MICHIGAN TECHNOLOGICAL UNIVERSITY

STATISTICAL METHODS

MA 5701

PROJECT REPORT

PROJECT TITLE: Analyzing Salary Determination in Tech Industry

Group- Pi Smarts

Name:

Karthik Kumar Cholleti

Hasibur Rashid Mahi

Shirisha Gajjela

1. Introduction

Nowadays, in the vast landscape of Job roles and their earnings, there exists a relationship between various factors that shape an individual's earnings. Our aspect is to know how these factors come together to affect a person's professional life and contribute to their income. We chose this data set because of its large impact on professional career goals. Understanding the factors that affect an individual's income rates would make some sense in choosing the right career path. This project aims to analyze the salaries of professionals working in the AI, ML, and Data Science fields across various countries from 2020 to 2024, focusing on the global salary levels of various roles in these fields. The dataset used for this study comprises 13,973 observations and 11 variables, including 8 categorical and 3 numerical variables.

1.1 Target Population:

The target population for this study is the "Salaries" earned by professionals depending on their performances in the world in their chosen career fields related to Data Science, who are working in several companies located across the globe.

1.2 Purpose of Study:

The primary purpose of the study is to analyze the relationship between Salary and each of the other independent variables that may influence the response variable. Specifically, this study focuses on the explanatory variables, which are strongly correlated with the response variable.

1.3 Motivation:

In this project, we are trying to look into the relationship between the salaries in the tech world employees like Machine Learning, Artificial, Data Science engineer, etc. These fields are rapidly growing filed with increasing Salary. We want to see the relations with Salary and other things like experience, residence, company size etc. By doing that, we will be able to understand how the pay difference will increase in the upcoming years.

2. Methods

A data set of 13,973 observations is collected from Kaggle which is an observational study since the data is sourced directly from a website without making any changes to the variables which are already cleaned and have no null values. Moreover, the researchers are not directly involved in this experiment which clearly states that this experiment is an observational study.

1.3 Sampling Unit:

The sampling units in this experiment are the specific individual persons within the population who are earning their salaries depending on various factors such as their Experience level, Employment Type, and Job Title. There are 146 professionals working in different companies from different places.

These sampling units are collected non-randomly, i.e., collected directly from the dataset available sourced from the Kaggle. These non-random samples are collected through a specific selection process, such as their career positions from different countries, rather than a random sampling method.

1.4 Variables:

The response variable in this study is the Salary of an individual in USD, which is measured on a continuous scale. This variable represents the annual compensation received by the professionals from different countries in AI, ML and Data Science space.

The Exploratory variables include Work Year, Experience Level, Employment Type, Job Title, Employee Residence, Remote Ratio, Company Location and Company Size.

There are a total of 13,972 observations in this dataset. The response variable is SalaryUsd within a range from $14000 to $ 30,400,0,00 with a mean of $1,66,001 and a standard deviation of $366154.5.

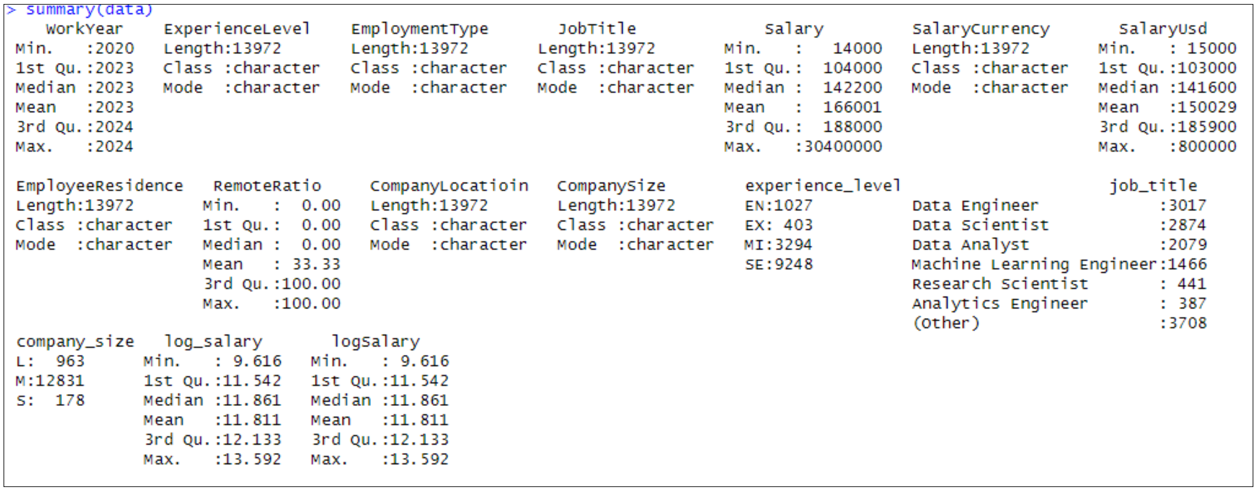
On the other hand, the predictor variables, like Work Year measured in Years, are taken within a range between 2020 and 2024. There are four categories in the Experience level: EN (Entry level), EX (Experienced level), SE (Senior Executive), and MI (Mid-level), with 403 observations. Coming to the Employment Type, there are four categories: 'FT' (Full Time), 'PT' (Part-Time), 'CT' (Contract), and 'FL' (Freelancer). Remote Ratio is measured as 0 (Work from Home) and 100 (Work from Office), and other dependent variables such as Employment Type and Employment Residence are also categorical in nature.

3. Exploratory Data Analysis

Analyzing the relationship between Salary and each of the other independent variables (Experience level, Remote Ratio, Employment Type, Company Size, and Location) that may have an effect on the response variable. Exploratory Data Analysis is performed to visualize the effect of the predictor variables on the response variable Salary (USD).

Before delving into the statistical analysis, let’s observe the summary of the dataset to understand the analyses better.

To get the values better, we performed a logarithmic function on our response variable.



```
> summary(data)
    WorkYear     ExperienceLevel    EmploymentType       JobTitle             Salary          SalaryCurrency        SalaryUsd
 Min.   :2020   Length:13972       Length:13972       Length:13972       Min.   :   14000   Length:13972       Min.   : 15000
 1st Qu.:2023   Class :character   Class :character   Class :character   1st Qu.:  104000   Class :character   1st Qu.:103000
 Median :2023   Mode  :character   Mode  :character   Mode  :character   Median :  142200   Mode  :character   Median :141600
 Mean   :2023                                                            Mean   :  166001                      Mean   :150029
 3rd Qu.:2024                                                            3rd Qu.:  188000                      3rd Qu.:185900
 Max.   :2024                                                            Max.   :30400000                      Max.   :800000

 EmployeeResidence   RemoteRatio      CompanyLocatioin   CompanySize        experience_level                     job_title
 Length:13972       Min.   :  0.00   Length:13972       Length:13972       EN:1027          Data Engineer            :3017
 Class :character   1st Qu.:  0.00   Class :character   Class :character   EX: 403          Data Scientist           :2874
 Mode  :character   Median :  0.00   Mode  :character   Mode  :character   MI:3294          Data Analyst             :2079
                    Mean   : 33.33                                         SE:9248          Machine Learning Engineer:1466
                    3rd Qu.:100.00                                                          Research Scientist       : 441
                    Max.   :100.00                                                          Analytics Engineer       : 387
                                                                                            (Other)                  :3708
 company_size   log_salary        logSalary
 L:  963      Min.   : 9.616   Min.   : 9.616
 M:12831      1st Qu.:11.542   1st Qu.:11.542
 S:  178      Median :11.861   Median :11.861
              Mean   :11.811   Mean   :11.811
              3rd Qu.:12.133   3rd Qu.:12.133
              Max.   :13.592   Max.   :13.592
```

Table1: Summary of the Data

Here, we have 13,972 observations with individuals having 146 unique Job Titles.

