Enron dataset description based on EDA result

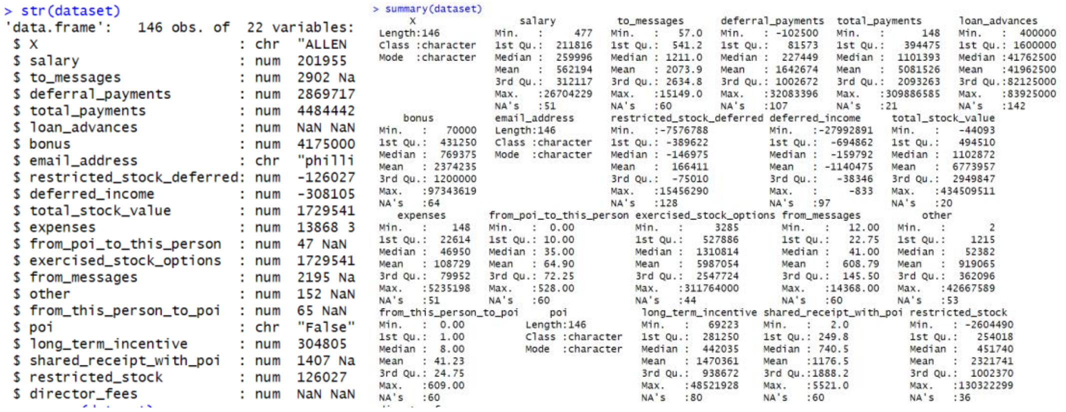
Name: Zheming Kang; UID: 3036195746;

1. Background Information

Enron dataset is from enron case, which is a famous financial fraud case. From Britannica, Enron company wrote unrealized future gains into current income statements and abused SPE(special purpose entities) distribution to hide loss.

