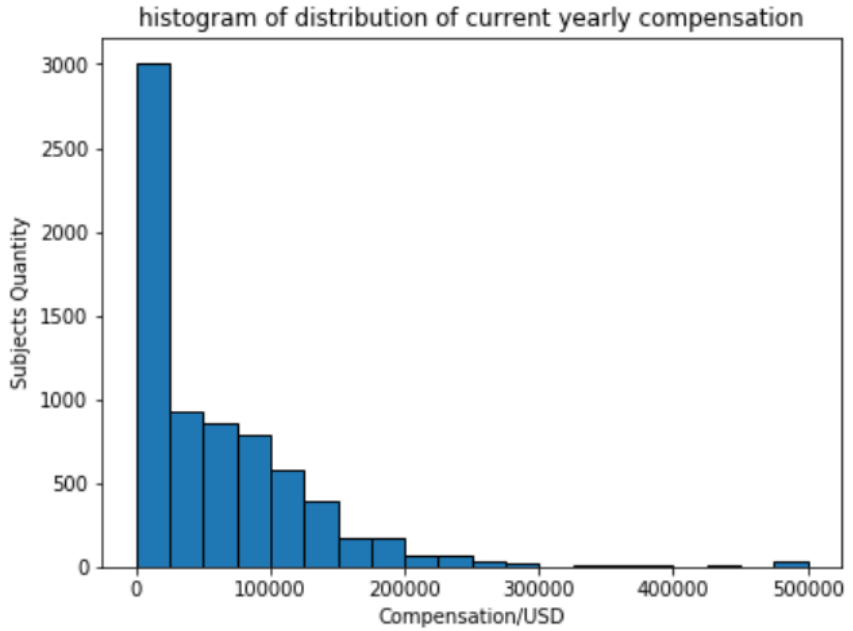
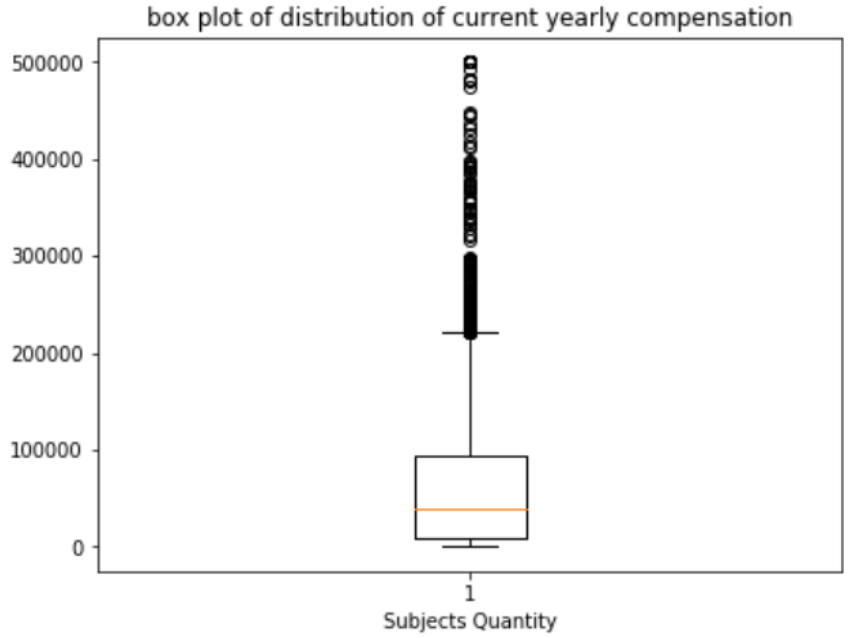
The distribution of the label values

The label can be consider as a very important feature of our dataset. Therefore I firstly visualize the distribution of the label values.

- We can see that most of the labels are in range (0,100,000) as yearly compensation. Meanwhile, there are only small portion of people whose yearly compensation are higher than 200,000 USD.

- This distribution indicate that our training model may only have good prediction ability on lower compensation instead of higher compensation because we don't have much data for higher compensation subjects.

