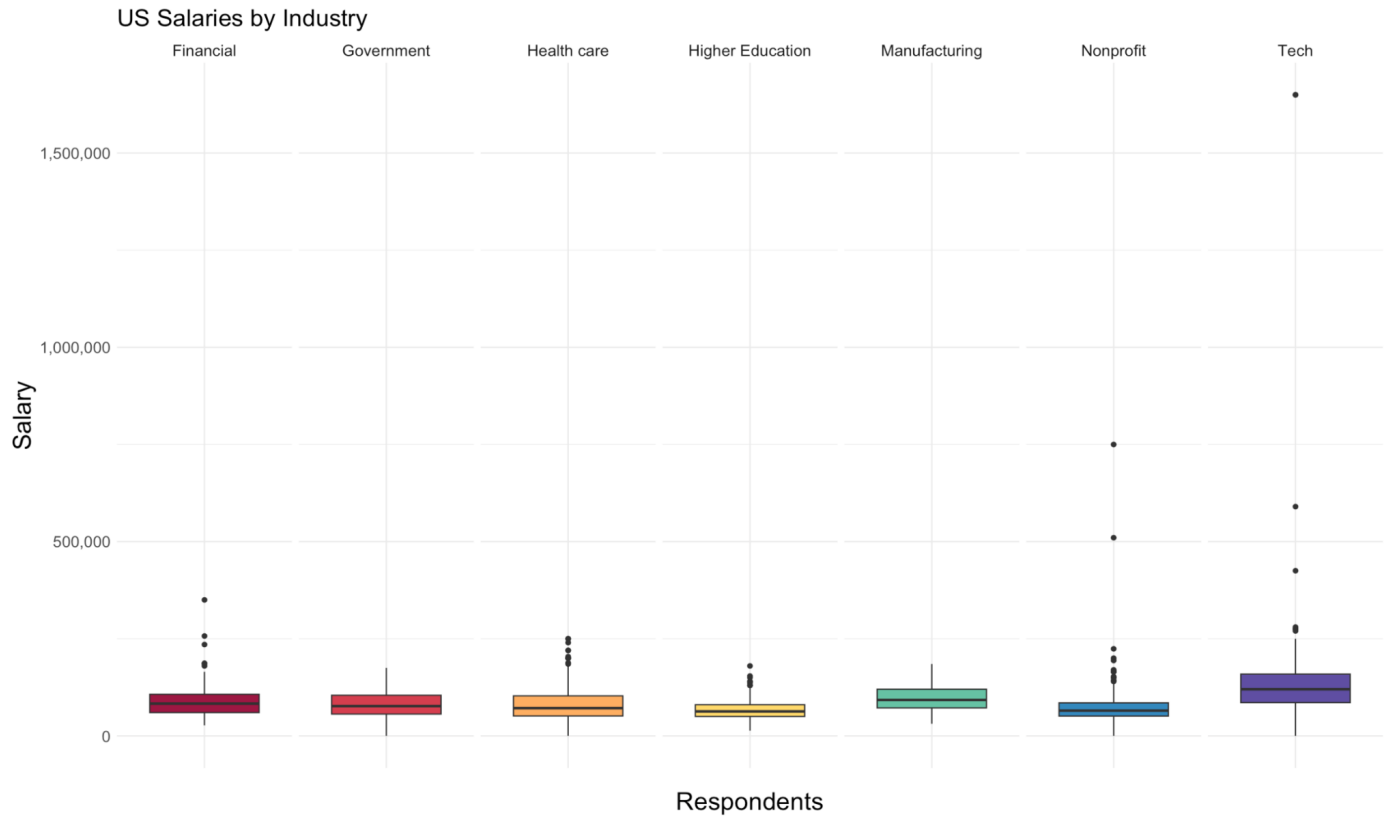
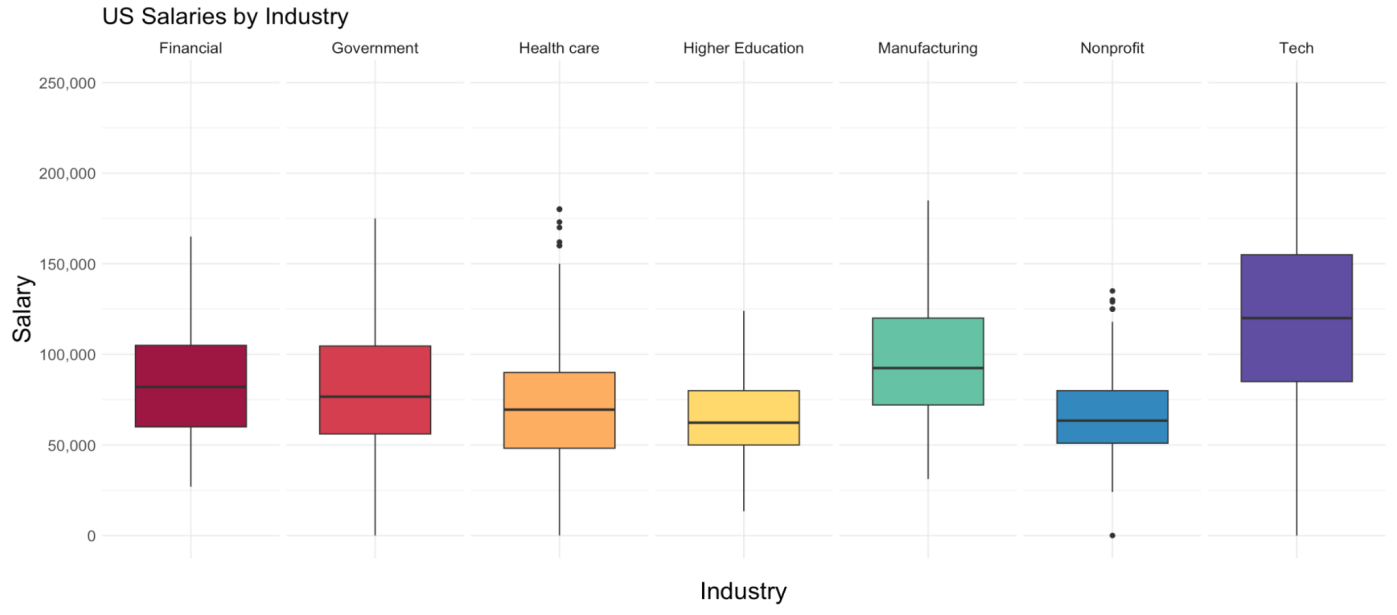
Explanation of visualizations