2. Summary description of the dataset

The dataset consists of 22 variables and 146 data points, which means that the dataset provides us 22 features and 146 people from Enron Company. Among the 22 variables, there are 3 classes containing characters and 19 classes holding numbers. X, email address and poi are the 3 characteristic classes which present name, email address and the status of if the person is a Person of Interest respectively. There are 128 FALSE pois and 18 TRUE pois, which means 87.67% and 12.33% of the dataset. Figure 1 presents details of the dataset.



```
> str(dataset)
'data.frame':	146 obs. of  22 variables:
 $ X                        : chr  "ALLEN
 $ salary                   : num  201955
 $ to_messages              : num  2902 Na
 $ deferral_payments        : num  2869717
 $ total_payments           : num  4484442
 $ loan_advances            : num  NaN NaN
 $ bonus                    : num  4175000
 $ email_address            : chr  "philli
 $ restricted_stock_deferred: num  -126027
 $ deferred_income          : num  -308105
 $ total_stock_value        : num  1729541
 $ expenses                 : num  13868 3
 $ from_poi_to_this_person  : num  47 NaN
 $ exercised_stock_options  : num  1729541
 $ from_messages            : num  2195 Na
 $ other                    : num  152 NaN
 $ from_this_person_to_poi  : num  65 NaN
 $ poi                      : chr  "False"
 $ long_term_incentive      : num  304805
 $ shared_receipt_with_poi  : num  1407 Na
 $ restricted_stock         : num  126027
 $ director_fees            : num  NaN NaN
```

```
> summary(dataset)
      X                 salary            to_messages      deferral_payments  total_payments       loan_advances
 Length:146         Min.   :     477   Min.   :   57.0   Min.   : -102500   Min.   :      148   Min.   :  400000
 Class :character   1st Qu.:  211816   1st Qu.:  541.2   1st Qu.:   81573   1st Qu.:   394475   1st Qu.: 1600000
 Mode  :character   Median :  259996   Median : 1211.0   Median :  227449   Median :  1101393   Median :41762500
                    Mean   :  562194   Mean   : 2073.9   Mean   : 1642674   Mean   :  5081526   Mean   :41962500
                    3rd Qu.:  312117   3rd Qu.: 2634.8   3rd Qu.: 1002672   3rd Qu.:  2093263   3rd Qu.:82125000
                    Max.   :26704229   Max.   :15149.0   Max.   :32083396   Max.   :309886585   Max.   :83925000
                    NA's   :51         NA's   :60        NA's   :107        NA's   :21          NA's   :142
     bonus            email_address      restricted_stock_deferred deferred_income       total_stock_value
 Min.   :   70000   Length:146         Min.   :-7576788          Min.   :-27992891   Min.   :   -44093
 1st Qu.:  431250   Class :character   1st Qu.: -389622          1st Qu.:  -694862   1st Qu.:   494510
 Median :  769375   Mode  :character   Median : -146975          Median :  -159792   Median :  1102872
 Mean   : 2374235                      Mean   :  166411          Mean   : -1140475   Mean   :  6773957
 3rd Qu.: 1200000                      3rd Qu.:  -75010          3rd Qu.:   -38346   3rd Qu.:  2949847
 Max.   :97343619                      Max.   :15456290          Max.   :     -833   Max.   :434509511
 NA's   :64                            NA's   :128               NA's   :97          NA's   :20
    expenses        from_poi_to_this_person exercised_stock_options from_messages        other
 Min.   :    148   Min.   :  0.00          Min.   :     3285       Min.   :   12.00   Min.   :       2
 1st Qu.:  22614   1st Qu.: 10.00          1st Qu.:   527886       1st Qu.:   22.75   1st Qu.:    1215
 Median :  46950   Median : 35.00          Median :  1310814       Median :   41.00   Median :   52382
 Mean   : 108729   Mean   : 64.90          Mean   :  5987054       Mean   :  608.79   Mean   :  919065
 3rd Qu.:  79952   3rd Qu.: 72.25          3rd Qu.:  2547724       3rd Qu.:  145.50   3rd Qu.:  362096
 Max.   :5235198   Max.   :528.00          Max.   :311764000       Max.   :14368.00   Max.   :42667589
 NA's   :51        NA's   :60              NA's   :44              NA's   :60         NA's   :53
 from_this_person_to_poi      poi            long_term_incentive shared_receipt_with_poi restricted_stock
 Min.   :  0.00          Length:146         Min.   :   69223    Min.   :   2.0         Min.   : -2604490
 1st Qu.:  1.00          Class :character   1st Qu.:  281250    1st Qu.: 249.8         1st Qu.:   254018
 Median :  8.00          Mode  :character   Median :  442035    Median : 740.5         Median :   451740
 Mean   : 41.23                             Mean   : 1470361    Mean   :1176.5         Mean   :  2321741
 3rd Qu.: 24.75                             3rd Qu.:  938672    3rd Qu.:1888.2         3rd Qu.:  1002370
 Max.   :609.00                             Max.   :48521928    Max.   :5521.0         Max.   :130322299
 NA's   :60                                 NA's   :80          NA's   :60             NA's   :36
```

Figure 1. summary & structure of dataset

We can distinguish the attributes by poi status. The variables present the information of the person. Salary, bonus, loan advances, total stock value, expenses, other, long term incentive, total payment, exercised stock options, director fees and restricted stock describe the individual financial status. Deferral payments, restricted stock deferred, and deferred income provide the information of delayed financial statements. To messages, from poi to this person, from messages, from this person to poi, and shared receipt with poi counts the amount of messages transmitted from the person to others.

3. Univariate Analysis

After step1: Distinguish Attributes, we can choose some attributes that may have potential relationships to analyse. For this part, I split the attributes into 3 groups and analysis them respectively.

The first group contains "from poi to this person", and "from this person to poi". I group them for their close relationship to poi. The second group contains salary, bonus, total stock value, and expenses since they can show the financial information of the person. As the Enron case was caused by counting delayed income and hiding current loss, I choose the third group attributes which are deferral payments, income, and restricted stock. Figure 2 is the histogram of total stock value. To get rid of the influence of outliers, I plot the attributes in a reasonable range, which reflects a smooth distribution just like in figure 2. Then, I plot the distribution of poi using bar chart. We can see the distribution of two types of people: poi and non-poi in figure 3. Finally, I plot the box plot to see the distribution of personal financial attributes and defer-relative attributes.

