Which programming language should I learn to secure a promising data science position?

Chaoran Yang

The University of California at Los Angeles

Which programming language should I learn to secure a promising data science position?

Abstract

This paper investigates the optimal programming language for aspiring data scientists seeking lucrative job opportunities by analyzing data sourced from the field of data science employment. It examines the programming languages frequently requested in data science job postings, the frequency of new job listings, and salary trends associated with proficiency in R programming. Additionally, it compares Python and R, two commonly taught languages in data science courses, exploring their correlation. Furthermore, a linear regression model is constructed to elucidate the impact of various programming languages on job salaries. Based on these findings, a strategic approach to programming language acquisition is proposed, tailored for individuals with limited time resources.

Key words: Data science, programming languages, job opportunities, comparative analysis, salary trends, linear regression model

Content

1 Topic Analysis

Whether you are a student of data science or an academic or worker in the field of data science, finding the ideal data science job may be one of the goals you are striving for. Here, we loosely define "ideal" jobs as those that pay well and support a high quality of life. To achieve this goal, we need to know what these jobs require of us and what skills are needed. Therefore, I collected data from a job listings website (https://www.seek.com.au/) of data science-related job listings and analyzed the most common words in data science jobs

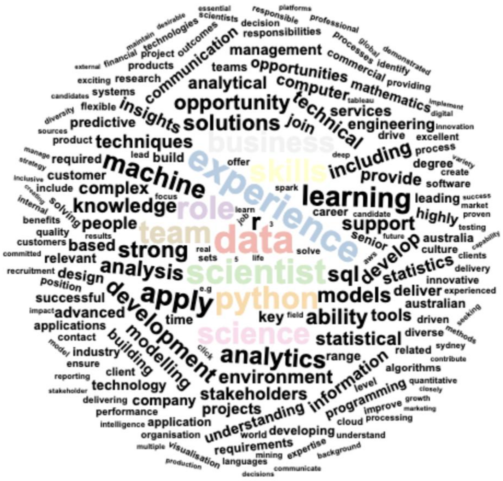


Figure 1. Most Common Words in Data Science Job

According to the word cloud map, proficiency in one or more programming languages is essential for data science jobs. The ability to process and analyze data through programming skills is an indispensable factor. However, with the plethora of programming languages available in the current era, it is challenging to choose which language to learn. To address this issue, this paper aims to analyze job listing data, identify the most commonly used programming languages in the data science field, and explore the correlation between these languages and the programming skills required to achieve an objective salary. The significance of this research lies in the guidance it can provide to students who wish to learn programming languages, as well as employers who aim to recruit and train employees in a targeted manner, which can ultimately lead to cost savings.

2 Data Preparation

The dataset "listings2019_2022.csv" has details about job vacancies related to data science in Australia from 2019 to 2022. Each job listing has 52 variables, and the dataset has a total of 3,902 observations. Table 1 shows some of the critical variables and their explanations.

Varible name	Varible Meaning
jobTitle	Name of the job
jobClassification	The job's specific area
companyRating	Evaluation of the company (0-5)
listingDate	When the job listing was posted
salary_string	Wage status (character type)
teaser	Company Info & Hiring Needs
workType	Full time work or Temp work
Python (Other programming languages)	Is knowledge of this programming language required for this job? (0=not required, 1=required)

Table 1. Explanation of the meaning of important variables

3 Data Cleaning

As we have obtained the data using a crawler, the dataset may contain formatting issues and missing values. Therefore, we need to preprocess the data before analyzing it.

3.1 Converting dataset formats

The collected data needs to be converted into an easily exportable format, so we need to do some processing. In addition, there is a very important point that the "salary_string" in the data is a character format, which not only contains the salary but also some descriptive text, so it is also necessary to extract the number of the data and convert it to a character type.

3.2 Dealing with jobDescription

The job description is a crucial element in a job listing, and we need to ensure we use it effectively. There are two fields that provide a job description: desktopAdTemplate and mobileAdTemplate. But which one should we use?

Upon a quick visual inspection, it appears that mobileAdTemplate has some lower quality data - some words are joined together without a linebreak, and it's unclear why. Therefore, it seems sensible to use desktopAdTemplate instead.

