****

**MIDTERM ASSIGNMENT**

Fundamental of Data Analytics

Group 6 – K20416C

Subject: Linear Regression

Data: CEO\_Salary

Members:

1. Trần Triệu Vy – K204162011

2. Bùi Thu Vân – K204162007

3. Lê Chí Hào – K204161988

4. Lê Thị Thùy Linh – K204161990

5. Hoàng Nguyễn Quỳnh Giang – K204161987

[**I.  State the requirements and objectives of the problem, develop research questions, propose a process - a diagram of research implementation.** 3](#_Toc88236836)

[**1.**  **Requirements and objectives of the problem:** 3](#_Toc88236837)

[**2**  **Research questions:** 3](#_Toc88236838)

[**3**  **Propose a processing diagam of research implementation on salary:** 3](#_Toc88236839)

[**II. Define Metrics/Variable** 4](#_Toc88236840)

[**1.** **The relationship between pcsalary and salary:** 5](#_Toc88236841)

[**2.** **The relationship between company performance and CEO salary:** 5](#_Toc88236842)

[**3.** **The relationship between Sales and CEO salary:** 5](#_Toc88236843)

[**4.** **The relationship between type, size of company and salary:** 6](#_Toc88236844)

[**5.** **The relationship between Issalary and Issale:** 6](#_Toc88236845)

[**III. Exploratory data analysis depedent variable with Descriptive Analytics method.** 6](#_Toc88236846)

[**1.** **Salary or Log(salary)?** 7](#_Toc88236847)

[**2.** **Data Size** 9](#_Toc88236848)

[**3.** **Statistics of each field (and histogram)** 10](#_Toc88236849)

[**3.1.** **Salary (Million dollars)** 11](#_Toc88236850)

[**3.2.** **Pcsalary** 13](#_Toc88236851)

[**3.3.** **Sales (Million dollars)** 16](#_Toc88236852)

[**3.4.** **Roe** 18](#_Toc88236853)

[**3.5.** **Pcroe** 21](#_Toc88236854)

[**3.6.** **Ros** 23](#_Toc88236855)

[**3.7.** **lsalary** 26](#_Toc88236856)

[**3.8.** **lsales** 28](#_Toc88236857)

[**3.9.** **indus** 31](#_Toc88236858)

[**3.10.** **finance** 31](#_Toc88236859)

[**3.11.** **consprod** 32](#_Toc88236860)

[**3.12.** **utility** 33](#_Toc88236861)

[**4.**  **Typical correlation between data fields** 34](#_Toc88236862)

[**5.** **Outlier points** 41](#_Toc88236863)

[**6.** **Box Plot:** 41](#_Toc88236864)

[**IV. Handle Outliers** 45](#_Toc88236865)

[**V. Liner Regression Model** 47](#_Toc88236866)

[1. Model (1) – 5 variables: salary, lsalary, lsales, sales, roe 47](#_Toc88236867)

[2. Model (2) – 6 variables: salary, pcroe, pcsalary, lsalary, lsales, sales 49](#_Toc88236868)

[3. Model (3) – 7 variables: salary, pcroe, pcsalary, lsalary, lsales, sale, roe 50](#_Toc88236869)

[4. Model (4) – 8 variables: salary, pcroe,pcsalary,lsalary,lsales,sale,roe,ros 51](#_Toc88236870)

[**VI. Linear Regression Model with Categorical Variables (or called Dummy Variables)** 52](#_Toc88236871)

[References: 55](#_Toc88236872)

# **I.  State the requirements and objectives of the problem, develop research questions, propose a process - a diagram of research implementation.**

**1.    Requirements and objectives of the problem:**

-       Explain the relationship between factors affecting remuneration.

-       Listen to the data

-       Catch mistakes in data

-       See patterns in the data

-       Find violations of statistical assumptions

-       Generate hypotheses

**2    Research questions:**

-       How do these factors affect CEO compensation?

-       Which is the biggest impact factor?

-       How does the correlation of these factors affect remuneration?

-       Proposed process - research implementation diagram:

**3 Propose a processing diagam of research implementation on salary:**

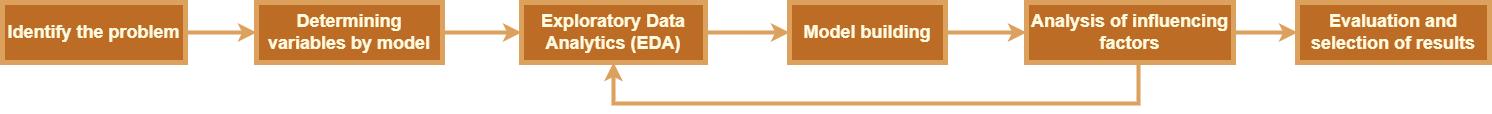


Figure 1: Processing diagram

* Identify the problem: Analyze problem requirements and objectives, ask questions and propose research processes
* Determining variables by model: Define variable, clean variable and analyze its influence on salary
* Exploratory Data Analytics (EDA): This EDA step gives us a first look at the data. You need to have a certain feel for what you have in hand before you have modeling strategies. EDA helps you visualize the complexity of the problem and outlines the first steps to take.
* Model building and Analysis of influencing factors: Build the models and explain the relationship between independent variables and dependent variable. Build the model with dummy variables and quantitve variables.
* Evaluation and selection of results: Select the most suitable model and evaluate the effectiveness of the model

# **II. Define Metrics/Variable**

|  |  |
| --- | --- |
| Variables Name | Measure |
| *Salary* | *CEO's total remuneration in 1990 (including salary and bonus)* |
| *Pcsalary* |  |
| *Sales* | *Revenue in 1990, in million dollars* |
| *Roe* | *Return on equity, 88-90 avg* |
| *Pcroe* |  |
| *Ros* | *Return on firm's stock, 88-90* |
| *Indus* | *= 1 if it's an industrial company, otherwise = 0* |
| *Finance* | *= 1 if it's an financial company, otherwise = 0* |
| *Consprod* | *= 1 if it's a consumer product company, otherwise = 0* |
| *Utility* | *= 1 if a utility company, otherwise = 0* |
| *lsalary* | *natural log of salary (ln(salary))* |
| *lsales* | *natural log of sales (ln(sales))* |

**Identify specific questions from the research question posed.**

* What is the correlation between each independent variable and the dependent variable?
* Does the type of company have any effect on CEO compensation?

1. **The relationship between pcsalary and salary:**

Theoretically, the higher the pcsalary percentage, the higher the remuneration received by the CEO. Therefore, the pcsalary has a positive correlation with the CEO's salary.

1. **The relationship between company performance and CEO salary:**

In the CEOSAL1 data file, there are two Metrics representing the effective operation of the company, ROE and ROS.

ROE is an indicator that reflects the ability to use capital to generate profits, reflecting the level of profit spending shown through the income statement and also the expenditure of average equity based on balance sheet.

ROS: Return on firm's stock. Shows how much profit 1 dollar of stock sales will generate (profit after tax).

Based on available data, we can see most ROE, high ROS will lead to higher CEO remuneration. That is, company performance can be positively correlated with CEO remuneration.

In fact, the degree of change in remuneration explained by ROE, ROS is not much. This explanation is not surprising because many other characteristics of both the company and the individual CEO will affect remuneration. These factors are necessarily included in the errors in a simple regression analysis.

1. **The relationship between Sales and CEO salary:**

We can estimate a regression model that relates CEO salary to company sales. Think of sales as the company's annual sales, measured in millions of dollars. A regression model is:

salary = β0 + β1\*sales + u

The coefficient of sales is the estimated elasticity of wages to sales. It implies that a 1 unit increase in the company's sales will increase the CEO's salary by about β1 units.

1. **The relationship between type, size of company and salary:**

In the CEOSAL1 data file, there are 4 types of companies including: Indus, Finance, Consprod, Utility. From the available data, we can know the average remuneration of CEOs by type of company, specifically as follows:

Indus: includes 67 CEOs and the average salary of each CEO is 1139.34

Finance: Including 46 CEOs and the average salary of each CEO is 1344.91

Consprod: Including 60 CEOs and the average salary of each CEO is 1722.42

Utility: Includes 46 CEOs and the average salary of each CEO is 727.97

From the above data, we can objectively comment that the CEO's remuneration will also depend on the type of company the CEO is managing. This, in fact, provides a theoretical explanation for a proportional relationship between size and remuneration.

In addition, research by Zhou (2000) shows that CEO compensation increases with company size, and total compensation is associated with firm performance.

1. **The relationship between Issalary and Issale:**

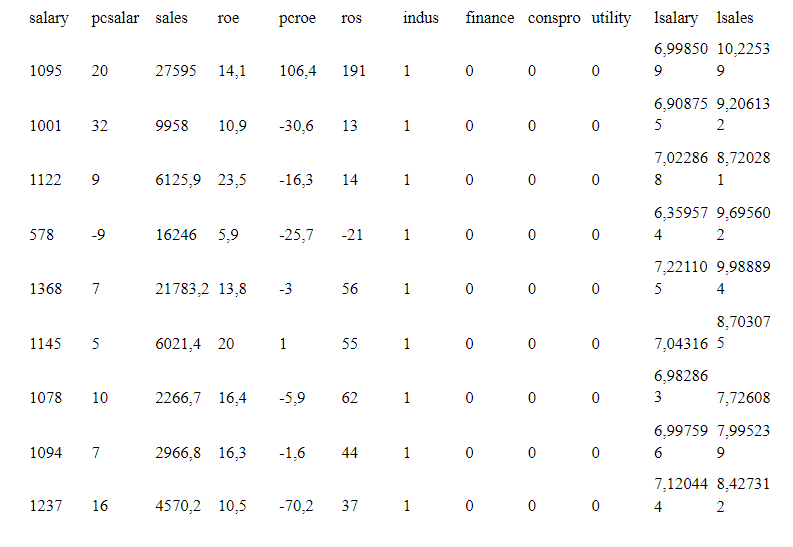
The same goes for sales' relationship with compensation. We can estimate a regression model that relates CEO salary to company sales. Think of sales as the company's annual sales, measured in millions of dollars. A regression model is:

log(salary) = β0 + β1\*log(sales) + u

The log(sales) coefficient is the estimated elasticity of wages to sales. It implies that a 1% increase in company sales will increase the CEO's salary by about β1 %.

# **III. Exploratory data analysis depedent variable with Descriptive Analytics method.**

Let's get acquainted with Data Salary (first 10 lines):



We can quickly see that:

* All columns are numeric.
* The meaning of each column in this dataset is relatively clear based on the column headings.
* Here, the label column is **salary or lsalary**. We need to build a model that represents this column correlation (salary or lsalar) based on the remaining columns. This means the model shows the factors that can affect the CEO's salary.

## **1. Salary or Log(salary)?**

It’s common in the economics literature to run studies trying to predict how various factors affect a person’s salary. When those studies run regressions, they don’t try to predict salary – they try to predict the logarithm of the salary.

