Unraveling the Trajectory of Data Science Salaries in the United States: A Comprehensive Analysis from 2020 to 2023 with Future Salary Projections

Jiayong Chen[1,a,*], Zhengbin Song[2,b], Ching Hoi Lam[3,c]

[1]Management of Information System, Guangdong University of Finance, Guangzhou, 510630, China

[2]Department of Information Systems, University of Kansas, Lawrence, KS, 66045, United States

[3]Alliance Manchester Business School, University of Manchester, Manchester, M13 9PL, Great Britain

a. 201541101@m.gduf.edu.cn, b. z554s369@ku.edu, ching, c. lam-5@student.manchester.ac.uk

**corresponding author*

Abstract: The purpose of this study is to analyze the impact of different positions, levels of expertise and company size on salary levels in the field of data science and to make salary projections for data science professionals in 2024. This research project can help data science professionals to understand what are the important factors that affect salary levels and to understand the salary environment and trends in the data science field in 2024. With the explosive growth of big data, the oversupply of data science jobs and the precise hiring needs have led to significant changes in the salaries of data science related careers from year to year. Data is thoroughly cleaned and preprocess sed to maintain data quality and consistency, including handling missing values and removing outliers. Descriptive analysis techniques were then used to understand the current state of data science salaries, calculating data such as mean, median and standard deviation. Time series modeling was used to determine how key factors affect pay levels over time. To further investigate salary trends, ARIMA was applied to visualize the evolution of data science salaries from 2020 to 2023, and then to forecast average salary levels for different positions in the data science field in 2024. In summary, the important factors affecting data science salaries and the trend of salaries for different careers in data science in 2024 are analyzed, and a detailed analysis is provided with salary as a key factor to provide valuable recommendations for data science stakeholders.

Keywords: data science salaries, salary projections, time series modeling, descriptive analysis

1. Introduction

Data science has become a transformative field that harnesses the power of data to discover valuable insights and results, make informed decisions, and drive innovation across industries. As companies become increasingly aware of the importance of data science, the demand for data science professionals has skyrocketed. At the same time, data science employment opportunities are growing, and data science talent is clustering in the U.S. as a global centre for technology and innovation.

Apply big data principles and methods into this field to improve business efficiency and make wiser decisions, which has become a frequent topic of discussion among scholars [1].

This study focuses on the evolution of Data Science salaries in the US from 2020 to 2023 and reveals the trends and patterns that emerge during this period. Understanding the evolution of salaries provides insights into the dynamics of the job market, the state of the economy, and the factors that influence the salary levels of data science professionals. The job market for data science professionals is dynamic and diverse, and we have categorized these different careers into four distinct areas, including engineers, analysts, scientists, and architects. These professionals are sought after by major companies for their expertise in working with big data, implementing advanced analytics, and building machine learning models to drive decision making.

Understanding salary trends is not only critical for job seekers and employees looking for fair compensation packages, but also for major organizations and businesses that need this data to give talent in the data science field a fair match that can be used to retain such talent. Analyzing salary trends can provide insights into the growth and maturity of the field, allow those working in data science to make better choices about the careers they want to pursue, and give others an understanding of the current state and future of data science as a discipline, as well as organizations and companies with a need for data science talent an understanding of salary trends.

In this study, we collected data from job postings and salary reports for different data science positions in the United States from 2020 to 2023. We first analyze the experience level (e.g., entry level, mid-level, senior level), and company size (small to large), as well as the impact of different Job Title on salary. Data analysts, data engineers, machine learning engineers, and data scientists are all job titles related to data science [2]. Linear regression was used to conduct exploratory analyses of the data to test the fit between different variables and salary, and control variables were used to test the fit between their variables and salary under different combinations of variables. We use these analyses to spread out the interactions between the various data variables, to distinguish between levels of experience, company size, and the relationship between different occupations and salary trends. In addition, to forecast future salary trends, we used time series modeling and historical salary data to produce a forecasting model that explains the dynamic nature of salaries over time. In this way, we were able to make a basic forecast of data science salaries in the United States in 2024. Through multiple linear regression and time series modeling, we hope to provide comprehensive insights that will lead to meaningful analytics for job seekers, employees, and employers in the data science field.

2. Literature Review

2.1. Factors Affecting Data Science Salaries

One of the factors why Data Scientists are highly demanded lies in its role, in which its broad and general, while being essential in each operational process within a business [3]. Tee Zhen Quan and Mafas Raheem’s journal on Data Science Salary Prediction conducted a study regarding the Top 10 Most Current Demanded Data Science Jobs and discovered a pattern that broader job tasks and a less job specific title result in higher demand, which in turn also shows that the number or range of tasks play a role in determining the salaries of data scientists. This is also normal since in general, higher job status result in higher salary and demand.

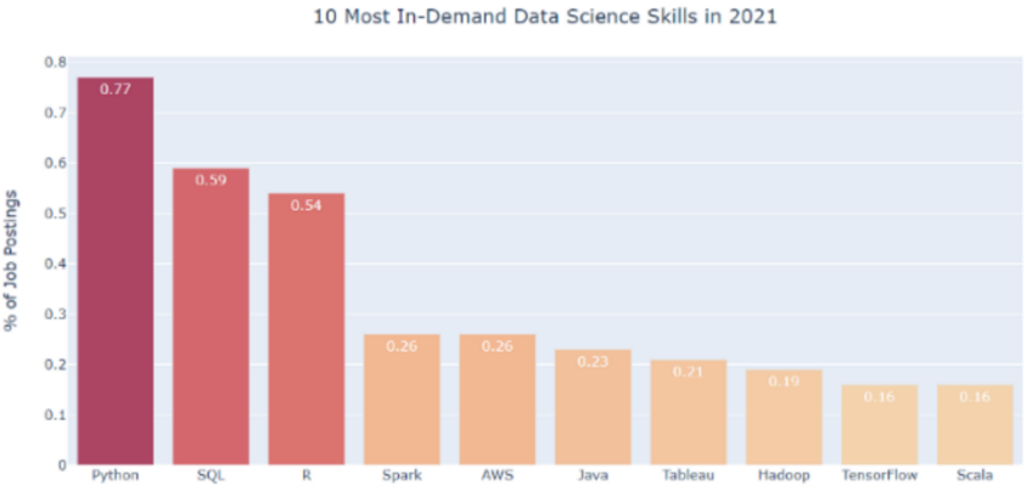


Figure 1: Top 10 Data Science Skills in 2021

A range of skills acquired by job titles and their relationship to salary amount were also conducted. The total demand in 2021 were documented, and figure 1 shows that Python, SQL, and R were the hottest skillsets demanded, while other programming languages share a similar low-demand trend.

