IDS702 Turquoise Final Project

Beibei Du, Wafiakmal Miftah, Suzanna Thompson, Alisa Tian

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(Bhatia, 2022). This dataset has no null or missing value.

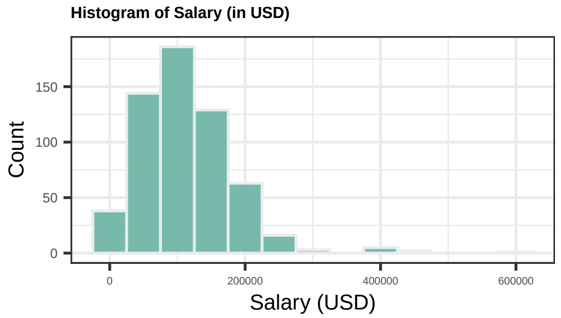
The variables in this dataset are:

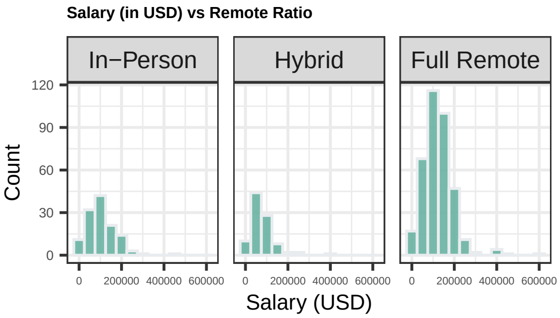
- Work Year = The year salary was paid.
- Experience Level = Level of experience in the current job, categorized into Executive Level (EX), Senior Level (SE), Mid Level (MI), Entry Level (EN)
- Employment Type = Employment type in the current job, categorized into Full-Time (FT), Part-Time (PT), Contract (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

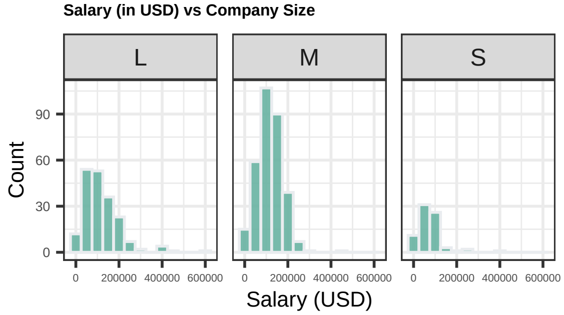
- Company Location = Country of company location with 50 unique values
- Company Size = Company size based on number of employees, categorized into Small (S), Medium (M), Large (L)

Methods

Data: EDA For Salary and Remote Ratio


Histogram of Salary (in USD)


Salary (in USD) vs Remote Ratio


Salary (in USD) vs Company Size

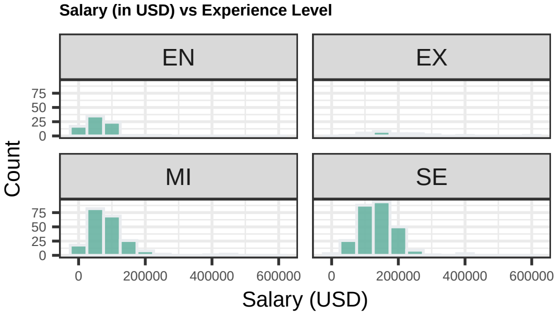

Salary (in USD) vs Experience Level

Table 1: Company Size vs Remote Ratio

Company Size / Remote Ratio	In-Person	Hybrid	Full Remote	Total
L	17% (32)	31% (60)	52% (101)	100% (193)
M	24% (78)	6% (18)	70% (223)	100% (319)
S	20% (15)	23% (17)	57% (43)	100% (75)
Total	21% (125)	16% (95)	63% (367)	100% (587)

Table 2: Employment Type vs Remote Ratio

Employment Type / Remote Ratio	In-Person	Hybrid	Full Remote	Total
CT	0% (0)	0% (0)	100% (4)	100% (4)
FL	25% (1)	25% (1)	50% (2)	100% (4)
FT	22% (124)	16% (89)	63% (358)	100% (571)
PT	0% (0)	62% (5)	38% (3)	100% (8)
Total	21% (125)	16% (95)	63% (367)	100% (587)

Our dataset has 11 variables, each already described in `introduction`. The first outcome variable, `salary (USD)`, has a right skewed distribution, which implied that our data has very few respondents with extremely high salary. This connects with the `experience level` variable that has small counts of `executive level`, as shown by the figure above. We decided to drop `salary` and `salary_currency` for obvious reason, and variable `work_year` because it is not relevant to answering the two research questions. We also collapsed some of the variable to lower unique value:

1. Job Title: 50 similar job titles into 4 job titles, Data Analyst, Data Scientist, Data Engineer, Machine Learning Engineer
2. Employee Residence: 50 countries into 3 different continents, Asia, North America, and Europe. We dropped countries from Africa and Oceania as both has number of respondents of less than 5.
3. Company Location: 50 countries into 3 different continents, Asia, North America, and Europe. We dropped countries from Africa and Oceania as both has number of respondents of less than 5.

