Factors Affecting Salaries in the IT Industry

Group members：
Yanqing Zhong 1155208169
Wei Wang 1155206916
Yifei Chen 1155208163
Chang LIU 1155209013
Zhenyu Xu 1155205225
Xinrui Song 1155208111

Data Description

The columns in the dataset include:
work_year, experience_level, employment_type, job_title, salary, salary_currency, salary_in_usd, employee_residence, remote_ratio, company_location, company_size. The dataset covers the years 2020 to 2023 and has 2054 observations, with the salary amount as the dependent variable (y) and the remaining columns as independent variables (X), which can be used to explore and analyze the relationship between salary and these factors.

Data Cleaning and Some Preparations

In this part, we processed the data in the original table and divide the data processing into several steps.

Handling Missing Values: Identify and handle missing values in the data, that can be used here is to delete rows or columns with 'NA' values.

Handling Duplicate Values: Detect and remove duplicate rows or columns in the data to avoid introducing bias during the analysis process.

Data Type Conversion: Ensure that each column of data has the correct type.

Handling Outliers: In this case, they are replaced with the median or mean value.

Data Standardization: Adjust all data to the same scale or range. Here, we standardize all salaries to the unit of US dollars.

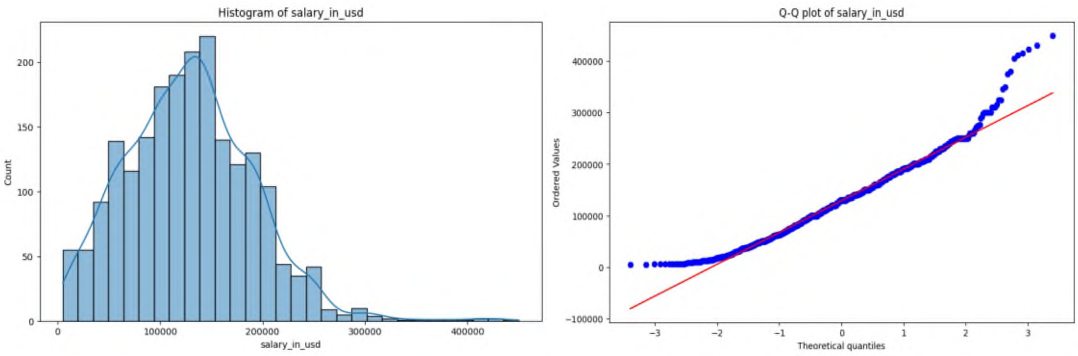
Data Analysis

We took three models to analyze the data, including OSL regression model, and ANOVA method.

OSL Regression Model

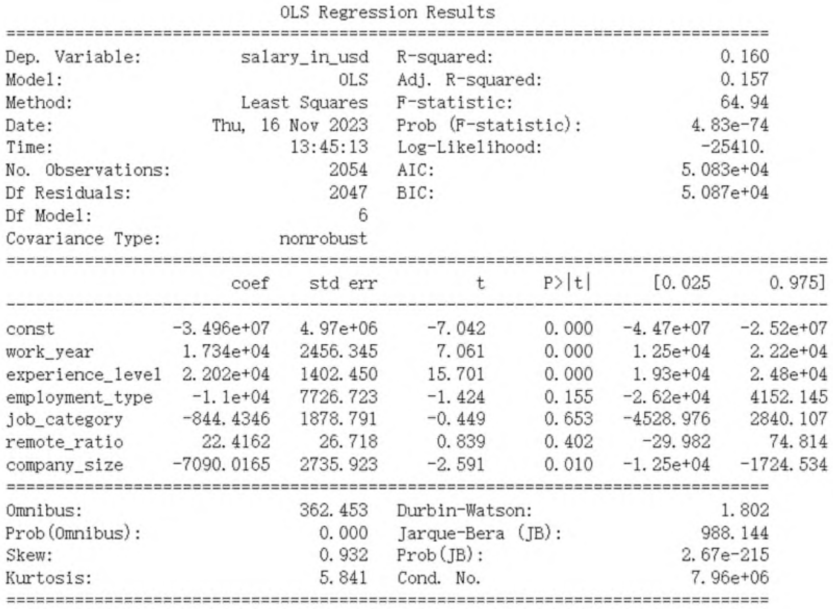
Since there are so many different job titles that it is difficult to analyze the data, we have divided the job titles into five categories for analysis.