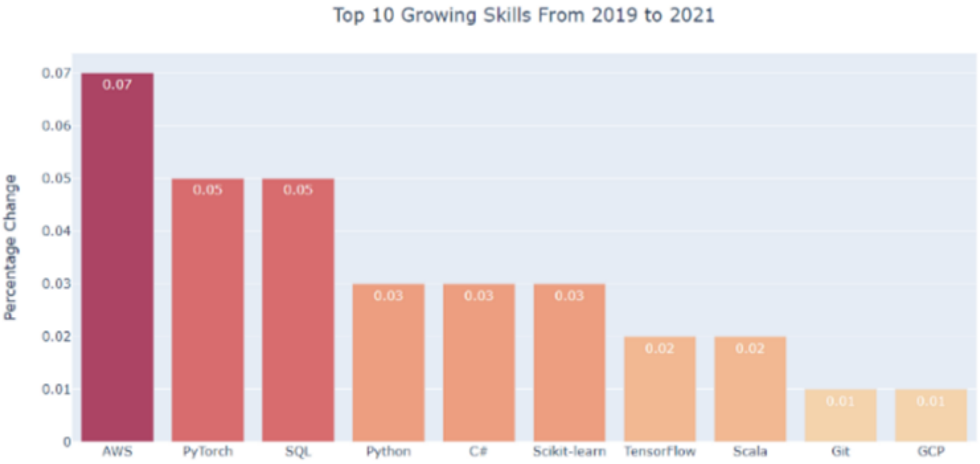


Figure 2: Top 10 Growing Data Science Skills between 2019 and 2021

Figure 2 shows that an impressive detail was that the rate of demand growth for each skill set was also considered, which could play an important role in influencing the different titles within Data Science and the Data Science Industry as a whole.

2.2. Models to Predict Data Science Salary

Unfortunately, the Salary Prediction Model by [4] Tee Zhen Quan and Mafas Raheem was developed through the SAS Enterprise Miner platform, lacking both transparency and could generate biased results. Moreover, not many details were stated for the data preparation except mentioning a few nodes containing "high-performance models and algorithms". For the results, salary increases across the "job hierarchy ladder", from the lowest junior level to the highest principal level, reaching expected results. Similarly, in our research, we also analyzed the level of "expertise level" on salary.

Most of the research paper conducted on Data Scientists' salary focus more on the skill aspects but didn't consider one of the key determinators of salary – the company. In order to work on this, we decided to include "company size", which could indirectly show the power or status of the company. With this information, future studies regarding data scientists' influence on company status or progression within its industry could be conducted after a certain amount of time, with sufficient develop for the data science field and given a noticeable influence by them.

2.3. Methods for Time-series Prediction

In time series analysis, ARIMA models are flexible and popular since it offers accurate short-run predictions for quantity-supported datasets [5]. Mohamed Reda Abonazel and Ahmed Ibrahim Abd-Elftah used ARIMA model in-depth to forecast Egyptian GDP and appreciated its "long-term memory" on using previous data of the year before. In our research about data scientists' salary and job titles within forecasts, we also noticed how the "long-term memory" benefit us to find complicated details which converged into a long-term trend, resulting in an essential key towards accurate predictions.

Computer Science is a well-developed field, thus various algorithms about salary prediction with different goals are already a popular subject. Unlike our goal of predicting the general trend of salary for Data Science jobs [6], this research paper aims to predict a concise salary for the specific CS position, due to its competitive nature and clear job performances. Three algorithms, namely Naive Bayes, Support Vector Machine (SVM), and Random Forest, are employed for salary prediction. These algorithms are good at real-time prediction, which satisfies their environment, while our model aims to give a general but accurate idea of the future development of Data Science.

