IDS702_Turquoise_Final_Project

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

2022-12-03

Abstract

This analysis aims to find factors that affects data science job salaries and factors associated with working remotely from office. This factors can be found by analyzing a dataset consisting of data science job salaries around the globe with a few variables characterizing the respondent's current job obtained from Kaggle . We found that experience level (executive, senior, mid-level, in reference to entry-level) and company location (Europe and North America in reference to Asia) have significant relationship and tends to increased data science job salaries, while job title (data analyst in reference to data engineer) also have a significant relationship but tends to decrease data science job salaries. As for remote working, we found that job title, employment type, and company location having a significant relationship with remote ratio.

Introduction

The data science field is sitting at the intersection of statistics and computer science. This intersection is proven by most data science jobs posting that require statistics analysis or data modeling using programming languages. In 2019, job postings for data science related jobs had risen by 256% (Davenport & Patil, 2022). This trend raised the question about the salary potential of data science jobs in the future. Another rising question following the Covid-19 virus in 2020 is remote working. Which also raised the question about what factors affecting remote working in the field of data science (Gifford, 2022).

This analysis aims to answer these questions:

1. Which factors are associated with an increase in salary for data science jobs? (Continuous outcome)
2. How do company size, company location, employment type, employee residence and job title affected the remote work ratio of a data scientist? (Discrete outcome) For this question, there are three possible values for the remote work ratio, 0, 50, and 100; these signify an in-person job, a hybrid job, and a fully remote job, respectively.

The data used in this project comes from Kaggle with 607 observation with 11 variables. This dataset has no null or missing value. This is the link to original post: Data Science Job Salaries The variables in this dataset are:

- Work Year = The year salary was paid.
- Experience Level = Level of experience in the current job, categorized into Executive Level (EX), Senior Level (SE), Mid Level (MI), Entry Level (EN)
- Employment Type = Employment type in the current job, categorized into Full-Time (FT), Part-Time (PT), Contract (CT), Freelance (FL)
- Job Title = Job title in the current company with 50 unique entries
- Salary = Annual gross salary in the specific currency
- Salary Currency = Currency of the annual salary variable

- Salary (in USD) = Normalized annual salary from the respective currency into USD
- Employee Residence = Country of residence of each respondents with 57 unique values
- Company Location = Country of company location with 50 unique values
- Company Size = Company size based on number of employees, categorized into Small (S), Medium (M), Large (L)

Table 1: Variable Names and Description

Column	Description
work__year	The year the salary was paid.
experience__level	The experience level in the job during the year with the following possible values: EN Entry-level / Junior MI Mid-level / Intermediate SE Senior-level / Expert EX Executive-level / Director
employment__type	The type of employement for the role: PT Part-time FT Full-time CT Contract FL Freelance
job__title	The role worked in during the year.
salary	The total gross salary amount paid.
salary__currency	The currency of the salary paid as an ISO 4217 currency code.
salaryinusd	The salary in USD (FX rate divided by avg. USD rate for the respective year via fxdata.foorilla.com).
employee__residence	Employee's primary country of residence in during the work year as an ISO 3166 country code.
remote__ratio	The overall amount of work done remotely, possible values are as follows: 0 No remote work (less than 20%) 50 Partially remote 100 Fully remote (more than 80%)
company__location	The country of the employer's main office or contracting branch as an ISO 3166 country code.
company__size	The average number of people that worked for the company during the year: S less than 50 employees (small) M 50 to 250 employees (medium) L more than 250 employees (large)

Methods

For one of our research questions we choose the outcome variable to be salary__in__usd to ensure all the data is converted to the same unit (usd). As shown by the table below, that there are other currency in this dataset.

Table 2: Salary Variable vs Salary in USD Variable

salary__currency	salary	salary__in__usd
EUR	70000	79833
USD	260000	260000
GBP	85000	109024
USD	20000	20000
USD	150000	150000
USD	72000	72000

We choose `salary_in_usd` over similar columns, `salary` and `salary_currency`, in order to standardize our outcome variable and reduce noise and colinearity.

3. Primary relationship of interest

Table 3 below is showing the descriptive statistic for each variable. Variable with asterisks are categorical variable that needs to be look into further in model building.

