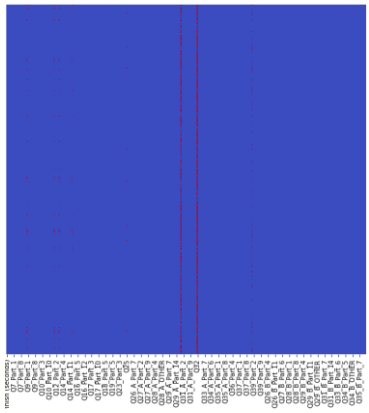
Assignment 2 Report

The objective of this assignment is to build, train and tune a multi-class ordinary classification model which could predict the yearly compensation bucket of each survey respondent.

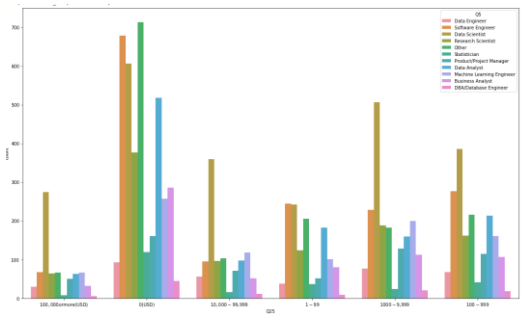
Data Cleaning

After loading the entire dataset, it can be found that the first row of the data includes each question of the survey. According to this row, it can be notice that the columns which includes "Part" and "OTHER" represent for different selections of each question, but there exist some Nan values representing for not select the answer. One way to avoid dropping these meaningful Null values is to fill these values by using “no selection”. This will help us clean the dataset without dropping them.

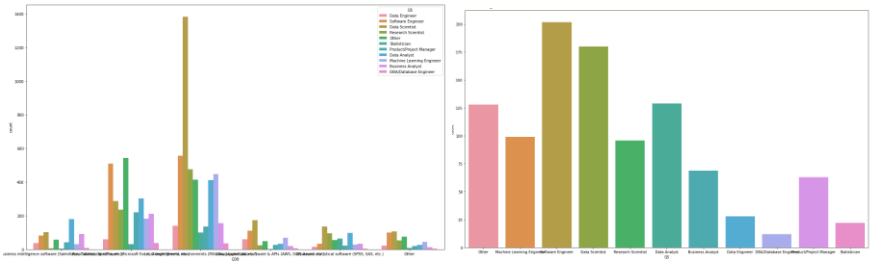
The figure below is the heatmap indicating the missing values’ locations. It shows that there are 2 columns with many missing elements. It is hard to fill such a information gap correctly, therefore dropping these two columns is a fair choice.



For Question 25, it asked about the money spending on machine learning service. Combining with different job titles, the plot below shows the relationship between these 2 variables. The highest bucket in the plot is the 0 USD, while the number of respondents of each job in 0 USD bucket is higher than the others. Therefore, it is reasonable to use 0 USD to fill the missing values.



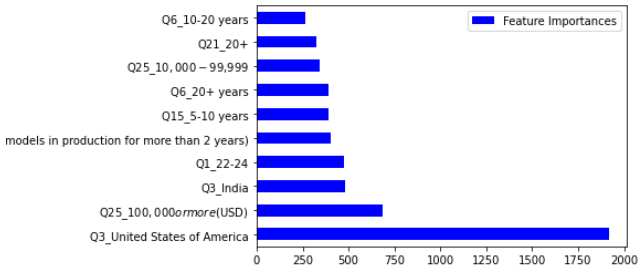
The same logic can be applied to the question 38. As shown in the first plot below, it is clear that most of the data scientist and Machine Learning Engineer use Local Development Environment, while some of the Statisticians and Software Engineers use Basic Statistic tools. More investigation can be conducted on the job titles for those people who did not answer this question. In the second plot, it can be discovered that most of the respondent who did not answer this question have a Data Scientist or Software Engineer job title. Therefore, it is possible to assume that the Local Development Environment is used if the job title is Engineer or Data related while the Basic Statistical Software is applied if the job tile anything else.



In the columns of Q8 Q11 Q13 Q15, they have the missing values in the same 561 rows. Comparing to the entire dataset, dropping those rows is an option, but it would also cause around 5 percent of information loss.

Feature Engineering & Feature Selection

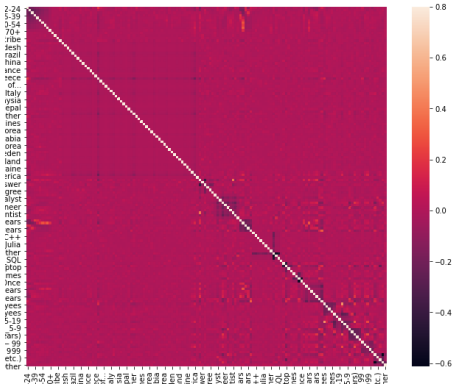
In terms of the exploratory data analysis, the figure below presents the highest importance 10 features. Since the most the features are categorical, the chi-square test is chosen for determined the feature importance.



Before conducting feature engineering, one effective method to reduce features is to investigate each question. It would be hard to fit a model if the number of features are high. Based on the question list, it can be seen that there are couple questions which are asking about people's opinion in 2 years. Since the goal of the model is trying to classify the data points into different current salary buckets, it can be assumed that those questions related to future have limited effect on the model. Therefore, Q26B to Q35B shall be dropped. Q24 should also be dropped since it is the target feature.

After that, a feature engineering technique can be applied to the dataset. Because most of the data type is categorical, feature encoding is an appropriate technique for classification model. Since those encoded columns which include “no selection” would cause some redundant information, these columns should be dropped as well.

Additionally, a correlation heat map can be carried out by calculating the correlation between each 2 columns. From the map, it can be seen that most the columns have around 0 correlation with each other. After selecting those columns with correlations smaller than 0.7, the number of features are finally reduced to 142.



Model Implementation & Model Tuning

After feature selection, the dataset can be applied to the ordinal logistic model using 10-fold cross validation. It is important to point out that it not necessary to scale the data, because the dataset is composed of 0 and 1. The metric used to measure the performance is accuracy. Although there are some imbalances of the number in each class, the number of each class is not highly skewed which might result a extremely high accuracy, Therefore, accuracy is an optional metric. The model accuracy across 10 folds are shown below, and it varies from 40.4% to 45.7%. While the average of the accuracy is 42.912%, the standard deviation of the accuracy is 1.638%. Based on this, the model has a relatively high bias. In order to find a best performance of the model, the hyperparameters need to be identified and tuned. In this ordinal logistic regression model, the hyperparameters are identical to the hyperparameters of the logistic regression model. Therefore, the hyperparameters are solver, penalty, and C value. The solver and C value are chosen to tune the model. The second figure blow shows the tuning result of the model. The best C value is found out to be 0.1, while the

solver is ‘liblinear’. Lastly, the accuracy of test set is found out to be 42.75%. Based on the training and test result, the model is not a well-performed model since it has a high bias and a high variance.

```
Fold 1: Accuracy: 42.135%
Fold 2: Accuracy: 41.854%
Fold 3: Accuracy: 42.135%
Fold 4: Accuracy: 44.101%
Fold 5: Accuracy: 45.225%
Fold 6: Accuracy: 41.011%
Fold 7: Accuracy: 40.449%
Fold 8: Accuracy: 43.179%
Fold 9: Accuracy: 45.71%
Fold 10: Accuracy: 43.319%
Average Score: 42.912%(1.638%)
```

Discussion

In terms of dataset, the biggest insight I gained from the dataset is that the location of the job truly affects the salary bucket. Based on the feature importance graph, the USA make a big contribution for identify the class. As shown above, the model performed not well on both training and testing set. The test accuracy is 0.2% lower than the training accuracy. Therefore, model is underfitting. The plot below shows the distribution of class. It can be seen that the distribution is skewed to the left, however, the classification model still cannot perform well on this dataset. One improvement would be the larger dataset. In this case, the training accuracy would be increasing. After tuning hyperparameters, the test accuracy would also be improved.