3.1 Descriptive Statistics:

Let's perform the exploratory data analysis for each predictor variable with respect to the response variable.

3.1.1 Visualization of the Salary distributions:

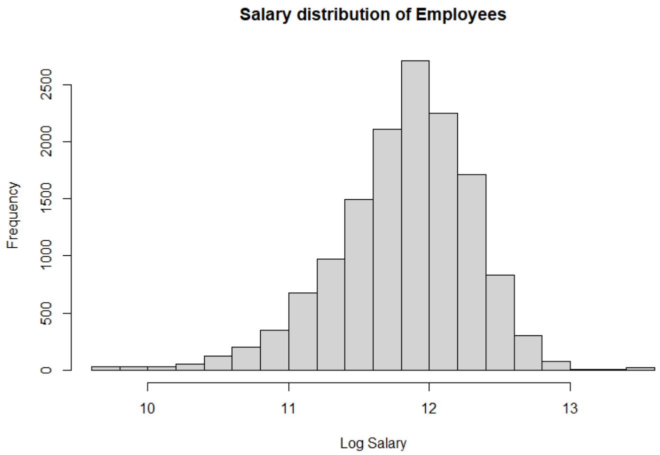


Fig1: Histogram of salary distribution of employees (USD)

The histogram of the salary variable suggests that most salaries of the individuals are clustered towards the lower end of the salary range, with a long right tail indicating some outliers.

This distribution implies that there may be a disparity in salaries within the dataset, with some individuals earning significantly higher salaries compared to the majority.

Additionally, it appears that the distribution is right-skewed, indicating that the mean salary may be higher than the median salary.

3.1.2 Salaries Vs. Experience levels:

The below box plot visualizes the distribution of the salaries across different experience levels.

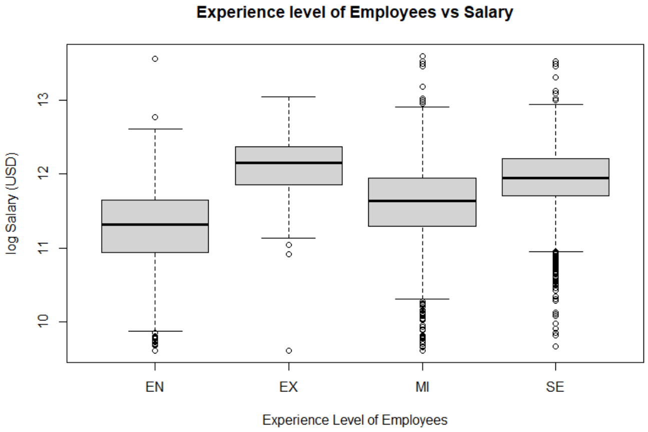


Fig2: Experience level vs Salary (USD)

From the above boxplot, it seems that there are considerable variations in salaries across different experience levels. It appears that there are some outliers present in the data, particularly for the Mid-level category and Senior executive. However, the individuals belonging to these levels seem to have higher salaries than compared to the other category levels.

Boxplot is appropriate in representing. Here Experience levels indicates-

EN- Entry Level

EX-Executive Level

MI-Middle Level

SE-Senior Executive Level

3.1.3 Employment Type vs. Salary:

The below box plot visualizes the distribution of the salaries across different Employment Types.

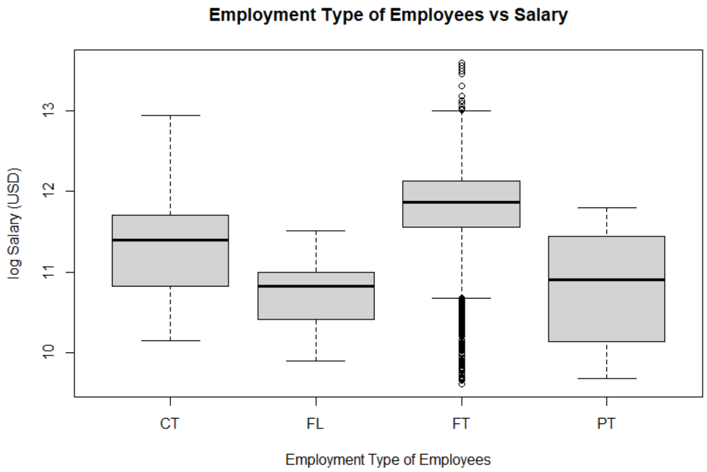


Fig3: Employment Type vs Salary (USD)

The above boxplot illustrates the distribution of salaries across different employment types Contract (CT), Freelances (FT), Full-Time (FT), and Part-Time (PT). It seems that there are considerable variations in salaries across different employment types.

It appears to be that the Full-Time employees have higher salaries than the other types followed by Contract levels (CT), Part-Time (PT) and Freelances (FL).

3.1.4 Employee Residence vs Salary

The below plot represents the distribution of salaries by specific employment residence.

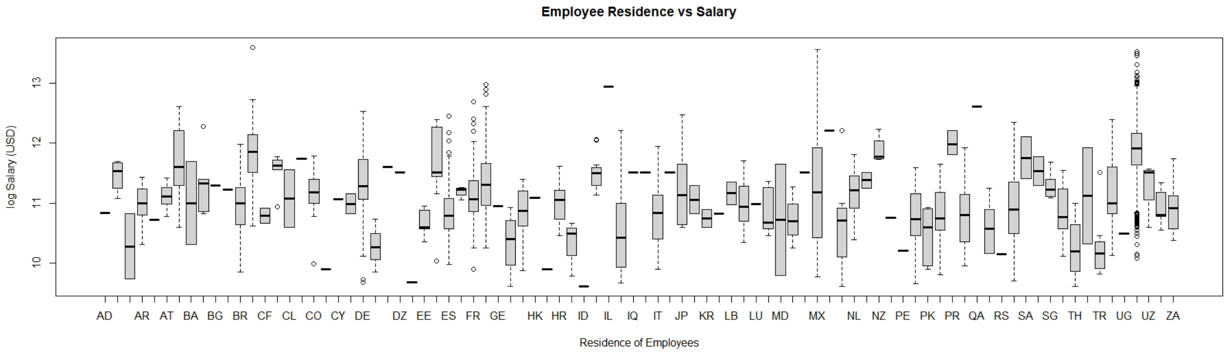


Fig4: Employment Residence vs Salary (USD)

From the plot we can observe that there is significant variation in salaries across the different countries. We can conclude that the individuals belonging to ‘US’ have more earnings than the individuals belonging to the other countries.

