A Statistical Exploration of Data Science Career Salaries in the United States

Paul Thomson, Emilio Velazquez-Carter, Dante Eschleman

04/04/2025

1 Introduction

Using the Data Science Salaries 2024 dataset (link [1] in references), we intend to analyze which variables affect data analytics career salaries across the world. The central question guiding this analysis is: What are the most important factors affecting salaries, and how do these trends vary across job roles, locations, work models, and other characteristics?

The dataset contains approximately 6600 rows, with data from the years 2020 to 2024. The primary outcome variable in this analysis is salary_in_usd. We will fit a linear regression model to the data. We intend to use stepwise regression to determine which predictors are most significant and then use those predictors to train the predictive model. In addition to linear regression, we will employ a random forest model. By comparing both models, we aim to identify important variables and also compare the effectiveness of each modeling technique.

The following is a description of each relevant variable:

job_title: The job title or role associated with the reported salary.

experience_level: The level of experience of the individual.

employment_type: Indicates whether the employment is full-time, part-time, etc.

work_models: Describes different working models (remote, on-site, hybrid).

work_year: The specific year in which the salary information was recorded.

employee_residence: The residence location of the employee. (country)

salary_in_usd: The converted salary in US dollars. <— This will be our outcome variable.

company_location: The geographic location of the employing organization.

company_size: The size of the company, categorized by the number of employees.

2 Data Processing

After reading the data and removing redundant columns (salary_currency and salary), we eliminated instances with missing values. In order to reduce outliers, we filtered out job titles with fewer than 50 occurrences and employee residences with fewer than 3 occurrences, and then removed any unused factor levels. In total we removed 1086 observations.

3 Data Exploration

The descriptive statistics were calculated using built in R functions and then displayed in the table using the kable function. The histogram was made using the built in hist function and r and shows that the majority of the salaries are within the 0-200,000 dollar range with outliers at 400k+ and some even around the 700k range.

	Salary (USD)
Minimum	15680.00
1st Quartile	100000.00
Median	140544.00
Mean	148848.31
Standard Deviation	70088.41
3rd Quartile	187415.00
Maximum	750000.00

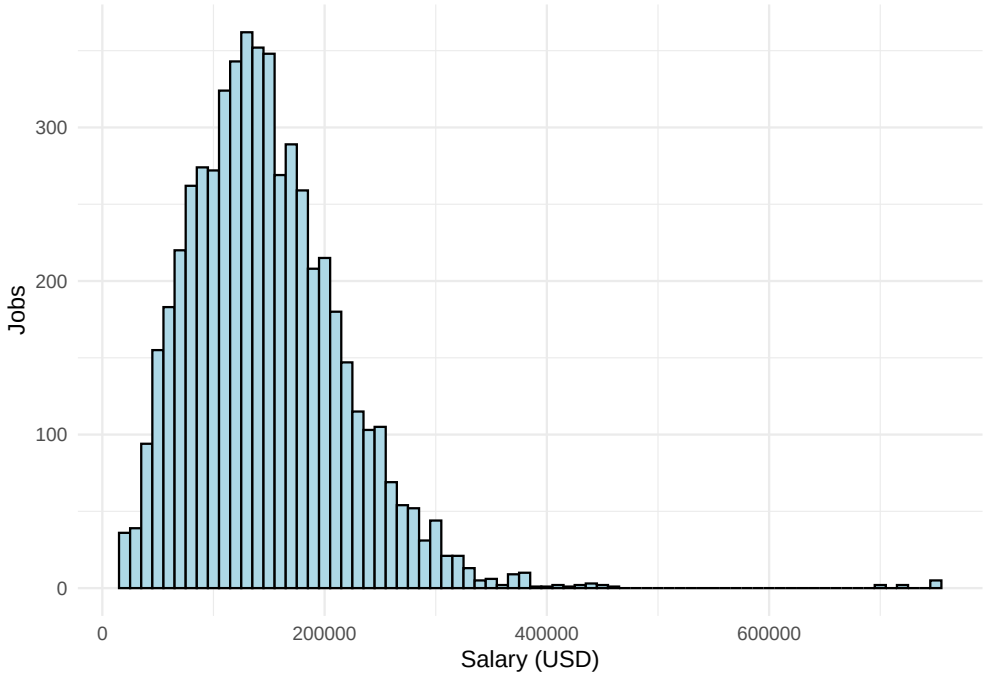


Figure 1: Descriptive Statistics for Data Science Salaries in USD (2020-2024)

4 Training and Testing Data Sets

We split the data into training and testing sets using random sampling. 80% of the data was used in the training set and 20% was used for the test set. In this project, we use 10-fold cross-validation for model training. Also, we changed salary_in_usd to a log function to reduce the variance and improve the fit performance. It also helps us measure the effects of each predictor as a percent change in salary.

