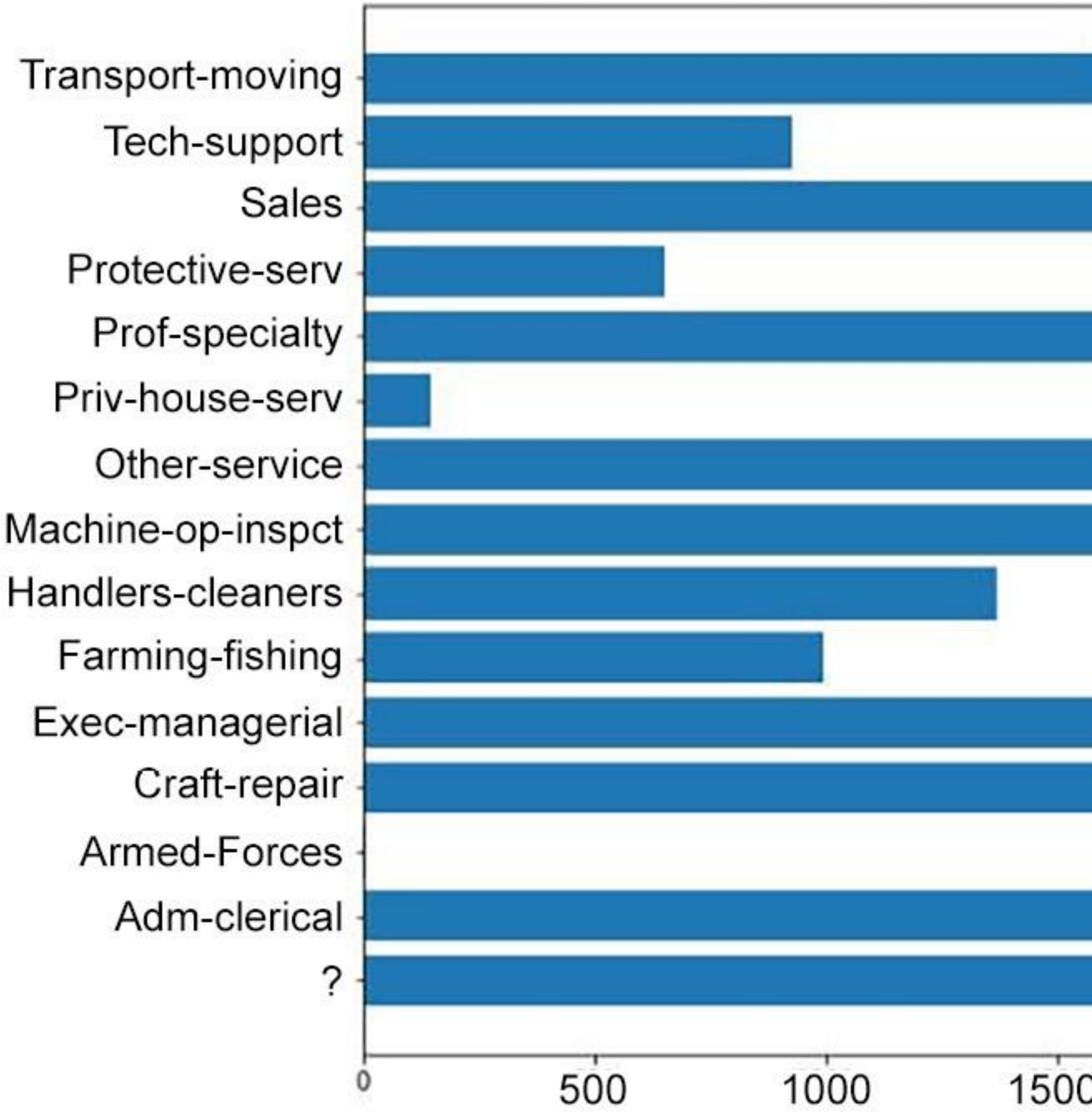
Working with the Adult Income Dataset (UCI)

In this activity, we will detect outliers in the Adult Income Dataset from the UCI machine learning (https://archive.ics.uci.edu/dataset/2/adult)
portal https://raw.githubusercontent.com/fenago/datasets/main/adult_income_data.csv.

We will use the concepts we've learned throughout this lecture, such as subsetting, applying user-defined functions, summary statistics, visualizations, boolean indexing, and group by to find a whole group of outliers in a dataset. We will create a bar plot to plot this group of outliers. Finally, we will merge two datasets by using a common key.

These are the steps that will help you solve this activity:

1. Load the necessary libraries.
2. Create a script that will read a text file line by line.
3. Add a name of `Income` for the response variable to the dataset.
4. Find the missing values.
5. Create a DataFrame with only age, education, and occupation by using subsetting.
6. Plot a histogram of age with a bin size of `20`.
7. Create a function to strip the whitespace characters.
8. Use the `apply` method to apply this function to all the columns with string values, create a new column, copy the values from this new column to the old column, and drop the new column. ***Consider this one a challenge***
9. Find the number of people who are aged between `30` and `50`.
10. Group the records based on age and education to find how the mean age is distributed.
11. Group by occupation and show the summary statistics of age. Find which profession has the oldest workers on average and which profession has its largest share of the workforce above the 75th percentile.
12. Use `subset` and `groupBy` to find the outliers.
13. Plot the outlier values on a bar chart. It should look something like this:


Transport-moving
Tech-support
Sales
Protective-serv
Prof-specialty
Priv-house-serv
Other-service
Machine-op-inspct
Handlers-cleaners
Farming-fishing
Exec-managerial
Craft-repair
Armed-Forces
Adm-clerical
?
0
500
1000
1500