However, after reviewing the data, it was observed that there are 695 entries for desktopAdTemplate that do not have any content. This was determined by counting the job description text, as shown in Figure 2.

To address this issue, we have createda new variable called jobDescription. This variable has a default value of desktopAdTemplate, unless it's of length zero, in which case it takes the value of mobileAdTemplate. (See Final.R for code)

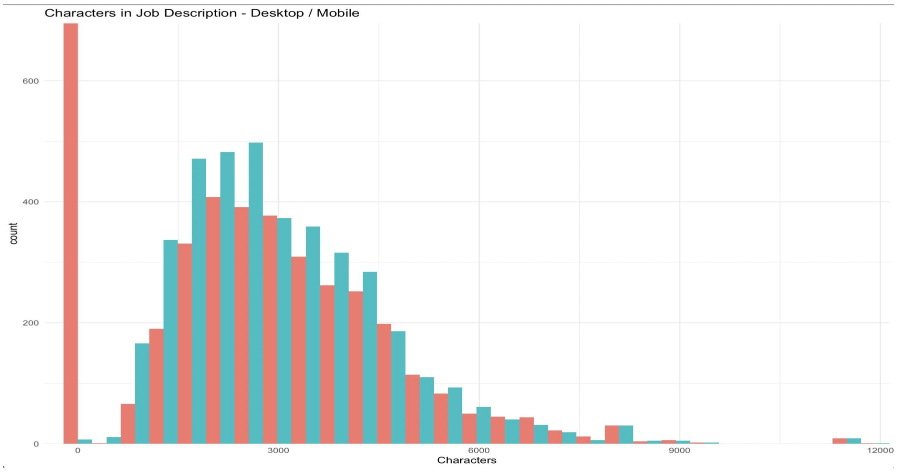


Figure 2. Characters in Job Description - Desktop / Mobile

4 Visualization and EDA

4.1 Most Used Programming Languages

To demonstrate the frequency of different programming languages used in data science jobs, we utilized R to tally the programming languages present in the dataset. After that, we generated a chart that displays the programming languages utilized in data scientist job listings, along with their corresponding number of job openings from March 2019 to January 2022.

Figure 3 illustrates that the most popular open source tools used in data science are Python, R and SQL, in that order. Additionally, Tableau, SAS, and Matlab have seen a significant rise in usage. Hadoop, Scala, and Spark are also prevalent among data scientists. While some academic tools like Stata, Matlab, and SPSS are not as popular as open source tools, they still have a considerable user base. Interestingly, Java is also high on the list, which could indicate that data scientists are increasingly working with software teams, or that more data science work is being integrated into software applications instead of being confined to interactive analysis or reports.

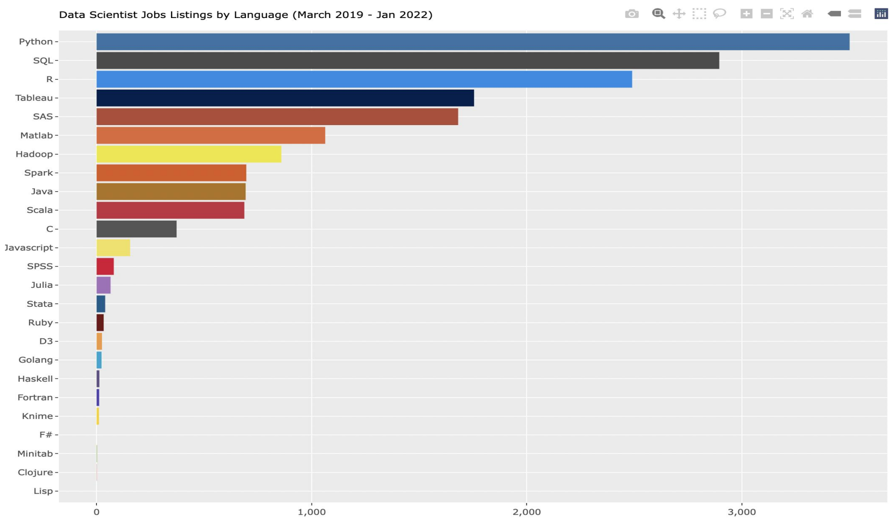


Figure 3. Data Scientist Jobs Listings by Language (March 2019 - Jan 2022)

4.2 What's the best time to find a job?

To increase the chances of securing a data science job, it is important to not only master common programming languages but also to understand the posting patterns of job listings. The chances of finding a job are higher during a period of high hiring demand. Figure 4 demonstrates the weekly variation in new job listings at a macro level.