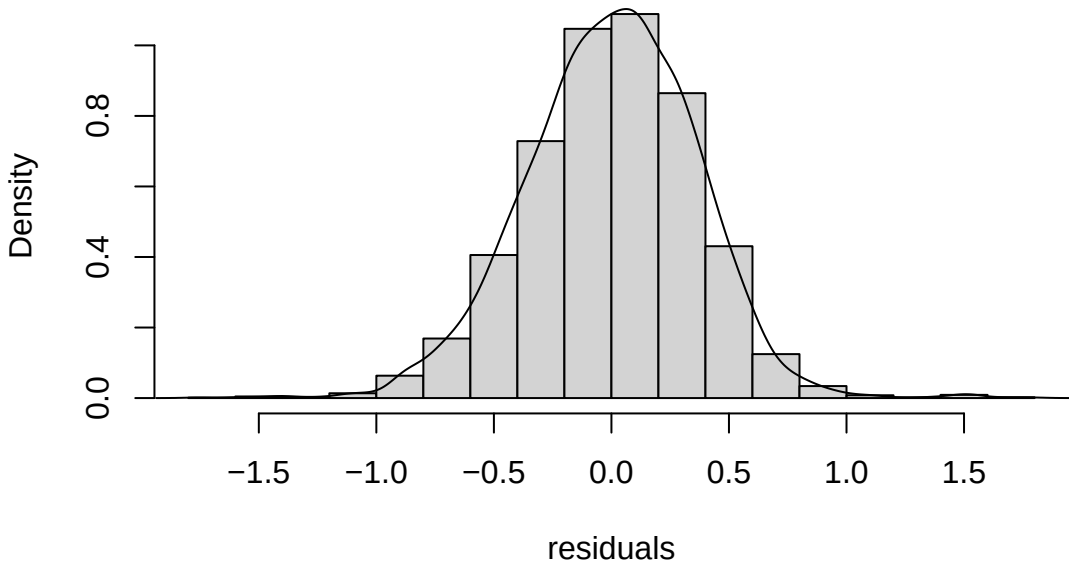
5 Modeling

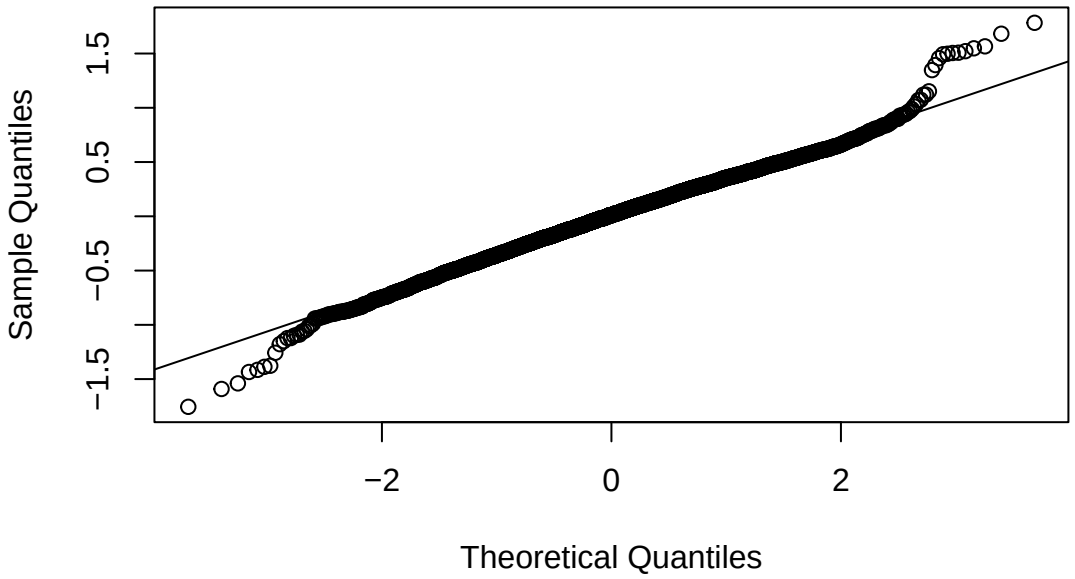
5.1 Linear Regression Model

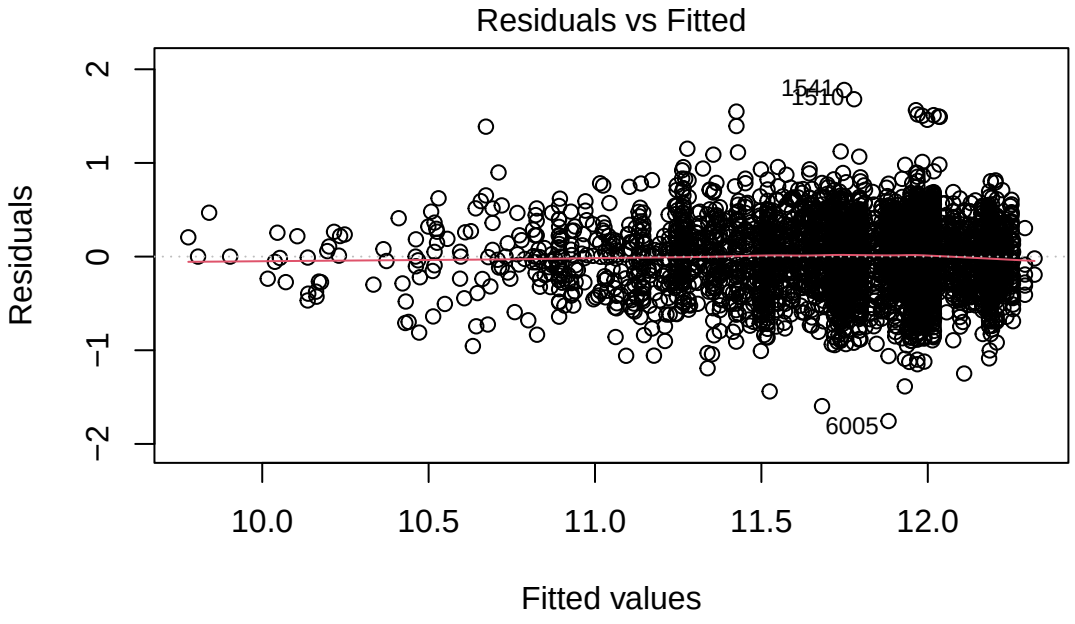
At first our model used all the possible predictors. We then used stepwise regression to discover the most important predictors to create a final model. It decided that each predictor was important to include, so then that model was then trained using 10-fold cross validation.

5.1.1 Model 1 Fit Results

The stepwise regression model identified the job_title, experience_level, employee_residence, work_models, work_year, and company_size predictors as the most statistically significant for predicting salary. A histogram, QQ plot, and a residuals plot were used to verify the model's fit.




Sample Quantiles
1.5
0.5
−0.5
−1.5
−2
0
2
Theoretical Quantiles


Residuals vs Fitted
Residuals
2
1
0
−1
−2
1541
1510
6005
10.0
10.5
11.0
11.5
12.0
Fitted values

5.1.2 Linear Regression Model Prediction Performance

In order to evaluate the performance of this linear regression model, we are going to be using Root Mean Squared Error (RMSE). If we were to use this model to predict salary, we would have to "un-log" it by exponentiating the value. An RMSE of 0.36 suggests that on the dollar scale (0.36 -> 1.43), the prediction error is around a 43% error. The proportion of variance explained by the data was around 47%.

Table 2: Linear Regression Prediction Performance

Model	RMSE	RSquared
Linear Regression	0.36	0.47

5.2 Random Forest Model

In the processing of optimizing the number of trees used in our random forest model, and the number of predictors considered in the split at each node, we decided to set our random forest model 200 trees and 6 predictors tried at each split. The initial number of trees used was set to 100, but as we see below in our plot displaying error rate compared to number of trees used in the model, anything below 150 trees leads to higher levels or error in our model.

