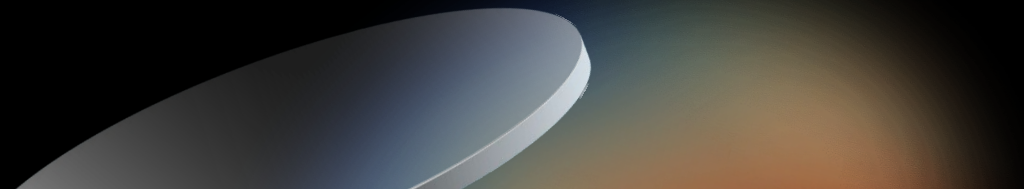
1994 Census Income

STAT362 – R for Data Science Group 18
Zain Parihar, Shrinidhi Thatahngudi
Sampath Krishnan,
Bartek Kowalski, Sefanit Zeray

Table of Contents



Introduction

- We wanted to see the relationship between a specific class of a person and their income so as to identify classifications of people that are more likely to enjoy higher income. This would help us identify socioeconomic trends, inequality and real world implications.
- Analysing better ways to address inequality, income disparities, quality of life and improve economic opportunities for different groups in society.
- Many studies have been done by universities, governments and and human rights activists.
- Ex. How Much Does Racial Bias Affect Mortgage Lending? Evidence from Human and Algorithmic Credit Decisions

Dataset

- “Census Income” Dataset, taken from the UCI ML Repository
- Data is scraped from the 1994 USA Census Database
- The dataset contains 14 features and 1 target variable
- Binary Classification task
- Researchers predict whether income exceeds $50K/yr

Data types of variables

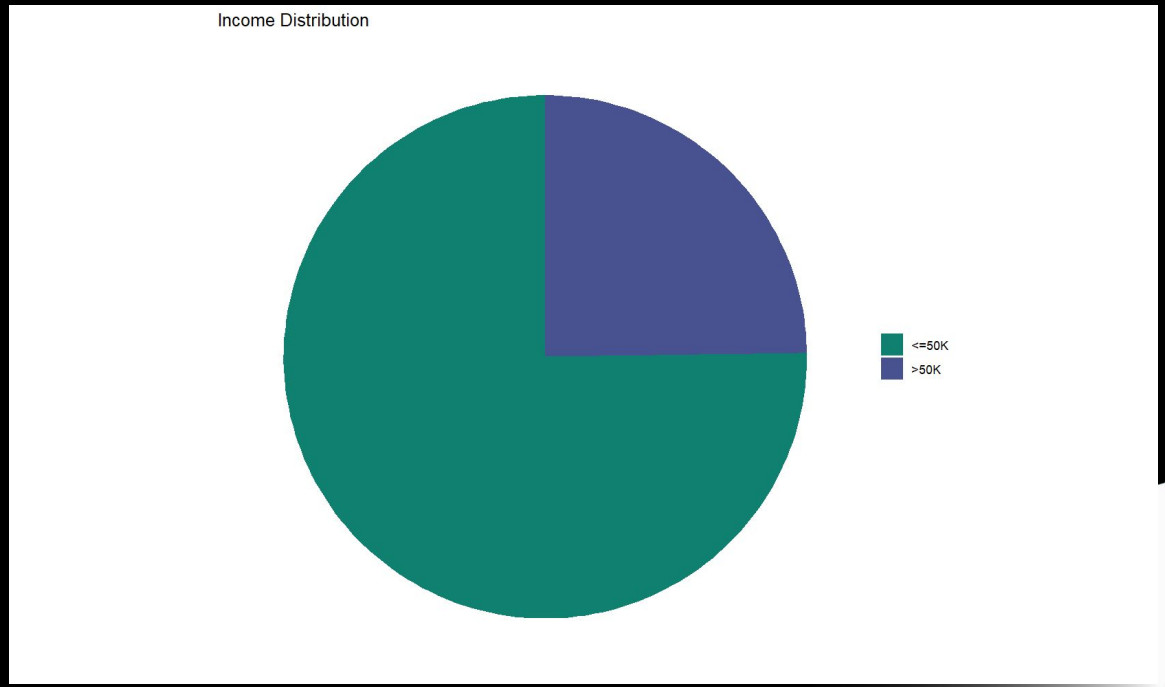
Income	Binary Categorical
Age	Continuous Numeric
Workclass	Categorical
Education	Categorical
Education-num	Continuous Numeric
Marital-Status	Categorical
Relationship	Categorical
fnlwgt	Continuous Numeric

Occupation	Categorical
Native-Countries	Categorical
Capital-Gain	Numeric
Capital-Loss	Numeric
Hours-per-week	Numeric
Sex	Binary Categorical
Race	Categorical

