

# US Census Data Analysis

**Members** Vasu Sharma  
Yan Zhan  
Sutianjie Zhou

**Github** [https://github.com/vash6618/data\\_mining\\_project](https://github.com/vash6618/data_mining_project)

# Project Description

Our project focuses on determining socioeconomic trends based on US-census dataset collected by IPUMS.

These trends are collected based on person level attributes like :-

- Education
- Age
- Gender
- Race
- Employment status
- Marital status
- Labor force participation



# Questions Sought to Answer

- Does the US have a larger income inequality now than a decade ago?
- How can a person's educational background and identity group predict their earning potential?
- What factors are generally found to correlate with higher income?

\* the estimation of the **rate of return to education** is simply the difference between earnings on **educational level k** minus earnings on **educational level k-1** divided by n years of **schooling** at **educational level k** and earnings on **educational level k-1**



# Dataset Overview

[IPUMS](#) (Integrated Public Use Microdata Series) is an individual-level population database that consists of microdata samples. IPUMS provides a consistent dataset with documentation.

We use the [U.S. Census Person-Level data](#) for our project. The dataset contains information such as education level, income, race, marital-status, age, gender, etc.


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
1	year	sample	serial	cbserial	hhwt	cluster	strata	gq	pernum	perwt	sex	age	marst	race	raced	educ	educd	empstat	empstatd	labforce	indnaics	inctot
2	2019	2019 ACS	1	2.019E+12	11	2.019E+12	220001	Other group	1	11	Male	39	Never marrie	Black/Africa	Black/Africa	Grade 10	Grade 10	Not in labor	Not in Labor	No, not in th	0	90C
3	2019	2019 ACS	2	2.019E+12	70	2.019E+12	100001	Group quart	1	70	Female	21	Never marrie	White	White	Grade 10	Grade 10	Not in labor	Not in Labor	No, not in th	0	15
4	2019	2019 ACS	3	2.019E+12	20	2.019E+12	110001	Other group	1	20	Male	19	Never marrie	Black/Africa	Black/Africa	1 year of coll	1 or more ye	Employed	At work	Yes, in the la	8131	14C
5	2019	2019 ACS	4	2.019E+12	79	2.019E+12	110001	Group quart	1	79	Male	77	Widowed	White	White	Grade 9	Grade 9	Not in labor	Not in Labor	No, not in th	0	227C
6	2019	2019 ACS	5	2.019E+12	53	2.019E+12	270101	Group quart	1	53	Male	41	Separated	Black/Africa	Black/Africa	Grade 9	Grade 9	Not in labor	Not in Labor	No, not in th	0	
7	2019	2019 ACS	6	2.019E+12	77	2.019E+12	200001	Other group	1	77	Male	18	Never marrie	Black/Africa	Black/Africa	Grade 12	Some colleg	Not in labor	Not in Labor	No, not in th	0	
8	2019	2019 ACS	7	2.019E+12	8	2.019E+12	270201	Group quart	1	8	Female	93	Widowed	White	White	Grade 12	Regular high	Not in labor	Not in Labor	No, not in th	0	360C
9	2019	2019 ACS	8	2.019E+12	15	2.019E+12	140001	Other group	1	15	Male	35	Never marrie	Black/Africa	Black/Africa	Grade 12	Regular high	Not in labor	Not in Labor	No, not in th	0	93C
10	2019	2019 ACS	9	2.019E+12	61	2.019E+12	210001	Group quart	1	61	Female	39	Divorced	White	White	4 years of co	Bachelor's di	Not in labor	Not in Labor	No, not in th	814	600C
11	2019	2019 ACS	10	2.019E+12	152	2.019E+12	130201	Other group	1	152	Female	18	Never marrie	Black/Africa	Black/Africa	Grade 12	Some colleg	Not in labor	Not in Labor	No, not in th	0	
12	2019	2019 ACS	11	2.019E+12	100	2.019E+12	260001	Group quart	1	100	Male	62	Divorced	Black/Africa	Black/Africa	1 year of coll	1 or more ye	Not in labor	Not in Labor	No, not in th	0	
13	2019	2019 ACS	12	2.019E+12	89	2.019E+12	30101	Other group	1	89	Male	19	Never marrie	Black/Africa	Black/Africa	1 year of coll	1 or more ye	Employed	At work	Yes, in the la	4481	39
14	2019	2019 ACS	13	2.019E+12	64	2.019E+12	240001	Other group	1	64	Female	19	Never marrie	Chinese	Chinese	Grade 12	Some colleg	Not in labor	Not in Labor	No, not in th	0	
15	2019	2019 ACS	14	2.019E+12	61	2.019E+12	210001	Group quart	1	61	Female	39	Divorced	White	White	4 years of co	Bachelor's di	Not in labor	Not in Labor	No, not in th	814	600C

