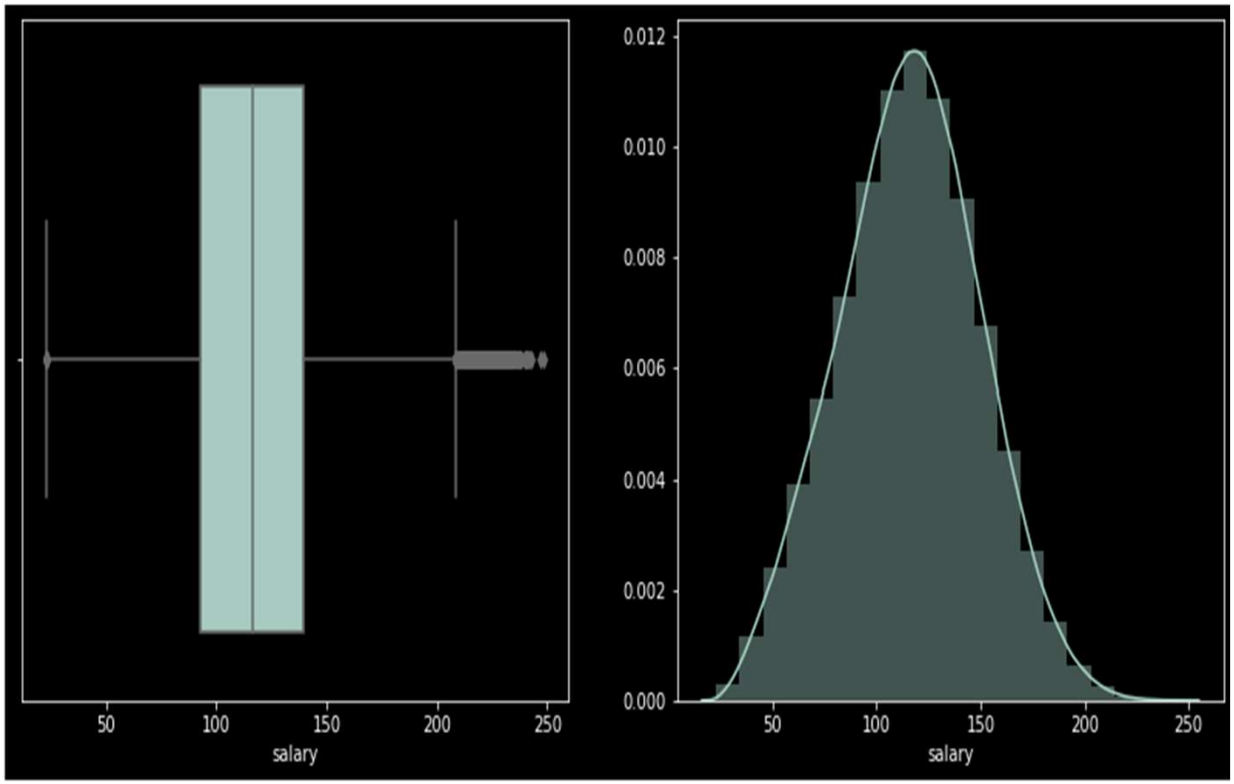
SALARY PREDICTION PROJECT

Laniya Oladapo

Train vs Test Salary Data Summary

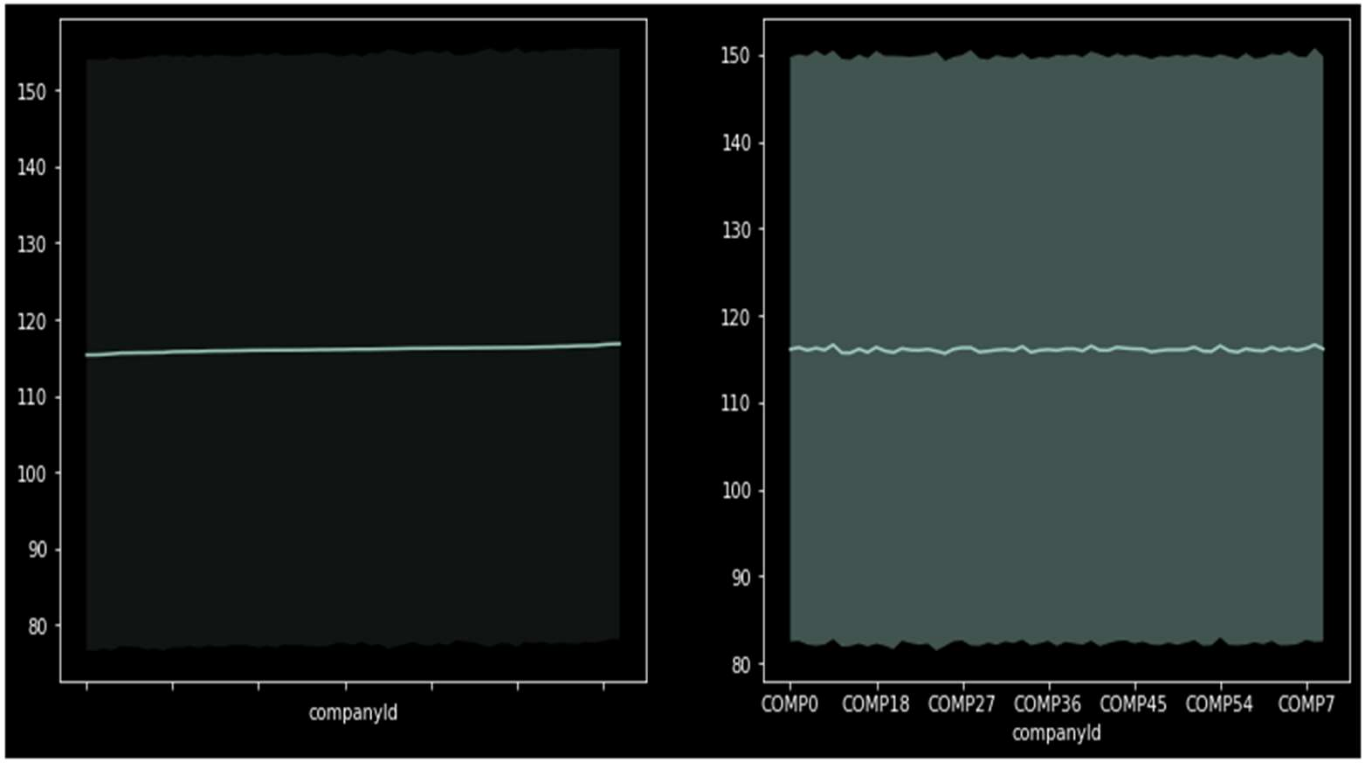
- Plots used to compare the data
  - Box Plot
  - Distribution Plot
  - Plots used to compare the data
- Training and Predicted Salary Data plotted against the respective features

