COMPUTATIONAL MATHS AND STATISTICS

Poster Presentation on Socioeconomic Dynamics Shaping Adult Income in the U.S

Anagh Sharma, Niyanta Pandey, Akshit Arora

Katz
Katz School
of Science and Health

Abstract

Income inequality in the United States demands solutions informed by a deep understanding of its root causes. This study tackles this challenge by analyzing anonymized census data to identify key demographic factors influencing adult income. We focus on whether an individual falls above or below a $50,000 income threshold.

By leveraging machine learning techniques, this research aims to illuminate the complex interplay between these demographics and income distribution.

In conclusion, this study utilizes machine learning to analyze census data and identify key demographic factors shaping income distribution. By providing policymakers with a deeper understanding of these factors, the research has the potential to inform the development of effective policies that promote a more equitable distribution of income in the US.

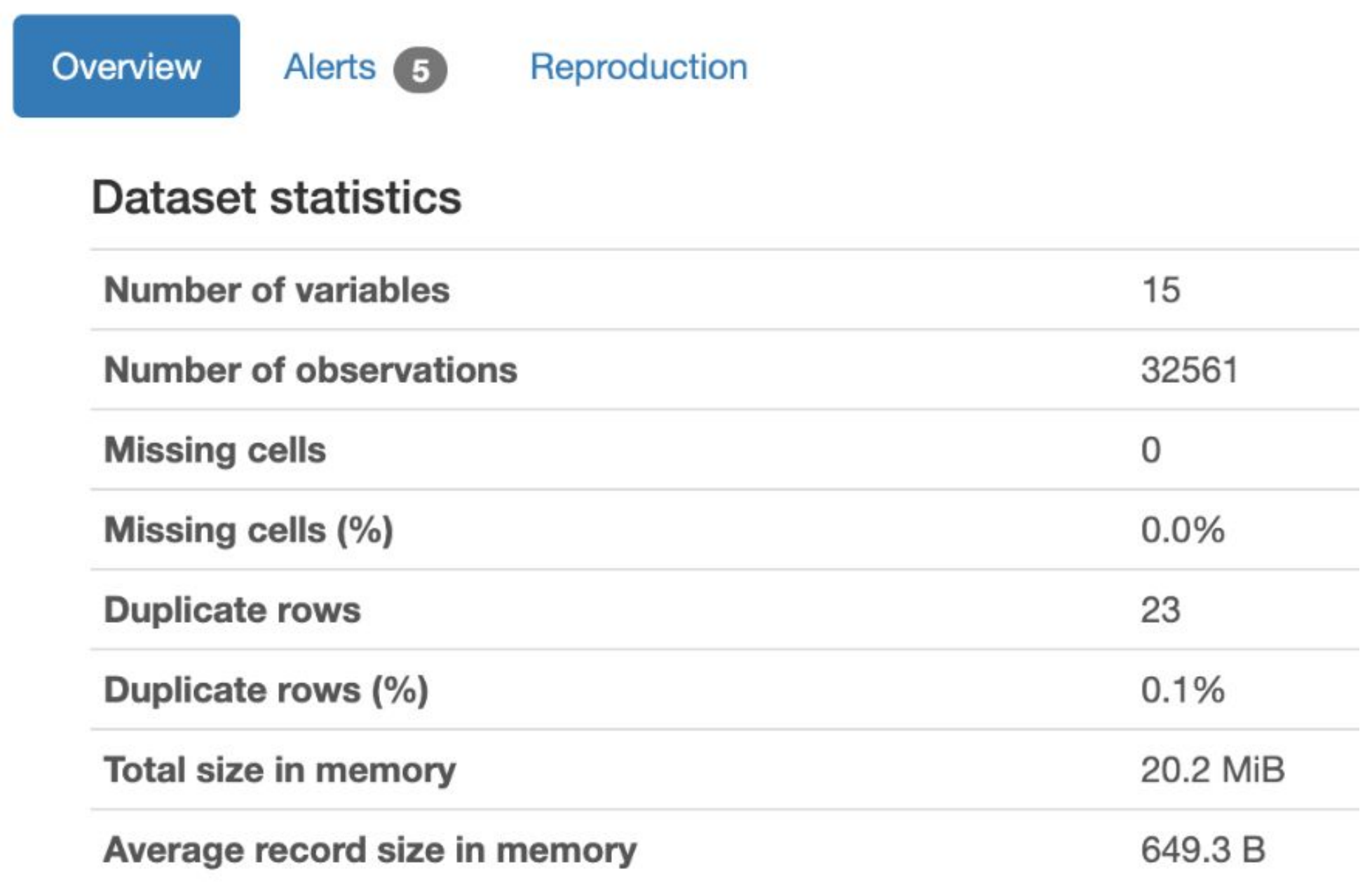
Introduction

The core question we aim to answer is: Which demographic characteristics most significantly influence whether an adult falls above or below a specific income threshold (e.g., $50,000)? By analyzing relationships between income and factors like age, education, and occupation, our research seeks to illuminate the underlying socio-economic landscape shaping income distribution in the US. This knowledge can then be harnessed by policymakers to develop targeted interventions that promote income equality and economic mobility.

Dataset

We used Google BigQuery to acquire this dataset which is a RESTful web service for analysing and downloading huge datasets.