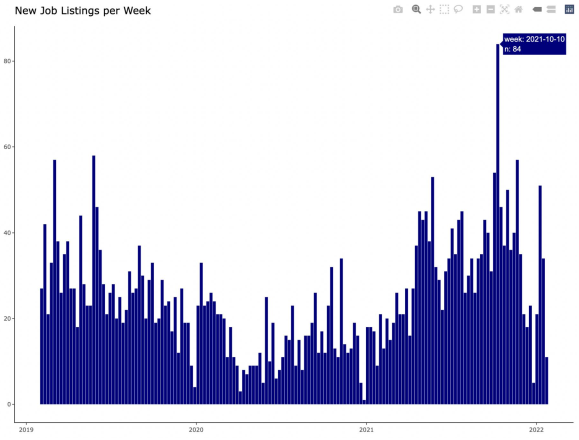


Figure 4. Data Scientist Jobs Listings by Language (March 2019 - Jan 2022)

In 2019, there was a busy job market in Australia, but from April 2020, the number of new job listings decreased drastically. This was due to the COVID-19 outbreak, which forced most people to work from home. In fact, during the week starting from April 5, 2020, there were only three new job listings for Data Scientists in the entire country! The market remained quiet throughout 2020, but in mid-2021, it came back with a bang and has continued to thrive since then.

Next, the posting time of the job listings is further analyzed. By using the release time of job listings in the data, we can plot the hot zones of job listings released in different time periods every day, as shown in Figure 5 (where the darker the color, the more job listings are released in this time period)

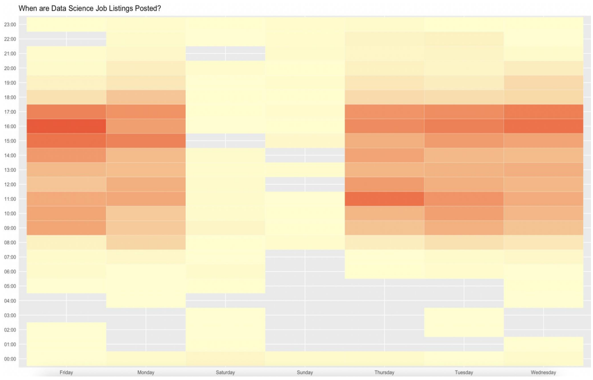


Figure 5. job listings posting hotspot map

It seems that friday afternoon is very popular - I wonder if that's simply because of deadlines or if it's a strategic decision to maximise weekend exposure.I suppose it could be a bit of both. People are likely rushing to finish up work before the weekend, but also posting on a Friday afternoon could mean more eyes on your content as people are starting to wind down and scroll through social media. It's definitely an interesting phenomenon to think about.

5 Model and result

5.1 Data Scientist Salary Analysis

After data cleansing, we now know the salaries for 1028 out of 3902 jobs and have calculated that the salaries for Data Scientists in Australia range from AU$58,000 to AU$360,000, with an average salary of AU$136,093 and a median salary of AU$125,416. Before analyzing the programming language in conjunction with the salary, let's analyze the trend of the average salary of a Data Scientist for the time period from May 2019 to January 2022.