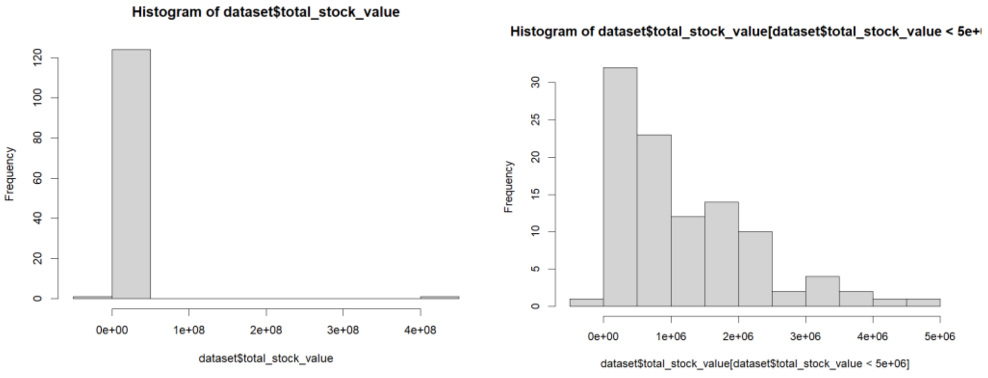


Figure 2. distribution of total stock value/ value < 5e6

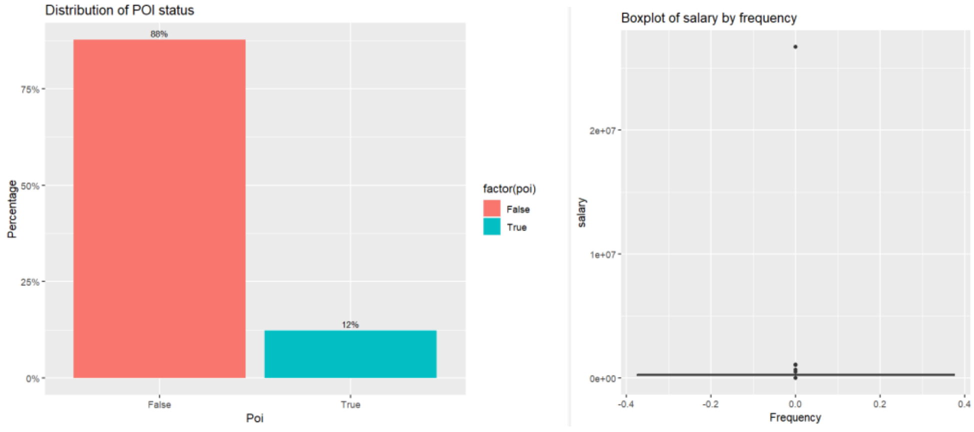


Figure 3. distribution by poi and box plot of salary

4. Bi-/Multi-variate analysis

Knowing the distribution of single attributes, we should move to bi-variate analysis. I plot the potentially useful attributes with poi feature and find “from poi to this person”, “total stock value”, “deferral payments”, “deferred income”, and “restricted stock deferred” have different distributions in poi and non-poi. Other attributes reflect similar distributions so we won’t keep them in multi-variate analysis.

Locking on these attributes, we could compute the correlation matrix and plot the matrix in heatmap. Figure 4 shows the correlation matrix. From the matrix, we could see the deferred income is highly negatively related with deferred payments, total stock value and restricted stock deferred. What’s more, the deferred income is slightly positive related to the “from poi to this person”. An interesting fact is, the total stock value is negatively related to from poi to this person. From the information above, we could infer that if a person who contains low stock, he might receive many emails from poi. We could also infer that if a person need to pay much back, if his deferred income is highly negative, he might have low total stock value, restricted stock deferred and deferral payment.

