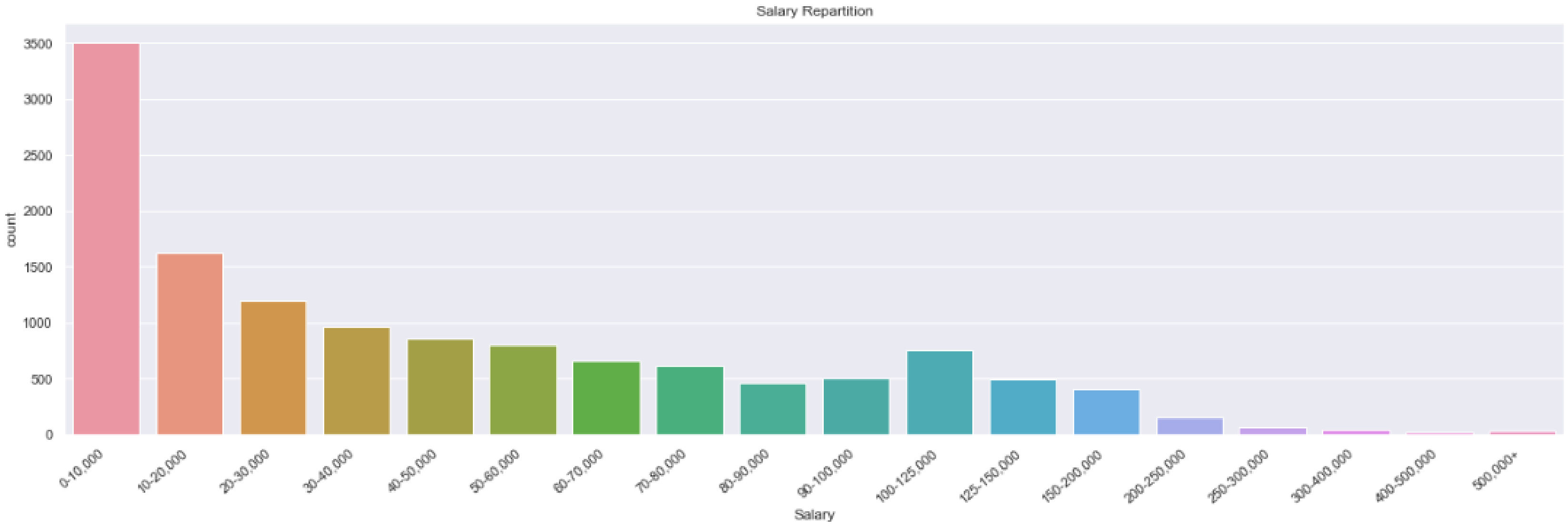
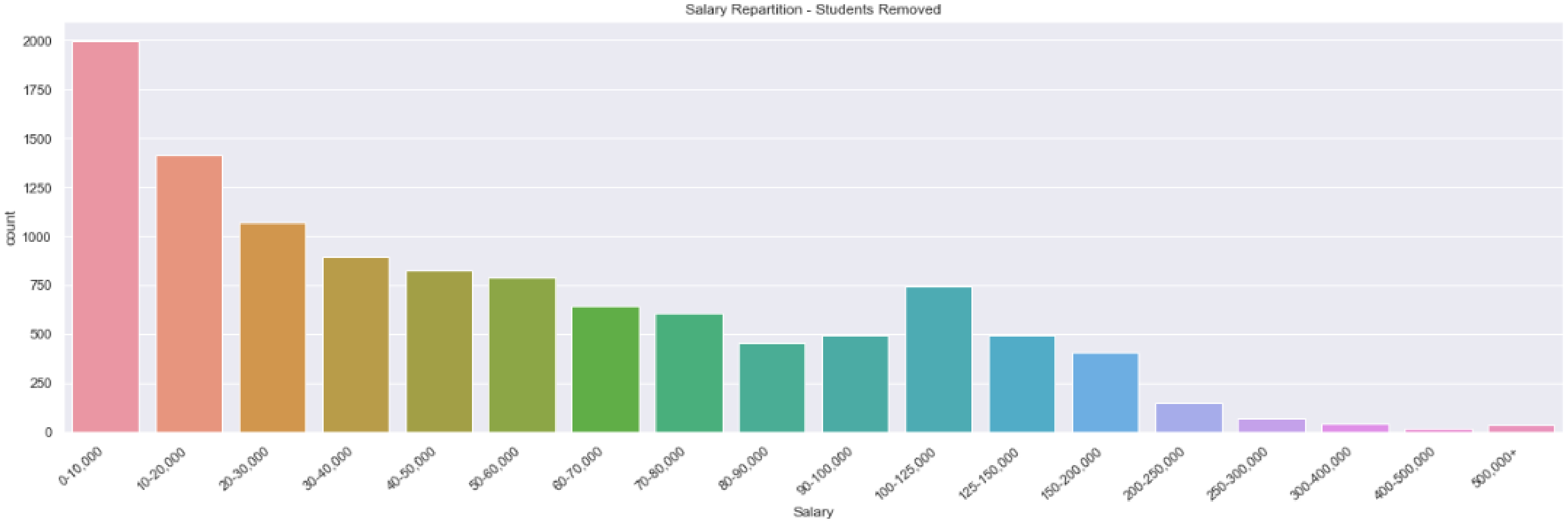
Exploratory Data Analysis

Salary Bracket Distribution



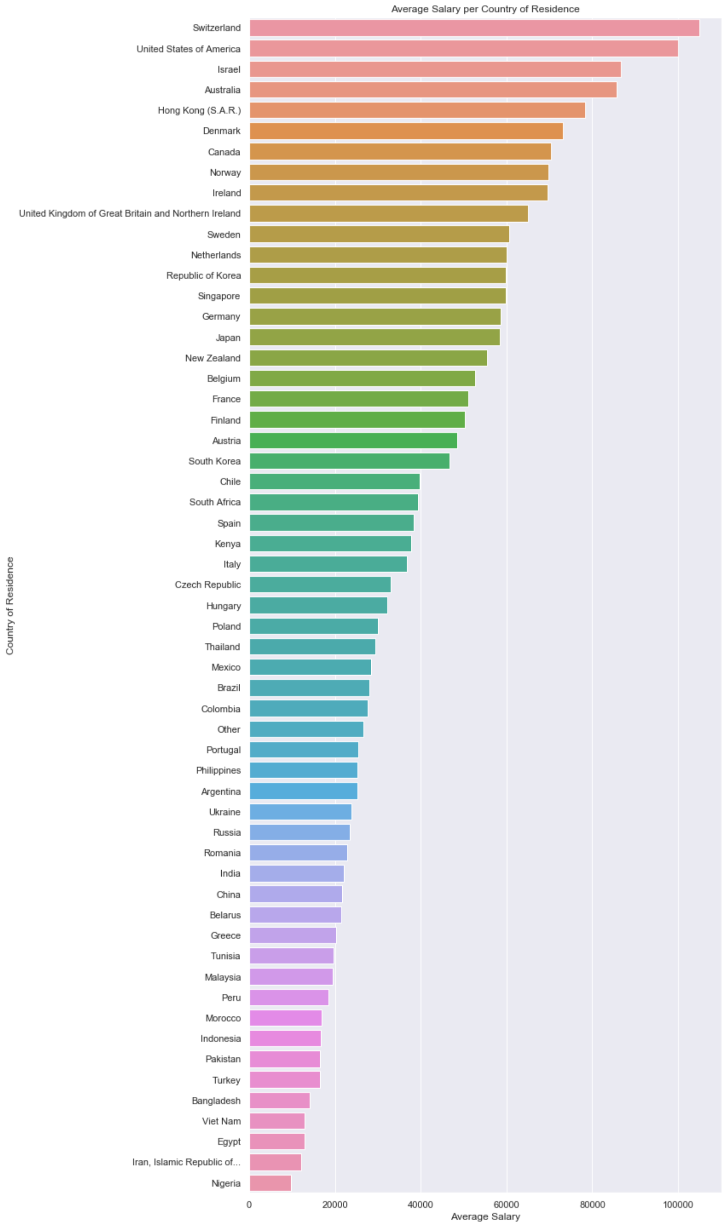
This figure represents the salary repartition or distribution. The distribution is right skewed meaning a lot of the respondents earn low salary. This seems odd since data scientists are known to get paid way more. This is because the large group of the people that took the survey are still a student and obviously students mostly have no income.



This figure represents the salary repartition or distribution with students removed. Even after removing the students, the distribution is still skewed, but less intense. This can be representation of respondents from countries with low salary standards. Also, the peak at the salary bracket,100-125,000 represents data scientists in United States

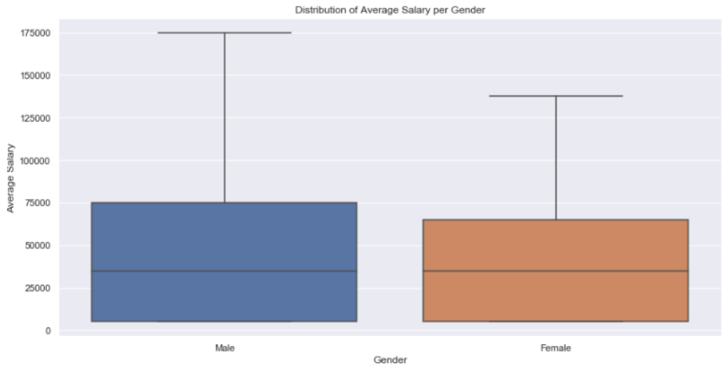
Exploratory Data Analysis

Q3: Country of Residence



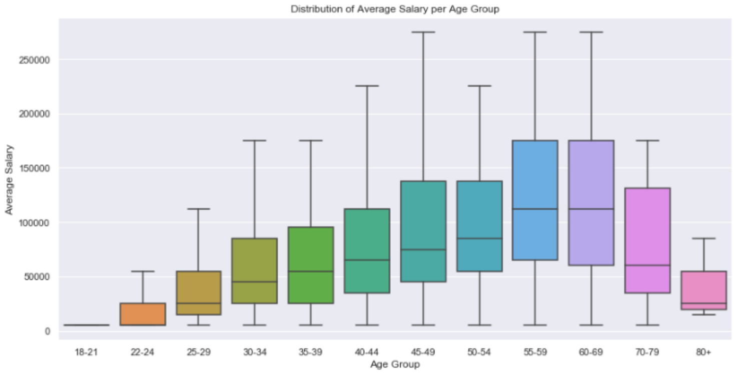
As expected, depending on the country of residence has large impact on the average compensation. The country with high cost of living tends to have higher salary standards. The country of residence show high correlation or trend between the salary.

Q1: Gender



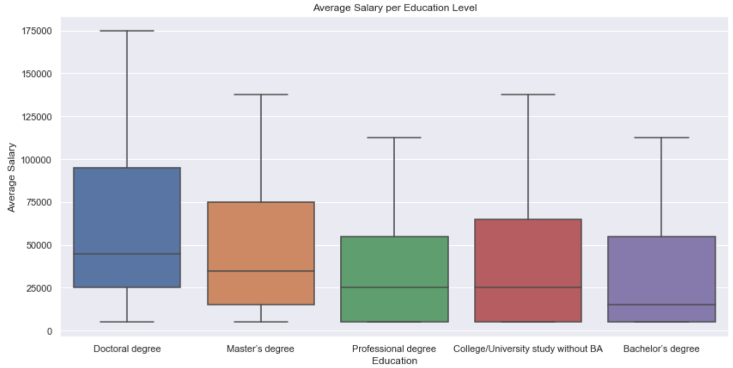
We see that the older you are, the higher the compensation. This makes sense, because older age group tends to have more experience which usually means higher salary.

Q2: Age Group



In average, male tends to earn more money than female. Also, based on the bar distribution curve, there are more male are distributed towards higher salary as opposed to female.

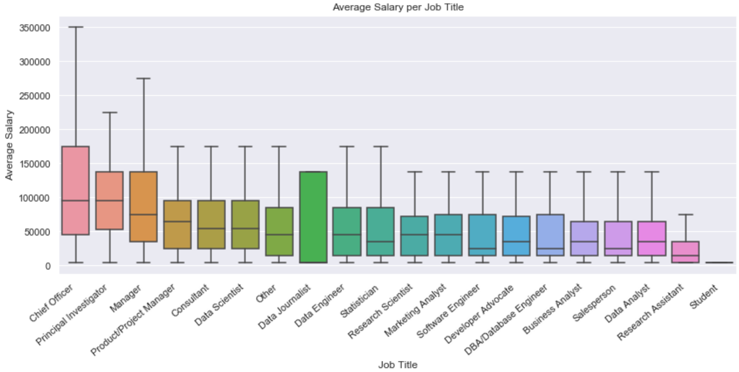
Q4: Education



Those with higher education, doctoral degree and master's degree do tend to earn more money compared to other education levels, but between professional degree, college/university study without BA and bachelor's degree have not much difference in salary.

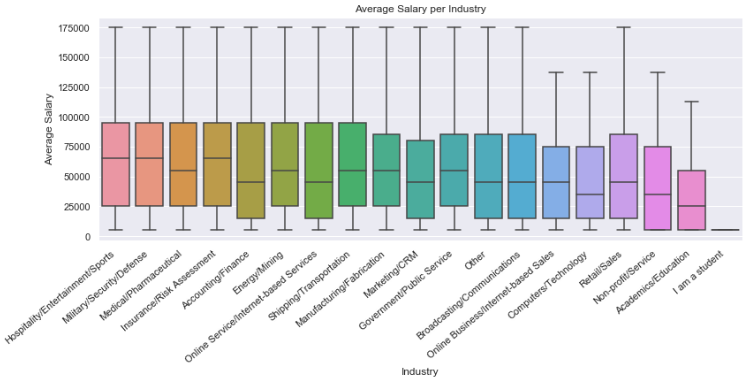
Exploratory Data Analysis

Q6: Job Title



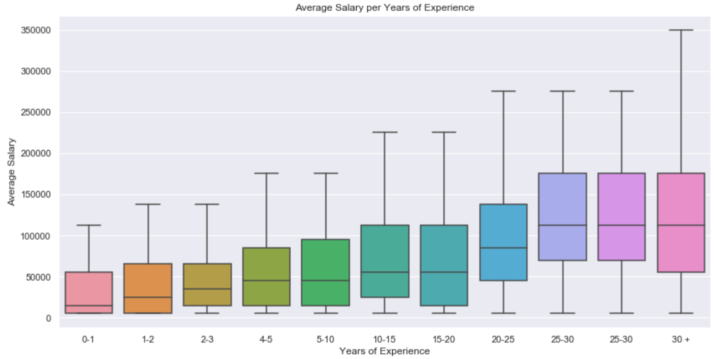
Managerial and c-level positions tend to earn the highest salary. Other positions tend to earn similar salary except for students where students earn less money. Strong salary correlation can be found only for managerial, c-level positions and students.

Q7: Industry



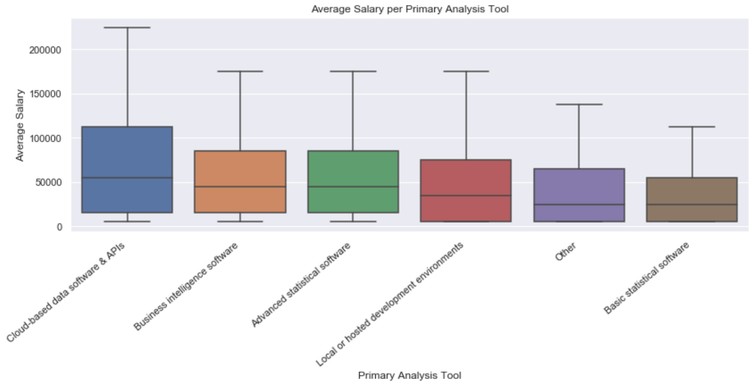
Based on the graph, unless you are a student, the industry and the salary does not have a strong relationship

Q8: Years of Experience



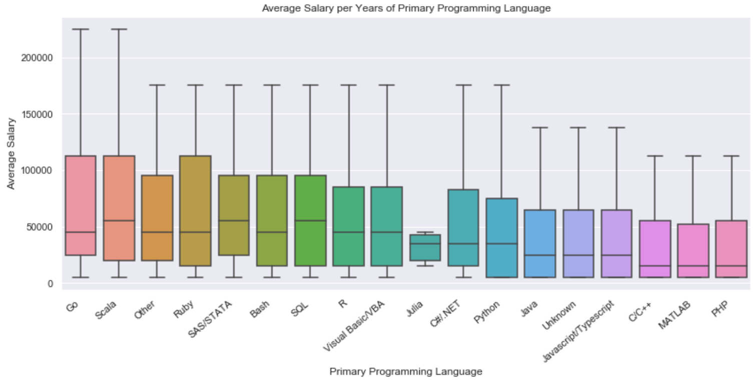
As expected, we see the trend o the more years of experience yo have, the more you earn. This feature shows a strong trend which will help us with the model

Q12: Primary Tool for Analysis



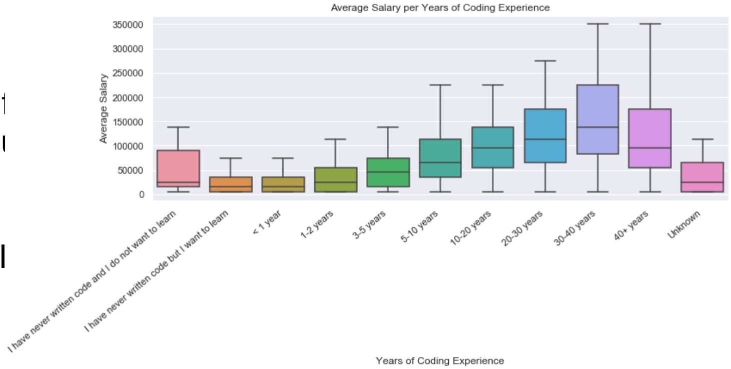
Based on the graph, if you use the cloud-based data software & APIs, you tend to earn more money, but there aren't significant relationship with the salary for other software users.

Q17: Primary Programming Coding Language



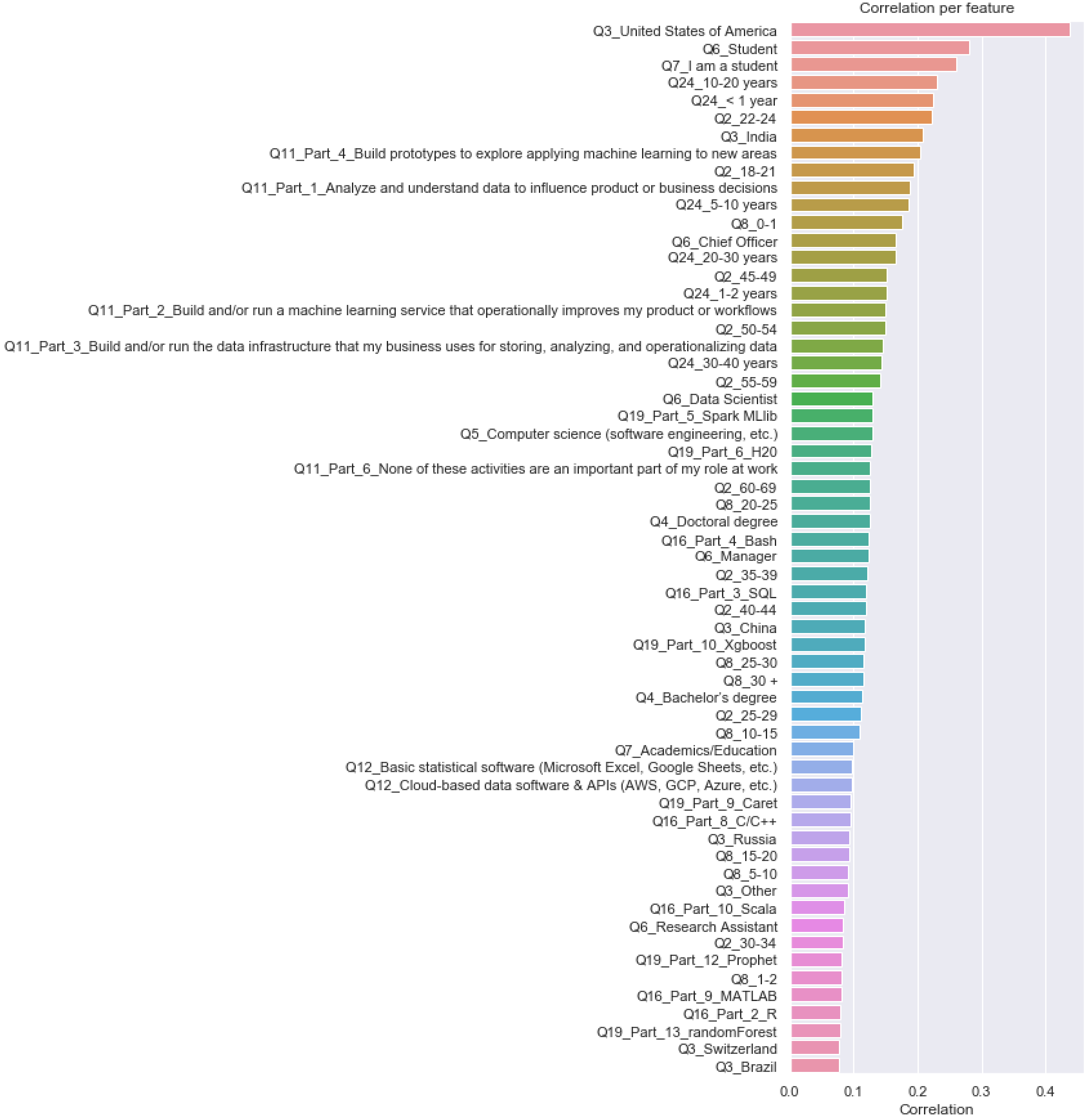
There are few coding language that could get you have strong correlation or trend with the salary, but overall, it seems like this attribute doesn't have strong trend or relationship with the salary.

Q24: Years of Coding Experience



As expected, the more years of coding experience you have, the more you earn.

Feature Importance



In the figure, it shows the feature rankings according to their Pearson correlation with the yearly compensation. Note that due to one hot encoding, it shows the correlation with each category or feature values rather than the feature or attribute itself.

Based on the rankings, we find the most correlated feature values or categories are as follows:

1. Q3_ United States of America
2. Q6_Student
3. Q7_I am a student
4. Q24_ 10-20 years
5. Q2_24 <1 year
6. Q2_ 22-24

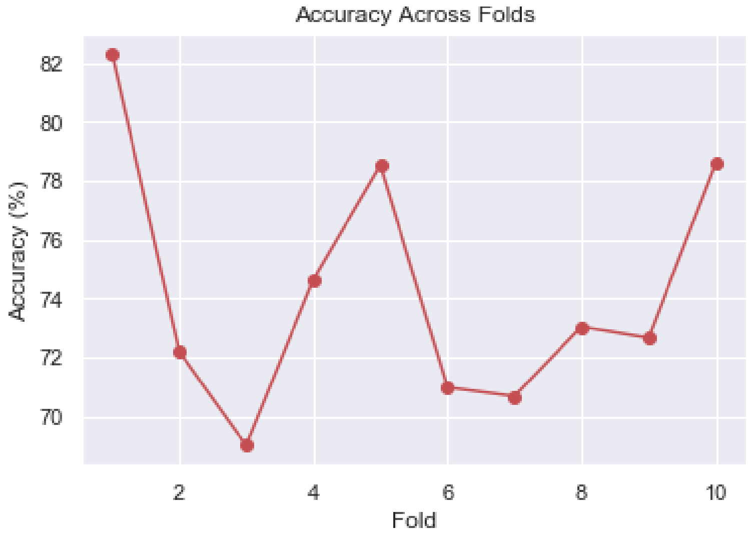
It is worth mentioning that being a student has strong correlation between salary as it is very likely that students earn very little to no money. Having students in our data makes our model bias towards lower salary!

Based on the figure, we find the most correlated features as follows:

1. Q3:Country of Residence
2. Q24: Years of Coding Experience
3. Q2: Age Group
4. Q11: Activities or most important role at work

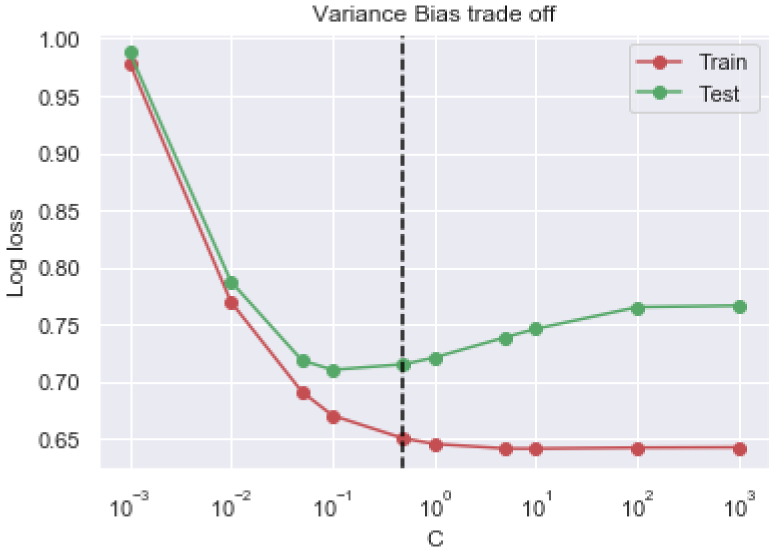
Model Tuning & Results

1st Implementation



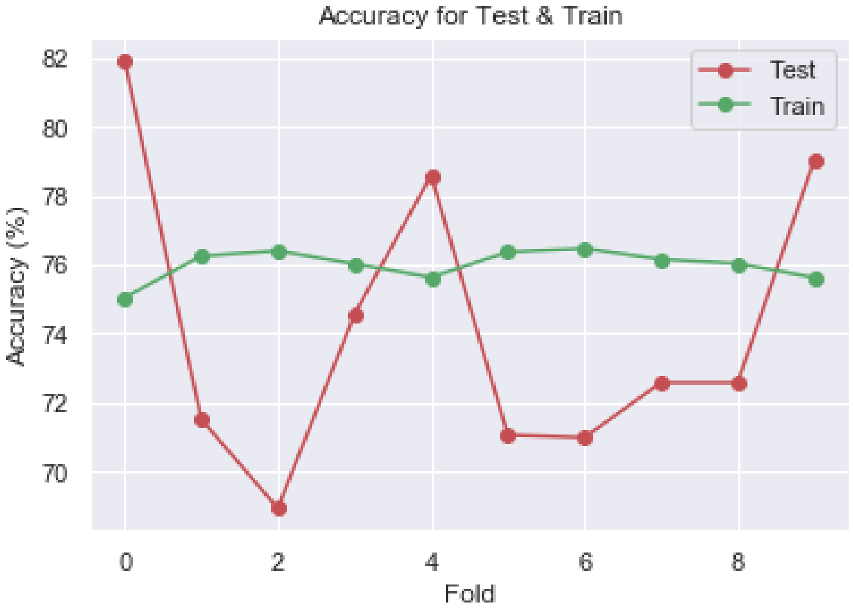
The first model implementation was done with arbitrary parameters, C = 0.01 and solver = liblinear. The accuracy do vary fold to fold. The difference between the highest and the lowest is around 18% which is quite high. This indicates that the test accuracy for each fold depends on which subset of the data are allocated to the test set. The average and variance of accuracy for folds are computed as 70.1% and 5.0% respectively.

Bias and Variance Tradeoff



This figure indicates the log loss across different parameter, C. Based on the figure, the optimal C value would be between 0.1 to 1 as it has optimal balance between bias and variance. Despite the minimum log loss at C = 0.1, we will say that the best performing model is when C = 0.5 as it had the highest average accuracy of 74.5%.

Model Tuning and Results



After running the Gridsearch, the optimal parameter that gave us the highest accuracy were C = 0.5 and solver = liblinear and the test and train average accuracy we got were 74.5% and 76.2 % respectively. For a classification predictive model with 5 classes, these accuracies are pretty good. Also, considering that the difference in accuracy between test and train set is small, we can say that our model is well trained and fitted, but as we saw in the salary bracket distribution, the dataset is highly imbalanced. There aren’t enough datapoints for high salary for our model to study on. Therefore, most of the predictions at higher salary may be incorrect. This causes our dataset to be overfitted. In order to increase the accuracy, we need to work with balanced data, reduce the complexity of the classification by reducing the number of classes, minimize bias, etc.