Application of model selection on factors affecting salaries of data scientists

Course:
STAT 306 201

Group: B2

Team members:
Wang Li 21984646
Cong Cheng 31070519
Alex Li 46338117
Tina Fu 86675303

Instructor:
Dr. Bruce Dunham

Date:
April 8, 2022

Introduction

In contemporary society, almost everything can be digitized. Thus, the importance of data related work has significantly increased. In addition, this kind of work is directly related to what we are studying. When we start our career as young data scientists, it's important for us to know what factors can impact the compensation. Having this information in advance can help us evaluate possible salary levels and have more negotiating power when we get hired. Therefore, we investigated which factors affect the salary of data related jobs.

We analyzed a Kaggle dataset (https://www.kaggle.com/nikhilbhathi/data-scientist-salary-us-glassdoor) gathered in 2021 by scraping from Glassdoor.com using Selenium for the search term "data scientist." The dataset included information about the job, including salary, job description, job title, and programming languages required, as well as information about the company, including age, location, size, and sector. We decided to start our analysis using company age, size, revenue, Python, SQL, and Excel, since we expect that these might have more explanatory power than other variables in the dataset.

We might expect companies that are older, larger, or have more revenue to also have more sources of robust cash flow that are required to pay larger salaries.
We decided to focus on Python, SQL, and Excel, since these three programming languages are commonly used in statistical study. Whether or not a job lists these as requirements may impact the associated salary.

Data Selection

The response variable is the average yearly salary of data scientists in thousands of dollars.

The explanatory variables we started with include:
Age (years): This is the company age, expressed as a continuous variable.
Company size (number of employees): This variable is categorized into different ranges.
Revenue (USD): Total revenue of the company per year. This is similar to the company size and is also split into different ranges and is also a categorical variable.
Python, SQL, and Excel: These are binary dummy variables that indicate whether or not they were required for the job.

We wanted to find which factors most influence data scientist salaries among company age, revenue, size, and three major programming languages or techniques, Python, SQL, and Excel. We wanted to explain the impact of those variables rather than find the best prediction model. Knowing which of these factors affects salary can provide guidance for data scientists about which companies to apply to and what skills they may want to have.

Data Transformations

Below is the summary of the raw data:

```
> summary(data)
      Age          Avg.Salary.K.                Size
 Min.   : -1.00   Min.   : 15.5   1001 - 5000  :150
 1st Qu.: 12.00   1st Qu.: 73.5   501 - 1000   :134
 Median : 25.00   Median : 97.5   10000+       :130
 Mean   : 47.52   Mean   :101.5   201 - 500    :117
 3rd Qu.: 60.00   3rd Qu.:122.5   51 - 200     : 94
 Max.   :277.00   Max.   :254.0   5001 - 10000 : 76
                                  (Other)      : 41
                          Revenue        Python            sql              excel
Unknown / Non-Applicable      :204   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
$10+ billion (USD)            :124   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
$100 to $500 million (USD)    : 91   Median :1.0000   Median :1.0000   Median :1.0000
$1 to $2 billion (USD)        : 60   Mean   :0.5283   Mean   :0.5121   Mean   :0.5229
$500 million to $1 billion (USD): 57   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
$50 to $100 million (USD)     : 46   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
(Other)                       :160
```

Figure 1: summary of raw data

The summary of the raw data above does not show it, but about ⅓ of the companies did not list their age. Only about 5% of the companies had their size listed as Other. Almost 20% of the companies did not list revenue, with about another 15% categorized as Other.
Python, SQL, and Excel are required in about half the jobs.

To sanitize the data (figure 1.1), we dropped the -1 values, which indicated failure to retrieve the values, from all parameters.

We releveled revenue, combining 13 levels into the following levels:
Low: less than $1 million to 25 million
Med: 25 to 1 billion
High: 1 billion to 10+ billion

We also releveled Size from 7 levels into the following levels:
Small: 1-200
Med: 201-5000
Large: 5001+

Below is the summary of our transformed data:

```
> summary(data)
      Age          Avg.Salary.K.      Size     Revenue        Python           sql             excel       
 Min.   :  2.00   Min.   : 15.5   Small: 63   High:241   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.: 18.00   1st Qu.: 71.5   Large:190   Low : 47   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median : 39.00   Median : 95.0   Med  :259   Med :224   Median :1.0000   Median :1.0000   Median :1.0000  
 Mean   : 60.56   Mean   :100.1                          Mean   :0.5254   Mean   :0.5215   Mean   :0.5098  
 3rd Qu.: 82.00   3rd Qu.:121.0                          3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :277.00   Max.   :254.0                          Max.   :1.0000   Max.   :1.0000   Max.   :1.0000  
> |
```

Figure 1.1: summary of sanitized data

As expected, by removing the -1 values, the quartiles for age have increased.