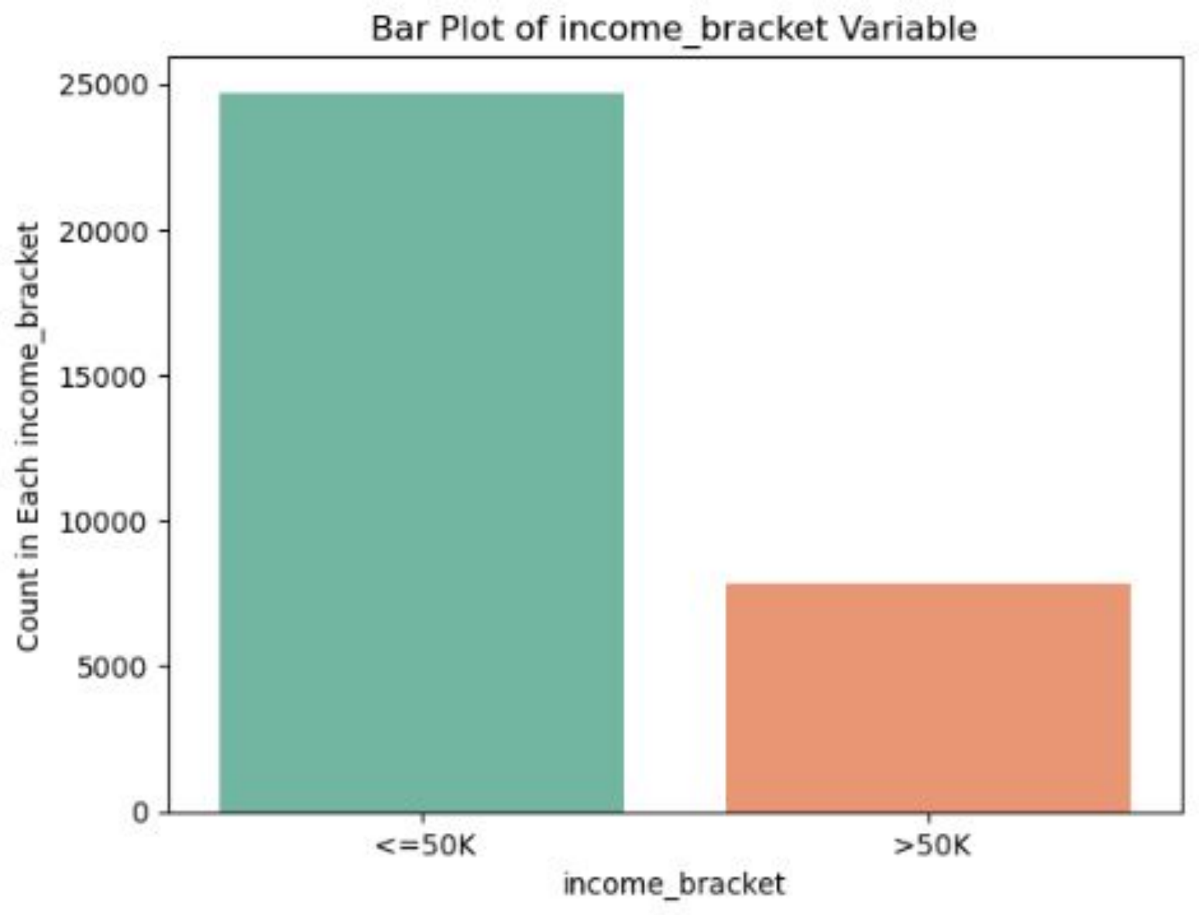
Data Profiling



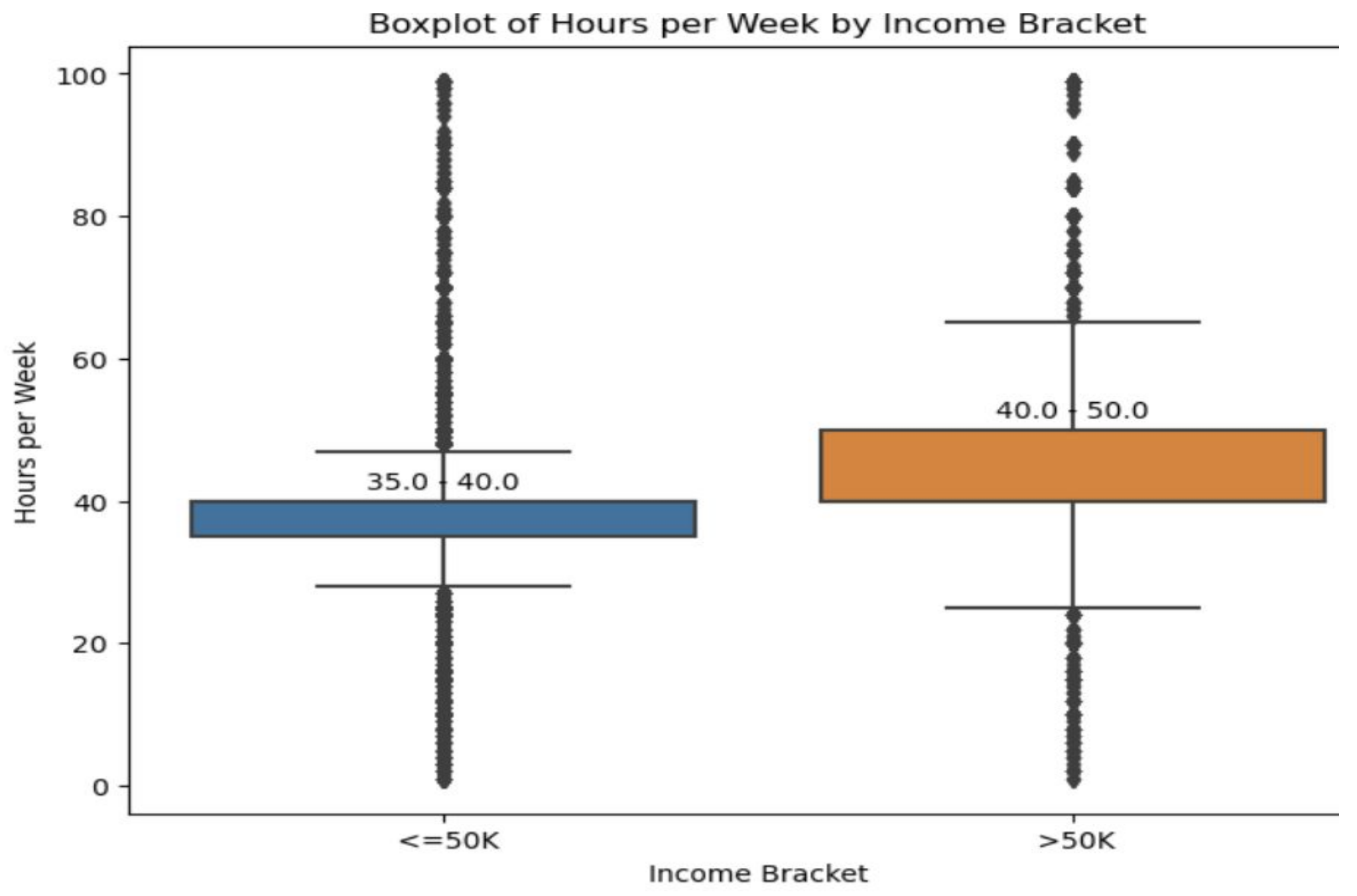
Dataset statistics

Number of variables	15
Number of observations	32561
Missing cells	0
Missing cells (%)	0.0%
Duplicate rows	23
Duplicate rows (%)	0.1%
Total size in memory	20.2 MiB
Average record size in memory	649.3 B