# Data Preparation Work


- **Data Cleaning**

- Finding Null or missing values in the dataset
- Missing or null values amounted to 0.05 % of the entire dataset

- **Data Preprocessing**

- Data binning was performed on INCTOT variable in the dataset
  - Objective was to classify the individual's earning potential into income bins
  - Bin size was based on the data range as well as the US tax bracket.
- 

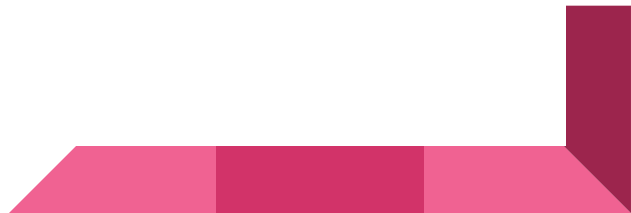
# Technologies to be used

- **Data collection, cleaning, preprocessing :-**
    - **Pandas** for efficiently carrying out data cleaning and transformation tasks on large datasets
    - **Numpy** for performing computational tasks on the data
  - **Data visualization :-**
    - Using **matplotlib** and **seaborn**: Box plots, bar charts, scatter plots
  - **Classification and Analysis**
    - Using Sklearn for classification tasks
    - Sklearn evaluation metrics like precision, accuracy and f1 score
- 

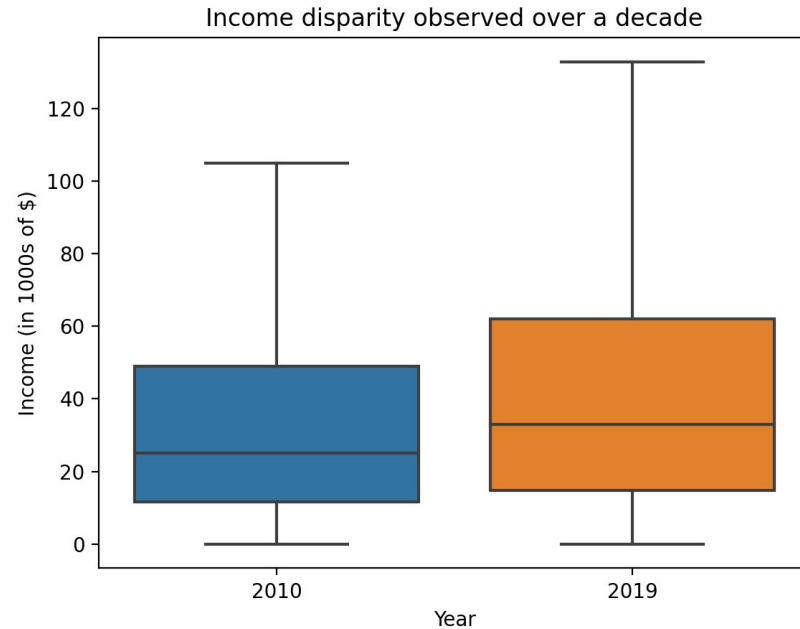
# Data Visualization: Income Disparity Analysis

- **Income Disparity Analysis**

- Boxplots were plotted on the INCTOT attribute since it is one of the better indicators to determine income inequality.
- Boxplots incredibly summarizes the numerical attribute of INCTOT and the inequality comparison becomes much easier to interpret and understand.

