PreProcessing:

Before using our data, we need to make sure to do the right proprocessing. Indeed, we need to drop the rows with missing values. Additionally, as mentioned in the HW PDF, we need to also drop the variables 1-2-3 and 5-6-7 for optimal results. I did not normalise my data as it does not make sense in the context of this assignment.

Question 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? -> LAB2 (multiple covariates)

The best predictor found is years of experience through using GridSearchCV learned in REC 2. We then use the multiple full regression that we learned in lab2 to get the R^2 value of 0.426 indicating that the multiple full regression is a better predictor for total annual compensation as it considers all the other variables.

Question 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?

I implemented the code that was given in Lab2 and i resulted in a lamda equal 20 and better coefficients for the predictors. The plot shows a line which is evident because of the lamda = 0 we got, this means that the model is the same as the OLS model implemented in the previous question. Indeed, this makes sense because there is no regularization penalty applied to the loss function.

Question 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 lamda now?

I also used the given code for the lasso regression from LAB2 in this question. Indeed, the results obtained where very confusing as I got a negative lamda which does not make sense at all mathematically. Indeed, I was able to compute the coefficients which show that a conisderable number of predictors were able to be shrunk to 0.

```
Out[299]: [('totalyearlycompensation', 138556.78421320952),
           ('yearsofexperience', 0.17496132898196493),
           ('yearsatcompany', -0.02778692299125396),
           ('Age', 0.00052004185914711 34),
           ('Height', 0.00010229422093743779),
           ('Zodiac', 0.004061539434189854),
           ('SAT', 0.11992493229672618),
           ('GPA', 0.003685701211452397)]
```

Question 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.

Referencing the code we used in REC4 I was able to run a logistic regression model on the dataset in order to predict the annual compensation from gender. I had various issues though with doing the model evaluation.

Question 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.

I also here used the code of REC4 indeed I was not able to do the model evaluation.

Extra credits:

a) **Is salary, height or age normally distributed? Does this surprise you? Why or why not?**
b) **Tell us something interesting about this dataset that is not already covered by the questions above and that is not obvious.**

In order to also include company, title, and location as part of our analysis we could convert them into dummy variables.