Data Scientists' Wages

André Filipe Silva - 20230972, João Maria Gonçalves - 20230560, Luís Queiroz - 20230584

NOVA IMS Information Management School

Introduction

As we get ready to enter the job market in the Data Science area, one of the things that is undoubtedly on our minds is our future income. Income in this field can be influenced by a complex interplay of variables, including your education, skills, experience, location… The goal of this project is to check what factors influence data scientists' wages according to our dataset. Our variables are if you code as a hobby (*Hobby*), where in the world you work in - United States, United Kingdom, Germany (*Country*), highest level of educational attainment (*Education*), size of the company (*OrgSize*), the field of your undergraduate degree (*Undergrad*) and the number of years you have been coding as a professional (*YearsCodePro*). Our cross-sectional data was gathered from processed data taken out of a survey from the Stack Overflow Website in 2020 [1].

Research Question

"What factors influence data scientists' wages according to our dataset?"

Methodology

To answer our research question, first we extracted data from Kaggle [2]. We then preprocessed the data using the help of the Python programming language. To give an example, the variable organization size was a numeric variable originally, but we decided to categorize it according to the EU recommendation 2003/361 [3]. It states that a small company is defined as having fewer than 50 employees, a medium-sized company has fewer than 250 employees, and a large company would presumably have 250 or more employees. All of the rest of this project was conducted using the R coding language. We performed a multiple linear regression model, having verified that the classical OLS assumptions hold. At the end we checked if our model is correctly specified using the RESET test.

Results

Analysis

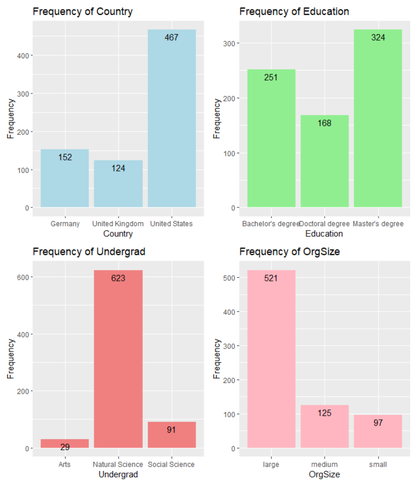

Figure 1: Bar plots for each categorical variable

Exploring our data, we can see that the majority of people work in the United States, have a Master's degree, have Natural Science as their Undergrad background, and belong to large organizations.

Model

Coefficient	Estimate	Std_Error	t_value	p_value
(Intercept)	10.6487	0.1247	85.3913	0.0000 ***
HobbyYes	-0.0555	0.0531	-1.0450	0.2964
CountryUnited Kingdom	0.1247	0.0673	1.8531	0.0643 .
CountryUnited States	0.5913	0.0536	11.0298	0.0000 ***
EducationDoctoral degree	0.2173	0.0562	3.8673	0.0001 ***
EducationMaster's degree	0.1230	0.0476	2.5848	0.0099 **
OrgSizemedium	-0.0619	0.0548	-1.1285	0.2595
OrgSizesmall	-0.2364	0.0607	-3.8946	0.0001 ***
UndergradNatural Science	0.1460	0.1053	1.3864	0.1660
UndergradSocial Science	0.2864	0.1173	2.4414	0.0149 *
YearsCodePro	0.0247	0.0028	8.8744	0.0000 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Figure 2: Model Table

From the model summary we can see that Hobby is not significant, CountryUnitedKingdom is only significant at 10% significance level. Orgsizemedium and UndergradNaturalScience are also not significant but also dummy variables. We do not remove any of these variables from our regression as they are dummy variables and the other categories are relevant towards our analysis.

R-squared: 0.2957, Adjusted R-squared: 0.2861 F-statistic: 30.73 on 10 and 732 DF, p-value: < 2.2e-16

We can also see the F-statistic for global significance of the regression, which shows us that the regression is globally significant at a 5% significance level.

The Adjusted R^2 sits at 0.2861 which indicates that 28.61% of the variation in ln(wages) can be explained by our current model.

Tests to check for OLS assumptions and Model specification

By performing tests and graphical analysis our OLS model has successfully met all the underlying assumptions.

Test	H0	P-value	Conclusion
Breusch-Pagan	Homoskedasticity	0.777	Homoskedasticity
Durbin Watson	No Autocorrelation	0.132	No Autocorrelation
RESET	Model is correctly specified	1.91E-05	Model is misspecified

Figure 3: Tests table

The p-value of the RESET test leads us towards concluding that our model is misspecified. However, we only have one variable in our dataset that is not a dummy variable. We tried to include exponentials of the numerical variable to see if it improves our specification but to no avail. Our model still shows as misspecified.

Interpretation

(5% significance level)

Note: We used the 100*(e^(B -1)) transformation to interpret the coefficients.

Having a job in the United States, increases the salary by 80.63% on average in relation to being in Germany, ceteris paribus.

Having a Doctoral degree increases the salary by 24.27% on average in relation to only having a Bachelor's Degree, ceteris paribus.

Having a Master's Degree increases the salary by 13.1% on average in relation to only having a Bachelor's Degree, ceteris paribus.

Working at a small sized organization (<19 employees) decreases the salary by 23.68% on average compared to working in a large organization, ceteris paribus.

Having an Undergrad in Social Sciences increases your salary by 33.16% on average compared to having an Undergrad in Arts, ceteris paribus.

For each year of work as a coding professional, salary increases by 2.53% on average, ceteris paribus.

All of our significant coefficients have the expected sign. For instance, it is expected that having higher education increases your wage, and that each year of experience will also increase your salary.

Next Steps

We would like to have a few more variables to help explain the wages of Data Scientists. Gender, Position (Junior, Mid, Senior, Executive), Freelancers. We could probably get a higher predictive power with said variables. These are some of the variables that the literature on the subject of wages in general reflects about. Furthermore, it would be interesting to turn this into a time-series study, to observe the trends over time for salaries and try to design causal experiments based on cut-off events (e.g. COVID-19 pandemic). An analysis of the trends and buzzwords in the media around this area could also prove important to check for some correlation with the interest of companies in the area.

Conclusion

This work has some clear recommendations to keep in mind. For instance, the wage increase in the USA is very big (since it is the country in the world with most tech companies, and probably bigger demand for Data Scientists, it makes sense that the salaries are the bigger there). If one is looking for a higher salary, the USA is likely a place to consider working in.

It also seems to pay off to have a Master's Degree and even a Doctorate Degree. This is quite remarkable in terms of the PhD, since there are not too many people who attain a PhD in Data Science. Further research into this finding is required.

We can also see that working for small organizations does not pay off. The 'why' of it would also be interesting to further investigate on.

Finally, having an Undergraduate degree in Social Sciences seems to be significant towards improving Data Scientist salaries. This is a surprising result, seeing as we would guess that Computer Science Majors would take the highest-paid spots. We can perhaps attribute this salary increase to the bigger domain knowledge that learners of Social Sciences acquire, but this would require further study.

References

[1] (n.d.). Retrieved from Stack Overflow: https://insights.stackoverflow.com/survey

[2] EUR-Lex. (n.d.). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32003H0361

[3] PhucNH. (n.d.). Salary and more-Data Scientist, Analyst, Engineer. Retrieved from Kaggle: https://www.kaggle.com/datasets/phuchuynguyen/salary-and-moredata-scientist-analyst-engineer/

King, J., & Magoulas, R. (2015). 2015 data science salary survey. O'Reilly Media, Incorporated.