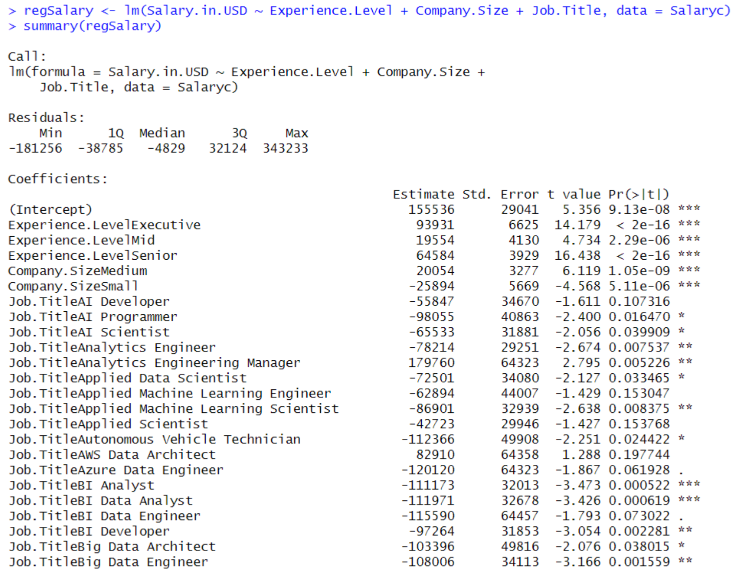
3. Methodology

This dataset was chosen for this study because it comes from a reliable data source and spans a long enough period of time to cover the salaries of data science professionals from 2020 to 2023. It also includes different job titles, levels of expertise, firm size, and experience levels, as well as firm location and employee nationality. Firstly, the data is cleaned and preprocessed. Since all the currencies in the data were standardized and one of the columns "Salary in USD" and USD as the salary currency, we deleted the columns "Salary" and "Salary Currency", and then deleted and processed the data for missing values and outliers. The data was then subjected to exploratory analyses to understand the distribution and relationship between the different variables. From this step, important factors were identified to understand their impact on the salary income of data science professionals. Linear Regression is an algorithm of machine learning based on supervised learning scheme. Linear regression carries out a task that may predict the value of a dependent variable (y) on basis of an independent variable (x) that is given. Therefore, this kind of regression technique looks for a linear type of relationship between input x and output y [7]. Linear regression models were then used to establish the relationship between experience level, job title, company size and salary, and these factors were used to build models to determine what factors influence salary levels in 2020-2023 and salary trends during this period, and diagnostic models were used to diagnose the feasibility of the regression models. We then used time series analysis to apply time series models such as ARIMA (Auto Regressive Aggregate Moving Average) to predict the salary trends for Data Science jobs in 2024 and used historical data on salaries to see where salaries were going for different occupations between 2020 and 2023.

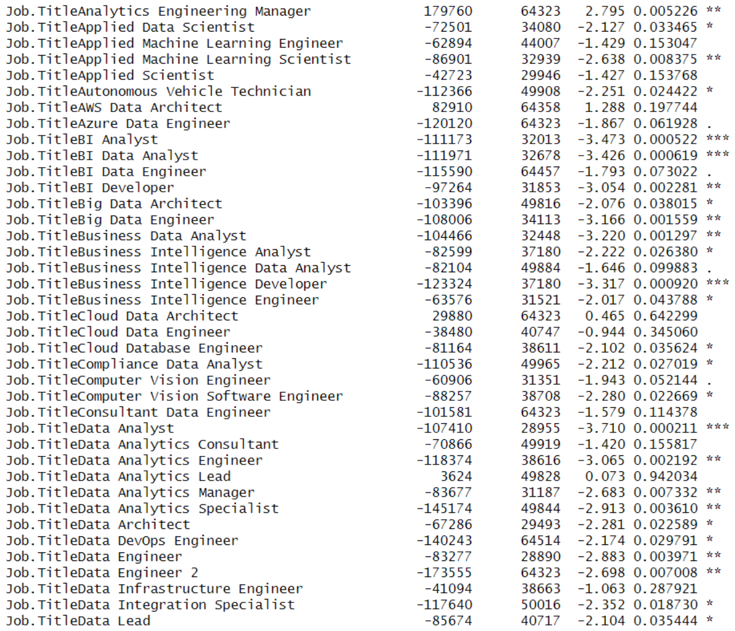
4. Results and Discussion

4.1. Correlation Analysis

The data were analyzed using a linear regression model based on three key factors: "Experience Level" and "Company Size" and "Job Title", and the coefficients and statistical significance of the predictor variables were assessed. The results show a significant association between data science salaries and the predictor variables.

