Datenanalyse in Physik und Astronomie

StackOverflow Developer Survey 2018
- Analysis -

Lars Gröber
October 7, 2019

Abstract

This analysis will look at the StackOverflow Developer Survey 2018 [3], present the data and will answer the following questions:

- Is there a difference in salary regarding which country a developer resides in?
- What is the correlation between a developers experience, education and their paycheck?
- What factors are relevant for the expected salary of a software developer?

Contents

1 Introduction

The StackOverflow Developer Survey 2018 (from now on only called *survey*) asked hundred of thousands of developers, which use StackOverflow on a regular basis, to answer 129 questions about their employment status, salary, thoughts on different tech-related questions and more.

98855 respondents answered their questions in 2018. StackOverflow decided to make the results of this survey public [3].

This report is accompanied by a Jupyter Notebook which is published on GitHub:
`https://github.com/Larsg7/dataScience/blob/master/analysis.ipynb`
All figures in this report are taken from there.

2 Simple Analysis

This chapter will present the survey answers in a visual form including a short analysis of the data.

2.1 Employment and Hobby

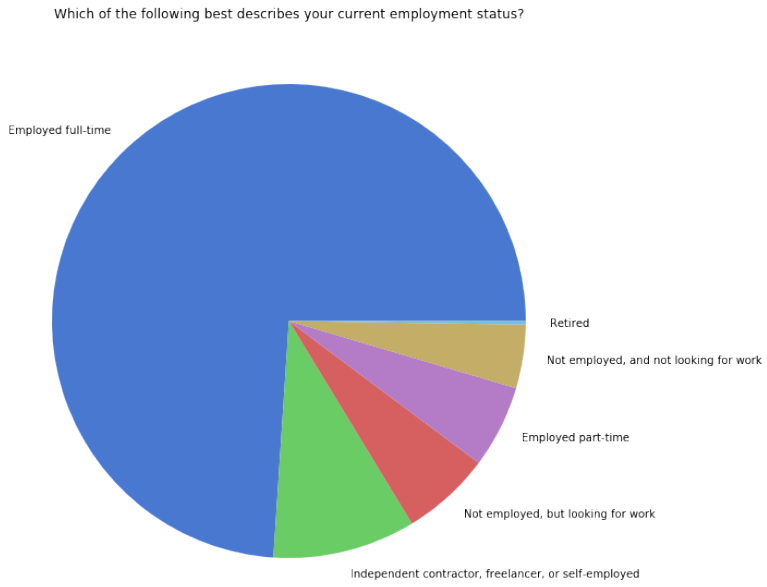

Figure 1: Employment status of all respondents

The question stated: "Which of the following best describes your current employment status?". About 70% of respondents are working full-time (we assume here full-time as some sort of software developer). Only a small number of respondents are unemployed. The question has been answered by around 96% of respondents.