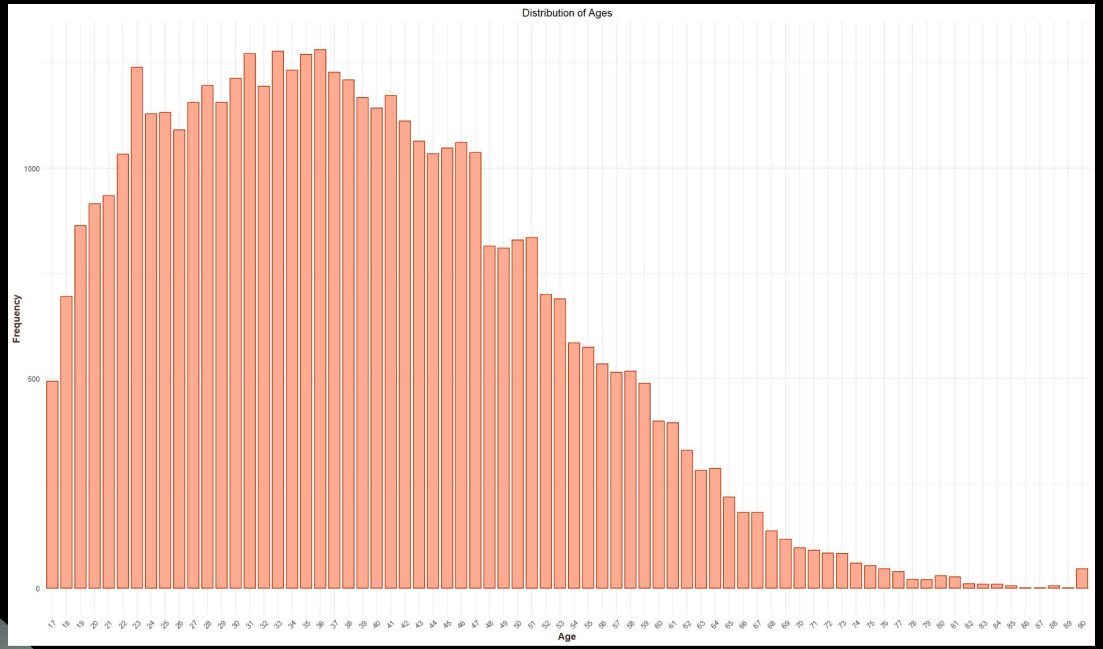
Preliminary Exploration

Income

- Target Variable
- Binary Category
- Approx 3:1



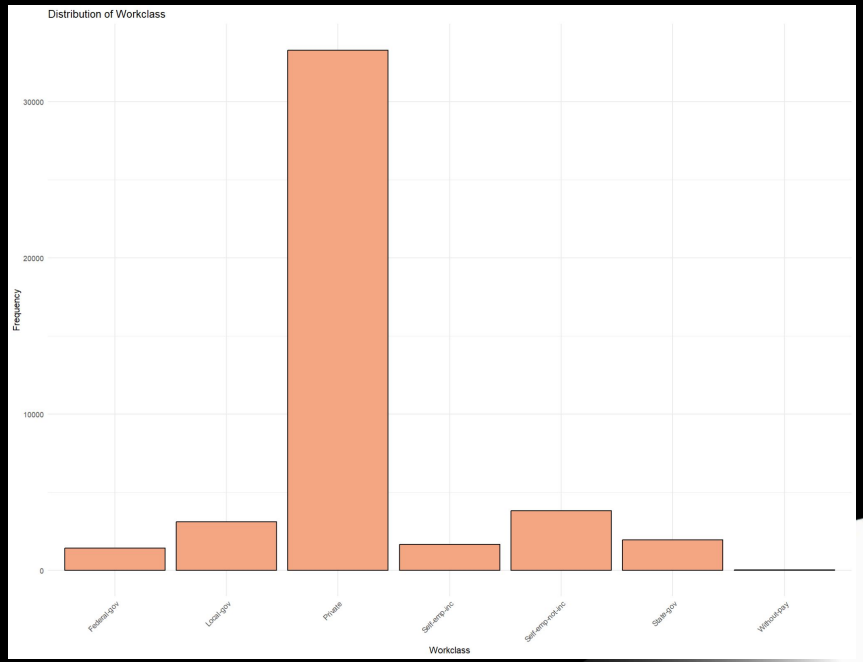
Age



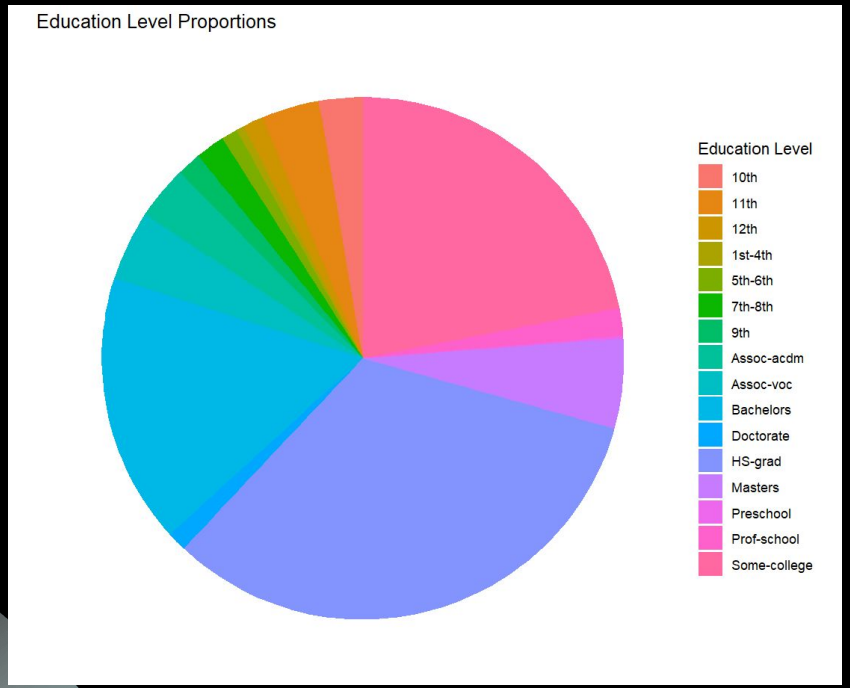
- Feature
- Continuous Variable
- Ranges from 17 - 90
- Right-Skewed
- Mean > Median > Mode

Workclass

- The type of employment
- Nominal Categorical Variable
- 7 Categories
- Government Jobs
  - Federal
  - State
  - Local
- Private
- Self-employed
  - “not-inc” implies that an individual’s business is not incorporated
- Employed without pay



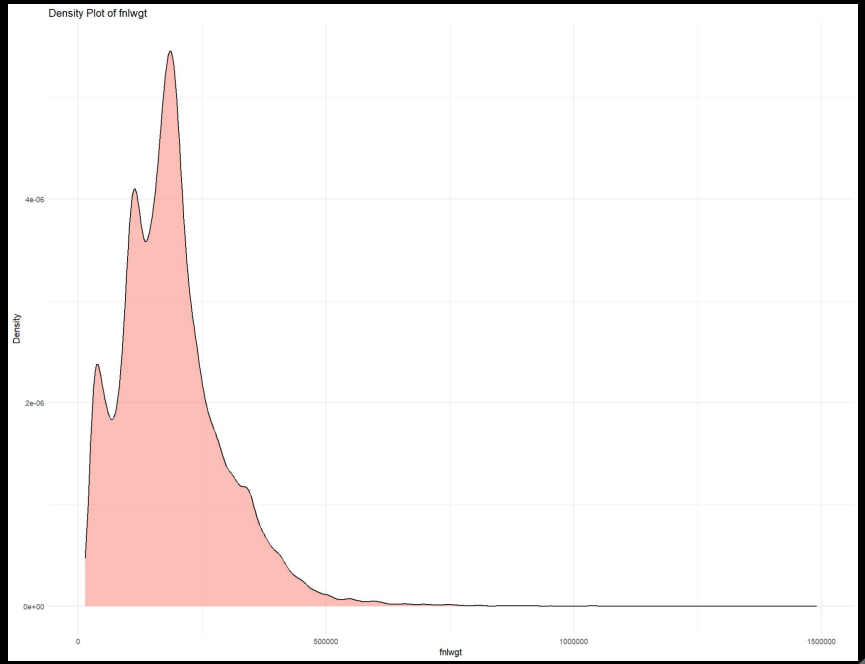
Education



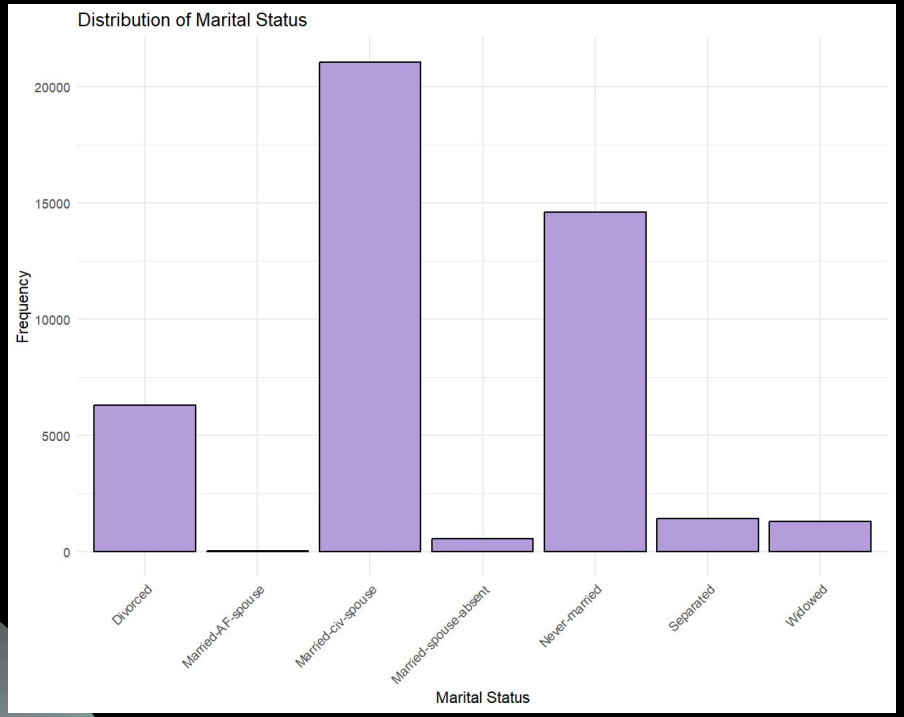
- Education - Nominal Categorical Variable
  - 16 Categories
- Education-num - Continuous variable, higher values correspond to better education
- Divided into grade levels or degree achieved
- Assoc-acdm - "Associate of Arts" or "Associate of Science"
- Assoc-acdm - "Associate of Vocational" or "Associate of Occupational"

fnlwgt

- Final Weight
- Continuous Variable
- The number of people that the census believes an entry represents
- Right-Skewed
- Mean > Median > Mode