5.2.1 Random Forest Fit Results

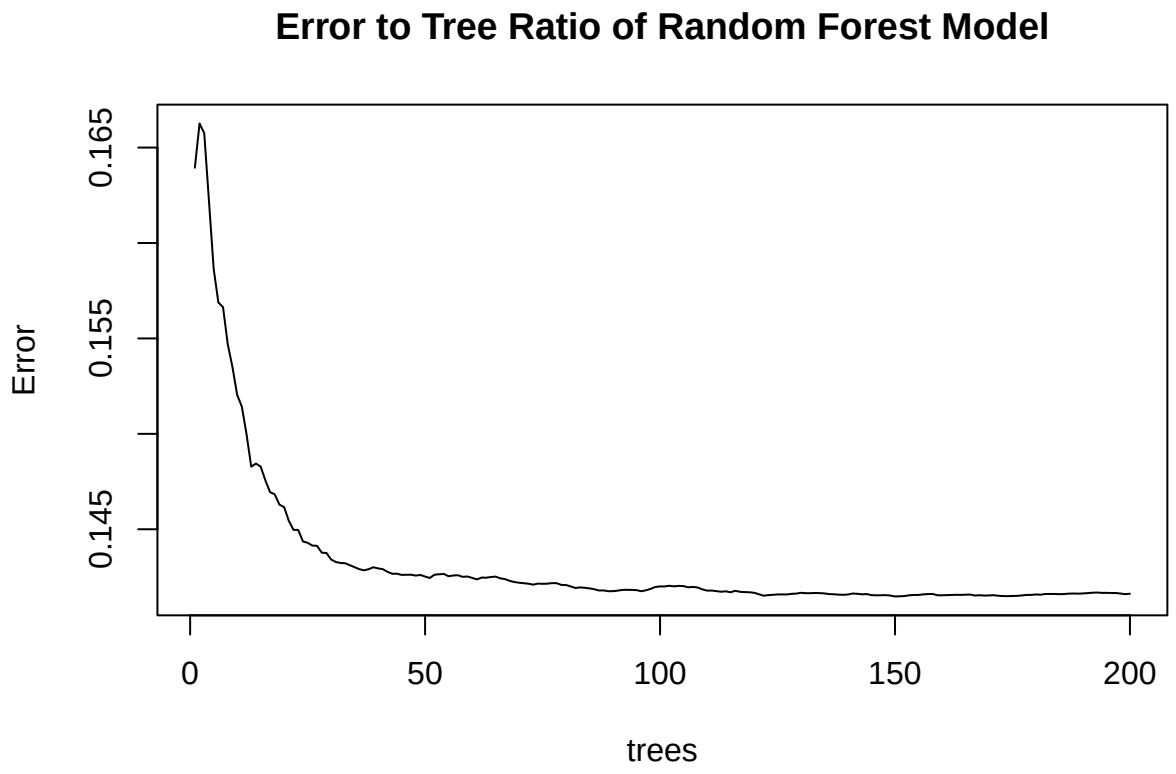


Figure 2: Random Forest Model Error

We see that as the number of trees start to increase, the out of bag error rate starts to progressively drop lower, until finally stabilizing from the number of trees equaling 150 onward. It would be a point of diminishing returns to keep our tree at 500, so we reduced the number of trees to a threshold in which the error rate does not change while leveraging other computational advantages to the lesser amount of trees (simpler model, more computationally efficient, less time to make predictions on new data).The proportion of variance explained by the model (similar to the R-squared value in our linear regression model -which was 47%-) is 45%.

5.2.2 Model 2 Prediction Performance

Table 3: Random Forest Prediction Performance

Model	RMSE
Random Forest	0.37

	%IncMSE	IncNodePurity
job_title	95.722887	173.119279
experience_level	101.407730	120.546541
employment_type	-2.685721	3.989125
work_models	14.586169	16.113327
work_year	20.559364	29.654356
employee_residence	32.090090	240.425244
company_location	18.239864	93.341672
company_size	15.352507	16.344549