After subsetting the data to focus on the US and industries that had at least 100 respondents, I began creating visualizations to answer the questions I defined earlier in this process. These first two box plots were used to confirm the existence of probable outliers and to subsequently confirm their removal.

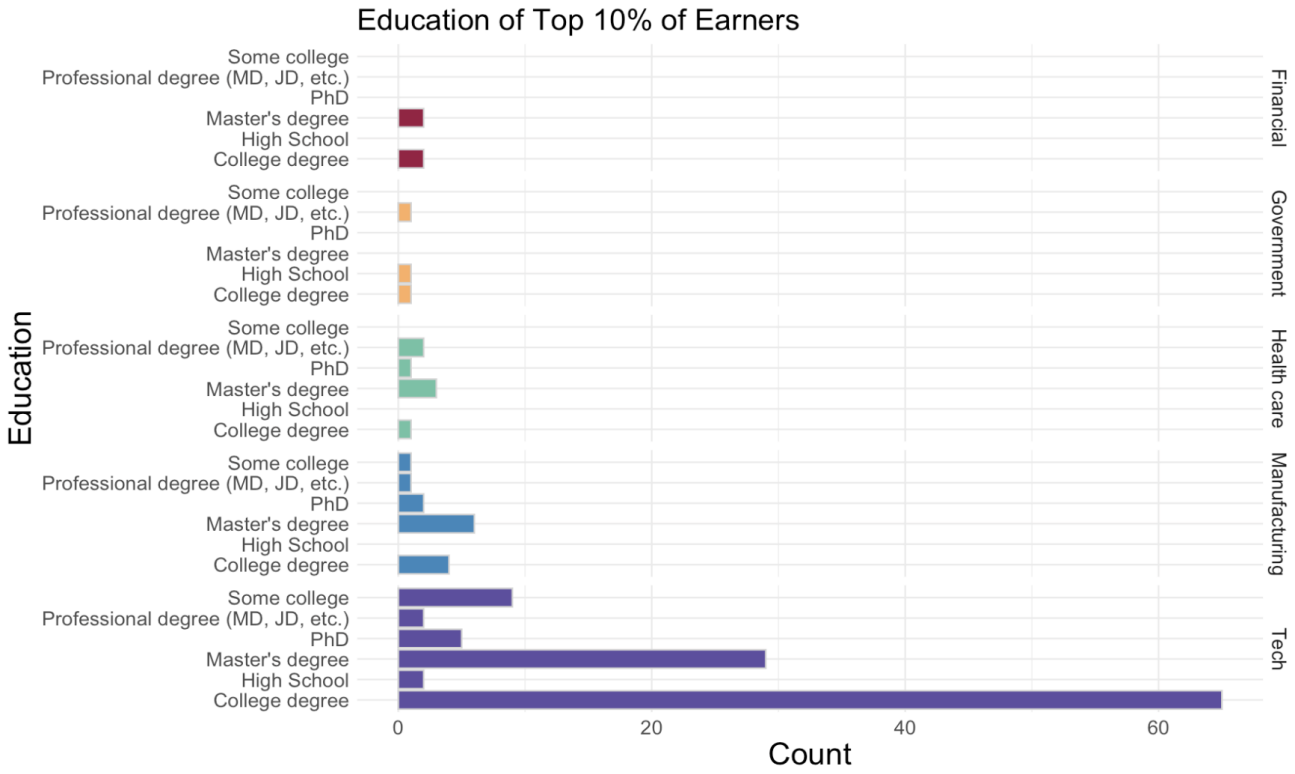
Initial plot containing outliers:

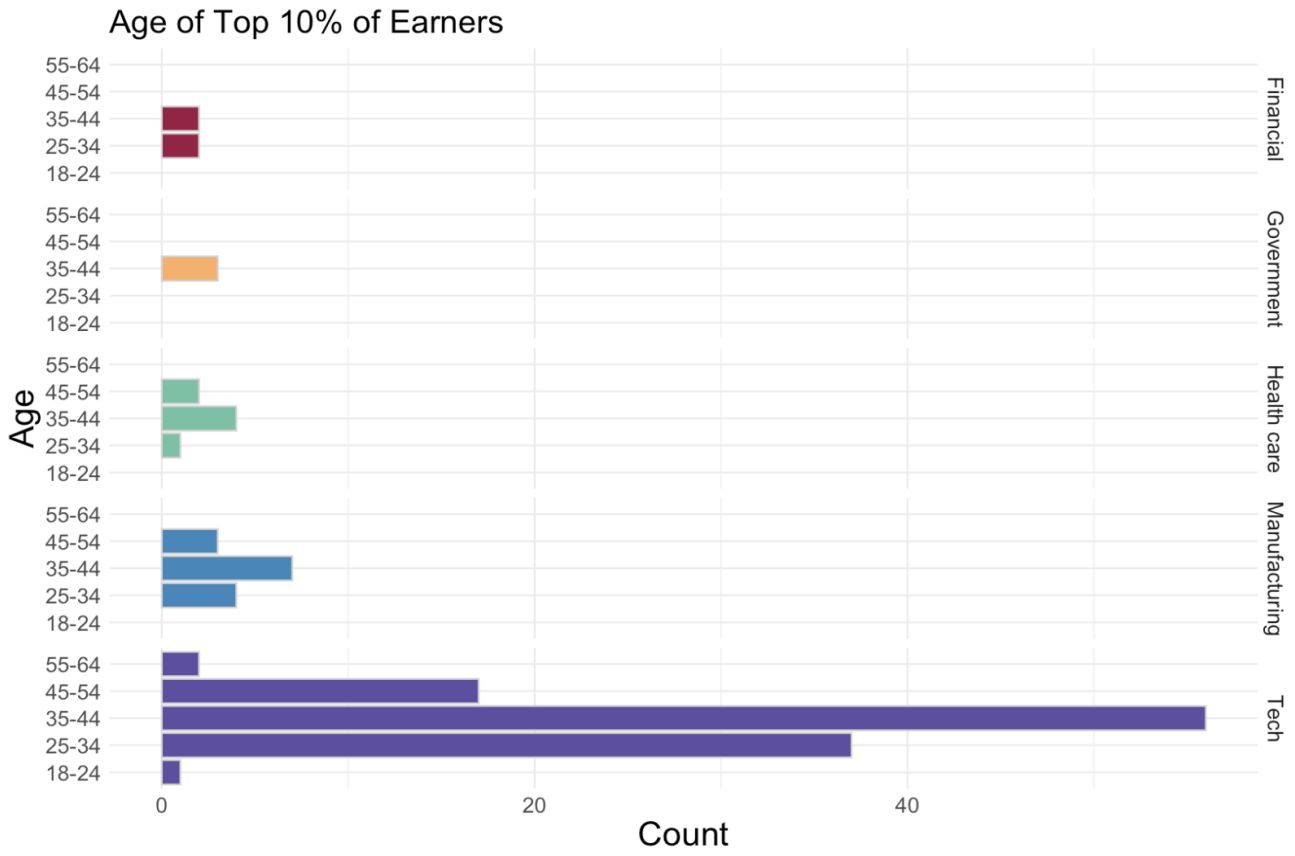


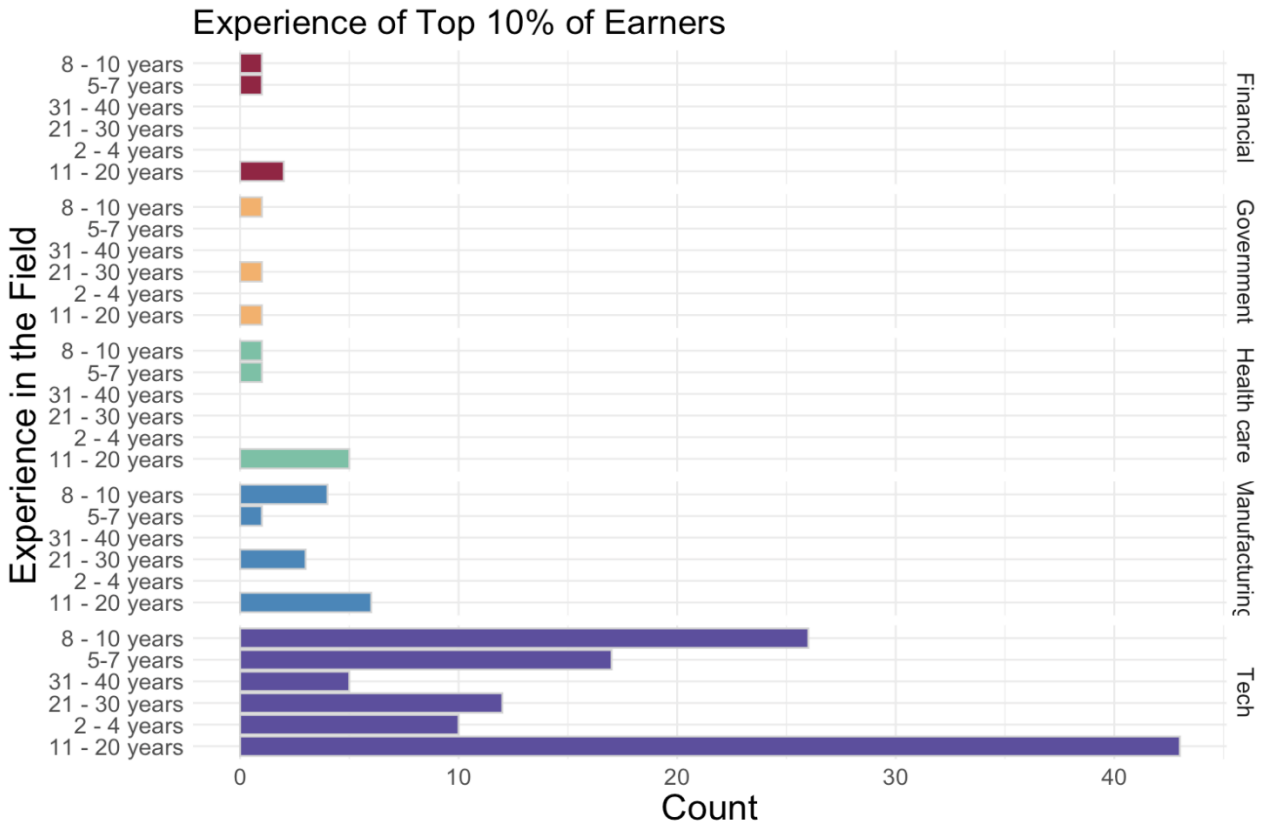
Data replotted with outliers removed:

