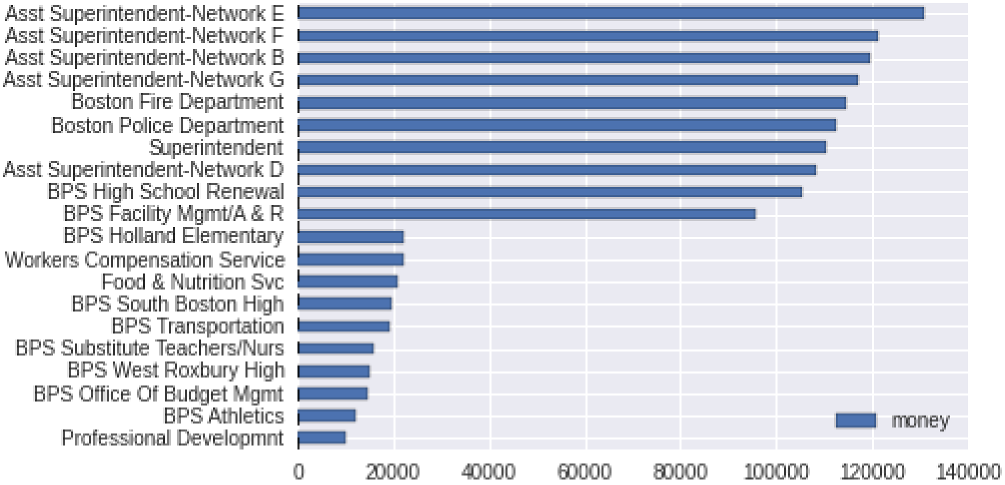
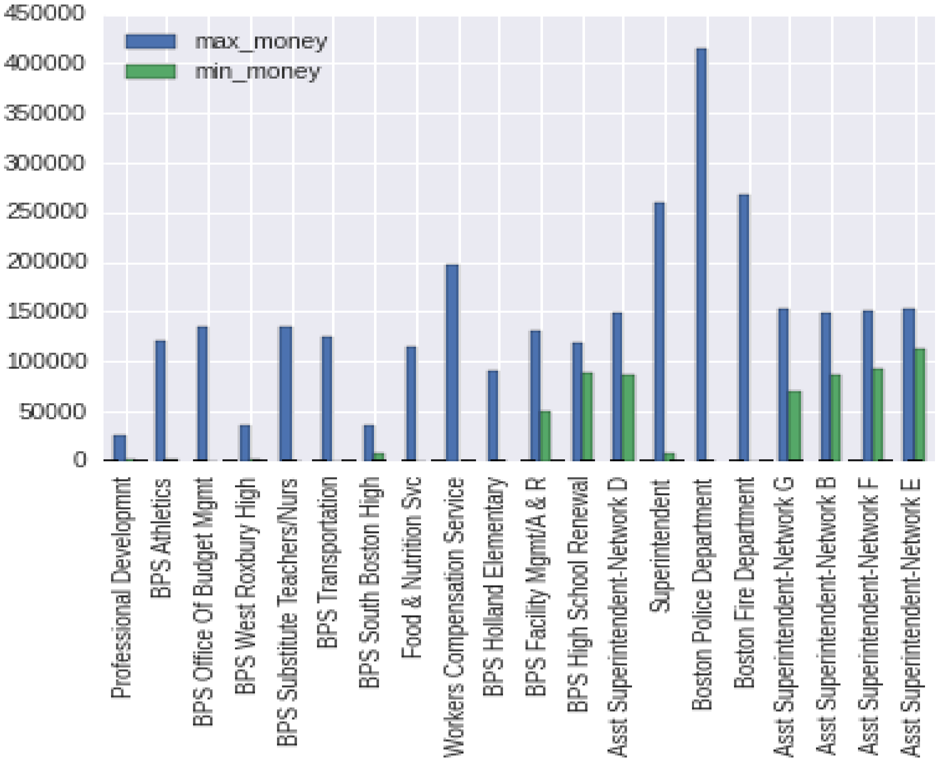
Homework 6

Initial data analysis:

In this project, we use the 2014 annual wage reporting datasheet. In this datasheet, it has columns: people's names, occupations, department and different salary components: retroactive salary, injury salary, others, regular salary and so on. We figure out that for different departments, there are great differences among people's salary as shown below:



In addition, we also find that even in each industry, different position has a great difference in amount of money they got.



First assumption:

We can witness that different industry has different components of salary. Some industries have a comparatively high threshold but a small space to increase the salary while others are vice versa like Boston Police Department. In the next step, we want to go to these departments to see what kind of positions will earn much more than any other position.

Second assumption:

Furthermore, different people tends to have different choices for industries. Later, we want to find that what part of salaries contribute more to the total salaries and based on that we infer the different characteristics. For example, if in the industry, overtime bonus is a main drive to the high salary, which means that this work will be so tired. In addition, if people working in these industries tend to get money because of injury, which means that this industry is risky. Thus, we will use the logistic regression to analyze the different characteristics of industries.

Third assumption:

Later, we will go through all the data in the dataset to figure out which one is more important to get a higher salary, the title level or the relevance to the main function of this industry.

Forth assumption:

As shown in the graph, we find that the working industry does matter to the amount of salary you obtain. Then we find out what kind of requirements can qualify you to go to the well-paid industry. In order to prove that, we will go to another datasheet 'Occupational Employment and Job Openning data'. In which, it has columns: industry names, working experience, academic level and training requirements. Based on these data, we will draw our conclusion.

Fifth assumption:

If time permits, still based on these data, we will find which one is important, academic level or working experience.