The reason you’d use log(salary) is that you expect that an arithmetic (additive) change in the inputs produce a geometric (multiplicative) change in outputs. For instance, leaving your money in a 5% savings account for one extra year (adding 1) to the amount of time your money stays in a 5% savings account increases your amount of money by five percent (multiplying by 1.05). Because of that, if you were to run a regression to predict log(balance) based on years, you’d get a perfect correlation of 1.00. But if you run the regression on just balance, instead of log(balance), it woudn’t work as well.

On the other hand, the relationship between wages and hours worked is just the opposite. If you work an extra day (at, say, $10 per hour), you’re going to increase your wages by a fixed $80. It’s addictive not multiplicative – add one day, add $80. In this case, using just plain “wages” would give the perfect correlation, and using log(wages) would be the less accurate method.

So sometimes “salary” is the better choice, and sometimes “log(salary)” is the better choice. Which to choose depends on whether the relationship is additive or multiplicative. Does adding one X change salary by a fixed number? If so, don’t use the log. Does adding one X change salary by a certain proportion? If so, then the logarithm is necessary.

Inshort, log values effectively captures the compounding effect. Natural log value is the more appropriate measure for computing the change in variable which follows normal distribution.

Statistically speaking, example in stock finance,  since simple stock returns is always assumed to follow Log Normal distribution, it follows that Log returns will behave as Normal distribution. It is therefore plausible to use properties of Normal distribution in statistical estimation for Log returns but not simple returns. Another reason is that stock returns analysis is time series analysis in which you also take care of stationarity which is normally obtained by Log returns but not simple returns.

In our exercise, it is seems reasonable to use natural log of salary. But there is still no guarantee that salary is completely inappropriate. We will build models for both.

## **2. Data Size**

**Dataset statistics**

|  |  |
| --- | --- |
| **Number of variables** | 12 |
| **Number of observations** | 209 |
| **Missing cells** | 0 |
| **Missing cells (%)** | 0.0% |
| **Duplicate rows** | 0 |
| **Duplicate rows (%)** | 0.0% |

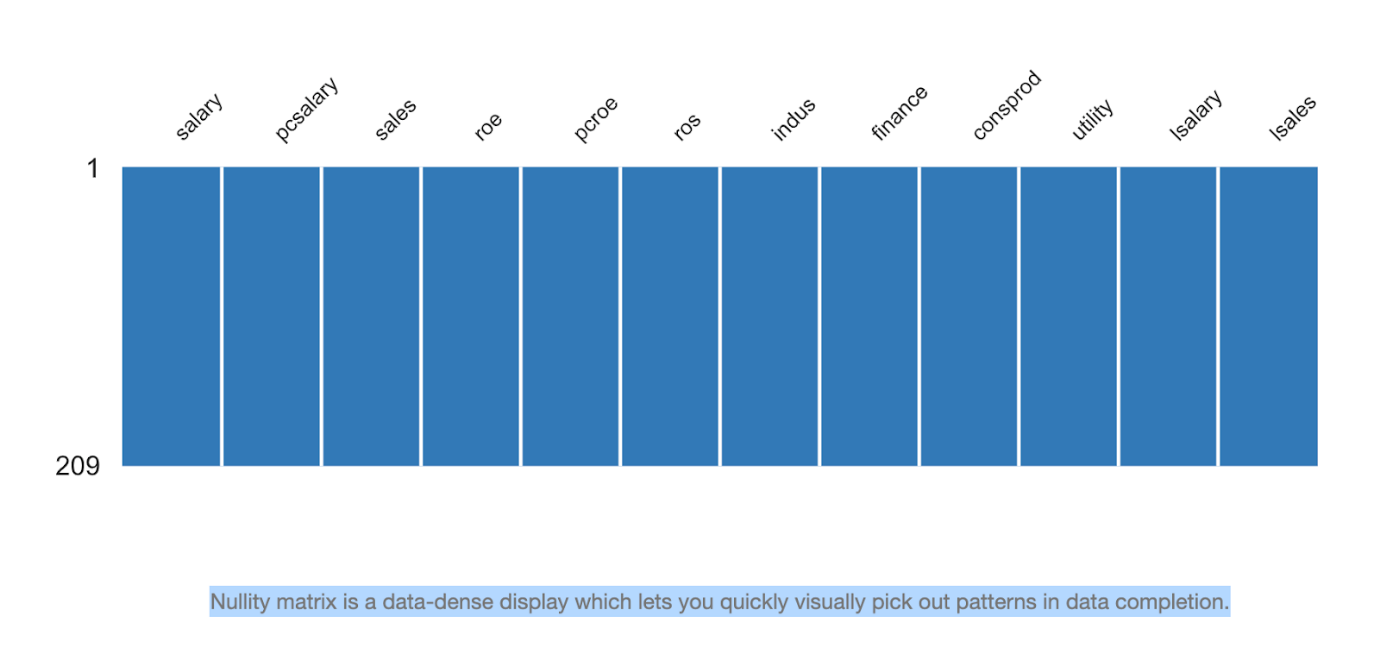
The table above shows an overview of the statistics about the data set.

* There are useful information such as the number of data fields (Number of variables), the number of data samples (Number of observations), the number of missing values ​​(Missing cells), the number of duplicate rows (Duplicate rows).
* There are 12 variables in total in the dataset.
* With 209 observations.
* No value cells are missing and no rows are duplicated.

**Variable types**

|  |  |
| --- | --- |
| **Numeric** | 8 |
| **Categorical** | 4 |

The number of columns in the category (Categorical) and the number of columns in the number (Numeric). That is, there are 8 quantitative variables and 4 categorical variables out of a total of 12 variables.



In all columns, no data is missing.

## **3. Statistics of each field (and histogram)**

In each information field, the statistics shown are:

* *Mean*: mean value
* *Median Absolute Deviation (MAD)*: absolute deviation
* *Standard Deviation*: Standard Deviation
* *Variance*: Variance
* *Coefficient of variation*: coefficient of variation
* *Kurtosis*: kurtosis compared to a normal distribution
* *Skewness*: deviation from the normal distribution
* *Minimum*: minimum value
* *Maximum*: maximum value
* *Median (50%):* median – value where exactly half of the elements in the column have a value less than or equal to it
* *Q1 (25%):* median of values ​​from min to 50%, ie exactly 25% of the elements in the column have a value less than or equal to it
* *Q3 (75%):* the median of the values ​​from 50% to max, that is, exactly 75% of the elements in the column have a value less than or equal to it
* *range:* range of variation

### **3.1. Salary (Million dollars)**

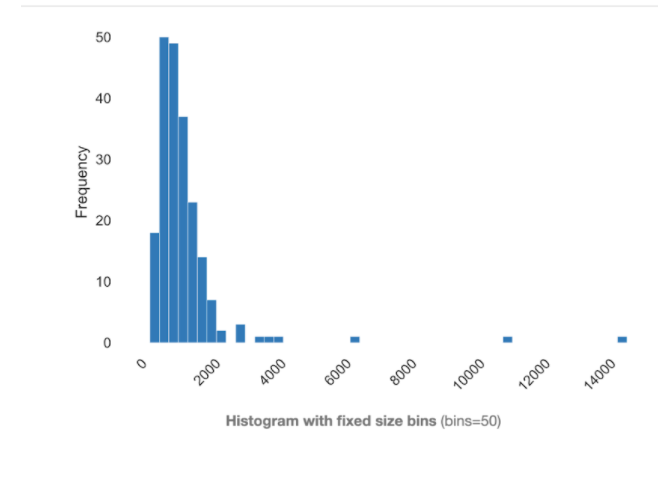
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | 223 |
| **5-th percentile** | 450 |
| **Q1** | 736 |
| **median** | 1039 |
| **Q3** | 1407 |
| **95-th percentile** | 2295.4 |
| **Maximum** | 14822 |
| **Range** | 14599 |
| **Interquartile range (IQR)** | 671 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 1372.345308 |
| **Standard Error** | 94,92710989 |
| **Mode** | 1368 |
| **Sample Variance** | 1883331,644 |
| **Coefficient of variation (CV)** | 1.071207785 |
| **Kurtosis** | 58.97122942 |
| **Mean** | 1281.119617 |
| **Skewness** | 6.904576803 |
| **Sum** | 267754 |
| **Variance** | 1883331.644 |
|  |  |

**Histogram**



For this column, which is the label column needed to build a model, the highest salary is $14822 thousand, the smallest salary is $223 thousand. Also, the median is less than the mean, which means that half of the elements in the column are below the mean.

Thus, there may be a clear disparity between the salary levels, which are unevenly distributed.

High salaries (which can be considered as outliers) have pushed the mean up (means), so more than 50% of the elements are below the mean.

The maximum value is far beyond the 3rd percentile (Q3).

The skew value must be more, the salary under 2000 thousand is the majority. Meanwhile, the kurtosis index is quite large, for which kurtosis only measures outliers. It shows that the outliers have quite large values.

### **3.2. Pcsalary**

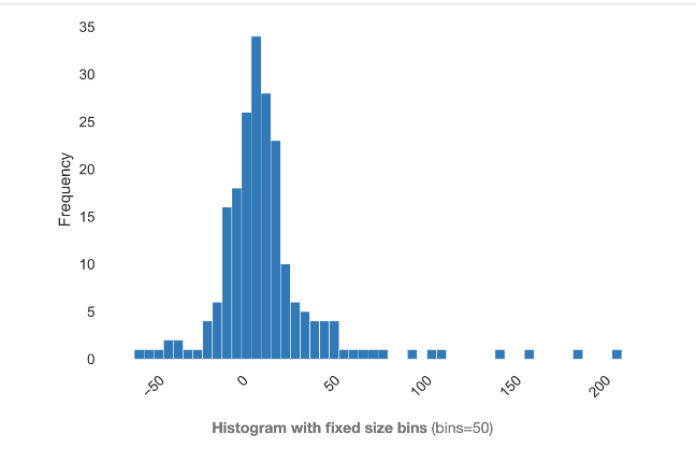
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | **-61** |
| **5-th percentile** | **-19** |
| **Q1** | **-1** |
| **median** | **9** |
| **Q3** | **20** |
| **95-th percentile** | **59.8** |
| **Maximum** | **212** |
| **Range** | **273** |
| **Interquartile range (IQR)** | **21** |
|  |  |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | **32.63392116** |
| **Standard Error** | **2,257335528** |
| **Mode** | **9** |
| **Sample Variance** | **1064,97281** |
| **Coefficient of variation (CV)** | **2.456948675** |
| **Kurtosis** | **12.61072602** |
| **Mean** | **13.28229665** |
| **Skewness** | **2.753377123** |
| **Sum** | **2776** |
| **Variance** | **1064.97281** |
|  |  |

**Histogram**



For the pcsalary column, is the change in salary as a percentage from the previous year. It is found that the smallest value is -61, which means that the salary is not ideal while not only increasing but also decreasing, and the maximum decrease is 61% compared to the previous year (1989).

In addition, the first percentile (Q1) still has a negative value, that is, up to 25% of the members may have a salary reduction compared to the previous year. This is not a large number, but it is not small, considering that in 1990 there may have been difficulties, or certain factors affecting their salary.