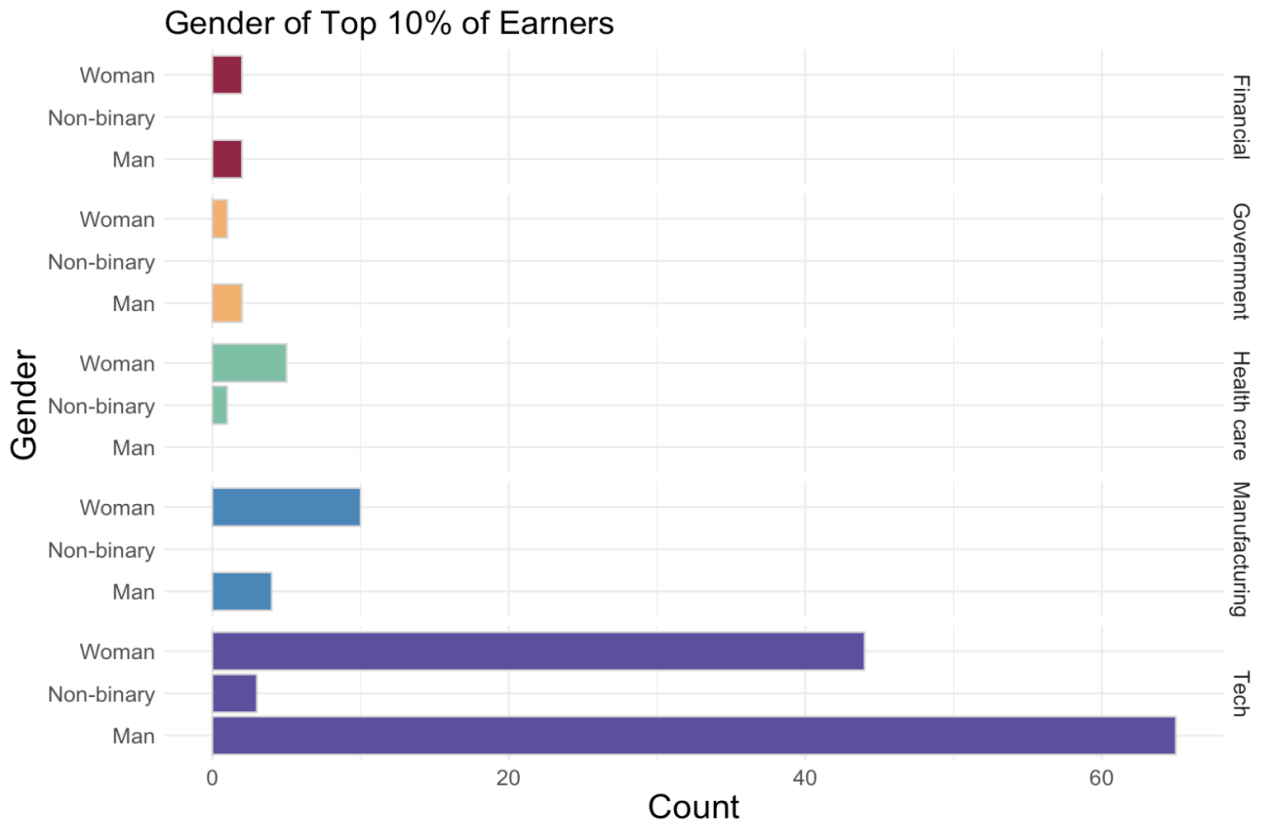


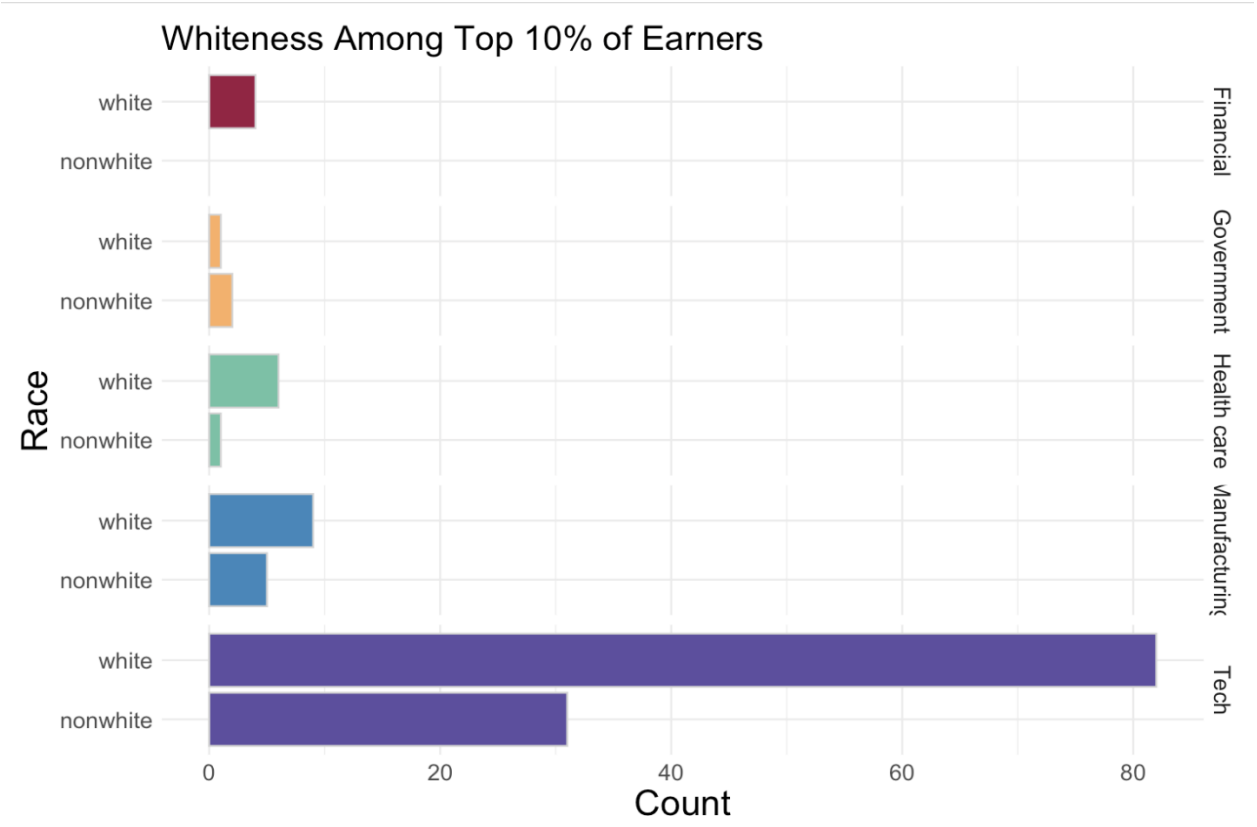
The following series of histograms was used to illustrate characteristics of the top 10% of earners in these five industries.



