Izaak King

Professor Eleish

Data Science

December 5, 2025

“Assignment 6”

For this Assignment I will conduct predictive and prescriptive data analytics by developing and validating predictive models for the provided census and income dataset. Particularly, we will look at quantitative variables for Age, Education Number (Education Level), and Hours Worked Per Week, as well as categorical variables for Sex, Race, and Workclass. Ultimately we will attempt to predict the categorical variable for income, which is given as either > 50k or <= 50k. To predict these models we will use Random Forest, kNN and Decision Tree models. The Census and Income data has also been divided randomly into a data file and a test file. The regular data will be used for analysis and training, and the test data will be for testing and validating the models.

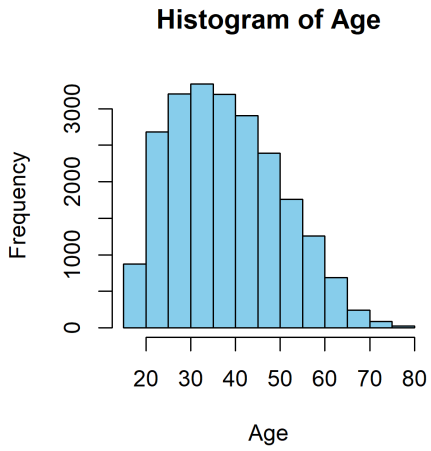
1. Exploratory Data Analysis

   I begin the analysis with basic Exploratory Data Analysis, looking at the distributions and summaries of the mentioned statistics. I first performed data cleaning, removing samples missing our desired categories. I also used the IQR to remove outliers in our dataset. Histograms, as well as box and density plots will be used for the quantitative variables, and bar plots and proportion bar plots will be used for the categorical variables.

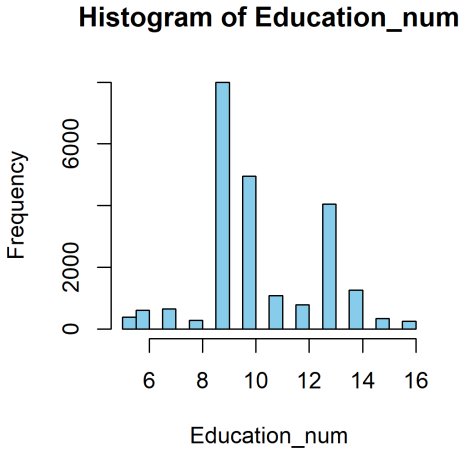
I. Summaries of Quantitative Variables

```
      age         education_num    hours_per_week
 Min.   :17.00   Min.   : 1.00   Min.   : 1.00
 1st Qu.:28.00   1st Qu.: 9.00   1st Qu.:40.00
 Median :37.00   Median :10.00   Median :40.00
 Mean   :38.58   Mean   :10.08   Mean   :40.44
 3rd Qu.:48.00   3rd Qu.:12.00   3rd Qu.:45.00
 Max.   :90.00   Max.   :16.00   Max.   :99.00
```

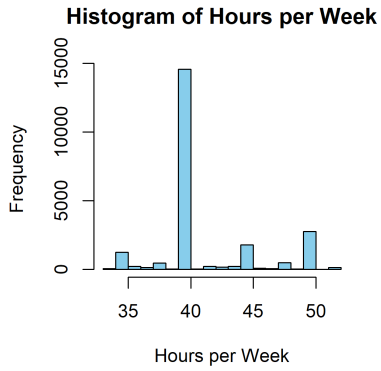
II. Histogram of Age



III. Histogram of Education Number

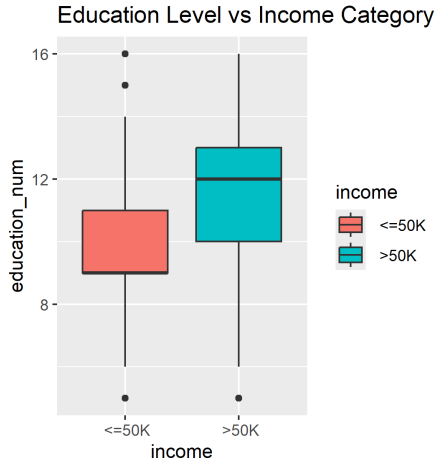


IV. Histogram of Hours per Week

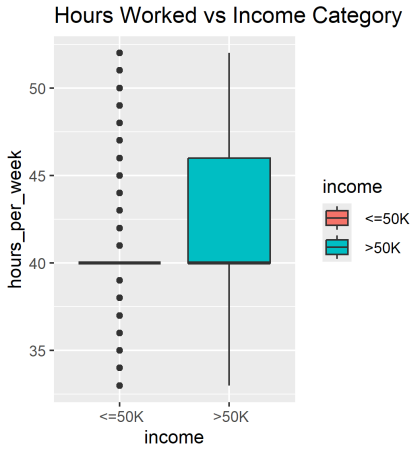


V. Box Plots of Age vs Income

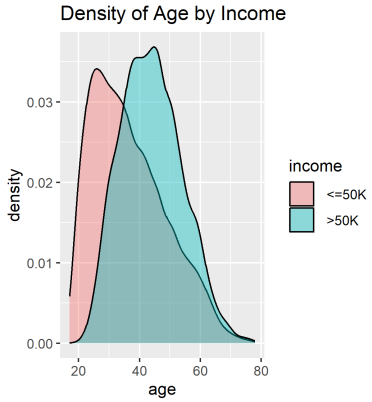
VI. Box Plots of Education Number vs Income



