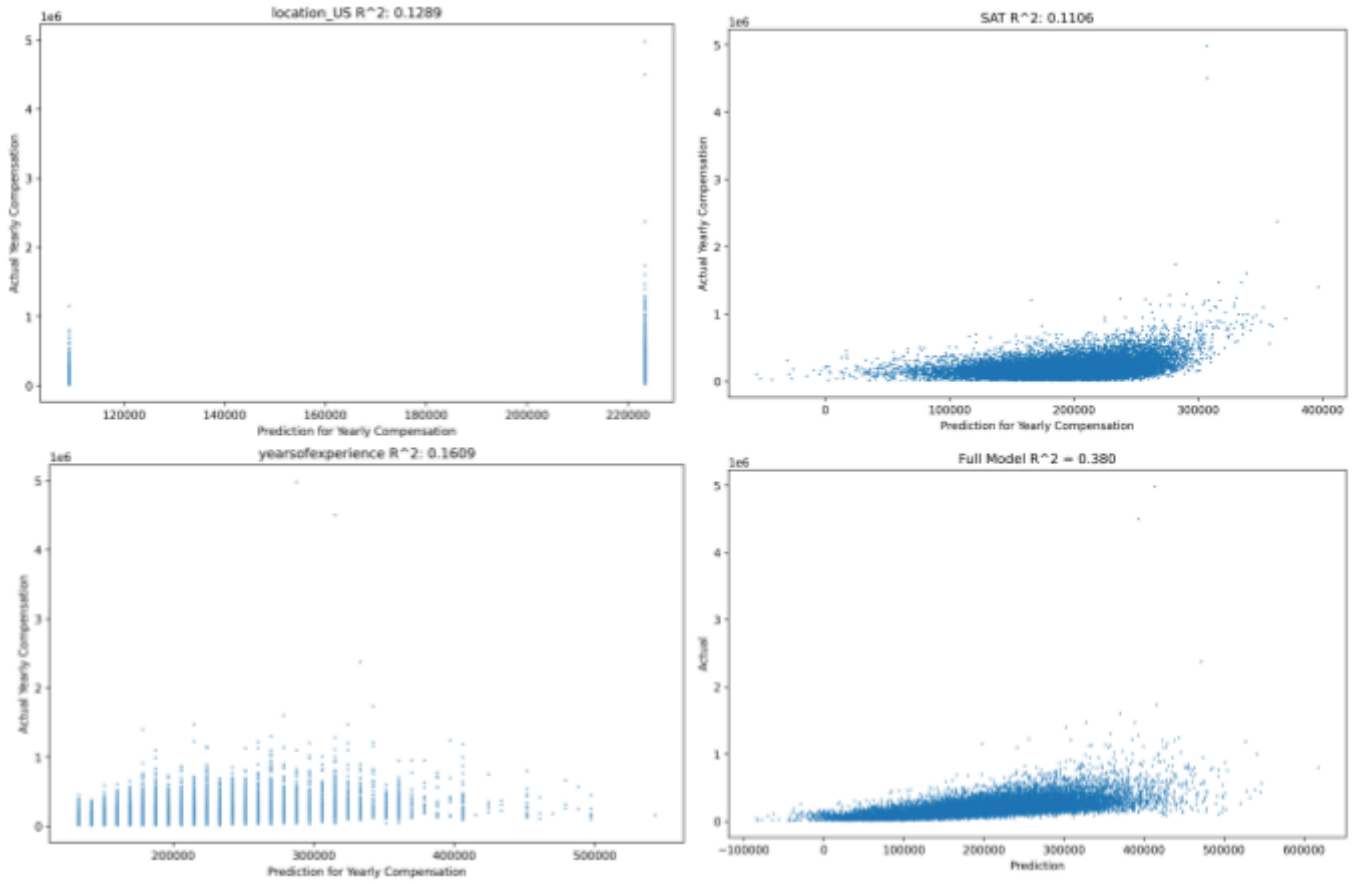
MULTIPLE LINEAR REGRESSION



To determine the best predictor of total annual compensation, I first built a full multiple linear regression model using all available predictors, excluding obvious features like total yearly compensation, base salary, stock grant value, and bonus as described by the homework specifications. Then, I systematically evaluated each individual predictor by fitting separate simple linear regression models and calculating their R^2 values, which measure the proportion of variance in yearly compensation explained by each predictor. The results I then visualized using scatter plots, comparing predicted vs. actual compensation.