Visualize Predicted Salary Data



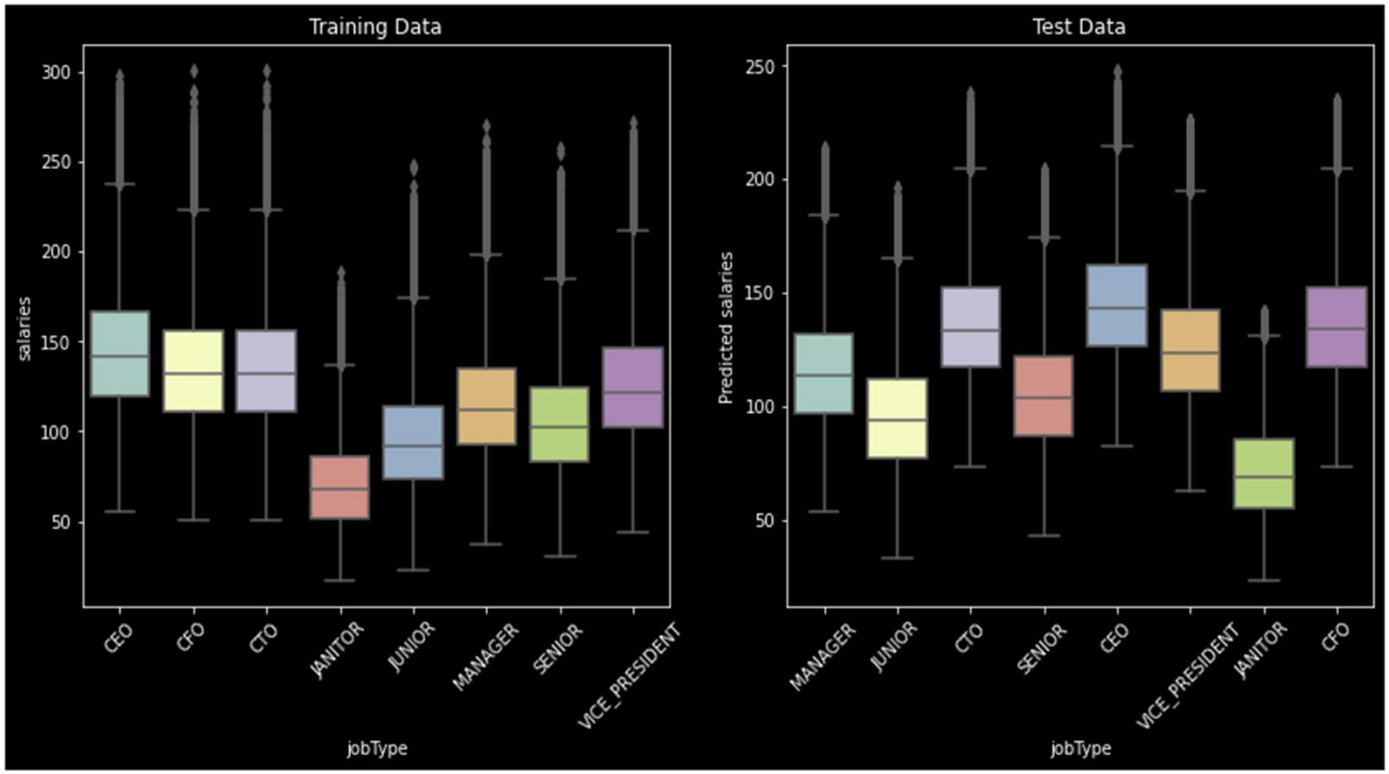
- Normally distributed data
- Upper bound outlier identified mostly employees with high years of service.