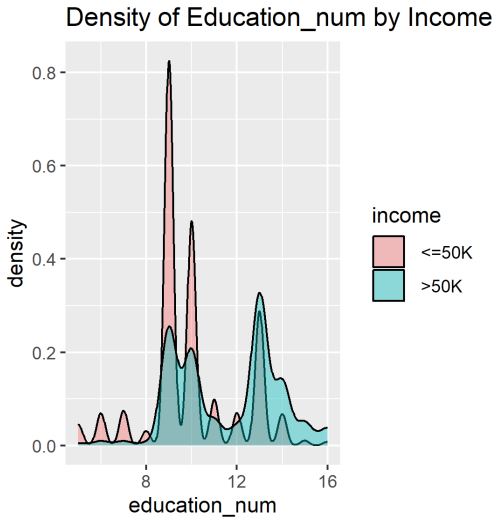
VII. Box Plots of Hours Worked vs Income



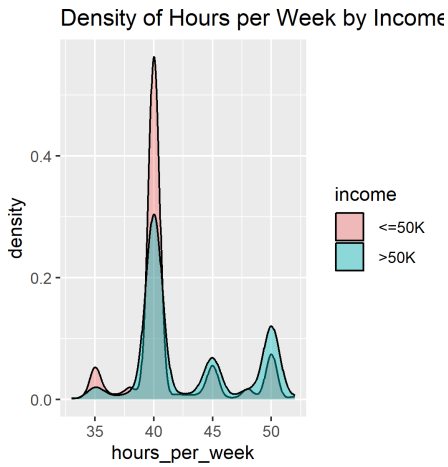
VIII. Density Plot of Age by Income



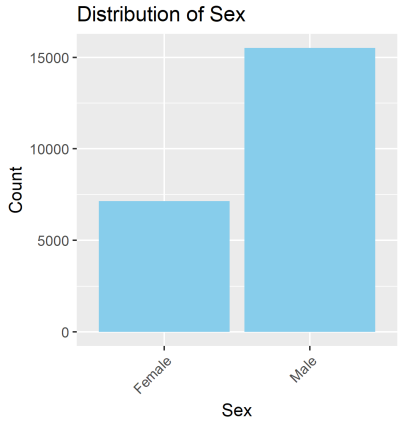
IX. Density Plot of Education Number by Income



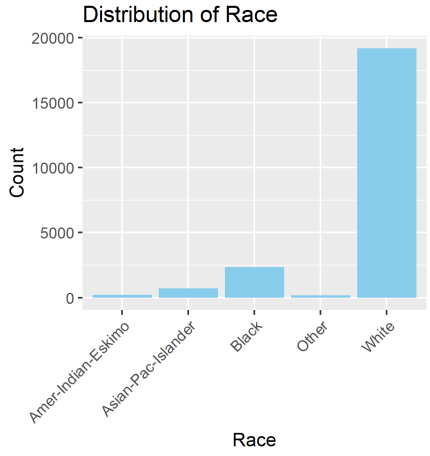
X. Density Plot of Hours Worked by Income



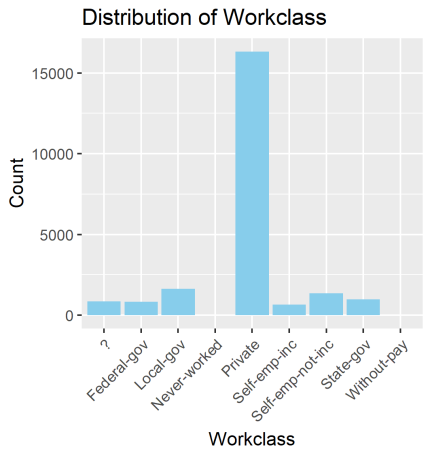
XI. Bar Plot for Distribution of Sex



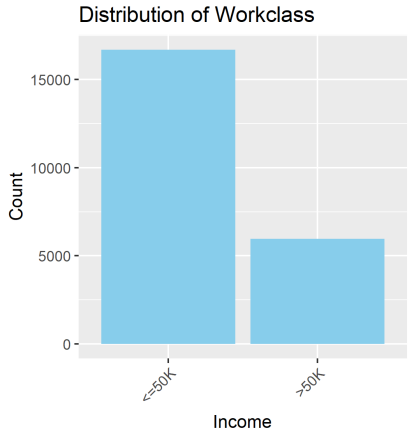
XII. Bar Plot for Distribution of Race



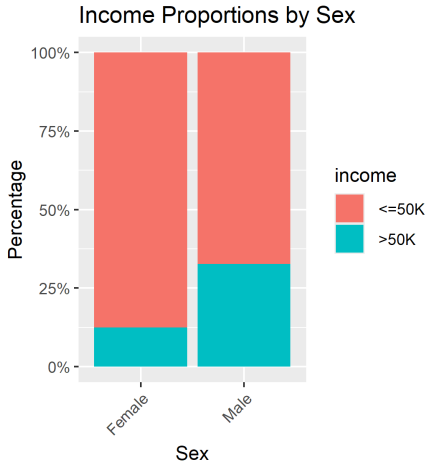
XIII. Bar Plot for Distribution of Workclass