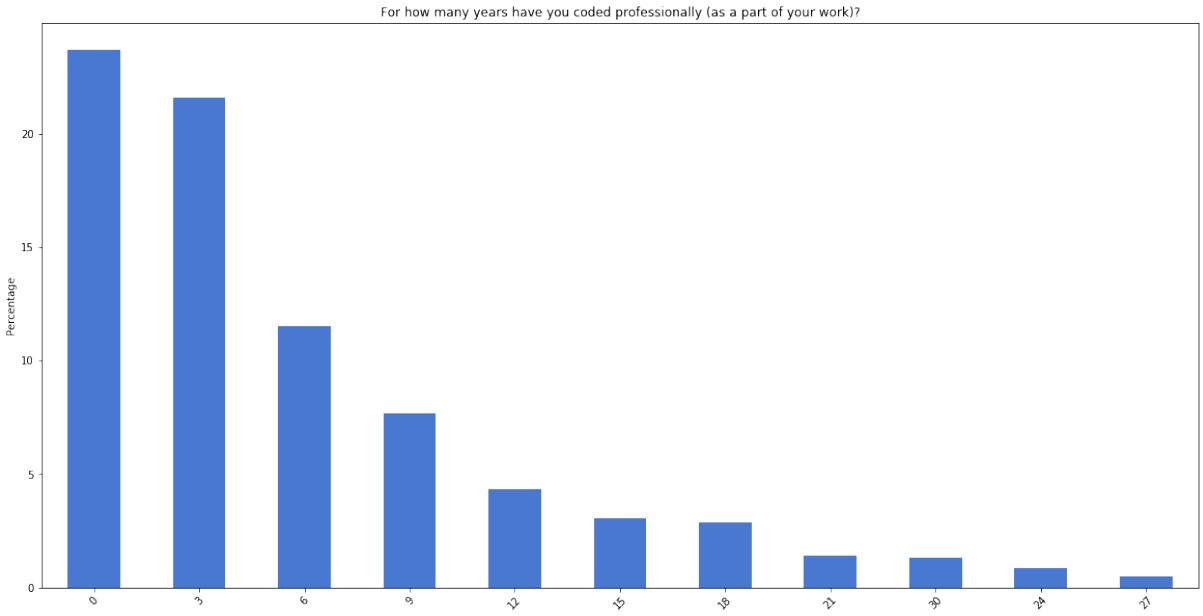

Figure 2: Years the respondents were working professionally for

For figure 2 the question stated: "For how many years have you coded professionally (as a part of your work)?". This of course does not exclude people who only work part-time. The data shows that most people who answered the survey are quite junior. Almost 25% of respondents have only 0-3 years of experience. StackOverflow is of course a site where people can ask questions. So there might be a higher percentage of people with little experience visiting the site.

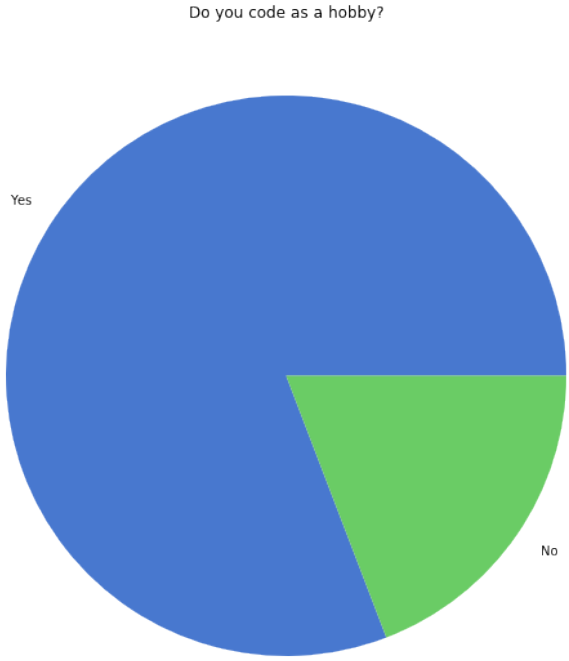

Figure 3: This graph shows whether the respondents were coding as a hobby

In figure 3 the respondents were asked if they "code as a hobby". Many software developers

also code in their free time, for example on smaller side projects or open-source code bases. The data shows that about 80% of respondents also code as their hobby.

In figure 4 we compare the number of respondents who "code as a hobby" for different employment statuses. People who work full-time are not unsurprisingly less likely to also "code as a hobby" in their free time. On the other hand people who are retired or unemployed, but not actively looking for a job see coding as their hobby more often.