After categorization, we tested for normality using the Shapiro-Wilk test. Then, we plotted a histogram of 'salary_in_usd' as well as a Q-Q plot (Quantile-Quantile plot). Histogram shows the distribution pattern of the data and by looking at it, we can see that the data conforms to a normal distribution.



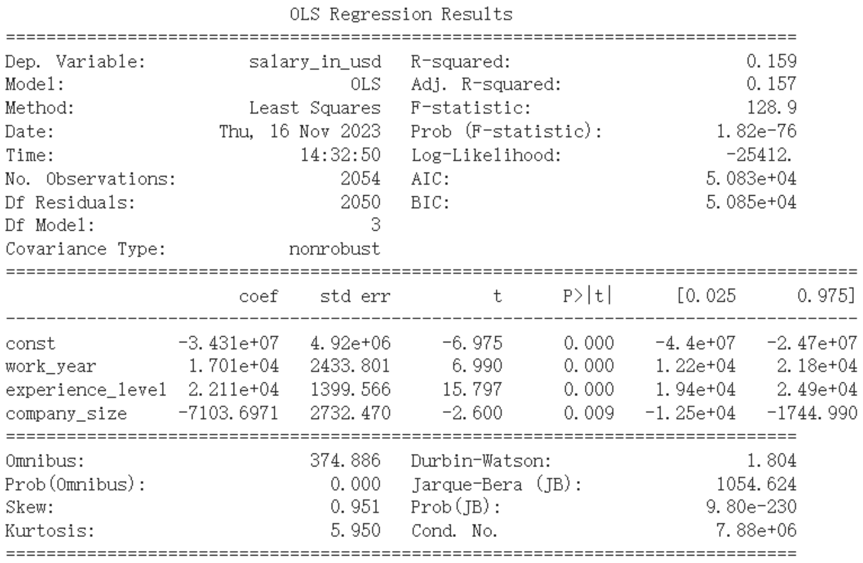
The dependent variable is 'salary_in_usd' and the independent variables are 'work_year', 'experience_level','company_size',’emplotment_type’,‘job_category’,’remote_ratio’.

