US Census Data Analysis

Members Vasu Sharma
Yan Zhan
Sutianjie Zhou

Github 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

Interesting Questions

How can a person’s educational background and identity group predict their earning potential?

What factors are generally found to correlate with higher income?

How does the rate of return on education* change in the past few decades?

Does the US have a larger income inequality now than twenty years ago?

* 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 plan to 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	900
3	2019	2019 ACS	2	2.019E+12	70	2.019E+12	100001	Group quarte	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	140
5	2019	2019 ACS	4	2.019E+12	79	2.019E+12	110001	Group quarte	1	79	Male	77	Widowed	White	White	Grade 9	Grade 9	Not in labor	Not in Labor	No, not in th	0	2270
6	2019	2019 ACS	5	2.019E+12	53	2.019E+12	270101	Group quarte	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 quarte	1	8	Female	93	Widowed	White	White	Grade 12	Regular high	Not in labor	Not in Labor	No, not in th	0	3600
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	930
10	2019	2019 ACS	9	2.019E+12	61	2.019E+12	210001	Group quarte	1	61	Female	39	Divorced	White	White	4 years of co	Bachelor's d	Not in labor	Not in Labor	No, not in th	814	6000
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 quarte	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 quarte	1	61	Female	39	Divorced	White	White	4 years of co	Bachelor's d	Not in labor	Not in Labor	No, not in th	814	6000

Previous Work

The federal and local governments use this dataset in order to:

- Identify where to build new infrastructure such as schools, hospitals and homes
- Assess economic wellbeing of communities
- Identify and assist low-income and marginalized populations
- Allocate funding to different social programs such as adult education
- Policy designing related to minority groups

Work Proposed

- **Data Cleaning and Preprocessing**
  - Looking for redundant, duplicate and Null values and applying cleaning techniques like removal or replacing with measures of central tendency
- **Data Transformation**
  - Binning the income data using tax brackets
  - Normalization of data set in order to facilitate correlation and prediction tasks
- **Classification Tasks**
  - Performing classification on earning potential considering a person’s socio-economic background
- **Correlation Analysis**
  - Understanding the various factors that correlate with income levels and education levels

Technologies to be used

- **Data collection and cleaning** :-
  - **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**
    - Histograms
    - Bar charts
    - Scatter plots
- **Classification and Analysis**
  - Using Sklearn or keras for classification tasks
  - Dimensionality reduction using PCAs or Autoencoders

Evaluation

Prediction model will be evaluated against the test split of the original dataset. The model will predict the income bins and the evaluation criteria will be based on the following parameters :-

- F1 score
- Precision
- Recall
- Accuracy