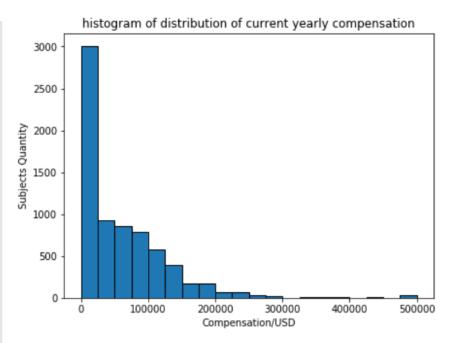
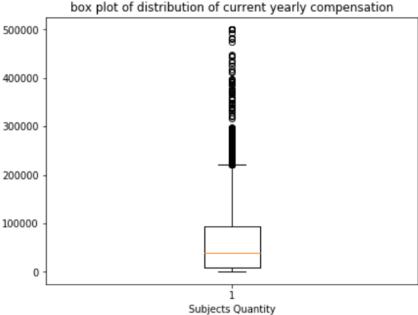
The distribution of the label values

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.

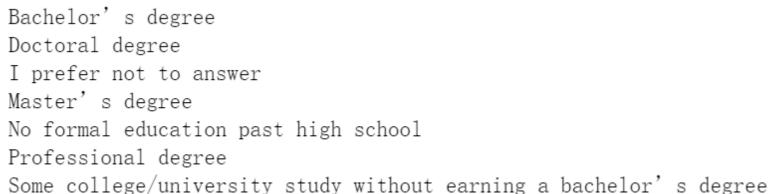


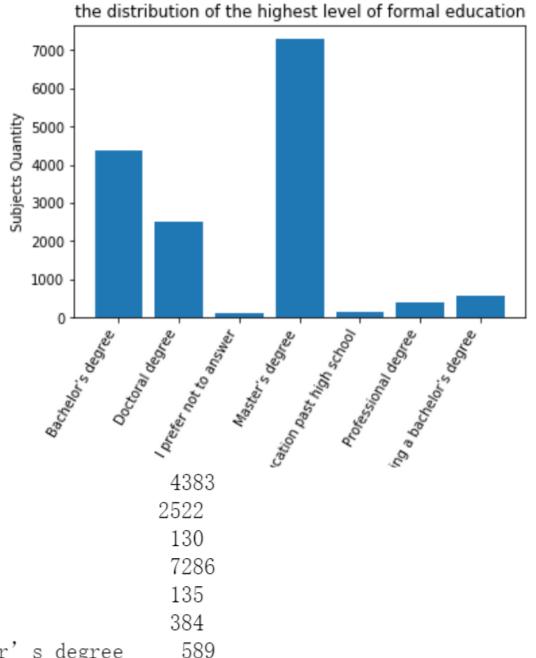


The distribution on level of eduction

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.

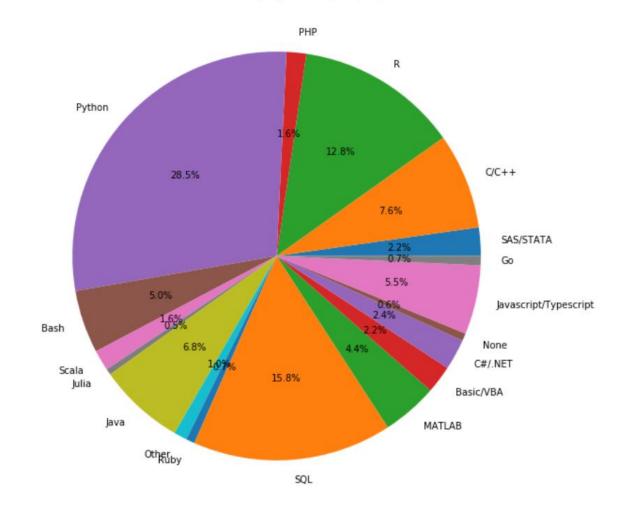




The distribution programming Language

- 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.

the distribution of programming languages



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

first 10 important features in total features with their correlation_coefficient

dummy_Does your current employer incorporate machine learning methods into their business?_We have well established ML methods (i.e., models in production for more than 2 years)

Which of the following relational database products have you used at work or school in the last 5 years? (Select all that apply) - Selected Choice - PostgresSQL

What data visualization libraries or tools have you used in the past 5 years? (Select all that apply) - Selected Choice - Shiny

dummy_How long have you been writing code to analyze data?_30-40 years

What methods do you prefer for explaining and/or interpreting decisions that are made by ML models? (Select all that apply) - Selected Choice - Examine feature correlations

Which of the following big data and analytics products have you used at work or school in the last 5 years? (Select all that apply) - Selected Choice - Databricks

dummy_For how many years have you used machine learning methods (at work or in school)?_4-5 years

What methods do you prefer for explaining and/or interpreting decisions that are made by ML models? (Select all that apply) - Selected Choice - Dimensionality reduction techniques

dummy Select the title most similar to your current role (or most recent title if retired): - Selected Choice Manager

Which of the following relational database products have you used at work or school in the last 5 years? (Select all that apply) - Selected Choice - MySQL

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 is 35206.721995 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.