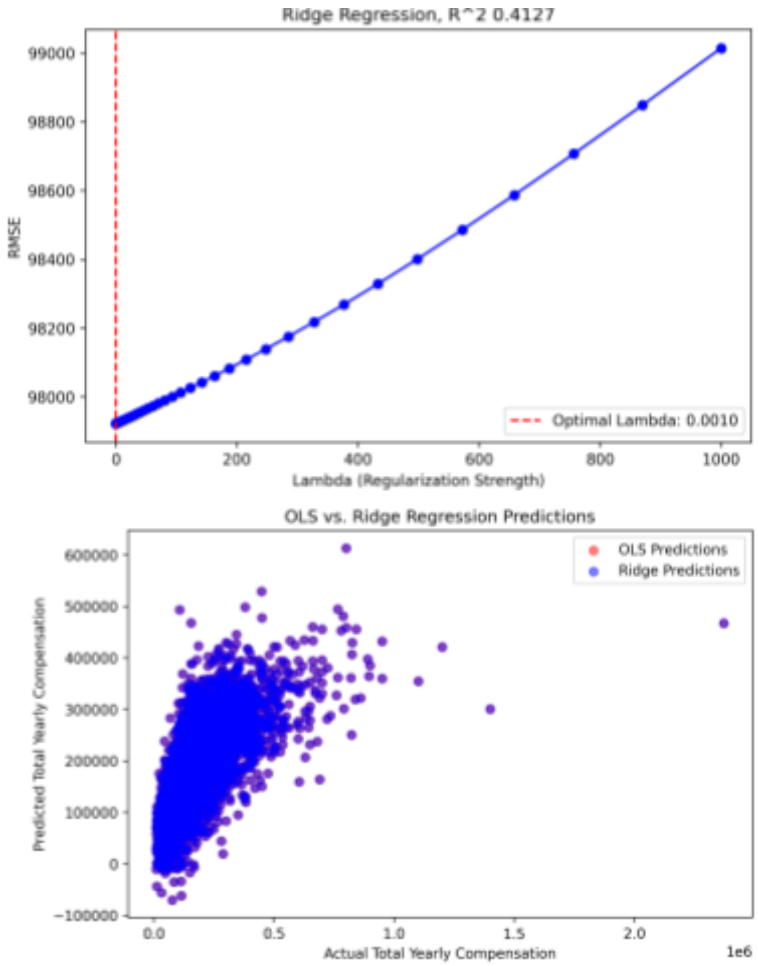
This approach was chosen because linear regression provides an interpretable measure of predictive strength through the R^2 statistic. By evaluating each predictor individually, I could directly assess its independent explanatory power. If I had used correlation coefficients instead, I might have missed the joint effect of multiple predictors.

The results show that "years of experience" was the best individual predictor of total yearly compensation, with an R^2 value of 0.1609. Other predictors, such as location ($R^2 = 0.1289$) and SAT score ($R^2 = 0.1106$), explained slightly less variance. However, the full multiple linear regression model performed significantly better, with an R^2 of 0.380, meaning it explains 38% of the variance in yearly

compensation. The scatter plots illustrate that while individual predictors exhibit a weak linear relationship with salary, combining multiple factors improves predictive power.

These findings suggest that while years of experience is the strongest single predictor, it alone does not fully explain salary variation. The full model's higher R^2 indicates that compensation is influenced by multiple factors simultaneously. This aligns with expectations in the tech industry, where salaries are affected by a combination of experience, location, education, and potentially unmeasured factors such as negotiation skills and company prestige. However, since even the full model only explains 38% of the variance, other complex, nonlinear influences likely play a significant role in determining salaries. This is apparent in the lack of linearity in the shape that the scatterplots take on.