Before we fit the model, all categorical variables is transformed from string to factor to ensure there's no typo and it will show up as its own level in the regression result.

With our data, there is inherent multicolinearity. It is clear that salary and experience level will have multicolinearity, as well as company location and employee location. However, our data only has one continuous variable and the rest of the variables are categorical, so we do not need to worry about multicolinearity with each categorical variable and the continuous variable. Between many categorical variables there is also correlation, as is the case with employee location and company location. Because our dataset is as limited as it is, we are chosing to include variables that have mulitcolinearity and correlation, and we keep this in mind as we do our analysis.

Models and Model Assesment

Question 1 To examine the changes in Salary (USD), a continuous variable as our outcome/response variable, we are considering to use Linear Regression Models, more specifically, Multiple Linear Regression Models. We will fit and assess various models considering experience level, employment type, employee residence, remote ratio, job title, company location and company size as variables. We will use Forward, backward, stepwise selections to pick the best features. Besides, taking AIC, BIC, Adjusted-R-Squared into consideration is highly valued as well. Since forward and backward selections have their own disadvantages and stepwise is the main selection we are taking into consideration.

Question 2 The response variable (remote ratio) is a categorical variable with 3 levels: In-Person, Hybrid, and Fully Remote. Thus the multiple linear regression model from first question is not applicable here. In this case, we can use multinomial logistic regression model. The log odds of remote ration will be calculated as a combination of all the predictor variables we are interested in (such as the experience level, job title, employee residence, company location, and company size). We are excluding employment type, as the data is highly concentrated to Full-Time. Multinomial model usually used use chi-squared test and change in deviance test to select the best model. But in this case, because we only have 6 predictors, we use all the predictors and assess the model accuracy using confusion matrix and ROC curve. Lastly, for the second model, we do a base relevel to suits our interests, for variable `job title` into Data Engineer and for variable `remote ratio` into In-Person.

Results

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.

Table 3: Multiple Linear Regression First Model Result for Salary (USD)

	Dependent variable:
	salary_in_usd
company_sizeM	−8,954.39
company_sizeS	−30,454.39***
employment_typeFL	−155,340.00***
employment_typeFT	−100,569.70***
employment_typePT	−143,985.90***
job_titleData Scientist	5,225.95
job_titleData Analyst	−19,575.73***
job_titleML Engineer	4,150.53
experience_levelEX	126,573.90***
experience_levelMI	21,769.26***
experience_levelSE	72,797.02***
Constant	178,560.50***
Observations	587
R^2	0.30
Adjusted R^2	0.28
Residual Std. Error	60,000.01 (df = 575)
F Statistic	22.21*** (df = 11; 575)
Note:	*p<0.1; **p<0.05; ***p<0.01

From the result of the baseline 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, we 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-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-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.

Final model is here:

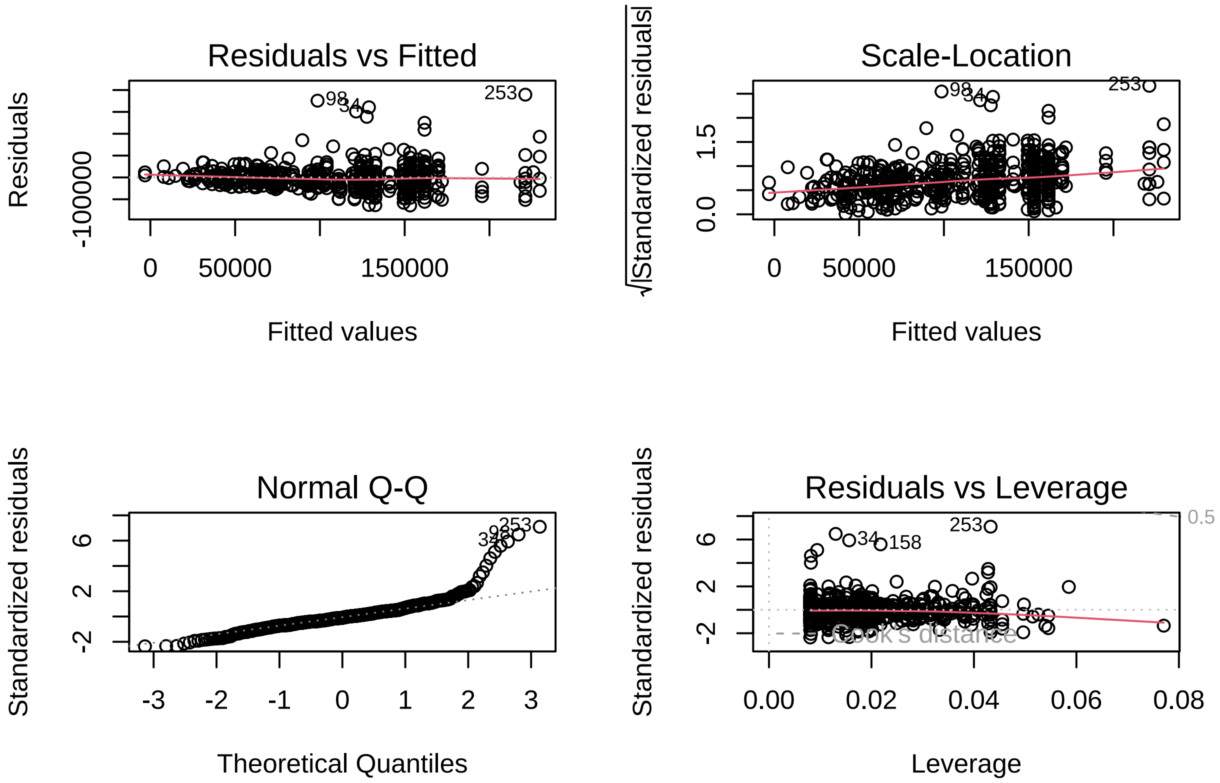
Table 4: Multiple Linear Regression Final Model Result for Salary (USD)

	Dependent variable:
	salary_in_usd
experience_levelEX	118,598.60***
experience_levelMI	21,604.76***
experience_levelSE	50,641.03***
job_titleData Scientist	8,514.97
job_titleData Analyst	−25,563.19***
job_titleML Engineer	16,606.31**
company_locationEU	18,712.63**
company_locationNA	76,434.83***
remote_ratioHybrid	−7,597.33
remote_ratioFull Remote	3,755.73
Constant	22,374.72**
Observations	587
R^2	0.42
Adjusted R^2	0.41
Residual Std. Error	54,569.31 (df = 576)
F Statistic	41.45*** (df = 10; 576)
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

We 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 is not 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. In the plot of leverage vs. standardized residuals, we can see that everything is inside of the range and nothing is outside of the cook's distance. Thus we can conclude that there are no influential points in this model. 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 5) No influential points.



Question 2

Table 5: Multinomial Logistic Regression Model Result for Remote Status

	Dependent variable:	
	Hybrid	Full Remote
	(1)	(2)
experience_levelEX	0.15 (0.88)	0.50 (0.71)
experience_levelMI	−0.64 (0.44)	−0.48 (0.37)
experience_levelSE	−0.54 (0.47)	−0.05 (0.38)
job_titleData Scientist	0.24 (0.38)	−0.47* (0.26)
job_titleData Analyst	0.07 (0.48)	0.06 (0.30)
job_titleML Engineer	1.06** (0.52)	−0.02 (0.42)
employee_residenceEU	−1.35 (1.32)	−1.87 (1.27)
employee_residenceNA	−2.37* (1.43)	−2.18* (1.32)
company_locationEU	2.05 (1.35)	2.25* (1.31)
company_locationNA	1.52 (1.43)	3.09** (1.34)
company_sizeM	−2.07*** (0.36)	−0.22 (0.25)
company_sizeS	−1.04** (0.45)	−0.04 (0.39)
Constant	0.95 (0.61)	0.81 (0.51)
Akaike Inf. Crit.	973.51	973.51
Note:	*p<0.1; **p<0.05; ***p<0.01	

Setting our p-value threshold to 0.05 for the result of multinomial logistic regression above, we found that the significant variables are job title, company location, and company size. The model interpretation are:

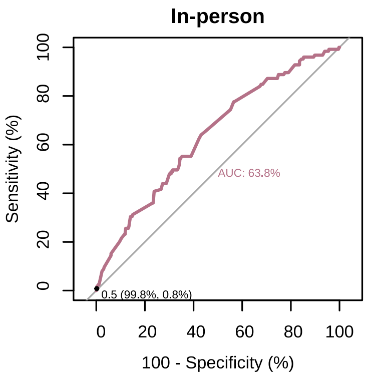
1. Machine learning engineer (compared to data engineer) is associated with a 1.06 increase in the log odds of hybrid (combination of remote and in-person) working compared to in-person working.
2. Company located in North America (compared to located in Asia) is associated with a 3.09 increase in the log odds of fully remote working compared to in-person working.
3. Medium company (compared to large company) is associated with a -2.07 decrease in the log odds of hybrid working compared to in-person working.
4. Small company (compared to large company) is associated with a -1.04 decrease in the log odds of hybrid working compared to in-person working.