3.1.5 Job Title vs Salary

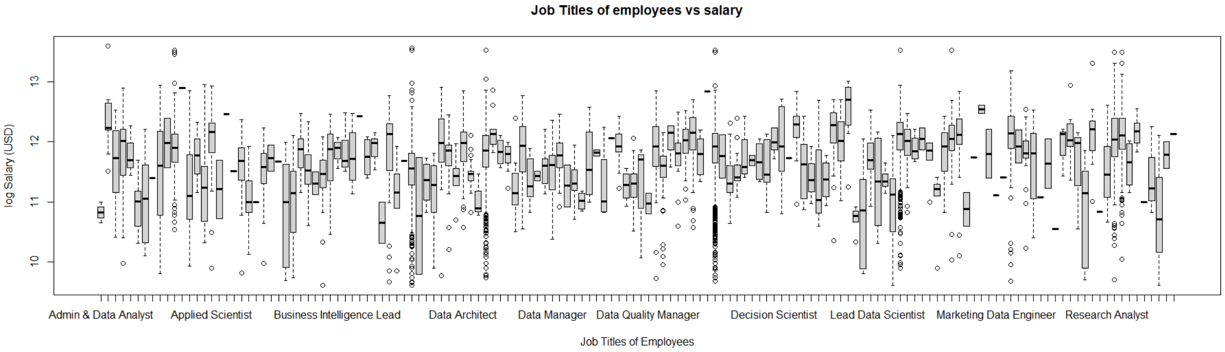
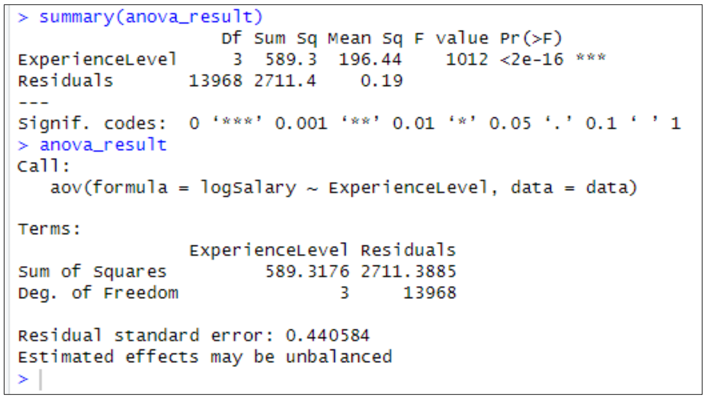


Fig5: Job Title vs Salary (USD)

In this plot we can observe many noticeable things. There are different job titles with a range of salaries where some positions have a tight salary range and some has a broader range. Like Research Analyst is one type of job where it has a significant range of salaries.



```
> summary(anova_result)
                   Df Sum Sq Mean Sq F value Pr(>F)
ExperienceLevel     3  589.3  196.44    1012 <2e-16 ***
Residuals       13968 2711.4    0.19
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova_result
Call:
   aov(formula = logSalary ~ ExperienceLevel, data = data)

Terms:
                ExperienceLevel Residuals
Sum of Squares         589.3176 2711.3885
Deg. of Freedom               3     13968

Residual standard error: 0.440584
Estimated effects may be unbalanced
> 
```

Table2: Anova Test between Salary and Experience Level

The ANOVA results between Experience level and Salary indicate that there is a significant effect of experience level on salary. As the p-value 2e-16 is less than the significance level 0.05, it suggests that there are statistically significant effects of Experience level on Salary.

4. Result

Simple Linear Regression models are fitted to analyze the relationship between Salary and each of the other independent variables (Experience level, Remote Ratio, Employment Type, Company size and its Location) that may influence the response variable. Further, a Multi Linear Regression is also fitted between the response variable and all the predictors to see how the effect of the predictors may influence the response variable.

4.1. Simple Linear Regression

As per the graphs below, we can see that the data points are normally distributed because most of the points follow the diagonal line. As we done log distribution of our response variable (Salary USD), we couldn't be able to observe the 'ab' line, but they are assessing the assumptions of homoscedasticity.

4.1.1 Salary vs Experience level

Coefficients	Point Estimate	Standard error	t-value	p-value
Intercept	11.27219	0.01375	819.91	<2e-16
ExperienceLevelEX	0.82438	0.02590	31.83	<2e-16
ExperienceLevelMI	0.34095	0.01575	21.65	<2e-16
ExperienceLevelSE	0.65707	0.01449	45.34	<2e-16

Table3: Summary of Parameter Estimates of Salary vs Experience Level

Multiple R-squared: 0.1785, Adjusted R-squared: 0.1784, F-Statistic: 1012,

P-value: 2.2e-16.

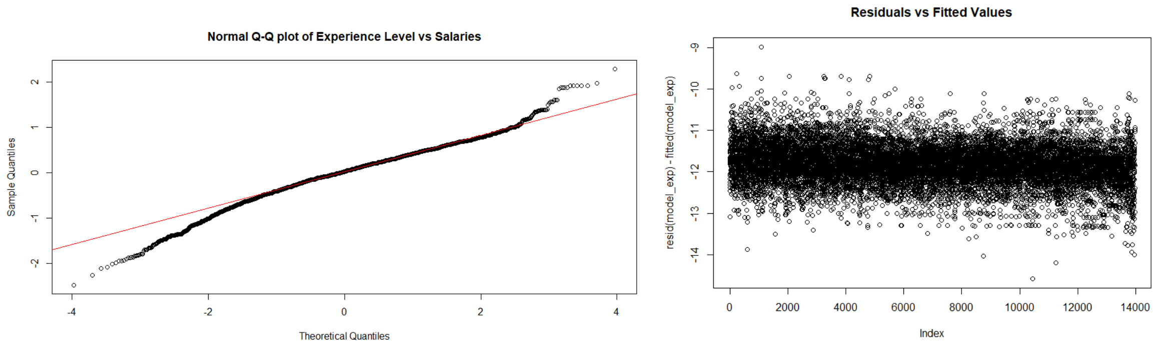


Fig6: Normal Q-Q plot and Residuals vs Fitted values of Salary vs Experience Level

From Table 1, the coefficients represent the average change in the Salary for each Experience Level. It seems that the estimated increase in salary for the individuals with Experience level ‘EX’ is $82,438. Similarly, the estimated increase in the Salary with Experience level ‘MI’ and experience level ‘SE’ is $34,095 and $65,707 respectively.

Thus, the coefficients indicate that as experience level increases from Entry level to mid-level to Senior Level, salaries tend to increase.

Also, as the p-value is less than the significance level, it indicates that there is a linear relationship between the Salary and Experience level. Thus, this analysis provides evidence that this model is significant in determining the Salary levels.

4.1.2 Salary vs. Employment Type

Coefficients	Point Estimate	Standard error	t-value	p-value
Intercept	11.35399	0.09473	119.860	<2e-16
Employment Type FL	-0.61321	0.16857	-3.638	0.000276
Employment Type FT	0.46065	0.09482	4.858	1.20e-06
	-0.55520	0.13992	-3.968	7.29e-05

Table4: Summary of Parameter Estimates of Salary vs Employment Type

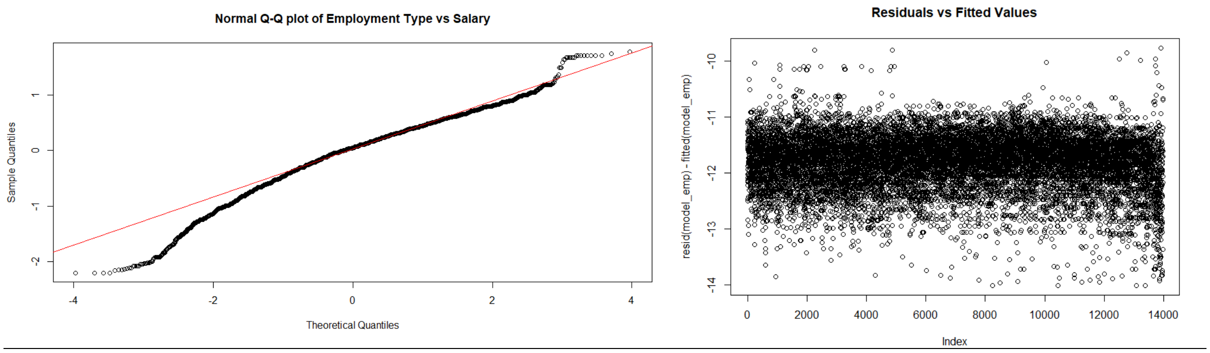


Fig7: Normal Q-Q plot and Residuals vs Fitted values of Salary vs Employment Type

Multiple R-squared:0.01269, Adjusted R-squared: 0.01248, F-Statistic: 59.86,

P-value: 2.2e-16.