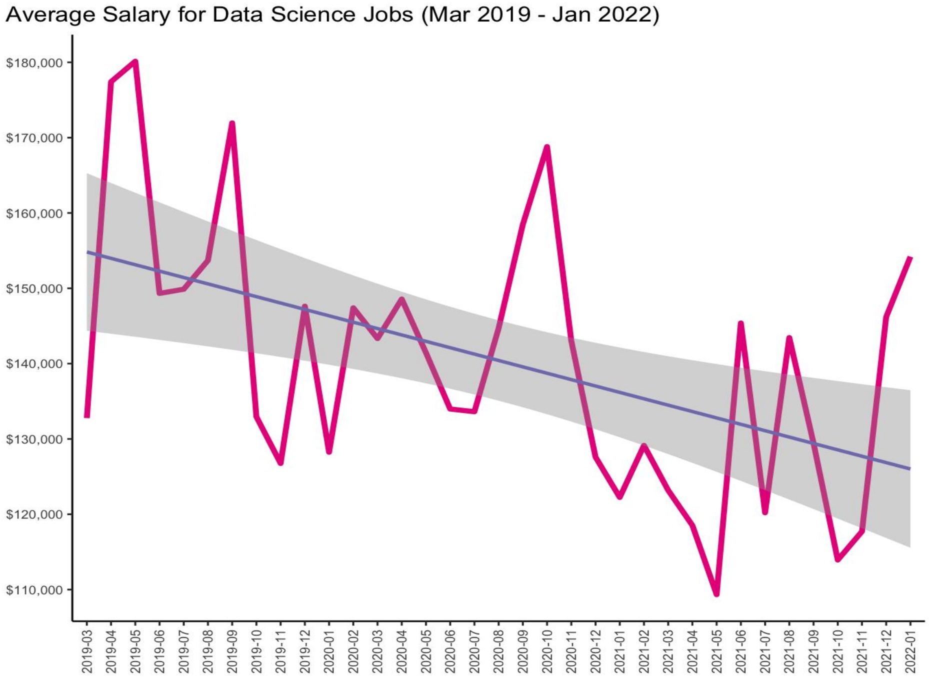


Figure 6. Average Salary for Data Science Jobs (Mar 2019 - Jan 2022)

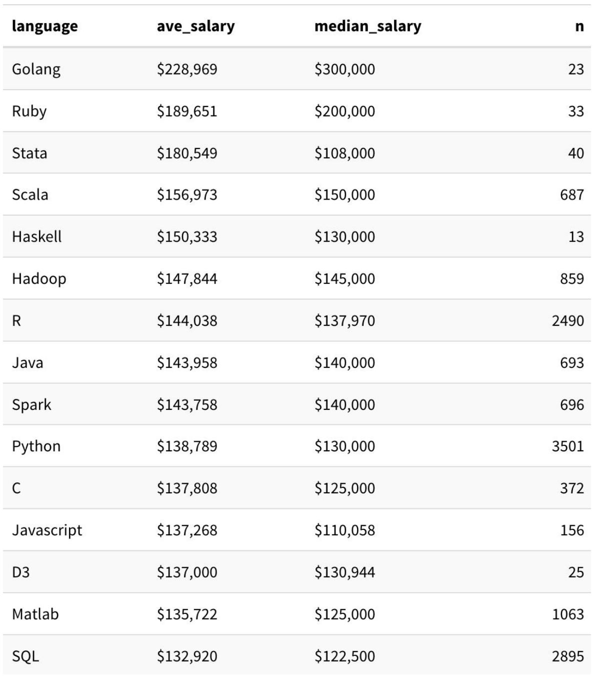
We can observe a decline in the average salary of data scientists, and there could be several reasons for this trend. One reason could be the maturation of the field - as companies become more proficient in understanding their specific data-related needs, it has led to more targeted hiring practices. This has resulted in lower salaries for certain positions based on the perceived value of the role within the organization.

Another factor could be economic downturns or periods of uncertainty, such as those experienced during the COVID-19 pandemic, which can impact overall hiring and salary trends across various industries, including data science.

Furthermore, as the field becomes more saturated with entry-level data science talent, companies may adjust their hiring strategies to focus more on mid-level or senior-level positions, which typically command higher salaries. This shift in hiring priorities can also contribute to a decrease in the average salary for data scientists.