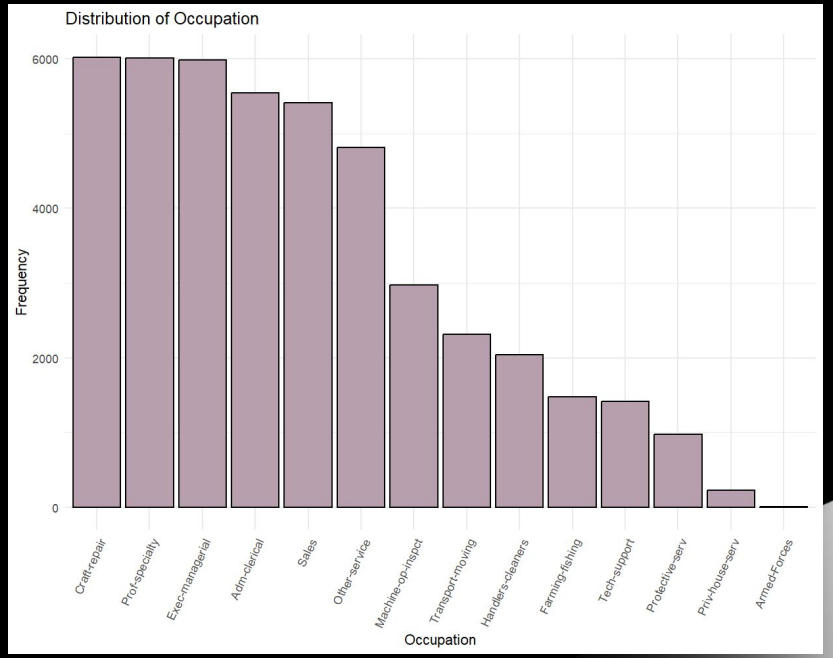
Marital-Status



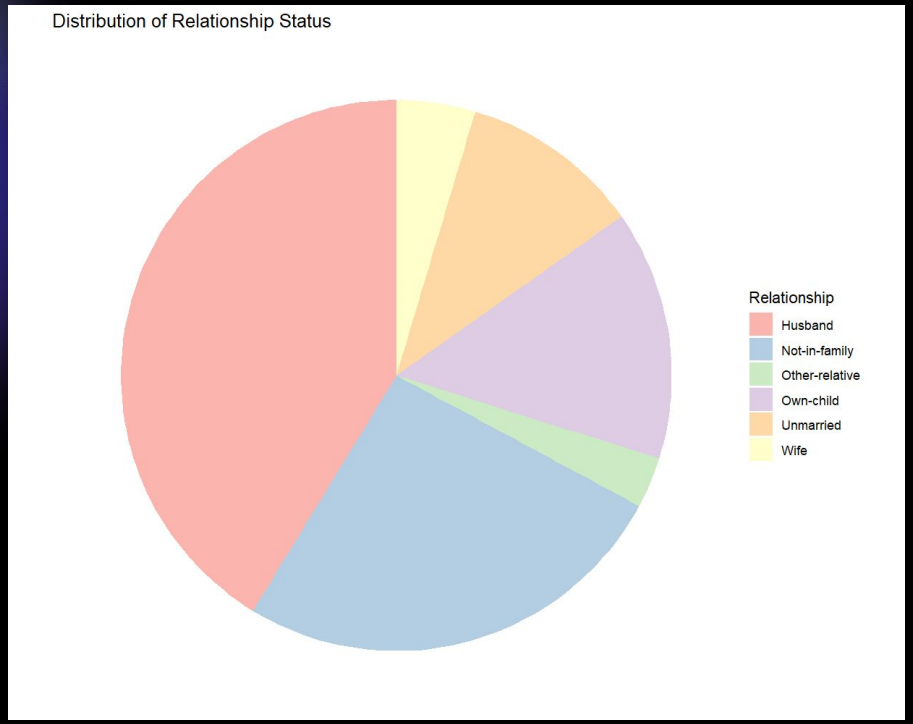
- Nominal Categorical Variable
- 7 Categories
- 3 Types of married
  - Civil
  - Spouse is absent
  - Armed Forces spouse

Occupation

- The type of Job
- Nominal Categorical Variable
- 14 Categories



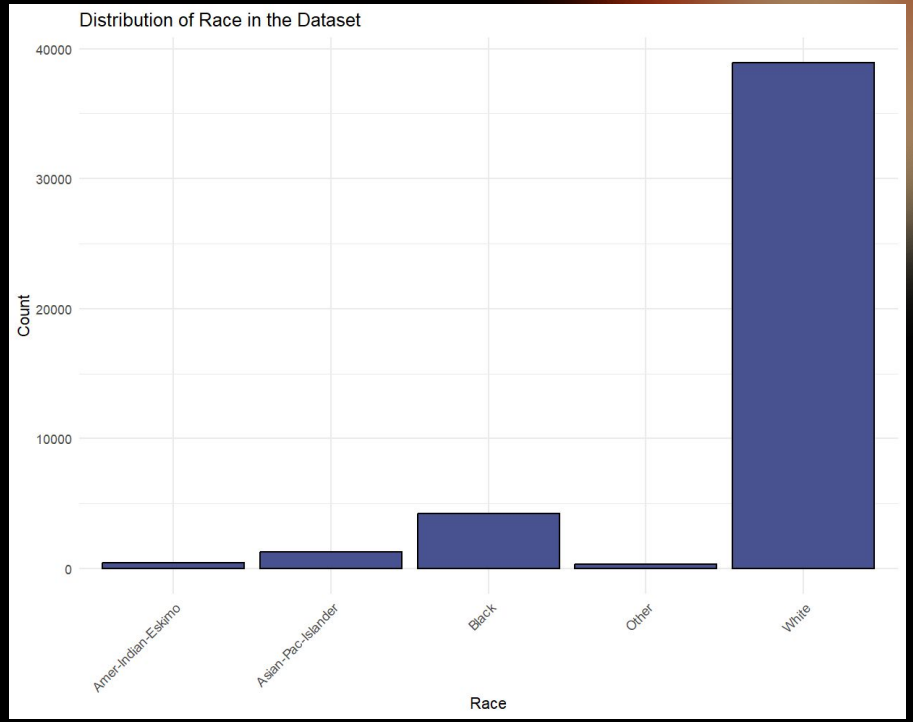
Relationship (Status)



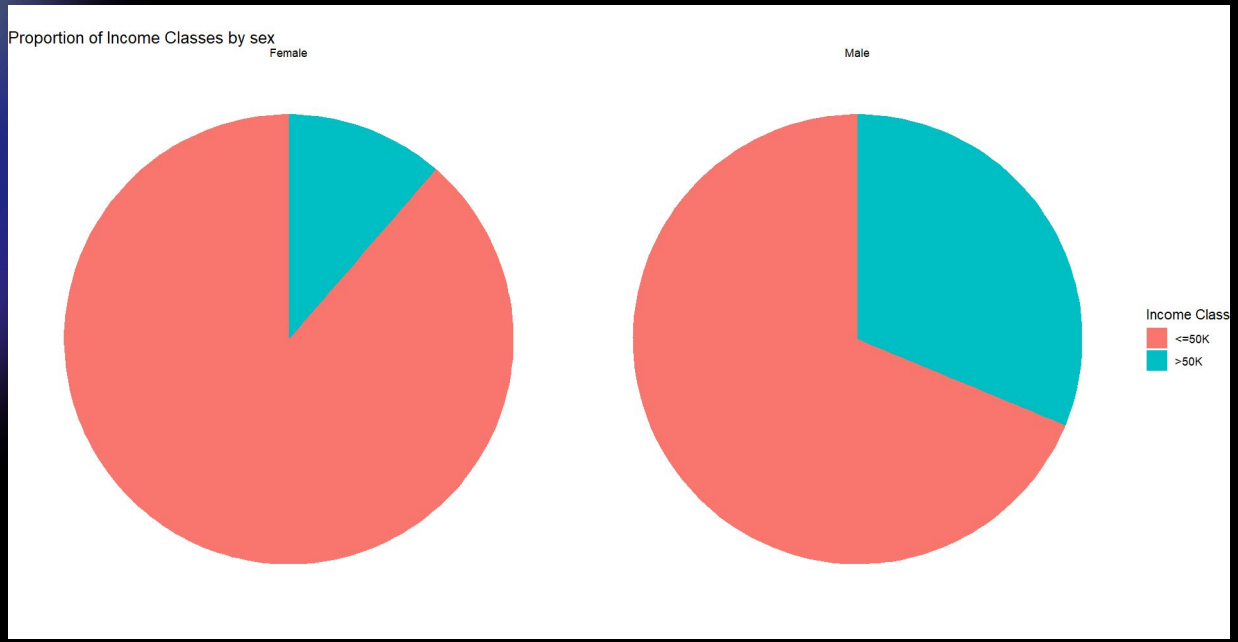
- Nominal Categorical Variable
- 6 Categories
- Not-in-family - individuals who are not related to the family reference person
- Own-child - child of the family reference person, unmarried and under a certain age

