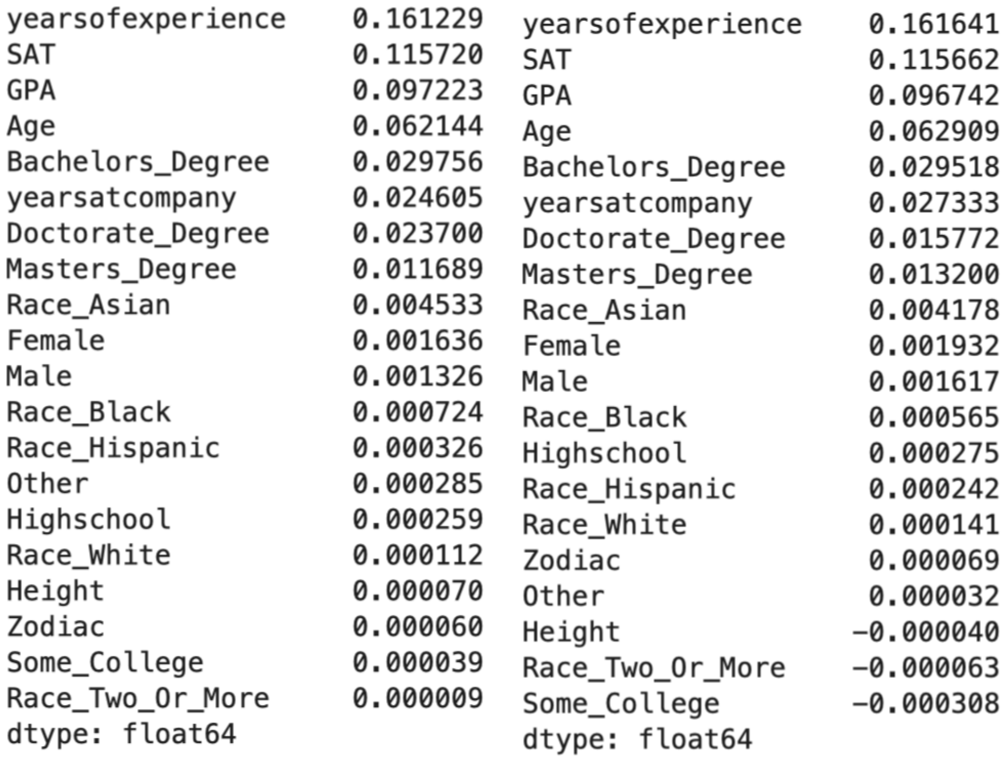
To carry out with the following questions, I first completed some data preprocessing and cleaning. I first determine the number of Nan values in the dataset on education variable. This is due to the reason that if education is Nan, variables that are dependent of education are not meaningful for interpretation. Dropping the rows where education variable is Nan will be the appropriate first step to proceed. This is then leaving us with almost half of the original dataset, from 62642 to 30370 observations. Although this is the half of the original data amount, it is still a fairly large data amount. Then I sliced the data and obtained the section of explanatory variables that are consisting of variable 8 to variable 20 and concatenated variable 23 to variable 27 to the dataframe. Continuing with the process, I concatenated the dependent variable (total yearly compensation) to the dataframe to ensure the dependent variable has the exactly same number of observations as all the explanatory variables. Then I drop all the rows with Nan values in the dataframe consisting all required variables making the number of observation reducing from 30370 to 27766. This is the most efficient way to reduce the number of observations loss to dropping Nan values and 27766 is also not consider a large different from 30370. Separating the dependent variable from the dataframe as y and drop the column from the dataframe to enable all explanatory variables as X. I then used the get_dummies function in pandas to encode observations in the gender variable to 3 different columns consisting observations in gender variable (female, male and other) with 1 or 0 corresponding to the correct value and concatenated into X which then allowed me to drop the gender variable. Hence, resulting in 20 explanatory variables and dependent variable with 27766 observations.