Analysis

We then used backward selection, removing the variable with the least significant coefficient in a stepwise manner, to arrive at our best model, which includes the parameters Age, Python, and Excel. The model summaries follow.

We begin with the summary of the full model:

```
> full_with_age <- lm(Avg.Salary.K.~.,data=data)
> summary(full_with_age)

Call:
lm(formula = Avg.Salary.K. ~ ., data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-74.949 -23.396  -5.762  20.689 125.760 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  97.80055    8.07049  12.118   <2e-16 ***
Age           0.02314    0.03119   0.742   0.4585    
SizeLarge     0.21414    7.65070   0.028   0.9777    
SizeMed      -9.74602    5.85621  -1.664   0.0967 .  
RevenueLow    3.96624    7.68286   0.516   0.6059    
RevenueMed  -10.14828    4.71837  -2.151   0.0319 *  
Python       27.70351    3.19115   8.681   <2e-16 ***
sql          -2.80784    3.24466  -0.865   0.3872    
excel        -6.55065    3.01159  -2.175   0.0301 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.9 on 529 degrees of freedom
Multiple R-squared:  0.2091,    Adjusted R-squared:  0.1971 
F-statistic: 17.48 on 8 and 529 DF,  p-value: < 2.2e-16
```

We can see from the result that the Size variable has the largest p-value on the “Large” level. Therefore, we get rid of the “Size” variable and refit the model:

```
> fit1 <- lm(Avg.Salary.K.~.-Size,data=data)
> summary(fit1)

Call:
lm(formula = Avg.Salary.K. ~ . - Size, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-73.082 -23.221  -7.914  19.461 127.204

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  94.70019    4.41259  21.461  < 2e-16 ***
Age           0.03758    0.03005   1.251   0.2116
RevenueLow    5.45092    5.32914   1.023   0.3068
RevenueMed  -15.24380    3.55169  -4.292 2.11e-05 ***
Python       27.99902    3.17789   8.811  < 2e-16 ***
sql          -3.84574    3.23522  -1.189   0.2351
excel        -6.74001    3.00411  -2.244   0.0253 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.06 on 531 degrees of freedom
Multiple R-squared:  0.1988,    Adjusted R-squared:  0.1897
F-statistic: 21.96 on 6 and 531 DF,  p-value: < 2.2e-16
```

Using the same logic, we then remove the “Revenue” variable and refit the model:

```
> fit2 <- lm(Avg.Salary.K.~.-Size-Revenue,data=data)
> summary(fit2)

Call:
lm(formula = Avg.Salary.K. ~ . - Size - Revenue, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-67.797 -24.696  -6.284  23.375 130.496

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 85.56689    3.26753  26.187  < 2e-16 ***
Age          0.07507    0.02645   2.838  0.00471 **
Python      29.75450    3.23295   9.204  < 2e-16 ***
sql         -4.07633    3.31680  -1.229  0.21962
excel       -6.39879    3.07540  -2.081  0.03794 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.92 on 533 degrees of freedom
Multiple R-squared:  0.1545,    Adjusted R-squared:  0.1482
F-statistic: 24.35 on 4 and 533 DF,  p-value: < 2.2e-16
```

Similarly, we deleted the “sql” variable, resulting in the reduced model with all variables being significant:

```
> fit3 <- lm(Avg.Salary.K.~.-Size-sql-Revenue,data=data)
> summary(fit3)

Call:
lm(formula = Avg.Salary.K. ~ . - Size - sql - Revenue, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-66.033 -23.576  -6.559  23.029 132.651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 84.20121    3.07425  27.389  < 2e-16 ***
Age          0.08046    0.02610   3.083  0.00216 **
Python      28.37846    3.03433   9.352  < 2e-16 ***
excel       -7.09325    3.02449  -2.345  0.01938 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.93 on 534 degrees of freedom
Multiple R-squared:  0.1521,    Adjusted R-squared:  0.1474
F-statistic: 31.94 on 3 and 534 DF,  p-value: < 2.2e-16
```