Race

- Nominal Categorical Variable
- 5 Categories



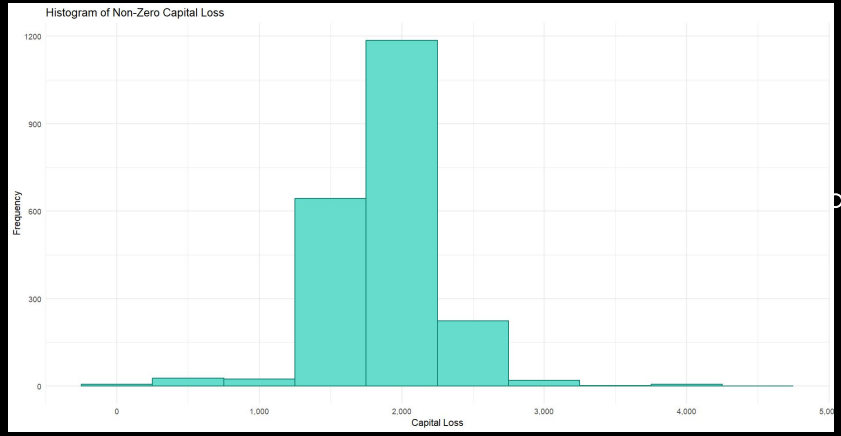
Sex

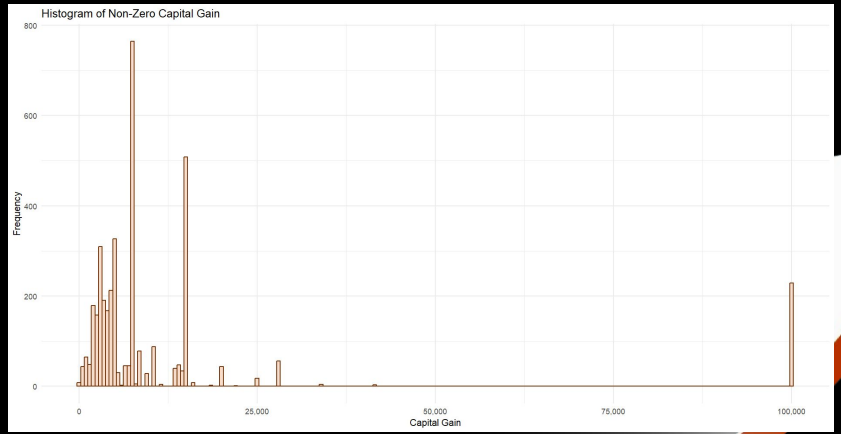


- Binary Categorical Variable
- 2 Categories
  - Male
  - Female
- Divided by Income Class

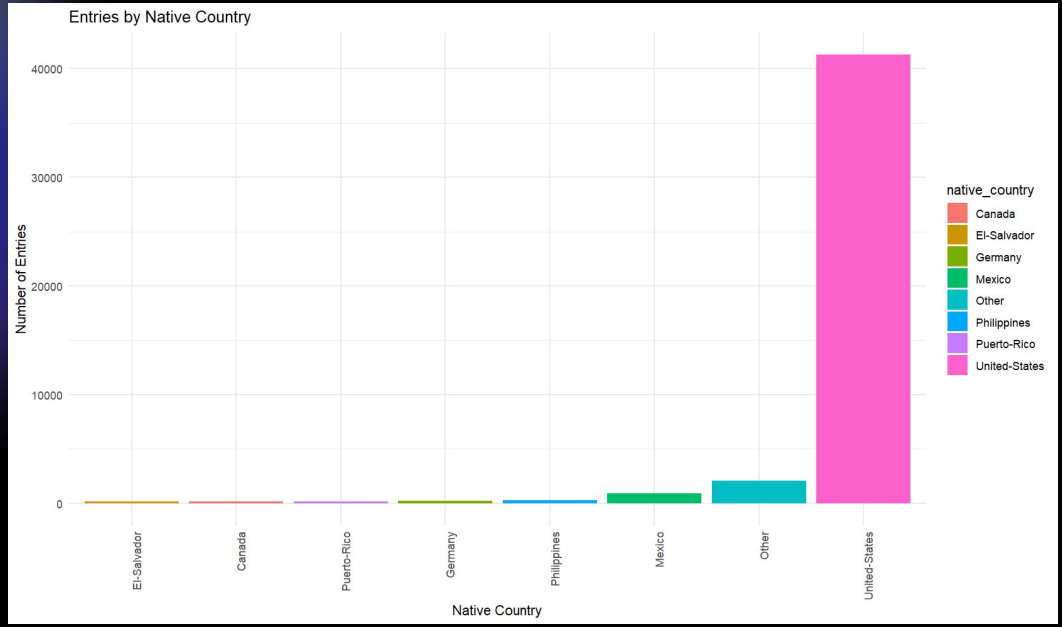
Capital Gains and Losses

- Continuous Variables
- Capital Gains - Profits from investments or property
- Capital Losses - Deficits from investments or property
- Unrelated to employment and salary



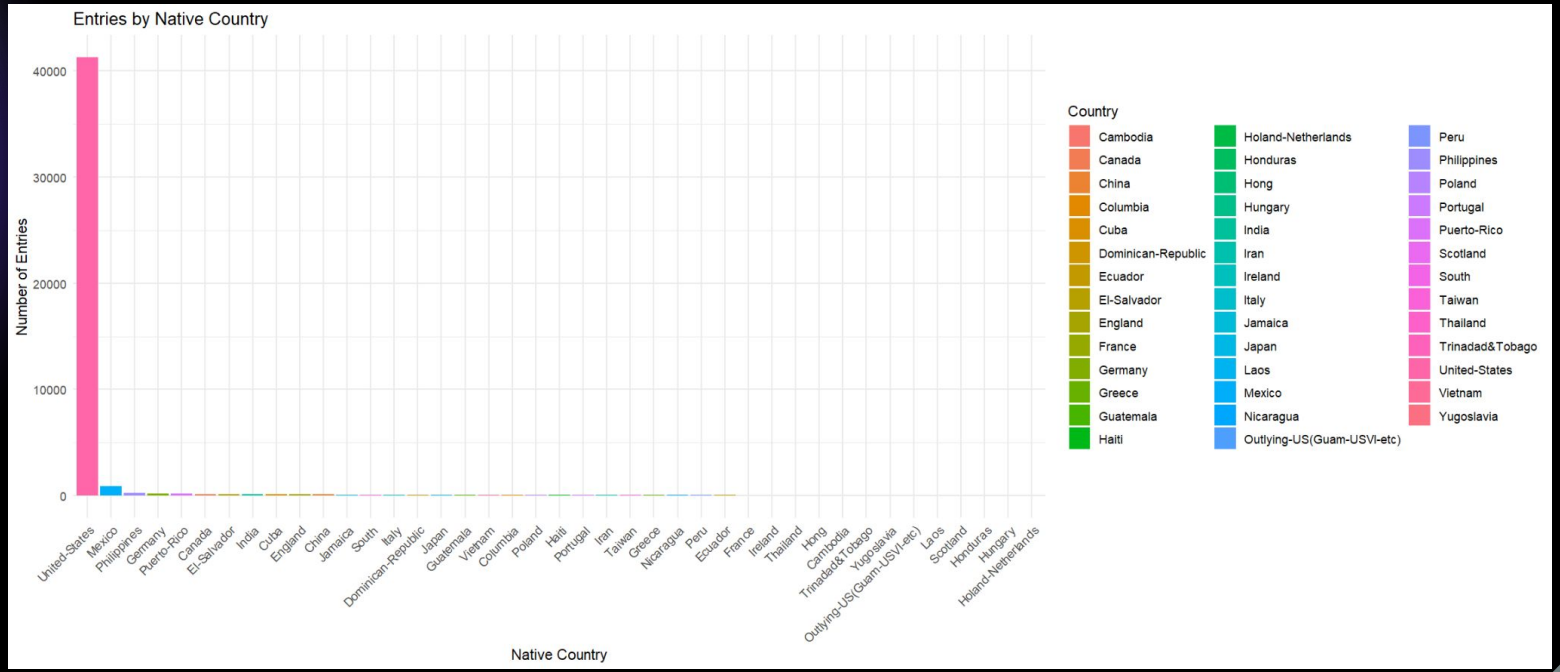


Native-Countries



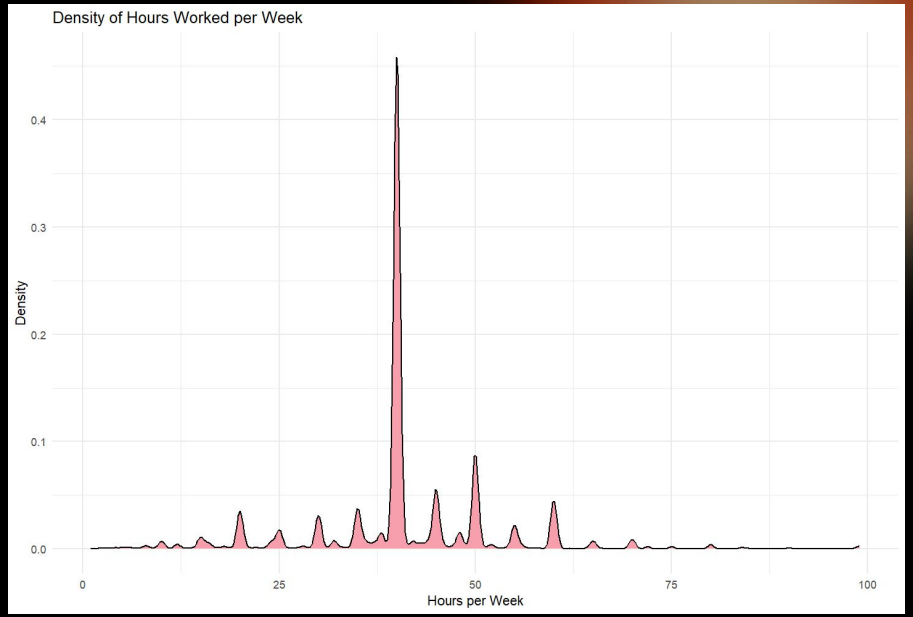
- An individual's country of origin
- Nominal Categorical Variable
- 41 Categories
- This graph has the top 7 countries