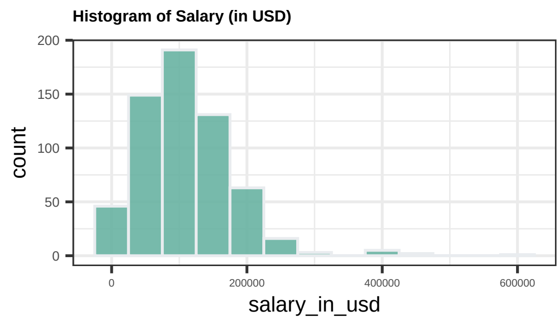
3.1. Descriptive stats and plots Answering question 1

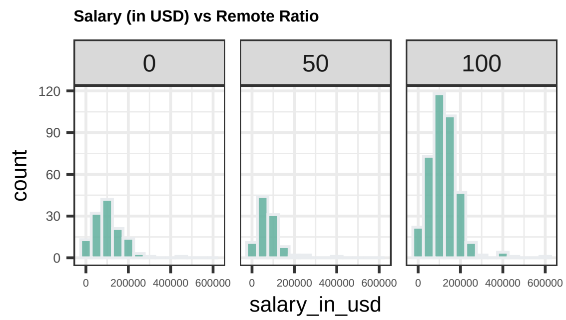
Our first outcome variable, salary in USD, ranged from USD2,859 to USD600,000. Diving into each variable, the majority of the respondents are Senior Level employee (46%), which majority working fully remote

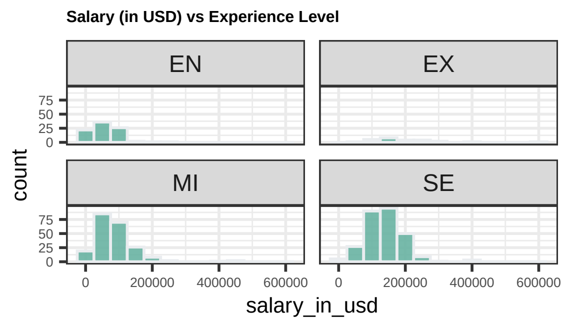
Table 3: Summary of the DS Salary Dataset

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
X	1	607	303.00	175.37	303	303.00	225.36	0	606	606	0.00	-1.21	7.12
work_year	2	607	2021.41	0.69	2022	2021.51	0.00	2020	2022	2	-0.73	-0.66	0.03
experience_level*	3	607	3.13	1.03	3	3.28	1.48	1	4	3	-1.04	-0.10	0.04
employment_type*	4	607	2.99	0.24	3	3.00	0.00	1	4	3	-4.14	45.81	0.01
job_title*	5	607	21.96	10.49	18	21.00	7.41	1	50	49	0.88	0.40	0.43
salary	6	607	324000.06	1544357.49	115000	118919.11	68706.65	4000	30400000	30396000	13.98	244.57	62683.54
salary_currency*	7	607	14.03	4.38	17	14.67	0.00	1	17	16	-1.03	-0.38	0.18
salary_in_usd	8	607	112297.87	70957.26	101570	106157.63	62906.72	2859	600000	597141	1.66	6.26	2880.07
employee_residence*	9	607	41.41	18.27	56	43.66	0.00	1	57	56	-0.67	-1.22	0.74
remote_ratio	10	607	70.92	40.71	100	76.08	0.00	0	100	100	-0.90	-0.90	1.65
company_location*	11	607	36.89	16.03	49	39.07	0.00	1	50	49	-0.77	-1.09	0.65
company_size*	12	607	1.81	0.65	2	1.76	0.00	1	3	2	0.21	-0.73	0.03

(71%), full in office (19%), or hybrid (10%). Most of the Senior Level employee works in medium size companies (66%), while the rest are working in large companies (26%) or small companies (8%). Most of their employment status are Full-Time (99%), while the rest are Contracts or Freelance. None of them are Part-Time employee. The second biggest respondents are Mid/Intermediate Level employee (35%), which majority working fully remote (54%), full in office (26%), or hybrid (20%). Most of the Mid/Intermediate Level employee works in medium size companies (46%), while the rest are working in large companies (40%) or small companies (14%). Most of their employment status are Full-Time (97%), while the rest are Contracts, Freelance, or Part-Time employee. The remaining respondents are Entry/Junior Level employee (15%) and Executive/Director level (4%). There are 50 different company location, which mostly in the US (58%), followed by Great Britain (8%), Canada (5%), and the rest of the world. While the employee residence data shows that respondents live in 57 different country, probably made possible by the ability to work remotely. Most of the respondents lived in the US (55%), followed by Great Britain (7%), India (5%), and the rest of the world. There are 50 different job titles in this dataset, but all of them are in the field of data science. The job title variable is dominated by data scientist (24%), data engineer (22%), data analyst (16%), while the rest are varied but mostly have “engineer” or “data” in the title. Based on the barplot shown below, there are certain pattern from Remote Ratio, Company Size and Experience level when plotted against salary (in USD).