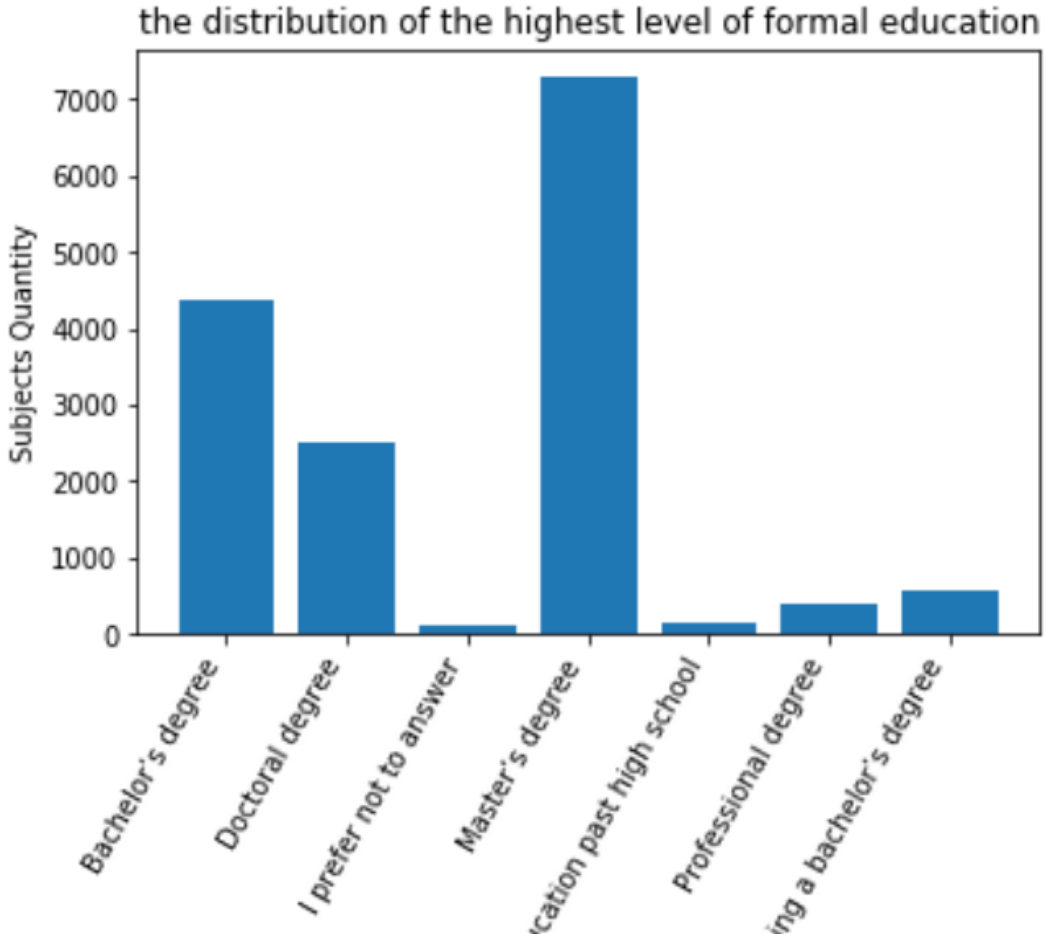




The distribution on level of eduction

Then I want to check wheather data analysis has a high enter level of eduction. So I visualize the “level of formal education”

