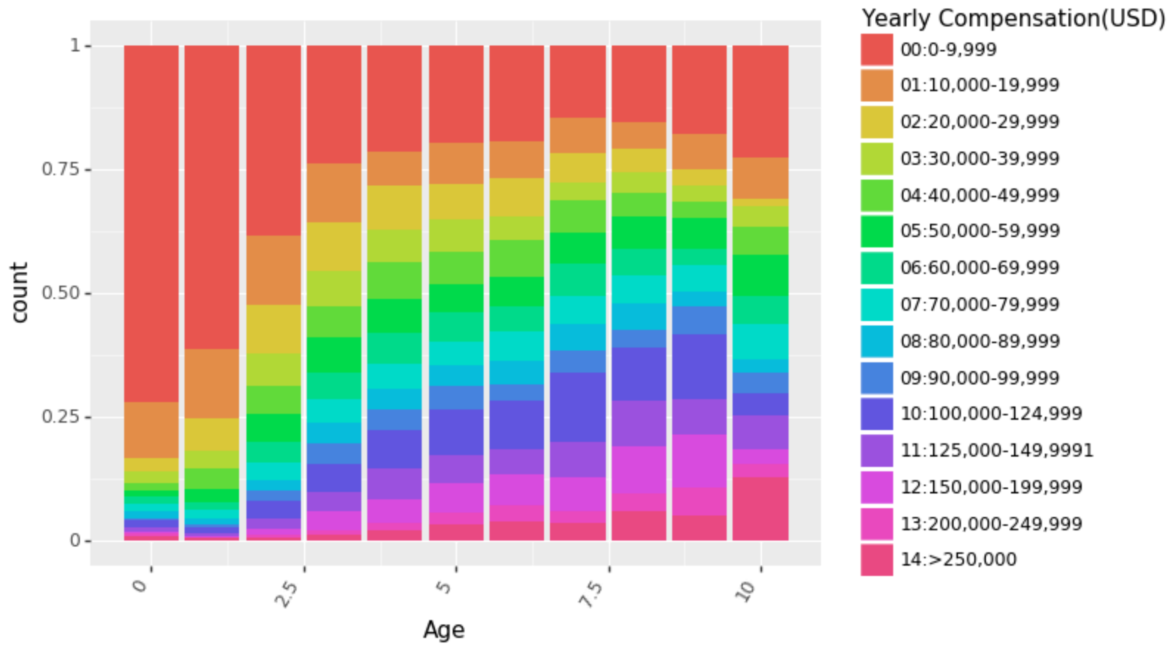
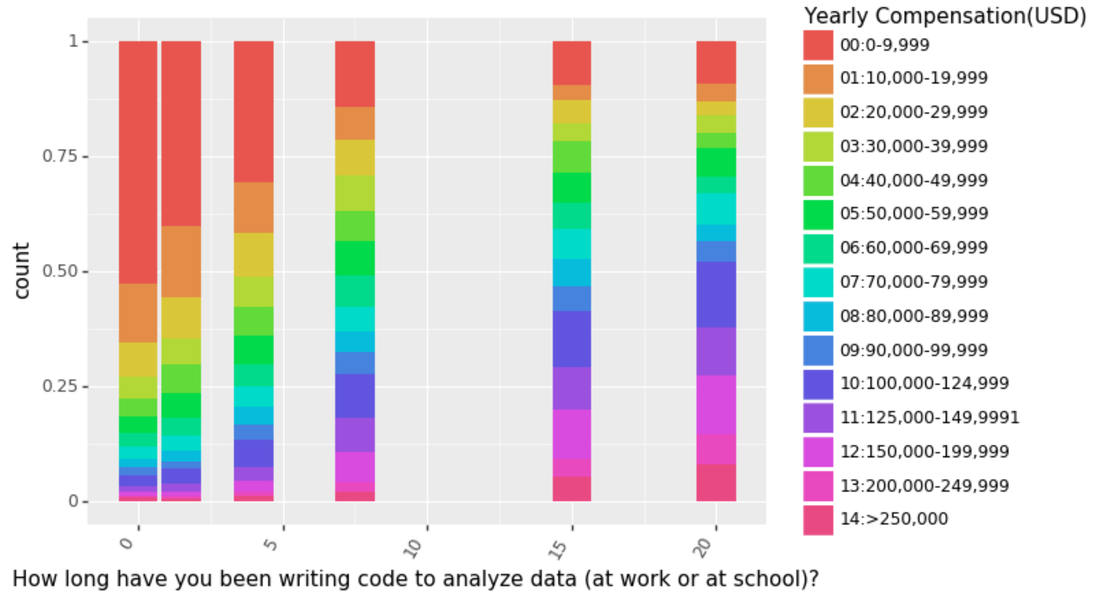
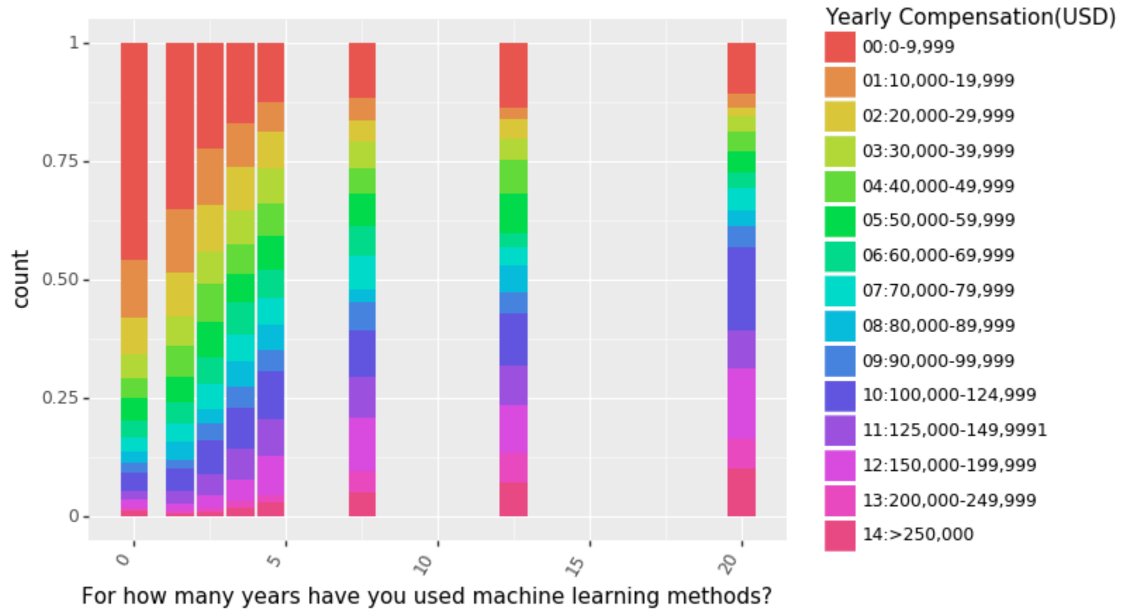
Exploratory data analysis

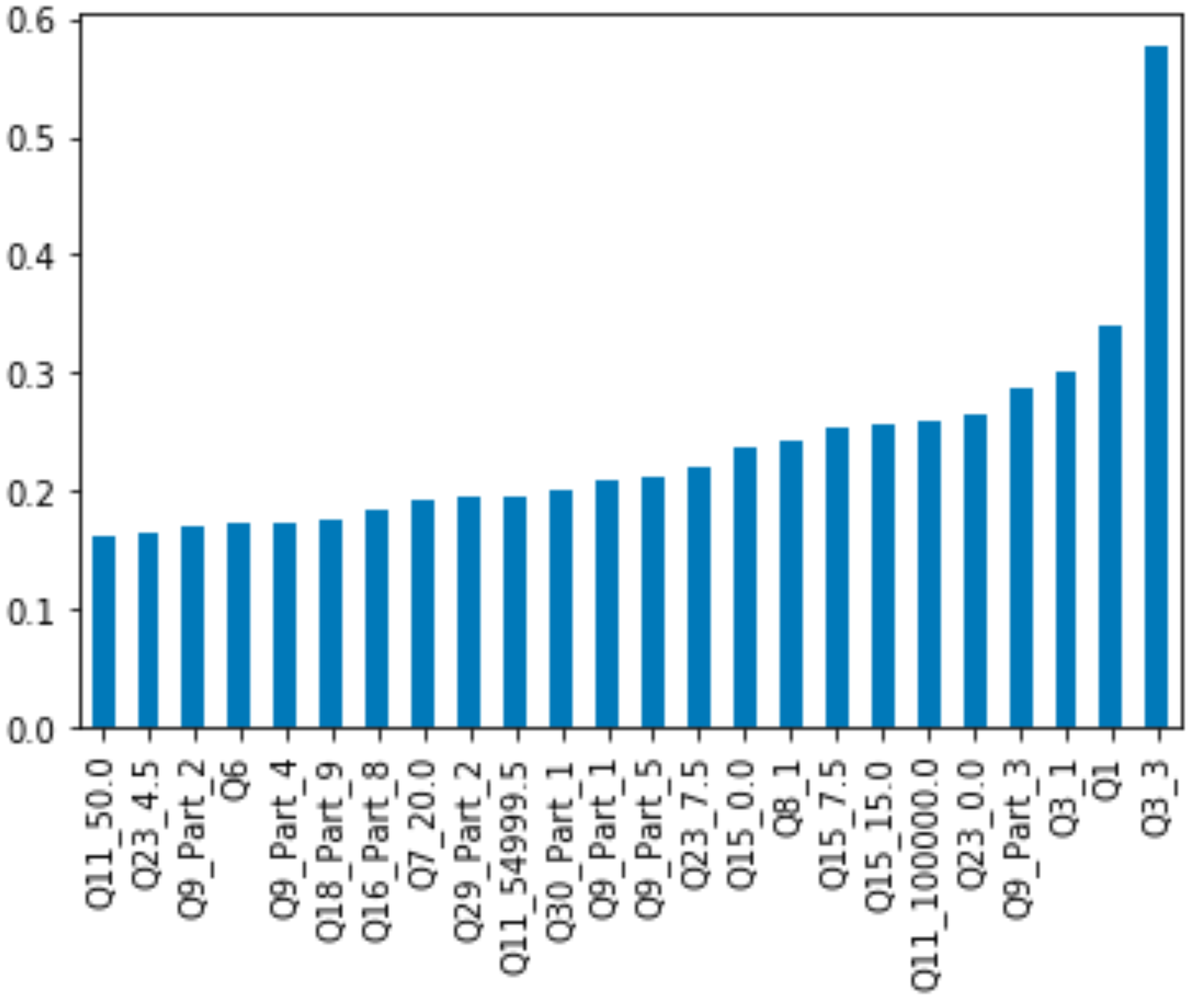






The trends of three figures imply dataset has continue features. And the dataset might be linearly separable. Thus, apply logistic regression model is possible.

Features



The figure shows correlation coefficients response with yearly compensation. The most important feature is Q3_3, which represents the people currently reside in US or not. However, there's no more features with coefficient higher than 0.5. Therefore, we need do more feature engineering like combine features.

Results

10-fold cross validation

1th	fold	accuracy:37.48%
2th	fold	accuracy:34.00%
3th	fold	accuracy:36.31%
4th	fold	accuracy:39.38%
5th	fold	accuracy:37.85%
6th	fold	accuracy:38.77%
7th	fold	accuracy:38.15%
8th	fold	accuracy:36.46%
9th	fold	accuracy:35.08%
10th	fold	accuracy:36.62%

Average accuracy: 37.01%

Accuracy std: 1.57%

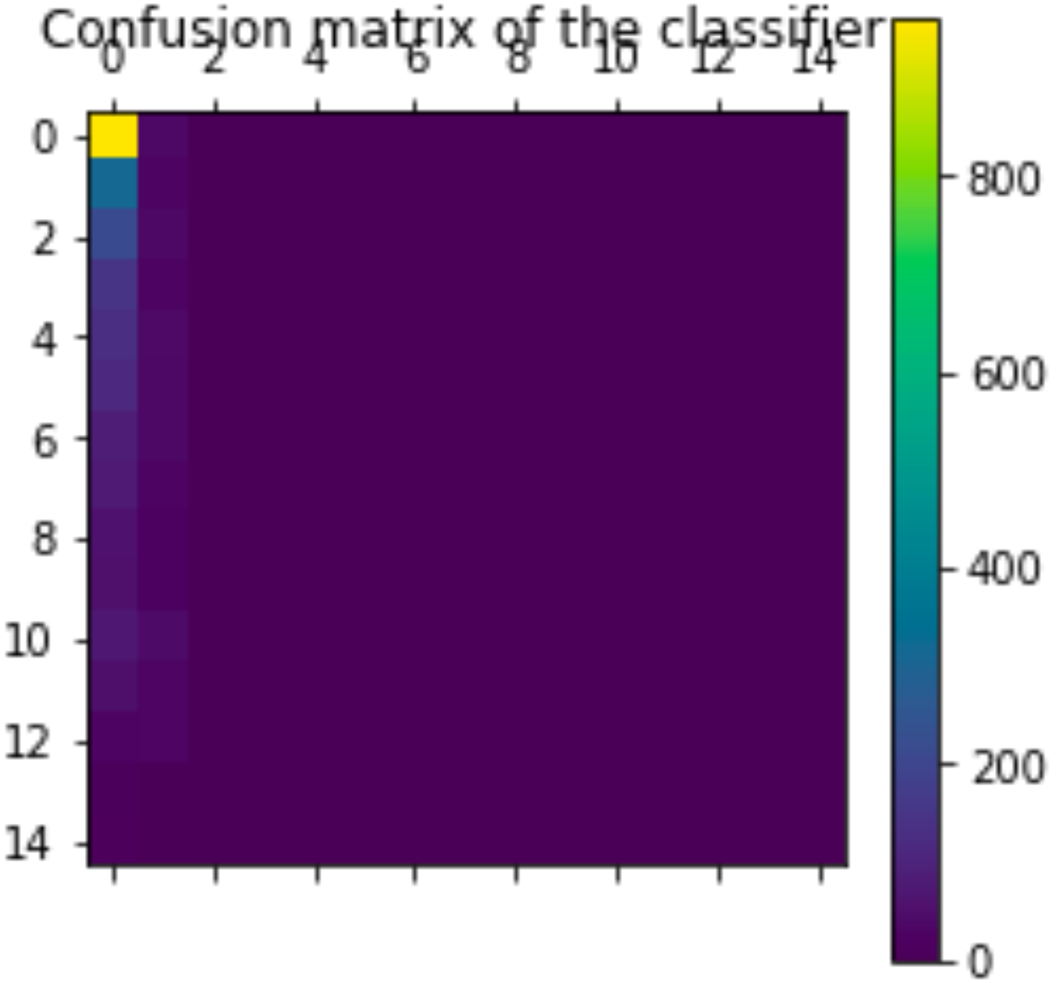
The accuracy of cross validation is poor. As we discussed before, the features are not very good for logistic regression.

Result

	C	Average accuracy	Accuracy Std
1	0.01	37.38%	1.44%
2	1	37.01%	1.57%
3	100	37.29%	1.41%

Then, I tuned model by varying parameter C. However, the results did not change much. So, the previous bad results are not caused by parameter C. The main reason is features do not meet the requirements of logistic regression.

Results



Confusion matrix implies that the model output “0”/ lowest income at most of times. Hence, the model was not well-trained