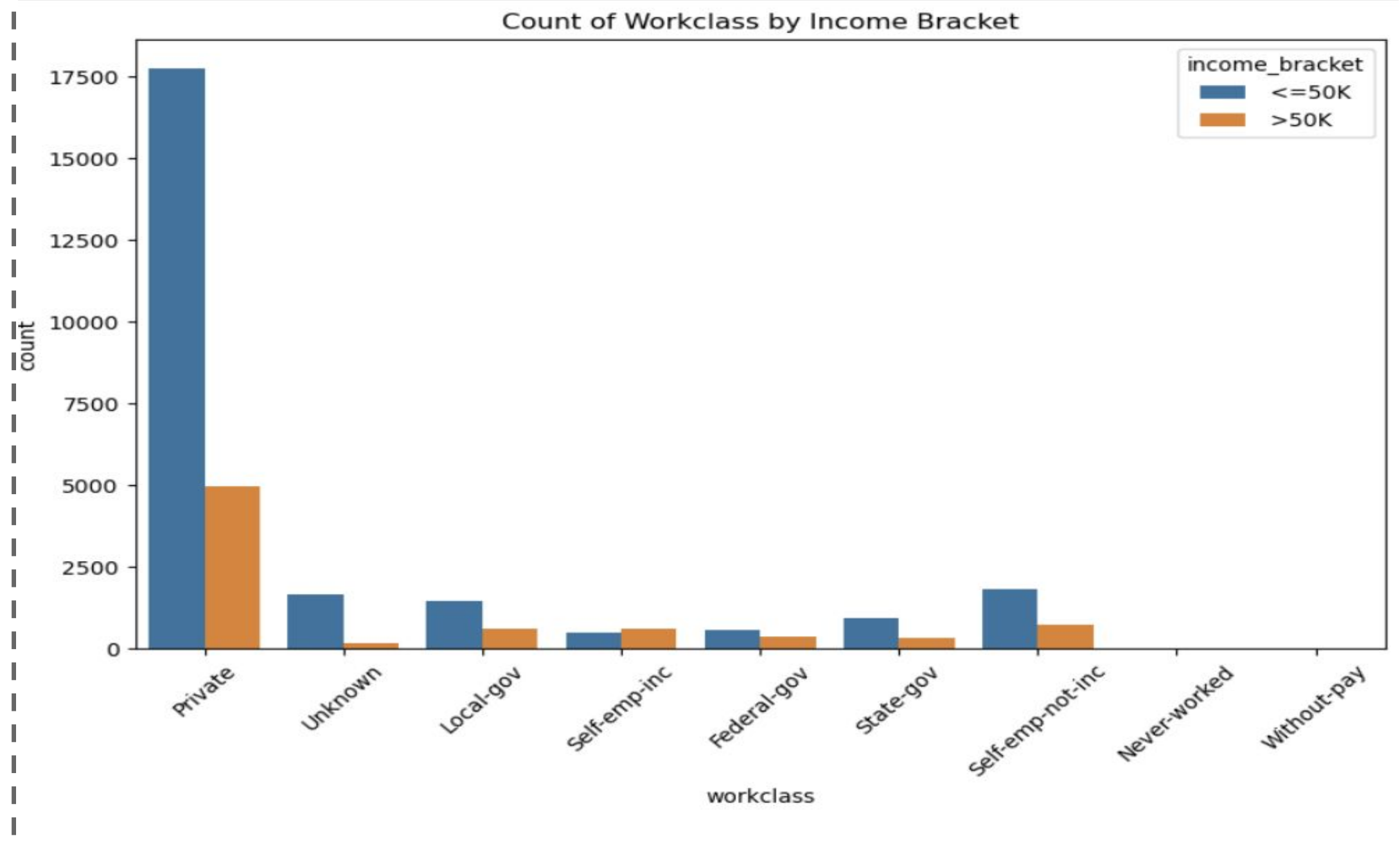
Exploratory Data Analysis



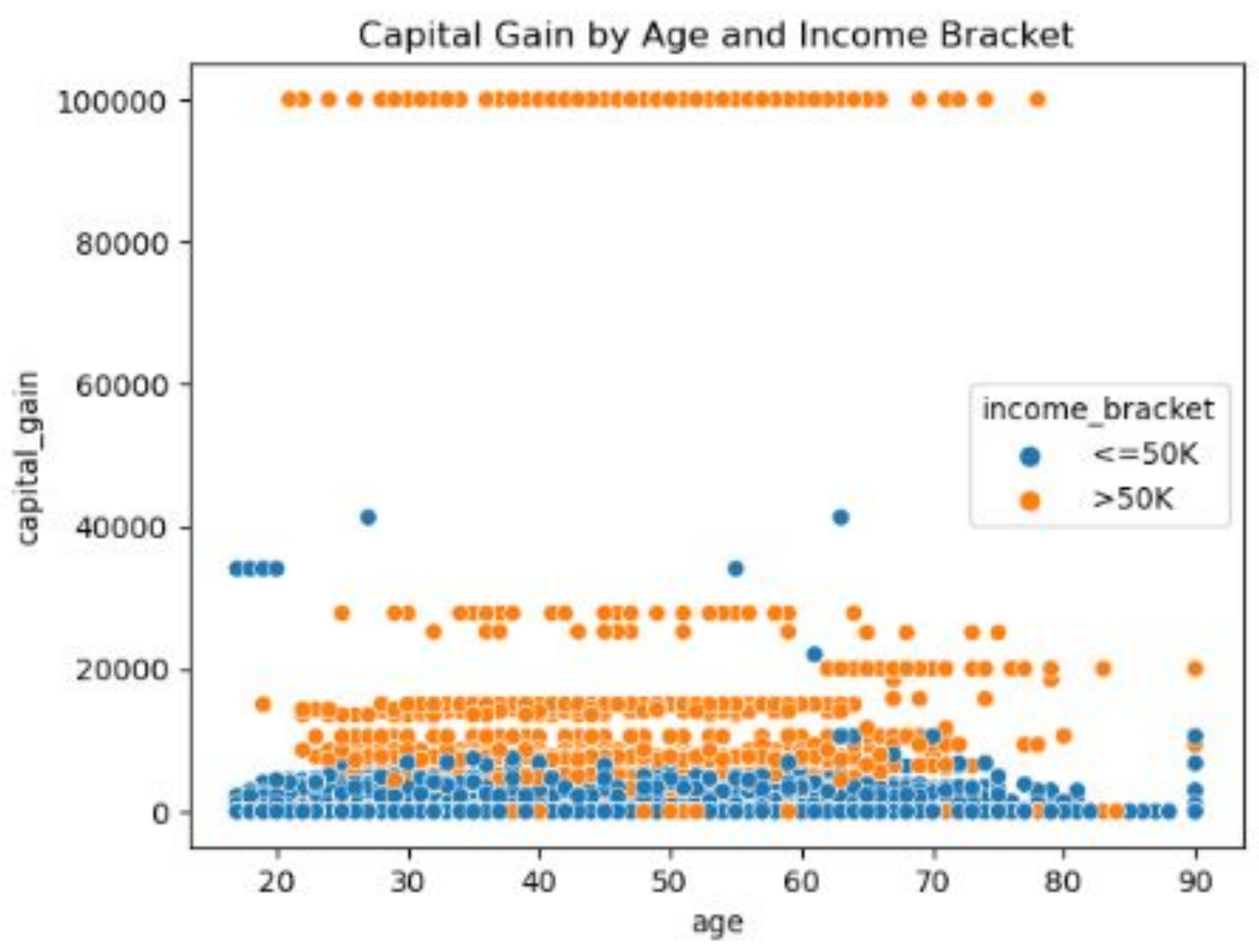
We can see that null error rate is approximately 75% . So if our model does not have accuracy greater than that it would not be considered a good model.