RIDGE REGRESSION



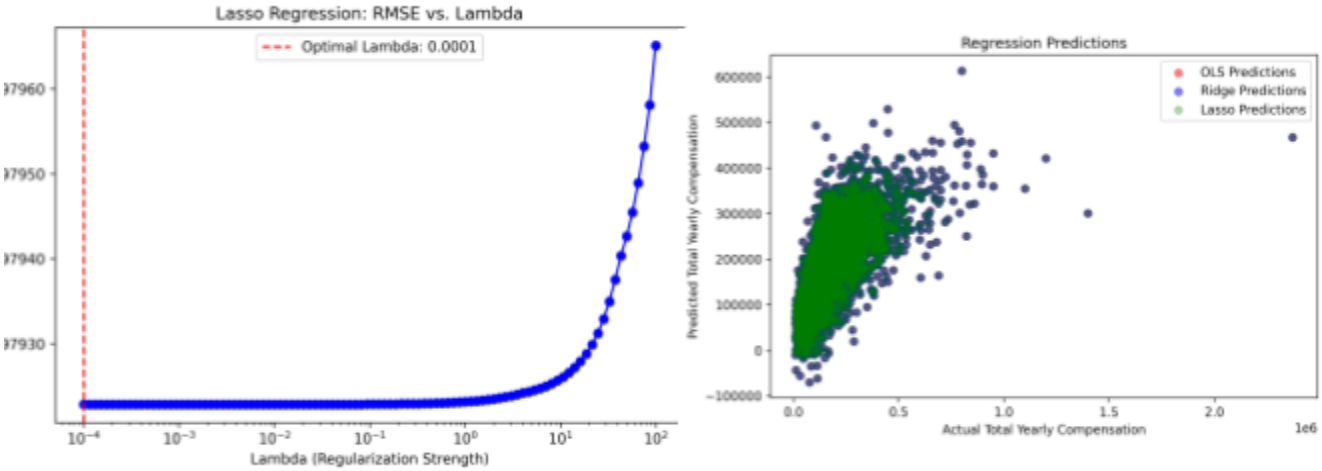
To assess the impact of Ridge Regression compared to Ordinary Least Squares (OLS), I trained both models on the same training and testing sets using a dataset of tech salaries. I retrained the linear regression model using the same split as Ridge Regression to make for a more sound comparison. I evaluated Ridge Regression across a range of lambda values, selecting the optimal lambda by minimizing Root Mean Squared Error (RMSE). I then compared Ridge and OLS using R^2 and RMSE, visualizing their predictive performance.

This approach was chosen because Ridge Regression applies L2 regularization, which helps address multicollinearity and prevents overfitting. By tuning lambda, I aimed to balance bias and variance, improving generalizability compared to OLS. Since OLS does not include a regularization term, it is more susceptible to large coefficient estimates when predictors are highly correlated. Evaluating both models using R^2 and RMSE ensures I capture differences in both explained variance and predictive accuracy.

Interestingly, the results showed that Ridge Regression and OLS performed identically, both achieving an R^2 of 0.4127 and an RMSE of 97,922.87. The optimal lambda was 0.0010, meaning Ridge applied only a very small penalty to the coefficients. The first plot shows how RMSE increases with higher lambda values, confirming that excessive regularization worsens predictions. The second plot illustrates that Ridge and OLS predictions are almost indistinguishable.

The reason Ridge and OLS yielded the same results is likely due to two main factors: (1) Minimal multicollinearity in the dataset, meaning the predictors are not strongly correlated, so Ridge's regularization had little effect on stabilizing coefficients. (2) A very small optimal lambda (0.0010), which applied only a negligible penalty, making Ridge's coefficient estimates almost identical to OLS. Since Ridge's primary advantage is reducing variance in highly correlated data, and the dataset does not exhibit strong collinearity, both models effectively performed the same. These findings suggest that in this case, OLS was sufficient, as Ridge Regression did not provide any meaningful improvement.

LASSO REGRESSION



To analyze how the model changes with the use of Lasso regression, I trained Lasso models across a range of lambda values and selected the optimal lambda based on the lowest Root Mean Squared Error (RMSE). I also examined how many predictor coefficients were shrunk to exactly zero, a key feature of Lasso regression. The first plot visualizes RMSE as a function of lambda, with the optimal lambda marked.

