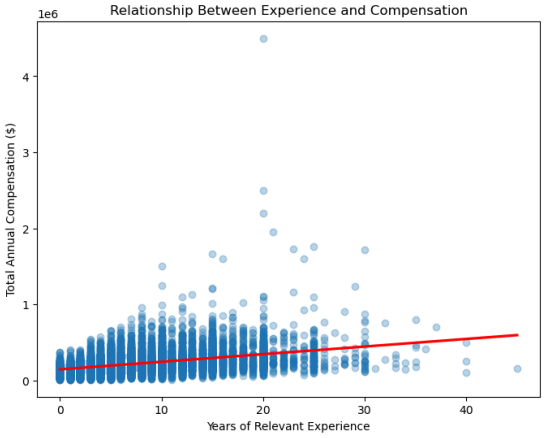
1. Using multiple linear regression: What is the best predictor of total annual compensation, how much variance is explained by this predictor vs. the full multiple regression model?

Methodology:
To determine the best predictor of total annual compensation, I ran a multiple linear regression model. The predictors included years of experience, tenure at the company, education levels, race, age, height, zodiac sign, SAT scores, and GPA. I cleaned the dataset by removing missing values, and an 80-20 train-test split was used for model evaluation. The best single predictor was identified by iteratively fitting simple linear regression models and comparing their scores.

Rationale:
I used multiple linear regression because it quantifies the effect of each predictor while controlling for others. This approach ensures that I can identify the best single predictor while also evaluating the performance of the full model.

Results:
I found years of relevant experience to be the best single predictor of total annual compensation. It explains 17.66% of the variance alone (R^2 = 0.1766). In contrast, the full multiple linear regression model, incorporating all predictors, explains 28.71% (R^2 =0.2871) of the variance. This indicates that additional variables contribute meaningfully to salary prediction, but years of experience is the most influential individual factor.



Interpretation:
Years of relevant experience is the strongest individual predictor of salary, which aligns with expectations in the tech industry, where compensation often scales with experience. However, the full model's higher R^2 suggests that other factors, such as education and company tenure, also play a role in determining salaries. While years of experience is a key determinant, a more comprehensive model is necessary for better salary prediction.

2. Using ridge regression to do the same as in 1): How does the model change or improve compared to OLS? What is the optimal lambda?

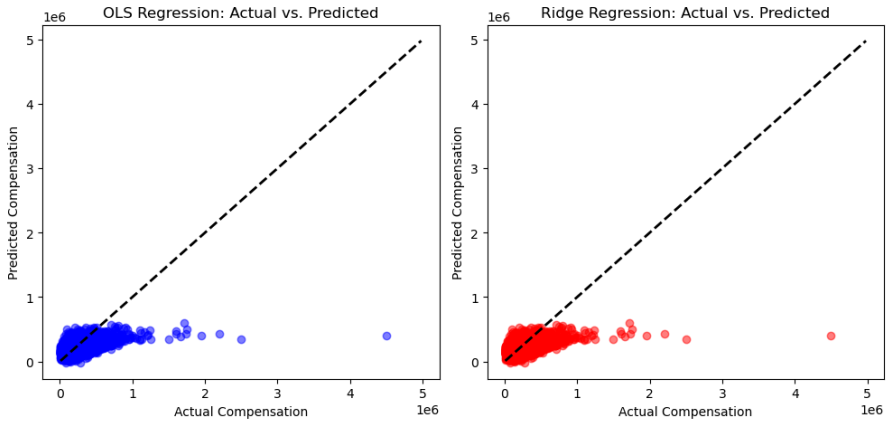
Methodology:
I ran a ridge regression model to compare its performance against ordinary least squares (OLS) regression. Using cross-validation, the optimal regularization parameter (lambda/alpha) was selected from a range of values. I then trained the ridge model with this optimal lambda and evaluated using the score.

Rational:
Ridge regression was chosen to address potential overfitting in the OLS model by introducing a penalty on large coefficients. This is possible because ridge regression shrinks betas towards 0 (but doesn't actually reach 0), which could potentially improve predictive accuracy. Additionally, cross-validation ensured that the optimal lambda value was selected to balance bias and variance.

Results:

The R^2 values for the OLS and Ridge models were identical: 0.287. This indicates that the ridge regression model did not improve the explained variance over OLS.

