Data Cleaning:

Columns deleted	Columns containing TEXT	These were the columns where presumable the surveyed had to type in some data. Is nothing was typed in, this column was left black, or a seemingly random number was present.
	How long it took to complete	This is unimportant to study relation
	Number of surveys	This is unimportant to study relation
Rows Deleted	Question containing row	Hinder effective data processing.
	Students with no formal education past high school	Presumed that the correlation between these individuals and the insight they would have to offer with regards to being a data scientist/analyst would be mostly invalid
	People who's current occupation is being a student are dropped.	These individuals are presumed to be in jobs that are not necessarily their careers (part time jobs) and their input on what makes a successful data scientist or engineer is minimal

Reasoning for Filling missing data:

% of missing data < 15%	If the data can be overlapped or asked as an approximation (eg.	Filled using Forward fill-
	Discrete values which only the surveyed could give insight into	New Variables added
% of missing information > 15%	In order to preserve any special relation the present data had with the Salaries column, these values were preserved	New variables added

Label encoding was used for data processing

Notes: There was only one column that was missing over 50% of the data, in retrospect this column should have been dropped and its effect on accuracy studied.

Exploratory Analysis :

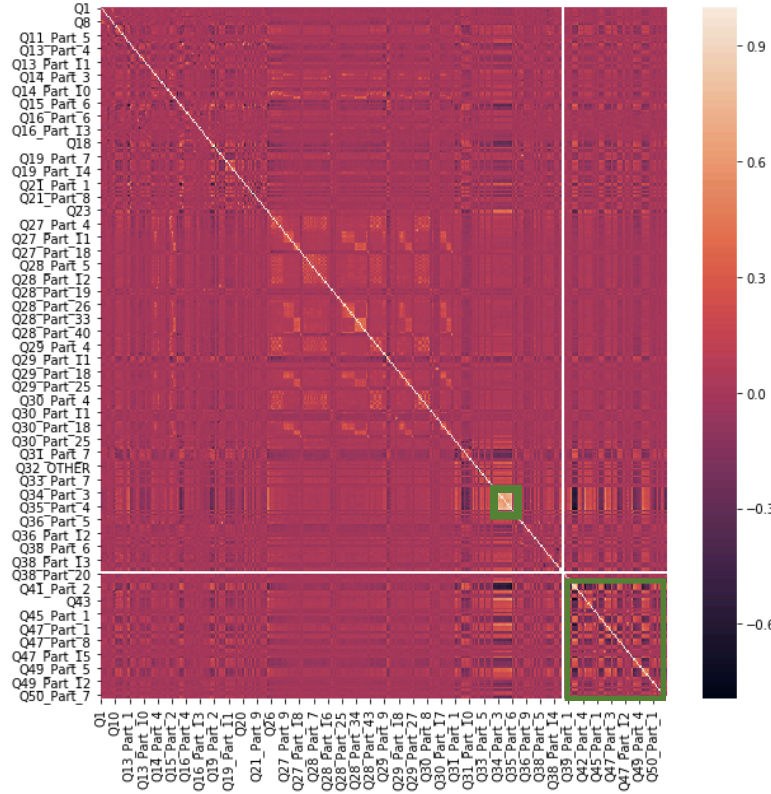
The most informative questions were the first few questions: Q1, Q2, Q3, Q4, Q5, Q6, Q7, Q8- these were the features that did not contain missing data therefore was not altered during the data cleaning process. The relations studied were:

1) Gender + education versus yearly salary
2) Education + Age group versus yearly salary
3) Industry of employment + age group versus yearly salary.

Since the goal of this project to study effect of each feature on the target variable (salary) – It was ensured that salary was one of the factors.

Another requirement was to produce a correlation plot where:

1) The diagonal line suggests that the features correlate with themselves the best
2) There are patters in the middle which show overlap between the different features that are similar
3) The darker the region – the more negative the correlation was
4) The lighter the region - the more positive the correlation was



The two regions outlined in black are regions of interest where there is strong overlap between Q34 and Q35 – where both these questions required the user to add a number that would add up to 100.