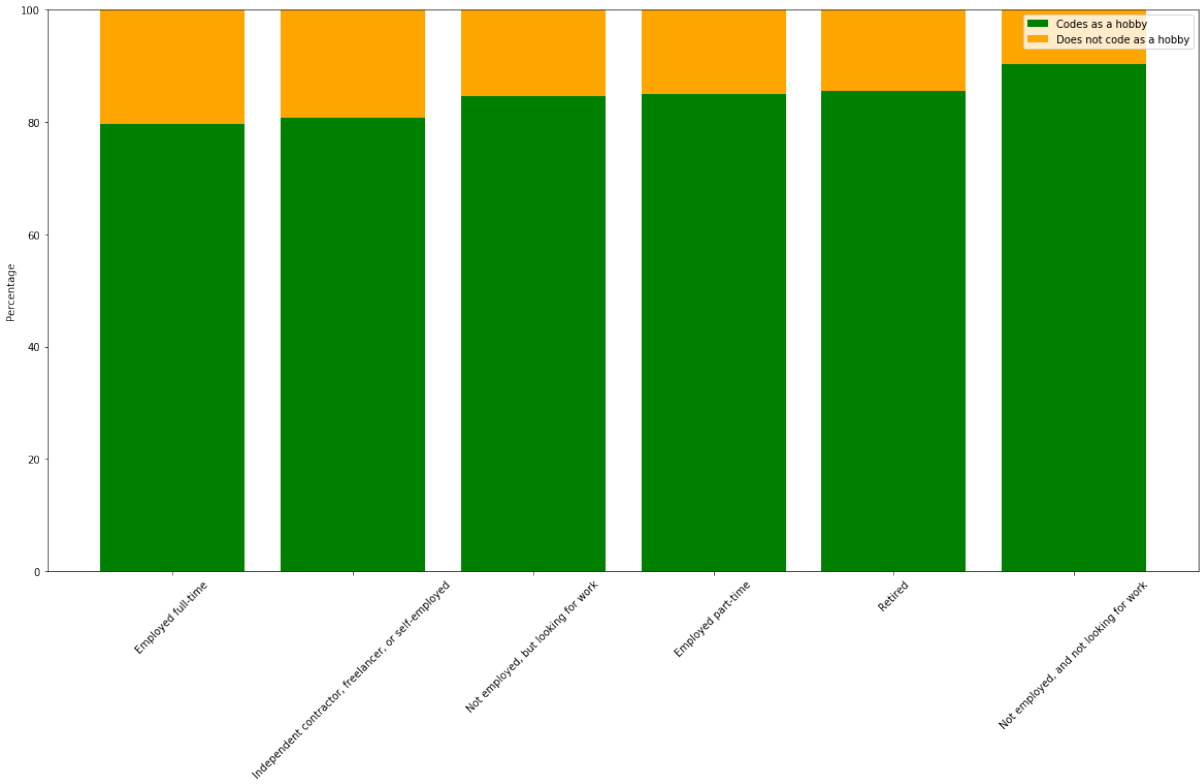

Figure 4: This graph shows the level of "coding as a hobby" for the different employment statuses.

2.2 Gender

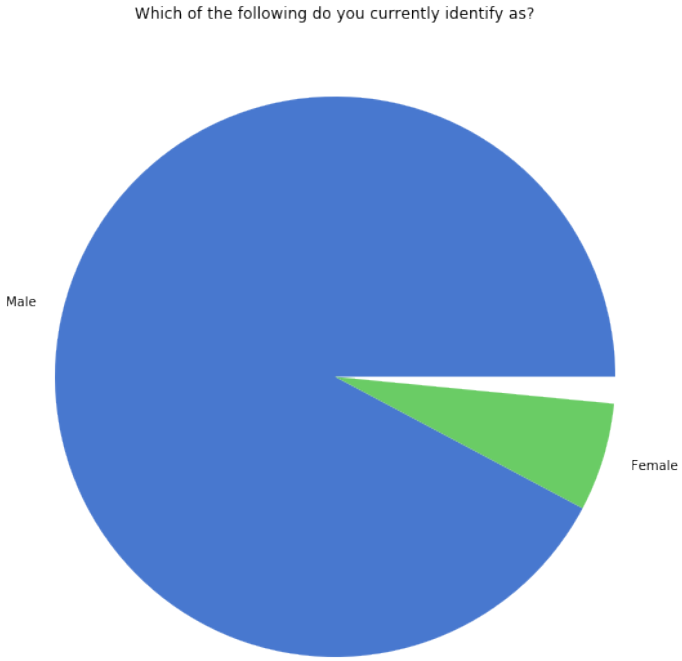

Figure 5: This graph shows which gender the respondents most identified with

Gender equality is especially in the field of software development a difficult topic. The survey shows that 92% of all respondents identify as male while only around 6% identify as female. The survey also offered a number of other options, we excluded them for the sake of simplicity.