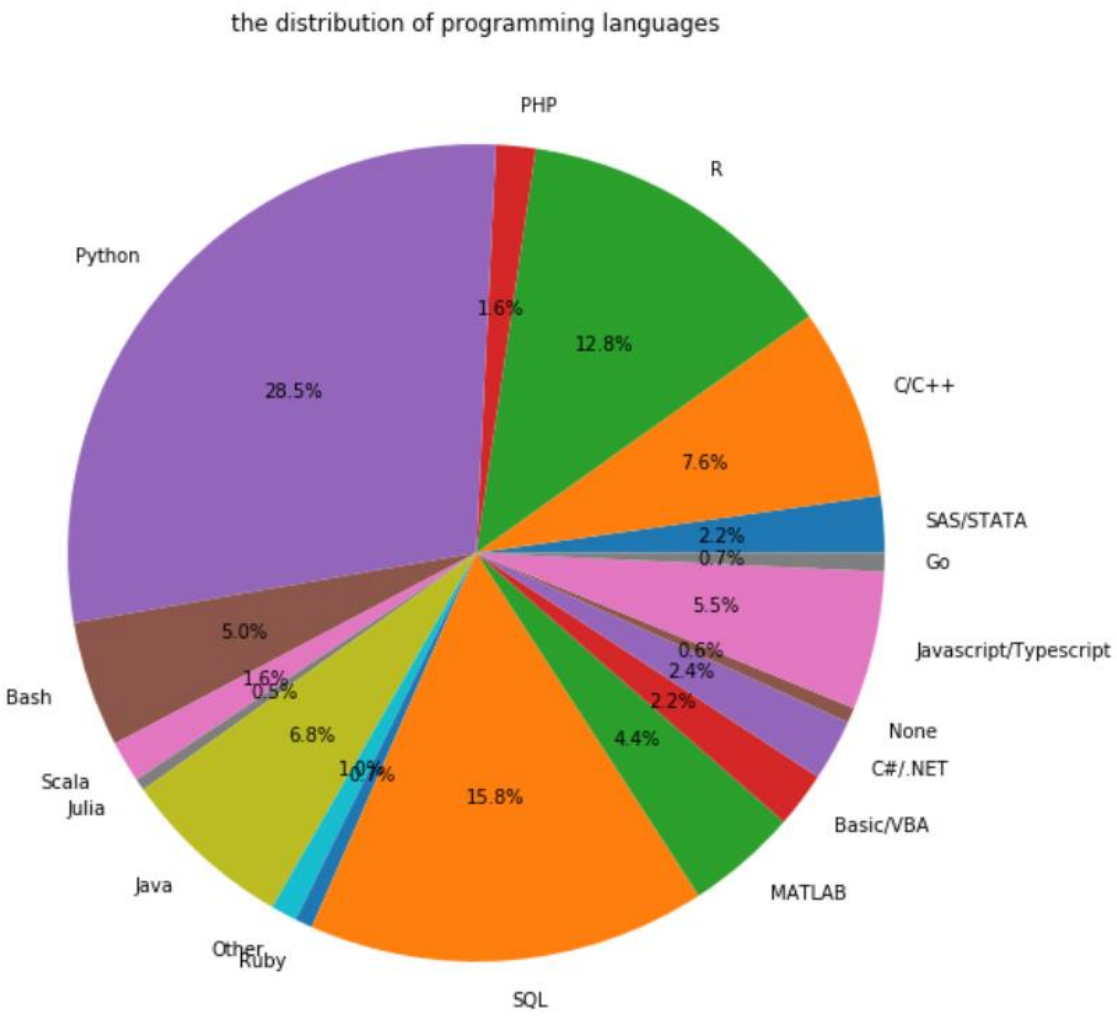
- In the plot, we can see that most of data analysists has at least Bachelor degree or high. Among them, almost half of them has a Master degree.
- So this indicate it has a relatively high enter level to have a job in data analysis. And master degree graduates are most prefer by companies compared with other degrees.



Bachelor’ s degree	4383
Doctoral degree	2522
I prefer not to answer	130
Master’ s degree	7286
No formal education past high school	135
Professional degree	384
Some college/university study without earning a bachelor’ s degree	589