Question 1

As the problem statement requires me to utilize multiple linear regression, to determine the best predictor of total annual compensation (dependent variable) I fit the each of the explanatory variable with total annual compensation into a linear regression model and determine the explanatory variable with the highest R-squared value. Then, by cross-validating with the test set that are obtained from train_test_split with a specification of test_size = 0.5 and random_state = 0 (holding constant across whole project), I am able to determine and validate the best predictor among the explanatory variables. This is due to the reason that R-squared score indicate the variation explained by the explanatory variable, the higher the R-squared score, the better the model fits the data. Also, I repeat the process doing a multiple linear regression with the complete explanatory variables. This enables me to compare the variation explained by the 2 different models. The result obtained from the single variable linear regression is that “years of experience” variable has the highest R-squared score, 0.161229 in the original data and 0.161641 in the split data.



```
yearsofexperience     0.161229
SAT                   0.115720
GPA                   0.097223
Age                   0.062144
Bachelors_Degree      0.029756
yearsatcompany        0.024605
Doctorate_Degree      0.023700
Masters_Degree        0.011689
Race_Asian            0.004533
Female                0.001636
Male                  0.001326
Race_Black            0.000724
Race_Hispanic         0.000326
Other                 0.000285
Highschool            0.000259
Race_White            0.000112
Height                0.000070
Zodiac                0.000060
Some_College          0.000039
Race_Two_Or_More      0.000009
dtype: float64
```

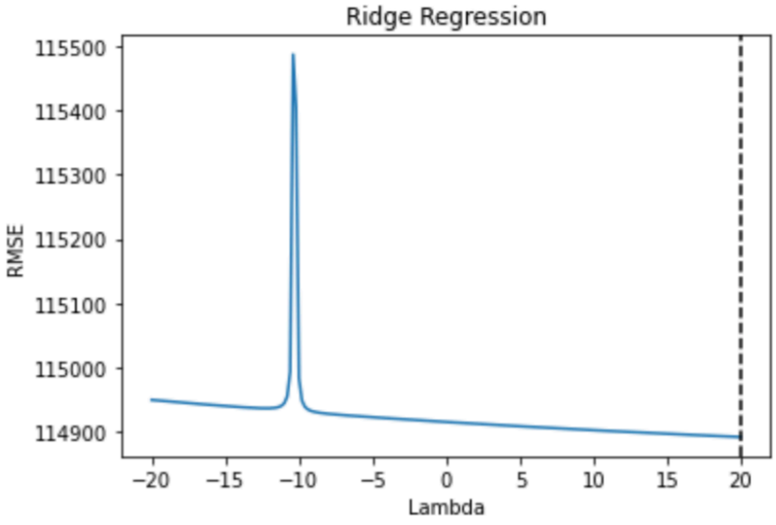
```
yearsofexperience     0.161641
SAT                   0.115662
GPA                   0.096742
Age                   0.062909
Bachelors_Degree      0.029518
yearsatcompany        0.027333
Doctorate_Degree      0.015772
Masters_Degree        0.013200
Race_Asian            0.004178
Female                0.001932
Male                  0.001617
Race_Black            0.000565
Highschool            0.000275
Race_Hispanic         0.000242
Race_White            0.000141
Zodiac                0.000069
Other                 0.000032
Height               -0.000040
Race_Two_Or_More     -0.000063
Some_College         -0.000308
dtype: float64
```

Original dataset vs. Split dataset

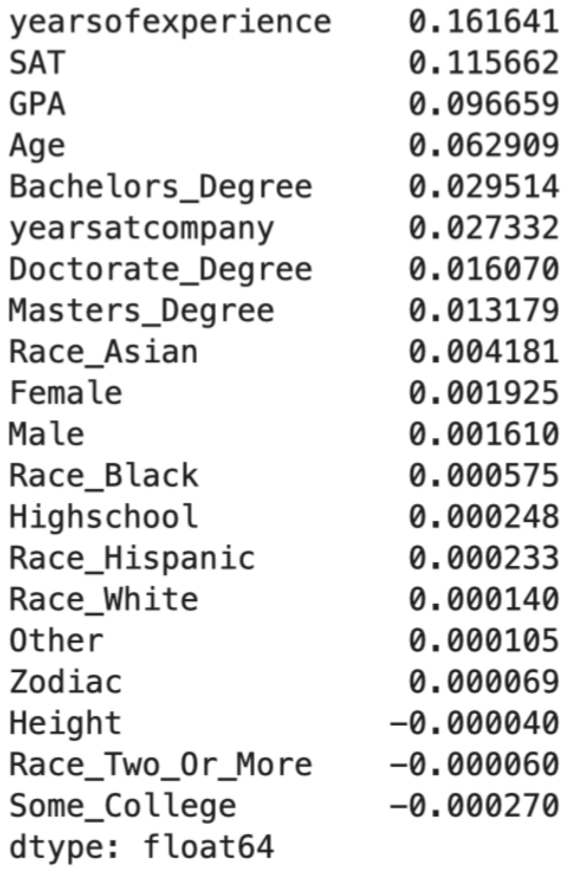
This not only indicates that “years of experience” is the best predictor of total annual compensation, it also validate that it is indeed the best predictor. For multiple linear regression, the R-squared score is approximately 0.2766 for the original data and 0.2662 for the split data. These results indicate that the best predictor successfully explained approximately 16.1% of the variation while the full multiple regression model successfully explained approximately 26.62% of the variation.