2.3 Country

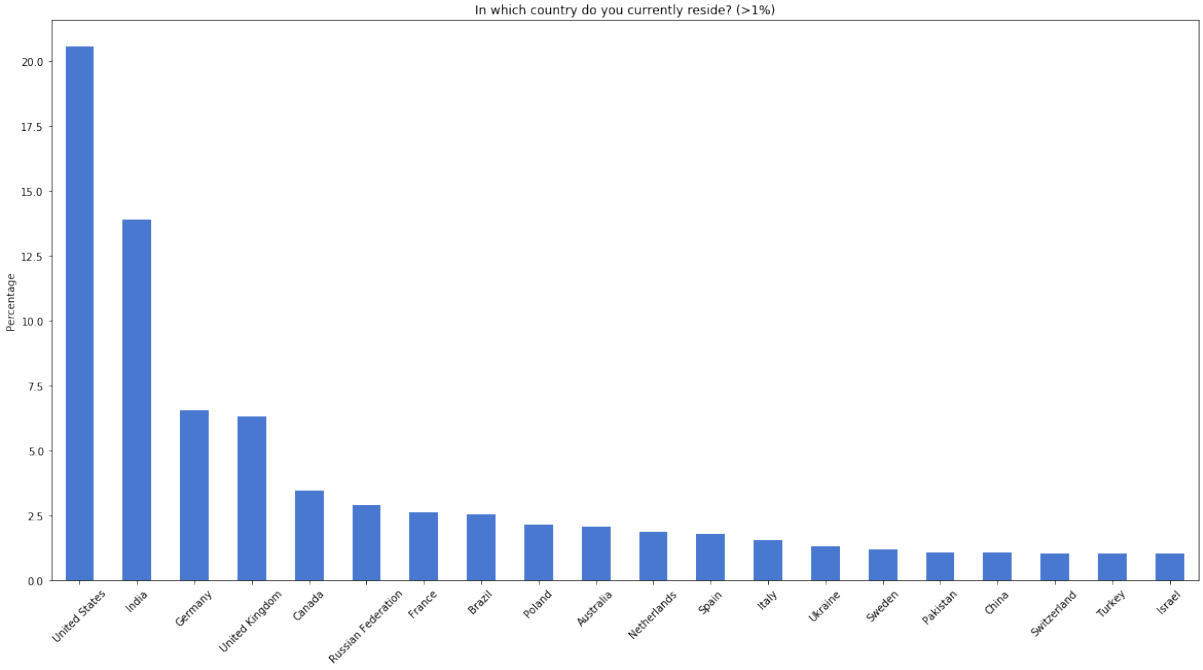

Figure 6: Top countries the respondents reside in

Figure 6 shows countries the respondents resided in while excluding countries with less than 1% of answers. The United States leads the graph followed by India, Germany, UK and Canada. India has of course a lot more inhabitants than the USA, figure 7 therefore looks at respondents per capita for the 5 most represented countries in this survey. In this ranking the USA falls behind Germany, UK and Canada while it shows that India has only a very small number of respondents per capita. All five countries have a high number of English speaking inhabitants, thus a language barrier would not explain this discrepancy. India has, of course, a much lower percentage of highly educated people than the other four countries. [1]