We then use the "regsubsets" command in R to select the "best" model for each size (size measured in terms of number of variables):

```
> ss2$which
  (Intercept)   Age Python excel
1        TRUE FALSE   TRUE FALSE
2        TRUE  TRUE   TRUE FALSE
3        TRUE  TRUE   TRUE  TRUE
```

The "regsubsets" command selects the reduced model as one of the "best" models. By comparing the C_p values to the number of parameters (including intercept parameter), we decided that the reduced model would be most appropriate here (figure 2), since this model has the lowest Cp value and it is the closest to the number of parameters in the model.

```
> ss2$cp
[1] 13.836197  7.500309  4.000000
```

From the summary of our "best" model, we predict that a one year increase in company age increases average salary by 80.46 USD. A job that requires Python is predicted to have a salary that is $28,378.46 greater than one that does not and a job that requires Excel is predicted to have a salary that is -$7,093.25 less than one that does not. This might result from the fact that jobs that list Excel as a requirement may involve less advanced data analysis skills and thus earn lower salaries than ones that do not and jobs that list Python as a requirement may have the reverse apply.

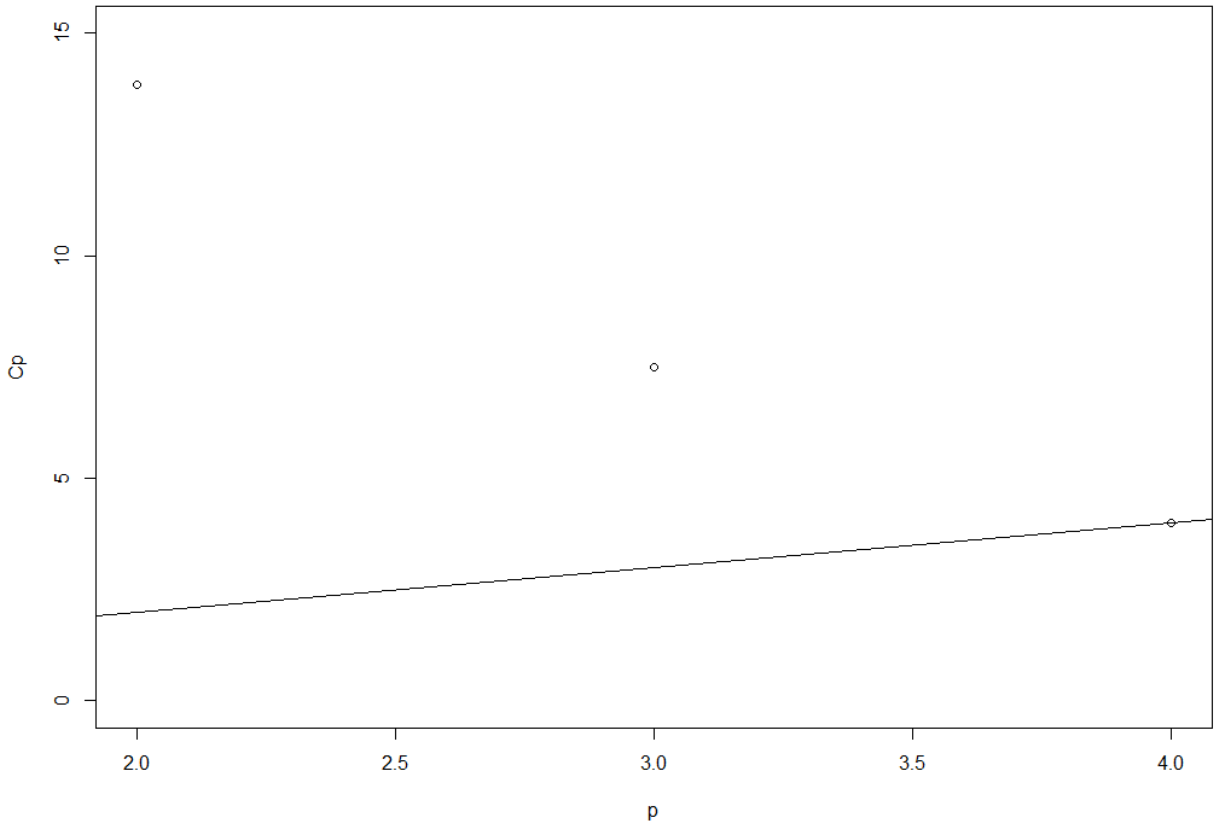

Figure 2: p vs. C_p

Relationship between variables

We also explored some relationships between explanatory variables to further support our decisions to remove certain variables from the full model.

