US Census Data Analysis

Vasu Sharma
Department of Computer Science
University of Colorado, Boulder
Boulder, Colorado
vasu.sharma@colorado.edu

Yan Zhan
Department of Economics
University of Colorado, Boulder
Boulder, Colorado
yan.zhan@colorado.edu

Sutianjie Zhou
Department of Economics
University of Colorado Boulder
Boulder, Colorado
sutianjie.zhou@colorado.edu

Abstract

Census data provides important information about the whole nation. Census data can provide socio-economic information for citizens of the country. IPUMS[1] is an organization that preserves and harmonizes U.S. census microdata and provides easy access to this data with enhanced documentation. In this project, we aim to focus on determining socio-economic trends based on the dataset collected by IPUMS. These trends are collected based on person-level attributes such as education, age, gender, race, employment status, marital status, labor force participation etc. Using data mining techniques, we have tried to identify certain socio-economic trends by analysing a person's educational background and their identity group to predict their earning potential, factors correlating to higher income, change in rate of return on education and income inequality over the past few decades. We were successfully able to achieve those objectives using data visualization, data cleaning and integration and classification algorithms.

1. Introduction

The questions that we are more interested in relate to the economic situation of an individual in the US. The IPUMS[1] census data extract that we generated includes person-level attributes such as age, sex, race, highest education level, region, marital status, labor force participation among others. We wanted to see the correlation of these attributes with income of an individual which we have demonstrated through data visualization. Some of these attributes play a major role in determining earning potential attributed to an individual. That is where we used data mining techniques like classification algorithms to see how well can one estimate the income given certain information related to the background of an individual. The other question that we were interested in was to see the change in income inequality over a decade which we demonstrated through the use of boxplots and data extracts which consisted of the same attributes but differentiated by year.

These questions are very important to understand as they form the backbone of various social and welfare programs that are adopted across the different states in the country. The federal and state governments continuously do analysis on the census data to improve or start programs that help the communities. Some of this analysis is used by the government to decide where to build new infrastructure such as schools, hospitals and homes. It is also used to assess the wellbeing of communities as well as identifying and assisting low-income and marginalized communities. One important use case of this analysis is to also determine which social programs need to be allocated more funding so that their effectiveness is widespread.

Moreover, studies related to the census data raise the level of political discourse and bring certain issues to public light that might have been swept under the rug.

2. Literature Survey

Paper using IPUMS dataset for analysis of social-economic questions have been published in many peer review journals in recent decades.

Sheng et al.[2] performed data mining on census data using CART which is a decision tree based classification model. Their paper was based on using this model to classify inhabitants in the provinces of Chengyang and Laixi.

Federal and local governments[3] use census dataset to make or tailor policies so as to adhere to various communities and minority groups present in the country.

Jaison R. et al [4] used the IPUMS U.S. dataset collection to identify the rate of return on college education. They used income after graduation as a parameter to judge rate of return and discussed the problems regarding stagnant wages and the decline of wages among people without a college degree. They concluded that return has remained high in spite of rising tuition and falling earnings because the wages of those without a college degree have also been falling, keeping the college wage premium near an all-time high while reducing the opportunity cost of going to school.

In Autor, Dorn, and Hanson (2013)[5], they use the PUMA-level (Public Use Microdata Area) data from IPUMS to explain the influence of increasing import from developing countries on the local U.S. labor market. They first generate local labor and employment indexes, then they map the increasing of U.S. imports from developing countries on each local labor market, to figure out the changing of the manufacturing employment indexes. They compare the local employment indexes before and after the increasing shock of import from developing countries (especially after China returned to the WTO in 2001), and explain that the increasing imports from developing countries reduce the U.S. local employment in manufacturing industries. In Autor, Dorn, and Hanson (2019)[7], they also use the IPUMS local data to show the relation between local employment situation of young people and their marital status. The increasing import of U.S. from developing countries reduces the local market manufacturing jobs and changes local employment structure. As the number of local manufacturing jobs decreases, the wage gap among young people increases (the reduction of manufacturing will make the local job market polarization), and the first marriage age of young people increases because more individuals want to spend more time for a potential ‘better’ partner.