Skewness (skew) is greater than 1, it is found that there is a lot of skew in the data.

On the other hand, the largest percentage increase in salary is 212%, but it can be as much as 95% of people with the most increase near 60% of salary. At the same time, the kurtosis index is quite large, for which kurtosis only measures outliers. Thus, such a high increase in CEOs is quite certain to be considered an outlier.

### **3.3. Sales (Million dollars)**

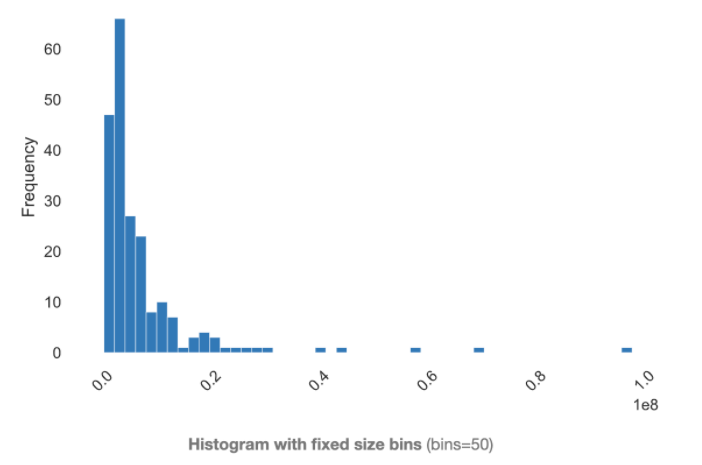
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | **175200** |
| **5-th percentile** | **772940** |
| **Q1** | **2210300** |
| **median** | **3705200** |
| **Q3** | **7177000** |
| **95-th percentile** | **20637000** |
| **Maximum** | **97649900** |
| **Range** | **97474700** |
| **Interquartile range (IQR)** | **4966700** |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 10633271.16 |
| **Standard Error** | 735518,7459 |
| **Mode** |  |
| **Sample Variance** | 1,13066E+14 |
| **Coefficient of variation (CV)** | 1.53575803 |
| **Kurtosis** | 33.11519374 |
| **Mean** | 6923793.301 |
| **Median Absolute Deviation (MAD)** | 2075900 |
| **Skewness** | 5.035335785 |
| **Sum** | 1447072800 |
| **Variance** | 1.130664556 × 1014 |

**Histogram**



In the sales column of this company in 1990, the range of variation is quite large, meaning that the difference between the minimum value and the maximum value is very large. If we ignore the difference between the size of the companies for which the data is collected, this is a field that is strongly related to the performance of the company in general, and to the salary of employees in particular.

Notice that the median is less than the mean for this field and that the difference between the 3rd percentile (Q3) and the mean is not high, i.e. more than 50% and less than 75% of the companies in data have the sales less than avarage sales. This demonstrates a clear impact of outliers on the mean.

Kurtosis has a high value.

Besides, the skew of the data is skewed to the right.

### **3.4. Roe**

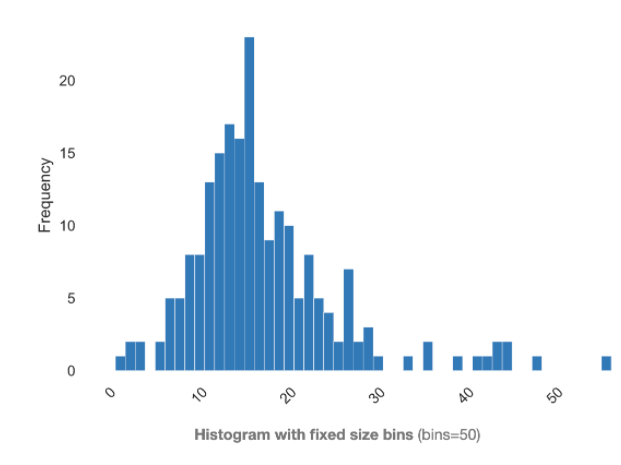
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | 0.5 |
| **5-th percentile** | 6.92 |
| **Q1** | 12.4 |
| **median** | 15.5 |
| **Q3** | 20 |
| **95-th percentile** | 34.38 |
| **Maximum** | 56.3 |
| **Range** | 55.8 |
| **Interquartile range (IQR)** | 7.6 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 8.518508665 |
| **Standard Error** | 0,589237565 |
| **Mode** | 15,1 |
| **Sample Variance** | 72,56498988 |
| **Coefficient of variation (CV)** | 0.4957171964 |
| **Kurtosis** | 3.797366054 |
| **Mean** | 17.18421053 |
| **Skewness** | 1.572125873 |
| **Sum** | 3591.5 |
| **Variance** | 72.56498988 |

**Histogram**



ROE is often the most significant among companies in the same industry. Assume that the companies in the data sheet are surveyed in the same industry. The difference between roe will have many reasons, but when comparing the roe index in the same field, it gives a high significance value.

The difference between the median and the mean is not too large. That is, the mean can relatively reflect the concentrated distribution of the data in this field. It can be seen that the data distribution ranges from 0.5 (minimum) to 56.3 (maximum) and with an approximate mean of 17 so the data concentration in this field is relatively even on both sides (still may have outliers).

Sometimes an extremely high ROE is a good thing if the net income is extremely large relative to equity because the company's performance is very strong. However, extremely high ROE is often caused by a small equity account relative to net income, which indicates risk. Therefore, it is necessary to consider this variable with other relationships.

### **3.5. Pcroe**

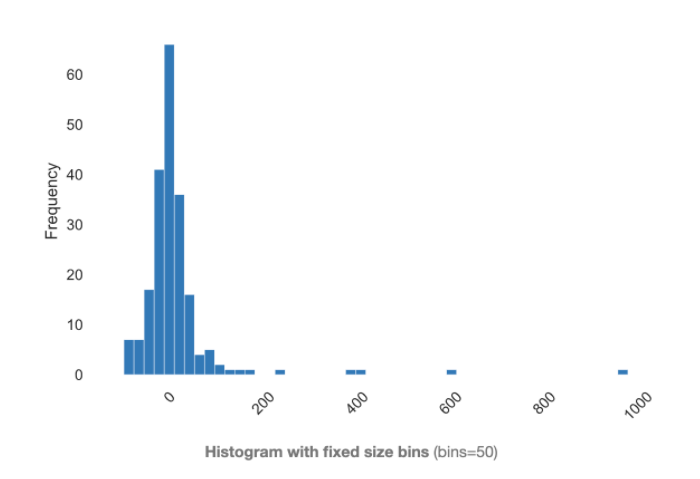
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | -98.9 |
| **5-th percentile** | -61.96 |
| **Q1** | -21.2 |
| **median** | -3 |
| **Q3** | 19.5 |
| **95-th percentile** | 91.38 |
| **Maximum** | 977 |
| **Range** | 1075.9 |
| **Interquartile range (IQR)** | 40.7 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 97.21940022 |
| **Coefficient of variation (CV)** | 9.00139753 |
| **Standard Error** | 6,72480653 |
| **Mode** | -7,8 |
| **Sample Variance** | 9451,611779 |
| **Kurtosis** | 54.42070584 |
| **Mean** | 10.80047847 |
| **Skewness** | 6.463321117 |
| **Sum** | 2257.3 |
| **Variance** | 9451.611779 |

**Histogram**



In this column reflecting the change of ROE from 1988-1990, it is clear that the median value has a negative value, which can be as high as 50% of the surveyed companies that have reduced ROE in this period. . And the highest is a decrease of approximately 99%.

When the ROE ratio decreases, it is a bad point of the business for investors when the company's ability to use capital is not efficient. With a nearly 50% drop, this may reflect a general event that has affected these businesses. We know that pcroe =  and the highest change is up to 977% compared to the old ROE, this is a group change I am too difficult to make. There may be errors or for some other reason, but still consider this a data worth noting for us.

The result is a large disparity between the median and the mean.

### **3.6. Ros**

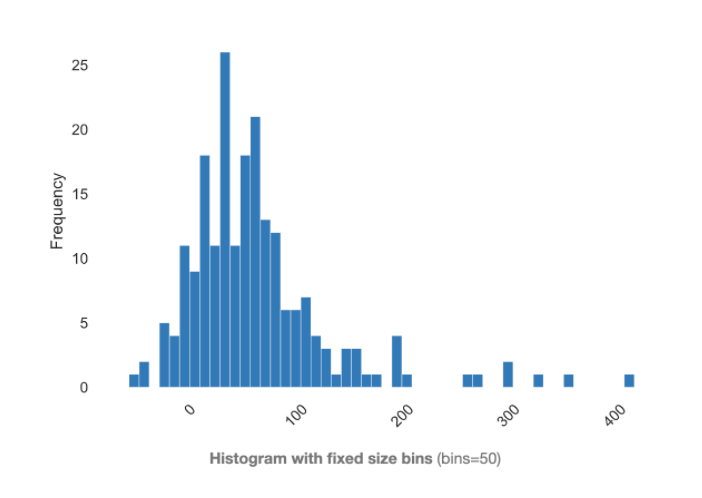
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | -58 |
| **5-th percentile** | -14.6 |
| **Q1** | 21 |
| **median** | 52 |
| **Q3** | 81 |
| **95-th percentile** | 191 |
| **Maximum** | 418 |
| **Range** | 476 |
| **Interquartile range (IQR)** | 60 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 68.17705164 |
| **Coefficient of variation (CV)** | 1.103120213 |
| **Standard Error** | 4,715905272 |
| **Mode** | 37 |
| **Sample Variance** | 4648,11037 |
| **Kurtosis** | 6.585102165 |
| **Mean** | 61.80382775 |
| **Skewness** | 2.094631134 |
| **Sum** | 12917 |
| **Variance** | 4648.11037 |

**Histogram**

****

In this column about the ROS index, which is meant to assess the profitability of a business, specifically here is from stock activities during the period 1988 - 1990. High ROS is a desire of any business. And also assuming these businesses are looking at the same industry aspect.

From the statistical indicators, the minimum value is -58% and possibly at most 5% of the surveyed companies have ROS of -14.6%. Thus, there is a small group of companies with negative ability, ie not profitable. This is largely due to stock devaluation or loss, but it is also undeniably the company's strategy.

In addition, the remaining companies have a positive ROS index, with the third percentile (Q3) corresponding to 81%, that is, at most 75% of the companies have the ability to turn a unit profit from the stock. brings 81% of profit units to the business.

However, the maximum value is up to 418%, which can be a note for research when in such a short period of time this company has such high returns on shares.

High kurtosis, showing that the impact of outliers is relatively large.