From Table 2, we can see that the salaries of Freelancers have the lowest estimated salary, followed by Part-time employees while Full-time employees have the highest estimated salary.

The adjusted R-squared value of 0.012 states that the 1.2% of the variance in salary is explained by the Employment type.

Overall, the low p-values in all three Employment types, it suggests that there is a statistically linear relationship between the Salaries and the Employment type.

4.1.3 Salary vs Job Title

Multiple R-squared: 0.2044, Adjusted R-squared: 0.2044, F-Statistic: 24.5,

P-value: <2.2e-16

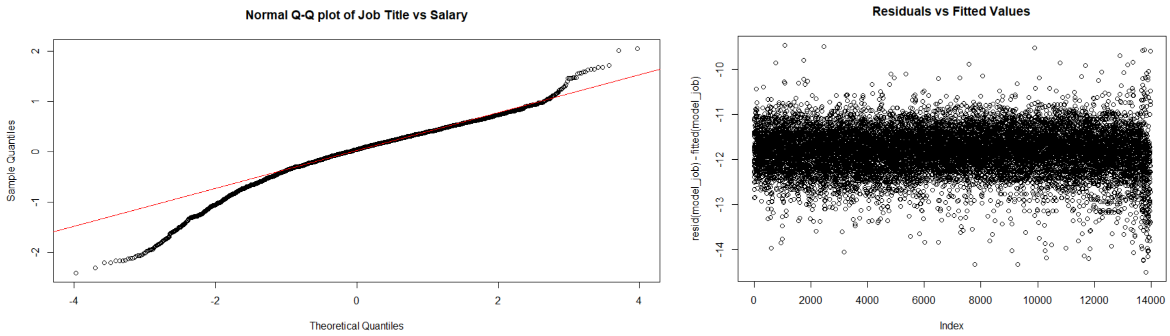


Fig8: Normal Q-Q plot and Residuals vs Fitted values of Salary vs Job Title

The analysis of this model based on the Job Title is an important factor influencing Salary variations. Here, some Job Titles show statistically significant effects on salary while others may not. Thus, different Job titles are having varying impact on the Salary distribution.

Overall, 20% of the variance in Salary is explained by the Job Title. However, it’s essential to recognize that this model may not capture all factors affecting Salary.

4.1.4 Salary vs Employee Residence

Multiple R-squared: 0.2592, Adjusted R-squared: 0.2545, F-Statistic: 55.83,

P-value: <2.2e-16

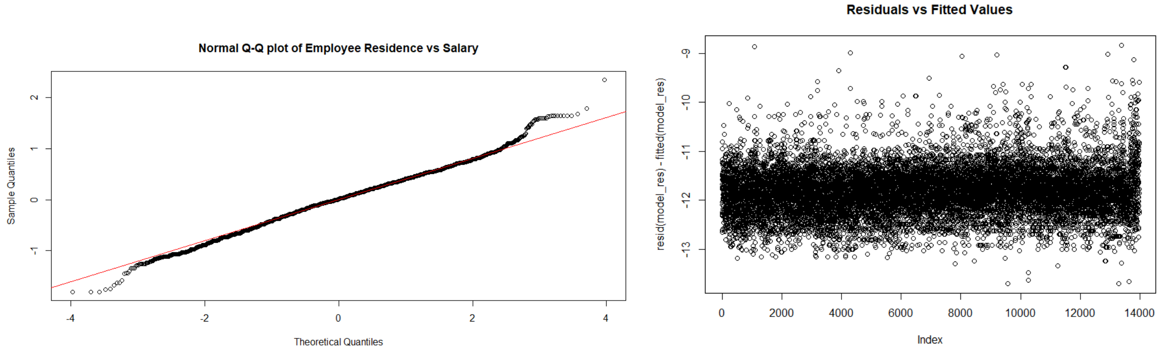


Fig9: Normal Q-Q plot and Residuals vs Fitted values of Salary vs Employee Residence

Here, Employee residence has a statistically significant relationship with the Salary. In this model, some countries like Israel (Employee residence IL) tend to have significantly higher salaries compared to the other countries.

However, some countries have coefficients with p-values greater than 0.05, which indicates that the residence in those countries does not affect Salary.

Overall, as per the adjusted R-squared, 25.4% of the variance in Salary is explained by the Employee Residence.

4.1.5 Salary vs Remote Ratio

Coefficients	Point Estimate	Standard error	t-value	p-value
Intercept	11.8340407	0.0050421	2347.023	<2e-16
Remote Ratio	-0.0006833	0.0000879	-7.774	8.14e-15

Table5: Summary of Parameter Estimates of Salary vs Employment Type

Multiple R-squared: 0.004307, Adjusted R-squared: 0.004236, F-Statistic: 60.43,

P-value: 8.137e-15

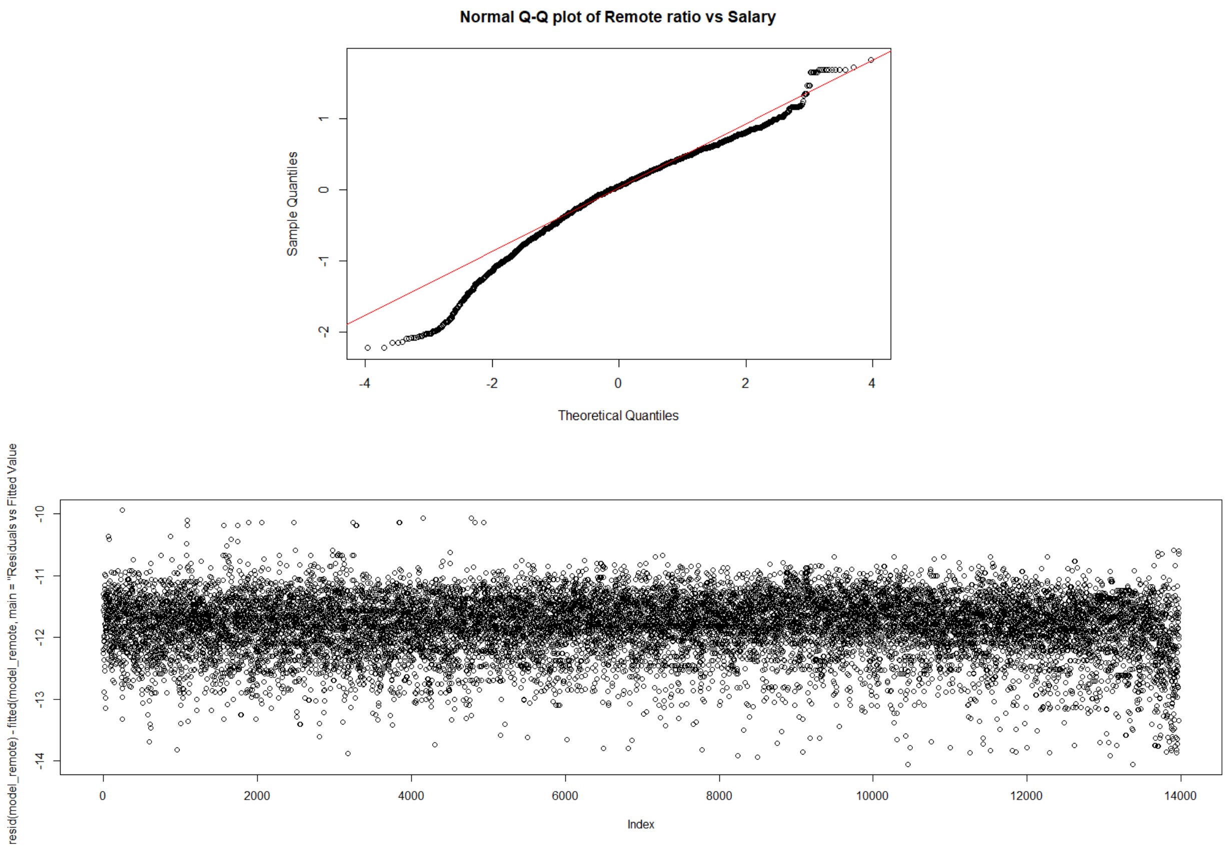


Fig10: Normal Q-Q plot and Residuals vs Fitted values of Salary vs Employee Residence