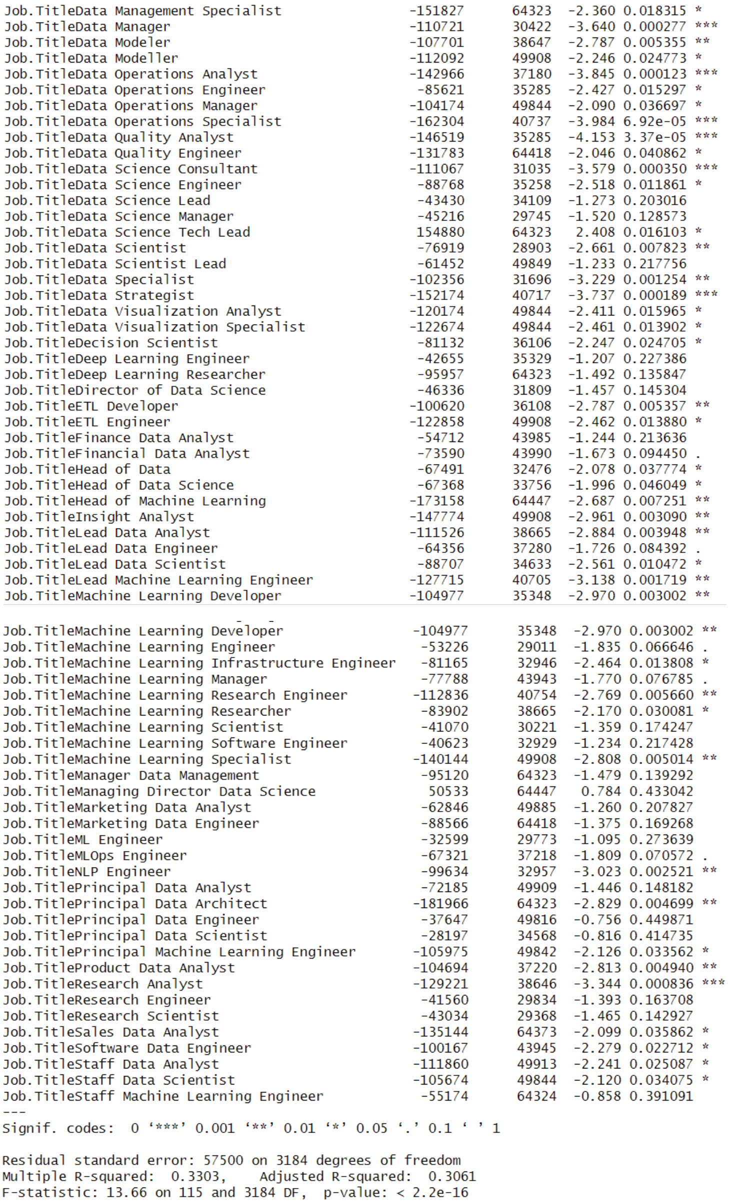


```
> regSalary <- lm(Salary.in.USD ~ Experience.Level + Company.Size + Job.Title, data = Salaryc)
> summary(regSalary)

Call:
lm(formula = Salary.in.USD ~ Experience.Level + Company.Size +
    Job.Title, data = Salaryc)

Residuals:
    Min      1Q  Median      3Q     Max
-181256  -38785   -4829   32124  343233

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                     155536      29041   5.356 9.13e-08 ***
Experience.LevelExecutive                        93931       6625  14.179  < 2e-16 ***
Experience.LevelMid                              19554       4130   4.734 2.29e-06 ***
Experience.LevelSenior                           64584       3929  16.438  < 2e-16 ***
Company.SizeMedium                               20054       3277   6.119 1.05e-09 ***
Company.SizeSmall                               -25894       5669  -4.568 5.11e-06 ***
Job.TitleAI Developer                           -55847      34670  -1.611 0.107316
Job.TitleAI Programmer                          -98055      40863  -2.400 0.016470 *
Job.TitleAI Scientist                           -65533      31881  -2.056 0.039909 *
Job.TitleAnalytics Engineer                     -78214      29251  -2.674 0.007537 **
Job.TitleAnalytics Engineering Manager          179760      64323   2.795 0.005226 **
Job.TitleApplied Data Scientist                 -72501      34080  -2.127 0.033465 *
Job.TitleApplied Machine Learning Engineer      -62894      44007  -1.429 0.153047
Job.TitleApplied Machine Learning Scientist     -86901      32939  -2.638 0.008375 **
Job.TitleApplied Scientist                      -42723      29946  -1.427 0.153768
Job.TitleAutonomous Vehicle Technician         -112366      49908  -2.251 0.024422 *
Job.TitleAWS Data Architect                      82910      64358   1.288 0.197744
Job.TitleAzure Data Engineer                   -120120      64323  -1.867 0.061928 .
Job.TitleBI Analyst                            -111173      32013  -3.473 0.000522 ***
Job.TitleBI Data Analyst                       -111971      32678  -3.426 0.000619 ***
Job.TitleBI Data Engineer                      -115590      64457  -1.793 0.073022 .
Job.TitleBI Developer                           -97264      31853  -3.054 0.002281 **
Job.TitleBig Data Architect                    -103396      49816  -2.076 0.038015 *
Job.TitleBig Data Engineer                     -108006      34113  -3.166 0.001559 **
```



```
Job.TitleAnalytics Engineering Manager             179760      64323   2.795 0.005226 **
Job.TitleApplied Data Scientist                    -72501      34080  -2.127 0.033465 *
Job.TitleApplied Machine Learning Engineer         -62894      44007  -1.429 0.153047
Job.TitleApplied Machine Learning Scientist        -86901      32939  -2.638 0.008375 **
Job.TitleApplied Scientist                         -42723      29946  -1.427 0.153768
Job.TitleAutonomous Vehicle Technician            -112366      49908  -2.251 0.024422 *
Job.TitleAWS Data Architect                         82910      64358   1.288 0.197744
Job.TitleAzure Data Engineer                      -120120      64323  -1.867 0.061928 .
Job.TitleBI Analyst                               -111173      32013  -3.473 0.000522 ***
Job.TitleBI Data Analyst                          -111971      32678  -3.426 0.000619 ***
Job.TitleBI Data Engineer                         -115590      64457  -1.793 0.073022 .
Job.TitleBI Developer                              -97264      31853  -3.054 0.002281 **
Job.TitleBig Data Architect                       -103396      49816  -2.076 0.038015 *
Job.TitleBig Data Engineer                        -108006      34113  -3.166 0.001559 **
```

```
Job.TitleBusiness Data Analyst                    -104466      32448  -3.220 0.001297 **
Job.TitleBusiness Intelligence Analyst             -82599      37180  -2.222 0.026380 *
Job.TitleBusiness Intelligence Data Analyst        -82104      49884  -1.646 0.099883 .
Job.TitleBusiness Intelligence Developer          -123324      37180  -3.317 0.000920 ***
Job.TitleBusiness Intelligence Engineer            -63576      31521  -2.017 0.043788 *
Job.TitleCloud Data Architect                       29880      64323   0.465 0.642299
Job.TitleCloud Data Engineer                       -38480      40747  -0.944 0.345060
Job.TitleCloud Database Engineer                   -81164      38611  -2.102 0.035624 *
Job.TitleCompliance Data Analyst                  -110536      49965  -2.212 0.027019 *
Job.TitleComputer Vision Engineer                  -60906      31351  -1.943 0.052144 .
Job.TitleComputer Vision Software Engineer         -88257      38708  -2.280 0.022669 *
Job.TitleConsultant Data Engineer                 -101581      64323  -1.579 0.114378
Job.TitleData Analyst                             -107410      28955  -3.710 0.000211 ***
Job.TitleData Analytics Consultant                 -70866      49919  -1.420 0.155817
Job.TitleData Analytics Engineer                  -118374      38616  -3.065 0.002192 **
Job.TitleData Analytics Lead                         3624      49828   0.073 0.942034
Job.TitleData Analytics Manager                    -83677      31187  -2.683 0.007332 **
Job.TitleData Analytics Specialist                -145174      49844  -2.913 0.003610 **
Job.TitleData Architect                            -67286      29493  -2.281 0.022589 *
Job.TitleData DevOps Engineer                     -140243      64514  -2.174 0.029791 *
Job.TitleData Engineer                             -83277      28890  -2.883 0.003971 **
Job.TitleData Engineer 2                          -173555      64323  -2.698 0.007008 **
Job.TitleData Infrastructure Engineer              -41094      38663  -1.063 0.287921
Job.TitleData Integration Specialist              -117640      50016  -2.352 0.018730 *
Job.TitleData Lead                                 -85674      40717  -2.104 0.035444 *
```



```
Job.TitleData Management Specialist               -151827      64323  -2.360 0.018315 *
Job.TitleData Manager                             -110721      30422  -3.640 0.000277 ***