5.2 Salary analysis based on programming languages

Overall, the overall salary level of data science jobs is a downward trend, so what is the salary difference between the different positions, we will analyze the following according to the need to master different programming languages



language	ave_salary	median_salary	n
Golang	$228,969	$300,000	23
Ruby	$189,651	$200,000	33
Stata	$180,549	$108,000	40
Scala	$156,973	$150,000	687
Haskell	$150,333	$130,000	13
Hadoop	$147,844	$145,000	859
R	$144,038	$137,970	2490
Java	$143,958	$140,000	693
Spark	$143,758	$140,000	696
Python	$138,789	$130,000	3501
C	$137,808	$125,000	372
Javascript	$137,268	$110,058	156
D3	$137,000	$130,944	25
Matlab	$135,722	$125,000	1063
SQL	$132,920	$122,500	2895

Table 2. Salary table based on programming languages

The results presented in Table 2 indicate that having expertise in mainstream languages commonly used in blockchain (Golang) and big data processing (Scala, Hadoop and Spark) can result in higher average wages. However, it should be noted that the sample size for languages outside Scala and Hadoop is limited, possibly due to the difficulty of programming or the limited applicability of these languages. On the other hand, while programming languages like Python, R, and SQL are widely used in data science, they may not command high salaries, but there is generally more demand for related jobs.

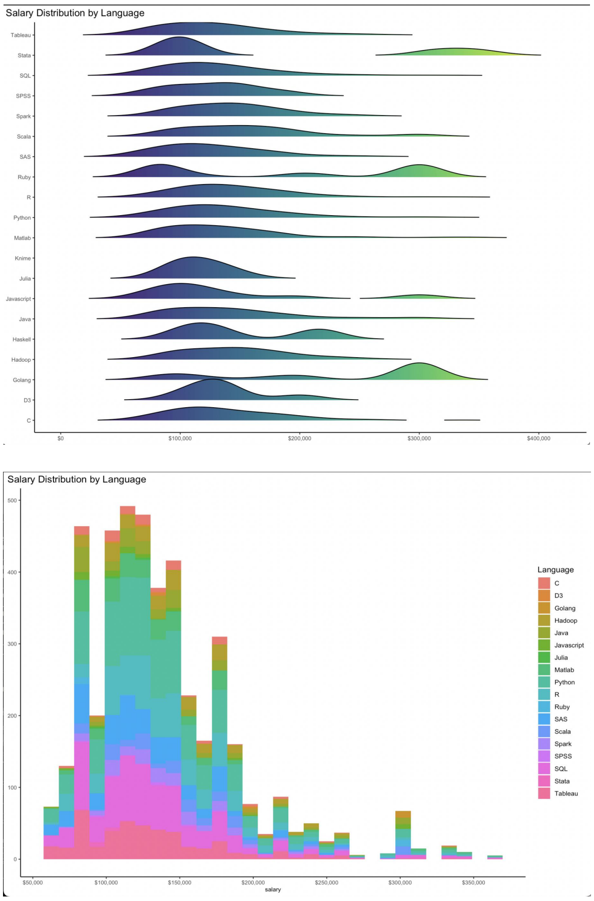


Figure 7-8. Average Salary for Data Science Jobs (Mar 2019 - Jan 2022)

Figure 7 and Figure 8 further explain the distribution of salaries by language, showing that the majority of data science jobs have a salary distribution within the $100,000 to $200,000 range. The three programming languages that clearly have a high salary distribution are also in the top three highest average salary levels. This may indicate that mastering these programming languages may be able to bring you more wealth. But again, due to the limited sample size, the results obtained from the analysis need to be interpreted more cautiously

5.3 Correlation analysis

In this section, we analyze the correlation of some key variables, includin g the correlation between programming languages and job titles and the correlation between individual programming languages.