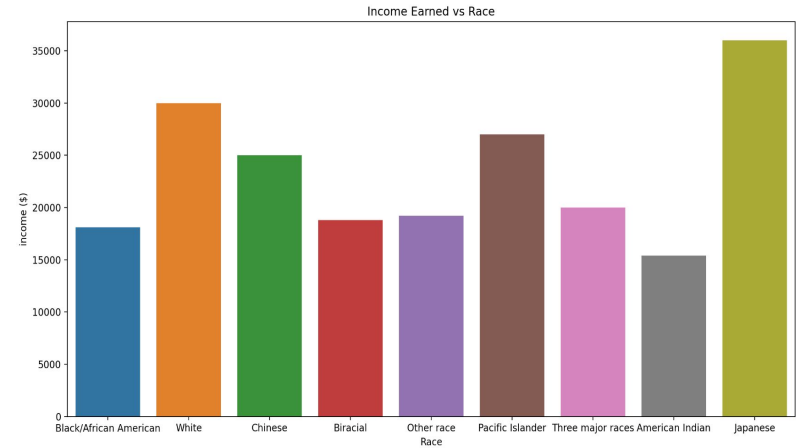
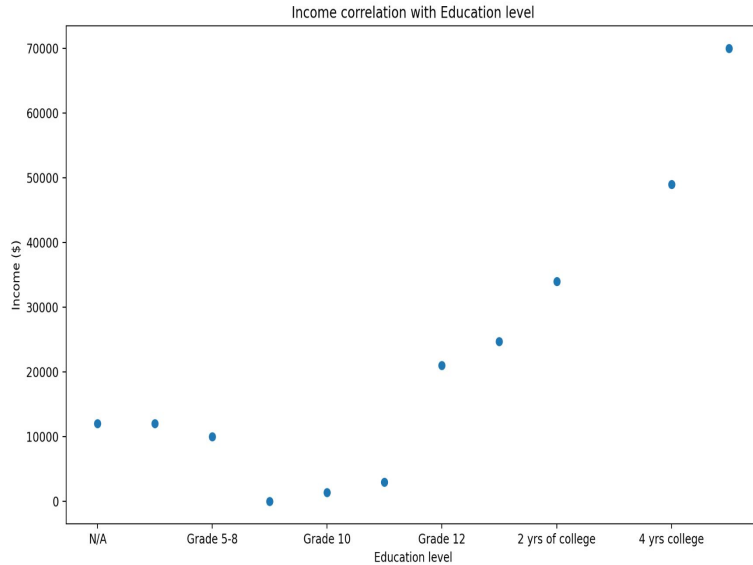