Company ID vs Salary Plot



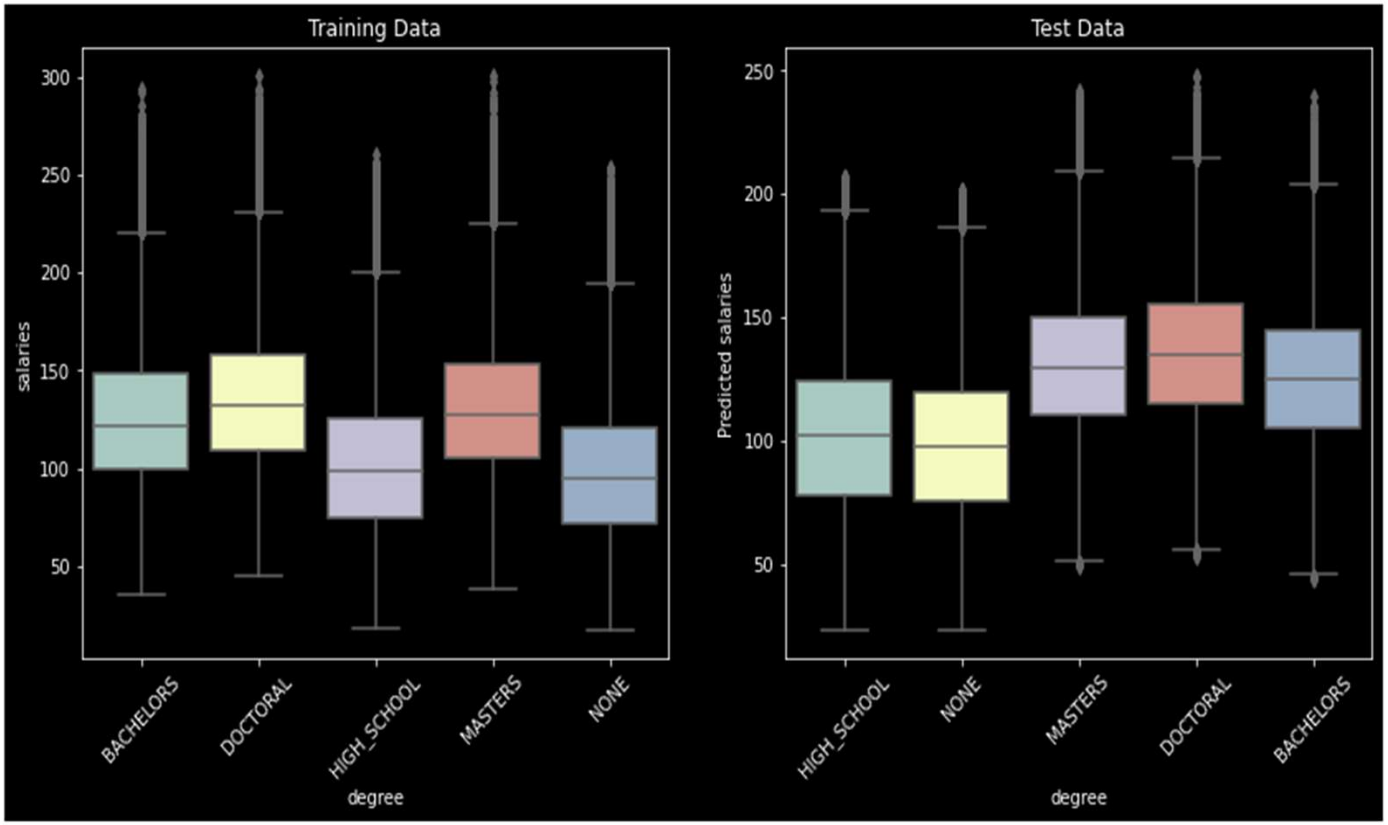
- This plot shows similarity of the generally weak correlation between company ID and salaries for both data sets.