All Countries

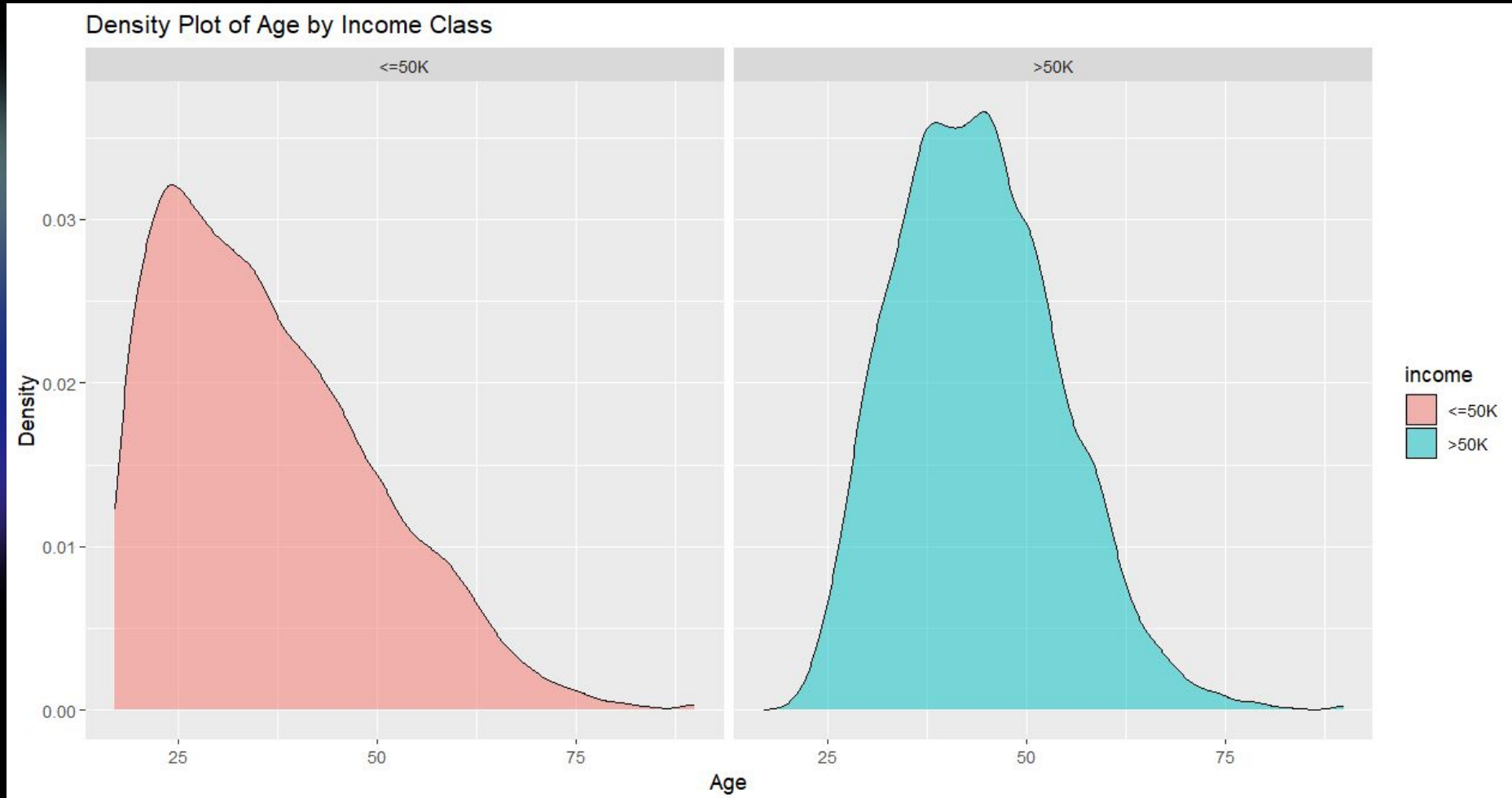


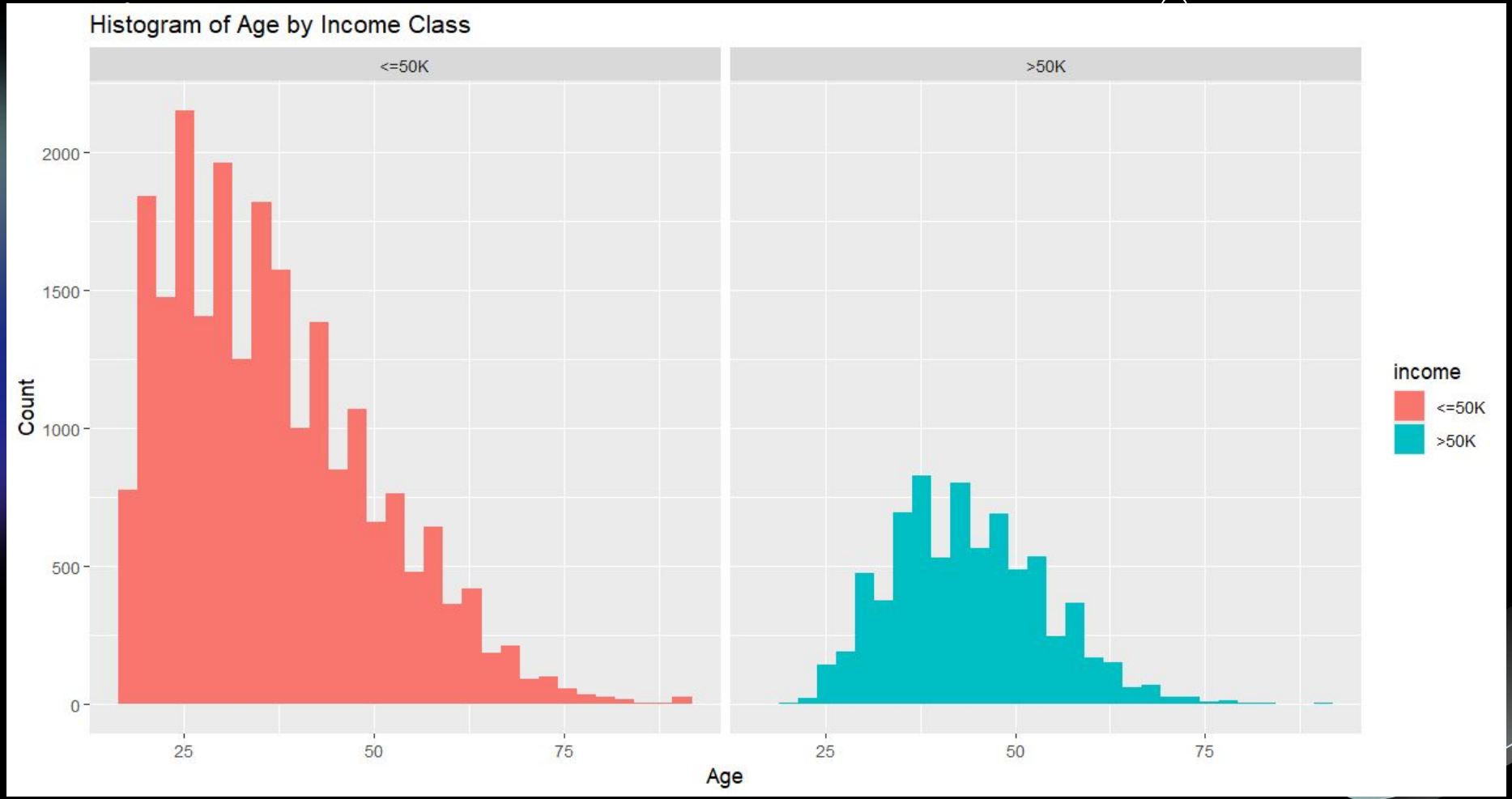
Hours-per-week

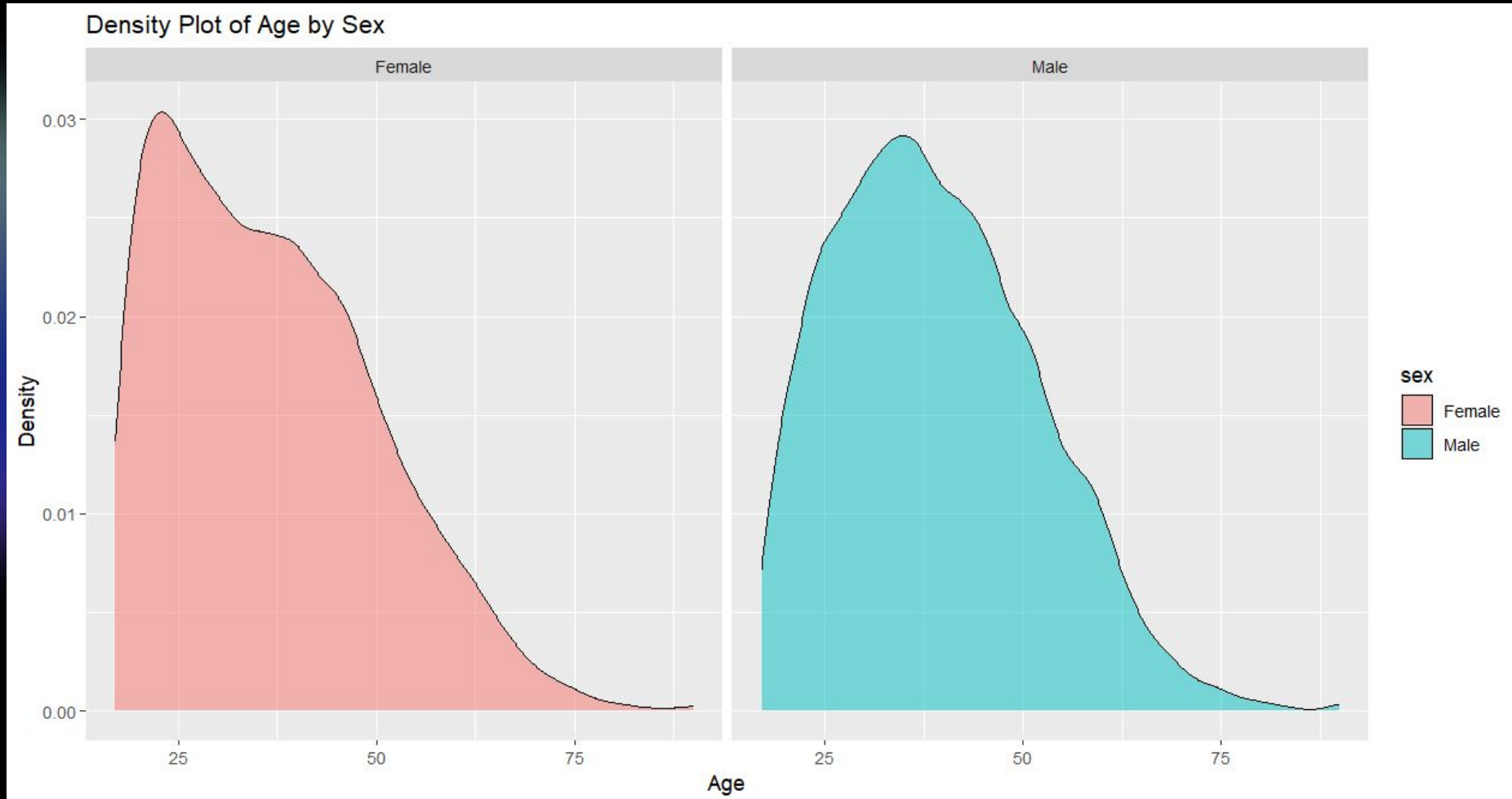
- Continuous Variable
- Number of hours worked every week
- For context, there are 168 hours in 1 week