The interquartile range (IQR) also widens with increasing hours per week, suggesting a greater spread of income data points among those with higher hours per week. This could indicate that working longer opens doors to a wider range of income possibilities.



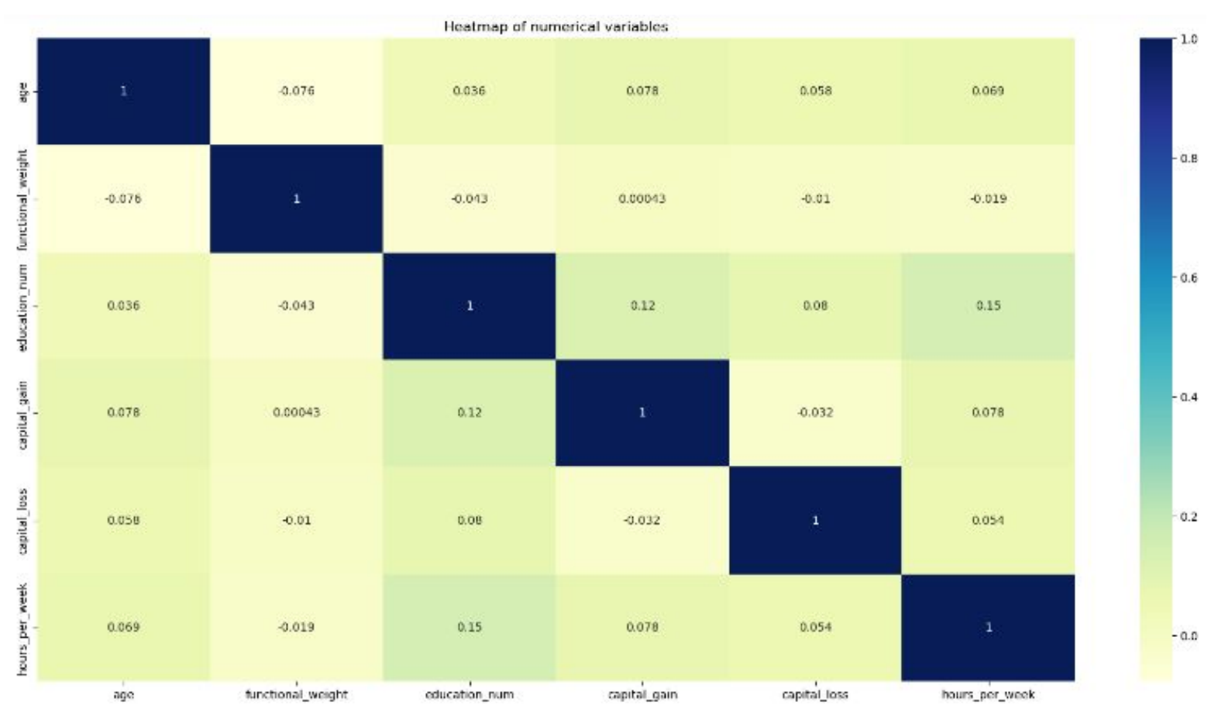
These observations suggest a potential association between workclass and income level. Occupations categorized as "Private" might generally offer lower wages or more limited work hours compared to government jobs ("Federal-Gov").



