Tech Salary 2016

Appendix

Appendix 1

The dataset has a lot of inconsistencies including empty strings, blanks, Nan values. I used regex (pattern matching), janitor package, filter and other operations to remove the erroneous values.

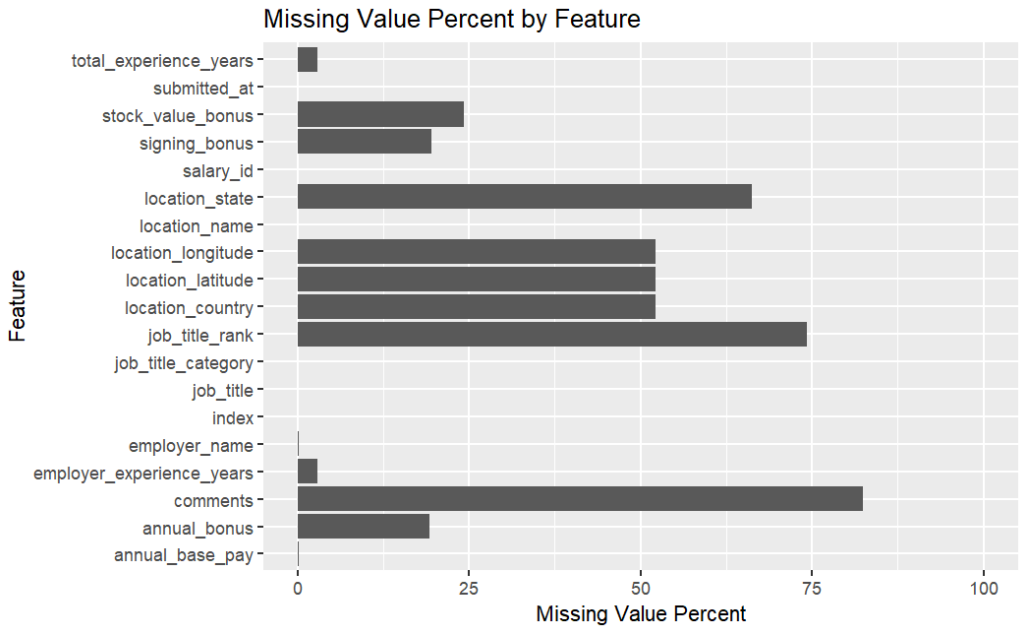


Figure 1: Missing Values

Appendix 2

Notably, the San Francisco area is characterized by a scarcity of individuals in the medium and low earning groups, whereas New York boasts a more varied distribution.

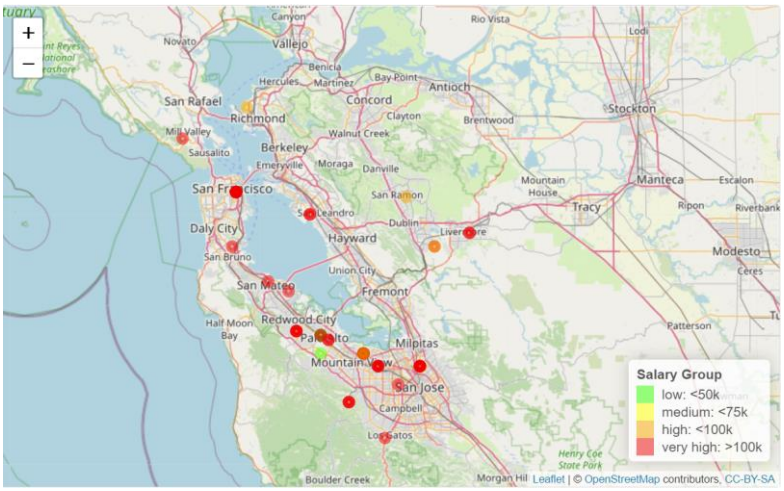


Figure 2: San Francisco Area

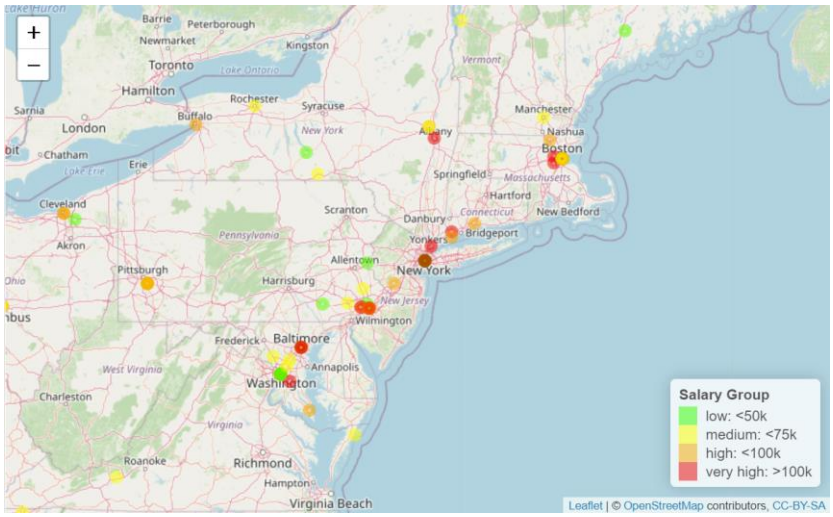
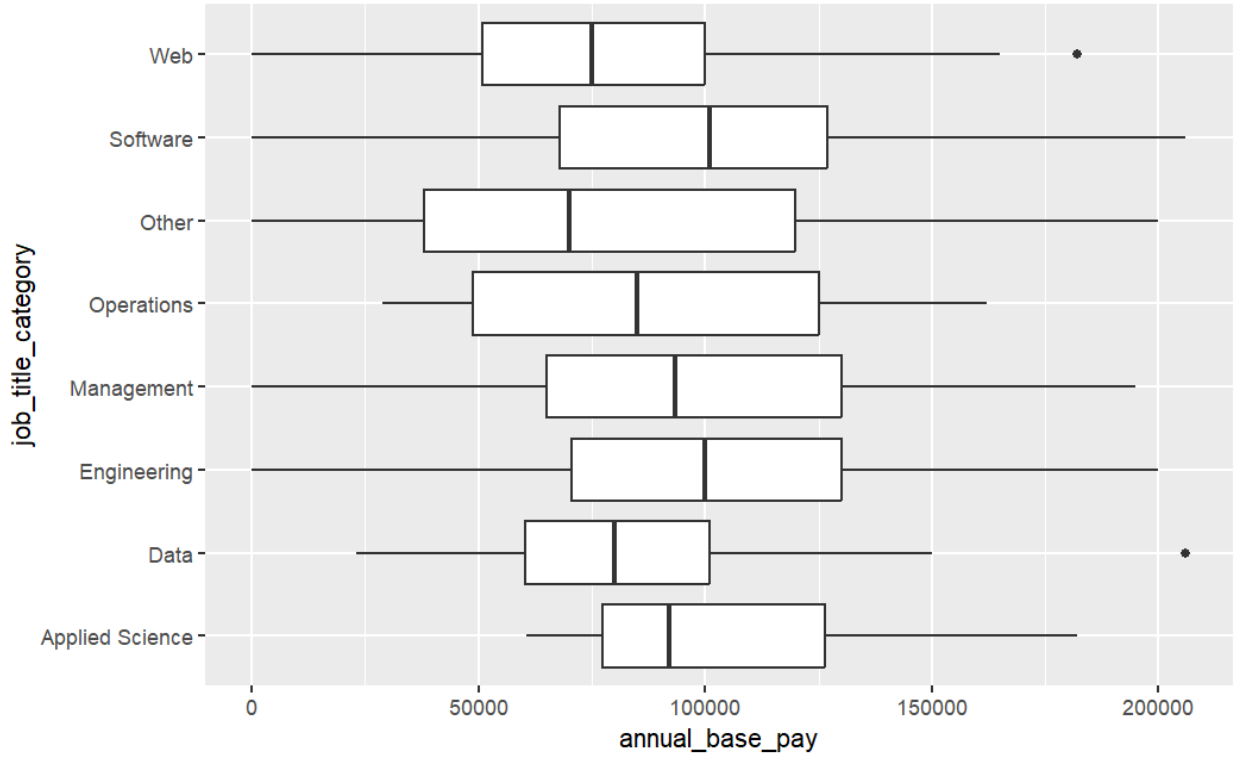


Figure 3: New York

Appendix 3

Annual base pay distribution based on job title.



Appendix 4

Annual base pay based on experience group.

We create a new column "experience_group" which groups the values in the "total_experience_years" column into different categories. Figure 5 illustrates the distribution of annual base pay across various experience levels using violin plots. The quantile lines for each group further reinforce the idea that each experience group possesses a distinct distribution. We use the Kruskal-Wallis rank sum test which is a non-parametric method for comparing the central tendency of two or more groups.

```
Kruskal-Wallis chi-squared = 117.57, df = 4, p-value < 2.2e-16
```

The low p-value indicates that there is a statistically significant difference in the annual base pay among the different experience groups.