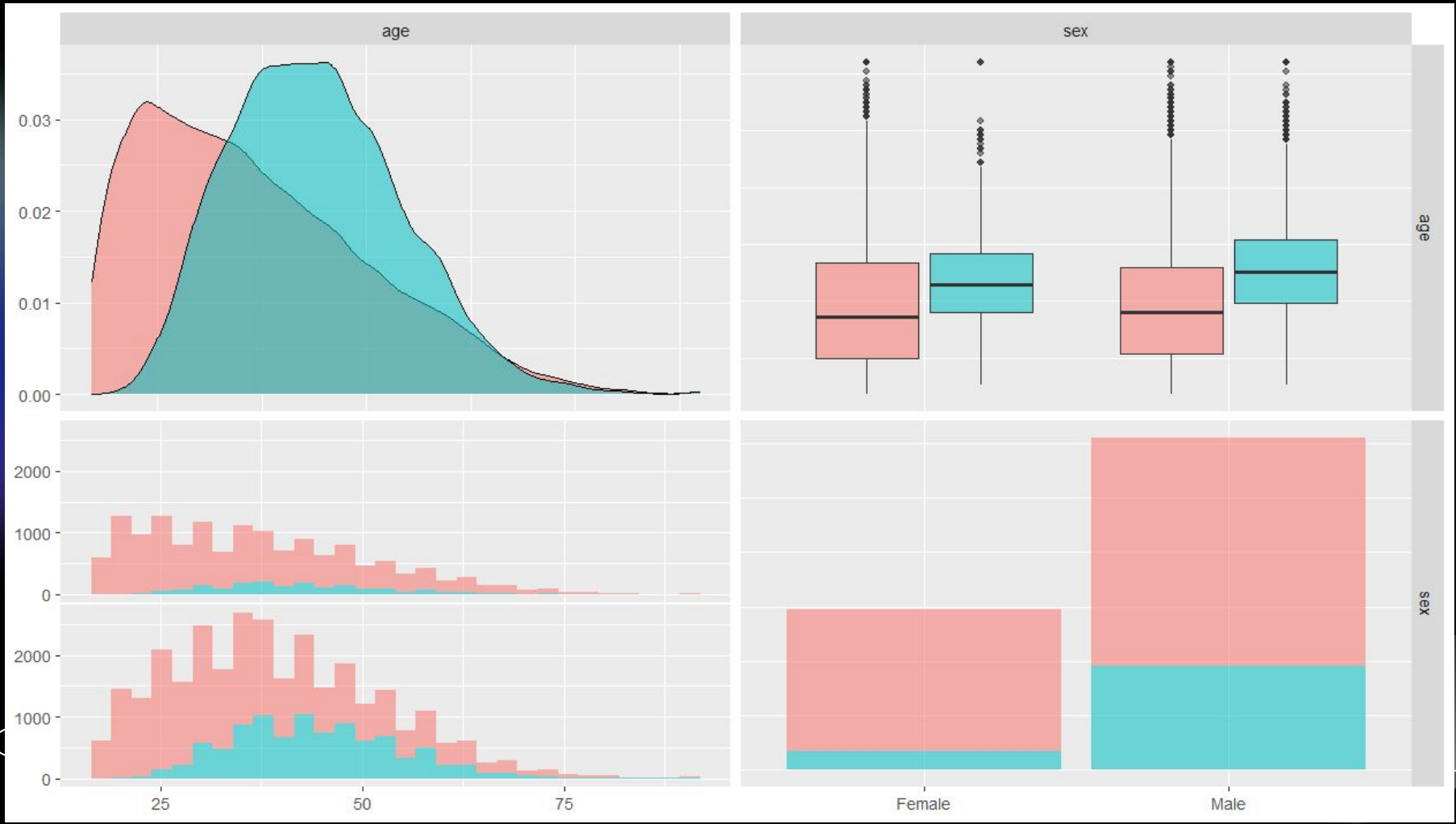


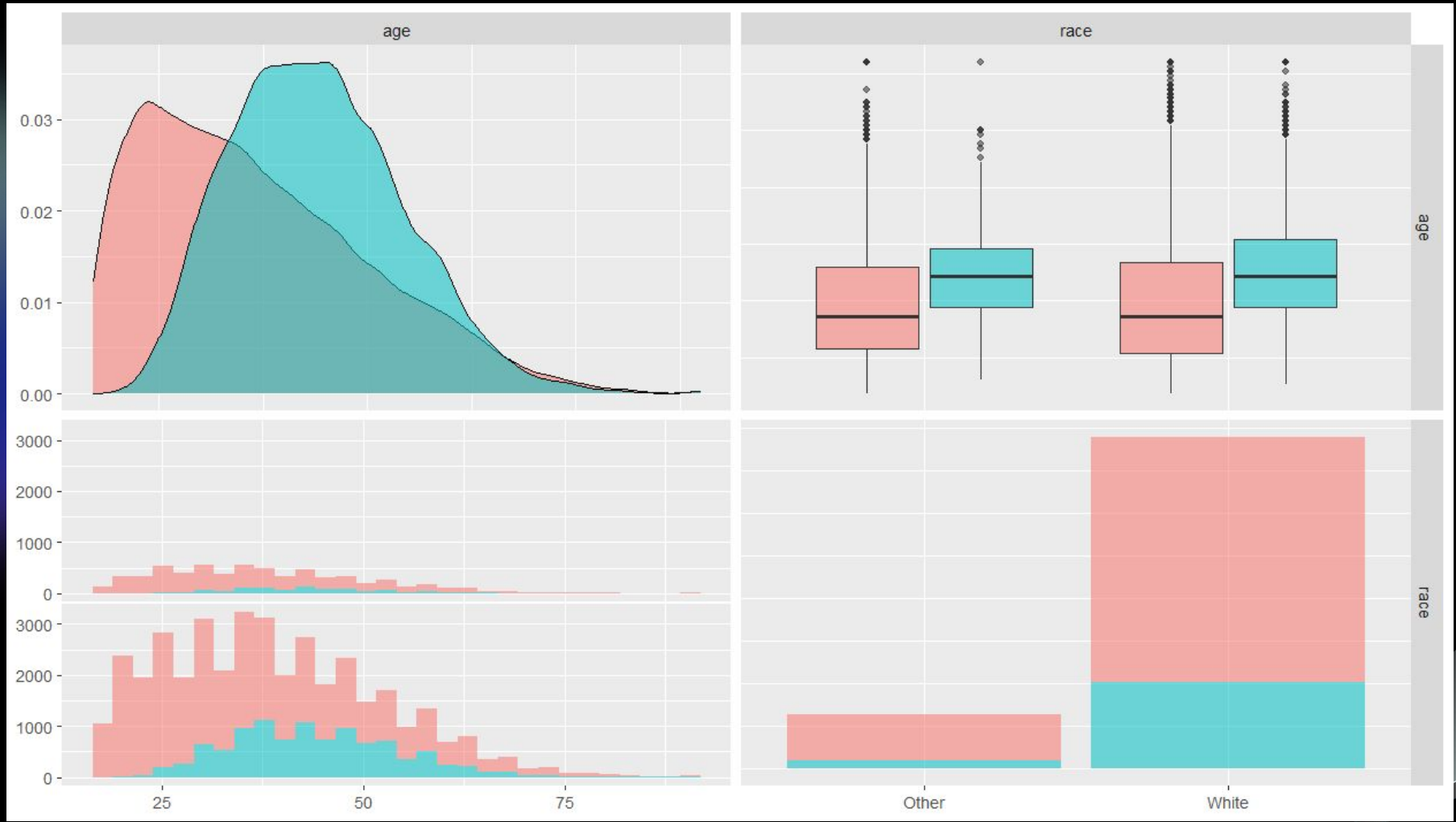
Preliminary Relationships


Density Plot of Age by Income Class
<=50K
>50K
Density
0.03
0.02
0.01
0.00
25
50
75
Age
income
<=50K
>50K


Histogram of Age by Income Class
<=50K
>50K
Count
2000
1500
1000
500
0
25
50
75
Age
income
<=50K
>50K


Density Plot of Age by Sex
Female
Male
0.03
0.02
0.01
0.00
Density
25
50
75
Age
sex
Female
Male


age
sex
0.03
0.02
0.01
0.00
2000
1000
0
2000
1000
0
25
50
75
Female
Male
age
sex


age
race
0.03
0.02
0.01
0.00
3000
2000
1000
0
25
50
75
Other
White
age
race

Methods

We used three main models to observe patterns in the data and classify accordingly:

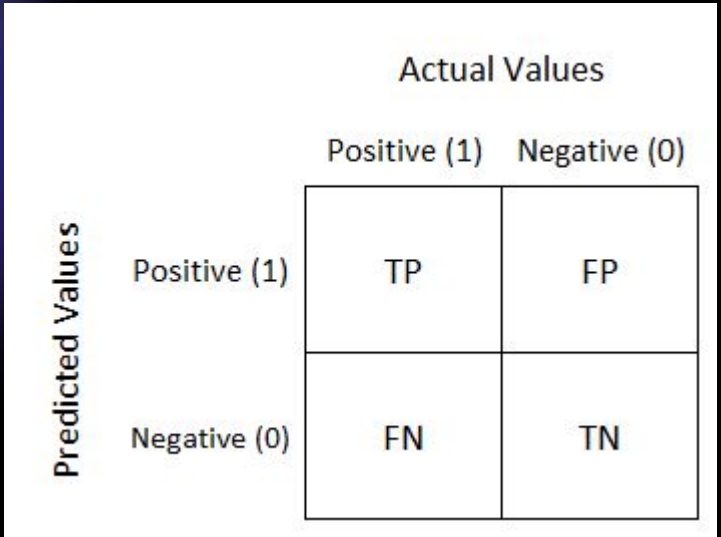
- KNN (K-nearest neighbors)
    - For KNN, we used K(7) nearest neighbors and used PCA (principal component analysis) to reduce the dimensionality of the data
- Binary Classification (Decision Trees)
    - We classified the tree according to the income (> 50k or <= 50k) as mentioned earlier
- Random Forest

Methods (contd.)

Confusion matrix was calculated for all the models that we trained where:

- **True Positive (TP)**: These are cases where the model predicted the individual's income to be above the threshold, and the actual income is indeed above the threshold.
- **True Negative (TN)**: These are cases where the model predicted the individual's income to be below the threshold, and the actual income is indeed below the threshold.
- **False Positive (FP)**: These are cases where the model predicted the individual's income to be above the threshold, but the actual income is below the threshold (Type I error).