Job Type vs Salary Plot



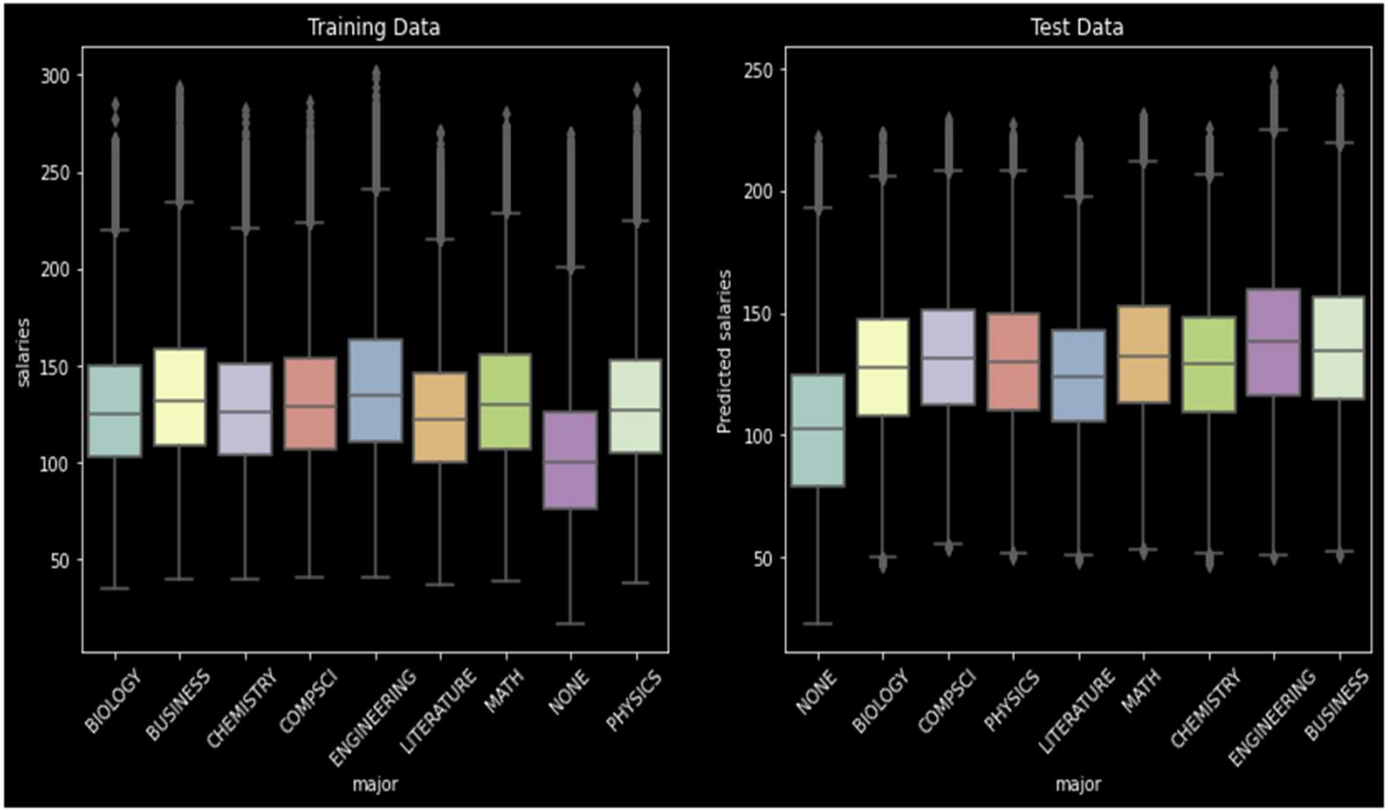
- Positive correlation between Job Type and Salary in both data sets
- Salary bracket increases relative to the level of seniority of roles in both data sets