# Data Visualization: Income Disparity Analysis

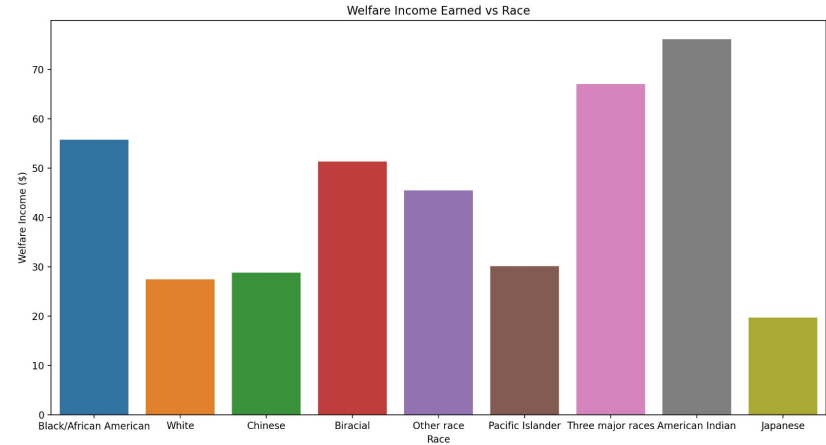
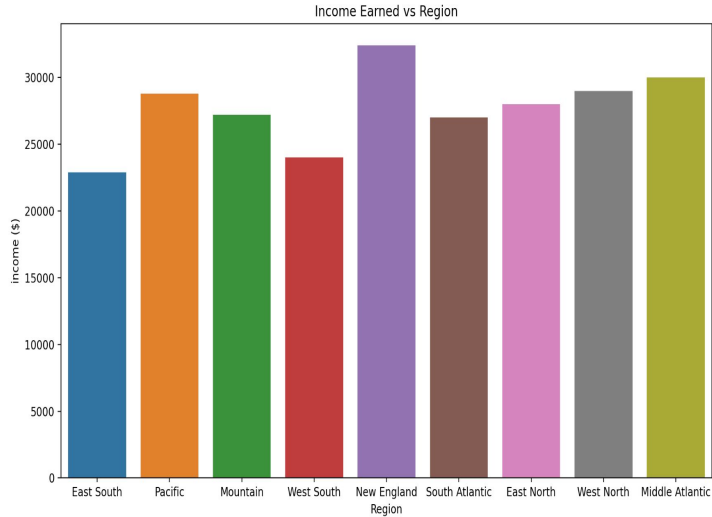




# Data Visualization: Correlation Analysis




# Data Visualization: Correlation Analysis

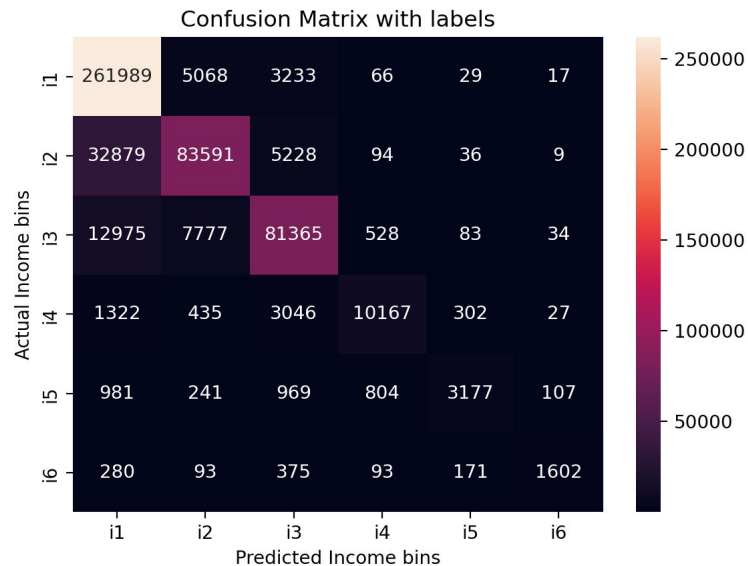


# Classification Analysis

- Decision Tree classification
- Aggregated F1-score
  - **0.8511 (using micro average)**

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



# Knowledge gained - 1

- The income disparity in the US has increased significantly during the last decade. This comes from the fast income growth among the upper middle class.
- The return to college education remains high in the US. In 2019, the average income of college graduates is more than two times higher than that of high school graduates.



## Knowledge gained - 2

- There are a strong correlation between income and gender, race and region.
- Using decision tree classifier, we calculate the precision values, recall values, and f1 scores for our classification of the income bins.



# Applications - 1

- **Federal and State governments**

- Assisting in creation or improving social and welfare programs
- Assess economic wellbeing of communities
- Allocate funding to different social programs such as adult education
- Policy designing related to minority groups



# Applications - 2

- **Business use-cases**

- Businesses manufacturing household items like home furnishings and washing machines benefit from analysing household trends across the country

- **Academic Research**

- Researching topics like income inequality, homelessness, trends over certain periods of time