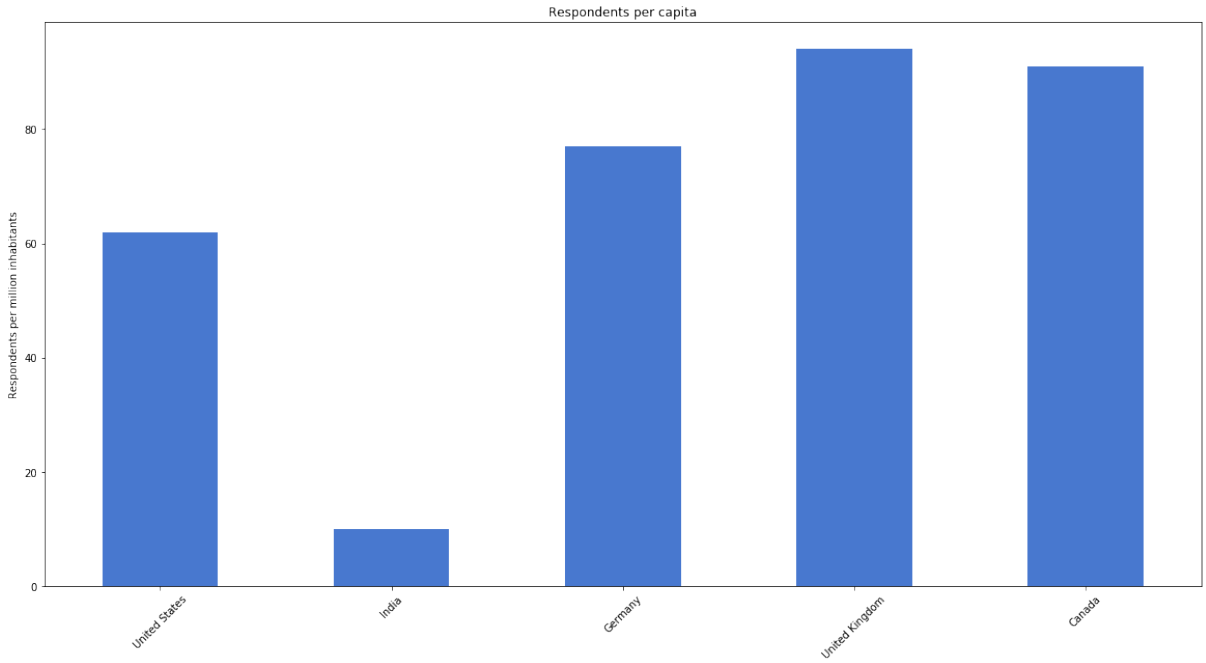

Figure 7: Respondents per capita for the top 5 countries

2.4 Education

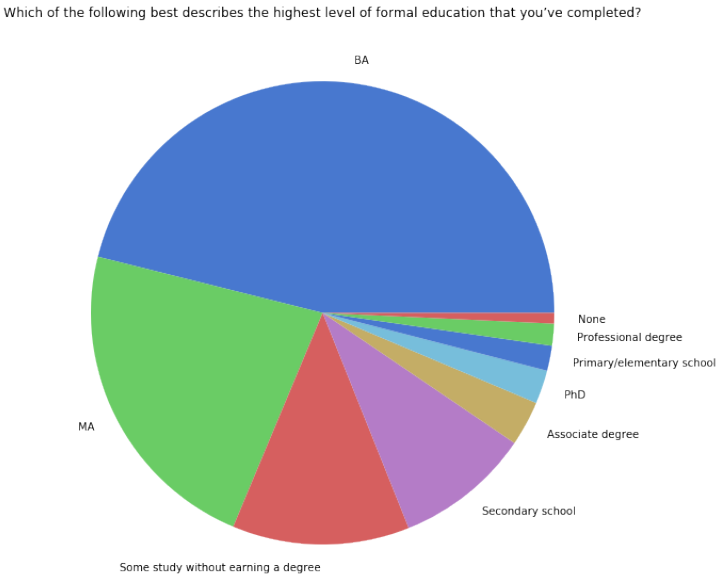

Figure 8: Education levels of all respondents

The question here stated: "Which of the following best describes the highest level of formal education that you've completed?".

Almost half of all respondents have completed some kind of Bachelor's degree while only about 22% had completed a Master's degree.

2.5 Salary

In this chapter we present an analysis of the reported salary by the respondents. We will look at the "Converted Salary", the survey states: "Salary converted to annual USD salaries using the exchange rate on 2018-01-18, assuming 12 working months and 50 working weeks." [3]

First we look at all the reported salaries in figure 9. The range goes up to 2,000,000 USD, we can clearly see that most peoples salary lies underneath 150,000 USD. Salaries which were exactly 0 were disregarded. In figure 10 we present salary information for values lower than 200,000 USD. This shows quite a flat distribution between 15,000 and 75,000 USD. In the following analyses we will only look at these lower salaries.

