Question 1

1. First, I dropped the NaN values in the dataframe to make it easier to use and made gender a binary variable so it's usable. I then fit a regression model with the predicor variables I wanted to use as inputs (all quantitative variables with the exception of "Some_College" and "Race_Two_Or_More" as including them after dropping the NaN values results in an overdetermined model and "basesalary", "stockgrantvalue", and "bonus" as you can predict total yearly compensation from a linear combination of these variables) and total yearly compensation as the output. I also calculated the r-squared value for the model. I then fit regression models with each of the predictor variables indivudall as the inputs and total yearly compensation as the output (for education and race I put all the binary dummy variables made for them into one regression model as inputs). I then calculated the r-squared value for each of the models. To visualize the models, I plotted the predicted values against the actual values on scatter plots for the regression model with all the predictor variables and the regression model with the highest r-squared among the individual predictor variables.
2. I did this to see how well all of the predictor variables together did at accurately predicting total yearly compensation using multiple linear regression and how well each predictor variable by itself did at predicting total yearly compensation using multiple linear regression.
3. I found that the linear regression model with years of experience as the input had the highest r-squared value among all the predictor variables with 0.162. I found that the linear regression model with all the predictor variables as inputs had an r-squared value of 0.266.
4. Since r-squared values tell us the percentage of variance explained by the model, we use these values to tell us how much variance in the total yearly compensation is explained by the predictor variable(s). Since years of experience has the highest r-squaed value among all of the individual predictor variables, it is the best predictor of total yearly compensation. Furthermore, we see that since the model with all the predictor variables has a relatively low r-squaed value of 0.266, it doesn't explain much of the variance in the total yearly compensation. However, the model with all the predictor variables as inputs still does notably better than the model with just the best predictor variable (which we've determined to be years of experience) as that model has an r-squaed value of 0.162.

Question 2

1. I fit a ridge regression model with the same predictor variables I used for Question 1 as inputs and total yearly compensation as the output (normalizing both the non-categorical inputs and the output before fitting the model. I then split the data into train and test sets with a 80/20 split. Then, I created the ridge regression model using a for loop that loops through a large number of lambdas. Then, I plotted the RMSE from all the different alphas I tested out to find the optimal lambda. I then plugged this optimal lambda into a ridge regression model and tested it against the test set to calculate an r-squaed value. I

repeated the same process with just the best predictor variable, years of experience, as the input.
2. I did this to see how well all of the predictor variables together did at accurately predicting total yearly compensation using a ridge regression model (as opposed to OLS) and how well the best predictor did by itself did at predicting total yearly compensation using a ridge regression model (as opposed to OLS). I also did this to find the optimal lambda that should be used when creating the models.
3. I found that the optimal lambda when using all of the predictor variables together is 10 and that the optimal lambda when using just the best predictor by itself is 0 (as these resulted in the lowest RMSEs). Furthermore, I found that the ridge regression model with all the predictor variables as inputs had an r-squaed value of 0.304 and that the ridge regression model with just the best predictor variable as input had an r-squaed value of 0.188.
4. Since the r-squaed values for both the ridge regression model that uses all the predictors and the ridge regression model that uses just the best predictor are better than their linear regression counterparts (0.304 vs. 0.266 and 0.188 vs. 0.162 respectively), we see that the model has improved since the ridge regression models explain a greater proportion of the variance in the dependent variable. Through using ridge regression instead of OLS to predict total yearly compensation, while we reduce variance which helps improve the generalizability of the models, we end up increasing bias. For our specific dataset, however, reducing variance and increasing bias through ridge regression results in higher r-squared values making this model better suited for our data.

Question 3

1. I fit a Lasso regression model with the same predictor variables I used for Question 1 and 2 as inputs and total yearly compensation as the output (normalizing both the non-categorical inputs and the output before fitting the model. I then split the data into train and test sets with a 80/20 split. Then, I created the Lasso regression model using a for loop that loops through a large number of lambdas. Then, I plotted the RMSE from all the different alphas I tested out to find the optimal lambda. I then plugged this optimal lambda into a Lasso regression model and tested it against the test set to calculate an r-squaed value. I repeated the same process with just the best predictor variable, years of experience, as the input.
2. I did this to see how well all of the predictor variables together did at accurately predicting total yearly compensation using a Lasso regression model (as opposed to OLS or ridge) and how well the best predictor did by itself did at predicting total yearly compensation using a Lasso regression model (as opposed to OLS or ridge). I also did this to find the optimal lambda that should be used when creating the models.
3. I found that the optimal lambda when using all of the predictor variables together is 0 and that the optimal lambda when using just the best predictor by itself is 0 (as these resulted in the lowest RMSEs). Furthermore, I found that both Lasso regression models

shrunk none of the predictor betas to 0. Finally, I found that the Lasso regression model with all the predictor variables as inputs had an r-squaed value of 0.304 and that the Lasso regression model with just the best predictor variable as input had an r-squaed value of 0.188.

4. Since the r-squaed values for both the Lasso regression model that uses all the predictors and the Lasso regression model that uses just the best predictor are better than their linear regression counterparts (0.304 vs. 0.266 and 0.188 vs. 0.162 respectively), we see that the model has improved from the initial model since the Lasso regression models explain a greater proportion of the variance in the dependent variable. The r-squaed values for the Lasso regression model, however, are basically the same as the ones for the ridge regression model. This is because no features were excluded shown by the fact that none of the predictor betas were shrunk to 0. Through using Lasso regression instead of OLS to predict total yearly compensation, while we reduce variance which helps improve the generalizability of the models, we end up increasing bias. For our specific dataset, however, reducing variance and increasing bias through Lasso regression results in higher r-squared values making this model better suited for our data compared to OLS. However, when comparing our Lasso regresion models to our ridge regression models, we can see that our ridge regression models are more suited for the data since no predictor variables ended up being excluded.