Younger individuals (on the left side of the axis) tend to have lower capital gains overall, regardless of income bracket. This could be due to factors like having less time for investment accumulation.
Also capital gains might increase with age (as we move to the right on the axis) for both income brackets. This could reflect accumulating investments over a longer career or lifespan.

Data Cleaning

In preprocessing the dataset, missing values in the 'native country' column were addressed by imputing them with mode values. Furthermore, to uphold data integrity, duplicate entries were identified and removed. Nominal categorical variables, including race and sex, underwent conversion into numeric variables through the application of the one-hot encoding technique. Meanwhile, ordinal categorical variables, such as education level, were subject to label encoding to preserve their inherent order. The target variable, income bracket, was also encoded via label encoding to facilitate subsequent analysis and modeling. Notably, during feature selection, variables displaying a high correlation with the target variable, namely 'marital_status_ Married-civ-spouse' and 'education_num', were identified for inclusion. Additionally, multicollinear features, such as 'workclass_Unknown' and 'occupation_Unknown', were recognized and addressed to mitigate redundancy in the dataset.



Our analysis revealed that features like 'marital_status_Married-civ-spouse' and 'education_num' exhibited high correlations with the income bracket. This suggests a potentially strong relationship between marital status (being married) and higher income, as well as between educational attainment (indicated by 'education_num') and income.

We also identified potential multicollinearity between 'workclass_Unknown' and 'occupation_Unknown'.To address this, we considered removing one of these features, potentially 'workclass_Unknown', as occupation might provide more specific information regarding income-influencing factors.