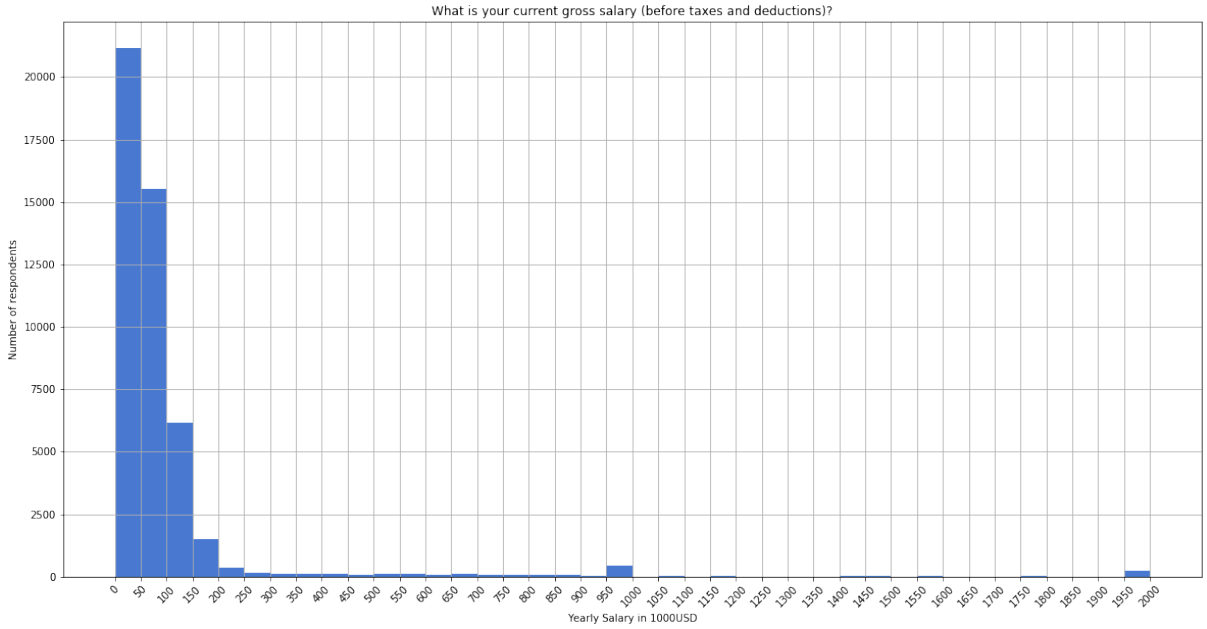

Figure 9: Total converted salary

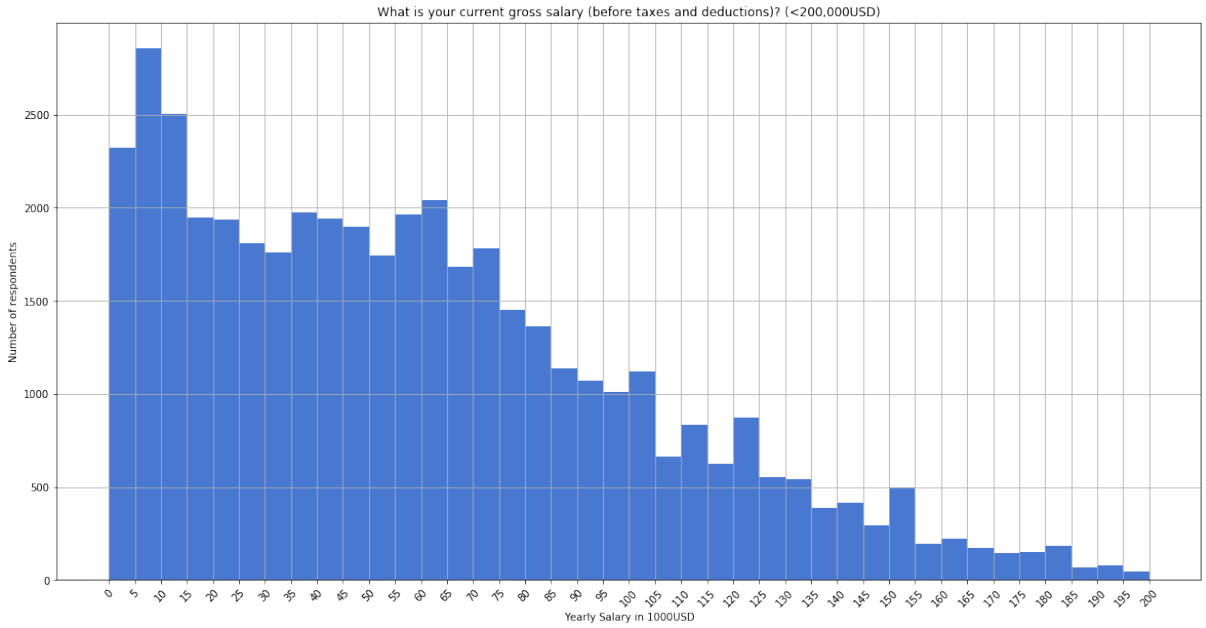

Figure 10: Converted salary under 200,000USD

Average salary of course depends on the living standards of the country one lives in. Hence in figure 11 we present the median salary per country (looking at the top 20 countries). We excluded all countries which had less than 5 respondents who answered this question. We used the median salary here to reduce the effect of very high or very low salaries.

The USA have the highest median salary at almost 100,000 USD followed by Switzerland and Israel. Germany has a median salary of around 60,000 USD (about 55,000 EUR).