The'experience_level', and‘company_size' are categorical variables which are converted to label codes.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          salary_in_usd   R-squared:                       0.160
Model:                            OLS   Adj. R-squared:                  0.157
Method:                 Least Squares   F-statistic:                     64.94
Date:                Thu, 16 Nov 2023   Prob (F-statistic):           4.83e-74
Time:                        13:45:13   Log-Likelihood:                -25410.
No. Observations:                2054   AIC:                         5.083e+04
Df Residuals:                    2047   BIC:                         5.087e+04
Df Model:                           6
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -3.496e+07   4.97e+06     -7.042      0.000   -4.47e+07   -2.52e+07
work_year         1.734e+04   2456.345      7.061      0.000    1.25e+04    2.22e+04
experience_level  2.202e+04   1402.450     15.701      0.000    1.93e+04    2.48e+04
employment_type    -1.1e+04   7726.723     -1.424      0.155   -2.62e+04    4152.145
job_category      -844.4346   1878.791     -0.449      0.653   -4528.976    2840.107
remote_ratio        22.4162     26.718      0.839      0.402     -29.982      74.814
company_size     -7090.0165   2735.923     -2.591      0.010   -1.25e+04   -1724.534
==============================================================================
Omnibus:                      362.453   Durbin-Watson:                   1.802
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              988.144
Skew:                           0.932   Prob(JB):                    2.67e-215
Kurtosis:                       5.841   Cond. No.                     7.96e+06
==============================================================================
```

Through OLS regression analysis, we found that the p-value of ‘emplotment_type’,‘job_category’,’remote_ratio’are greater than 0.05, they have no significant effect on the dependent variable, so we adjusted the independent variable to 'work_year', 'experience_level','company_size'. The p-value of each independent variable is less than 0.05, and it can be assumed that all these independent variables have a significant effect on the dependent variable.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          salary_in_usd   R-squared:                       0.159
Model:                            OLS   Adj. R-squared:                  0.157
Method:                 Least Squares   F-statistic:                     128.9
Date:                Thu, 16 Nov 2023   Prob (F-statistic):           1.82e-76
Time:                        14:32:50   Log-Likelihood:                -25412.
No. Observations:                2054   AIC:                         5.083e+04
Df Residuals:                    2050   BIC:                         5.085e+04
Df Model:                           3
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const            -3.431e+07   4.92e+06     -6.975      0.000    -4.4e+07   -2.47e+07
work_year         1.701e+04   2433.801      6.990      0.000    1.22e+04    2.18e+04
experience_level  2.211e+04   1399.566     15.797      0.000    1.94e+04    2.49e+04
company_size     -7103.6971   2732.470     -2.600      0.009   -1.25e+04   -1744.990
==============================================================================
Omnibus:                      374.886   Durbin-Watson:                   1.804
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1054.624
Skew:                           0.951   Prob(JB):                    9.80e-230
Kurtosis:                       5.950   Cond. No.                     7.88e+06
==============================================================================
```

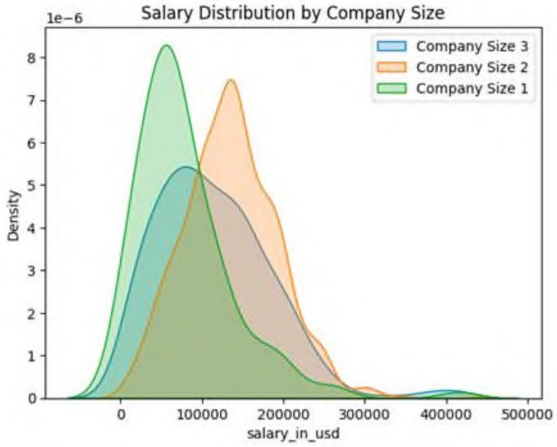
The relationship between the independent variables and the dependent variable can be expressed as follows:

Salary_in_usd = **-3.496e+07+1.734e+04 work_year** + **2.202e+04 experience_level - 7090.0165 company_size**

The formula shows a positive relationship between 'work_year', 'experience_level' and 'salary_in_usd' and a negative relationship between 'company_size' and 'salary_in_usd'.

ANOVA Method

Before conducting ANOVA statistical methods, it is necessary to test the data for compliance with normal distribution conditions. By conducting Shapiro-Wilk normality tests on the data and drawing KDE graphs, (By weighted the outlier)the actual salary distribution shape was observed, and the data conforms to a normal distribution.



A one-way ANOVA is used to compare the means of multiple groups, but focuses on whether one of the groups is significantly higher than the others. This is common in real-world problems, such as examining whether company size has a significant positive effect on salary.

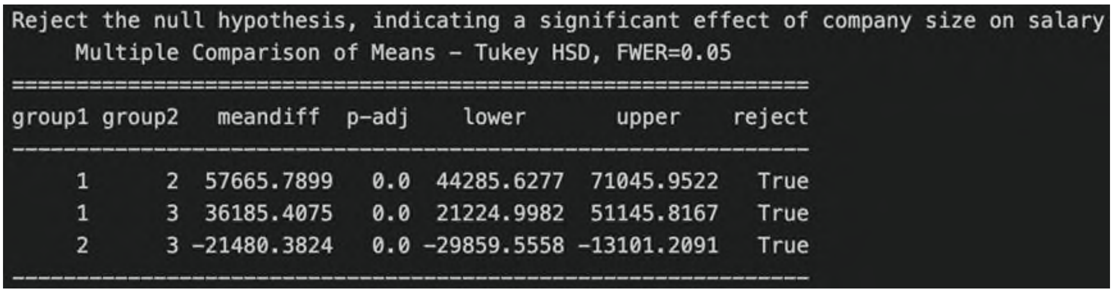
Then we mapping the column 'company_size' as a numeric value: Mapping to values preserves the company size relationship and simplify the analysis process.