Similarly, the latter part of the survey (Q42-Q50) shows a large number of questions that show high positive and negative correlation. The nature of these questions was such that the user was to choose from the information already provided, therefore some correlation is excepted and seen

** Due to the large number of features, pin pointing the exact questions was not possible and estimations needed to be drawn.

Feature Selection:

Three models were used to ensure that the same questions were being picked up as the most important features.

Models	Results:
Correlation Features	15% correlation: 'Q2', 'Q3', 'Q9', 'Q10, 'Q11_Part_4', 'Q15_Part_2', 'Q27_Part_1', 'Q30_Part_9', 'Q38_Part_10']
RFE – Recursive Feature elimination	Q1', 'Q11_Part_1', 'Q11_Part_3', 'Q11_Part_4', 'Q11_Part_5', 'Q11_Part_6', 'Q13_Part_1', 'Q13_Part_11', 'Q13_Part_12', 'Q13_Part_13', 'Q13_Part_14', 'Q13_Part_15', 'Q13_Part_2', 'Q13_Part_4', 'Q13_Part_6', 'Q13_Part_7', 'Q13_Part_8', 'Q13_Part_9', 'Q14_Part_4', 'Q14_Part_6', 'Q15_Part_1', 'Q15_Part_5', 'Q15_Part_6', 'Q2', 'Q3', 'Q4', 'Q5', 'Q7', 'Q8', 'Q13_Part_3', 'Q14_Part_2', 'Q15_Part_4', 'Q14_Part_1', 'Q12_MULTIPLE_CHOICE', 'Q14_Part_8', 'Q14_Part_10', 'Q13_Part_5', 'Q14_Part_9', 'Q14_Part_3', 'Q15_Part_2', 'Q11_Part_7', 'Q6', 'Q14_Part_11', 'Q10', 'Q14_Part_7', 'Q14_Part_5', 'Q11_Part_2', 'Q15_Part_3', 'Q13_Part_10']
ExtraTreesClassifier	Q10 Q34_Part_1 Q35_Part_3 Q34_Part_4 1 Q2 250 Q34_Part_2 7 Q8 105 Q24 4 Q5 251 Q34_Part_3 2 Q3 254 Q34_Part_6 255 Q35_Part_1 106 Q25 253 Q34_Part_5 5 Q6 258 Q35_Part_4 107 Q26 104 Q23 296 Q39_Part_1 297 Q39_Part_2 89 Q20 236 Q32 307 Q43 298 Q40 6 Q7 320 Q46 358 index 259 Q35_Part_5 3 Q4 68 Q17 103 Q22 337 Q48 16 Q12_MULTIPLE_CHOICE 300 Q41_Part_2 25 Q13_Part_9 26 Q13_Part_10 299 Q41_Part_1 317 Q45_Part_4 241 Q33_Part_4 12 Q11_Part_4 260 Q35_Part_6 343 Q49_Part_6 354 Q50_Part_5 29 Q13_Part_13 9 Q11_Part_1 232 Q31_Part_9 231 Q31_Part_8 246 Q33_Part_9