### **3.7. lsalary**

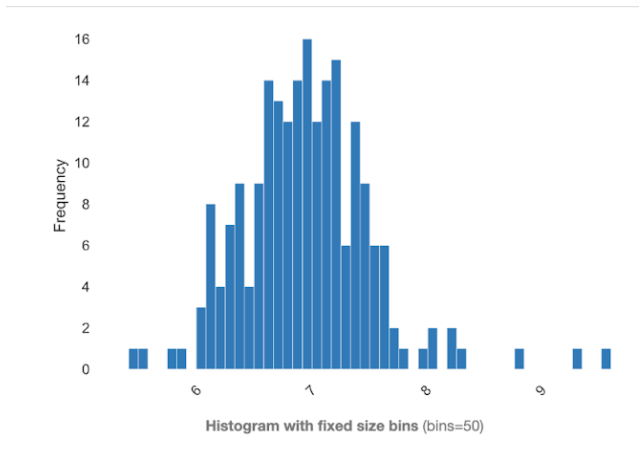
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | 5.407172 |
| **5-th percentile** | 6.1092326 |
| **Q1** | 6.60123 |
| **median** | 6.946014 |
| **Q3** | 7.249215 |
| **95-th percentile** | 7.7385194 |
| **Maximum** | 9.603868 |
| **Range** | 4.196696 |
| **Interquartile range (IQR)** | 0.647985 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 0.5663741498 |
| **Standard Error** | 0,039176919 |
| **Mode** | 7,221105 |
| **Sample Variance** | 0,320779678 |
| **Coefficient of variation (CV)** | 0.08148815631 |
| **Kurtosis** | 3.376628036 |
| **Mean** | 6.95038611 |
| **Skewness** | 0.8862121141 |
| **Sum** | 1452.630697 |
| **Variance** | 0.3207796775 |

**Histogram**

****

With this column, the label column which is needed to build the model. Compared to the salary column, the distribution of data in this column is somehow more "standard".

With the mean approximately equal to the median, the distribution of the data is relatively on both sides.

However, the high kurtosis and relative bias also reflect the influence of the extreme points of the data set.

Low standard deviation.

### **3.8. lsales**

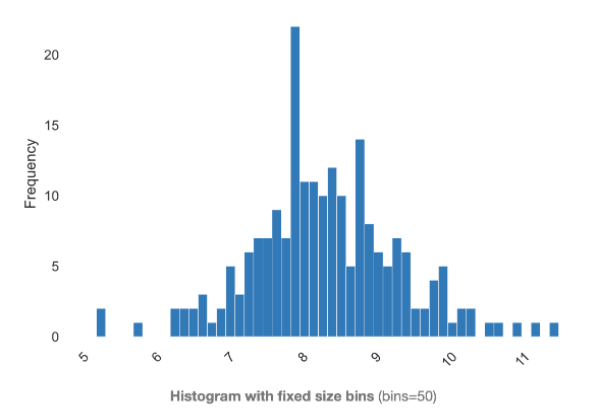
**Quantile statistics**

|  |  |
| --- | --- |
| **Minimum** | 5.165928 |
| **5-th percentile** | 6.6501162 |
| **Q1** | 7.700883 |
| **median** | 8.217492 |
| **Q3** | 8.878636 |
| **95-th percentile** | 9.93484 |
| **Maximum** | 11.48914 |
| **Range** | 6.323212 |
| **Interquartile range (IQR)** | 1.177753 |

**Descriptive statistics**

|  |  |
| --- | --- |
| **Standard deviation** | 1.013160462 |
| **Standard Error** | 0,070081784 |
| **Mode** |  |
| **Sample Variance** | 1,026494396 |
| **Coefficient of variation (CV)** | 0.1221813929 |
| **Kurtosis** | 0.8022010073 |
| **Mean** | 8.292264785 |
| **Skewness** | 0.09488930197 |
| **Sum** | 1733.08334 |
| **Variance** | 1.026494122 |

**Histogram**

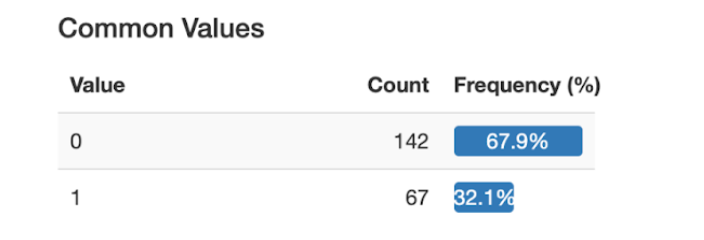


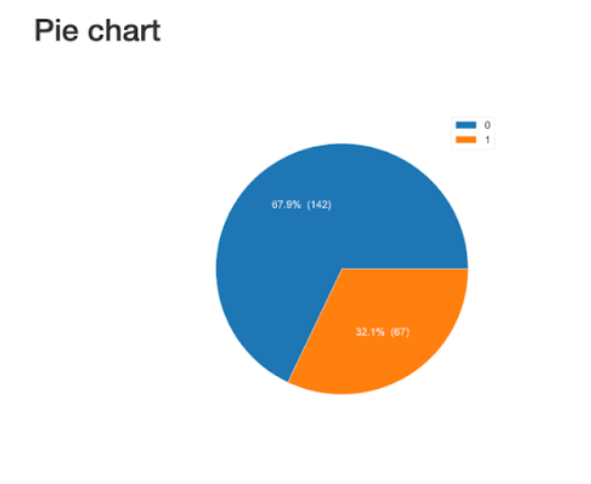
Similar to the lsalary column, the distribution of sales when using lsales is much more "standard". The distribution of the data is somewhat more even. Instead of the obvious right deviation like in sales, it is no longer seen here.

In addition, it also overcomes the unit difficulty when sales are calculated in units of thousands of dollars and because they will have a considerably long tail for high values ​​and a shorter tail for small values. It is very difficult to deduce if it remains the same.