Question 2

As the problem statement requires to repeat the same process in Question 1 with ridge regression. I first determine the optimal lambda using ridge regression, then display the plot of lambda against RMSE to visualize the optimal lambda where maximum value of mean squared error is equal to minimum value of lambda. Then, use the obtained optimal lambda to fit a ridge regression for each of the explanatory variables and a ridge regression for full variables regression. This is because using a numpy linespace enables us to determine the optimal value for lambda by testing multiple (201) values between -20 and 20. This action is called hyperparameter tuning. Difference between ridge regression and multiple linear regression is that ridge regression is a regularization technique that are able to prevent overfitting by adding penalty term to the least-squares objective function in the linear regression model. The ridge regression’s penalty term shrinks coefficients towards 0 and hence reducing the variance of the estimates. The optimal lambda is 20.0 displayed in the plot below:



According to the single variable ridge regression, the best predictor for total annual compensation is same as in Question 1, “years of experience” variable with a R-squared score of 0.161641 whereas the full variable ridge regression has a R-squared score of approximately 0.2665.

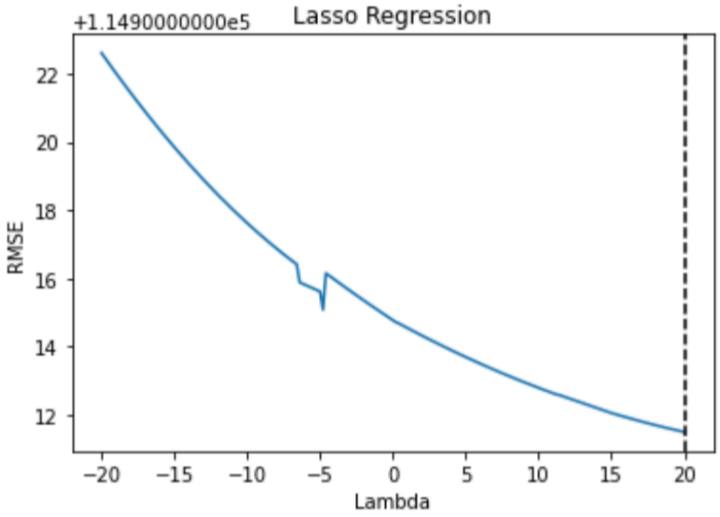


```
yearsofexperience    0.161641
SAT                  0.115662
GPA                  0.096659
Age                  0.062909
Bachelors_Degree     0.029514
yearsatcompany       0.027332
Doctorate_Degree     0.016070
Masters_Degree       0.013179
Race_Asian           0.004181
Female               0.001925
Male                 0.001610
Race_Black           0.000575
Highschool           0.000248
Race_Hispanic        0.000233
Race_White           0.000140
Other                0.000105
Zodiac               0.000069
Height              -0.000040
Race_Two_Or_More    -0.000060
Some_College        -0.000270
dtype: float64
```

