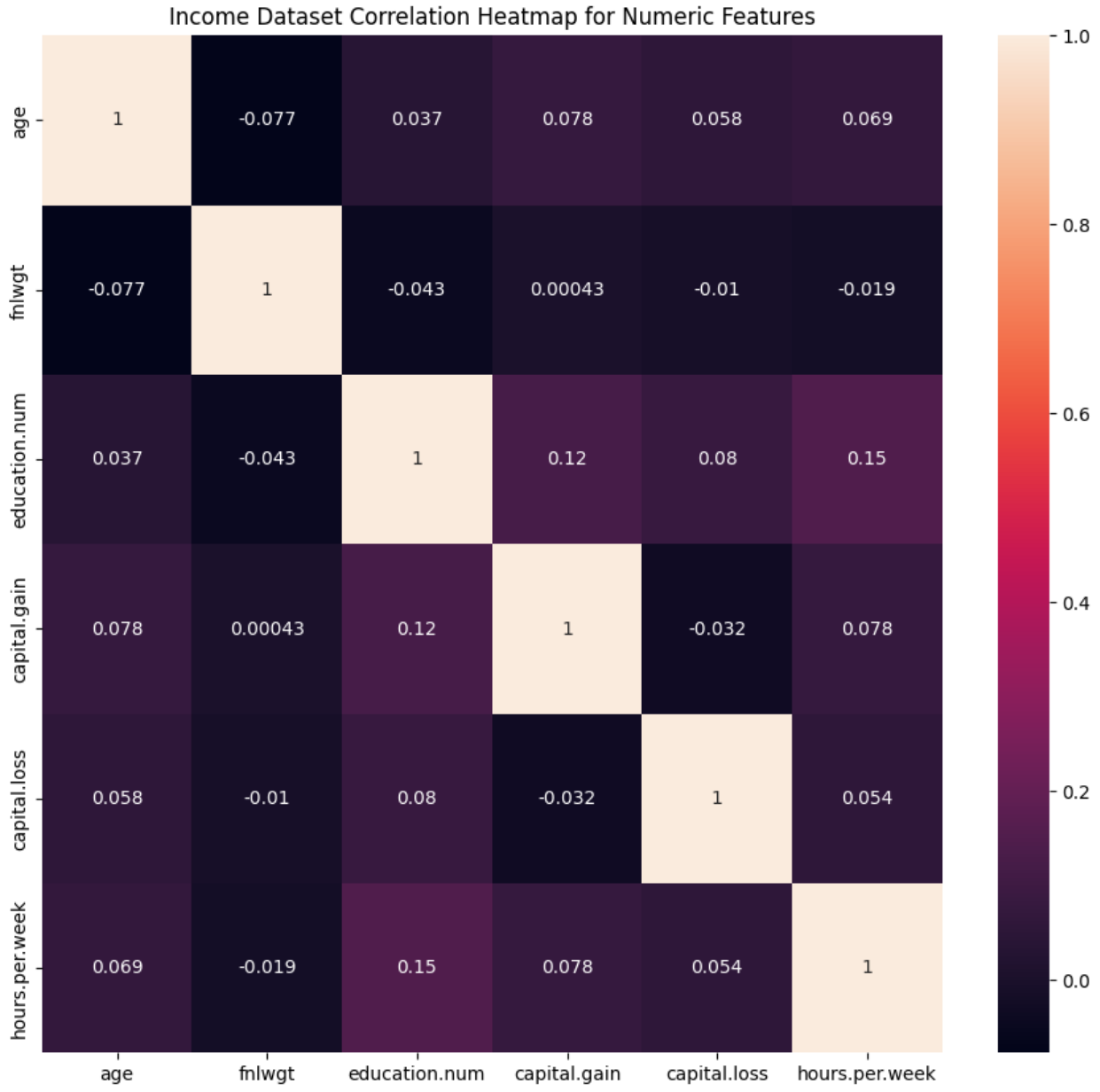
Dataset Description and Visualizations

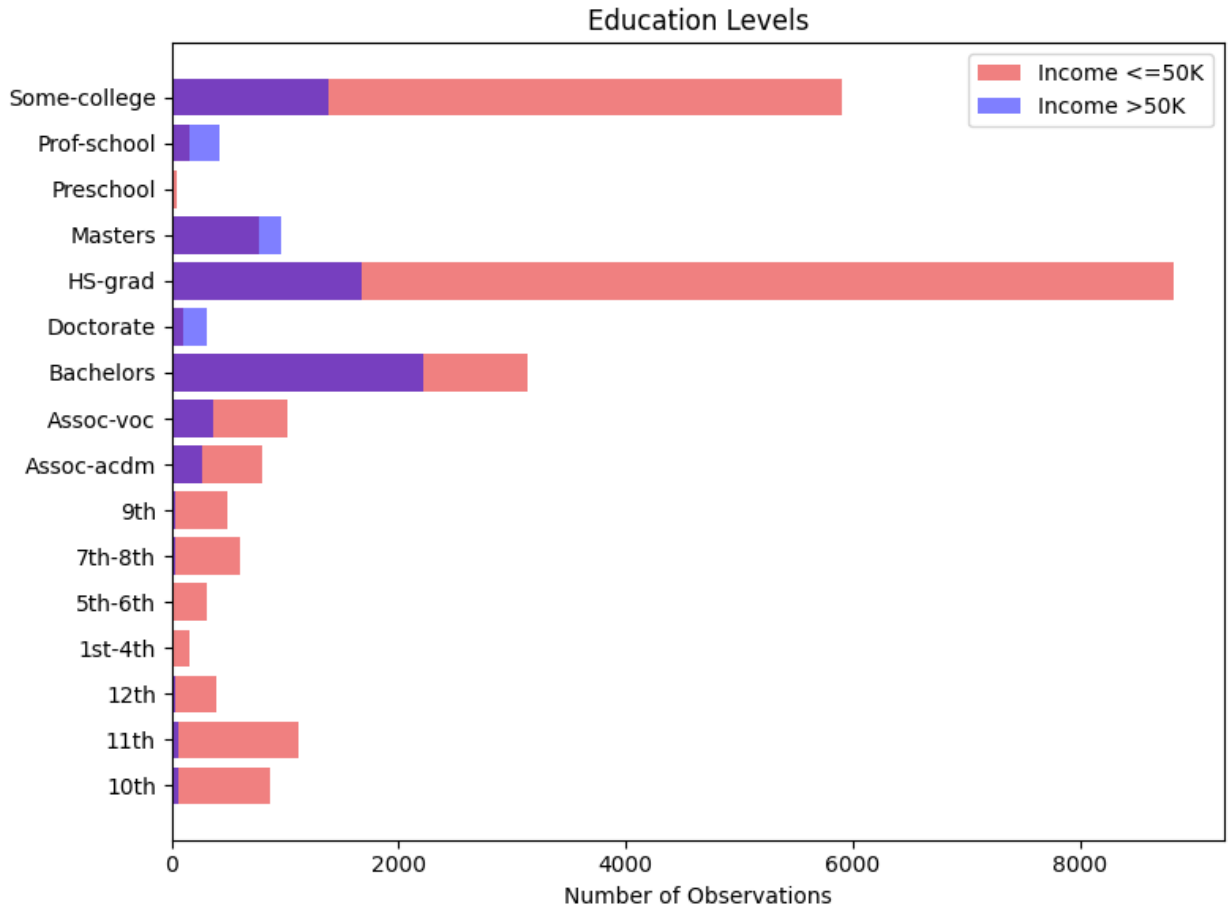
The Income dataset contains demographic and income data for individuals. The raw data set is comprised of 32,561 rows and 15 columns. The data are a combination of integers and strings. There are 2,399 rows that have at least one unknown value.

The main statistics of the numeric features are provided in the table below. Of the 32,561 entries in the dataset, 7,841 have income >50K; the remaining 24,720 observations have income <50K. So, this dataset is imbalanced.