Degree vs Salary Plot



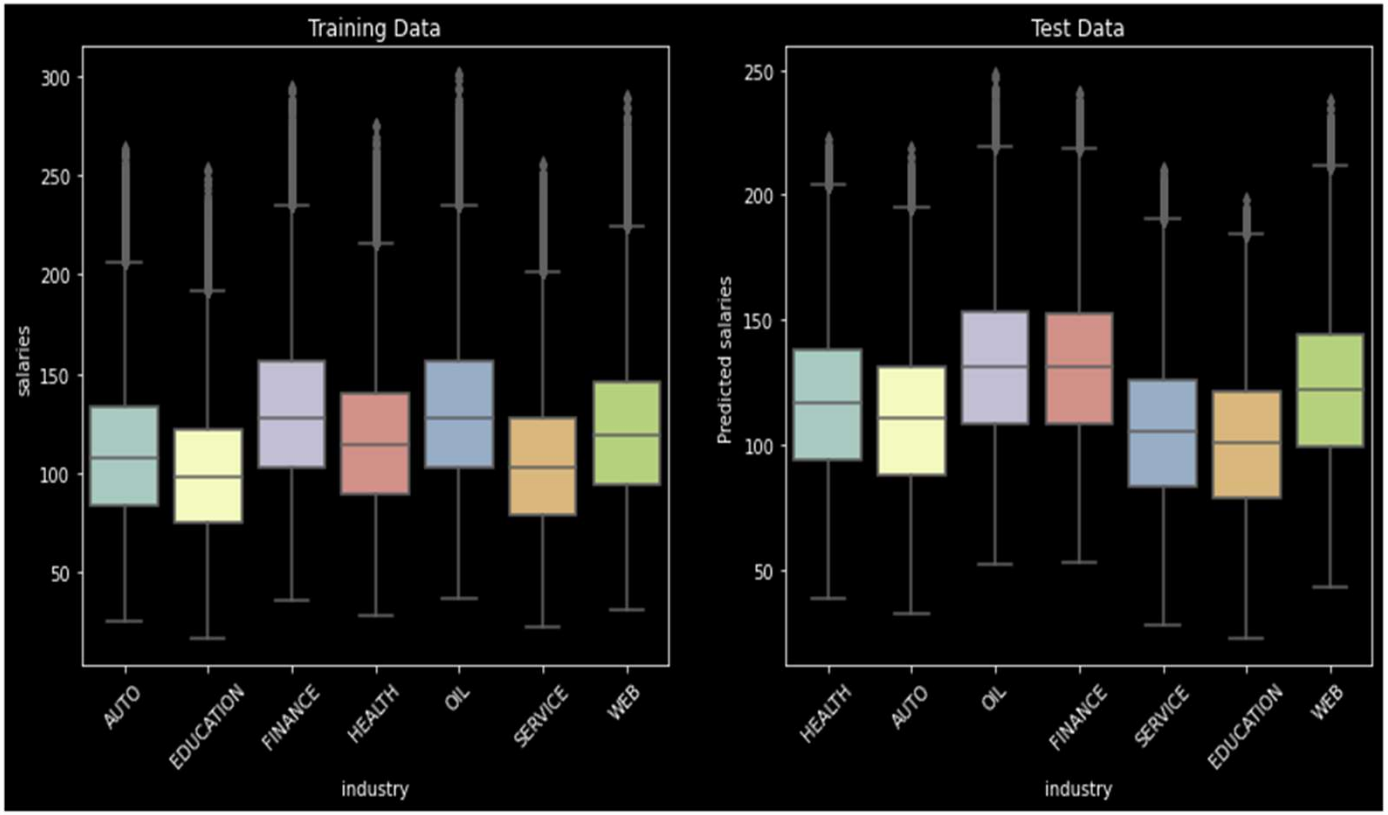
- Positive correlation between Degree and Salary in both data sets
- Salary bracket increases relative to the level of degree in both data sets