- **False Negative (FN)**: These are cases where the model predicted the individual's income to be below the threshold, but the actual income is above the threshold (Type II error).

Confusion matrix



$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Cleaning the data

- We factorized all of the text data so it will be useable with our machine learning methods
- Simplified 'native_country' to a binary variable ('United-States' vs. 'Other') to reduce complexity and focus on significant predictive information.
- Made the text data consistent and easier to work with:
    - Trimmed whitespace from all character variables to ensure data consistency and accuracy.
    - Removed trailing periods from the 'income' variable for consistency across training and test datasets.

KNN

- For our KNN model, we decided to use a K value of 7.
- Our accuracy was 80.04%, while our balanced accuracy was 68.87%. This points to the fact that we were able to discern classes with high amounts of data accurately, while struggling on datasets with lower amounts of data. This is further followed up by our Specificity (True Negative Rate for >50K) being 46.92%.
- The model shows overall good accuracy, however struggles with the minority class.

PCA with KNN

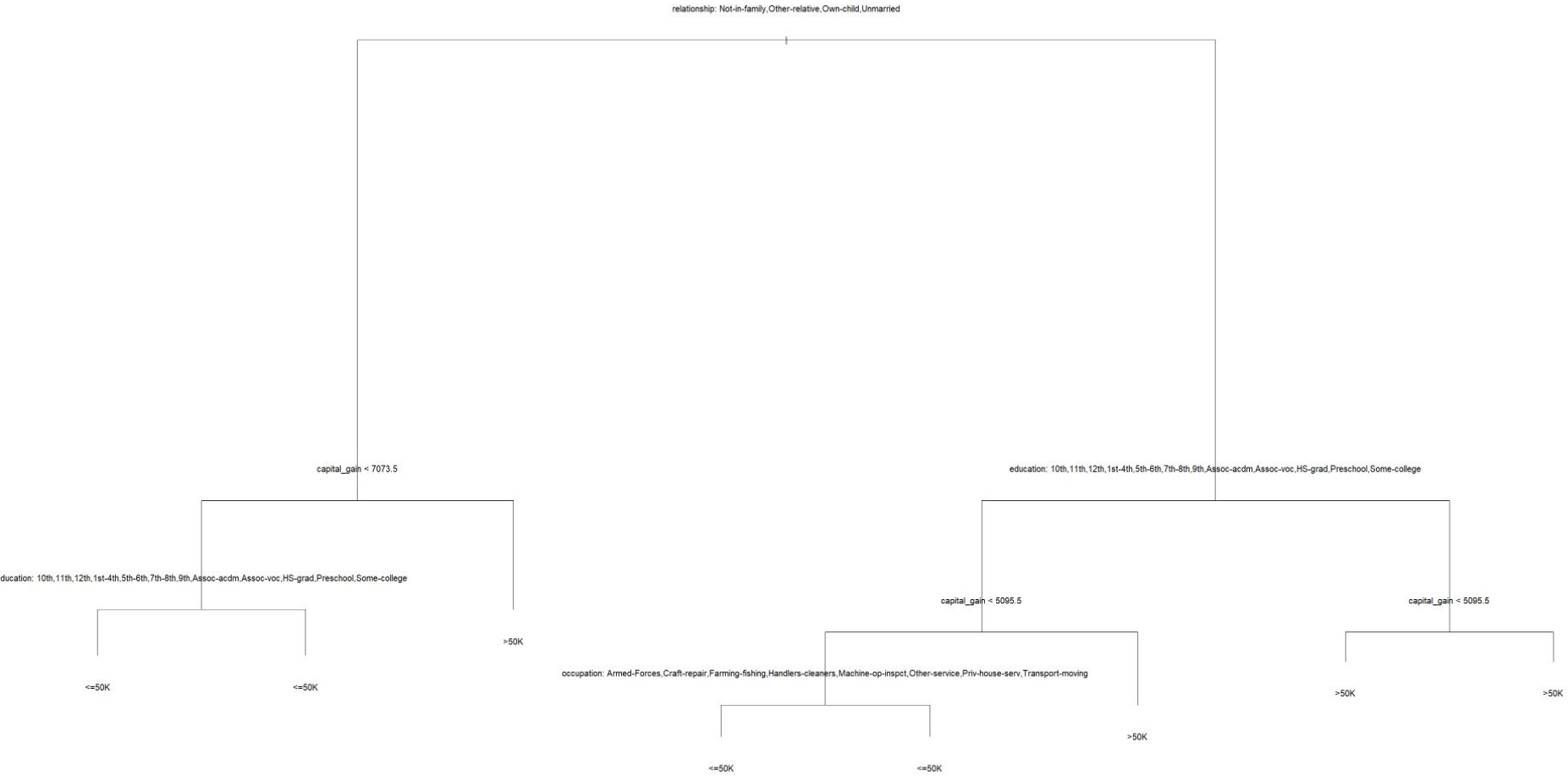
- Implemented PCA before running KNN to reduce dimensionality.
- The integration of PCA with KNN resulted in an accuracy of 80.01% and a balanced accuracy of 68.84%, maintaining model performance while simplifying data complexity.
- Did not really improve/degrade model performance when compared to KNN without PCA.