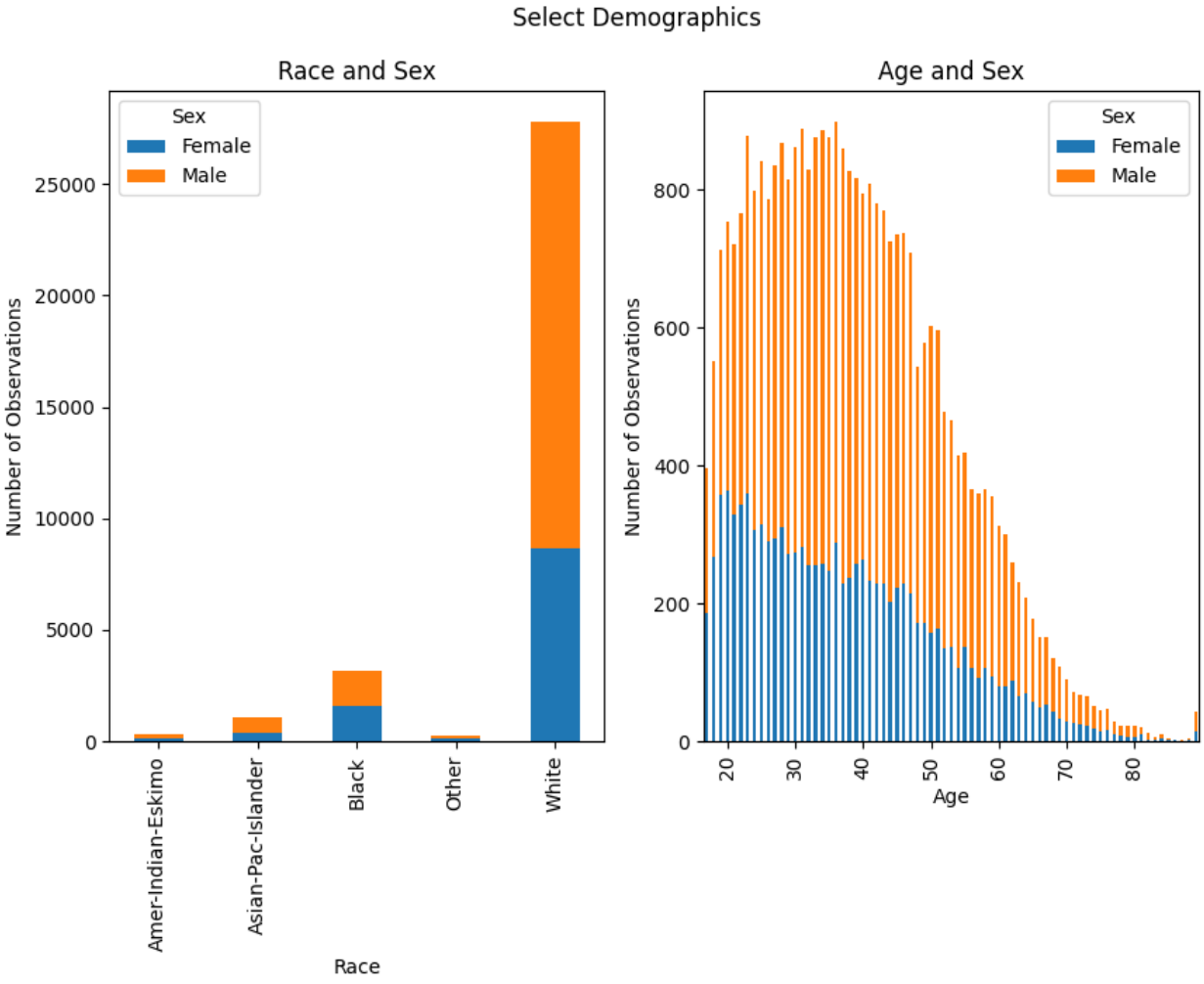
	age	fnlwgt	education.num	capital.gain	capital.loss	hours.per.week
count	32561	3.26E+04	32561	32561	32561	32561
mean	38.581647	1.90E+05	10.080679	1077.648844	87.30383	40.437456
std	13.640433	1.06E+05	2.57272	7385.292085	402.960219	12.347429
min	17	1.23E+04	1	0	0	1
25%	28	1.18E+05	9	0	0	40
50%	37	1.78E+05	10	0	0	40
75%	48	2.37E+05	12	0	0	45
max	90	1.48E+06	16	99999	4356	99



The correlations of the numeric features are visualized in the above heatmap. Most of these variables are not correlated. The highest coefficient is 0.15 between the numeric representation of education level and hours worked per week, but this maximum coefficient is too small to be considered even weak correlation. Additional visualizations were employed to get a sense of how the data is distributed across the dataset, including looking at the categorical variables.



The bar graph above examines the breakdown of education level within both of the two income groups. As expected, there are far more observations for the lower income group, but their general distribution differs from that of the higher income group. High school graduates are by far the most common in the lower income group, followed by individuals with some college education. In the higher income bracket, it is most common for individuals to possess a bachelors degree; high school is the second most common education level. Further, despite the overall greater number of observations in the lower income group, there are more total observations belonging to the higher income group for professional school, masters, and doctorate education levels.