The coefficient for remote work ratio is 0.06% which indicates for every 1% increase in the remote work ratio, there is an estimated decrease of $0.06 in the Salary of the employees. However, as the p value is less than the significance level, remote work ratio has statistically significant impact on the Salary.

4.2 Multi Linear Regression Model

In this analysis, we will conduct a Multi Linear regression model to identify the key factors influencing Salary in the dataset. Here, the aim was to understand how various variables such as work experience, experience level, job title, company location, and company size together affect salary.

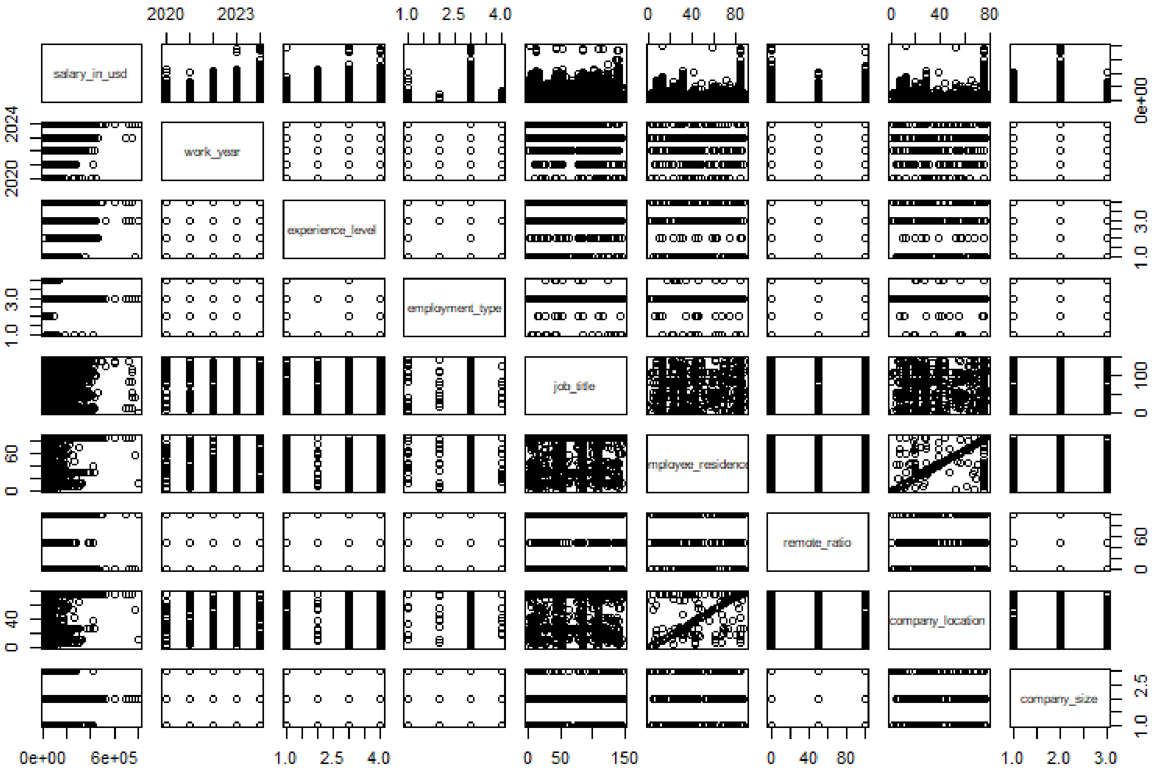
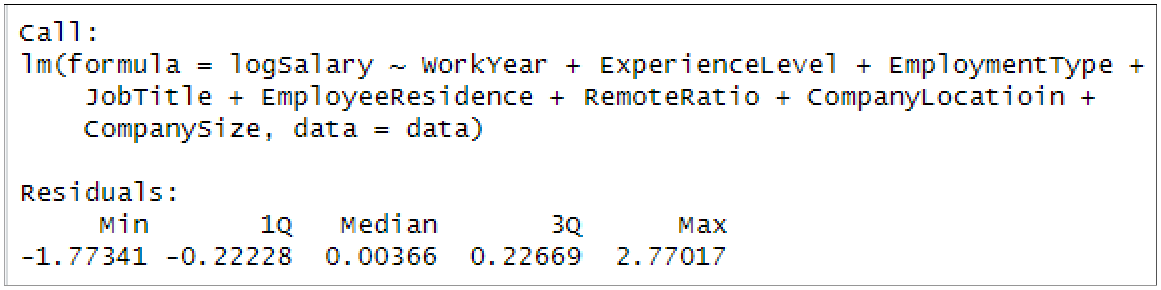
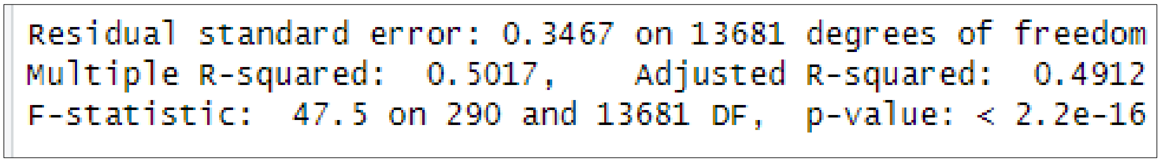


Fig11: Pairwise Scatter plot of the response variable and explanatory variables.