3. Dataset

The dataset is collected from the organization IPUMS[1] which provides consistent data with documentation. We are using the U.S. Census Person-Level data[8] for our project. The 2019 sample is used for the visualization as well as for building the classification model. The size of this sample is around 806 MB and it has 3239553 rows and 23 columns. Some of these columns contain important person-level information like race, age, gender, income, education, marital status and labor force participation. These columns are chosen from a wide range of columns and are specifically focused towards an individual rather than an entire household or family. The attributes we are interested in and will study include gender, age, race, education level, location, occupation, marital status, wage

income, and total income. We are also making use of the 2010 extract of the Person-Level data to do income disparity analysis. All the attributes involved in the 2010 extract are the same when compared with the 2019 extract.

4. Proposed Work

In this project, based on the dataset, we will analyze the following questions

1. What is the relation among a person's educational background, their identity group and earning potential?
2. What factors are generally found to correlate with higher income?
3. What is the rate of change in income inequality in the U.S. over the past decade?

To answer these questions, we will first perform data cleaning and preprocessing where we will remove redundant data or apply cleaning techniques to replace duplicate or Null values with measures of central tendency. Then we will transform the dataset by performing binning on income attributes using tax brackets and normalize the data set. After data cleaning and transformation, we will work on building the classification model to identify the income bin for an individual based on their socioeconomic background. We will start with support and confidence calculation and analysis between the interesting attributes (whether to receive higher education, married or not, low income or not) and individuals' demographic characteristics (gender, age, race, etc.). Then, we will apply the accuracy measurement in this class to study the relation between education and never being married before. Bayesian classification will also be applied to study the marital status and income gap when we mark different individuals as low-income, middle-income and high-income. After we compare the dependent attributes with objects in different groups based on different attribute standards, we will use an ordinary linear regression model to check the effect of gender, race, and education level on individuals' income. As for education and marital status, we will use a logit/probit model to check the effect of education, gender, racial, and other demographic attributes(in this case, we will set each age from 25 to 45 as fixed effect).

5. Evaluation Methods

We created a machine learning model which acts as a classifier for predicting earning potential by taking into account a person's socio-economic background. This prediction model is evaluated against the test split of the original dataset. In the preprocessing phase, earning potential is binned based on the income attribute, so the model predicts the income bin and it is evaluated based on confusion matrix, precision, recall, accuracy and F1 score.

6. Tools

The project work is divided into certain parts and all of those parts require specific tools. The tools for some of those parts are as follows :-

1. **Data Collection and Cleaning** :- We made use of pandas and numpy. Pandas is used for efficiently carrying out data cleaning and transformation tasks while numpy is used for performing computational tasks on the data.
2. **Data Visualization** :- Matplotlib and seaborn is used to visualize data in the form of histograms, bar charts, scatter plots and confusion matrix.
3. **Classification and Analysis** :- Sklearn is used for the classification aspect of the

project. For analysis on the classification process, sklearn's evaluation metrics like f1 score, precision and recall are used.

7. Main Techniques Applied

As part of our data mining project, we applied the following techniques to achieve our results :-

1. **Data Cleaning** :- Certain attributes have missing or no information when it comes to our data extract from IPUMS[1]. Every attribute description along with its codes is elaborately mentioned in the IPUMS documentation. The code mentioned gives us an idea of what sort of values to expect while parsing the data extract. When there are missing or no values, codes seem to identify it with a certain value and that value is unique for every attribute. In our case, the dataset contained missing information for around 0.05% of the total rows across all attributes. We chose to ignore those percentage of rows instead of using other techniques like filling in the statistical mean or median since that amount of rows seemed to be quite insignificant compared to the rest of the dataset.

2. **Data Preprocessing** :- INCTOT is one of the attributes which signified the total amount of income earned by an individual over the last year. It combined the income earned as wage, interests earned on assets as well as income received from certain welfare programs to calculate that variable. We wanted to do a classification task so we binned that INCTOT variable to classify individual person total income to be classified to one of the bins. We decided to keep the number of bins to six and they were decided roughly according to the U.S tax brackets.