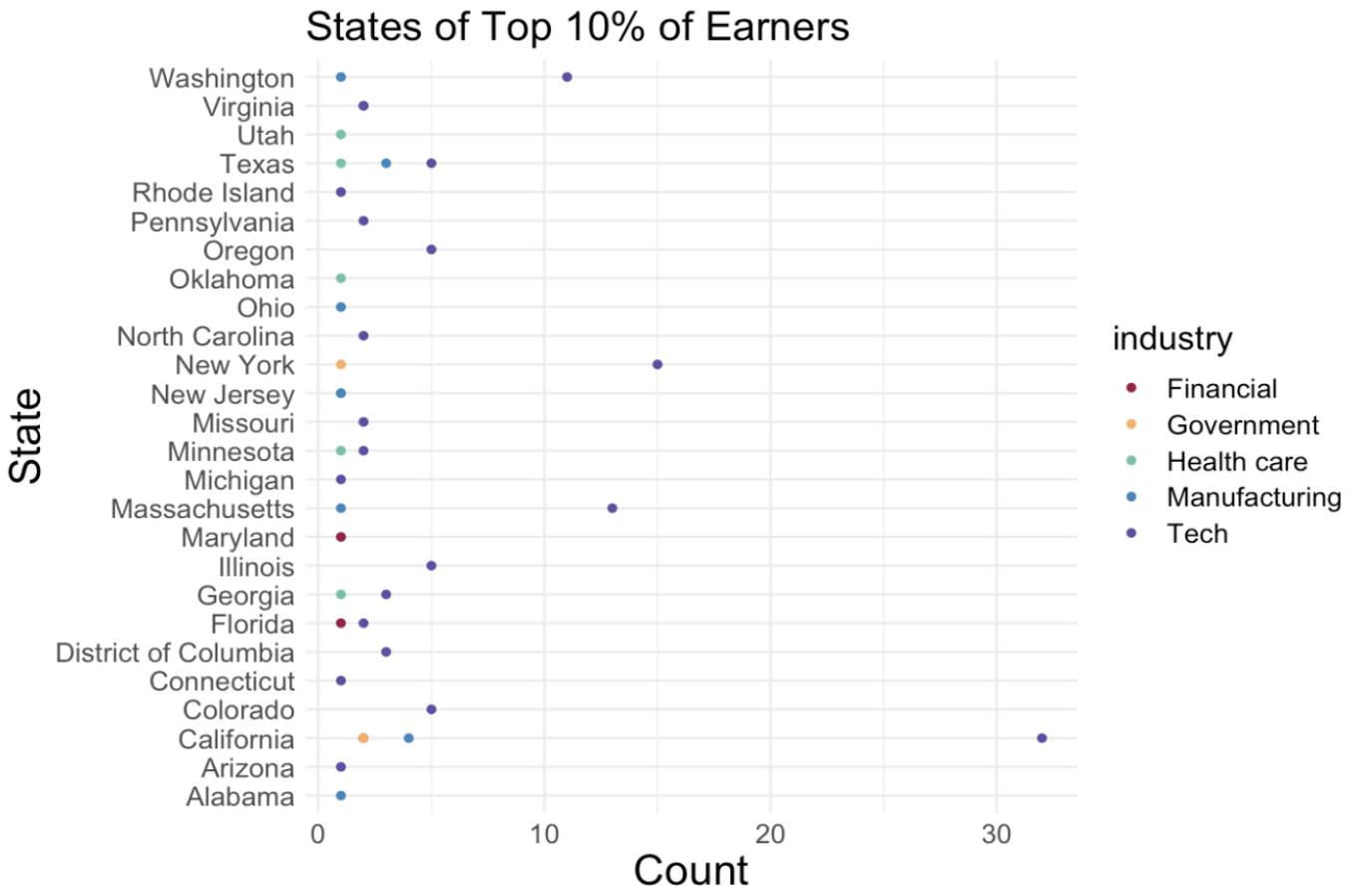



Experience of Top 10% of Earners
Experience in the Field
8 - 10 years
5-7 years
31 - 40 years
21 - 30 years
2 - 4 years
11 - 20 years
Financial
Government
Health care
Manufacturing
Tech
0
10
20
30
40
Count

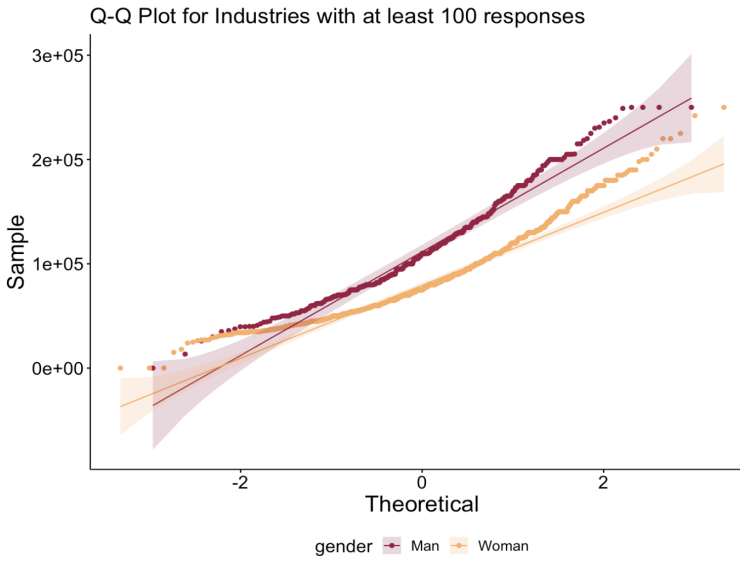

Gender of Top 10% of Earners
Gender
Woman
Non-binary
Man
Financial
Government
Health care
Manufacturing
Tech
0
20
40
60
Count

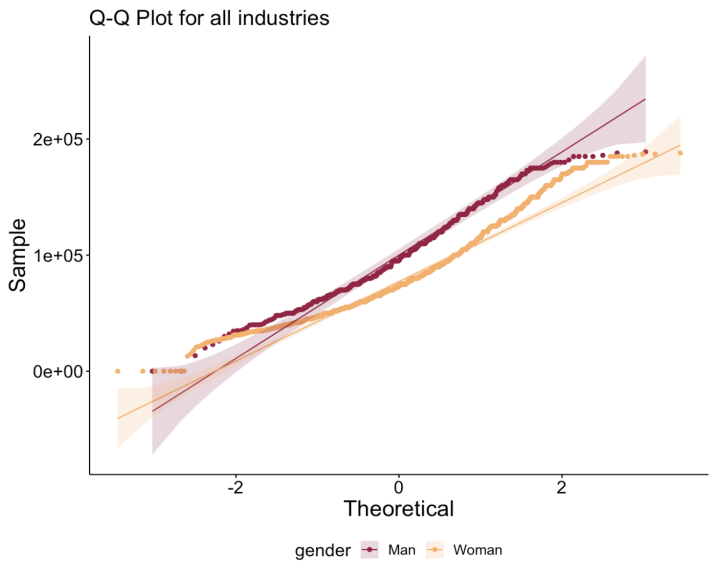


I later created a scatter plot to summarize where the top 10% of earners are located in the US.