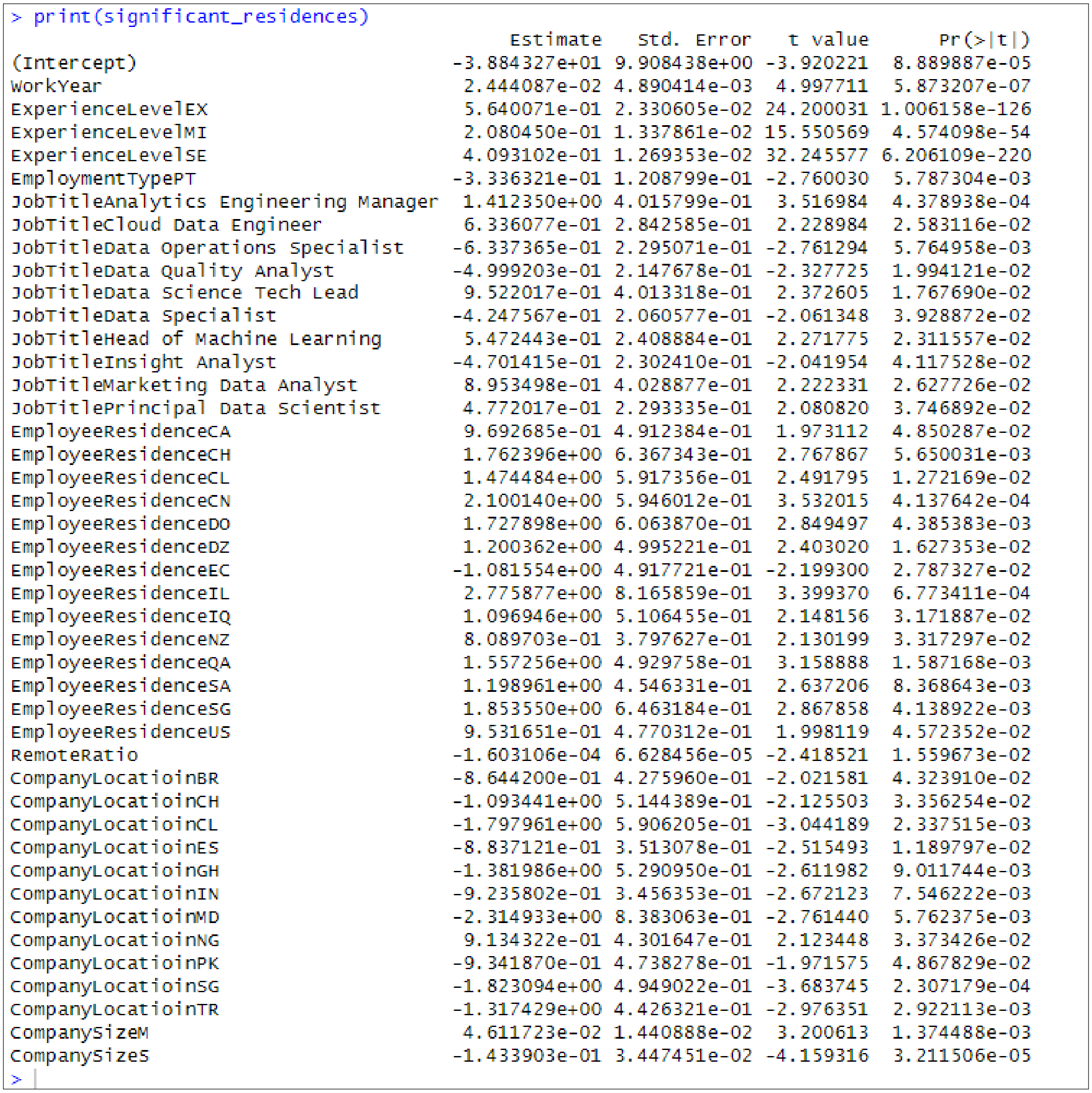
```
Call:
lm(formula = logSalary ~ WorkYear + ExperienceLevel + EmploymentType +
    JobTitle + EmployeeResidence + RemoteRatio + CompanyLocatioin +
    CompanySize, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.77341 -0.22228  0.00366  0.22669  2.77017
```



```
Residual standard error: 0.3467 on 13681 degrees of freedom
Multiple R-squared:  0.5017,    Adjusted R-squared:  0.4912
F-statistic:  47.5 on 290 and 13681 DF,  p-value: < 2.2e-16
```

Based on adjusted R squared, we can see that we only explain 49.12% of the variance in Salary using these explanatory variables.

However, from the above table, we can see only a few of the explanatory variables show some effect on the response variable, whose p-value is less than the significance level of 0.05. Those variables will be



```
> print(significant_residences)
                                           Estimate   Std. Error   t value      Pr(>|t|)
(Intercept)                           -3.884327e+01 9.908438e+00 -3.920221  8.889887e-05
WorkYear                               2.444087e-02 4.890414e-03  4.997711  5.873207e-07
ExperienceLevelEX                      5.640071e-01 2.330605e-02 24.200031 1.006158e-126
ExperienceLevelMI                      2.080450e-01 1.337861e-02 15.550569  4.574098e-54
ExperienceLevelSE                      4.093102e-01 1.269353e-02 32.245577 6.206109e-220
EmploymentTypePT                      -3.336321e-01 1.208799e-01 -2.760030  5.787304e-03
JobTitleAnalytics Engineering Manager  1.412350e+00 4.015799e-01  3.516984  4.378938e-04
JobTitleCloud Data Engineer            6.336077e-01 2.842585e-01  2.228984  2.583116e-02
JobTitleData Operations Specialist    -6.337365e-01 2.295071e-01 -2.761294  5.764958e-03
JobTitleData Quality Analyst          -4.999203e-01 2.147678e-01 -2.327725  1.994121e-02
JobTitleData Science Tech Lead         9.522017e-01 4.013318e-01  2.372605  1.767690e-02
JobTitleData Specialist               -4.247567e-01 2.060577e-01 -2.061348  3.928872e-02
JobTitleHead of Machine Learning       5.472443e-01 2.408884e-01  2.271775  2.311557e-02
JobTitleInsight Analyst               -4.701415e-01 2.302410e-01 -2.041954  4.117528e-02
JobTitleMarketing Data Analyst         8.953498e-01 4.028877e-01  2.222331  2.627726e-02
JobTitlePrincipal Data Scientist       4.772017e-01 2.293335e-01  2.080820  3.746892e-02
EmployeeResidenceCA                    9.692685e-01 4.912384e-01  1.973112  4.850287e-02
EmployeeResidenceCH                    1.762396e+00 6.367343e-01  2.767867  5.650031e-03
EmployeeResidenceCL                    1.474484e+00 5.917356e-01  2.491795  1.272169e-02
EmployeeResidenceCN                    2.100140e+00 5.946012e-01  3.532015  4.137642e-04
EmployeeResidenceDO                    1.727898e+00 6.063870e-01  2.849497  4.385383e-03
EmployeeResidenceDZ                    1.200362e+00 4.995221e-01  2.403020  1.627353e-02
EmployeeResidenceEC                   -1.081554e+00 4.917721e-01 -2.199300  2.787327e-02
EmployeeResidenceIL                    2.775877e+00 8.165859e-01  3.399370  6.773411e-04
EmployeeResidenceIQ                    1.096946e+00 5.106455e-01  2.148156  3.171887e-02
EmployeeResidenceNZ                    8.089703e-01 3.797627e-01  2.130199  3.317297e-02
EmployeeResidenceQA                    1.557256e+00 4.929758e-01  3.158888  1.587168e-03
EmployeeResidenceSA                    1.198961e+00 4.546331e-01  2.637206  8.368643e-03
EmployeeResidenceSG                    1.853550e+00 6.463184e-01  2.867858  4.138922e-03
EmployeeResidenceUS                    9.531651e-01 4.770312e-01  1.998119  4.572352e-02
RemoteRatio                           -1.603106e-04 6.628456e-05 -2.418521  1.559673e-02
CompanyLocatioinBR                    -8.644200e-01 4.275960e-01 -2.021581  4.323910e-02
CompanyLocatioinCH                    -1.093441e+00 5.144389e-01 -2.125503  3.356254e-02
CompanyLocatioinCL                    -1.797961e+00 5.906205e-01 -3.044189  2.337515e-03
CompanyLocatioinES                    -8.837121e-01 3.513078e-01 -2.515493  1.189797e-02
CompanyLocatioinGH                    -1.381986e+00 5.290950e-01 -2.611982  9.011744e-03
CompanyLocatioinIN                    -9.235802e-01 3.456353e-01 -2.672123  7.546222e-03
CompanyLocatioinMD                    -2.314933e+00 8.383063e-01 -2.761440  5.762375e-03
CompanyLocatioinNG                     9.134322e-01 4.301647e-01  2.123448  3.373426e-02
CompanyLocatioinPK                    -9.341870e-01 4.738278e-01 -1.971575  4.867829e-02
CompanyLocatioinSG                    -1.823094e+00 4.949022e-01 -3.683745  2.307179e-04
CompanyLocatioinTR                    -1.317429e+00 4.426321e-01 -2.976351  2.922113e-03
CompanySizeM                           4.611723e-02 1.440888e-02  3.200613  1.374488e-03
CompanySizeS                          -1.433903e-01 3.447451e-02 -4.159316  3.211506e-05
> 
```

Table6: Explanatory variables variance

The above table explains the significant explanatory variables that explain the variance in Salary.