### **3.9. indus**

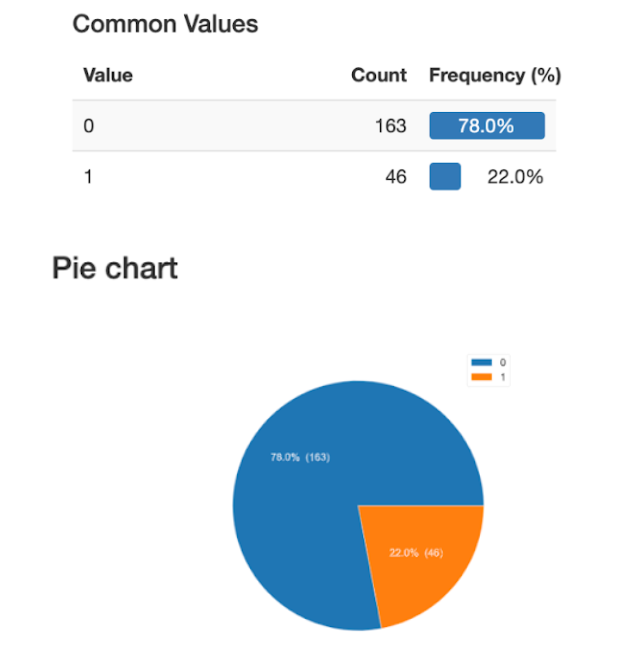
Total characters: 209



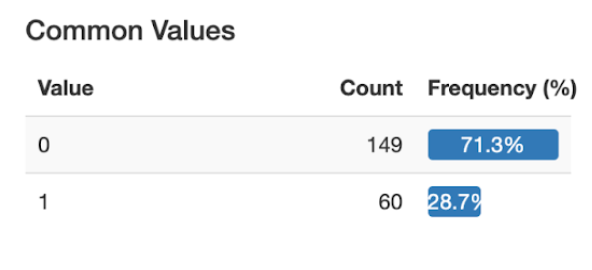


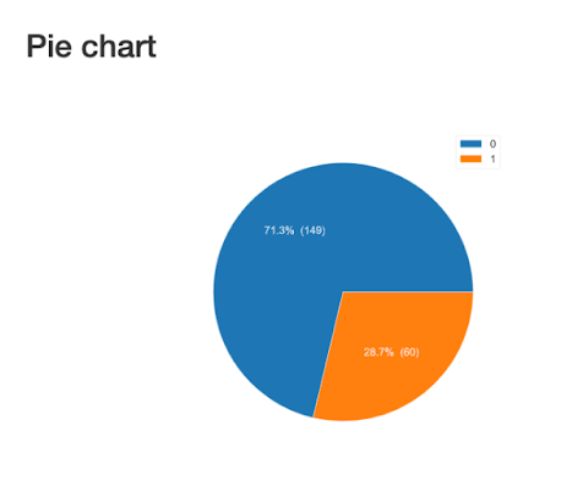
### **3.10. finance**

Total characters: 209

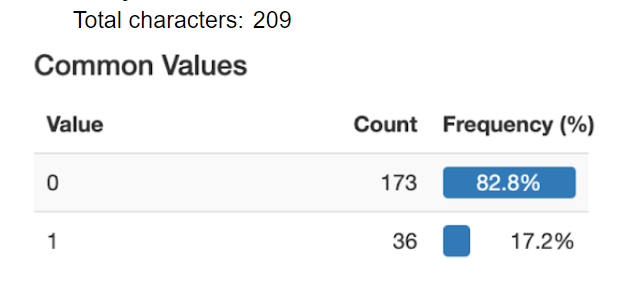


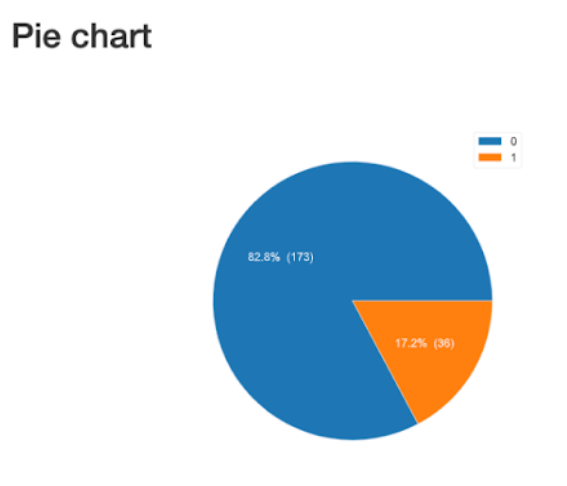
### **3.11. consprod**





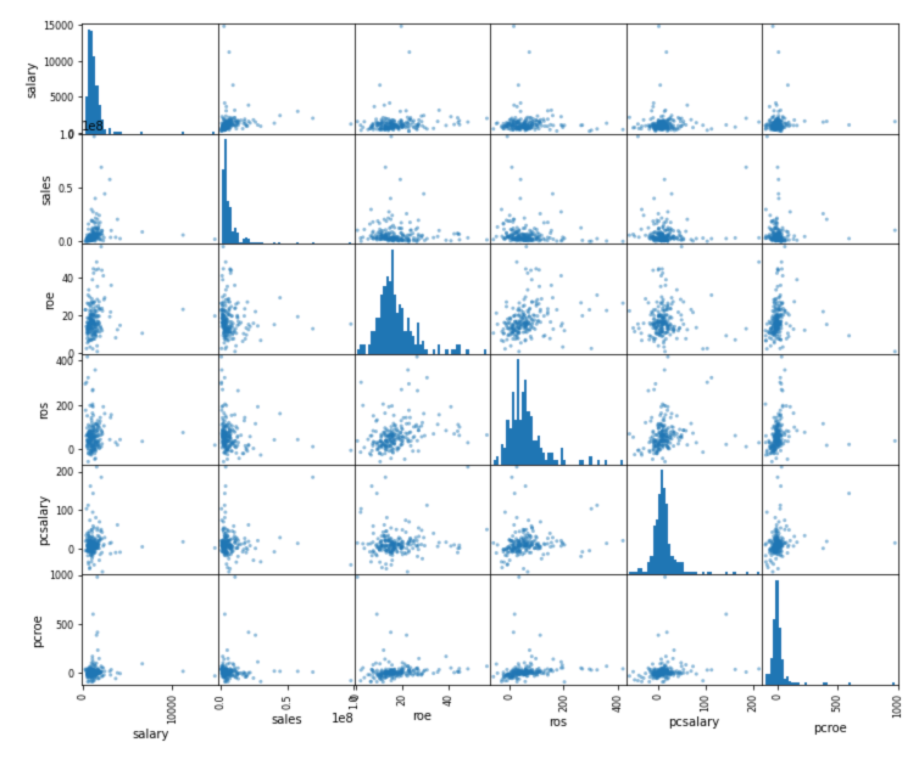
### **3.12. utility**





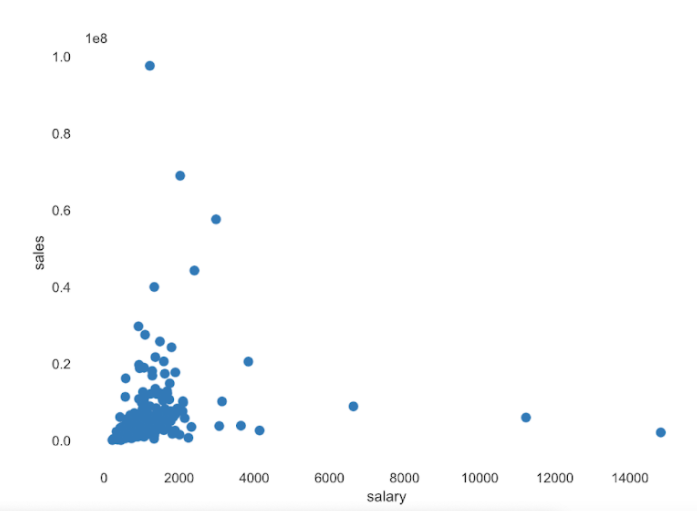
## **4. Typical correlation between data fields**

When excuting EDA, we also need to calculate the correlation between the data fields, especially between the prediction label and the remaining fields.



The figures on the diagonal represent the histogram of each column.

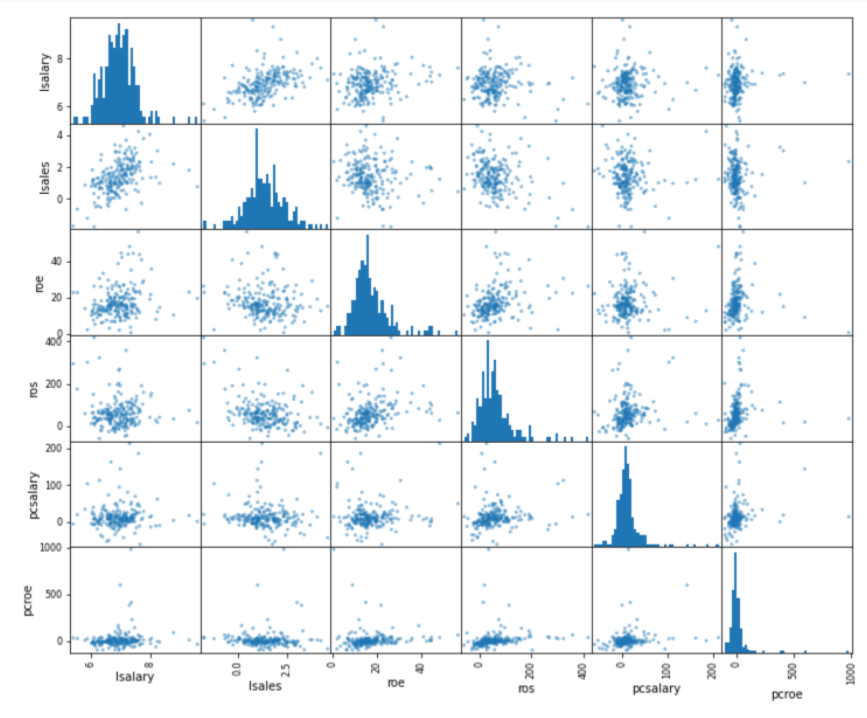
The remaining figures show the scatter plot of the data points. Specifically, with the cell in the upper left corner corresponding to the pair of columns (salary, sales), we take the two corresponding columns as abscissa and ordinate for these points. Zooming on this box we can see:



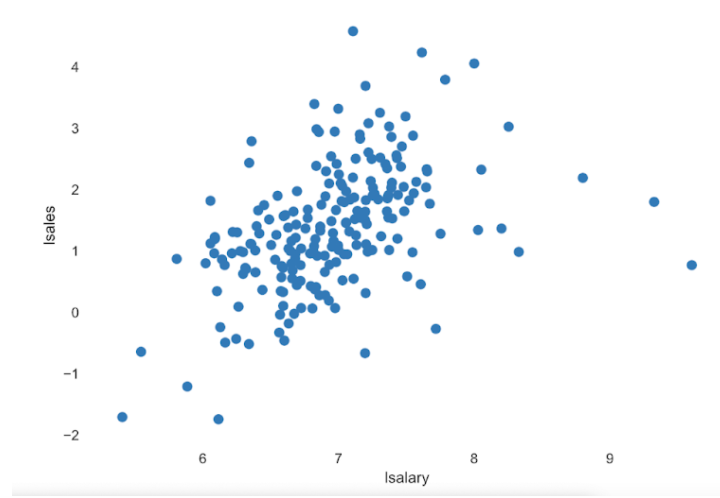
Here, we observe that companies with high sales tend to have higher salaries than companies near the bottom left corner. This is not quite the case at the edge of the right part and the points above.

Although, at the point where the revenue is at the highest level, the salaries of those points are not high. In contrast, at the point where the salary is at the highest level, but the sales revenue is not high.

The data points are mainly located in the lower left part, the density of the remaining cells is quite sparse.



With the cell in the upper left corner corresponding to the pair of columns (lsalary, lsales), we take the two corresponding columns as abscissa and ordinate for these points. Zooming on this box we can see:



Here, we observe that firms with high natural logarithms of sales revenue tend to have higher natural logarithms of salary. Since the logarithm simply reflects the change in proportion, an increase in the percentage of sales revenue will probably increase the salary ratio. This also makes sense given the fact that an increase in sales has a positive effect on salary.

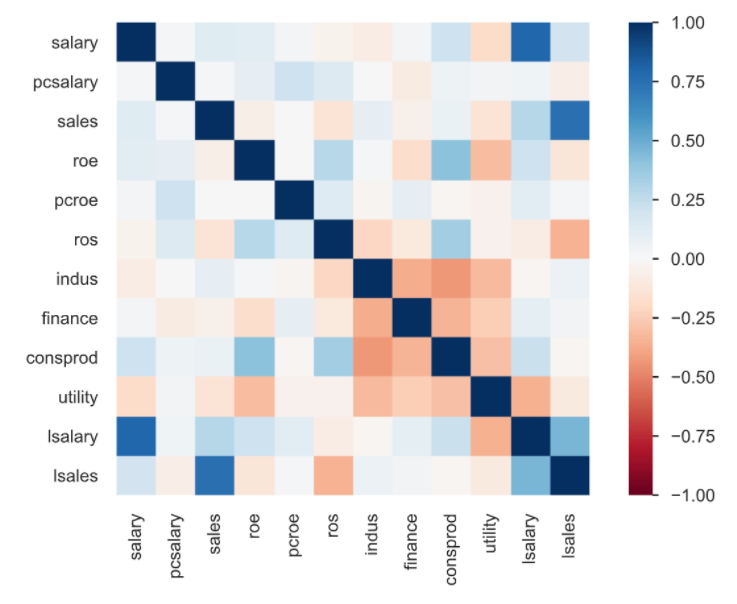
However, the point with the highest salary is still not the highest level of revenue.

In addition, our team also introduced a matrix to show the correlation: Pearson's r

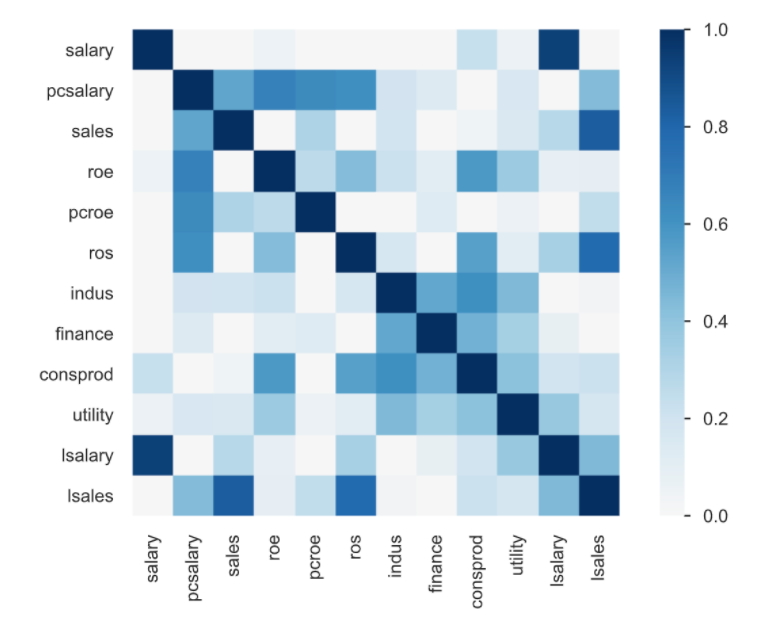
The Pearson correlation coefficient (r) is a measure of the linear correlation between two variables. Its value is between -1 and +1, -1 indicates total negative linear correlation, 0 indicates no linear correlation, and 1 indicates positive total linear correlation. Furthermore, r is invariant when there are separate changes in the positions and sizes of the two variables, implying that for a linear function, the angle to the x-axis has no effect on r.

To calculate r for two variables X and Y, one divides the covariance of X and Y by the product of their standard deviations.

Specifically in this exercise:



Here, we see that the column “lsalary” and column “lsales” have a medium blue correlation (about 0.5), that is, the smaller “lsalary” is, the smaller “lsales” will be and in contrast. This can be seen in the scatter chart.



This is the Phik matrix (φk):

Phik(φk) is a new and real correlation coefficient that active consistently across categorical, ordinal, and timescale variables, capturing the non-linear and reverting dependence on the Pearson correlation coefficient in the field in case there is normal distribution for two input variables. In this matrix we clearly see that in the column “pcsalary”, that is, the change of salary has a dark blue correlation with ros, roe, with the change of roe (pcroe) and the column “sales”.

In addition, there is also a high correlation between “lsales” and the “ros” column.