3. **Data Visualization** :- Since one of our main objectives was to correlate certain attributes to the earning potential of an individual, we used bar plots and scatter plots to show that correlation. We handpicked some of the attributes that we thought would play a key role in this correlation. The attributes which seemed to show no correlation during visualization were rejected to be a part of the classification process.

4. **Income Disparity Analysis** :- The disparity analysis is compared over a decade by plotting boxplots to see the differences in income gap over that time period. Boxplots were plotted on the INCTOT attribute since it is one of the better indicators to determine income inequality. Boxplots were chosen as the visualization tool because it incredibly summarizes the numerical attribute of INCTOT and the inequality comparison becomes much easier to interpret and understand.

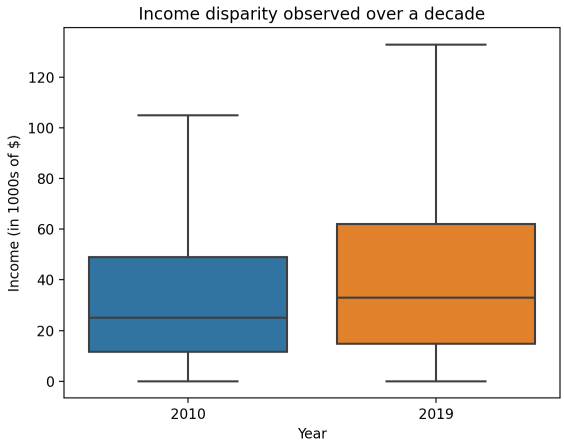
5. **Classification Analysis** :- One of our objectives was to determine the earning potential of an individual based on certain information regarding their background. The variable on which the classification is performed is INCTOT, which is already binned in the data preprocessing step. So our classifier's aim was to classify the individuals in the dataset in one of the bins. We went ahead with a decision tree classifier since it is quite easy to interpret and tends to give good results. Since we had already identified some of the key attributes in our data visualization process, this proved to be a good set of input to our classifier model to predict the INCTOT variable. The classifier model was evaluated based on its F1 score, recall and precision. Since our classification included multiple classes due to the presence of six bins in the INCTOT attribute, the scores were calculated on a class level. Along with this, F1 score was also calculated at an aggregated level using 'micro' as value to the parameter 'average' to get a single value which can determine the effectiveness of the classifier as a whole. The confusion matrix was plotted to showcase what is going on in the classifier.

8. Key Results

The key results of our project are separated in three categories. These three categories align with what we aimed to achieve by working on this project. They also help in providing a summarized view of the project.

8.1 Income Disparity Analysis

We wanted to see how the income gap has increased over a long period of time. Income gap can explain some of the most important economic and social issues prevalent in a country. We have plotted the income gap using a boxplot which is observed over a decade.

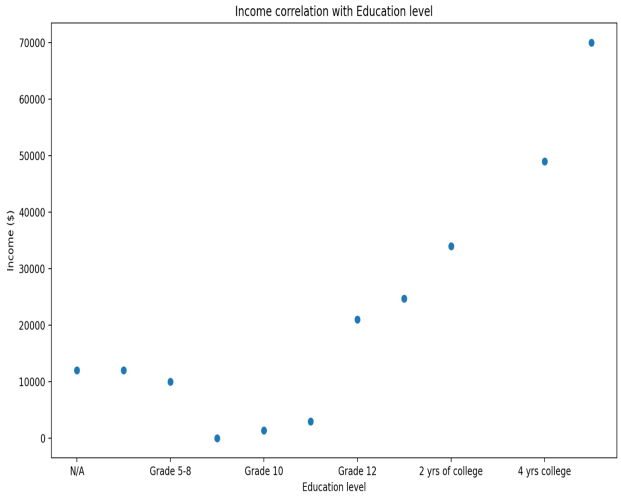


From the boxplot of income in 2010 and 2019, we can see that the wage-gap in 2019 is larger than that in 2010. From the boxplot we can also observe that the median income of individuals hasn't increased by a significant percentage but the income among the upper middle class to rich individuals has seen a significant growth. This is one of the reasons for the ever increasing income gap in the country.