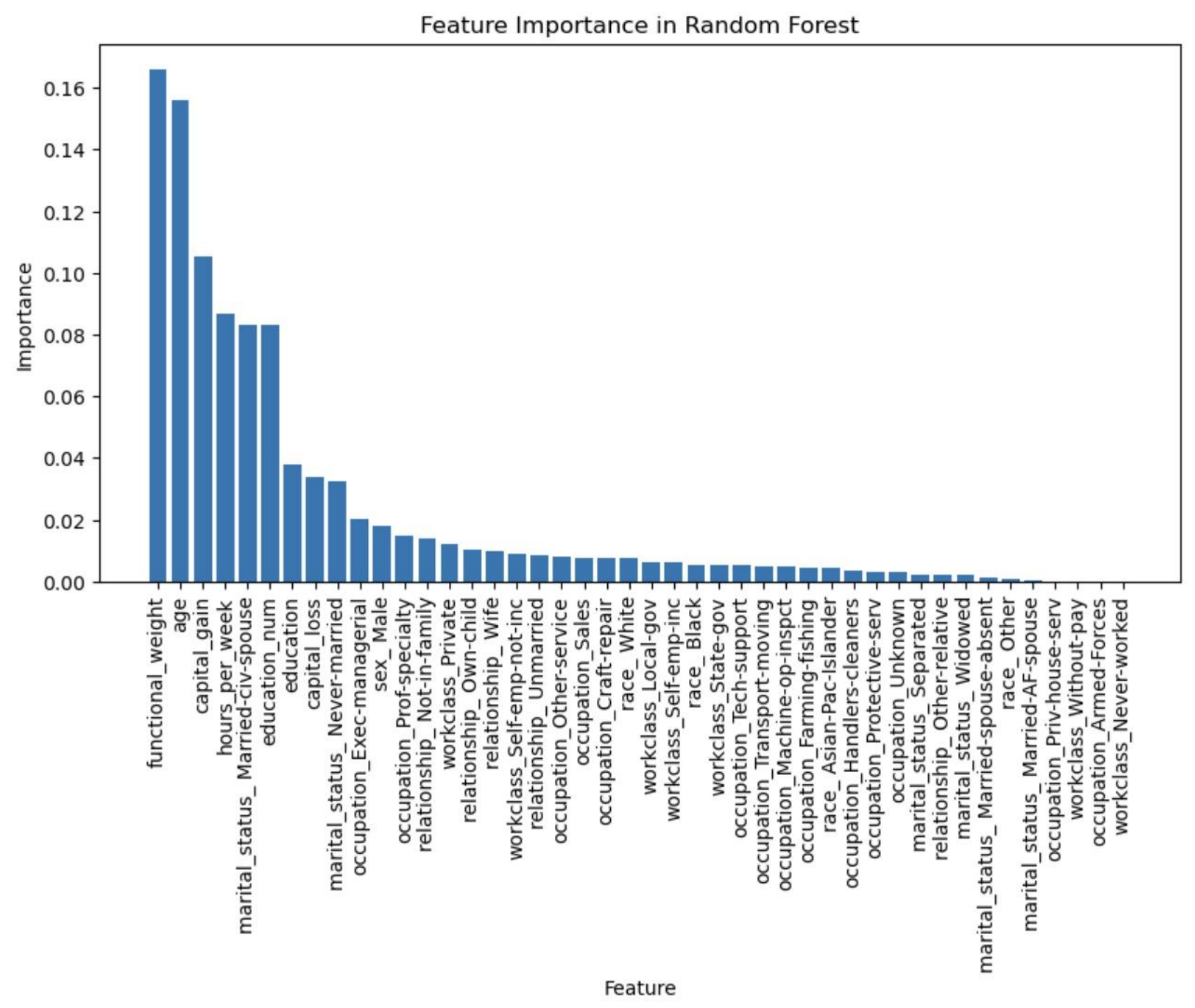
Modelling

Our study involved the development of four machine learning models: two logistic regression and two random forest models. To train and evaluate these models, we partitioned the dataset into training and testing subsets using a 70:30 ratio.

In our logistic regression modeling, the first model exclusively incorporated numeric variables, while the second model utilized both numeric and one-hot encoded categorical variables.

For our random forest models, the first one was trained solely on numeric features. In contrast, for the second random forest model, we conducted hyper-parameter tuning to optimize performance and identified crucial features through feature importance analysis. Subsequently, we trained the model using these key features.

Finally, we assessed the performance of logistic regression Model I and random forest Model II on the testing data.



Model Comparison and Analysis

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression Model 1	0.8	0.76	0.62	0.63
Logistic Regression Model 2	0.8	0.76	0.62	0.64
Random Forest Model 1	0.84	0.84	0.72	0.76
Random Forest Model 2	0.87	0.85	0.78	0.81
Logistic Regression Model 1 Testing	0.8	0.76	0.61	0.63
Random Forest Model 2 Testing	0.86	0.83	0.77	0.79

Conclusion

Random Forest Model 2 exhibited outstanding performance (accuracy = 0.87), effectively discerning between income brackets. These findings resonate with existing research on income determination factors. While the study highlights the significance of human capital and financial factors, future research could explore potential interactions between these features and consider income granularity beyond $50,000 thresholds (e.g., $25,000 increments). This might involve employing a regression-based approach, such as linear regression, to model income as a continuous variable. Overall, this study demonstrates that a combination of education, work experience, financial factors, work hours, and marital status significantly affects adult income in the US.One limitation to consider is that the study relied on self-reported data from a national survey, which might be susceptible to reporting bias. However, limitations like data source bias warrant further investigation using the proposed future research avenues.

References

BigQuery. (https://cloud.google.com/bigquery/docs)

Verma, Nitika. "Machine Learning application — Census Income Prediction." Medium, Alien Status, 23 Aug. 2019, https://medium.com/analytics-vidhya/interpreting-machine-learning-models-in-census-income-data-set-56c9dc7e0f27

Matz, Brian D., et al. "Prediction of Individual Level Income: A Machine Learning Approach." 2019 IEEE International Conference on Big Data (Big Data), Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 2547-2556. https://digitalcommons.bryant.edu/honors_economics/39/

Heinze, Christina, et al. "Fairness and Bias in AI-Based Income Prediction." arXiv preprint arXiv:2105.00667, 2020. https://arxiv.org/pdf/2212.09868