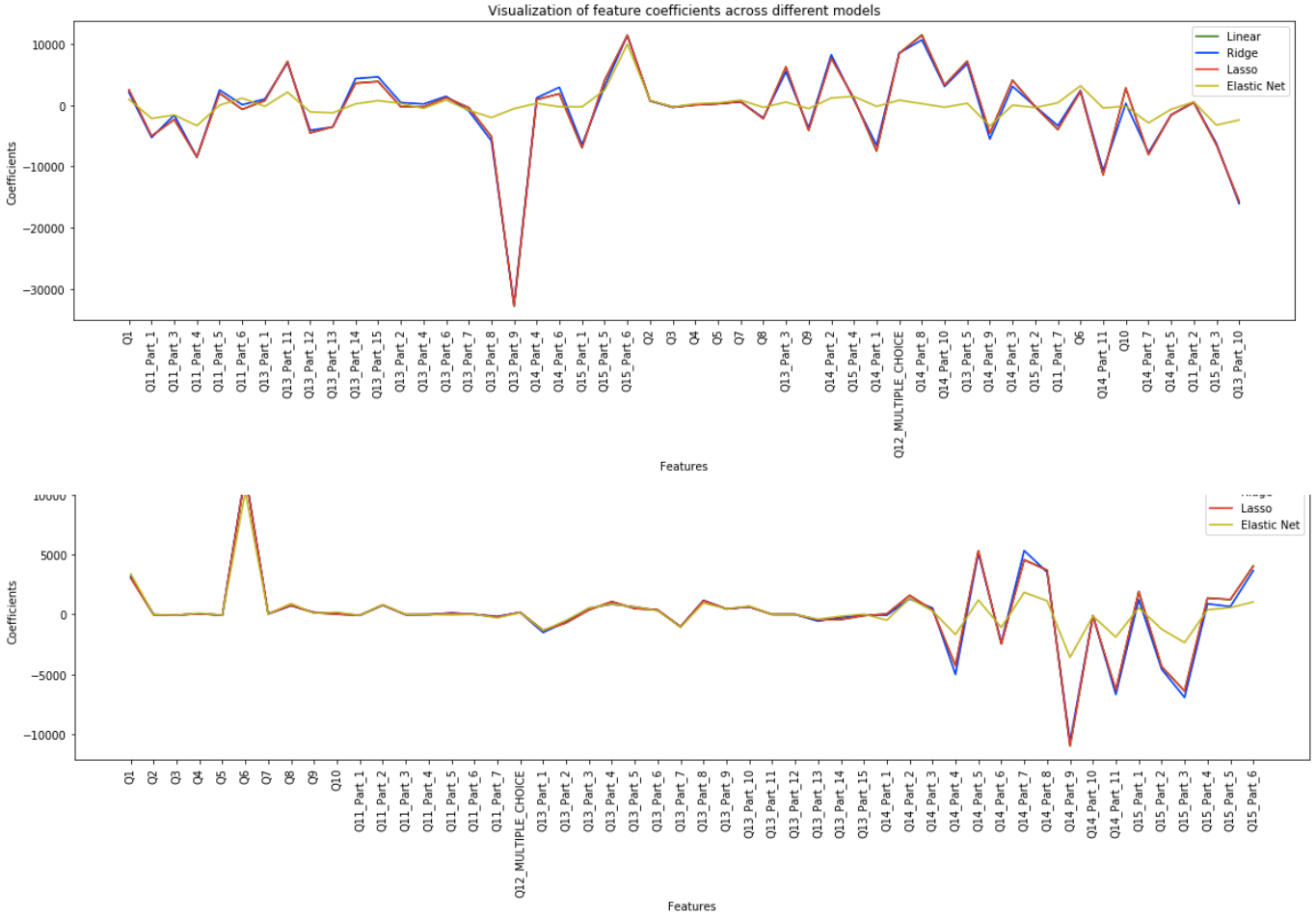
All preliminary questions were picked up – i.e. Q1-Q10. These questions were given the utmost importance as they were not altered. This was automatically feature selected but also manually checked.

The model was forced to provide the top 50 words that were highly ranked – this was an issue later on because some of the features selected were not informative and added more error to the model.

Final features chosen: Was ExtraTreesClassifier and correlation features. However, all the models were tested on both REF+15% correlation results and ExtraTreesClassifier+15% correlation results and there was minimum difference and in the interest of saving time, only the results for the former was reported and discussed.

Model Results and Visualizations – part 1

- All the visualizations of the models was best understood by plotting the coefficients of the selected features. This gives insight into the hyper parameter and model tuning. This is also able to visualize which parameters needed to be damped the most. The following diagrams show the coefficients of the models based on the RFE features and the models based on the tree classifier.



Q 13- part 9: Both questions were present in the lists of features where the question was "Which of the following integrated development environments (IDE's) have you used at work or school in the last 5 years? (Select all that apply) - Selected Choice - Notepad++" . The first sets of models needed to provide a large coefficient to this feature where as the second model did not do so. There with respect to the other parameters, this question was well parameterized in the the second set of features.