Then we choose Tukey's HSD for multiple comparisons: It is a method of overall comparison to determine which groups have differences.It reduce the risk of Type I errors, effectively handling multiple comparisons, avoiding cumulative significance issues.Tukey's HSD for multiple comparisons is an effective method, but attention needs to be paid to the issue of cumulative significance. When making multiple comparisons, it may be necessary to consider adjusting the significance level to control the overall error rate.

In summary, through the analysis of One-Way ANOVA, we observe a significant impact of company size on salary.

Further conducting Tukey's HSD multiple comparison, we identify significant salary differences among different company sizes.

- Company size 1 and company size 2 is $57,665.79, rejecting the null hypothesis, indicating a significant difference between these two groups.
- Company size 1 and company size 3 is $36,185.41, also rejecting the null hypothesis, indicating a significant difference between these two groups.
- Company size 2 and company size 3 is -$21,480.38, similarly rejecting the null hypothesis, indicating a significant difference between these two groups.



```
Reject the null hypothesis, indicating a significant effect of company size on salary
     Multiple Comparison of Means - Tukey HSD, FWER=0.05
===============================================================
group1 group2   meandiff  p-adj    lower       upper    reject
---------------------------------------------------------------
     1      2  57665.7899   0.0  44285.6277  71045.9522   True
     1      3  36185.4075   0.0  21224.9982  51145.8167   True
     2      3 -21480.3824   0.0 -29859.5558 -13101.2091   True
---------------------------------------------------------------
```

These results suggest that company size does indeed have a significant impact on salary levels. Specifically, compared to small companies, medium-sized companies exhibit higher average salaries, while large companies tend to have lower average salaries. This may reflect advantages in competitive salary offerings for medium-sized companies, whereas large companies might prioritize other benefits or development opportunities.

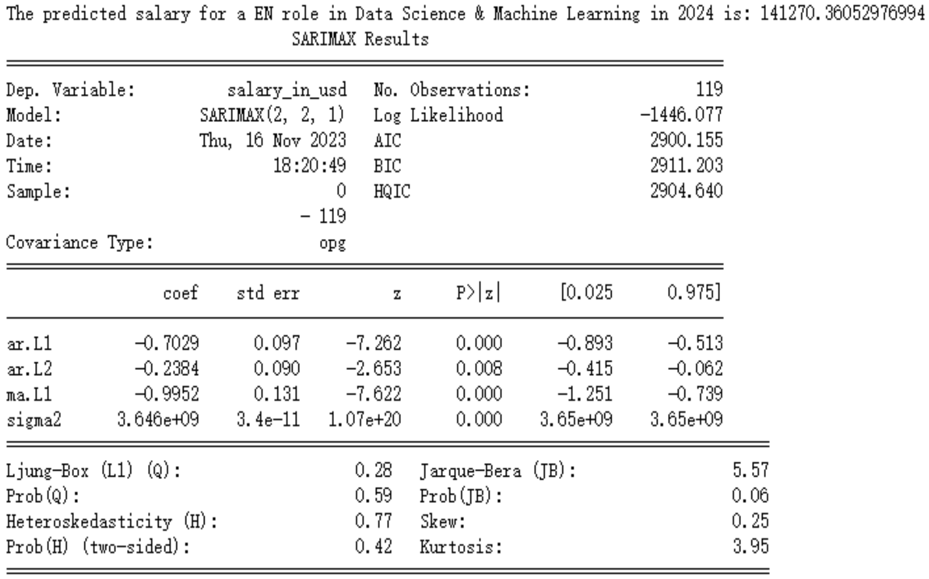
In conclusion, this algorithm indicate that company size significantly influences salary levels. This discovery not only holds statistical significance but also carries substantial practical implications. Further exploration into the dynamics of company size can assist businesses in tailoring compensation strategies to meet the diverse needs and expectations across different-sized enterprises.