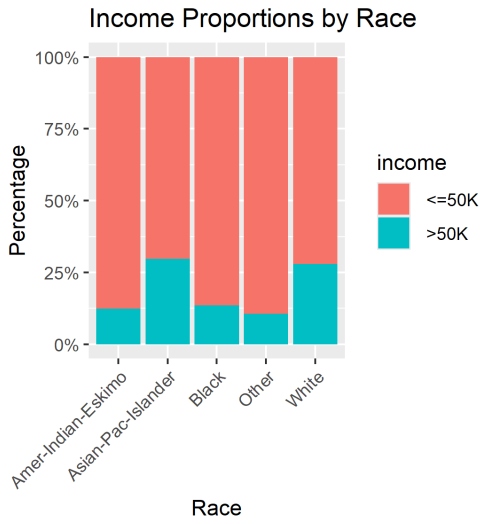
XIV. Bar Plot for Distribution of Income



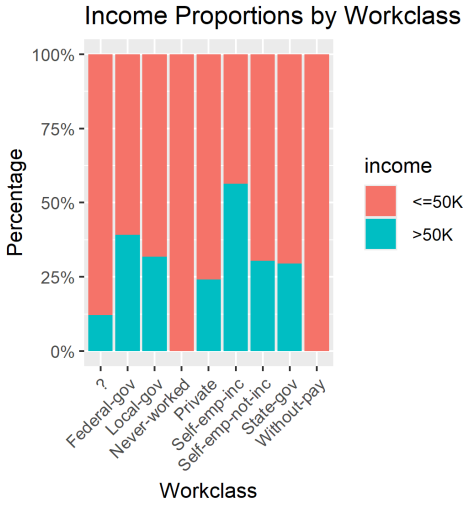
XV. Income Proportions by Sex



XVI. Income Proportions by Race



XVII. Income Proportions by Workclass



Based on the exploratory analysis, there do seem to be significant correlations between Age, Education Number, Sex, Race, and Workclass for income level. Meanwhile, Hours per Week seems relatively similar across both income levels. Thus, we will continue our analysis using Random Forest, kNN and Decision

Tree models with Age, Education Number, Sex, Race, and Workclass for inputs to predict Income Level.

2. Model Development, Validation, Optimization

The next section of this assignment will focus on Model Development, Validation and Optimization. In this I will compare performances of the Random Forest, kNN and Decision Tree Models in how well they can predict income level based on the Census and Income dataset. Particularly I will compute, accuracy, precision, recall and f1 scores, as well as the corresponding confusion matrices.

I. Random Forest Model Confusion Matrix

	actual	
predicted	<=50K	>50K
<=50K	8078	618
>50K	4357	3228

II. Random Forest Model Metrics

	Accuracy	Precision	Recall	F1
1	0.6944291	0.4255768	0.8393136	0.56478

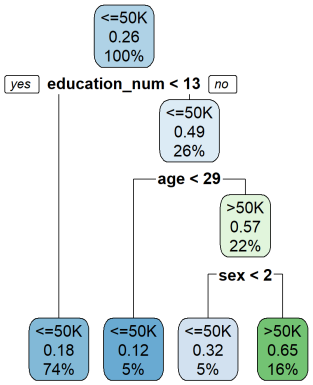
III. Decision Tree Confusion Matrix

	actual	
predicted	<=50K	>50K
<=50K	11528	2312
>50K	907	1534

IV. Decision Tree Metrics

	Accuracy	Precision	Recall	F1
1	0.8022849	0.628431	0.398856	0.4879911

V. Decision Tree Plot



VI. kNN Model Confusion Matrix

	actual	
predicted	<=50K	>50K
<=50K	10793	2158
>50K	1642	1688

VII. kNN Model Metric

	Accuracy	Precision	Recall	F1
1	0.7665991	0.5069069	0.4388976	0.4704571

3. Decisions, Conclusions

When looking at the performance metrics of the models I trained, the Decision Tree has the best accuracy, with kNN a close second. The Decision Tree has a moderate Precision score of about 0.63, but had Recall and F1 scores of .40, .49. In fact, this seemed to be common across all models, with each model having relatively low Precision, Recall and F1 scores. It's worth noting that the Random Forest model did have a surprisingly high Recall of about 0.84. Overall, none of these models performed significantly well with our test validation dataset.

These models do have relatively good accuracy, but the performance in the other metrics show that these models are likely not a good option to make decisions.