4.2.1 Interpretation of the Multi-Linear Regression Model

From the above table, the coefficients explain how much the predictor variables are expected to change when the response variable increases by one unit.

For example, for each additional increase in the year of the work experience of an individual, the Salary is expected to be increased by $2444.

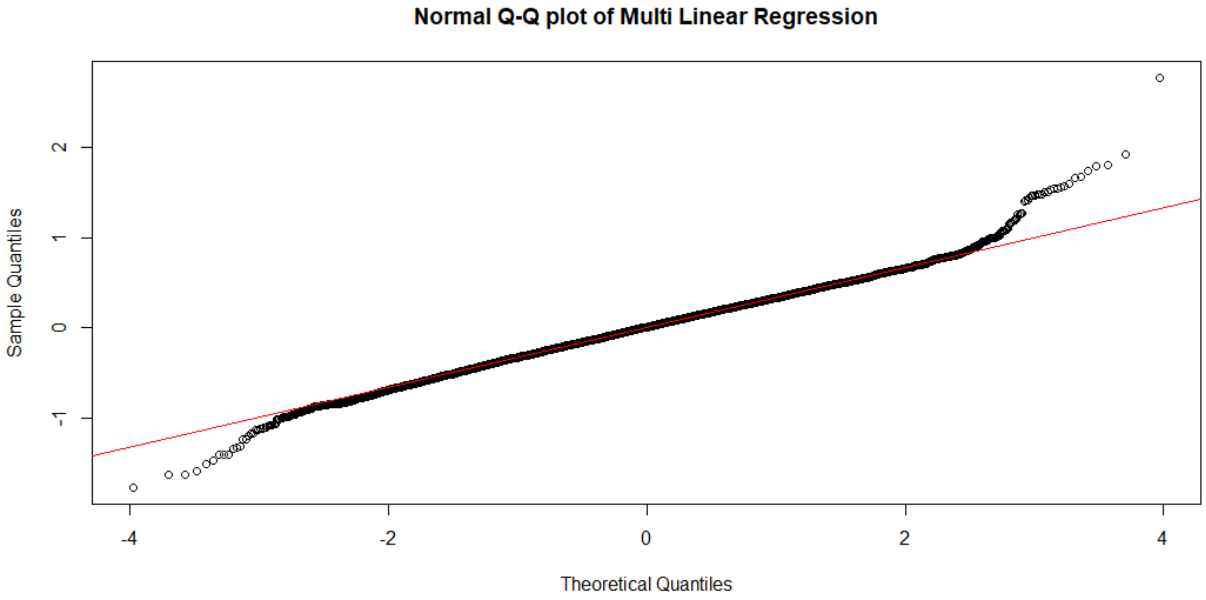


Fig12: Normal Q-Q plot of Multi Linear regression

A simple linear regression is said to be robust to violations of the normality when the sample size is large but, based on the above normal Q-Q plot for multi-linear regression, the normality assumption is not violated. The points on the plot approximately closely follow the diagonal line concluding that the residuals are normally distributed. However, there are some deviations in the tails.

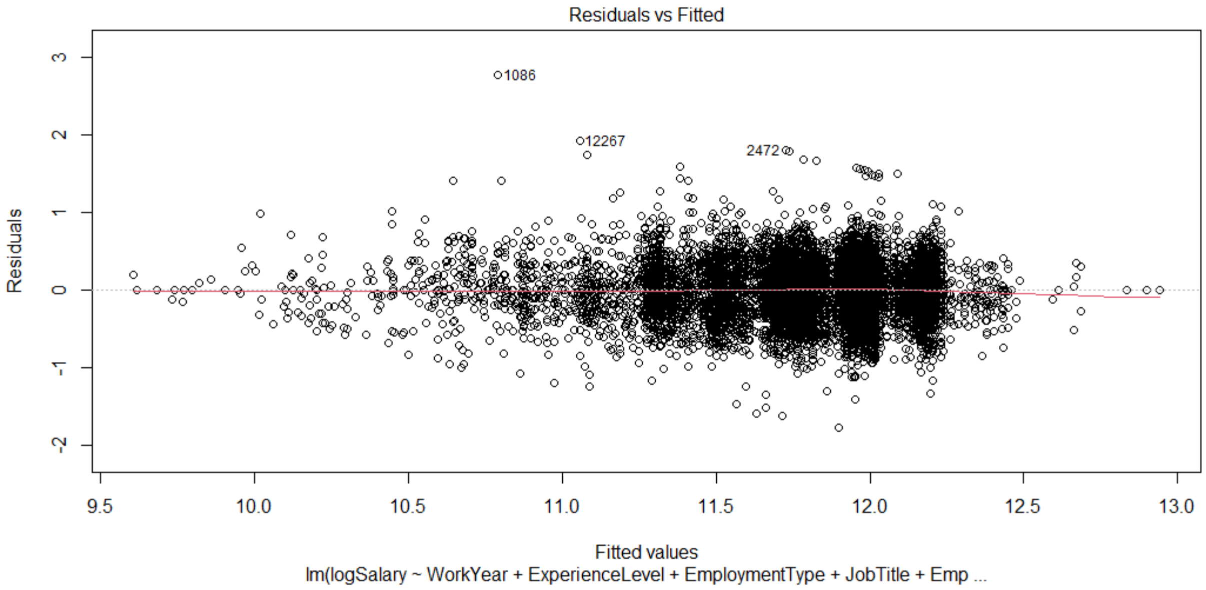


Fig13: Residuals vs Fitted values of Multi-Linear Regression

From the Residuals vs Fitted plot, we can conclude that there is a linear relationship between the residuals and fitted values as the residuals are scattered around the horizontal line 0. However, some of the points deviate far away from the horizontal line that indicates those points have no linearity in their regions and violating homogeneity of variances.

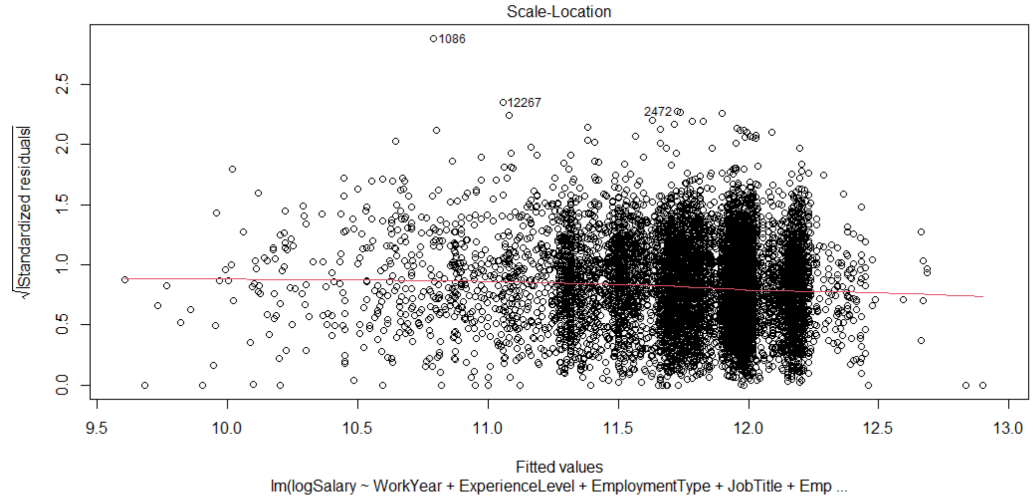


Fig14: Scale Location plot for multi-linear regression

From the above Scale location plot, the points are fairly distributed across the fitted values, suggesting Homoscedasticity. However, as the sample size is large, there is variation in the spread of the residuals with the increase in the Fitted values, which concludes that the variance of the residuals is not uniform and violates the assumptions of homoscedasticity.

