary, this whitepaper, and documentation discussing the racial wealth gap in greater detail.

This is an effort that could be standardized and completed on a yearly basis, as the Census is updated yearly with new estimates as is incarceration data. However, there are obstacles to recreating the dataset that have little to do with process and more to do with data collection issues stemming from the 2020 pandemic and controversy surrounding the Census more broadly.

How the pandemic changed 2020 Census data collection
The 2020 Census encountered unique challenges both political and otherwise; the Trump administration attempted to add a citizenship question, and the pandemic reduced the capacity of door-to-door counting. Together, these threaten an accurate population count and foreshadow a misallocation of elected representatives and funding.

One of the biggest issues with conducting the Census during the pandemic is the slower enumeration of data collection. Rather than collecting responses door-to-door, there was a heavy reliance on the internet and mail-in forms. This caused tremendous problems for those lacking internet access, leaving out a large number of people. Compounding this, the administration placed a deadline on data collection for the end of October, an unprecedented target in the history of Census data collection and production.

These are issues that mainly affect the short-term data collection, however, the possibility of an undercount has severe implications for underrepresented communities throughout the country.

Collecting population data through the Census allows for the distribution of federal funding to state and local governments. If there is an undercount in a particular state, the federal funding for that state diminishes significantly. The fear of an undercount has raised significantly due to the pandemic, as people living in rural communities, minority urban communities, and other vulnerable groups may not have completed the 2020 Census for various reasons, including lack of online access. This problem is compounded by the fact that these groups are historically difficult to count.

The U.S. Census had been treated as an unbiased microcosm of the true characteristics of the United States—until now. It appears that more and more individuals are doubting the data, whether due to discrepancies in the 2020 count, the racist history of certain questions, or fringe conspiracy theories. Discussions about how these topics should be accounted for in the Census without creating distrust in government data collection procedures are still up for debate and will affect the future of Census data collection for decades to come.

Lessons Learned

In creating the dataset we ran into several challenges of note. The first regards the systematization of PUMA codes across the Census. The Census, being a government entity, is extremely well documented and easy to access, however, there are inconsistencies with how these codes are represented across datasets. The PUMS data (individualized Census responses) had PUMA codes that were truncated, without the state code at the beginning. Our supplemental data (educational attainment, insurance coverage rates, etc) all came from the American Community Survey (ACS), and the PUMA codes for that data were included in full. This presented a challenge to the initial data exploration and joins, as the PUMS data needed to be modified so that PUMA codes would match across all states, and to the ACS data. While the Census includes monumental documentation, this small difference was not codified and resulted in null values after the first join.

Establishing the baseline models and measuring our succession dataset enhancements posed its own set of challenges. In order to fairly evaluate the value in each change to the dataset, we needed to keep all else equal, including our methods of evaluation. This meant several iterations of our baseline models, in order to ensure that the same models could be used across all versions of the dataset. Further, the early model evaluations must necessarily leave some room for improvement, which meant models that over-performed (such as Random Forest) were not ideal for the task at hand. Ultimately, a logistic regression model proved to be the most pragmatic model to deploy for our needs. We did, however, use several models in the final evaluation of our master dataset to ensure its utility as a machine learning teaching tool.

Aside from dealing with government data sets and comparing the effects of new variables on the baseline model, there is a wider perspective gained on the impact of this dataset and the use of socio-demographic data in general. Importantly, we did not apply machine learning to a social justice problem, instead we incorporated a social justice framework into machine learning pedagogy. The changes we made to the original dataset improved the surface-level problem of a persistent, outdated dataset, but did not alter the potential for false or superficial conclusions.

This was never the goal; models do not explain all of the variance in an attribute. Rather, the changes we made defined a structure and process around dataset creation while engaging with social justice. We hope that students accessing this dataset in future will think critically about the conclusions they draw from modelling, and be meticulous in the language they use to describe their findings. It is dangerous to be casual with causality; these models suggest relationships between variables but they do not explain why these relationships exist. This matters especially when dealing with data that involves real people, where implicit bias is always present: conclusions arising from machine learning are a means to an end, but not the end itself.

Appendix A - US Census Data and Links

PUMS Technical Documentation
https://www.census.gov/programs-surveys/acs/technical-documentation/pums/documentation.html

PUMS Data Dictionary
https://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMS_Data_Dictionary_2019.pdf

PUMS File Structure
https://www.census.gov/programs-surveys/acs/technical-documentation/pums/filestructure.html

PUMA Map Viewer
https://tigerweb.geo.census.gov/tigerweb/

2019 PUMS 1-year Readme
https://www2.census.gov/programs-surveys/acs/tech_docs/pums/ACS2019_PUMS_README.pdf

Appendix B - Papers/scholarly articles citing the original UCI "Adult" dataset

(All links below come from the UCI archive - authors of this paper are not aware of new links)

Manuel Oliveira. Library Release Form Name of Author: Stanley Robson de Medeiros Oliveira Title of Thesis: Data Transformation For Privacy-Preserving Data Mining Degree: Doctor of Philosophy Year this Degree Granted. University of Alberta Library. 2005.

Aristides Gionis and Heikki Mannila and Panayiotis Tsaparas. Clustering Aggregation. ICDE. 2005.

Rakesh Agrawal and Ramakrishnan ikant and Dilys Thomas. Privacy Preserving OLAP. SIGMOD Conference. 2005.

Dan Pelleg. Scalable and Practical Probability Density Estimators for Scientific Anomaly Detection. School of Computer Science Carnegie Mellon University. 2004.

Ke Wang and Shiyu Zhou and Ada Wai-Chee Fu and Jeffrey Xu Yu. Mining Changes of Classification by Correspondence Tracing. SDM. 2003.

Douglas Burdick and Manuel Calimlim and Jason Flannick and Johannes Gehrke and Tomi Yiu. MAFIA: A Performance Study of Mining Maximal Frequent Itemsets. FIMI. 2003.

Bart Hamers and J. A. K Suykens. Coupled Transductive Ensemble Learning of Kernel Models. Bart De Moor. 2003.

James Bailey and Thomas Manoukian and Kotagiri Ramamohanarao. Fast Algorithms for Mining Emerging Patterns. PKDD. 2002.

Dennis P. Groth and Edward L. Robertson. An Entropy-based Approach to Visualizing Database Structure. VDB. 2002.

Eibe Frank and Geoffrey Holmes and Richard Kirkby and Mark A. Hall. Racing Committees for Large Datasets. Discovery Science. 2002.

Stephen D. Bay. Multivariate Discretization for Set Mining. Knowl. Inf. Syst, 3. 2001.

Zhiyuan Chen and Johannes Gehrke and Flip Korn. Query Optimization In Compressed Database Systems. SIGMOD Conference. 2001.

Stephen D. Bay and Michael J. Pazzani. Detecting Group Differences: Mining Contrast Sets. Data Min. Knowl. Discov, 5. 2001.

Nikunj C. Oza and Stuart J. Russell. Experimental comparisons of online and batch versions of bagging and boosting. KDD. 2001.

Jinyan Li and Guozhu Dong and Kotagiri Ramamohanarao and Limsoon Wong. DeEPs: A New Instance-based Discovery and Classification System. Proceedings of the Fourth European Conference on Principles and Practice of Knowledge Discovery in Databases. 2001.

Dan Pelleg and Andrew W. Moore. Mixtures of Rectangles: Interpretable Soft Clustering. ICML. 2001.

Jie Cheng and Russell Greiner. Comparing Bayesian Network Classifiers. UAI. 1999.

John C. Platt. Using Analytic QP and Sparseness to Speed Training of Support Vector Machines. NIPS. 1998.

Ron Kohavi. Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. KDD. 1996.

Gabor Melli. A Lazy Model-Based Approach to On-Line Classification. University of British Columbia. 1989.

David R. Musicant and Alexander Feinberg. Active Set Support Vector Regression.

David R. Musicant. DATA MINING VIA MATHEMATICAL PROGRAMMING AND MACHINE LEARNING. Doctor of Philosophy (Computer Sciences) UNIVERSITY.

Chris Giannella and Bassem Sayrafi. An Information Theoretic Histogram for Single Dimensional Selectivity Estimation. Department of Computer Science, Indiana University Bloomington.

Masahiro Terabe and Takashi Washio and Hiroshi Motoda. The Effect of Subsampling Rate on S 3 Bagging Performance. Mitsubishi Research Institute.

References

Aliprantis, Dionissi, and Daniel R. Carroll. “What Is Behind the Persistence of the Racial Wealth Gap?” website. Federal Reserve Bank of Cleveland, February 28, 2019. https://www.clevelandfed.org/newsroom-and-events/publications/economic-commentary/2019-economic-commentaries/ec-201903-what-is-behind-the-persistence-of-the-racial-wealth-gap.aspx

Aprosio, Alessio Palmero, Stefano Menini, and Sara Tonelli. “Creating a Multimodal Dataset of Images and Text to Study Abusive Language.” arXiv Labs: Cornell University, May 5, 2020. https://arxiv.org/abs/2005.02235

Ashman, Hero, and Seth Neumuller. “Can Income Differences Explain the Racial Wealth Gap? A Quantitative Analysis.” Review of Economic Dynamics 35 (January 2020): 220–39. https://doi.org/10.1016/j.red.2019.06.004

Brown, Anna. “The Changing Categories the U.S. Census Has Used to Measure Race.” Pew Research Center. Pew Research Center, September 3, 2020. https://www.pewresearch.org/fact-tank/2020/02/25/the-changing-categories-the-u-s-has-used-to-measure-race/.

Darity, William, Darrick Hamilton, Mark Paul, Alan Aja, Anne Price, Antonio Moore, and Caterina Chiopris. “What We Get Wrong About Closing the Racial Wealth Gap.” Durham, NC: Samuel DuBois Cook Center on Social Equity, Duke University, 2018.