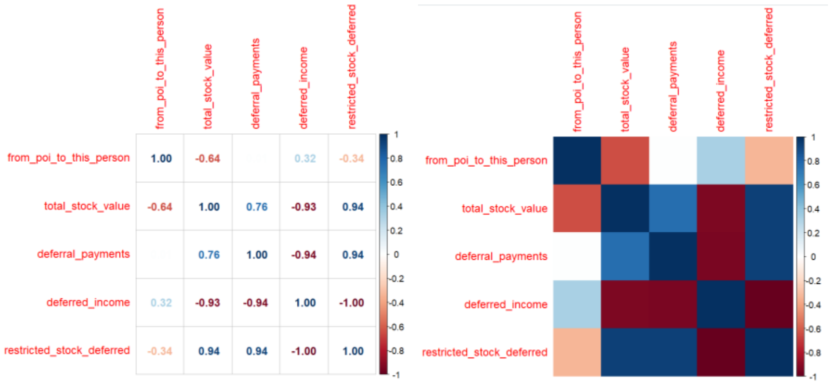


Figure 4. correlation matrix

5. Missing data/ Outlier analysis

I counted the missing value distribution of various indicators, and the image is as shown in the figure 5.

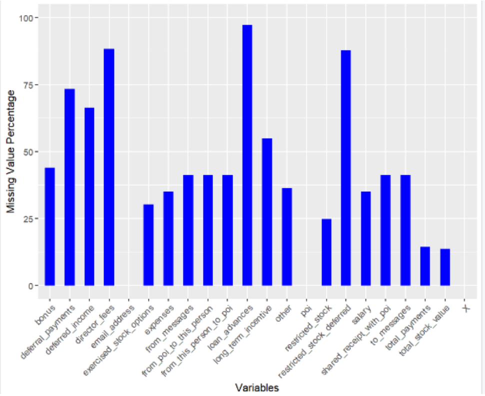
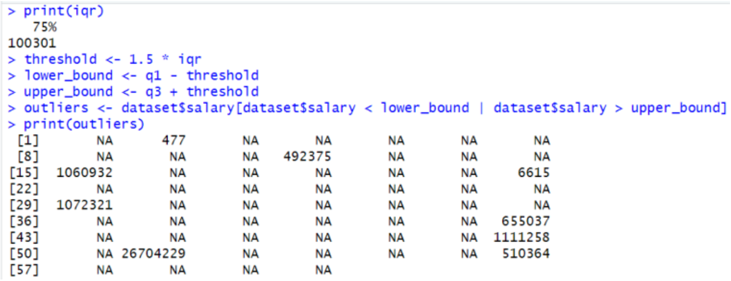


Figure 5. distribution of missing value

For univariate outliers, I compute the IQR of salary as an instance. In figure 5 we could see the iqr is 100301. Then, let's assume threshold is 1.5*iqr, and find all outliers.



```
> print(iqr)
   75%
100301
> threshold <- 1.5 * iqr
> lower_bound <- q1 - threshold
> upper_bound <- q3 + threshold
> outliers <- dataset$salary[dataset$salary < lower_bound | dataset$salary > upper_bound]
> print(outliers)
 [1]       NA      477       NA       NA       NA       NA       NA
 [8]       NA       NA       NA   492375       NA       NA       NA
[15]  1060932       NA       NA       NA       NA       NA     6615
[22]       NA       NA       NA       NA       NA       NA       NA
[29]  1072321       NA       NA       NA       NA       NA       NA
[36]       NA       NA       NA       NA       NA       NA   655037
[43]       NA       NA       NA       NA       NA       NA  1111258
[50]       NA 26704229       NA       NA       NA       NA   510364
[57]       NA       NA       NA       NA
```

Figure 6. IQR and outliers of salary

For Bi-/Multi-varite outliers, we could find them by histogram in session 4. From figure 2 left part, we could see the outlier in the right hand side. Actually the outlier is data "total" in salary. In figure 3 right hand side, we could see the outlier from the box plot. The outlier is far away from the data piece.

There are a lot of figures when I doing EDA, the figures above are just examples.

6. References

Enron scandal | Summary, Explained, History, & Facts | Britannica

Enron Person of Interest Dataset (kaggle.com)