8.2 Data Visualization: Income and Education, Gender, Race and Region

We have tried to obtain multiple visualizations of the IPUMS[1] dataset to achieve a better understanding and representation of the data. These visualizations make the job easier when it comes to milestone tasks like classification because we know about the major attributes influencing the classification variable which is the income bin of a person.

We wanted to see some correlations between income and other person level attributes in the IPUMS dataset.

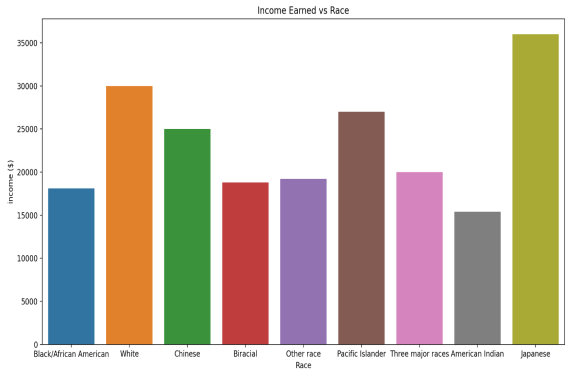


We first decided to do a scatter plot between income and educational background of a person. From the above visualization, we can see that income is positively correlated with the education level of a person. Another interesting fact that we observed is that having a college degree makes a significant difference in the earning potential of an individual.

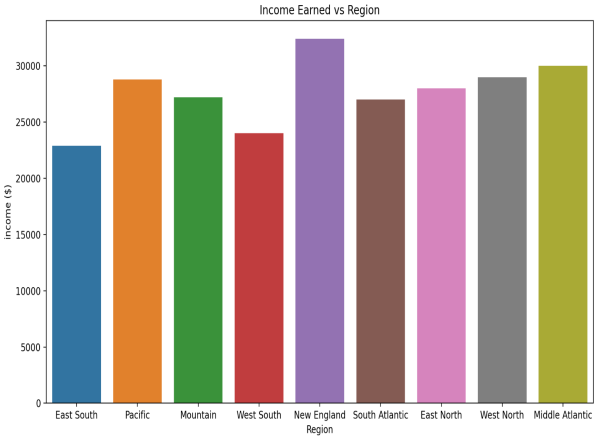
Our first step is to divide all individuals between 25 and 45 into 5 groups according to their education level: high-school dropout, high-school graduate, some college, college graduate, and post college. All income numbers are in dollars. The mean wage income of high-school dropout individuals is around 17000, that of high-school

graduates is around 26000, while that of college graduates is around 60000 and lastly that of post graduates is around 85000. We can infer from these results that education plays a significant role in income at a personal level. This shows that a college degree is really valuable. In our data, the mean income of high school dropout female is around 11000 compared with high school dropout male which is around 21000, that of high school graduate female is around 18000 compared with high school graduate male which is around 32200, that of a graduate female is around 46000 compared with graduate male which is around 76000. We can infer from these observations that the gender wage gap exists at each step of education level, especially at the higher education level where it tends to be much larger. The gender wage gap at high school dropout level is around 10000, while that at high school graduate level is around 14000 and finally at the college graduate level is around 30000.

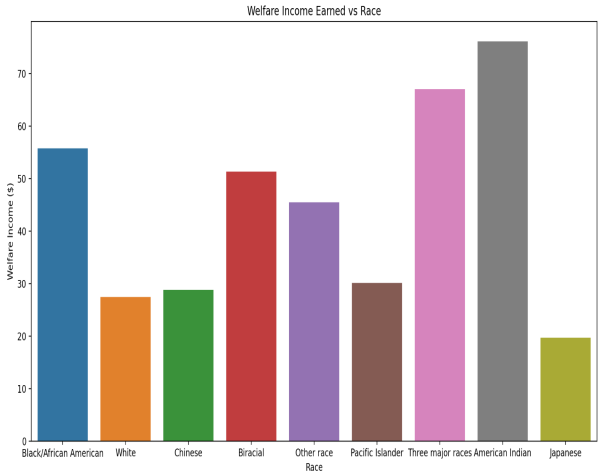
Along with the gender wage gap, we also try to figure out a potential racial wage gap that might exist. To figure out the details about the potential racial wage gap, we generated dummy variables of white and black and observed that the mean income of black high school dropout individuals is around 10000 compared with that of white high school dropout individuals which is around 17855.56, that of black high school graduates is around 19000 compared with white high school graduates which is around 28000. The mean value wage income of black some college education individuals is 27593.23, for the white counterpart is 36361.67. The mean value wage income of balck college graduation education level individuals is 45989.15, the white counterpart is 61614.39. The mean value wage income of black post college education level is 66128.77, that counterpart of white individuals is 85537.81. These observations do hint at a potential racial wage gap prevalent in the country.