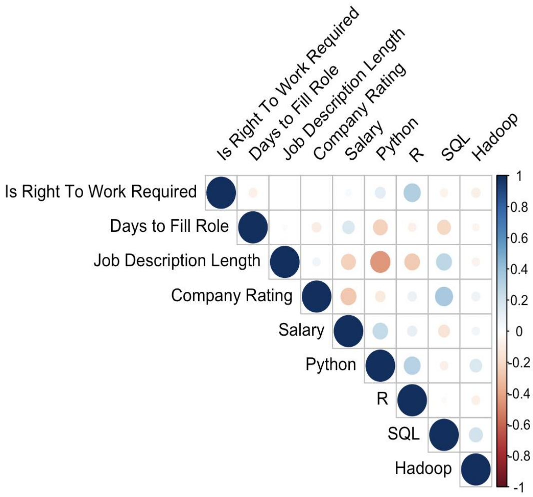


Figure 9. Correlation Matrix of Key Variables

With the above chart, we are able to draw the following conclusions:

- There are positive correlations between salaries and jobs that require proficiency in both Python and R programming languages.
- Companies that pay lower salaries usually receive lower ratings.
- Python jobs tend to fill quickly
- R jobs are positively correlated with Company Rating, but are negative correlated with salary, meaning R jobs tend to be slightly lower paid, but at good companies.
- SQL jobs tend to be at great companies.
- Higher paying roles tend to take longer to fill, possibly because they’re more senior (and hence more important) roles.

Further, we can analyze the correlation between programming languages in terms of both absolute and relative values by exploring whether two languages appear in the same job listings. The relative value co-occurrence graph (Fig.10) indicates that for the language on the vertical axis, this indicates the proportion of job postings that also specify the language on the horizontal axis. For instance, out of the positions that demand Java proficiency, 95.4% of them also necessitate Python skills.

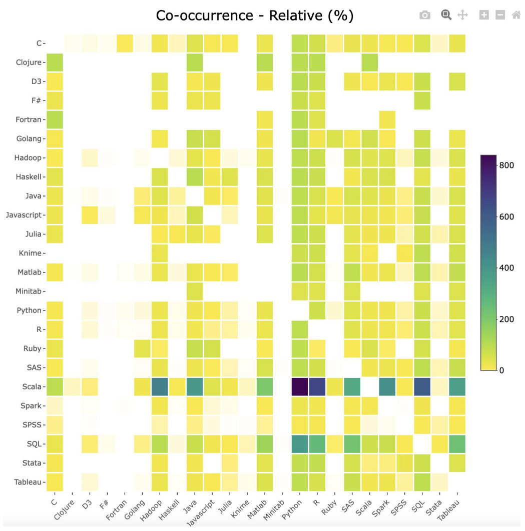


Figure 10. Relative co-occurrence graph based on languages

While the absolute value co-occurence graph (Fig.11) shows that for each language on the left-hand side, this chart shows the number of job listings that mentioned each language along the bottom axis. For instance, 1059 Python job listings also mentioned MATLAB.

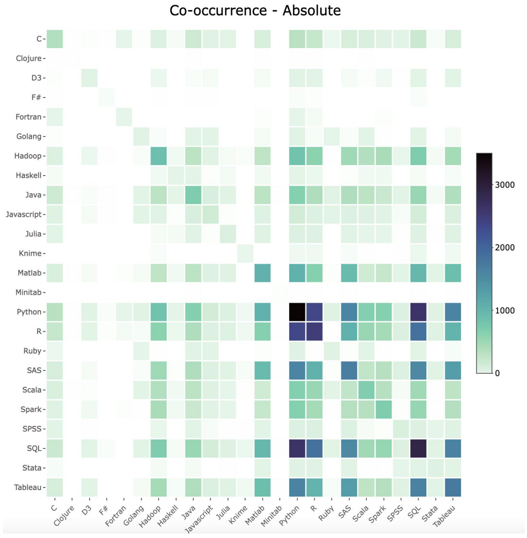


Figure 11. Relative co-occurrence graph based on languages

5.4 Comparison: Python and R

Since R and python are the programming languages we're learning in this two-week course, and both languages also have a relatively larger number of learners, I was curious about the salary level distribution of the two.

When comparing salaries, it is evident that the distributions are quite similar. However, there are more Python jobs available in the mid-range, and a scattering of more jobs in the mid-to-high salary range. And in general, there is a relatively even distribution of companies that hire candidates skilled in both Python and R programming languages.