The distribution programming Language

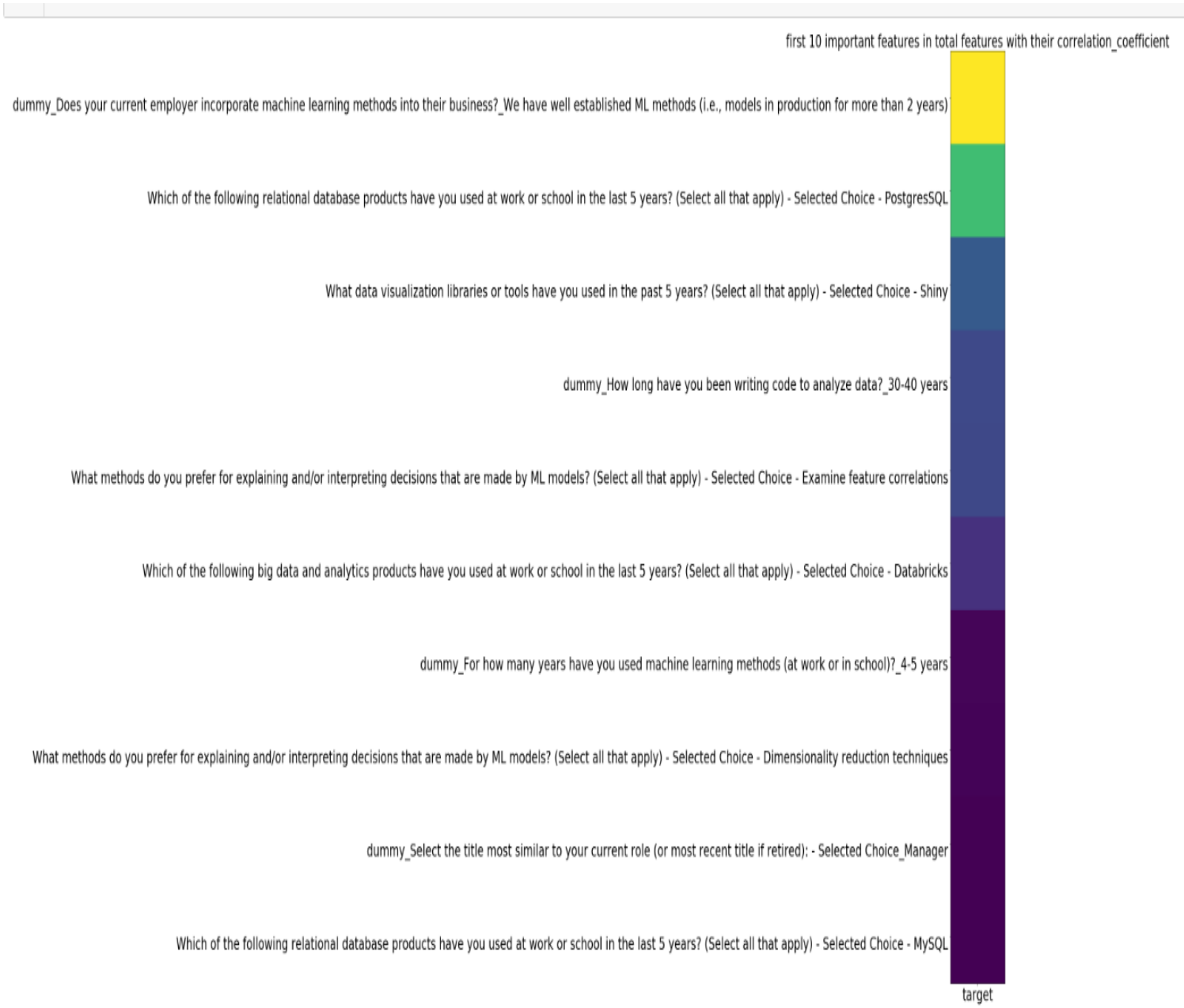
- The data shows that python is a most preferred programming language for data analysis. 28.5% subjects use it to do the job.
- The second places of most popular programming language is SQL(15.8%). This indicate data storing are also one of important features of data science
- Although python, R, and SQL are the most popular one, there are still lots of choice you can make for learning data science.



Feature Importance

- I use correlation matrix between the label vectors and features to indicate feature importance

- Since I have almost 800 features in data set, I can't show them all here. The plot shows the first 10 important features



Performance on Best Model

	Bias	R2	RMSE	Variance
Train	6.231759e+08	0.823512	24963.491782	2.467954e+09
Test	1.239513e+09	0.646788	35206.721995	2.333225e+09

1. The gradient boosting regression model results are shown in form above. The results indicate a relatively high bias and high variance model. R2 score looks good on training set, but drop a lot on test set..The RMSE on test set is35206.721995 while only 24963.491782 on training set.

2. Is it overfitting or underfitting?

From above form, we can see that the test bias, test R2 score and test RMSE are much higher than the train bias, train R2 score and train RMSE, respactively. That indicate overfitting may occur on the training processing. Because both indicators point out that training performance better that testing performance. That means our gradient boosting regression model fitting better on training data, but when we give new data, such as test data to it, its accuracy drops. Our model learning to much details about training data, lead itself not generalize enough for new data.