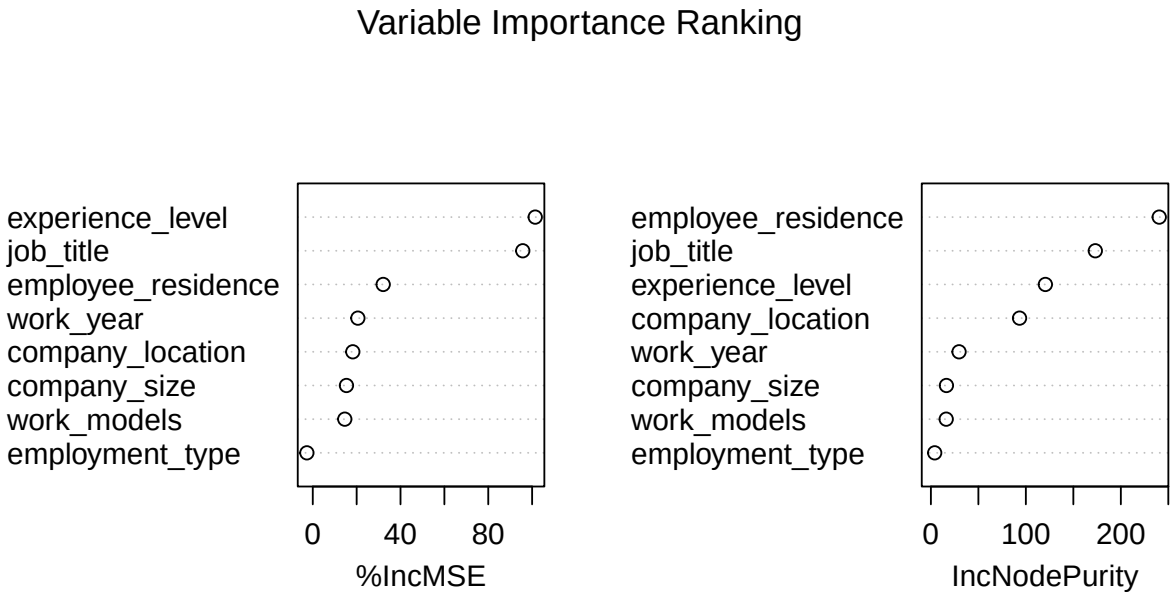


Figure 3: Variable Importance - Random Forest Model

The ranking displayed by the results of our model show that the most important variables based on our metrics of IncNodePurity (higher value depicts greater importance for a certain variable in our model) and %IncMSE (higher value depicts greater importance for a certain variable in our model), that affect the predicted salary are: job title, experience level, and employee residence. From greatest to least importance for node purity we have job_title (157.35982), experience level (158.78542), and employee residence (240.189459). For our %IncMSE we have: job title (100.31562), experience level (99.77138), and employee residence (30.23758). Both metrics labeled the same set of variables as most important (job title, experience level, and employee residence), but in different order.

5.3 Results Summary

Table 5: Model RMSE for both models

Model	RMSE
rmse.rf	0.3700459
linregrmse	0.3621927

This table shows the RMSE (root mean squared error) values for the random forest model and the linear regression model. RMSE is a useful metric to compare error rates between models. The RMSE for the random forest model is manually calculated by testing the predictions the model makes on unseen data, and subtracting the error rate by one. The RMSE for the linear regression model is also calculated by using the same method. The typical error rate metric for the random forest is the out of bag error (rate of error from model predictions on unseen data).

Describe how each model performed. Use tables and plots. For regression, summarize in a table and/or plot. For classifiers, summarize confusion matrix results in a table, combine both models in an ROC curve.

() - INSERT DESCRIPTION HERE

6 Conclusions

When comparing the percent of variance explained by the models on the test set, we see that the random forest model is able to explain at a higher rate than the linear regression model. But only slightly. The proportion of variance explained by the models were both fairly low, (45% and 47%) thus leading us to conclude that both models failed to accurately capture much of the trend in the data. However, these models still provided valuable insights. Our initial goal was to discover what the most important factors affected data science job salaries and these models assisted in discovering them. The random forest model showed that job title, experience level, and the location of the residence of the employee was the most important predictors of yearly salary. And the linear regression model agreed.

The fact that variables like employment type and company size were not as influential surprised us. These models suggest that while company-related factors may play some role, personal qualifications and career-specific factors may carry more weight in determining salary. If we were to continue this analysis we would likely try out different modeling techniques or perhaps use a different dataset entirely. One with more expansive data or a larger time-range might be a more effective dataset to use.

7 References

This section contains a numbered list of any data sets or data sources you used, and any other references you accessed.

[1] https://www.kaggle.com/datasets/sazidthe1/data-science-salaries?resource=download