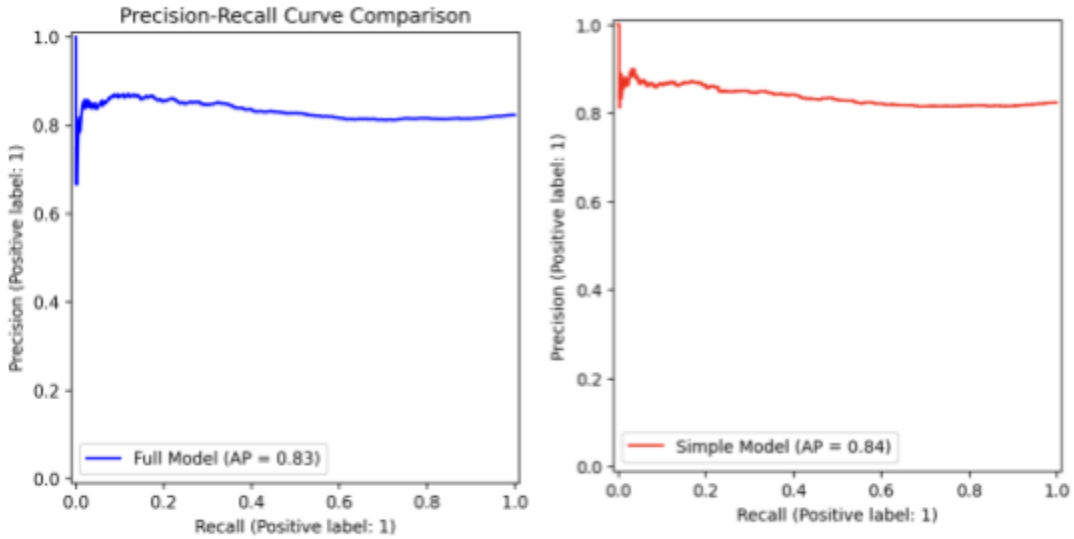
Lasso regression helps identify the most relevant predictors while potentially improving generalization. Unlike Ridge regression, which shrinks coefficients but retains all predictors, Lasso can

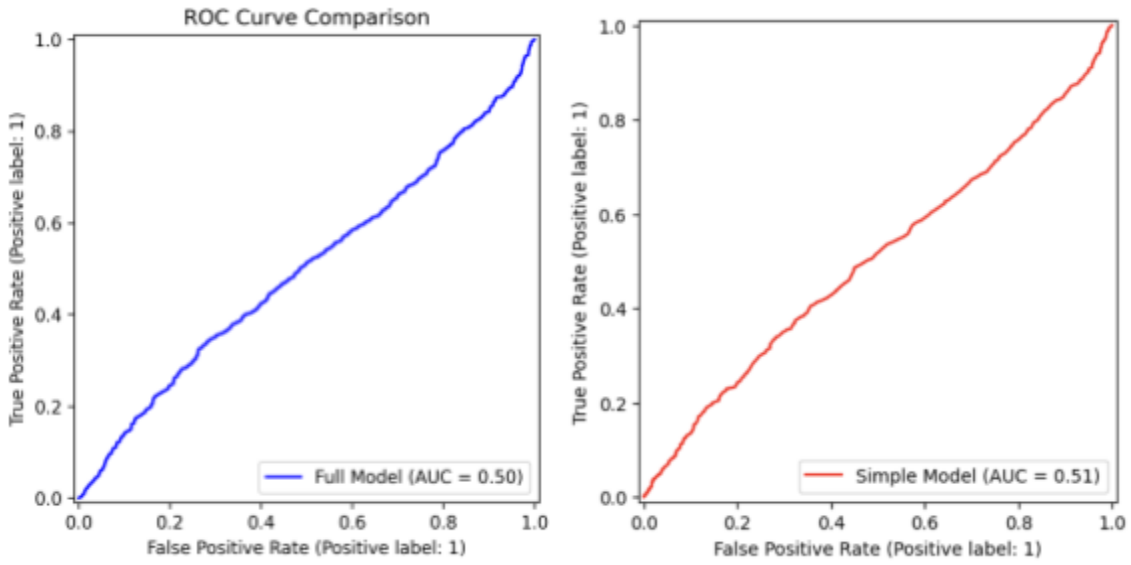
simplify the model by removing less important variables. By plotting RMSE versus lambda, I aimed to balance predictive accuracy and model sparsity.