Interpretation:
Since ridge regression did not increase R^2, this suggests that the original OLS model was not suffering from severe overfitting. The optimal lambda of 15.199 applied some shrinkage to the coefficients, but it did not meaningfully change predictive performance. While ridge regression is useful for improving generalization, in this case, it did not provide a tangible advantage over OLS. This is demonstrated in the identical depictions of actual compensation versus predicted compensation for each model below.

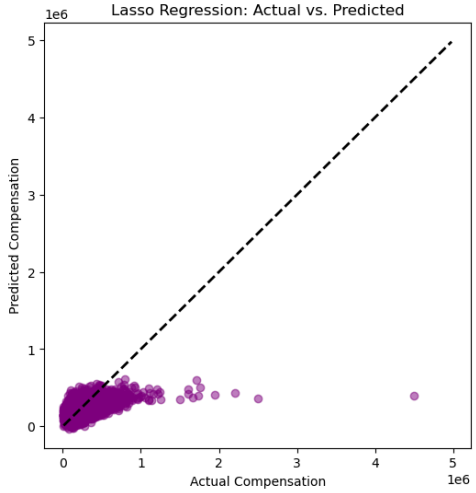


3. Using Lasso regression to do the same as in 1): How does the model change now? How many of the predictor betas are shrunk to exactly 0? What is the optimal lambda now?

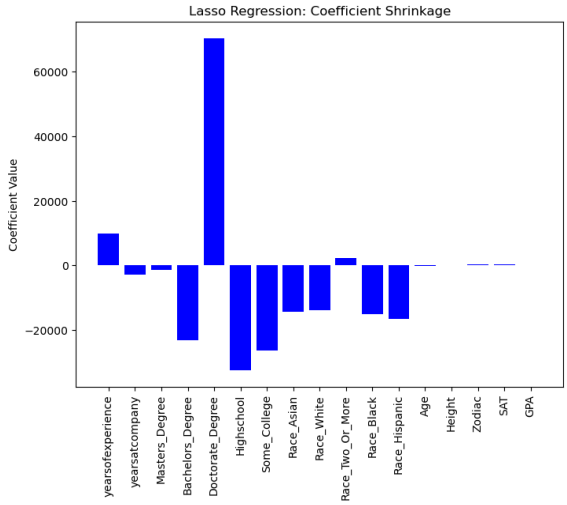
Methodology**:**
Utilizing lasso regression, I determined the optimal lambda value for the model and used this to fit the model. The model's performance was evaluated with R^2, and the number of coefficients shrunk to zero was counted. A coefficient shrinkage plot visualized feature selection, and a scatterplot compared actual vs. predicted compensation.

Rationale:
Lasso regression was chosen because it allows some betas to actually shrink to 0, helping identify the most relevant predictors. Cross-validation ensured an optimal balance between bias and variance.

Results:
The optimal lambda value was 10.723, resulting in an R^2 of 0.287, indicating that the model explains about 28.71% of the variance. Only one coefficient was shrunk to zero, meaning most predictors contributed to the model. The scatterplot suggests a wide spread in prediction errors, particularly for higher salaries.



Interpretation:
Lasso regression did not eliminate many predictors, implying that most features have some relevance to compensation. Additionally, since the R^2 score of this model was identical to those of the previous models, Lasso did not provide a tangible advantage over OLS or even Ridge regression. You can see how the plot of actual compensation to predicted compensation is identical to those in the previous question, as well as how Lasso did not shrink most of the predictors in the figures below.



4. There is controversy as to the existence of a male/female gender pay gap in tech job compensation. Build a logistic regression model (with gender as the outcome variable) to see if there is an appreciable beta associated with total annual compensation with and without controlling for other factors.

Methodology:

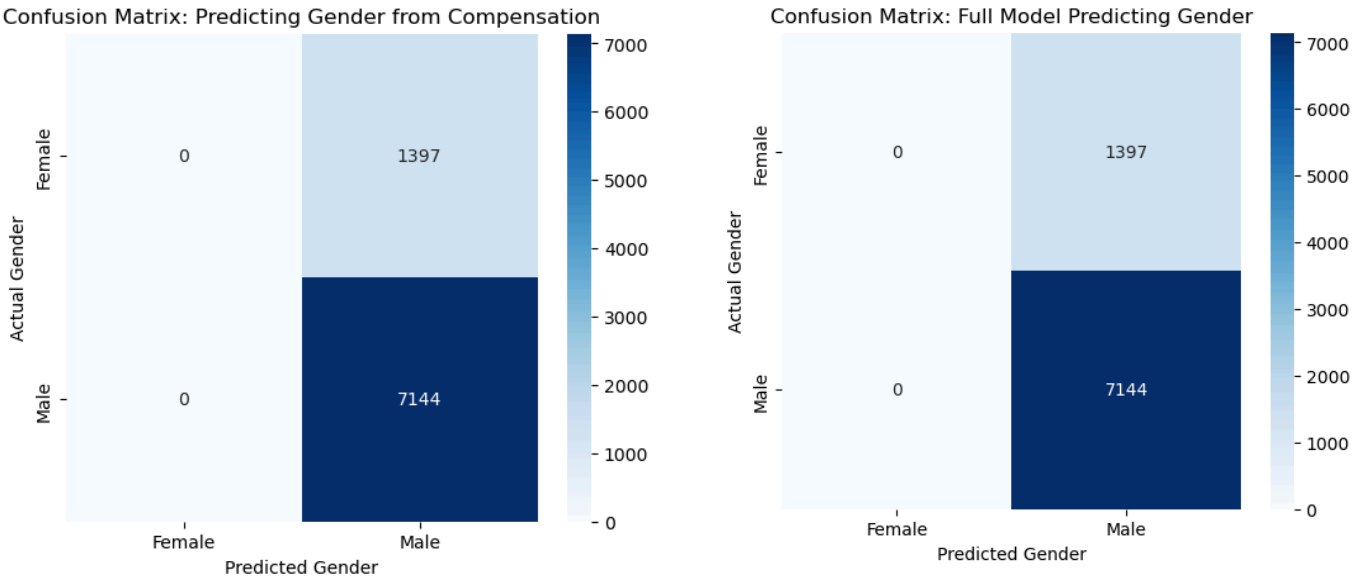
I built two logistic regression models to determine whether total compensation is a significant predictor of gender in tech job compensation. The first model used only total compensation as a predictor, while the second model included all available factors. I applied L1 regularization to each model to encourage meaningful coefficient selection, and total compensation was standardized to ensure proper scaling.

Rationale:
Logistic regression is an appropriate method for binary classification problems like this one. Extracting the beta coefficient allows us to determine whether compensation meaningfully influences gender classification. Including additional variables in the second model helps

identify whether compensation remains a significant predictor alone after controlling for other factors.

Results:
The betas for total compensation in both models was 0.00, indicating that compensation had no direct impact on gender classification. The accuracy for both models remained at 83.64%, and the confusion matrices showed that the model classified all individuals as male, failing to predict any female respondents correctly.

Interpretation:
I concluded that total compensation alone does not have a significant, independent effect on predicting gender within this dataset. The model's complete inability to classify female respondents correctly implies that salary distributions are highly imbalanced, potentially reflecting gender disparities in pay. Controlling for additional factors did not change the outcome, reinforcing that salary alone is not a sufficient determinant of gender classification in this dataset.



5. Build a logistic regression model to see if you can predict high and low pay from years of relevant experience, age, height, SAT score and GPA, respectively.

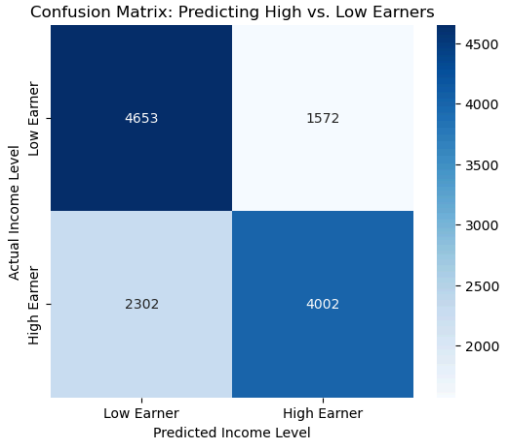
Methodology:

I built a logistic regression model to classify individuals as high or low earners based on years of experience, age, height, SAT score, and GPA. The model's performance was evaluated using accuracy, sensitivity (recall), specificity, and precision. A confusion matrix was generated to analyze classification errors.

Rationale:
Logistic regression is appropriate for binary classification tasks. Beyond accuracy, sensitivity and specificity provide deeper insights into how well the model distinguishes between high and low earners. Precision helps assess how reliable high earner predictions are.

Results:
Accuracy: 69.08% - 69.08% of all predictions were correct
Sensitivity (Recall): 63.48% – 63.48% of actual high earners were correctly identified.
Specificity: 74.75% – 74.75% of actual low earners were correctly classified.
Precision: 71.80% – 71.80% of predicted high earners were actually high earners.
The confusion matrix below shows the actual number of correct/incorrect predictions per class.
The most influential predictor was years of experience (β = 0.1470), while age had almost no effect (β = -0.0006). SAT scores (β = 0.0040), GPA (β = 0.0454), and height (β = 0.0024) had small but positive effects.