We plotted some other attributes against the total income of a person. One of them being the race of an individual. We tried to see how income is visualized among the various communities in the US. This is important because it indicates how different social and cultural communities are faring in the economic progression. These stats and visualizations help state and federal governments to create social programs that are aimed at helping certain communities which are in dire need of those assistance programs.



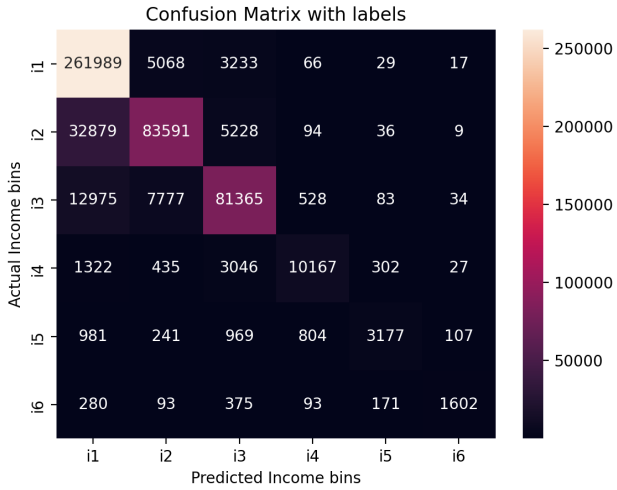
Location is one of the key attributes in determining the earning potential of an individual and that can be seen from the above visualization as well. The above distribution is around the median income earned by person and distributed by region. The New England region seems to have the highest median income earners among all the regions based on the IPUMS[1] dataset of year 2019.



People who are usually in the low income brackets of a country are one of the key portions of the total number of people on welfare and assistance programs. In our dataset, we utilized the variable, INCWELFR that indicates the amount of income obtained over a year from welfare or social assistance programs. We wanted to see the relationship between this particular variable and the race of an individual. This information tells about some specific communities that generally require good types of assistance programs either at the federal level or the state level. As is apparent from the above visualization, on average American Indians usually receive the most amount of income through welfare programs, followed by people who classify themselves as part of three major races and then immediately followed by Black Americans. The INCWELFR is also one of the key contributors in determining the income bin of an individual since people who generally receive more welfare from the government tend to have incomes in the lower brackets of the income split.

8.3 Classification Analysis

The classifier used for identifying the earning potential of an individual is the decision tree.



The above figure showcases the confusion matrix after running the decision tree classifier on our dataset. We experimented with gini impurity and entropy as the criterion parameter for the decision tree to decide the optimum split. We experienced slightly better results with entropy as the criterion value. This can be explained by the fact that entropy is more complex since it makes use of logarithms while gini is computationally less expensive. Although the time required for training under entropy was slightly higher than training under gini, it was still not a significant factor to move towards gini impurity.

The income bins as showcases in the figure correspond to the following :

i1 -> [0, 25000]
i2 -> [25000, 50000]
i3 -> [50000, 125000]
i4 -> [125000, 200000]
i5 -> [200000, 350000]
i6 -> [350000 and above]

All these numerical values are represented in U.S. dollars ($). These bins are a result of the data preprocessing step and they are created on top of the INCTOT attribute. All the rows with missing or no values as part of the dataset were excluded after the data cleaning step since they formed an insignificant percentage when compared with the entirety of the dataset. Bining was one of the key factors for converting this problem into a classification problem. These income bins are based on the distribution of the INCTOT variable in the dataset as well as the U.S. tax brackets.