Major vs Salary Plot



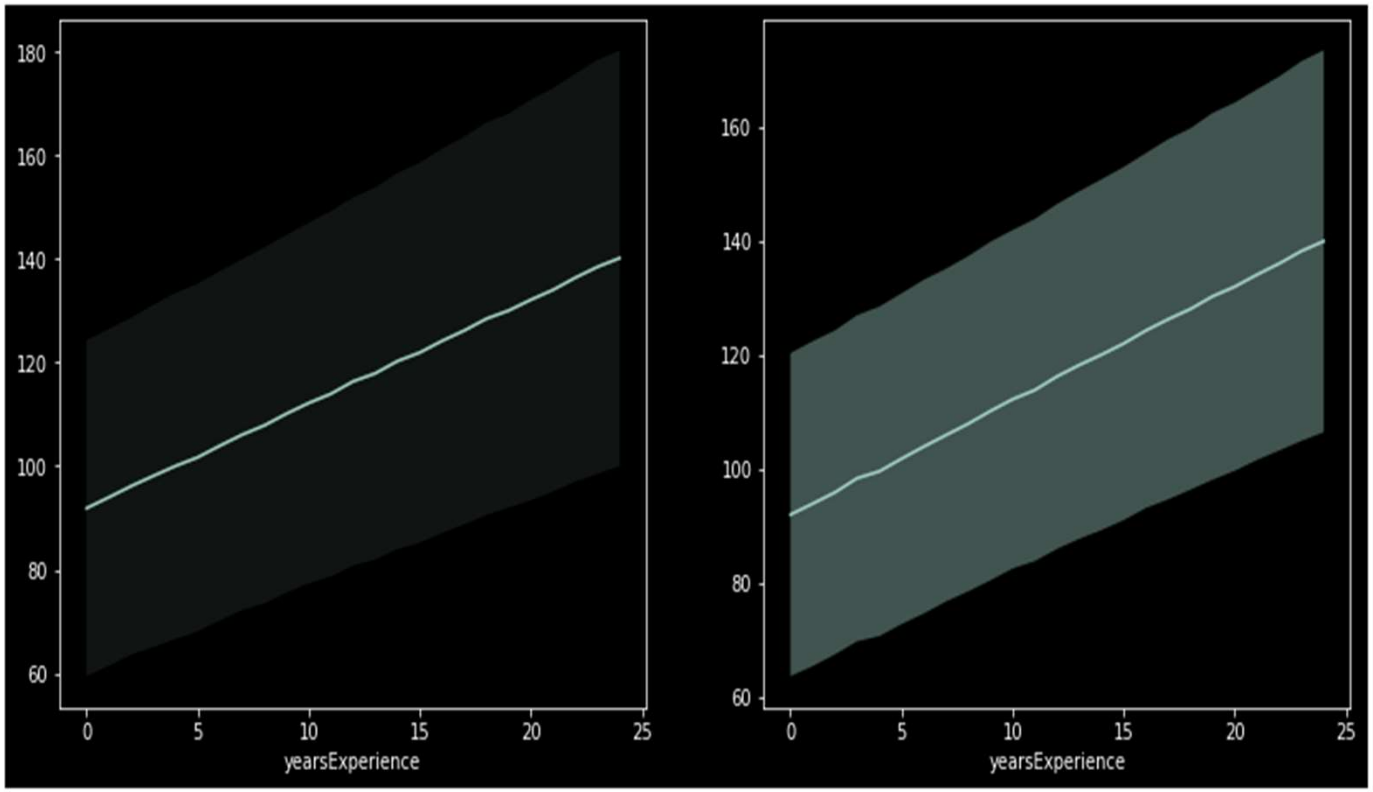
- Weak correlation between Major and Salary in both data sets
- None major group have the lowest salary range.