Interpretation:
I found that the model is more effective at identifying low earners than high earners, as indicated by the higher specificity. Precision is fairly strong, meaning that when the model does predict a high earner, it is often correct. Years of experience is the strongest predictor of being a high earner, while age is almost irrelevant, likely because experience better captures career progression. SAT and GPA show minor positive correlations, suggesting a weak link between academic performance and salary.



Extra Credit A: Is salary, height or age normally distributed? Does this surprise you? Why or why not?

Methodology:

I conducted a Shapiro-Wilk test to assess whether salary, height, and age are normally distributed. I also generated histograms with KDE plots to visualize the distributions. I used a significance level of 0.05 to determine normality.

Rationale:
The Shapiro-Wilk test is a standard method for testing normality in small to moderately large datasets. The histogram and KDE plots help visually confirm the results.

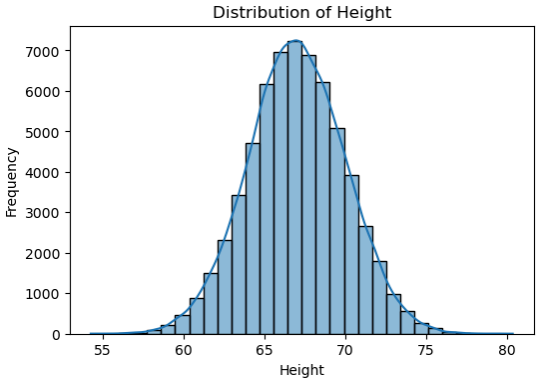
Results:

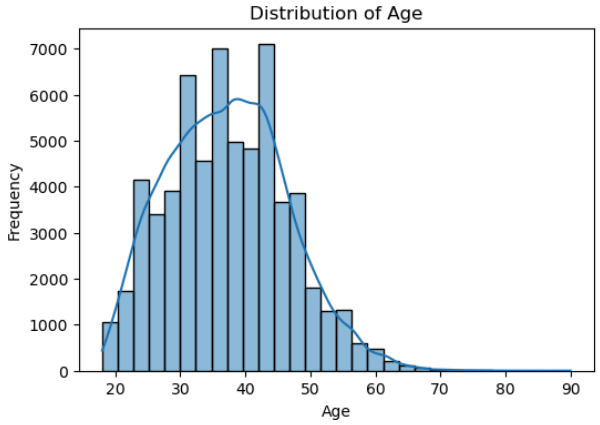
Salary: Not normally distributed ($W = 0.7916$, $p < 0.0001$). The distribution is highly skewed to the right, with most salaries concentrated at lower values and a long tail extending to very high salaries.

Height: Appears normally distributed (W = 1.0000, p = 0.5426). The histogram confirms a bell-shaped curve, indicating that height follows an approximately normal distribution.

Age: Not normally distributed (W = 0.9887, p < 0.0001). The distribution is slightly right-skewed, with more younger workers and a gradual decline in frequency as age increases.

Interpretation:
I expected the non-normal distribution of salary, as compensation often follows a skewed pattern where a few individuals earn significantly more than the majority. Height's normal distribution aligns with biological expectations, as human height tends to follow a Gaussian distribution. Age's slight skewness suggests that the tech industry might have a younger workforce, which is consistent with industry trends favoring early-career professionals.





Extra Credit B: Tell us something interesting about this dataset that is not already covered by the questions above and that is not obvious.

Methodology:

I generated 2 violin plots for job titles and their total yearly compensation to see any interesting trends in the distribution. One plot includes outliers while the other plot excludes outliers.

Rationale:

The violin plots visually highlight how compensation varies across roles. Including outliers allows us to see the highest earning potential for each job, while excluding outliers allows us to make easier interquartile comparisons between total yearly compensation across the different jobs.

Results:
The violin plot reveals significant salary differences across job titles. Roles like Product Manager and Software Engineering Manager tend to have the highest median salaries, while others exhibit lower median salaries and a tighter distribution, such as Business Analyst & Recruiter. Some job titles show a few extreme outliers with exceptionally high compensation, which skews the distribution.

Interpretation:
The presence of extreme outliers suggests that within high-paying roles (Product Manager, Software Engineer, Software Engineering Manager), a subset of individuals (possibly executives or employees at top-tier firms) earn disproportionately more.