In the data visualization, we showcased certain attributes that seem to have some correlation with the income earned by an individual. Among the ones that we picked for the classification task were state, region, labor force participation, marital status, highest level of education, employee status, sex, age, race and income earned from welfare. This is a good mix of attributes to represent the socio-economic background of an individual. We also experimented with other attributes during the data visualization phase but they didn't seem to correlate much with income of an individual. So we finalized on this set of attributes for our decision tree classifier.
The following table represents the results achieved with our classifier. These results are evaluated on the test split of the dataset which amounts to 519193 data records. Since this is a multiclass classification problem, the table represents the scores across each class.

Score metrics across classes

scores	i1	i2	i3	i4	i5	i6
precision	0.84	0.85	0.86	0.86	0.83	0.89
recall	0.96	0.68	0.79	0.66	0.50	0.61
f1	0.90	0.76	0.82	0.75	0.63	0.72

Certain observations that we can infer from the table above include the precision values which seem to be higher across all classes while recall seems to vary from very high for certain classes to moderate for others.

We also calculated an aggregated representation of the f1 score using 'micro' as the value for the parameter 'average' in the f1 score function. Micro is a better representation of the aggregation when there exists imbalance among the classes. That is why we went ahead with this parameter. The aggregated f1 score value turned out to be :- **0.8511**

These results were achieved after doing multiple experiments with the type of attributes to consider, adjusting the error parameters and tuning the decision tree classifier. As we can see that decision tree seems to work quite well considering the bins are significantly imbalanced in the training dataset which is why this classifier performs really well for some classes while performing moderately for the others.

The unemployment and labor force participation are another two important attributes which reflect individuals' situation. From our data, high school dropout individuals' labor force participation rate is 59.86%, high school graduate individuals' labor force participation rate is 74.86%, some college individuals' labor force participation rate is 83.57%, college graduate individuals' labor force participation rate is 89.56%, post college individuals' labor force participation rate is 92.52, so there is a clear positive correlation of education level with labor force participation. Then we classified the individuals of our data not only based on their education level but also their gender. The high school dropout female's labor force participation rate is 51.12%, the high school dropout male's labor force participation rate is 66.12%, the high school graduate female's labor force participation rate is 69.22%, the high school graduate male's labor force participation rate is 78.85%, the some-college female's labor force participation rate is 79.24%, the some-college male labor force participation rate is 88.08%, the college graduate female's labor force participation rate is 84.96%, the college graduate male's labor force participation rate is 94.94%, the post-college female's labor force participation rate is 89.88%, and the post-college male's labor force participation rate is 96.20%. From all these results, although we control the education level, we still can conclude that the male's labor force participation is higher than female's in each education level group. In other words, the gender gap in labor force participation exists. Another point is whether the racial gender exists in labor force participation, so we classify individuals in our data set not only based on their education level but also whether the individual is black or white. From our results, black high school dropout individuals' labor force participation rate is 41.79%, white high school dropout individuals' labor force participation rate is 61.00%, black high school graduate individuals' labor force participation rate is 65.63%, white high school graduate individuals' labor force participation rate is 76.57%, black some-college individuals' labor force participation rate is 81.20%, white some-college labor force participation rate is 84.37%, black college graduate individuals' labor force participation rate is 91.55%, white college graduate individuals' labor force participation rate is 90.29%, black post-college individuals' labor force participation rate is 94.34%, white post-college individuals' labor force participation rate is 93.47%. These results show that, for individuals without a college degree, the black-white gap in labor force participation exists, however, for individuals with a college degree or advanced education degree, the labor force participation is similar between black individuals and white individuals.