Job.TitleData Modeler                             -107701      38647  -2.787 0.005355 **
Job.TitleData Modeller                            -112092      49908  -2.246 0.024773 *
Job.TitleData Operations Analyst                  -142966      37180  -3.845 0.000123 ***
Job.TitleData Operations Engineer                  -85621      35285  -2.427 0.015297 *
Job.TitleData Operations Manager                  -104174      49844  -2.090 0.036697 *
Job.TitleData Operations Specialist               -162304      40737  -3.984 6.92e-05 ***
Job.TitleData Quality Analyst                     -146519      35285  -4.153 3.37e-05 ***
Job.TitleData Quality Engineer                    -131783      64418  -2.046 0.040862 *
Job.TitleData Science Consultant                  -111067      31035  -3.579 0.000350 ***
Job.TitleData Science Engineer                     -88768      35258  -2.518 0.011861 *
Job.TitleData Science Lead                         -43430      34109  -1.273 0.203016
Job.TitleData Science Manager                      -45216      29745  -1.520 0.128573
Job.TitleData Science Tech Lead                    154880      64323   2.408 0.016103 *
Job.TitleData Scientist                            -76919      28903  -2.661 0.007823 **
Job.TitleData Scientist Lead                       -61452      49849  -1.233 0.217756
Job.TitleData Specialist                          -102356      31696  -3.229 0.001254 **
Job.TitleData Strategist                          -152174      40717  -3.737 0.000189 ***
Job.TitleData Visualization Analyst               -120174      49844  -2.411 0.015965 *
Job.TitleData Visualization Specialist            -122674      49844  -2.461 0.013902 *
Job.TitleDecision Scientist                        -81132      36106  -2.247 0.024705 *
Job.TitleDeep Learning Engineer                    -42655      35329  -1.207 0.227386
Job.TitleDeep Learning Researcher                  -95957      64323  -1.492 0.135847
Job.TitleDirector of Data Science                  -46336      31809  -1.457 0.145304
Job.TitleETL Developer                            -100620      36108  -2.787 0.005357 **
Job.TitleETL Engineer                             -122858      49908  -2.462 0.013880 *
Job.TitleFinance Data Analyst                      -54712      43985  -1.244 0.213636
Job.TitleFinancial Data Analyst                    -73590      43990  -1.673 0.094450 .
Job.TitleHead of Data                              -67491      32476  -2.078 0.037774 *
Job.TitleHead of Data Science                      -67368      33756  -1.996 0.046049 *
Job.TitleHead of Machine Learning                 -173158      64447  -2.687 0.007251 **
Job.TitleInsight Analyst                          -147774      49908  -2.961 0.003090 **
Job.TitleLead Data Analyst                        -111526      38665  -2.884 0.003948 **
Job.TitleLead Data Engineer                        -64356      37280  -1.726 0.084392 .
Job.TitleLead Data Scientist                       -88707      34633  -2.561 0.010472 *
Job.TitleLead Machine Learning Engineer           -127715      40705  -3.138 0.001719 **
Job.TitleMachine Learning Developer               -104977      35348  -2.970 0.003002 **
```

```
Job.TitleMachine Learning Developer                -104977      35348  -2.970 0.003002 **
Job.TitleMachine Learning Engineer                  -53226      29011  -1.835 0.066646 .
Job.TitleMachine Learning Infrastructure Engineer   -81165      32946  -2.464 0.013808 *
Job.TitleMachine Learning Manager                   -77788      43943  -1.770 0.076785 .
Job.TitleMachine Learning Research Engineer        -112836      40754  -2.769 0.005660 **
Job.TitleMachine Learning Researcher                -83902      38665  -2.170 0.030081 *
Job.TitleMachine Learning Scientist                 -41070      30221  -1.359 0.174247
Job.TitleMachine Learning Software Engineer         -40623      32929  -1.234 0.217428
Job.TitleMachine Learning Specialist               -140144      49908  -2.808 0.005014 **
Job.TitleManager Data Management                    -95120      64323  -1.479 0.139292
Job.TitleManaging Director Data Science              50533      64447   0.784 0.433042
Job.TitleMarketing Data Analyst                     -62846      49885  -1.260 0.207827
Job.TitleMarketing Data Engineer                    -88566      64418  -1.375 0.169268
Job.TitleML Engineer                                -32599      29773  -1.095 0.273639
Job.TitleMLOps Engineer                             -67321      37218  -1.809 0.070572 .
Job.TitleNLP Engineer                               -99634      32957  -3.023 0.002521 **
Job.TitlePrincipal Data Analyst                     -72185      49909  -1.446 0.148182
Job.TitlePrincipal Data Architect                  -181966      64323  -2.829 0.004699 **
Job.TitlePrincipal Data Engineer                    -37647      49816  -0.756 0.449871
Job.TitlePrincipal Data Scientist                   -28197      34568  -0.816 0.414735
Job.TitlePrincipal Machine Learning Engineer       -105975      49842  -2.126 0.033562 *
Job.TitleProduct Data Analyst                      -104694      37220  -2.813 0.004940 **
Job.TitleResearch Analyst                          -129221      38646  -3.344 0.000836 ***
Job.TitleResearch Engineer                          -41560      29834  -1.393 0.163708
Job.TitleResearch Scientist                         -43034      29368  -1.465 0.142927
Job.TitleSales Data Analyst                        -135144      64373  -2.099 0.035862 *
Job.TitleSoftware Data Engineer                    -100167      43945  -2.279 0.022712 *
Job.TitleStaff Data Analyst                        -111860      49913  -2.241 0.025087 *
Job.TitleStaff Data Scientist                      -105674      49844  -2.120 0.034075 *
Job.TitleStaff Machine Learning Engineer            -55174      64324  -0.858 0.391091
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 57500 on 3184 degrees of freedom
Multiple R-squared:  0.3303,	Adjusted R-squared:  0.3061 
F-statistic: 13.66 on 115 and 3184 DF,  p-value: < 2.2e-16
```