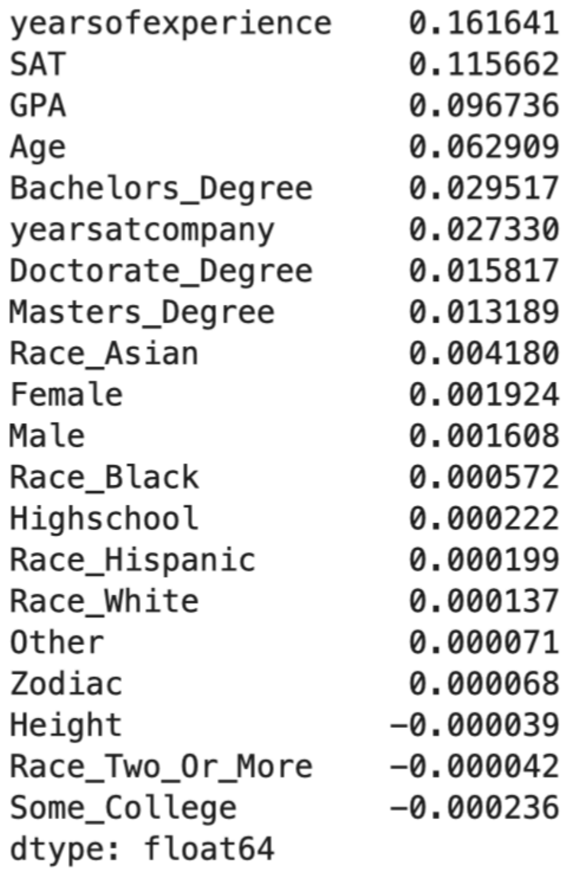
This indicates that the best predictor best predictor successfully explained approximately 16.1% of the variation while the full multiple regression model successfully explained approximately 26.65% of the variation. Comparing to OLS, ridge regression has a slightly higher R-squared score as it capture more of the systematic variation in the data but it also slightly increases bias. This will resulting in a more generalizable model on new data than the OLS model.

Question 3

As the problem statement requires to repeat the same process in Question 1 with lasso regression. I first determine the optimal lambda using lasso regression, then display the plot of lambda against RMSE to visualize the optimal lambda where maximum value of mean squared error is equal to minimum value of lambda. Then, use the obtained optimal lambda to fit a lasso regression for each of the explanatory variables and a ridge regression for full variables regression. This is because using a numpy linespace enables us to determine the optimal value for lambda by testing multiple (201) values between -20 and 20. This action is called hyperparameter tuning. Difference between lasso regression and multiple linear regression is that ridge regression is a regularization technique that are able to prevent overfitting by adding L1 penalty term to shrink coefficients towards 0. The optimal lambda is 20.0 displayed in the plot below:



According to the single variable lasso regression, the best predictor for total annual compensation is same as in Question 1, “years of experience” variable with a R-squared score of 0.161641 whereas the full variable ridge regression has a R-squared score of approximately 0.2663. This indicates that the best predictor best predictor successfully explained approximately 16.1% of the variation while the full multiple regression model successfully explained approximately 26.63% of the variation.



```
yearsofexperience    0.161641
SAT                  0.115662
GPA                  0.096736
Age                  0.062909
Bachelors_Degree     0.029517
yearsatcompany       0.027330
Doctorate_Degree     0.015817
Masters_Degree       0.013189
Race_Asian           0.004180
Female               0.001924
Male                 0.001608
Race_Black           0.000572
Highschool           0.000222
Race_Hispanic        0.000199
Race_White           0.000137
Other                0.000071
Zodiac               0.000068
Height              -0.000039
Race_Two_Or_More    -0.000042
Some_College        -0.000236
dtype: float64
```

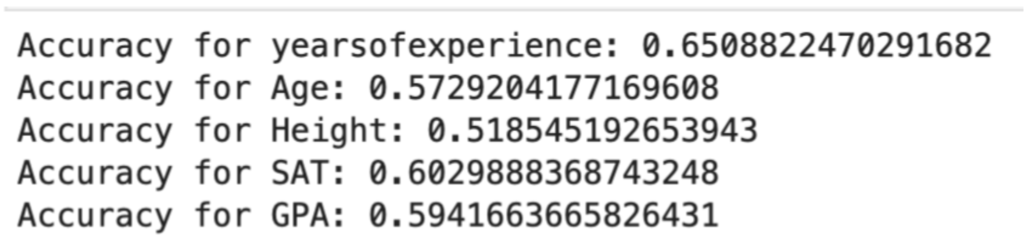
It is also observed that 1 of the predictor beta are shrunk to exactly 0 which is the gender (male) variable. Comparing to OLS, lasso regression has a slightly larger R-squared score in this case as there exist predictor (male) in the model that are not useful to predict the total annual compensation.