The results showed that the optimal lambda for Lasso was 0.0001, yet none of the predictors were shrunk to zero. This is visible in the second plot, where Lasso predictions (green) completely overlap with Ridge (blue) and OLS (red) predictions, indicating that Lasso did not simplify the model by removing features. Additionally, Lasso performed identically to Ridge and OLS, achieving an R^2 of 0.4127 and an RMSE of 97,922.8674.

This suggests that Lasso's feature selection capability was not utilized because the dataset does not contain strongly redundant or irrelevant predictors. The small optimal lambda value (0.0001) imposed only a minimal penalty, meaning coefficients were barely affected. In this case, all three models (OLS, Ridge, and Lasso) performed the same, indicating that regularization did not significantly improve predictive power. This suggests that the dataset may not suffer from multicollinearity, and OLS was already performing optimally without needing additional constraints.

GENDER PAY GAP IN TECH COMPENSATION – LOGISTIC REGRESSION





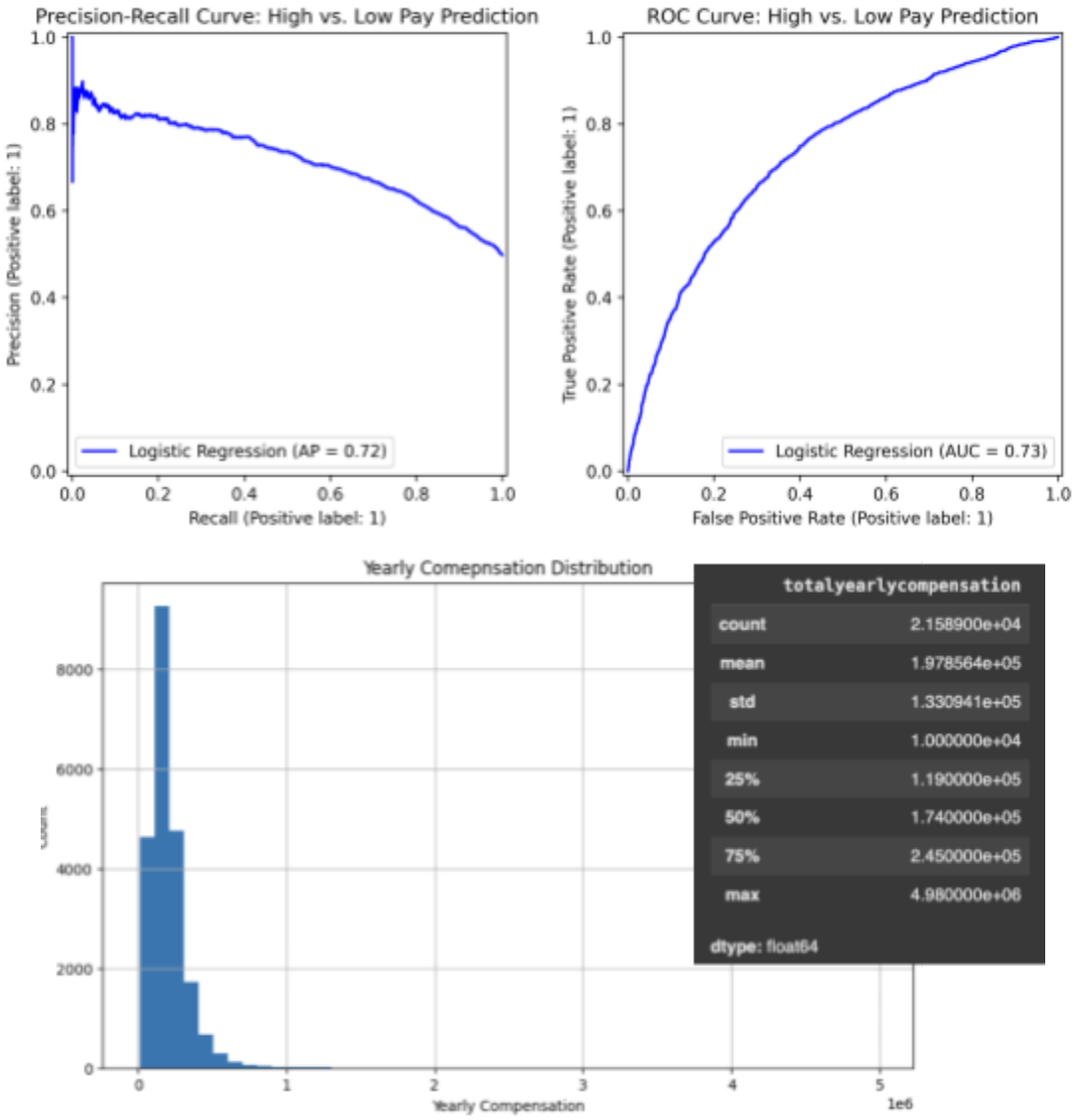
To answer the question of whether the data reflects a male/female gender pay gap, I fit a simple model using only total yearly compensation as a predictor. Then, I fit a full model, incorporating additional factors to control for potential confounders. Both models were evaluated using accuracy, precision, recall, and confusion matrices. Additionally, I plotted ROC and Precision-Recall (PR) curves to compare classification performance.