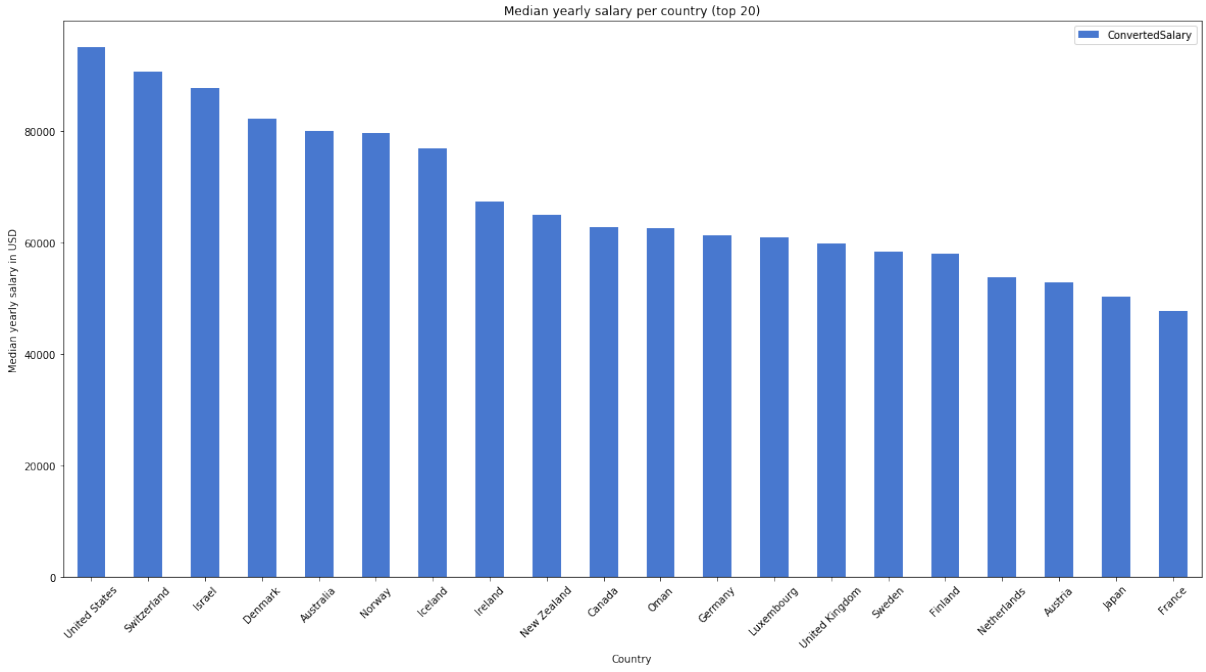

Figure 11: Converted salary by country (top 20)