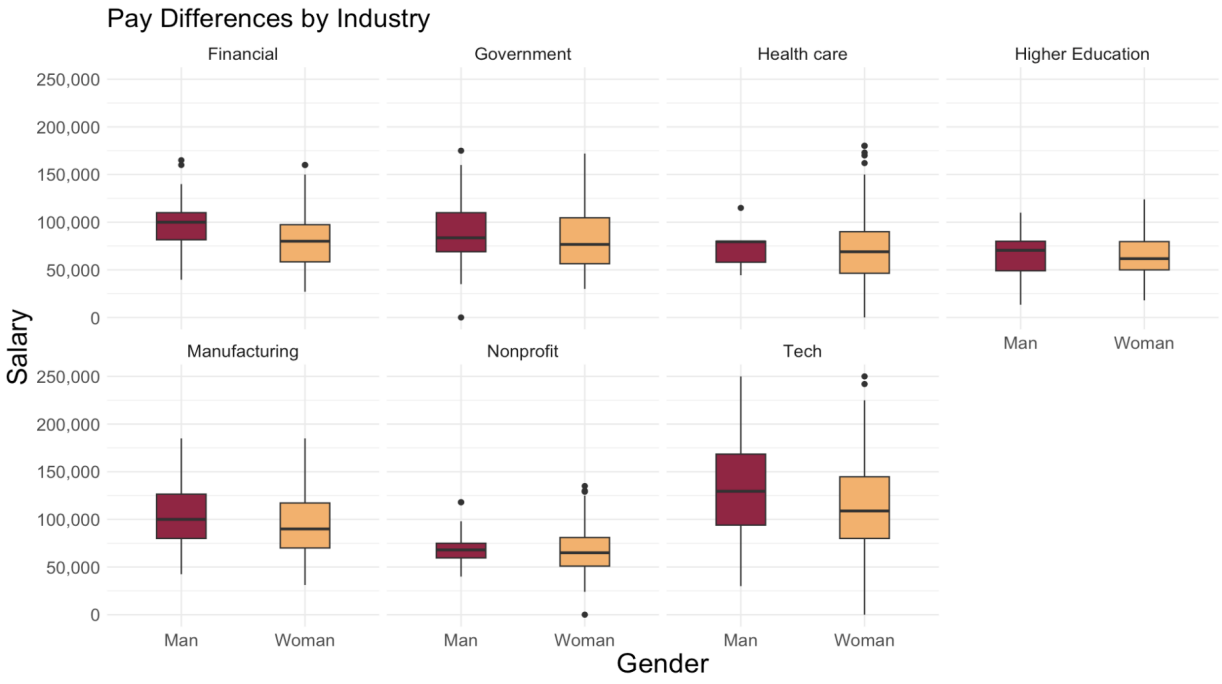


After creating the summary histograms and scatter plot I shifted my focus to the gender pay gap seen in the dataset. I intended to run sets of two-sample t-tests to determine whether or not the pay differences observed in the dataset are statistically significant, however, in order to use the t-test the data would need to be normally distributed and there would need to be no significant difference in variance between the two comparison groups. In addition to running Shapiro-Wilks tests and F-tests, I created quantile-quantile plots (Q-Q plots) of the data to test for normality. Both the plots and the statistical tests showed that normality could not be assumed and that there was significant difference in variance between the two comparison groups. As a result, I could not use two-sample t-tests to assess the significance of the differences seen in the dataset. The first Q-Q plot shows only industries with at least 100 respondents while the second includes the full dataset.