These correlations are meaningful for building research models.

## **5. Outlier points**

There are two groups of outliers:

The values ​​are not in the specified range of the data. For example, age, income or distance cannot be negative.

Values ​​are likely to occur but the probability is very low. For example, 120 years old, income 1 million dollars/month. These values ​​are likely to occur but are actually rare.

How to handle:

With data belonging to the first group, we can consider it as a missing value. Sometimes the missing values ​​are encoded with a special value that is not in the possible range of the data.

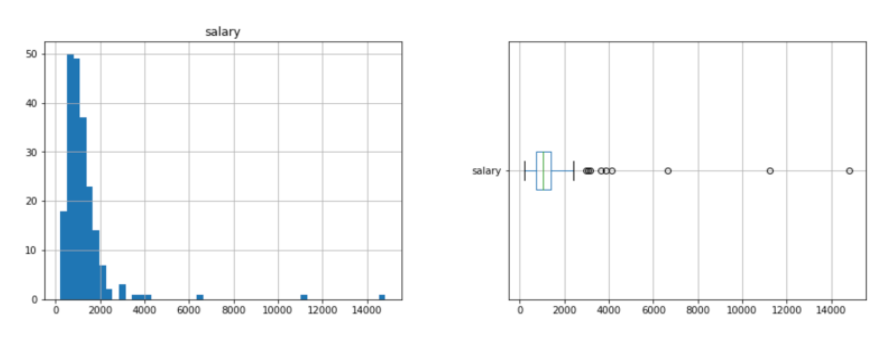
For data belonging to the second group, people often use the method of upper or lower bound (clipping or capping). That is, when a value is too large or too small, we reduce it to the maximum/smallest value that is considered normal points. For example, given a value of age of 120, we can bring it back to 70 and assume that this data point has the common characteristics of "elderly". One point worth noting is that choosing the largest/smallest value also depends on the data. If the data only includes elderly people aged 65 and over, it is clear that blocking above by 70 is not reasonable because 70 is still too young in this dataset.

As analyzed in the section “Statistics of each field” with histogram, we have noticed that variables such as “salary”, ”pcsalary”, ”sales”, ”pcoe”, ”ros” have skewed data. right (or left) pretty much. It may be due to the effect of external variables or outliers.

## **6. Box Plot:**

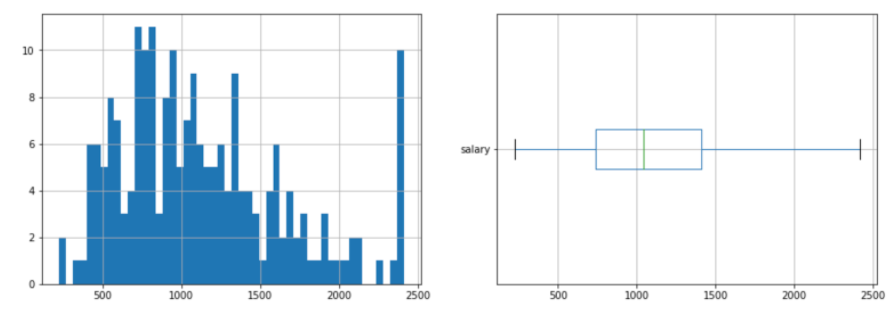
The box plot not only helps to determine if the data has outliers, but also helps to find the maximum and minimum thresholds to be intercept.

Let's start with the "salary" column:



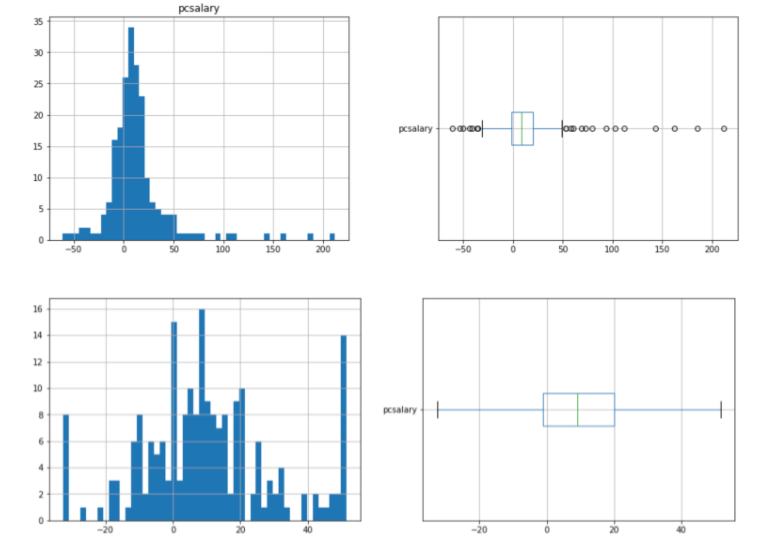
From the histogram we see that the data is skewed to the right (there are outliers that skew a lot to the right, or the "tail" of the histogram is to the right). From the boxplot we see that there are many points that are considered outliers. Outliers can be handled by clipping the minimum and maximum values ​​of the Box plot.

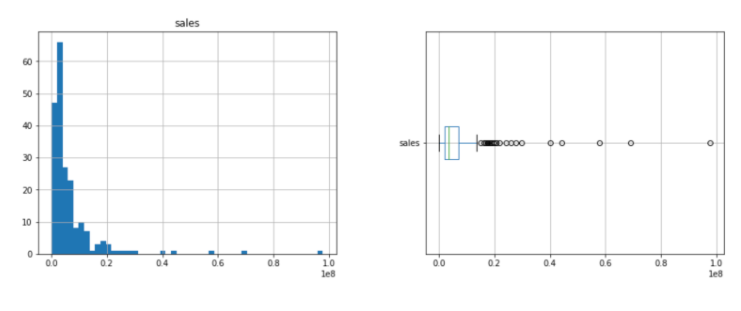
This processor can be implemented as a sklearn API. Applying again to the data of the "salary" column, we have a new histogram and boxplot as follows:

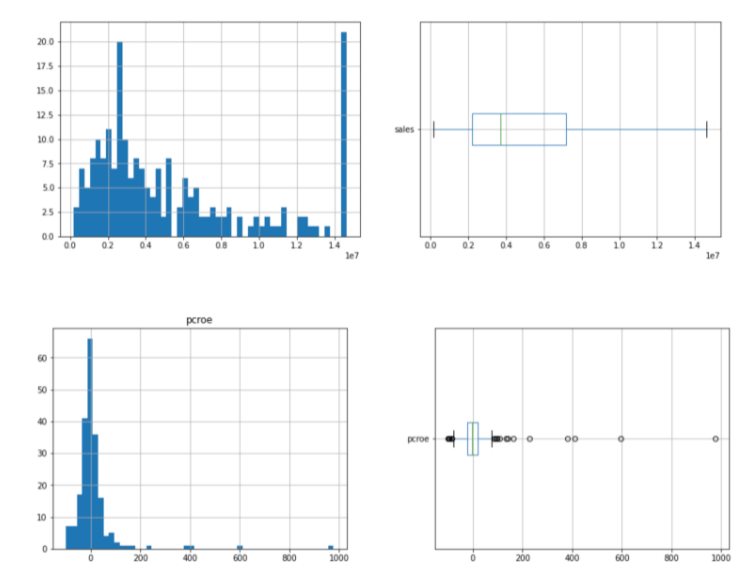


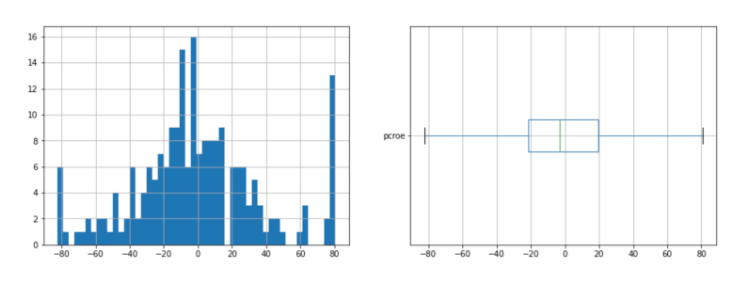
After clipping the data to the minimum and maximum of the box plot, we see that the data is less skewed. The box plot also shows that there are no more outliers. New data is always has no outliners. This is achieved because the clip transform does not change the quartiles of the data. The "valid" range of the boxplot before and after the clip does not change.

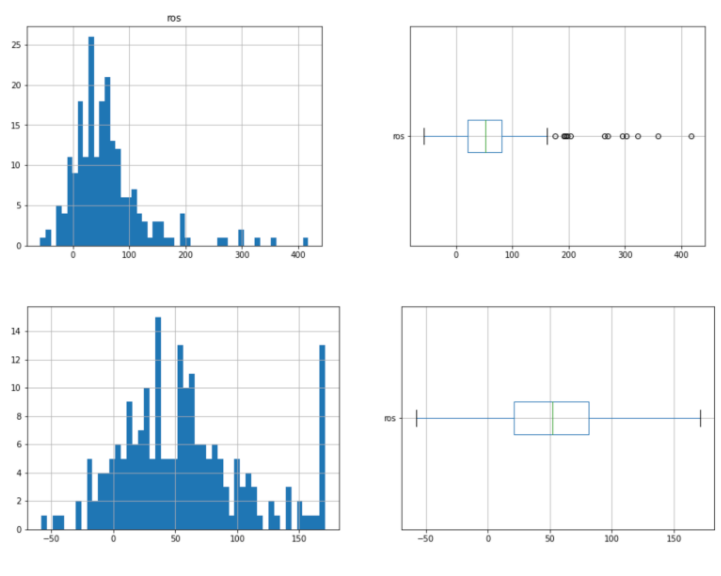
Same goes for the columns “pcsalary”, “sales”, “pcoe”, “ros”:





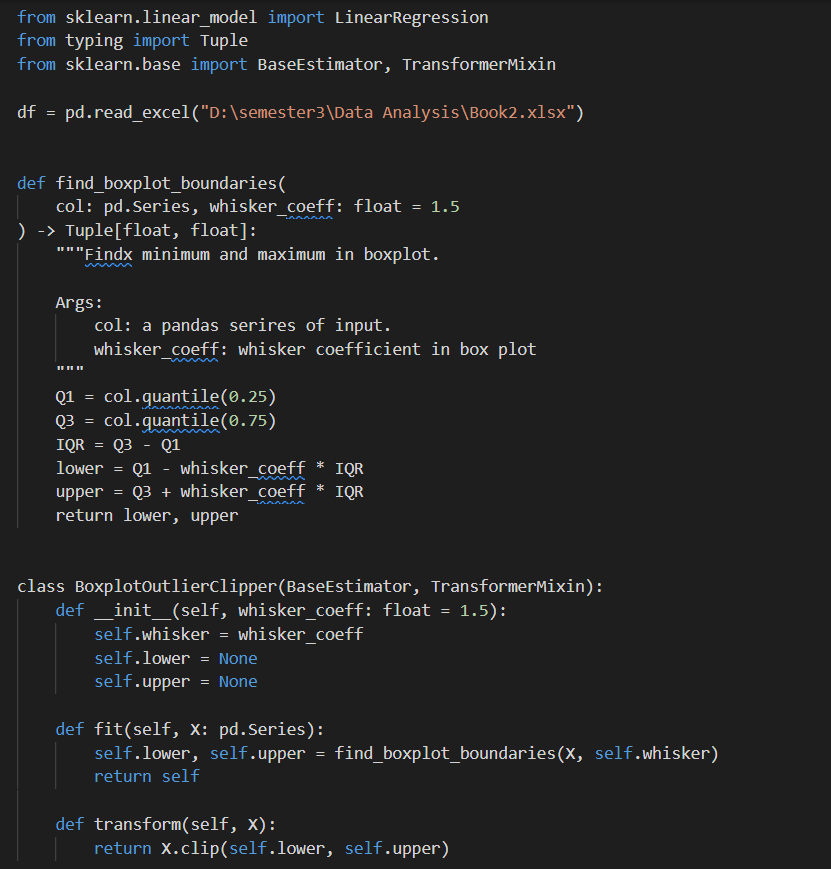


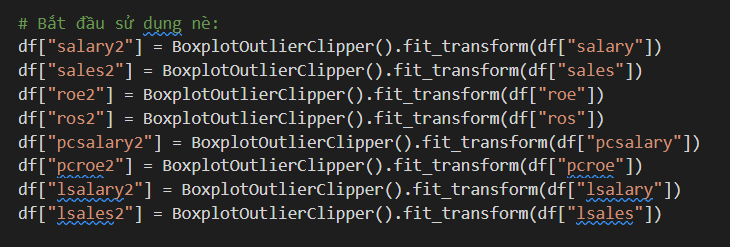




# **IV. Handle Outliers**

Before building a regression model, we decided to handle the outliers and build the right dataset. We use code like this:





# **V. Liner Regression Model**

First, review 4 models, respectively 5 variables, 6 variables, 7 variables and 8 variables.

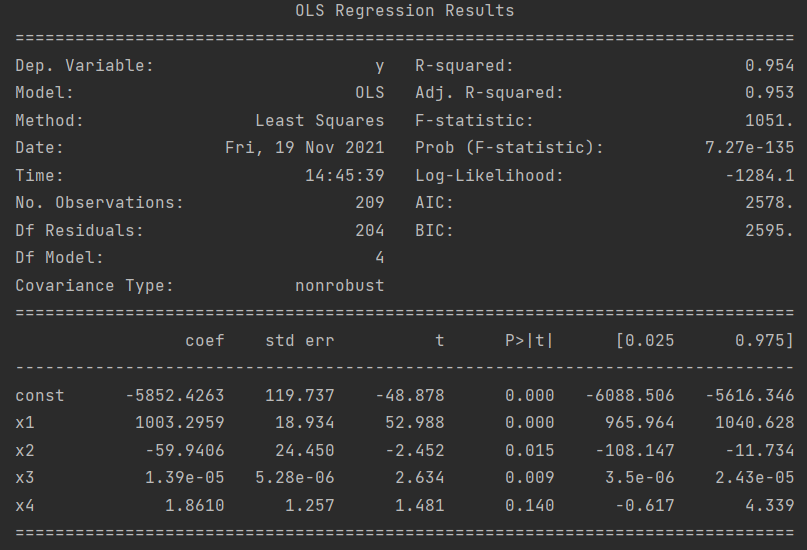
These model will be build in this format:

Y = β0 + β1\*X1 + β2\*X2 +....+ βn\*Xn

Y: dependent variable

Xn: independent variables

## 1. Model (1) – 5 variables: salary, lsalary, lsales, sales, roe



⇒ salary = -5852.4263 + 1003.2959\*lsalary – 59.9406\*lsales + 1.39e-05\*sales + 1.8610\*roe

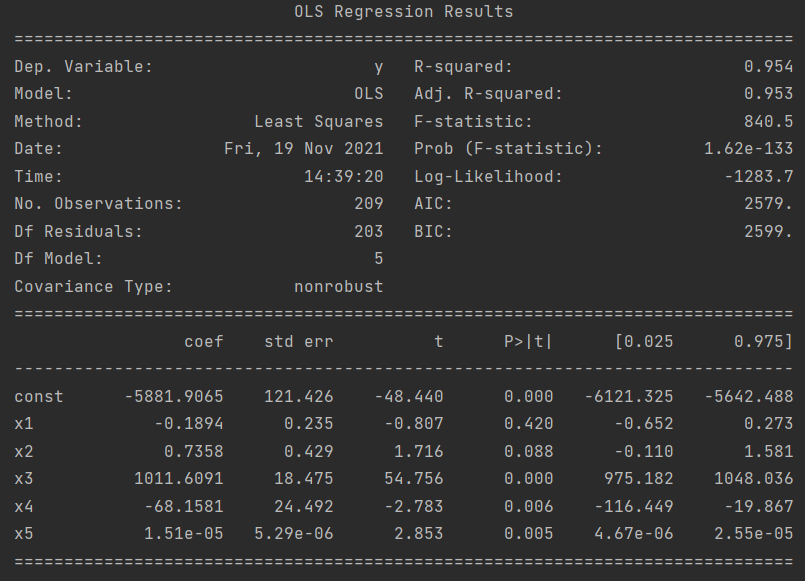
* When lsalary increases by 1 percent, salary will increase by 1003.2959 percent
* When lsales increases by 1 percent, salary will decrease 59.9406 percent
* When sales increase to 1 thousand dollars, salary will increase by 1.39e-05 thousand dollars
* When roe increases by 1 percent, salary will increase by 1.8610 thousand dollars

We can see Adj. R-squared < R-squared, which mean we use Adj. R-squared to evaluate the suitability of model because it will not exaggerate the suitability.

Adj. R-squared = 0.953 meaning that independent variables can explain 95.3%

variation of the dependent variable. Other 4.7% independent variables is explained by outliner variables and random errors.

## 2. Model (2) – 6 variables: salary, pcroe, pcsalary, lsalary, lsales, sales

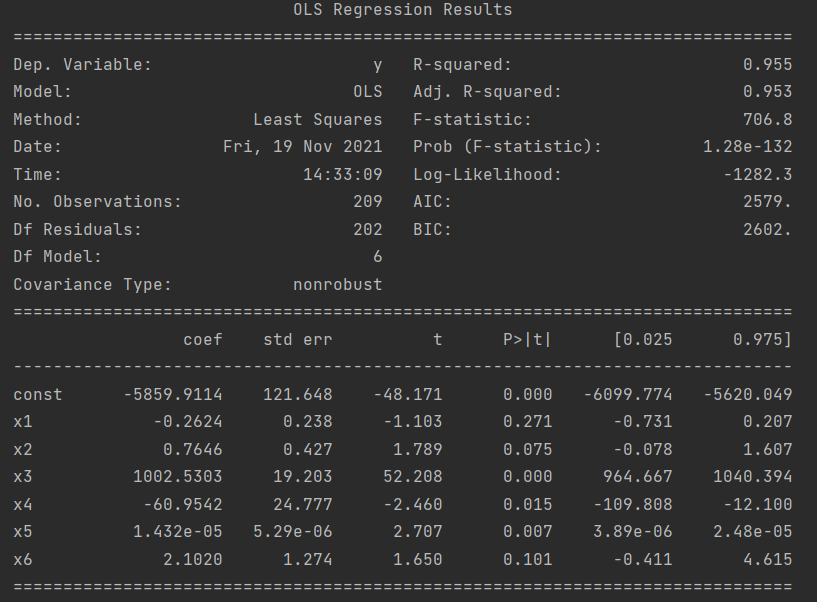


⇒ salary = -5881.9065 – 0.1894\*pcroe + 0.7358\*pcsalary + 1011.6091\*lsalary -68.1581\*lsales + 1.51e-05\*sales

* When pcroe increases by 1 percent, salary will decrease 0.1894 thousand dollars
* When pcsalary increases by 1 percent, salary will increase by 0.7358 thousand dollars
* When lsalary increases by 1 percent, salary will increase by 1011.6091 percent
* When lsales increases by 1 percent, salary will decrease 68.1581 percent
* When sales increase to 1 thousand dollars, salary will increase by 1.51e-05 thousand dollars

Adj. R-squared = 0.953 meaning that independent variables can explain 95.3% variation of the dependent variable. Other 4.7% independent variables is explained by outliner variables and random errors.

3. Model (3) – 7 variables: salary, pcroe, pcsalary, lsalary, lsales, sale, roe

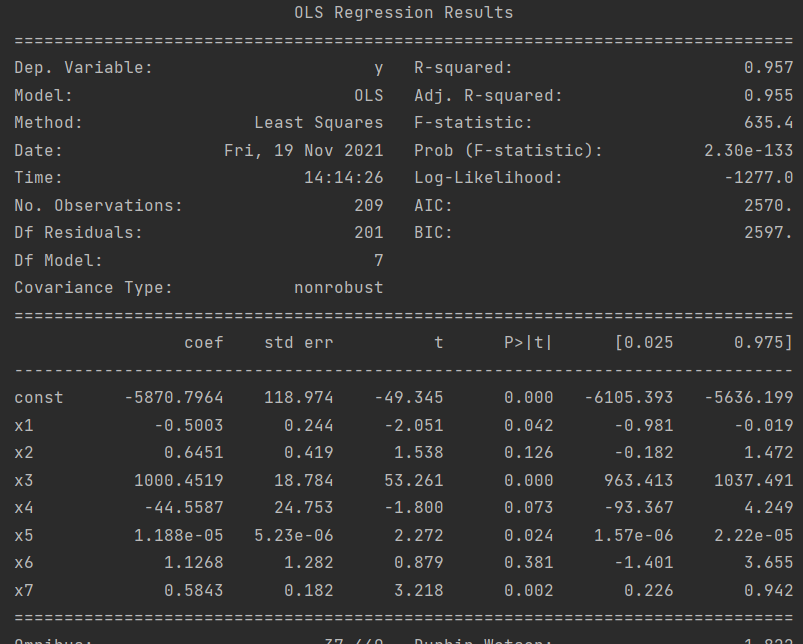


⇒ salary = -5859.9114 - 0.2624\*pcroe + 0.7646\*pcsalary + 1002.5303\*lsalary -60.9542\*lsales + 1.432e-05\*sales + 2.1020\*roe

* When pcroe increases by 1 percent, salary will decrease 0.2624 thousand dollars
* When pcsalary increases by 1 percent, salary will increase by 0.7646 thousand dollars
* When lsalary increases by 1 percent, salary will increase by 1002.5303 percent
* When lsales increases by 1 percent, salary will decrease 60.9542 percent
* When sales increase to 1 thousand dollars, salary will increase by 1.432e-05 thousand dollars
* When roe increases by 1 percent, salary will increase by 2.1020 thousand dollars

Adj. R-squared = 0.953 meaning that independent variables can explain 95.3% variation of the dependent variable. Other 4.7% independent variables is explained by outliner variables and random errors.