By comparing a model with only compensation to a model that controls for other factors, I knew I would be able to assess whether total yearly compensation alone predicts gender or if other factors account for differences. I examined model performance through accuracy, precision, recall, and ROC/PR curves to ensure robustness.

The results indicate that both the simple and full models performed identically, with an accuracy of 82.3%, precision of 82.3%, and recall of 100.0%. The confusion matrices show that the models classified all males correctly, but failed to classify any females correctly (zero true negatives). The ROC and PR curves demonstrate similar classification performance betIen the two models, indicating that controlling for additional factors did not significantly alter predictions. This suggests that total yearly compensation does not meaningfully separate genders in this dataset.

These findings suggest that, in this dataset, logistic regression does not provide evidence of a strong gender-based compensation difference. The inability of the models to classify females correctly suggests that other factors beyond salary are driving the classification outcome. This could indicate data imbalance, omitted variables, or systemic biases affecting salary reporting. While the debate on the gender pay gap in tech is complex, these results imply that, based on this dataset alone, total yearly compensation alone does not serve as a strong predictor of gender.

PREDICTING HIGH/LOW PAY



To determine whether years of relevant experience, age, height, SAT score, and GPA can predict high vs. low pay, I built a logistic regression model. First, I defined “high pay” and “low pay” using the median total yearly compensation ($174,000). Individuals earning above the median Ire classified as high pay (1), while those earning below Ire labeled low pay (0). The predictors were standardized, and the dataset was split into training (80%) and validation (20%) sets. A logistic regression model was trained, and performance was evaluated using accuracy, precision, confusion matrices, and ROC/PR curves.

The median salary was chosen as the threshold because it evenly splits the dataset into two balanced groups, ensuring that the classification task is not biased toward one category. As shown in the salary distribution plot (Figure 1), yearly compensation is highly right-skeId, with a small number of extreme salaries pulling the mean ($197,856) higher than the median ($174,000). If I had used the mean salary as a cutoff, I would have created an imbalanced dataset where feIr individuals fall into the “high

pay” category. The 25th and 75th percentiles ($119,000 and $245,000, respectively) show that most salaries are concentrated below $250,000, reinforcing why the median provides a fair split.

The model achieved an accuracy of 67.6% and a precision of 69.6%. The confusion matrix shoId that 1,573 low-pay individuals were correctly classified, but 582 were misclassified as high-pay. Similarly, 1,331 high-pay individuals were correctly classified, while 811 were misclassified as low-pay. The ROC curve resulted in an AUC of 0.73, indicating moderate predictive performance, while the Precision-Recall curve showed an average precision (AP) of 0.72, suggesting that the model balances precision and recall but struggles with clear classification.

These findings suggest that years of experience, age, height, SAT score, and GPA alone provide only moderate predictive poIr for salary classification. The ROC AUC of 0.73 indicates that the model performs better than random guessing but is not highly reliable. The significant misclassification rate suggests that salary is likely influenced by additional factors such as job title, company, location, and industry trends, which were not included in the model. The choice of the median as a cutoff ensured a balanced classification, but future models incorporating more economic and career-related variables could improve prediction accuracy.