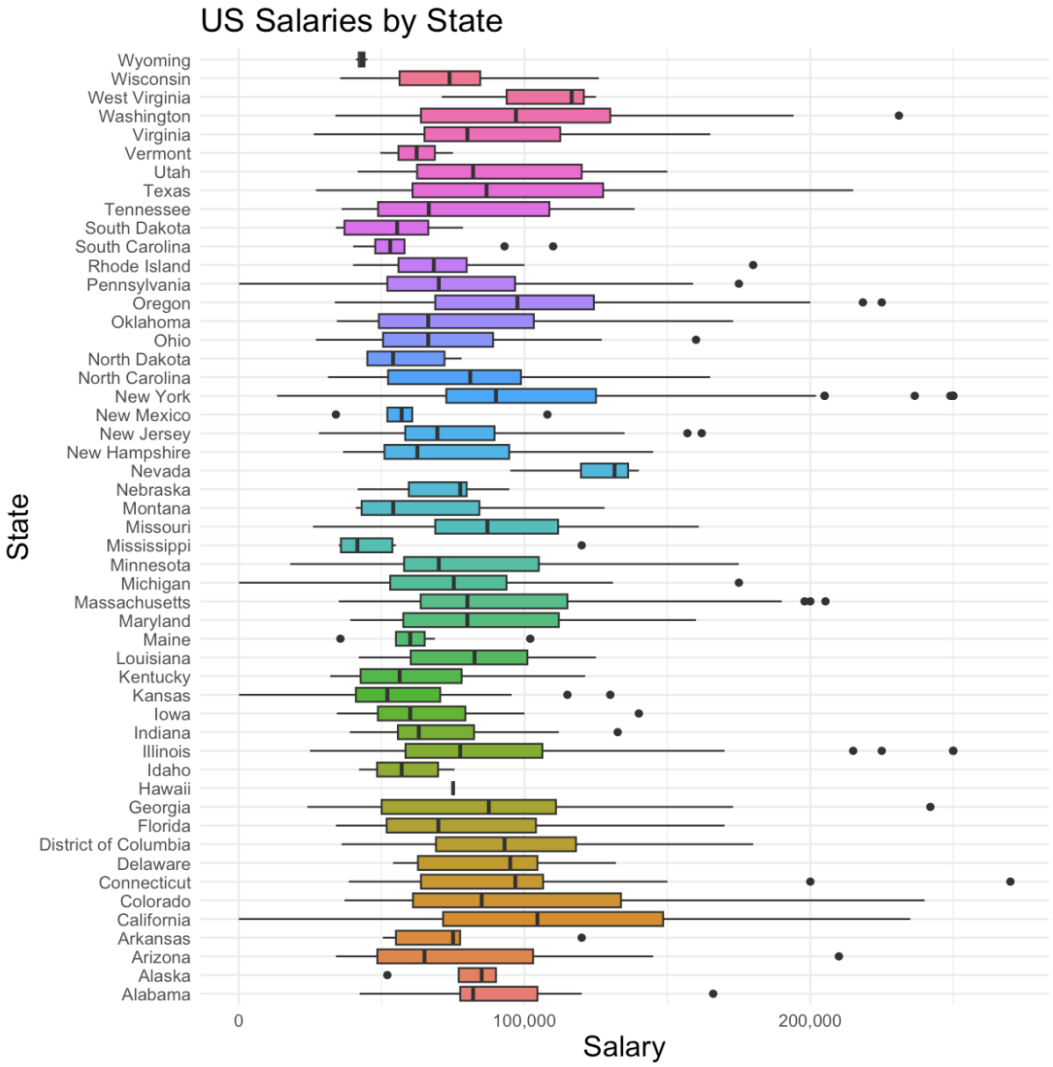


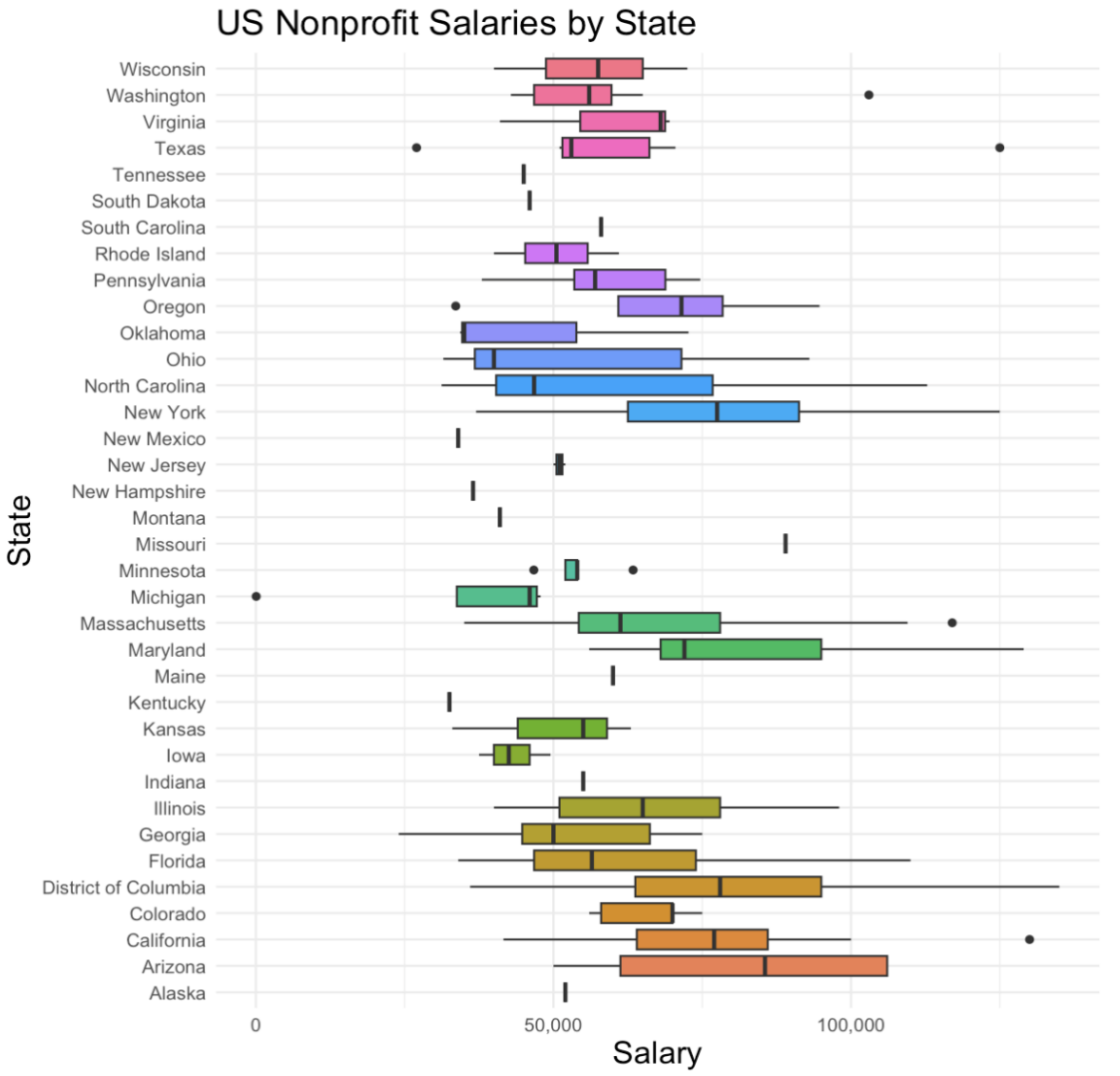
Although I could not run more statistical tests, I made the next visualization to summarize observable pay gaps. From this plot you can see that men have higher salaries on average than women but the distribution of salaries between the two groups is similar within each industry. The largest disparity is found in Tech.



I made one more plot to show the pay differences between the two groups and added white points to mark the mean salary for each. The mean salary is also printed to the right of each point. Overall the average salary of a man is $20,693.80 more than the average woman's salary.

The final two boxplots I made show how salaries differ between states. The first of these plots shows all industries and the second shows this relationship for only nonprofits.




US Nonprofit Salaries by State
Wisconsin
Washington
Virginia
Texas
Tennessee
South Dakota
South Carolina
Rhode Island
Pennsylvania
Oregon
Oklahoma
Ohio
North Carolina
New York
New Mexico
New Jersey
New Hampshire
Montana
Missouri
Minnesota
Michigan
Massachusetts
Maryland
Maine
Kentucky
Kansas
Iowa
Indiana
Illinois
Georgia
Florida
District of Columbia
Colorado
California
Arizona
Alaska
State
0
50,000
100,000
Salary