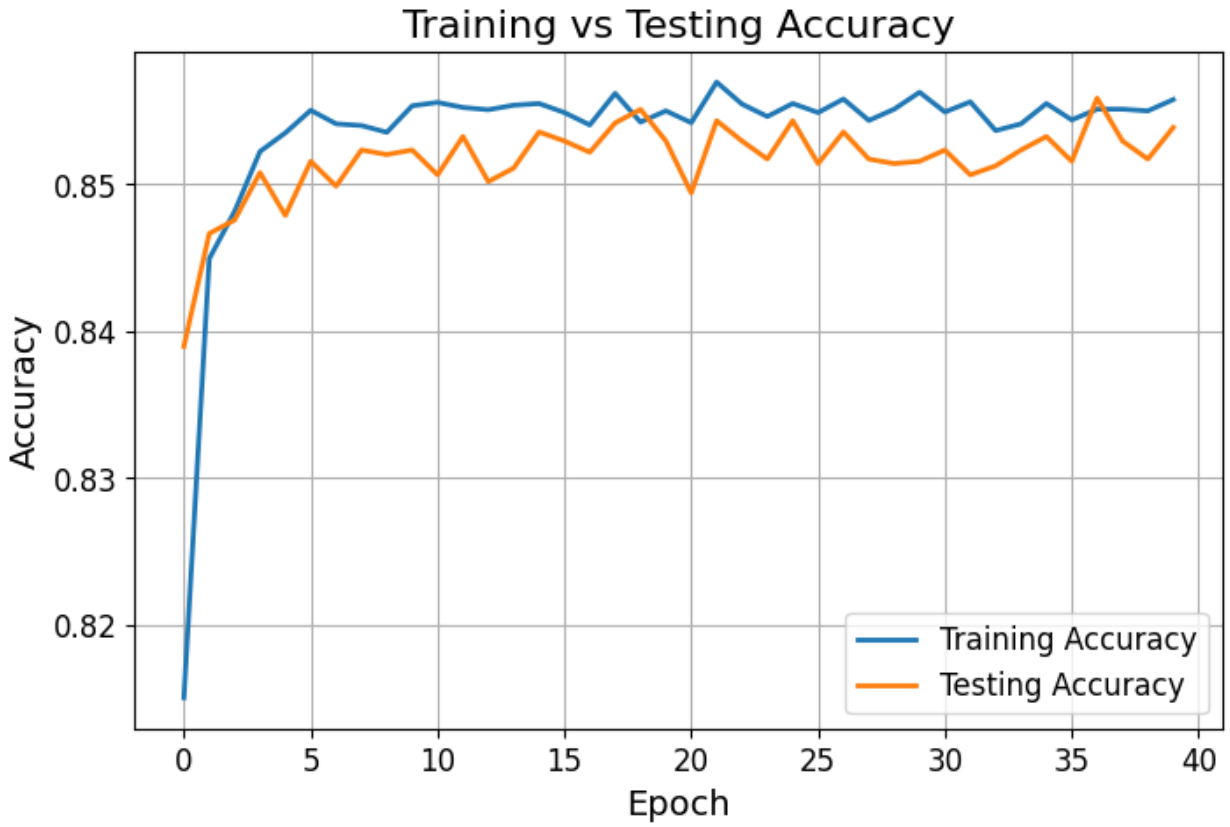
Additional demographic information is visualized in the bar charts above. The first panel shows the male/female split within the distribution of different races noted in the dataset. This is another area of imbalance: white individuals make up the vast majority of the dataset. The participants also tend to be male, though the sex split is more even for the Black group. Age is pretty well distributed across the range of typical working ages. Most observations occur within the 25-45 range and drop off as age continues to increase. This is compatible with our knowledge of overall workforce trends, where the number of working individuals decreases at higher ages due to retirement. (Other factors may play a role, but retirement is a key driving force.) The proportion of male to female participants is reasonably consistent across the age range.

Preprocessing and Neural Network Construction

Preprocessing needed to address how to handle the missing values in the dataset. Dropping all rows with missing data results in a loss of over 7% of the available data. In order to retain this data, the unknown values in each column were replaced by the most common value for their respective columns. During preprocessing, columns with string data types were transformed to one hot encoded vectors. The data were also standardized so that each feature has a mean of 0 and a standard deviation of 1.

A two-layer neural network was built using Keras. The network has 96 input neurons (plus 1 bias). The hidden layer has 6 neurons and uses the ReLU activation function. Since this is a binary classification problem, the output layer has 2 neurons and uses the sigmoid activation function. L2 regularization was applied to the hidden and output layers, and the Adam optimizer was used. The model was trained for 40 epochs.

The resulting testing accuracy was 85.38%.




Training vs Testing Loss
Training Loss
Testing Loss
Loss
Epoch
0.46
0.44
0.42
0.40
0.38
0.36
0.34
0
5
10
15
20
25
30
35
40