Figure 3: Results of a linear regression analysis of dollar salary level versus experience level, company size, and job title.

Figure 3 shows that positive salary trends were observed in terms of “level of experience,” which tends to increase as professionals lengthen and progress in their careers. And, over the study period, Figure 3 also shows study found that data science professionals hired by midsize companies experienced higher salary growth compared to their counterparts at smaller companies. This result is consistent with the conventional wisdom that experience brings additional expertise, making experienced data scientists more sought after by employers.

Interestingly, we found that Figure 3 is shown some occupations have a strong correlation with salary levels. For example, Business Intelligence Developer, Data Analyst, Data Manager, Data Science Consultant, Data Strategist and so on. When job seekers choose a career in the field of data science, they can consider the salary levels of different occupations to choose.

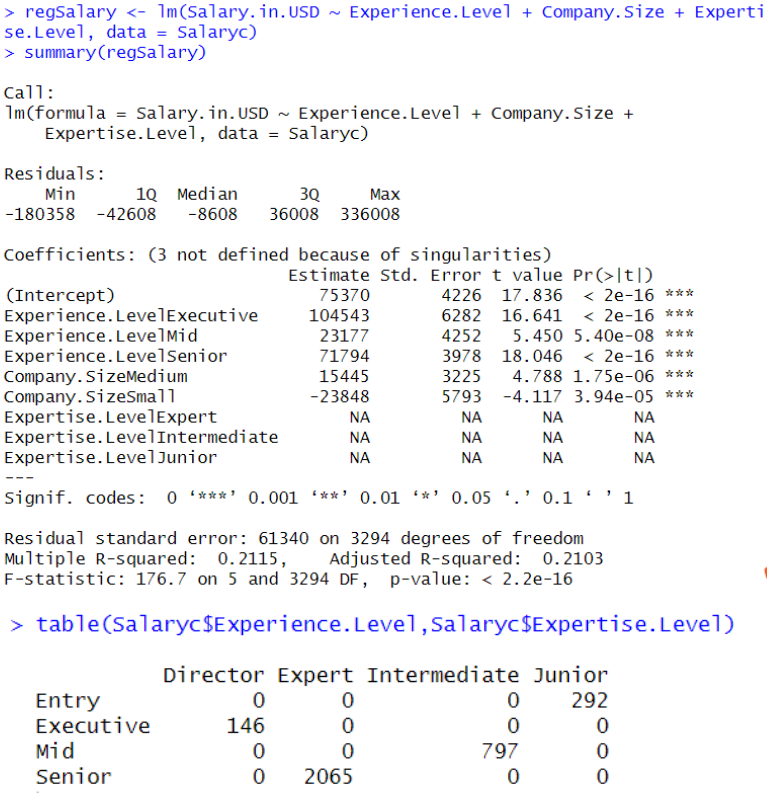
These findings provide data science professionals, as well as employers, with insights into the factors that influence salaries in the U.S. data science field and help organizations make informed decisions about compensation strategy and planning.

Overall, this analysis helps to enrich the existing knowledge of data science salaries and provides a basis for further investigation of salary trends and trends in the data science job market. As the field of data science continues to evolve, understanding the significant factors that affect salaries is important for organizations to retain talent and make competitive compensation decisions. Among them, occupation selection, company size and experience level can be an important factor for relevant personnel to look for jobs in the job market and can also become an important factor for the organization to make relevant decisions.

4.2. Model Diagnosis

However, in the diagnosis of the multiple linear regression model, we found that the model is not suitable for this data set, for the following four reasons:

Firstly, when analyzing the linearity of relationships, significance of variables and multicollinearity, we found the data of model include β coefficients, the Rsq, and the F-Stats, which show that the experience level and the expertise level in the data set have a completely linear relationship as shown in Figure 4, so these two dependent variables cannot be correlated with the salary at the same time. And in order to quantify the performance of a linear model, we consider the coefficient of determination (R2), which provides an assessment of the variability of the output any linear model is able to capture and depends on the variance of the data and the sum of the squared errors of the model [8]. Although our selected influencing factors are significant such as experience level, company size and occupation, the Adjusted R-square is 0.3061 obtained by the model which does not mean that it's an excellent fitting regression model.



```
> regSalary <- lm(Salary.in.USD ~ Experience.Level + Company.Size + Experti
se.Level, data = Salaryc)
> summary(regSalary)

Call:
lm(formula = Salary.in.USD ~ Experience.Level + Company.Size + 
    Expertise.Level, data = Salaryc)

Residuals:
    Min      1Q  Median      3Q     Max 
-180358  -42608   -8608   36008  336008 

Coefficients: (3 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    75370       4226  17.836  < 2e-16 ***
Experience.LevelExecutive     104543       6282  16.641  < 2e-16 ***
Experience.LevelMid            23177       4252   5.450 5.40e-08 ***
Experience.LevelSenior         71794       3978  18.046  < 2e-16 ***
Company.SizeMedium             15445       3225   4.788 1.75e-06 ***
Company.SizeSmall             -23848       5793  -4.117 3.94e-05 ***
Expertise.LevelExpert             NA         NA      NA       NA    
Expertise.LevelIntermediate       NA         NA      NA       NA    
Expertise.LevelJunior             NA         NA      NA       NA    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61340 on 3294 degrees of freedom
Multiple R-squared:  0.2115,	Adjusted R-squared:  0.2103 
F-statistic: 176.7 on 5 and 3294 DF,  p-value: < 2.2e-16

> table(Salaryc$Experience.Level,Salaryc$Expertise.Level)
           
            Director Expert Intermediate Junior
  Entry            0      0            0    292
  Executive      146      0            0      0
  Mid              0      0          797      0
  Senior           0   2065            0      0
```

Figure 4: Summary of Linear Regression Results

Secondly, when we analyzed whether the variances of errors are constant for predicted values which also referred to as homoscedasticity, we conducted the quick test to explore the correlation between the absolute value of the residuals and the dependent variable (in this case, the independent variable). The results indicated that the absolute value of the residuals is strongly correlated with the independent variable, because the p-value obtained by the model is less than 0.05, so we can reject the null hypothesis. In other words, the residuals will change with the dependent variable, directly indicating that the residuals will vary with the independent variable X, but the null hypothesis is that the residuals will not vary with the independent variable X. The result violates the model rules, which shows the model is not a fitting linear regression model.