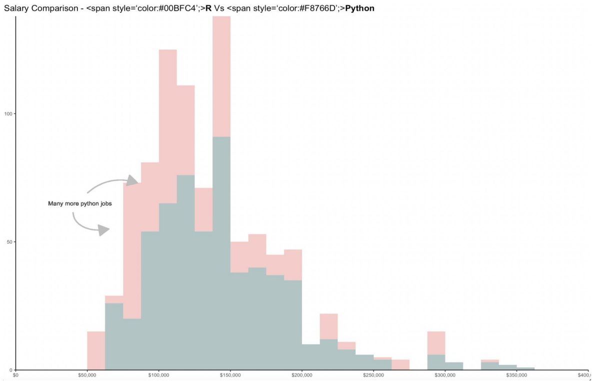


Figure 12. Salary Comparison between Python and R

5.5 Regression analysis

Although the data collected was not suitable for a linear regression analysis (programming languages are dummy variables), I made a rough attempt: obviously, salary level is the target variable, and since the question to be explored is which programming language I should learn in the future, all the programming language variables are the independent variables of the model. The results obtained from the regression are shown in Table 3:

	Dependent variable:
	salary
R	17,235,540,560,984,186.000***
	t = 3.498
Python	-7,562,228,761,899,511.000
	t = -0.992
SQL	-3,447,551,010,275,344.000
	t = -0.602
Java	40,869,433,799,341,184.000***
	t = 6.186
C	-11,896,284,830,905,444.000
	t = -1.485
Scala	-36,210,711,039,371,792.000***
	t = -5.138
Tableau	6,904,260,170,175,498.000
	t = 1.166
Hadoop	26,895,291,360,857,500.000***
	t = 4.462
SAS	3,107,963,807,625,850.000
	t = 0.535
Julia	-27,181,859,731,623,364.000
	t = -1.069
Fortran	
Constant	-5,474,363,105,314,308.000
	t = -0.704
Observations	1,131
R2	0.069
Adjusted R2	0.061
Residual Std. Error	74,735,168,196,258,704.000 (df = 1120)
F Statistic	8.317*** (df = 10; 1120)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 3. Result of the regression model

Based on the results of the model, we can draw the following results:

- The value of the F-statistic is 8.317, corresponding to a p-value of 3.994e-13, indicating that the model is significant as a whole and that the effect of programming language on wages is significant.
- The model is poorly fitted with a low multiple R-square, indicating that the model explains only a small portion (6%) of the variation in wages.
- Some of the independent variables in the model have significant coefficients, such as R, Java, Scala and Hadoop, which may have a greater impact on wages.

It is not difficult to find that the R-square of the model is not high and the coefficients before variables such as python, for example, are negative. This is very strange (because the conclusion that mastering a programming language decreases income instead is just not common sense). The reasons for this situation are as follows:

- Collinearity: due to the strong correlation between some programming languages. Resulting in covariance in the model, the coefficient estimation may be inaccurate or even contrary to the actual situation.
- Omitted Variable: There may be omitted important variables in the model, such as job_description or jobSubClassification. These variables may have influenced wages along with the programming language, but the fact that they were not taken into account leads to a bias in the programming language coefficients (this is probably the most significant cause)

In summary, we cannot simply explain the relationship between programming language and salary using linear regression, but perhaps some variables with positive coefficients and large absolute values (R, Java, and Hadoop) can have a more significant impact on salary

6 Conclusions

The takeaways of the report are as follows:

- The most commonly used open source tools in data science are Python, R, and SQL, so it's well worth learning them!

- Jobs that utilize Python, R, and SQL may not command high salaries on average, but there is generally more demand for related jobs.

- Learning a programming language that is more specialized, difficult or in the early stages of development in the field of data science (Golang, Ruby, Scala) can lead to a better paying job.

- The salary distribution of jobs in python and R is similar to the distribution of company ratings, and mastering both languages has a positive effect on salary levels

- Most jobs require python skills, even if the dominant language in the field is not python, so learn python!

- Even though the average salary for jobs in the data field is on a downward trend due to factors such as COVID-19, it is gradually recovering since October 2021, so there is no need to be overly concerned, as this is still a promising field!