The boxplot of company age categorized by company size showed a rather strong relationship between the two variables. In general, larger companies are also older than smaller ones (figure 3). This phenomenon could potentially explain the insignificance of the coefficients in the full model, and therefore served as one of the major reasons why we deleted the “Size” variable.

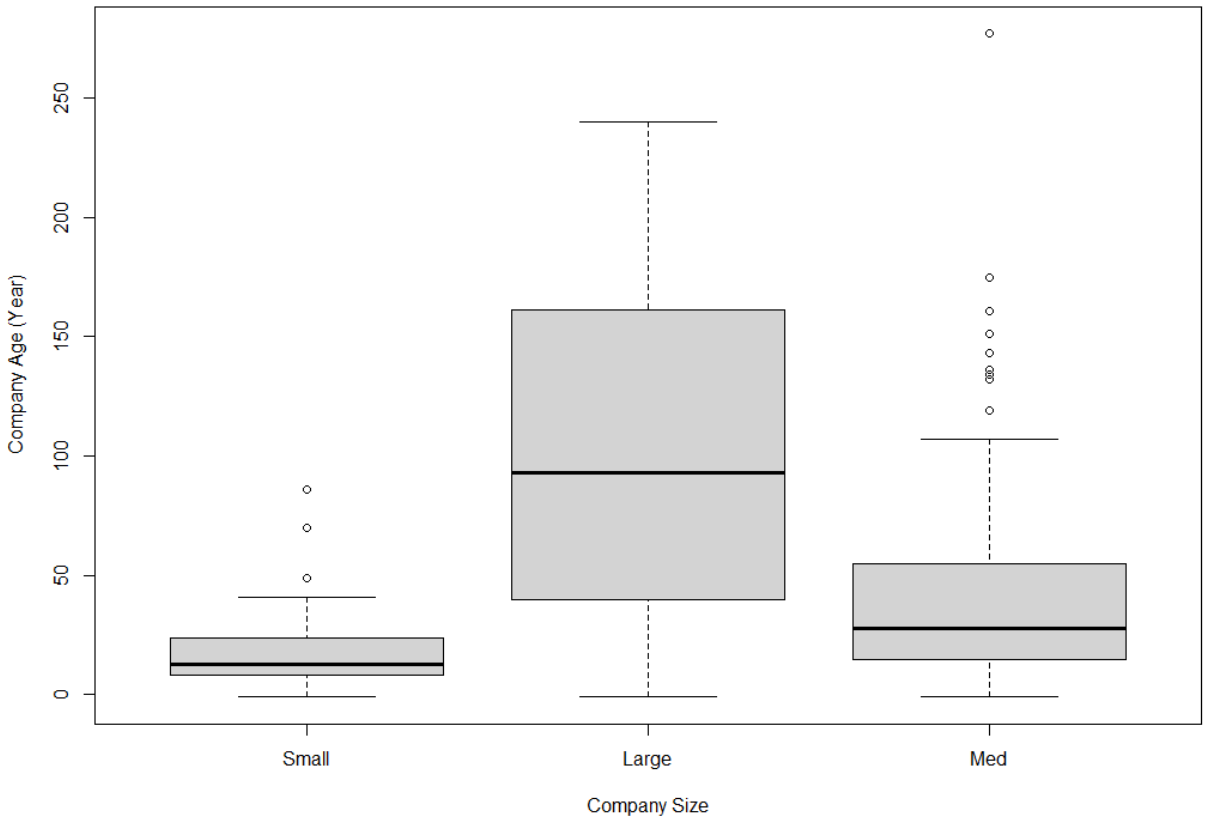

Figure 3: Company Age vs. Company Size

Similarly, we also found company age to vary with company revenue (figure 4). Following the logic stated above, we also removed the “Revenue” variable.