Question 4

There is always a controversy regarding the gender pay gap especially in tech companies. To build a logistic model for both controlling or not controlling for other factors, I first filter out the all the observations under gender variable that are classified as “other” and only consider male and female in this scenario as I am targeting to observe between only male and female. Then, to be controlling for other factors, X_gender should be considering all variables dropping out 3 selections of gender (female, male and other) variable as y_gender whereas to not be controlling for other factors, y_gender will remain the same while Xnew_gen will only consist of the “total yearly compensation” variable. Then fit combination of (X_gender and y_gender) and (Xnew_gen and y_gender) to a logistic regression with specification of solver = newton-cg and class-weight = balanced after conducting a train_test_split. The newton-cg specification is a well-suited specification for logistic problems with sparse data which there are many more observation that the explanatory variables whereas the balanced specification allow leveraging of the values of y to adjust the weights inversely proportional to class frequencies to address the class imbalance issue in the gender variables. These series of actions leads up to result of 50% accuracy controlling for other factors which indicates that there are possibly no bias in pay between male and female in this dataset. Moreover, since all the betas obtained are relatively small, I suggest that it is difficult to determine whether the betas we obtained is appreciable. For not controlling other factors, the accuracy obtained is slightly smaller than controlling for which is 46%. This also indicates that there exist no bias on pay rate between male and female.

Question 5

To predict high and low pay from the required explanatory variables (years of relevant experience, age, height, SAT score and GPA respectively, I first have to classify the “total yearly compensation” variable into 2 different category by first determining the median of the column and if the target observation are greater than the median value, I classify it under high pay otherwise low pay and place the classified observation under a new column named “paylevel”. Then I drop the original “total yearly compensation” column from the dataset and the newly created “paylevel” variable as y_log. Split the data and create list of the targeted variables stated above to run logistic regression respectively. The result is as follow:



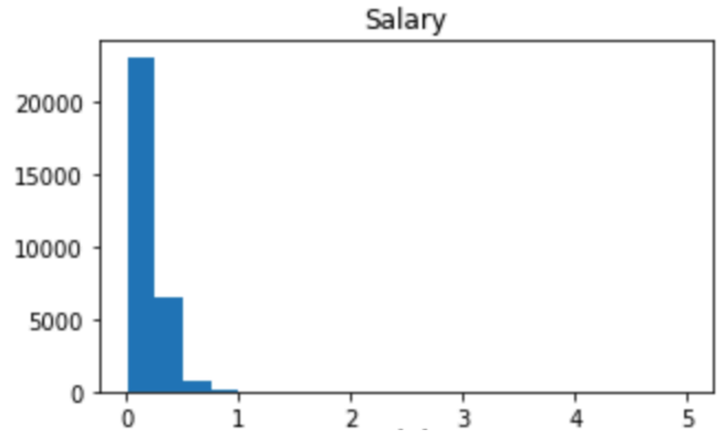
```
Accuracy for yearsofexperience: 0.6508822470291682
Accuracy for Age: 0.5729204177169608
Accuracy for Height: 0.518545192653943
Accuracy for SAT: 0.6029888368743248
Accuracy for GPA: 0.5941663665826431
```

According to the accuracy score displayed above, I can conclude that accuracy for years of experience is about 65% which is the highest among all targeted variables implying that this is the most influencing variable to whether an individual receives high pay or low pay. The accuracy for SAT and GPA are fairly similar which are approximately 60% whereas the accuracy for age is 57% and the worst accuracy is resulted from height. Since the accuracy for height is approximately 51% the model is as good as just pure guessing indicating that height is not exactly correlating with receiving high pay or low pay.