Industry vs Salary Plot



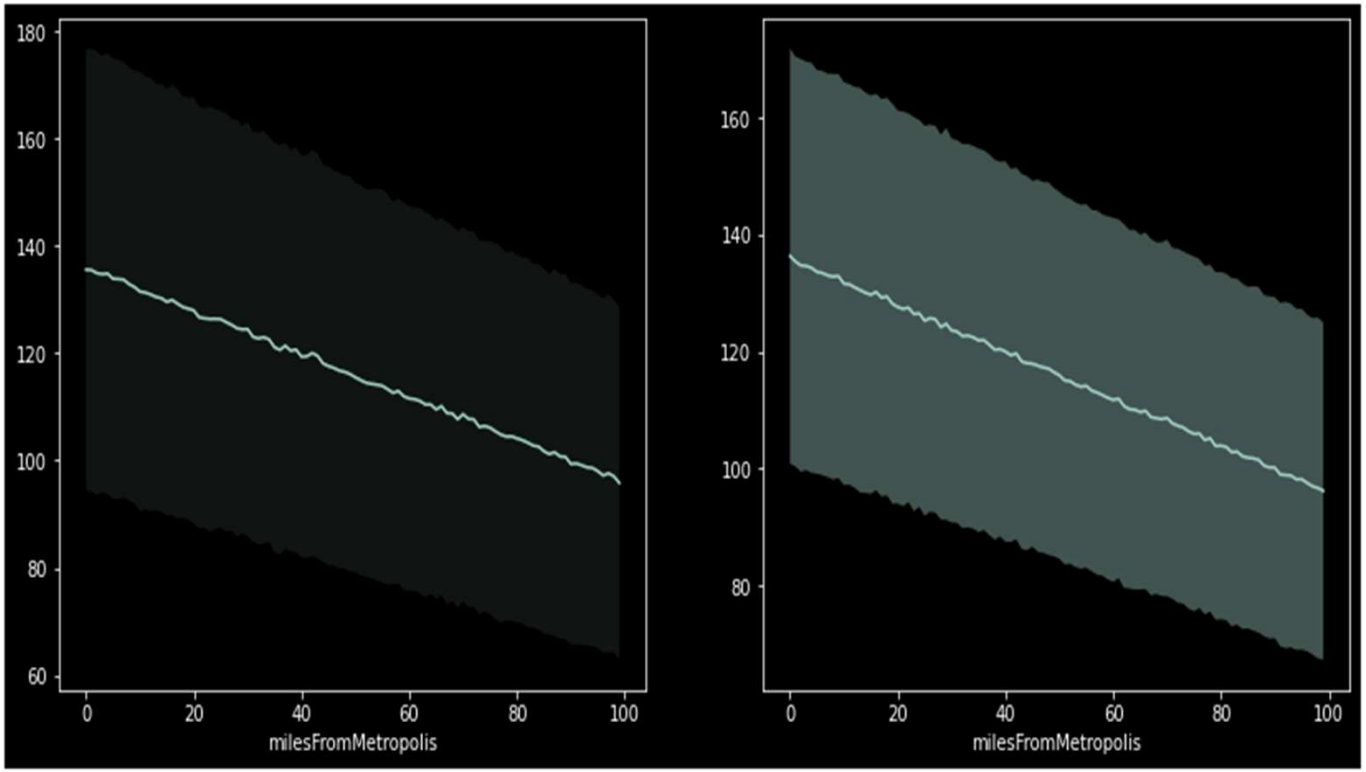
- Weak correlation between Industry and Salary in both data sets
- Oil and Finance industries have higher compensation band compared to the rest.

YearsExperience vs Salary Plot



- Strong positive correlation between Years of Experience and Salary in both data sets.
- Outliers in the data sets are mostly employees with high years of experience.

MilesFromMetropolis vs Salary Plot



- Strong negative correlation between Miles From Metropolis and Salary in both data sets.

Conclusions

- **JobType** is the most important feature for this prediction model. Perhaps gathering more granular data to further understand the salary ranges in each job type could be considered for future models.
- **MilesFromMetropolis** has a negative correlation with salary, is there something here the company might consider leveraging on to lower compensation cost in the P&L.
- Utilizing summary statistics should be considered as an improvement opportunity for the metrics.
- Multicollinearity is not much of an issue considering our main focus is on the salary prediction but the high correlation between **Major** and **Degree** should be investigated for any effect on the model.