Summary of Findings

To answer our proposed questions, we have found that our response salary_in_usd is in fact statistically associated with both the experience_level and remote_ratio values by looking at their coefficients. We found that employees with higher experience levels had higher salaries than those with less experience, which concurs with the findings in the study by Schlee and Karns with respect to the marketing industry. As such they will be suitable predictors to use in a predictive model.

Improvements

One improvement we could make to address the heteroscedasticity is to apply a transformation to our current model, such as creating a logistic regression or poisson regression. Another improvement that can be made is to gather more data, where a larger sample may be able to help us better rely on CLT. To better explain any covariates of the two used in our model (experience_level and remote_ratio), we could also include other variables which were not included in the original data set, such as level of education. To further improve the predictive power of our model, we could use LASSO to exclude any unnecessary variables and improve the accuracy of our predictions.

Future Plans

As the data used in this project was from 2020-2022, remote work in this time period was the preferred method of working due to the pandemic. As such it does not really give us a good idea on how remote work affects the salary of data science employees. To better investigate the effect of remote work we could compare the salaries of data scientists in another time period to the data set in this project. In regards to one of the improvements we can make, where we introduce another covariate of experience level like education level, we can make a model including education level to see if we can better explain the effect of experience level and create more accurate predictions.