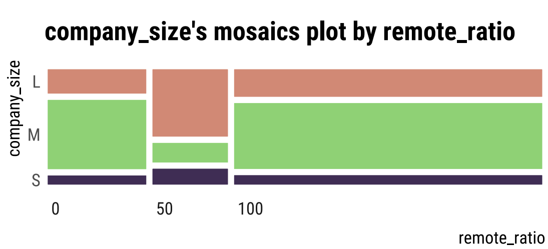


3.2. Descriptive stats and plots Answering question 2

We are interested in the correlation between company size and remote ratio. From the mosaic plot of company size and remote ratio, it is obvious that most companies from this survey are medium-sized, and most companies allow their employees to work from home. However, more large-sized companies (60) adopt hybrid working than medium-sized ones (21). According to the table, for these three types of companies, the percentage of supporting 100% remote work is the highest (54%, 69% and 59% respectively), when compared to those supporting 50% remote work and working on-site. According to the figure of remote ratio and employment type, most employees that work from home are full-time. It is interesting to note that the number of freelancers with these three types of work is approximately the same. All of the contract employees in this survey work remotely.

Table 4: Company Size vs Remote Ratio

company_size/remote_ratio	0	50	100	Total
L	16% (32)	30% (60)	54% (106)	100% (198)
M	24% (79)	6% (21)	69% (226)	100% (326)
S	19% (16)	22% (18)	59% (49)	100% (83)
Total	21% (127)	16% (99)	63% (381)	100% (607)



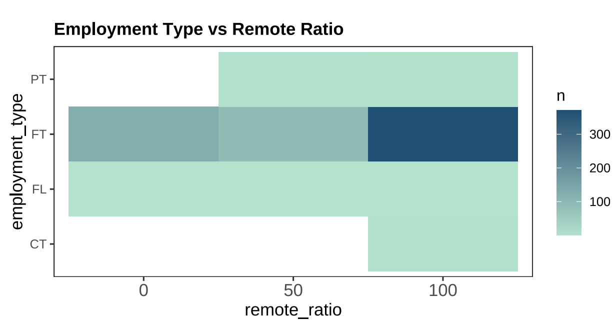


Table 5: Employment Type vs Remote Ratio

employment_type/remote_ratio	0	50	100	Total
CT	0% (0)	0% (0)	100% (5)	100% (5)
FL	25% (1)	25% (1)	50% (2)	100% (4)
FT	21% (126)	16% (92)	63% (370)	100% (588)
PT	0% (0)	60% (6)	40% (4)	100% (10)
Total	21% (127)	16% (99)	63% (381)	100% (607)

4. Other characteristics Our dataset includes work_year, which is the year that the data was collected. This variables contains three unique values, 2020, 2021, and 2022, which contain 11.86%, 35.75%, and 52.39% respectively. Similarly, job_title has 50 unique values. Below is a table showing the unique values in the variable and their respective counts.

Table 6: Summary of Job Titles

Var1	Freq
3D Computer Vision Researcher	1
AI Scientist	7
Analytics Engineer	4
Applied Data Scientist	5
Applied Machine Learning Scientist	4
BI Data Analyst	6
Big Data Architect	1
Big Data Engineer	8
Business Data Analyst	5
Cloud Data Engineer	2

Var1	Freq
Computer Vision Engineer	6
Computer Vision Software Engineer	3
Data Analyst	97
Data Analytics Engineer	4
Data Analytics Lead	1
Data Analytics Manager	7
Data Architect	11
Data Engineer	132
Data Engineering Manager	5
Data Science Consultant	7

Var1	Freq
Data Science Engineer	3
Data Science Manager	12
Data Scientist	143
Data Specialist	1
Director of Data Engineering	2
Director of Data Science	7
ETL Developer	2
Finance Data Analyst	1
Financial Data Analyst	2
Head of Data	5

Var1	Freq
Head of Data Science	4
Head of Machine Learning	1
Lead Data Analyst	3
Lead Data Engineer	6
Lead Data Scientist	3
Lead Machine Learning Engineer	1
Machine Learning Developer	3
Machine Learning Engineer	41
Machine Learning Infrastructure Engineer	3
Machine Learning Manager	1

Var1	Freq
Machine Learning Scientist	8
Marketing Data Analyst	1
ML Engineer	6
NLP Engineer	1
Principal Data Analyst	2
Principal Data Engineer	3
Principal Data Scientist	7
Product Data Analyst	2
Research Scientist	16
Staff Data Scientist	1

From table 4, we can see that the job titles related to data science are not in short supply. Furthermore, company_location has 50 unique values and employee_residence has 57 unique values. They are each listed from table 5 to table 8.

Table 7: Summary of Employee Residence(1)

AE	AR	AT	AU	BE	BG	BO	BR	CA	CH	CL	CN	CO	CZ	DE	DK	DZ	EE	ES	FR	GB	GR	HK	HN	HR
3	1	3	3	2	1	1	6	29	1	1	1	1	1	25	2	1	1	15	18	44	13	1	1	1

Table 8: Summary of Employee Residence(2)

HU	IE	IN	IQ	IR	IT	JE	JP	KE	LU	MD	MT	MX	MY	NG	NL	NZ	PH	PK	PL	PR	PT	RO	RS	RU	SG	SI	TN	TR	UA	US	VN
2	1	30	1	1	4	1	7	1	1	1	1	2	1	2	5	1	1	6	4	1	6	2	1	4	2	2	1	3	1	332	3

The company_size variable is also of note: it has three potential variables, S for small, M for medium, and L for large. Its distribution of counts is listed in Table 9.

Table 9: Summary of Company Location(1)

AE	AS	AT	AU	BE	BR	CA	CH	CL	CN	CO	CZ	DE	DK	DZ	EE	ES	FR	GB	GR	HN	HR	HU	IE	IL
3	1	4	3	2	3	30	2	1	2	1	2	28	3	1	1	14	15	47	11	1	1	1	1	1

Table 10: Summary of Company Location(2)

IN	IQ	IR	IT	JP	KE	LU	MD	MT	MX	MY	NG	NL	NZ	PK	PL	PT	RO	RU	SG	SI	TR	UA	US	VN
24	1	1	2	6	1	3	1	1	3	1	2	4	1	3	4	4	1	2	1	2	3	1	355	1

Table 11: Summary of Company Size

L	M	S
198	326	83

5. Modeling

5.1 Multivariate Linear Regression Model to answer Question #1

- Based on our guesses, we assume that a larger `company_size`, full time as the `employment_type`, higher `experience_level`, and a more "hardcore" `employment_type` will lead to a higher `salary_in_usd`. Thus the `model1` is our preliminary guess on the predictors.
- From the baselines model above, we can see that there are something that are statistically significant. For example, if the job title is "Principal Data Engineer", "Financial Data Analyst", "Data Analytics Lead", "Data Analytics Engineer", then these will be the effective predictors of `salary_in_usd` that will likely be the predictors that drastically increase the salary.
- This `model1` is statistically significant. The adjusted R-squared that we got is 0.2793, which is not too high. Thus we are trying to explore better linear models to fit the data.
- In the next few models, I will try out more combinations before using forward,backward,and stepwise selection to select the features.
- In the model above, we added two extra features into the model: `remote_ratio` and `salary_currency`. The reason why we consider these two variables are that: If the position is remote, we consider that job should play an important role in the company and expect a higher salaries. The currency matters because in some more underdeveloped countries with their own currency will lead to a lower salary in terms of their national economic status. The p-value we get here is $< 2.2e\text{-}16$, which means that this model is statistically significant with an adjusted R-squared of 0.4347.
- Something to notice is that none of the `company_location` has a p-value that is smaller than the threshold, same to the `remote_ratio`.
- We added two more predictors in the `model3`, `employee_residence` and `company_location`.
- The p-value we get from `model3` is $< 2.2e\text{-}16$, which means that this model is statistically significant with an adjusted R-squared of 0.4601, which explains 46.01% of the variation in `model3` can be accounted by these predictors.

```
##            (Intercept)      experience_levelEX      experience_levelMI
##               56943.91               133076.38                26049.23
##     experience_levelSE   job_titleData Analyst remote_ratioFull Remote
##               74659.06               -23620.26                20098.15
##     company_locationSA
##                   0.00
```

```r
coef(forward.model$finalModel, 8)
```

```
##            (Intercept)      experience_levelEX      experience_levelMI 
##              39480.887              118804.472               22493.125 
##     experience_levelSE   job_titleData Analyst remote_ratioFull Remote 
##              51537.071              -31315.920                5819.275 
##     company_locationNA      company_locationSA      company_locationAF 
##              62255.906                   0.000                   0.000
```

Final model is here:

Table 12:

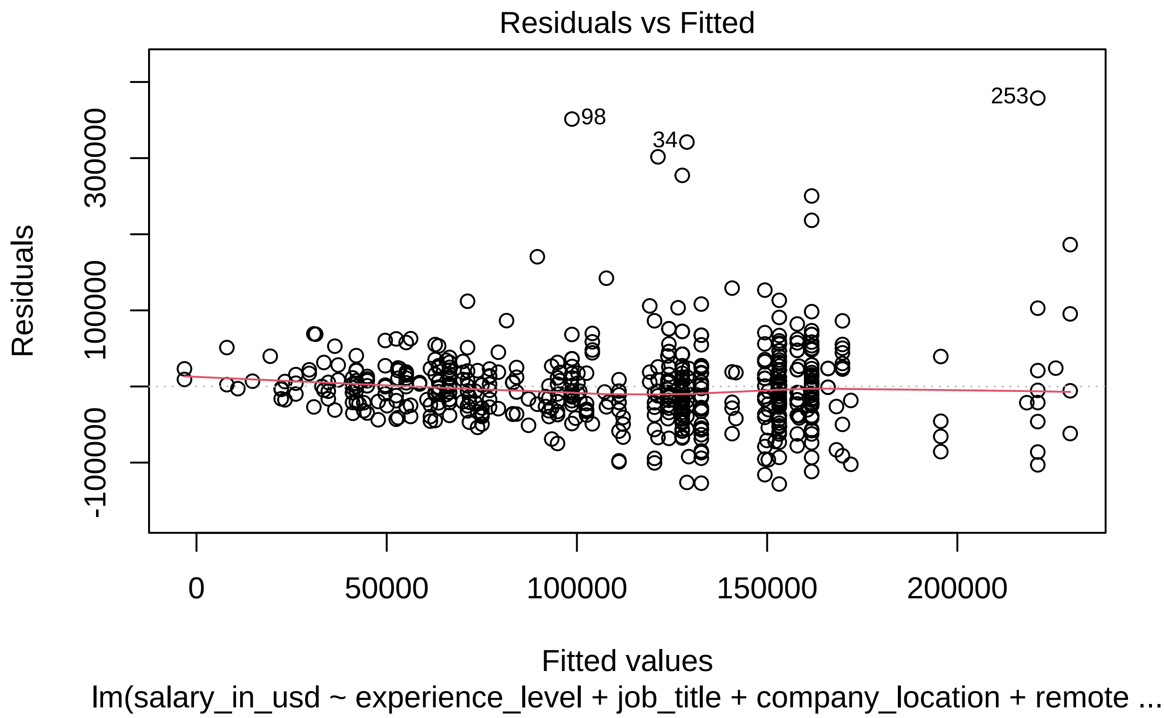
	Dependent variable:
	salary_in_usd
experience_levelEX	118,604.30***
experience_levelMI	21,642.55***
experience_levelSE	50,650.79***
job_titleData Scientist	8,520.92
job_titleData Analyst	−25,494.46***
job_titleMachine Learning Engineer	16,609.94**
company_locationEU	18,759.60**
company_locationNA	76,415.42***
company_locationEE	10,704.39
remote_ratioHybrid	−7,611.04
remote_ratioFull Remote	3,785.59
Constant	22,335.92**
Observations	587
R^2	0.42
Adjusted R^2	0.41
Residual Std. Error	54,615.72 (df = 575)
F Statistic	37.62*** (df = 11; 575)
Note:	*p<0.1; **p<0.05; ***p<0.01

- To check the assumptions for Linear Regression, we need to make sure the following:

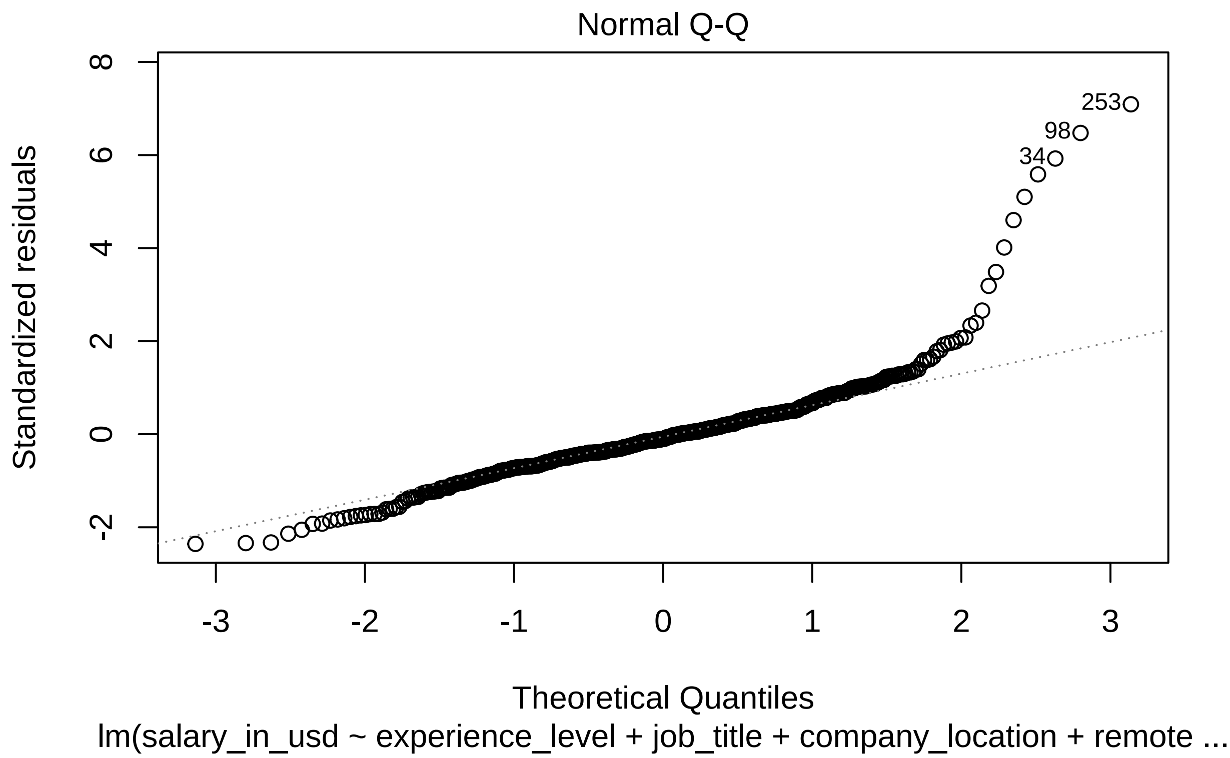
1. Linearity
2. Equal Variance of Error Terms (Heteroscedasticity)
3. Independence of Error Terms
4. Normality of Error Terms
5. Leverage Points and Outliers –> Influential Points

- I have plotted the residuals vs. each predictor, residual vs. fitted values, Q-Q plots. These three files helped us to check if the assumptions are met. In the residual vs. fitted values, We find out that there is no obvious pattern/trend in the plot, which satisfy the assumption of Independence of Error Terms. There are an obvious funnel-shape trend in the plot. Thus, it meets the assumption of Equal Variance of Error Terms because the points are pretty equally spread out around the y = 0 line.
- In the plot of residual vs. the predictor, it has no obvious trend/curve that is slightly not linear, and it agrees the assumption of linearity. Lastly, when we checked the Q-Q Plot, the points are very close to the 45-degree line except for a few points. Thus the last assumption of normality of error terms are checked.

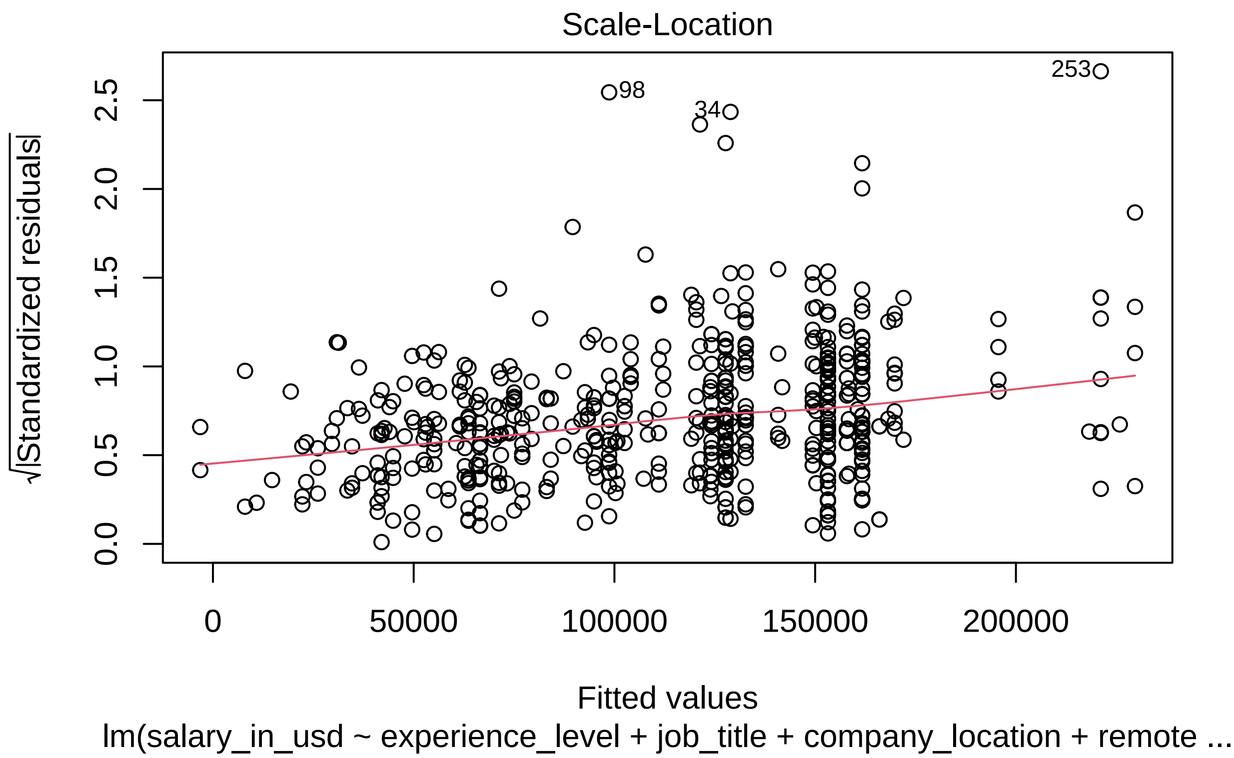
- The regression assumptions are plausible because 1) Linearity match rigorously; 2)Independence of Error Terms; 3)the residual plots does not show any heteroscedasticity; 4) Normality of Error Terms.




```
## Warning: not plotting observations with leverage one:
##   486
```




```
## Warning: not plotting observations with leverage one:
##   486
```



Scale-Location
98
34
253
0.0
0.5
1.0
1.5
2.0
2.5
√|Standardized residuals|
0
50000
100000
150000
200000
Fitted values
lm(salary_in_usd ~ experience_level + job_title + company_location + remote ...

6. Potential challenges

A challenge we'll need to address before modeling is the amount of job titles that are present in our data. To solve this, we will either collaspe certain job titles into relevant categories or exclude job titles based on some metric. In that, if the job title "NLP Engineer" only shows up once, we may decide to drop it and all other job titles with similar counts. This decision will be informed by domain knowledge and field experts. We face a similar problem with employee_residence and company_location. Not only is more than half of our data from the United States, Great Britian, and Canada, but the two variables are also highly correlated. Because we're investigating the relationship between work place attributes and the remote-ratio, we will include both location-based variables. There are various common data issues that need to be addressed before modeling. We consider them below. Messy data is defined to be a dataset where values are unstandardized, unorganized, or bias. Largely, our data is clean. The only potential messiness in our data comes from the large number of job titles as mentioned above. Another common data issue is a lack of data. As our data has 607 entries, we do not face this problem. Finally, confounding variables can cause modeling issues as there may be variables that are related to our questions at hand that are not present in our data. Because our research questions are somewhat poignant in their phrasing, we will not encounter this as a data issue. One might suppose that a level of education may have an effect on a data scientist's salary as it tends to have an effect in other fields. However, because data science is a relatively new field, there are not much data on whether or not a higher-education degree has an effect on the ensuing salary. For this reason, we find no issue in assuming it has marginal effect.