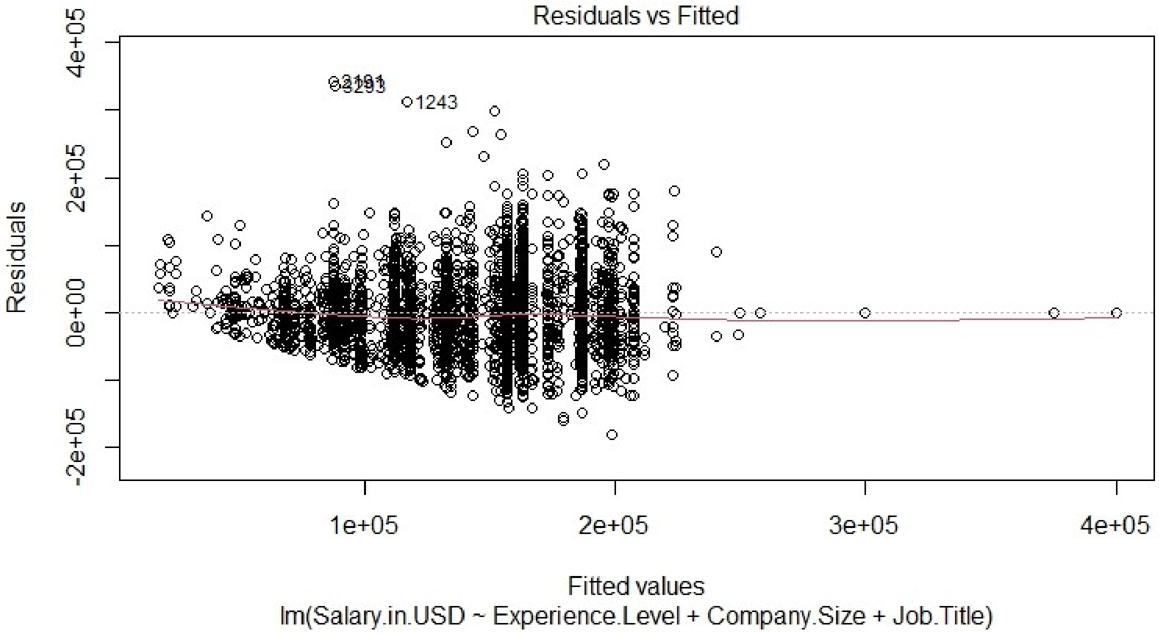


Figure 5: Linear Regression Model Residual Plot

Thirdly, when checking for the normality of residues, we use the following methods: See histogram as shown in Figure 6 and summary stats as shown in Figure 7 to visually check for skew, check normal quantile plot of the residuals and use shapiro.test(residuals(regressedModel)) for normality. The results indicate the data have many outliers as shown in Figure 5, and for normality, the plot should lie close to the normal line. However, regSalary is not normal because of the long-tails. Though the p-value obtained by the Shapiro-Wilk test, we can reject the null hypothesis which means the residuals do not conform to the normal distribution. Therefore, the model is not suitable for analysis of this data set.

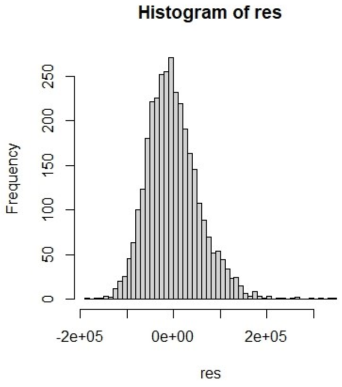


Figure 6: Linear Regression Model Histogram

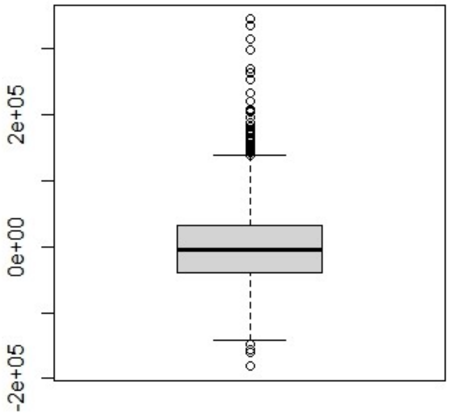
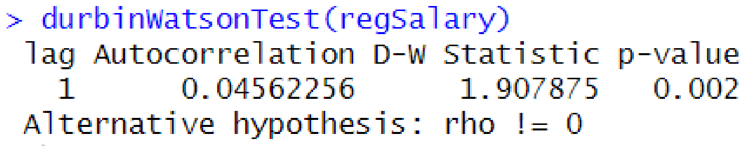


Figure 7: Linear Regression Model Histogram and summary stats

Last but not least, when we check the independence of observations which means the residuals are not autocorrelated, we obtained the p-value and D-W value from the Durbin-Watson test as shown in Figure 8 which indicated residuals are positively autocorrelated. That is to say, the model is not suitable for the data set.



```
> durbinWatsonTest(regSalary)
 lag Autocorrelation D-W Statistic p-value
   1      0.04562256      1.907875   0.002
 Alternative hypothesis: rho != 0
```

Figure 8: Durbin-Watson test results

4.3. Predictions

Results obtained revealed that the ARIMA model has a strong potential for short-term prediction [9]. Based on the analysis of time series models and Predictions (ARIMA), several very key results were obtained. From 2020 to 2021, the average salary of "Scientist" related professions experienced a decline, but from 2021 onwards, it rebounded and has been rising beyond the highest level of the average salary in 2020, showing a very good upward trend. Despite this, the average salary directly related to "Scientist" as shown in Figure 9 is expected to decline from its peak in 2023 to $162,415.60 in the 2024 forecast.

Data science occupations related to "Architect" began to decline in 2021, with the average salary reaching its lowest point in 2022. However, from the lowest point in 2022, it rose to 2023, increasing by a third of the total decline level. For 2024, the average salary for "Architect" related data science careers will increase slightly to $179,248.70 in figure 10.

Salaries for "Engineer" and "Analyst" related occupations show a continuous upward trend from 2020 to 2023. The average of Engineers' salaries will see a small decrease in 2024 to $173,616.20 in figure 11, while the average of analysts' salaries will see a significant decrease to $106,192.80 in figure 12.