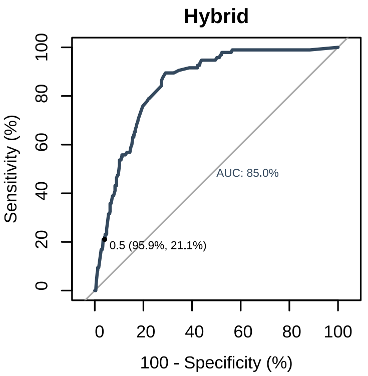
In order to assess our model, our discussion with experts indicated that we don't have to do a deviance test in choosing the best model, so we can go straight to calculating accuracy using Confusion Matrix by using the current model. This confusion matrix yields an accuracy of `66.10%`.

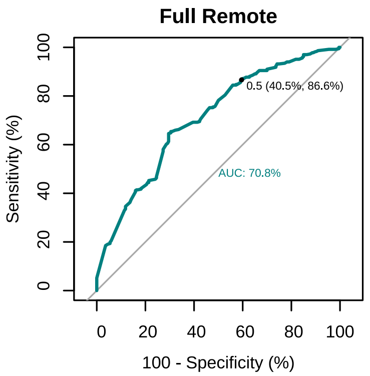
Table 6: Confusion Matrix for Multinomial Model

	True In-Person	True Hybrid	True Full Remote
Predicted In-Person	3	0	2
Predicted Hybrid	11	42	22
Predicted Full Remote	111	53	343

In addition, to measure our prediction we can use Receiver Operator Characteristic (ROC) curve with the axis as true positive rate (Sensitivity) and true negative rate (1 - Specificity). We use the standard cut-off 0.5, and we can see that for in-person, the area under the curve is 63.9%, for hybrid, the area under the curve is 85%, and for full remote, the area under the curve is 70.8%. Even though the result for predicting in-person is low with 122 mistakes in predicting no remote, we have a good prediction for full remote, with only 24 mistakes.







Conclusion

Based on the two optimized final models we had previously, we could see that a correct career path choice will positively affect the salaries received. Although salaries should not be the only consideration when choosing a career path, it still plays an important role in people's life choices. Additionally, the company location and size should be considered comprehensively as the factors of the salaries. For example, in some locations, the salaries are higher. However, at the same time, the commodity prices are higher accordingly. Thus, when making decisions, people should also look beyond the model. The remote ratio is something that just happened and has been trending in the recent three years, mainly because of the covid-19 pandemic. We are not sure if that's continuously happening in the future in the post-covid19 era. Although Company Location, Employment Type, and Job Title are the statistically significant predictors impacting whether an employee works remotely, hybrid, or in person, we should also take the current covid-19 cases into consideration and make adjustments for the future data science job remote ratio.

Key Takeaways

As mentioned in the introduction, this research aims to explain the factor that affects higher salary and working remote for data science jobs. This research will be useful for human resource professionals, job seeker, or journalist that will keep track and update the data from time to time. Though there are some limitations that will be mentioned below.

Limitations

From interpreting and assessing the result of both multiple linear regression and multinomial regression, we found some limitation: 1. Relatively low number of entries (only 607) 2. Relatively low number of variables (only 8 valid variables after dropping salary, salary currency, and year) 3. Relatively high percentage of data coming from North America (65% for company location, 62% for employee residence)

We believe that this study can still be improved by focusing to resolve these limitations.

Future Potential Study

In order to increase the accuracy of our model, some of the potential for future study are: 1. Adding interaction term for job title and experience level 2. Focusing the study in one continent, such as the United States and adding detail for each states and county. 3. Trying an ordinal model and assess the model using multinomial model.

References

Bhatia, R. (2022, May 1). *Data Science Job Salaries.* Retrieved September 30, 2022, from Kaggle: https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries

Davenport, T. H., & Patil, D. J. (2022, July 21). Is data scientist still the sexiest job of the 21st century? *Harvard Business Review.* Retrieved December 3, 2022, from https://hbr.org/2022/07/is-data-scientist-still-the-sexiest-job-of-the-21st-century

Gifford, J. (2022, March 15). Remote Working: Unprecedented Increase. *Human Resource Development International*, 10.