Q14: were selected from both the feature selections and it was noted that and it was the set of features that were highly erratic which begs the question as to why the feature selection model chose these sets of features. This might again come back to the cleaning as the model was forced to select the 50 of the best features, perhaps there were so few relevant features that the model had to choose the less impressive features as well.

In retrospect more analysis could have been done to study the effect of different numbers of features on the models.

The features do not line up – the first graph is based off of the RFE feature selection and the second graph is based off of tree classifier feature selection

Model results and visualizations part 2

	Bias	MSE from 10 folds	R2	Variance	Post parameter optimization
Linear Regression	5.364351e+09	34917.250000	0.260204	1.182696e+09	0.276928
Ridge	5.355220e+09	34917.644531	0.260652	1.173799e+09	0.276005
Lasso	5.363920e+09	34916.003906	0.260236	1.182263e+09	0.274863
Elastic Net	5.047986e+09	35847.984375	0.241625	8.660464e+08	0.276417

Results from the features selected by the RFE model

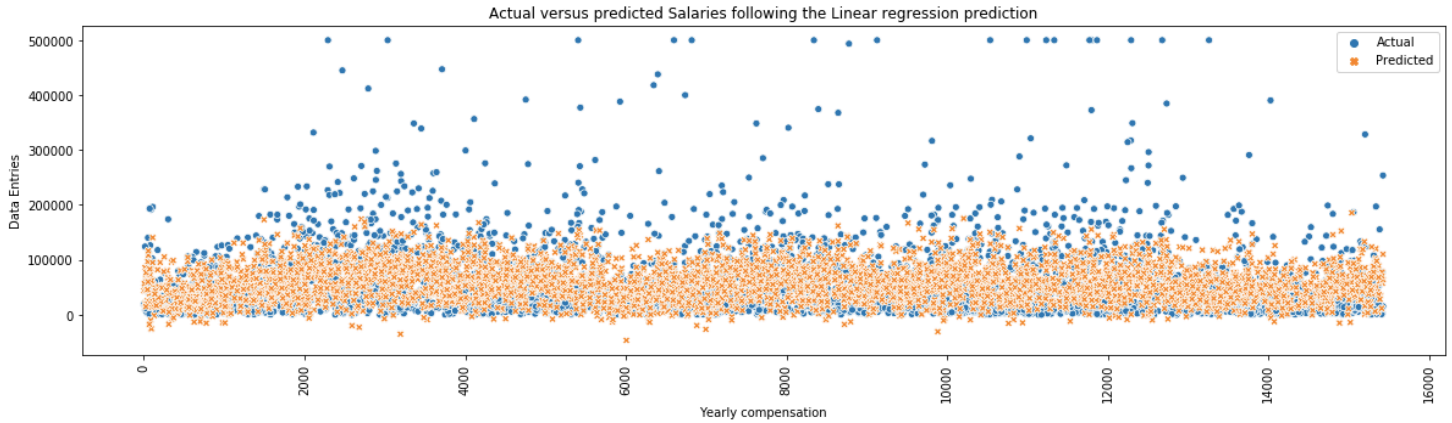
	Bias	MSE from 10 folds	R2	Variance	Post parameter optimization	Train Accuracy
Linear Regression	3.203628e+09	35609.997417	0.233720	1.108263e+09	0.275205	0.262690
Ridge	3.197483e+09	35607.407693	0.235190	1.096467e+09	0.274264	0.263380
Lasso	3.203448e+09	35607.087195	0.233763	1.108068e+09	0.273944	0.262701
Elastic Net	3.215956e+09	36029.720457	0.230771	8.833241e+08	0.274684	0.252479

Final results from tree classifier feature

Based off of marginal increase in the R2 score after model optimization, it was noted that linear regression performed the best.

The Accuracies remained constant with the training accuracy slightly above the testing accuracy. This goes to show that the modes were very well generalized despite their large mean squared error and the small prediction score.

It can be noted that the accuracy does not change across different models even after model optimization. This would imply that the issue is probably with the way that the data was cleaned and the encoding. I used label encoding for my data, however in retrospect, it would have been more beneficial to have used the one hot encoding method.



High level visualization of the actual versus predicted scores of the linear regression model: