Salary Prediction Initiative

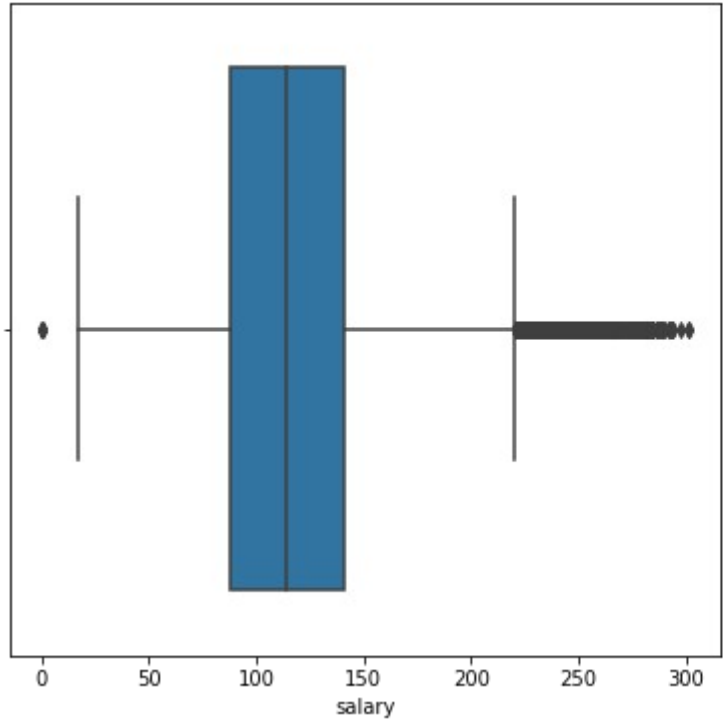
Graphical Analysis

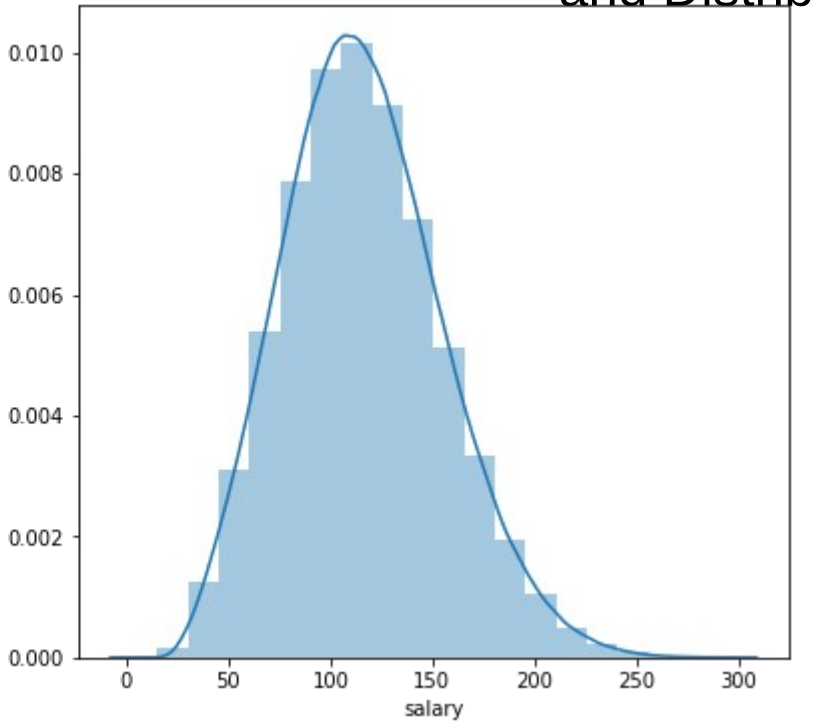
John Freeman

EDA Overview

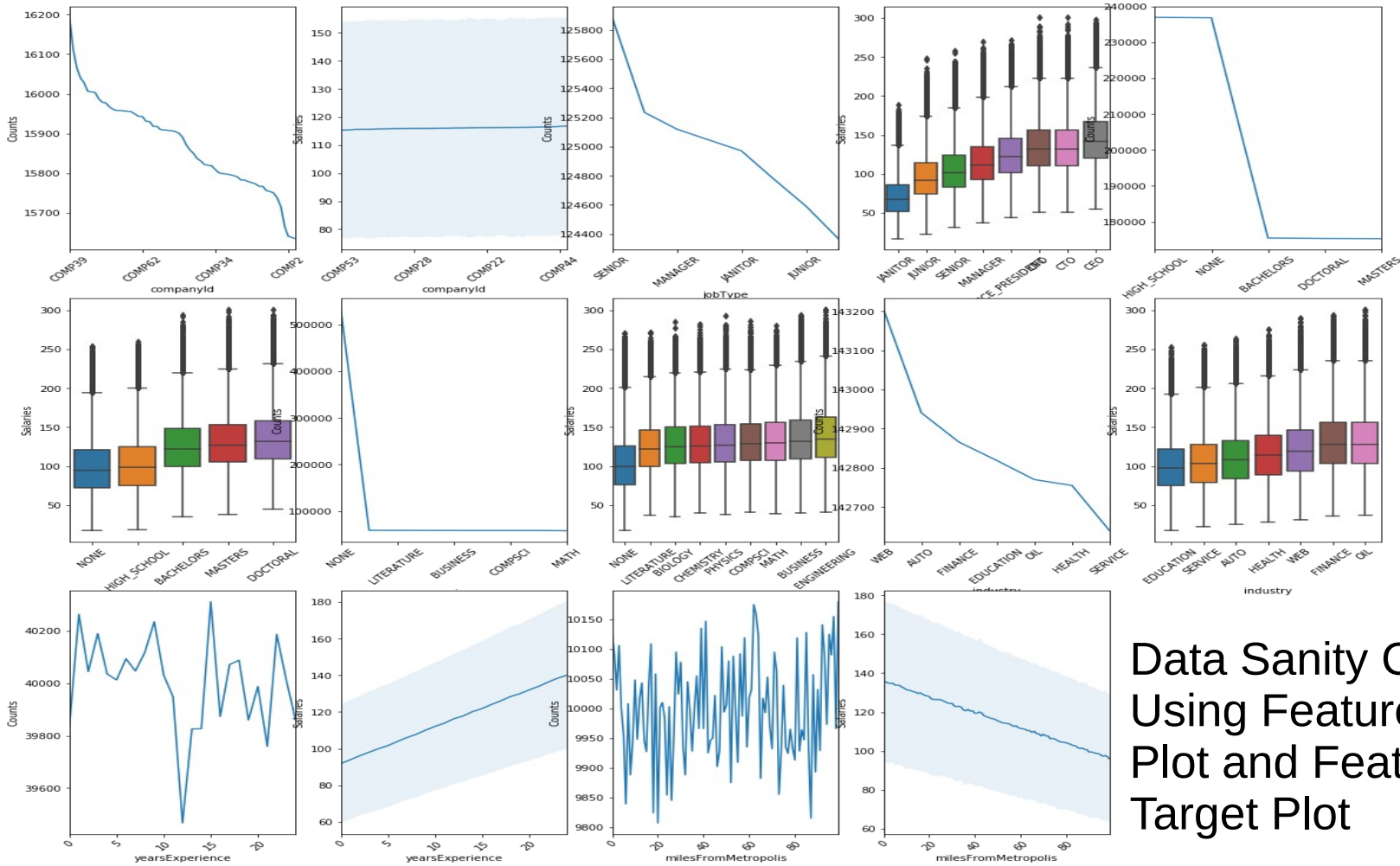
- Several plots were made to explore the relationships between various features
- This will guide our approach to modeling the data, especially categorical features
- The data covers over one million records (with no duplicates)
- Generally appears to be uniformly distributed among many features
- Several rows of outlier data were removed to prevent skewing of results

Target Check Using Box and Distribution Plot





As we can see in the boxplot, there are some outliers both above and below the "Maximum" and "Minimum" respectively. These are investigated and resolved in the EDA. In the distribution plot, it shows a very nice distribution of the salary around a bell-shaped curve.



Data Sanity Check Using Feature Count Plot and Feature vs Target Plot

The graphs on the previous page show the following correlation between the listed feature and the target (salary):

Company - Shows a weak association with Salary; Salaries don't seem to vary much between companies

JobType - Salary is clearly positively associated; more senior and executive positions are higher paid

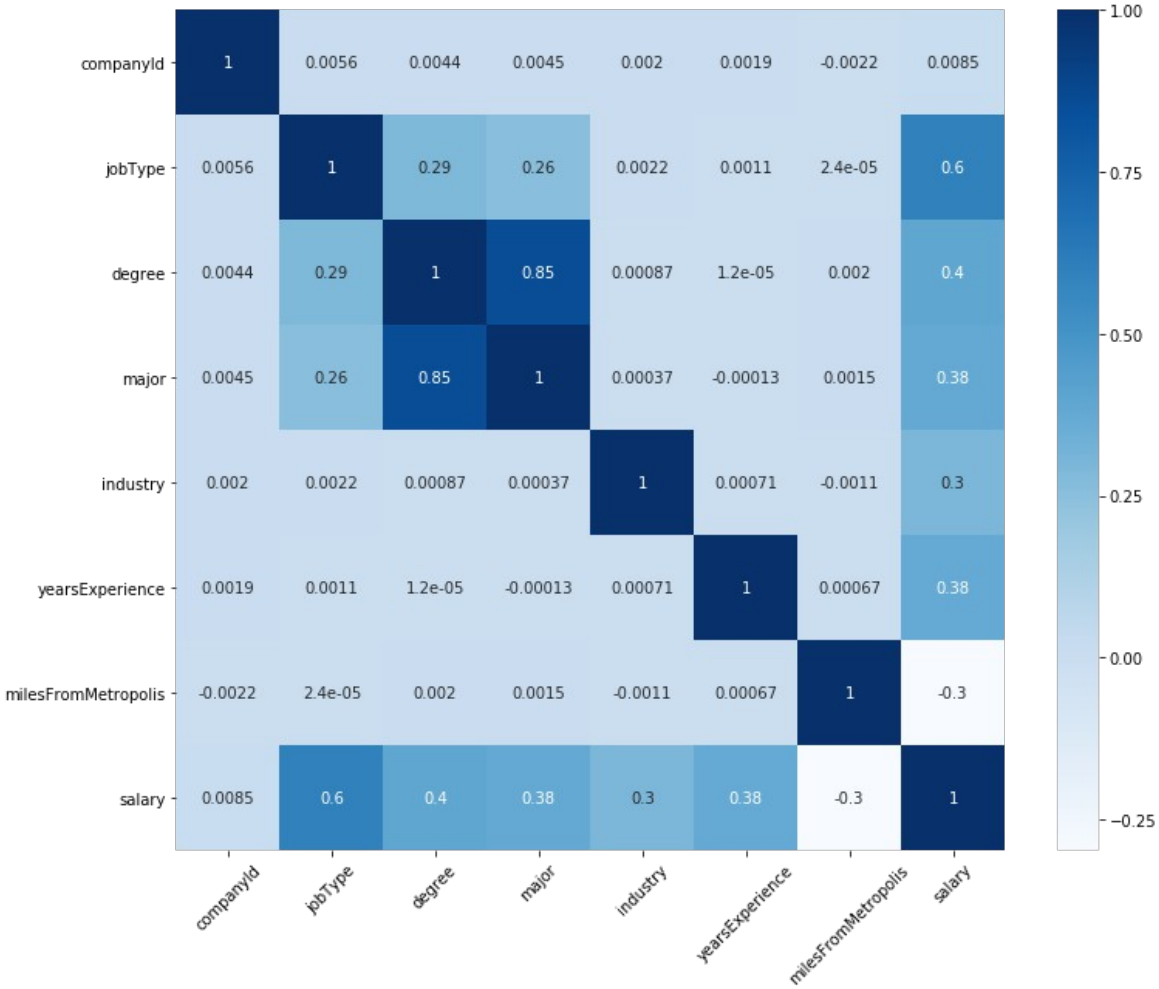
Degree - Salary is associated with degrees; higher degrees are paid more

Major - Business, Engineering and Math degree holders receive higher salaries

Industry - Oil and Finance industries are higher paid.

Years of Experience - There is a clear correlation with higher salaries given based on years of experience.

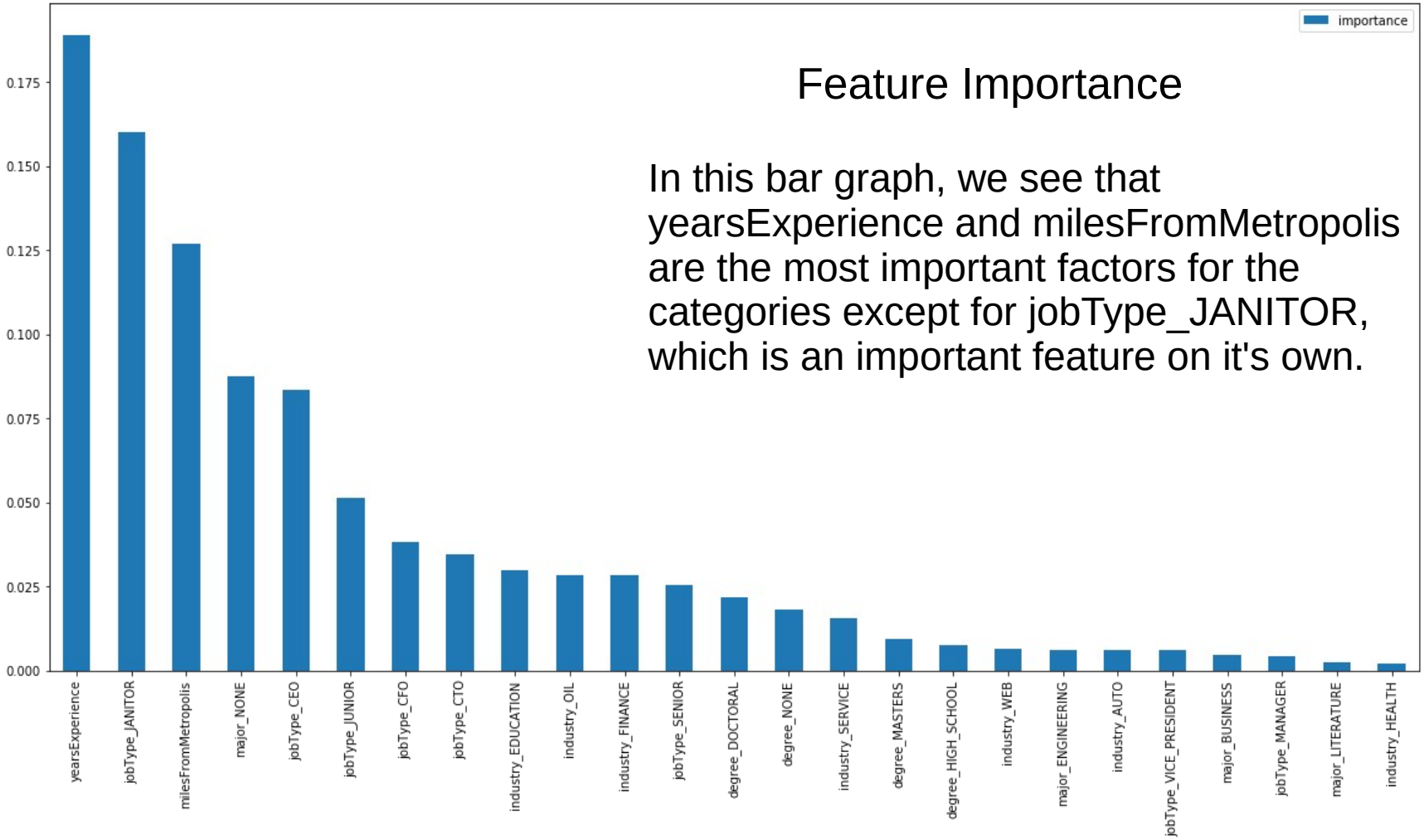
Miles from Metropolis - Generally, salaries decrease as the miles from Metropolis increases.



Correlation Heat Map

From this heatmap, there is shown a greater degree of correlation of jobType to salary, followed by degree, major, and yearsExperience. Using this analysis enables determination of col linearity, and so, determine if dimension reduction is needed. Also, there is some correlation with industry to salary, and a negative correlation between milesfromMetropolis to salary. This is because salaries are less in rural and suburban areas than metropolitan areas.

Overall, this shows that this is a clean dataset that can be modeled as is.



Feature Importance

In this bar graph, we see that yearsExperience and milesFromMetropolis are the most important factors for the categories except for jobType_JANITOR, which is an important feature on it's own.