Question 4

1. I fit a logistic regression model with total yearly compensation as the input and gender as the output. I then split the data into train and test sets with a 70/30 split. After running a logistic regression, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. Furthermore, I got the coefficient(s) of the model. I repeated the same process with total yearly compensation and the other predictor variables as the inputs.
2. I did this to see how accurately I could predict gender when just looking at total yearly compensation versus total yearly compensation with other predictor variables. I also did this to analyze if there is a gender-based pay gap in the company.
3. I found that for both logistic regression models, the AUC = 0.5 and the accuracy = 0.82. The curve that comes from the model with just total yearly compensation as the input is around half the time below the red dotted line and half the time above it while the curve that comes from the model that controls for other factors is almost always either at or above the red dotted line. Furthermore, I found that the coefficient of the logistic regression model with just total yearly compensation as input is -6.77043834e-06 and the coefficients of the logistic regression model that controls for other factors are -8.57818040e-07, -6.08184031e-06, -3.47021104e-06, -2.33255286e-07, -8.53271175e-07, -2.47171689e-08, -6.20531365e-08, -7.09185662e-07, -4.36386954e-07, 1.84250281e-08, -1.00650664e-07, -4.23277335e-05, -8.12448726e-05, -7.76908676e-06, -1.21372230e-03, and -3.32382462e-06.
4. As the AUC = 0.5 and the accuracy = 0.82, we can conclude that total yearly compensation both by itself and with other factors is not able to predict gender (even

though when controlling for other factors we get a slightly better result as shown by the comparison of the curves). While the coefficient(s) being negative means that being male is associated with a higher log-odds of earning a higher wage, the coefficient(s) overall being very close to 0 indicate that there may not be a gender-based pay gap in the company.

Question 5

1. I started with finding the median of the total yearly compensation column, creating a new column based on this median that represents a binary variable where compensations equal to or above this median result in a value of 1 and less than this median result in a value of 0. I then fit a logistic regression model with years of experience as the input and this new binary variable I just created as the output. I then split the data into train and test sets with a 70/30 split. After running a logistic regression, I ran the predictions on the test set to calculate my accuracy. Then, I plotted the graph to show the ROC curve and the AUC. I repeated the same process with age, height, SAT score, and and GPA as inputs individually.
2. I did this to see if any of these factors individually could predict if an individual received high or low pay, categorizing anything equal to or above the median as “high pay” and anything below the median as “low pay”.
3. I found that years of experience had accuracy = 0.65 and AUC = 0.65, age had accuracy = 0.58 and AUC = 0.58, height had accuracy = 0.49 and AUC = 0.50, SAT score had accuracy = 0.60 and AUC = 0.60, and GPA had accuracy = 0.59 and AUC = 0.59.
4. From these results, we see that years of experience has both the highest accuracy and AUC making it the best predictor of high/low pay among these factors. Furthermore, we see that height has the lowest accuracy and AUC making is the worst predictor of high/low pay among these factors.

Extra Credit 1

1. I plotted the distribution of salary, height, and age using histograms.
2. I did this to compare how the distribution of all the variables looked to the shape of a normal distribution function.
3. I found that the salary distribution is heavily skewed to the right (with the majority of salaries falling between 0 and 200k), the height distribution having most of the data points cluster around the middle of the range with the rest of the points falling on the extremes symmetrically, and the age distribution being skewed to the right (with the majority of ages falling between 20 and 50).
4. Since a normal distribution is described as having most of the data points falling in the middle of the range of values (mean and median generally in the center) with the rest of the data points symmetrically tapering off to the lower and higher extremes (resulting in a bell-shaped symmetrical curve), only the height predictor variable is normally distributed. This does not surprise me because this dataset has a large number of samples and is a good representation of the population in regards to height as jobs in tech don’t

discriminate based on heights (as opposed to certain occupations like NBA players). Furthermore, salary not being normally distributed also doesn't surprise me because where the majority of the data points lie for this distribution matches the average software engineer salaries with the skew coming from the few outliers in the company with incredibly high salaries (likely executives). Finally, age not being normally distributed also doesn't surprise me because most software engineers tend to be younger and people older than the age of 50 tend to begin retiring.

Extra Credit 2

1. I looked at everything I coded for previous questions to look for anything interesting about the dataset that may have already been semi-explored by work I've done for previous questions.
2. I did this to see if there's anything particularly interesting about this dataset that hasn't been explored in-depth in this assignment.
3. I found that in the first question, when using SAT score as the input and total yearly compensation as the output for a linear regression model, the r-squared value is surprisingly high at 0.1107, making it the second best predictor of total yearly compensation.
4. This surprises me as the SAT is a test taken in high school, making it somewhat strange that it is the second best predictor for total yearly compensation for a job that someone does likely after university (which is well after high school). A possible theory of why SAT is a relatively good predictor of total yearly compensation is it having an effect on where a person goes to college which does have a more direct effect on what kind of job prospects and therefore salaries are available to a person.