Prediction

we used a time series model to predict salaries based on specific job types, years worked, employment types, and experience levels.

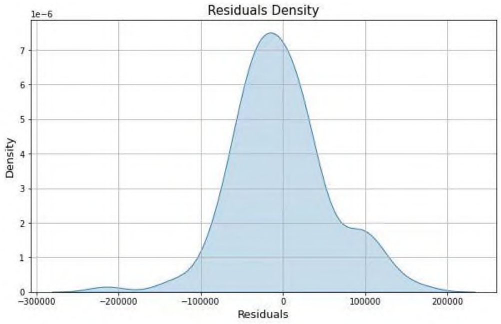
The function first filters data from the entire dataset that matches the specified job category, employment type, and experience level. If no data matches these criteria (i.e., there are no historical salary data that meet the conditions), the function will return a message indicating there are no available historical salary data. Then, we set ‘work_year’ as the index to facilitate subsequent time series analysis. A stationarity check function is used here. If the data is not stationary, the function will difference the data until it becomes stationary. Finally, we trained the model using the SARIMAX model. If the input ‘work_year’ exists in the data, the function will return the salary data for that year. If the predicted ‘work_year’ exceeds the ‘work_year’ in the dataset, the trained model will be used for prediction.

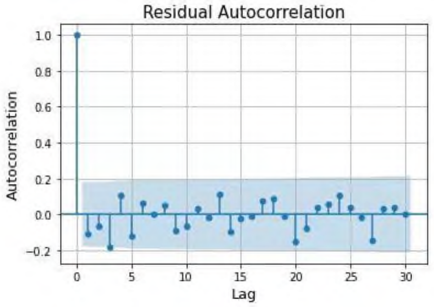
Here are our prediction results. In 2024, the anticipated salary for a data science and machine learning role at the "EN" experience level is $141,270. It can be seen that all the parameter P values are significantly less than 0.05, indicating that these parameters are significant in the model.



```
The predicted salary for a EN role in Data Science & Machine Learning in 2024 is: 141270.36052976994
                               SARIMAX Results
==============================================================================
Dep. Variable:           salary_in_usd   No. Observations:                  119
Model:                SARIMAX(2, 2, 1)   Log Likelihood               -1446.077
Date:                 Thu, 16 Nov 2023   AIC                           2900.155
Time:                         18:20:49   BIC                           2911.203
Sample:                              0   HQIC                          2904.640
                                 - 119
Covariance Type:                   opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1         -0.7029      0.097     -7.262      0.000      -0.893      -0.513
ar.L2         -0.2384      0.090     -2.653      0.008      -0.415      -0.062
ma.L1         -0.9952      0.131     -7.622      0.000      -1.251      -0.739
sigma2      3.646e+09    3.4e-11   1.07e+20      0.000    3.65e+09    3.65e+09
===================================================================================
Ljung-Box (L1) (Q):                   0.28   Jarque-Bera (JB):                 5.57
Prob(Q):                              0.59   Prob(JB):                         0.06
Heteroskedasticity (H):               0.77   Skew:                             0.25
Prob(H) (two-sided):                  0.42   Kurtosis:                         3.95
===================================================================================
```

The y-axis of the autocorrelation plot represents the autocorrelation coefficient, and the x-axis represents the lag value.The blue box represents the confidence interval, and all the autocorrelation coefficients of lag values in the residual autocorrelation plot are within the confidence interval. This means that the model has captured all correlations in the data and therefore the model residuals do not show significant autocorrelation.


Residuals Density
1e−6
Density
Residuals


Residual Autocorrelation
Autocorrelation
Lag