Then we will focus on unemployment status. To better figure out the unemployment situation among individual groups with different gender, race, and education levels, we classify the individuals based on their gender, race, and education level. The high-school dropout individuals' unemployment rate is 5.06%, the high-school graduate individuals' unemployment rate is 4.32%, the some-college individuals unemployment rate is 3.32%, the college graduate individuals' unemployment rate is 2.10%, the post-college individuals unemployment rate is 1.57%. Therefore, there is a clear negative correlation between unemployment and education level. The high school dropout female unemployment rate is 5.39%, the high school dropout male unemployment rate is 4.83%, the high school graduate female unemployment rate is 4.41%, the high school graduate male unemployment rate is 4.26%, the some-college female unemployment rate is 3.30%, the some-college male unemployment rate is 3.35%, the college graduate female unemployment rate is 2.03%,

the college graduate male unemployment rate is 2.18%, the post-college female unemployment rate is 1.65%, the post-college male unemployment rate is 1.46%. From the results, we could find that the gap between female and male in unemployment rate is narrow as the education level increases, the lower the education level, the larger the gender gap in unemployment rate. As for the racial analysis, the high school dropout balck individuals' unemployment is 7.60%, the high school dropout white individuals' unemployment rate is 4.87%, the high school graduate balck individuals' unemployment rate is 6.71%, the high school graduate white individuals' unemployment rate is 3.90%, the some-college black individuals' unemployment rate is 5.56%, the some-college white individuals' unemployment rate is 2.90%, the college graduate black individuals' unemployment rate is 3.76%, the college graduate white individuals' unemployment rate is 1.88%, the post-college balck individuals' unemployment rate is 2.62%, the post-college white individuals' unemployment rate is 1.37%. All these shows that the racial gap exists in unemployment even though we control the education level of individuals.

9. APPLICATIONS

Through our project we aimed to see various socio economic trends which are quite important to consider when evaluating the quality of life led by the individuals of that country.

One of the important observations to point out is the income disparity analysis. The boxplots clearly show how the income gap has significantly increased over a decade while the median income has seen just a slight increase. This is one of the key factors why the poor to lower middle class in the country has remained stagnant while the upper middle to rich class has seen significant prosperity. We also saw how some of the attributes related to a person's socio economic background becomes a key factor in determining their earning potential.

This sort of analysis on census data can provide the following uses :-

1. **Federal and State governments** :- Governments need to constantly monitor and analyze the trends using the census data to make effective policy decisions for the citizens of the country. Currently, governments at the state and federal level do rely on census data analysis to come up with social and welfare programs as well as for designing policies when it comes to minority and underrepresented groups. Assessing economic well being, creating specialized programs for elderly and veterans are all reliant on the accuracy of the census data and its analytics.

2. **Business use cases** :- Businesses need to keep track of the trends using census data to market their products and services. For example, businesses that manufacture household items like home furnishings and washing machines would benefit from analysing household trends across the country and businesses in the real estate domain would significantly benefit from knowing the composition of households and region as well as the rate at which people are selling their homes.

3. **Academic Research** :- Census data is the main source of truth when it comes to researching on topics like income inequality, increasing homelessness as well as identifying certain trends over a period of time.

REFERENCES

[1] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

[2] Bin Sheng and Sun Gengxin, "Data Mining in census data with CART," *2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, 2010, pp. V3-260-V3-264, doi: 10.1109/ICACTE.2010.5579631.

[3] https://www.census.gov/econ/overview/go0100.html

[4] Abel, Jaison R. and Deitz, Richard, Do the Benefits of College Still Outweigh the Costs? (August 1, 2014). Current Issues in Economics and Finance, Vol. 20, No. 3, 2014, Available at SSRN: https://ssrn.com/abstract=2477864

[5] David, H., David Dorn, and Gordon H. Hanson. "The China syndrome: Local labor market effects of import competition in the United States." American Economic Review 103.6 (2013): 2121-68.

[6] Steven Ruggles, Sarah Flood, Sophia Foster, Ronald Goeken, Jose Pacas, Megan Schouweiler and Matthew Sobek. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0

[7] Dorn, David, and Gordon Hanson. "When work disappears: Manufacturing decline and the falling marriage.market value of young men." American Economic Review: Insights 1.2 (2019): 161-78.

[8] https://usa.ipums.org/usa-action/variables/group