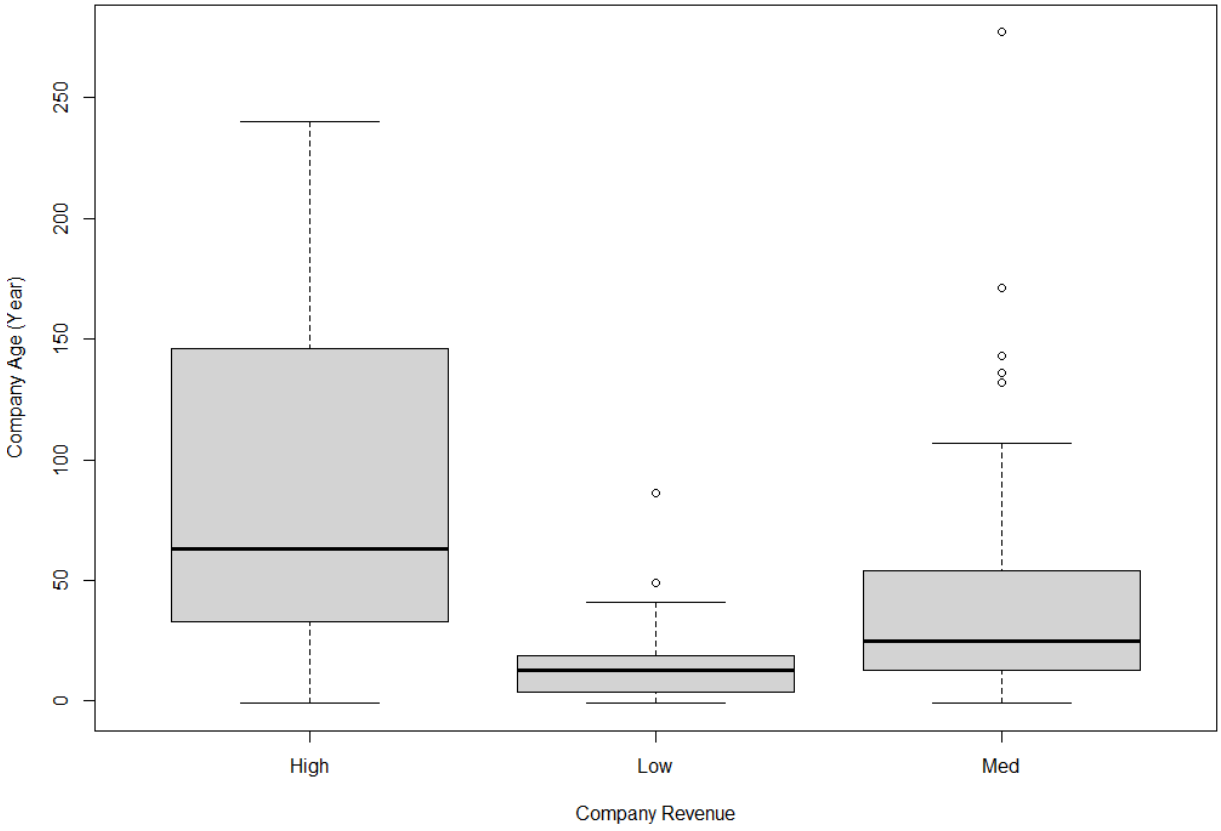

Figure 4: Company Revenue vs. Company Age

Conclusion

Among the variables we selected, we found that company age, Python, and Excel were most predictive of average salary, with company age and Python increasing average salary and Excel decreasing it.

There are some limitations of this report. First of all, the explanatory variables in the raw data are limited, which resulted in the small R^2 value of the models we fitted. In reality, there are more factors that affect the salaries of a company. Job candidates should not limit themselves to the variables discussed in this study in their job searching process. It is also difficult to relate the individual skills required for data scientist jobs, Python and Excel, to average salaries of all data scientists in a company and the company age. Since there are a large number of data scientists in each company with different job titles, the average salary is affected by the upper salary and lower salary in each company. When performing backwards selection, some levels of size and

revenue had insignificant P-values while others had significant P-values. We removed the entire variable to fit the model, which may have resulted in a non-representative model. In addition, data points with variable status “unknown” were dropped to clean the data, which may have caused loss of valuable information.