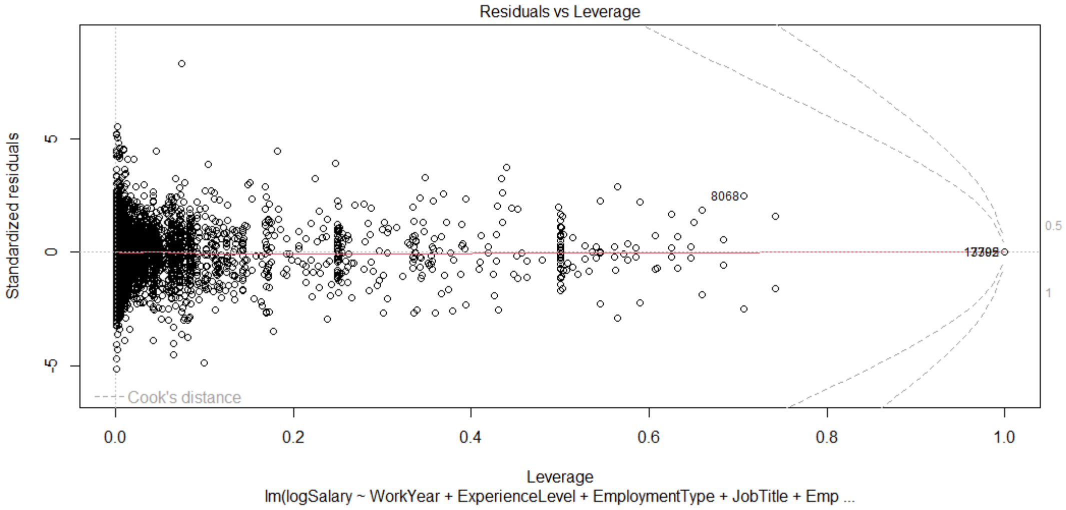
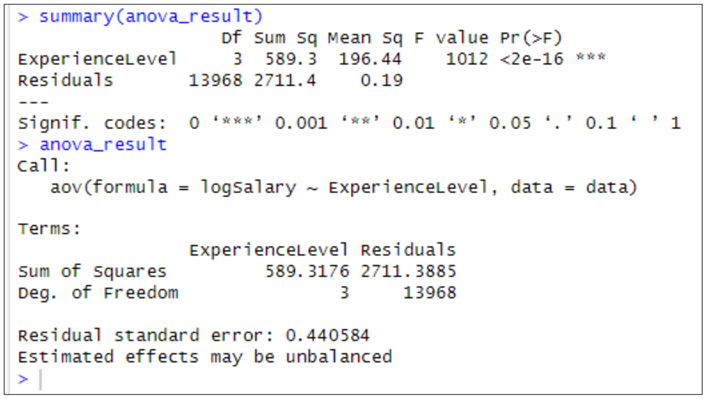


Fig15: Residuals Vs Leverage of Multi-linear regression

The above plot, Residuals vs Leverage, does not reveal any clear patterns.



```
> summary(anova_result)
                   Df Sum Sq Mean Sq F value Pr(>F)    
ExperienceLevel     3  589.3  196.44    1012 <2e-16 ***
Residuals       13968 2711.4    0.19                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova_result
Call:
   aov(formula = logSalary ~ ExperienceLevel, data = data)

Terms:
                ExperienceLevel Residuals
Sum of Squares         589.3176 2711.3885
Deg. of Freedom               3     13968

Residual standard error: 0.440584
Estimated effects may be unbalanced
> 
```

Table7: Anova Test between Salary and Experience Level

The ANOVA results between Experience level and Salary indicate that there is a significant effect of experience level on salary. As the p-value 2e-16 is less than the significance level 0.05, it suggests that there are statistically significant effects of Experience level on Salary.

4.3 Hypothesis and F-test for Multi Linear regression

4.3.1 Hypothesis Test for Remote Ratio:

	Point Estimate	Std. Error	t value	Pr(>|t|)
Intercept	11.8340407	5.042150e-03	2347.022636	0.000000e+00
Remote ratio	-0.0006833	8.790144e-05	-7.773956	8.136943e-15

Table8: Hypothesis Table for Remote ratio

From the above table, the p-value for the remote work ratio is approximately 8.136943e-15, which is significantly smaller than the chosen significance level of 0.05; we reject the null hypothesis and conclude that there is a statistically significant relationship between the remote ratio and Salary.

4.3.2 Hypothesis Test for Company Size

	Point Estimate	Std. Error	t value	Pr(>|t|)
Intercept	11.66805	0.01542931	756.22654	0.000000e+00
Company Size M	0.1628695	0.01599784	10.18072	2.938246e-14
Company Size S	-0.4995042	0.03906424	-12.78674	3.144404e-37

Table9: Hypothesis Table for Company Size

We performed hypothesis tests for company size where M- Medium sized, S- Small Sized, this test concludes that there is statistically significant relationship between the Salary and company size as both the P-values are less than the chosen significance level.

4.4 Conducting F-test for the predictor variables against the response variable.

The analysis of variance is conducted to assess the significance of the predictor variable against the response variable.

The results are summarized in the Anova table below.

Variable	**Df**	**Sum Sq**	**Mean Sq**	**F value**	**Pr(>F)**
work_year	1	49.44	49.444	411.2920	< 2.2e-16
experience_level	3	603.59	201.196	1673.6184	< 2.2e-16
employment_type	3	13.51	4.504	37.4657	4.655e-05
job_title	145	413.41	2.851	23.7161	< 2.2e-16
employee_residence	87	550.74	6.330	52.6575	< 2.2e-16
remote_ratio	1	0.62	0.622	5.1769	0.0004027
company_location	48	20.04	0.417	3.4595	0.8890365
company_size	2	4.68	2.339	19.4595	3.638e-09
Residuals	13681	1644.68	0.120		

Table10: Anova table of Multi-linear Regression

The F-test concludes that the predictor variables significantly affect the response variable (Salary).

Work Year, experience level, job title, employee residence, and remote work ratio have a significant impact on predicting the Salary of the individuals.

While Employment type and company size also show some significance, they are weaker compared to other predictors.

On the other hand, company location does not appear to significantly influence the response variable.

5. Conclusion

Based on the Statistical analysis provided, we can say that we have strong significant evidence to conclude that there is a relationship between the response variable and the other predictor variables. Based on the multi-linear regression, we found that even if there is a linear relationship between the variables, there are some violations in the graphs that state that the data points are heteroscedastic.

Based on the additional information, we may say that further conclusions can be interpreted. There may be other factors that strongly affect the relationship between the salaries of an individual working in the Data Science field. By incorporating this new factor, we can then conclude the relations between them so that it would be very useful to the present job seekers to choose the right path in their career. Overall, this study has the potential to make a significant impact on the way that we conclude the statistical relationship between the predictors and the response variables.