## 4. Model (4) – 8 variables: salary, pcroe,pcsalary,lsalary,lsales,sale,roe,ros

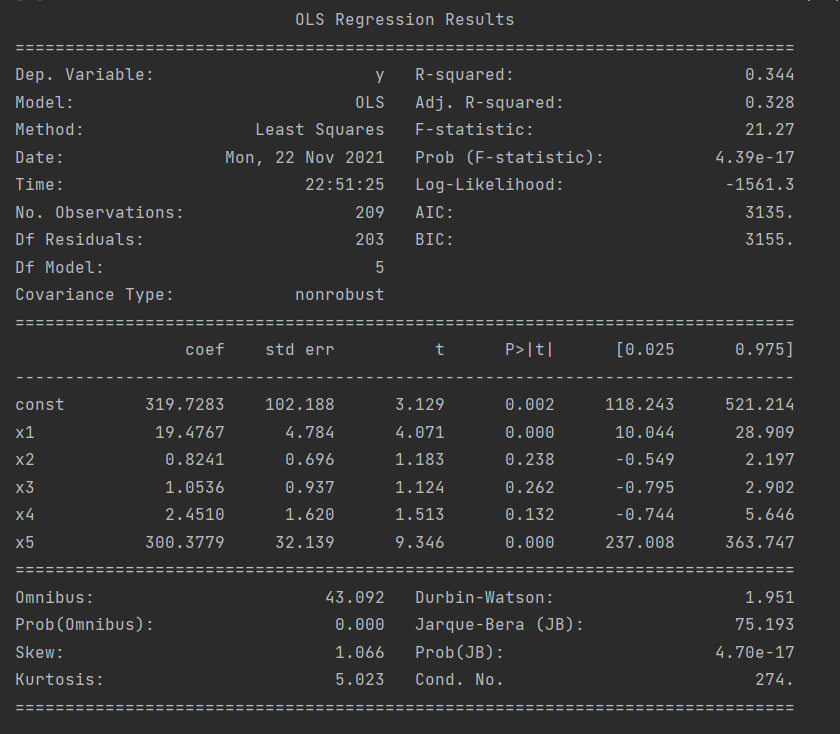


⇒ salary = -5870.7964 - 0.5003\*pcroe + 0.6451\*pcsalary + 1000.4519\*lsalary -44.5587\*lsales + 1.188e-05\*sales + 1.1268\*roe + 0.5843\*ros

* When pcroe increases by 1 percent, salary will decrease 0.5003 thousand dollars
* When pcsalary increases by 1 percent, salary will increase by 0.6451 thousand dollars
* When lsalary increases by 1 percent, salary will increase by 1000.4519 percent
* When lsales increases by 1 percent, salary will decrease 44.5587 percent
* When sales increase to 1 thousand dollars, salary will increase by 1.188e-05 thousand dollars
* When roe increases by 1 percent, salary will increase by 1.1268 thousand dollars
* When ros increases by 1 percent, salary will increase 0.5843 thousand dollars

Adj. R-squared = 0.955 meaning that independent variables can explain 95.3% variation of the dependent variable. Other 4.5% independent variables is explained by outliner variables and random errors.

In addtion, the phenomenon of multicollinearity strong between independent variables usually occurs when the coefficient determine high R\_Squared (≥ 0.8) while t.statistic is low. To handle this situation, we cut out some variables that connected to each other and can occur the redundancy of information. We call that model (4.1) – 6 variables after cut out the lsalary and lsales:



salary = 319.7283 + 19.4767\*pcroe + 0.8241\*pcsalary + 1.0536\*sales + 2.4510\*roe + 300.3779\*ros

After testing, we will choose model (4.1) with 6 quantitative variables

Check whether the model is suitable or not:

H0: **β1** = **β2** = **β3** = **β4** = **β5** = **β6** =  **β7** = 0

H1: **β12**+ **β22**+ **β32**+ **β42** + **β52**+ **β62**+  **β72** ≠ 0

We rely on the P-value of Prob (F-statistic) with the test level a = 5%

P-value = 4.39e-17 < 0.05 eligible to reject H0, our model is suitable.

# **VI. Linear Regression Model with Categorical Variables (or called Dummy Variables)**

From here, we will call Dummy Variables as D.

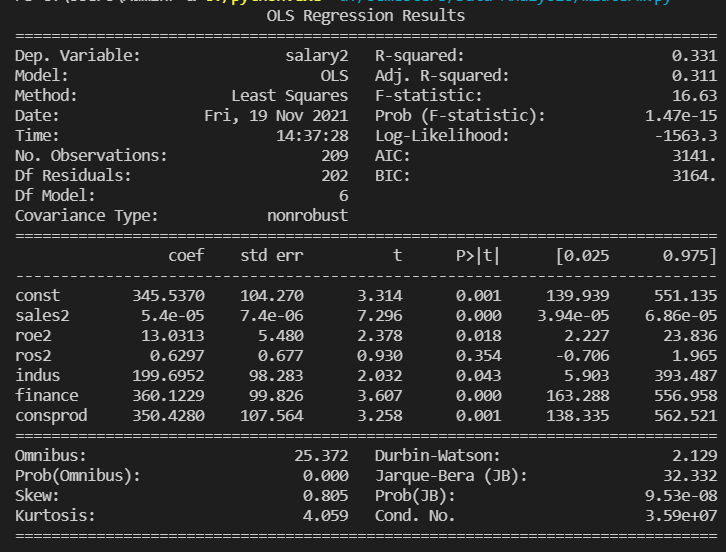
**Y = β0 + β1 sales + β2 roe + β3 ros + β4Z1 + β5Z2 + β6Z3 + ui (1)**

**Z1**: Dummy Variable. This variable = 1 if it is a indus company. If it is other type of company, the variable = 0.

**Z2**: Dummy Variable. This variable = 1 if it is a finance company. If it is other type of company, the variable = 0.

**Z3**: Dummy Variable. This variable = 1 if it is a consprod company. If it is other type of company, the variable = 0.

Variable *utility* will be reference, compared to 3 types of company above.



Y = 345.5370 + 5.4e-05sales + 13.0313 roe + 0.6297 ros + 199.6952Z1 + 360.1229Z2 + 350.4280Z3

In explaination for coefficents, other factors remain the same.

**β4** = 0.6297: coefficient of **Z1**, means that average salary of CEO in indus company will higher than average salary of CEO in utility company 0.6297 thousand dollars. Otherwise, if this number is negative, CEO’s average salary in indus company will have an amount of β4 less than average salary of others.

Similar to finance and consprod companies.

**β5 =** 199.6952 **:**coefficient of **Z2**, means that average salary of CEO in finance company will higher than average salary of CEO in utility company 199.6952 thousand dollars. Otherwise, if this number is negative, CEO’s average salary in finance company will have an amount of β5 less than average salary of others.

**β6** = 360.1229: coefficient of **Z3**, means that average salary of CEO in consprod company will higher than average salary of CEO in utility company 360.1229 thousand dollars. Otherwise, if this number is negative, CEO’s average salary in consprod company will have an amount of β6 less than average salary of others.

We can see Adj. R-squared here is 0.311, means that 31.1% salary can be explained by sales, roe, ros and types of company.

**Hypothesis review:**

With the model (1), if sales increase an unit, CEO’s salary will increase an amount of β1 (thousand dollars) with all 4 types of company. However, we consider the posibility if sales increase, CEO’s salary will increase differently according to type of company. In this hypothesis we use interactive dummy variables between sales and types of company.

In particular, coefficient of sales will be called **βi:**

**Y = β0 + βi sales + β4Z1 + β5Z2 + β6Z3 + ui**

Build a regression model:

βi = β2 + β3 Z1 + β7Z2 + β8Z3 (2)

If we take model (2) into model (1):

Y = β0  +β2sales + (β3 + β4)\*Z1\*sales + (β7 + β5) \*Z2\*sales + (β8 + β6)\* Z3\*sales + ui

There are three different interactive dummies: Z1\*sales, Z2\*sales, Z3\*sales.

Z1 = 1 -> This is indus company -> Y = (β0 + β4) + (β2 + β3 )sales + ui

Z2 = 1 -> This is finance company -> Y = (β0 + β7) + (β2 + β5)sales + ui

Z3 = 1 -> This is consprod company -> Y = (β0 + β6) + (β2 + β8 )sales + ui

We can explain the coefficient here:

**β0**: If sales = 0 then average CEO’s salary in utility company = β0.

**β4**: when there is no sales, the average salary of CEO in indus company is different from average salary of CEO in utility company an amount of β4.

**Β7**: when there is no sales, the average salary of CEO in finace company is different from average salary of CEO in utility company an amount of β7.

**Β6**: when there is no sales, the average salary of CEO in consprod company is different from average salary of CEO in utility company an amount of β4.

**β2**: When sales increase an unit then CEO’s salary in utility company will increase an amount of β2 (because all Z = 0).

**β3**: when sales in indus company increase an unit then CEO’s salary in indus company increase more than CEO’s salary in utility company an amount of β3 (if β3 > 0), increase less with an amount of β3 if β3 < 0.

**Β5**: when sales in finance company increase an unit then CEO’s salary in finance company increase more than CEO’s salary in utility company an amount of β5 (if β5 > 0), increase less with an amount of β5 if β5 < 0.

**Β8**: when sales in consprod company increase an unit then CEO’s salary in consprod company increase more than CEO’s salary in utility company an amount of β8 (if β8 > 0), increase less with an amount of β8 if β8 < 0.

# References:

1. Jeffrey M.Wooldridge (2012), Introductory Econometrics: A Modern Approach, <https://economics.ut.ac.ir/documents/3030266/14100645/Jeffrey_M._Wooldridge_Introductory_Econometrics_A_Modern_Approach__2012.pdf?fbclid=IwAR0V4Umd179Wn10PcQQiwLlCFxLVmwQI1yFeBdz03ngvOO01liIpEQmQxXc>

2. Andy (2019), DEV Community, Logarithmic Transformation in Linear Regression Models: Why & When, <https://dev.to/rokaandy/logarithmic-transformation-in-linear-regression-models-why-when-3a7c>

3. Cao Minh Ngoc (2019), VIBLO, Feature Engineering (Phần 3): Feature engineering với dữ liệu dạng phân loại (Categorical Data), <https://viblo.asia/p/feature-engineering-phan-3-feature-engineering-voi-du-lieu-dang-phan-loai-categorical-data-GrLZDQx2lk0>

4. Toankinhte UEL(2021), Biến giả và hồi quy với biến định tính <https://www.youtube.com/watch?v=qfW80NYlgN0&t=2235s&ab_channel=ToankinhteUEL>

5. Phùng Thanh Bình (2011), Giới thiệu mô hình hồi quy tuyến tính, <https://vi.vnp.edu.vn/wp-content/uploads/securepdfs/2020/01/Gujarati-2011-Ch%C6%B0%C6%A1ng-1-_-Gi%E1%BB%9Bi-thi%E1%BB%87u-m%C3%B4-h%C3%ACnh-h%E1%BB%93i-quy-tuy%E1%BA%BFn-t%C3%ADnh.pdf>

--------------------------------------------END--------------------------------------------------