3 Complex Analysis

3.1 Correlation between Experience, Education and salary

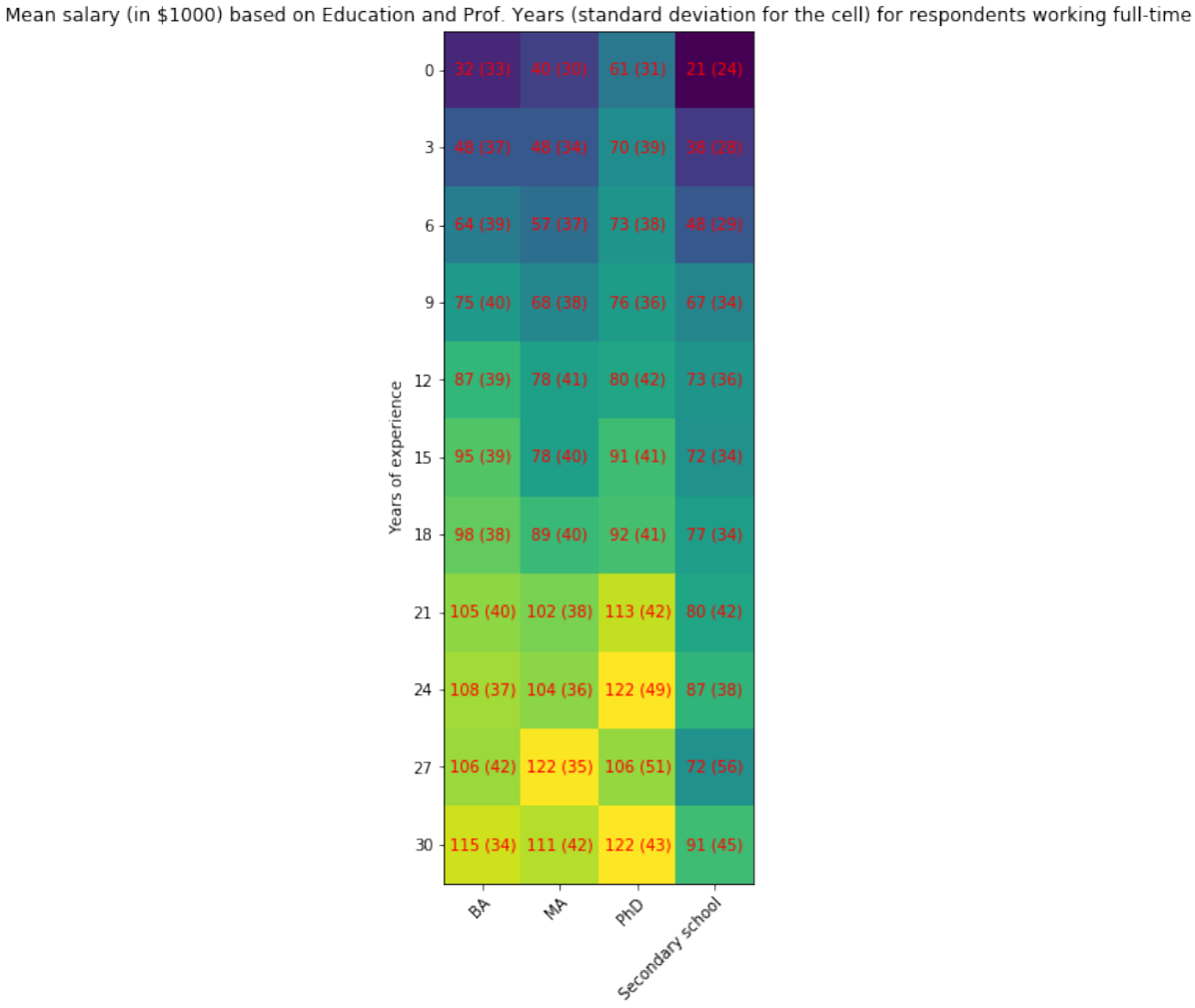

Figure 12: Median salaries for given years of experience and education. The standard deviation is added in parenthesis. Lighter colors show a higher salary.

Figure 12 shows a heatmap for the median (again excluding extreme salaries) salaries and their standard deviations for four education levels and all the experience ranges.

This plot shows a few clear correlations:

- Starting salary is highly dependent on education, a PhD will earn more than double as a BA absolvent.
- For BAs, MAs and PhDs more years of experience correspond to a higher salary.
- Respondents who answered with "Secondary school" have an overall lower salary.

After 24 years the second correlation does not apply anymore. This is probably caused by a low number of respondents in this experience range.

The standard deviation is for all cells pretty large (between 50 and 100%), this matches the overall standard deviation on our analysis of all salaries, though.

3.2 Salary prediction

In this chapter we will attempt to train a regression model in order to predict the salary of a given respondents given their:

- Education
- Age
- Country
- Years of experience
- Gender

We will again look at the top five countries regarding number of respondents.

3.3 Linear model

For this prediction we will use an Elastic net regression model. Elastic net uses a combined L1 and L2 regularizer which enables it to not only keep weights small but also potentially rule out certain parameters.

It uses the following model [2]:

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1) \tag{1}$$

Here y are the target values, X the input and β the parameters. λ_1 and λ_2 are regularization parameters.

We will also do some preprocessing on the data which is being fed into the regression model:

- The two scalar features Age and Years of experience in addition to the label (the salary) will be scaled to unit variance with zero mean.
- The remaining three categorical features will be "one-hot-encoded". This means for each category a new feature will be created which is 1 only if the category is present otherwise it is set to 0.

The input data has 18015 rows and will be split randomly 80:20 in training and test set.

3.4 Results

After training the model had a R^2 score of 0.57 on the training and test set. This indicates a moderate to low model suitability. Since the R^2 score of training and test data lie very close together, this speaks for a good generalisation of the model. We used a value of 0.001 as regularization strength (called α in scipy).

The parameters of the linear model can be seen in table 1. They are of course normalized as described in the chapter above. The zero values stem from the use of a L2 regularizer.

Feature	Value
YearsCodingProf	0.07775
Age	-0.00659
BA	0.00000
MA	0.02429
PhD	0.03045
Secondary school	-0.07787
Canada	0.00000
Germany	-0.00467
India	-0.19886
United Kingdom	0.00000
United States	0.16203
Female	-0.01092
Male	0.00821

Table 1: Regression parameters

Still this model gives us a nice intuition for the importance of all properties looking at the coefficients.

Let us first look at both of the numeric properties:

- As expected is experience an important factor and increases overall salary by roughly 2000USD.
- Age reduces salary, this is probably the case because experience and age tend to be correlated but salaries usually don't grow strict linearly over time.

The one-hot-encoded other properties also show interesting features:

- The model takes having a BA degree as the baseline, MA and PhDs are of cause a plus, while having no degree reduces salary.
- Countries show a similar picture: Canada, UK and Germany are roughly equal, while living in India reduces salaries by a lot and living in the US increases salary.
- Gender plays a small but not negligible role: being male raises salaries slighly.

4 References

[1] *Education stats for India*. http://uis.unesco.org/en/country/in. [Online; accessed 07-October-2019].

[2] *Elastic net regularization*. https://en.wikipedia.org/wiki/Elastic_net_regularization. [Online; accessed 03-October-2019].

[3] *StackOverflow Developer Survey 2018*. https://insights.stackoverflow.com/survey. [Online; accessed 09-September-2019]. 2018.