Escobar, Natalie, and Leah Donnella. “Who Are We? We're Finding Out Together.” NPR. NPR, March 31, 2020. https://www.npr.org/sections/codeswitch/2020/03/31/823881121/who-are-we-were-finding-out-together.

Fabusuyi, Tayo, Robert Hampshire, and Zhen (Sean) Qian. “Profiling Commuters’ Travel Behavior in the Pacific States of the Continental U.S.” *World Conference on Transport Research*, Mumbai, 2019.

Hamilton, Darrick, and William A. Darity. “The Political Economy of Education, Financial Literacy, and the Racial Wealth Gap.” Review 99, no. 1 (February 2017): 59–76. https://doi.org/10.20955/r.2017.59-76

Jacobsen, Linda A, Mark Mather, and Beth Jarosz. “Coronavirus and the 2020 Census: Where Should College Students Be Counted?” Population Reference Bureau, October 28, 2019. https://www.prb.org/covid-19-and-the-2020-census-where-should-college-students-be-counted/.

Jason Gauthier, History Staff. “Decennial Census - History - U.S. Census Bureau.” Accessed November 15, 2020. https://www.census.gov/history/www/programs/demographic/decennial_census.html.

Kinney, Satkartar K. and Karr, Alan, Public-Use vs. Restricted-Use: An Analysis Using the American Community Survey (February 1, 2017). US Census Bureau Center for Economic Studies Paper No. CES-WP-17-12. http://dx.doi.org/10.2139/ssrn.2909935

Kohavi, Ronny, and Barry Becker. “UCI Machine Learning Repository - Adult Data Set.” UCI Machine Learning Repository: Adult Data Set. UC Irvine, 1996. https://archive.ics.uci.edu/ml/datasets/adult.

Lynn, Stuart. “US Census + Machine Learning to Map Entirely New Populations Using CARTO.” CARTO Blog, March 7, 2016. https://carto.com/blog/creating-segments-from-the-census/.

Mather, Mark, and Paola Scommegna. “Why Is the U.S. Census So Important?” Population Reference Bureau, September 17, 2019. https://www.prb.org/importance-of-us-census/.

McKernan, Signe-Mary, Caleb Quakenbush, Caroline Ratcliffe, Emma Kalish, and C. Eugene Steuerle. Nine Charts about Wealth Inequality in America (Updated), October 4, 2017. https://apps.urban.org/features/wealth-inequality-charts/

Meseguer-Brocal, Gabriel, Alice Cohen-Hadria, and Geoffroy Peeters. “Creating DALI, a Large Dataset of Synchronized Audio, Lyrics, and Notes.” Transactions of the International Society for Music Information Retrieval 3, no. 1 (June 11, 2020): 55–67. https://doi.org/10.5334/tismir.30.

Noel, Nick, Duwain Pinder, Shelley Stewart, and Jason Wright. “The Economic Impact of Closing the Racial Wealth Gap,” August 7, 2020. https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-economic-impact-of-closing-the-racial-wealth-gap

Policy Network, Asset Building. “Overall - Racial Wealth Gap Infographic (2019).” Asset Building Policy Network, June 14, 2019. http://assetbuildingpolicynetwork.org/resources/racial-wealth-gap-infographic-2019/

Kimberly Proctor, Shondelle M. Wilson-Frederick, and Samuel C. Haffer. Health Equity. Dec 2018.82-89.http://doi.org/10.1089/heq.2017.0036

Qurishee, Murad Al, Weidong Wu, Babatunde Atolagbe, Joseph Owino, Ignatius Fomunung, and Mbakisya Onyango. “Creating a Dataset to Boost Civil Engineering Deep Learning Research and Application.” Engineering 12, no. 03 (March 2020): 151–65. https://doi.org/10.4236/eng.2020.123013.

Schmid, Eric. “The 2020 Census Is Underway, But Nonbinary And Gender-Nonconforming Respondents Feel Counted Out.” St. Louis Public Radio, March 17, 2020. https://news.stlpublicradio.org/politics-issues/2020-03-17/the-2020-census-is-underway-but-nonbinary-and-gender-nonconforming-respondents-feel-counted-out.

Seamster, Louise, and Raphaël Charron-Chénier. “Predatory Inclusion and Education Debt: Rethinking the Racial Wealth Gap.” Social Currents 4, no. 3 (June 2017): 199–207. https://doi.org/10.1177/2329496516686620.

Shapiro, Thomas, Tatjana Meschede, and Sam Osoro. “The Roots of the Widening Racial Wealth Gap: Explaining the Black-White Economic Divide.” Waltham, MA: Institute on Assets and Social Policy, Brandeis University, 2013.

Wang, Hansi Lo. “What You Need To Know About The 2020 Census.” NPR. NPR, April 1, 2019. https://www.npr.org/2019/03/31/707899218/what-you-need-to-know-about-the-2020-census.