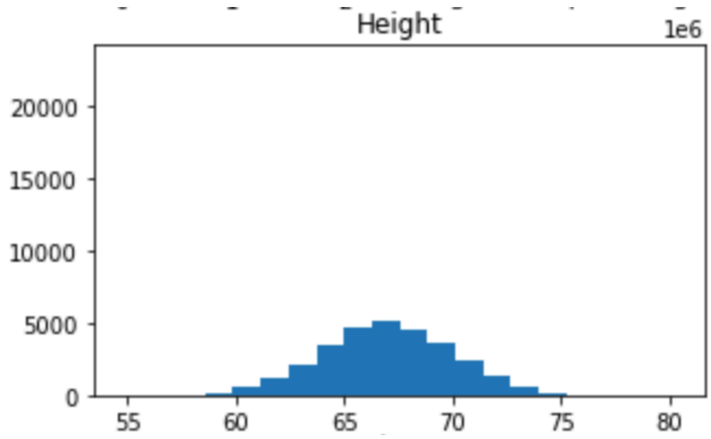
Extra Credit 1

To identify if the data is normally distributed, I separated the required data (salary, height and age) from the dataset and plot the histogram of each of the targeted data. Plotting out histogram allow me to display if the distribution of the required data is a bell-shaped curve. If the displayed curves are indeed bell-shaped, I am able to conclude that the data is normally distributed. The founding are as displayed below in plots:

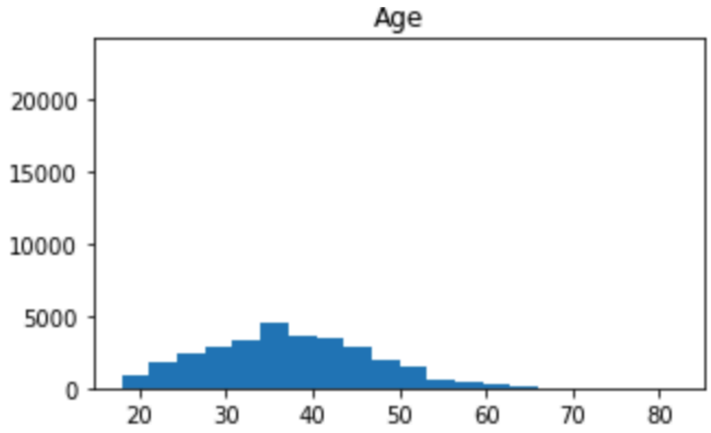
For Salary:



For Height:



For Age:



From the plots above, I am able to conclude that the Height variable is normally distributed. This is not surprising as age is a variable measured in years which is a continuous variable making it a natural consequence that age will be normally distributed. On the other hands, the fact that the distribution of the Age and the Salary variable are not normally distributed are not surprising as well. There are a possible reason that there are relatively few individuals who will earn very high salaries but there will be a large number of individuals who earn low to moderate salaries. For age, the distribution can be skewed in wither direction depending on the population targeted. In this instance, the population targeted are consisting of more young adults and adults that are in employment making the distribution skewed more to towards the left.

Extra Credit 2

To determine if there are something interesting about this dataset, I have to observe the value counts for the 3 variables (company name, position title and location of company) that are not given attention to in earlier phase. Looking at the count for company name, it is observed that bigger tech companies have a greater capacities for employment. However, I also observed that among the bigger tech companies, Amazon has the greatest count of observations whereas Apple has the lowest count of observations. This is possibly due to the reason that Amazon required a bigger employment capacity as their business is diversified across a wide range of industries and services such as e-commerce, cloud computing and digital streaming. This indicates that Amazon requires larger and more diverse workforce to support their operations in these areas which also implying that they would require a large number of software engineer. On the other hand, even if Apple is also developing software are services, its primary focus in the industry is on hardware products resulting that their employment capacity is lower in terms of bigger tech companies. Moreover, from the count of position title, I observed that software engineer is the pinnacle of the position offered in tech companies. This is not very surprising as most of the tech companies' core business is software making software engineers essential in their companies. According to the company location count, tech companies seems to be more prone to locate themselves in city like Seattle, New York, Redmond and San Jose. The main reason is that these locations are the major tech hubs in United States. In these locations, the tech companies' proximity to access fresh talents from universities and larger pool of highly skilled workforce are more convenient to themselves. Also, the closer the distance to a bigger city, the higher the possibility that a company would be able to access to funding and investment which are essential for tech companies to grow as the amount of cost needed to produce a software is high. Last but not least, the attention is given to the gender variable. As observed from Question 4, there is a huge imbalance for the male and female. This is not a surprising environment for tech industry due to stereotypes that tech jobs are male-dominated jobs and biases in hiring practices that are not welcoming to female.