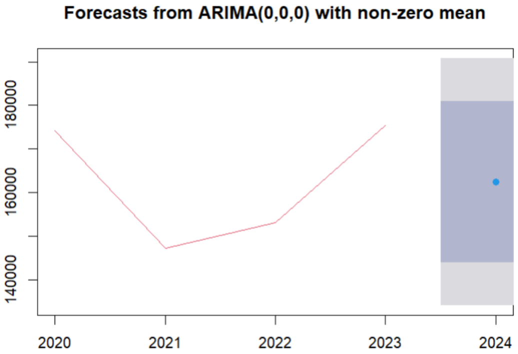


Figure 9: Scientist Salary Prediction

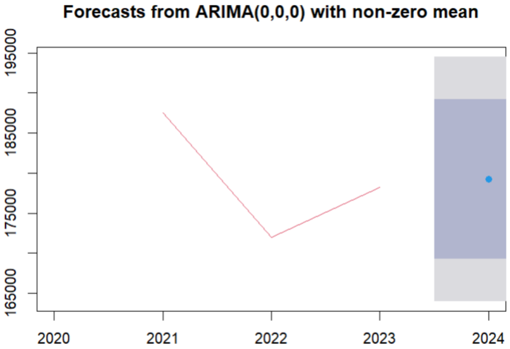


Figure 10: Architect Salary Prediction

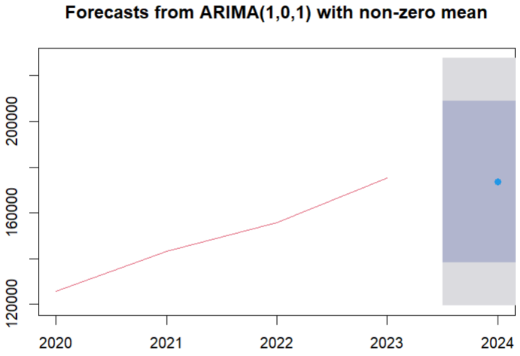


Figure 11: Engineer Salary Prediction

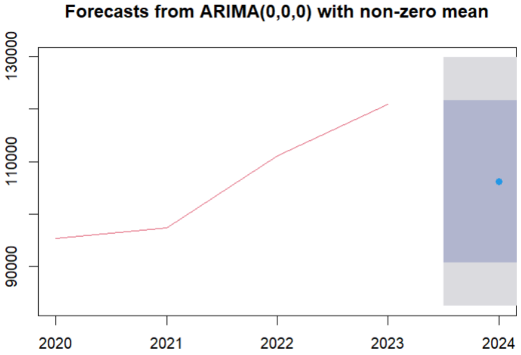


Figure 12: Analyst Salary Prediction

These results highlight the dynamic nature of wage trends across occupations. Although scientists' average salaries are expected to face a temporary setback in 2024, they remain relatively high. "Architect" related occupations will emerge from the recession. Although the average salary of data science occupations related to "Engineer" has decreased slightly, it is still the most stable occupation, which also helps this profession to maintain its advantage. However, the decline was most pronounced in data science occupations related to "Analyst," which also had a lower average, indicating potential challenges or changes in this job market and may cause those intending to choose this field to reconsider.

In summary, the salary trends forecast and analysis for "Scientist," "Architect," "Engineer," and "Analyst" provide analytical insights into the dynamics of the job market in these fields. Predicting trends indicates both opportunities and challenges, urging professionals and organizations to constantly adapt to change, and can likewise use the results of analysis to make their own changes and decisions.

5. Conclusion

Predicting salary trend in Data Science is a very sophisticated task considering that some decisive factors are not available for access and some latent yet crucial variables are difficult to be captured. Over-all, our model approves the significance of experience level, company size and different types of job title on the salary trend and possesses a relatively decent predictive ability. However, by being trained and tested based on data gathered from various types of experience level, company size and job title, our model is proven not to be an excellent fitting regression model. The main problem is that a few job titles in data science are not well-fitted, in other words, some job titles do not play a crucial role in influencing salary in data science. Therefore, our subsequent research will focus on a more appropriate regression model to figure out the relationship between the factors and the salary trends on data science. Besides the multiple linear regression model, we can explore other nonlinear models or machine learning models and see if a higher accuracy could be achieved. In addition, the research can add more objective measures and indexes to better value salary trends. For example, the working locations, the tools used, education and academic background, as well as specializations and skills.

References

[1] *Li, B. (2021, April). Quantitative Analysis of Salary Data in the Big Data Era. In Journal of Physics: Conference Series (Vol. 1881, No. 3, p. 032022). IOP Publishing. ISO 690*

[2] *Erpapalemlah, M. A. (2023). PREDICTING SALARY OUTCOME IN THE FIELD OF DATA SCIENCES WITH EXTREME GRADIENT BOOSTING ALGORITHM (Doctoral dissertation, UNIVERSITAS SRIWIJAYA).*

[3] *Quan, T., & Raheem, M. (2023). Human Resource Analytics on Data Science Employment Based on Specialized Skill Sets with Salary Prediction. International Journal of Data Science, 4(1), 40-59.*

[4] *Quan, T. Z., & Raheem, M. (2022). Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature. Journal of Applied Technology and Innovation, 6(3), 70-74.*

[5] *Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting Egyptian GDP using ARIMA models. Reports on Economics and Finance, 5(1), 35-47. Abonazel, M. R., & Abd-Elftah, A. I. (2019). Forecasting Egyptian GDP using ARIMA models. Reports on Economics and Finance, 5(1), 35-47.*

[6] *Saeed, A. K. M., Abdullah, P. Y., & Tahir, A. T. (2023). Salary Prediction for Computer Engineering Positions in India. Journal of Applied Science and Technology Trends, 4(01), 13-18.*

[7] *Das, S., Barik, R., & Mukherjee, A. (2020). Salary prediction using regression techniques. Proceedings of Industry Interactive Innovations in Science, Engineering & Technology (I3SET2K19).*

[8] *Martín, I., Mariello, A., Battiti, R., & Hernández, J. A. (2018). Salary prediction in the it job market with few high-dimensional samples: A Spanish case study. International Journal of Computational Intelligence Systems, 11(1), 1192-1209.*

[9] *Ariyo, A. A., Adewumi, A. O., & Ayo, C. K. (2014, March). Stock price prediction using the ARIMA model. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation (pp. 106-112). IEEE.*