- Cross-validation was also performed on the K-nearest-neighbors algorithm
  - It is useful for predicting the performance of a model prior to testing
  - It helps identify whether outliers are present
  - It can help reduce overfitting
- The results show that there are few outliers in the dataset, as the expected performance is the same as the actual results

KNN Cross Validation

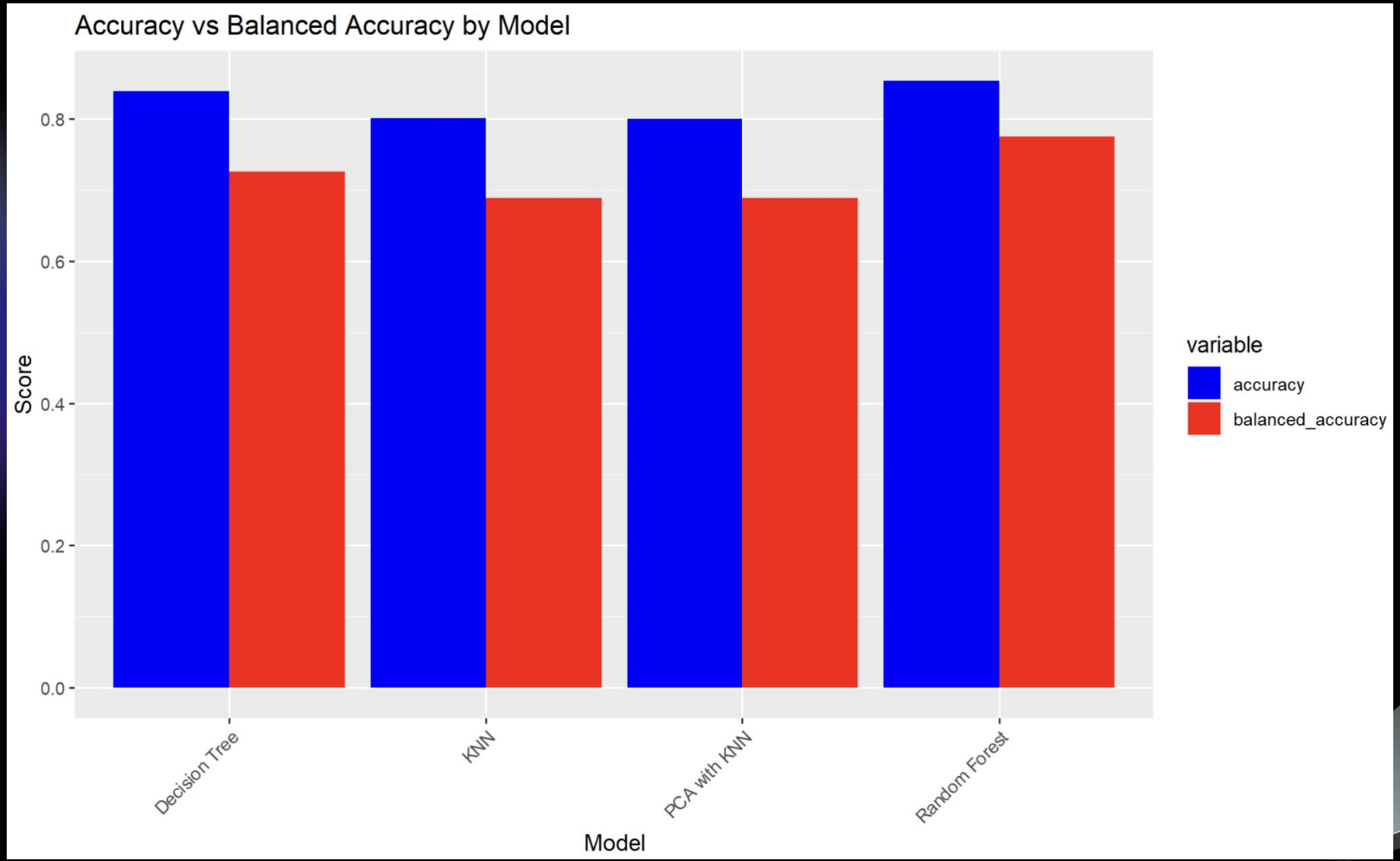
Binary Classification (decision tree)

- Accuracy of 83.9%, with a balanced accuracy of 72.59%, pointing to the fact that the decision tree still struggles with classifying the minority class. However, this result improves upon the previous attempt using KNN.
- Sensitivity: 94.82%, Specificity, 50.35%. This furthermore points to the fact that the model struggles to classify the True Negative Rate for >50K.
- Better than both KNN, KNN with PCA.


relationship: Not-in-family,Other-relative,Own-child,Unmarried
capital_gain < 7073.5
education: 10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Preschool,Some-college
ducation: 10th,11th,12th,1st-4th,5th-6th,7th-8th,9th,Assoc-acdm,Assoc-voc,HS-grad,Preschool,Some-college
>50K
<=50K
<=50K
capital_gain < 5095.5
capital_gain < 5095.5
occupation: Armed-Forces,Craft-repair,Farming-fishing,Handlers-cleaners,Machine-op-inspct,Other-service,Priv-house-serv,Transport-moving
>50K
>50K
<=50K
<=50K

Random Forest

- Utilized 1000 trees and set the number of variables considered at each split to one-third of the total.
- The Random Forest model achieved an accuracy of 85.29% and a balanced accuracy of 77.46%, indicating superior handling of both classes and a marked improvement over the decision tree model, especially in addressing the minority class.
- Sensitivity of 92.86% and specificity of 62.05% underscore the Random Forest model's proficiency in identifying lower income earners while displaying ongoing challenges in accurately classifying individuals earning more than 50K.


Accuracy vs Balanced Accuracy by Model
Score
0.0
0.2
0.4
0.6
0.8
Decision Tree
KNN
PCA with KNN
Random Forest
Model
variable
accuracy
balanced_accuracy

Discussion

- We focused solely on Decision Trees, KNN, and Random Forest models, potentially missing the advantages offered by other algorithms such as Support Vector Machines or neural networks.
- Throughout all of our experiments, classifying the minority class was still difficult. Employing techniques to potential subsidize this would likely improve the performance of our models.
- To improve on our results, we can apply techniques like SMOTE to try and remedy class imbalance, as well as trying different methods such as XGBoost.

Results

- In our testing using KNN, Binary Classification, and Random Forest, we found that Random Forest had both the best Accuracy and Balanced Accuracy.
- This seems to be because the model was better at predicting the minority class compared to the other two methods.

Why is our analysis useful

Our analysis can be used to measure the changes in the structure of society either positive or negative. As well as used to measure the impact of programs that are aimed at decreasing inequality.

We also identify key factors which predict income inequality, which can be used to guide decisions.