Predicting salaries using StackOverflow Data

Ishan Gupta
https://github.com/ishansgupta/ids-702-final-project

Introduction

Stack Overflow's annual Developer Survey is the largest and most comprehensive survey of people who code around the world. For the year 2019 more than 90,000 developers participated in this 20 minute survey.

The survey consists of a range of questions related to work satisfaction, employment type, salaries, programming languages, databases, frameworks etc. The total question set conists of 85 questions.

Number of programmers are increasing day by day, and so are programming languages. Very often the question arises, learning which language will have the most imapct. In this project, stackoverflow's salary data is used to see the impact different languages have on salaries. The impact of working in different countries is also being analyzed. From a traditional software engineer, the number of coding positions has expanded over time. We have different positions like Developer, Designer, Data Scientist, Business Analyst, Engineering Managers etc. We try to see which positions help your salary the most.

Apart from the answering the questions, we try to find more interesting insights.

Data and Exploratory Data Analysis

The data used in this study is the same set of the data in the original study. The dictionary is not included in this document for the sake of brevity. More details about the original study can be found at https://insights.stackoverflow.com/survey/2019. Data can be downloaded at https://insights.stackoverflow.com/survey. This dataset consists of the data dictionary as well.

In the data cleaning process, three major challenges had to be resolved.

1. Cleaning up different techonologies - We were given a respondent's programming languages in one column as ; seperated values, with the language being the primary language of the person. We split the values and stored top 10 languages for a user. In this project, we will be using the first language (primary) for our analysis.
2. Job Title Sanitization - Using domain knowledge, developer - Backend, developer - Frontend and other type of developers were grouped together.
3. While predicting salaries, number of years of work experience might seem a very legitimate metric. The survey question was "How many years have you coded or how many years have you coded in work?". The question is a bit vague, and as expected, we didn't see an upward trend with this variable to the compensation. We tried standardizing the compensation in different brackets but the results didn't look promising.

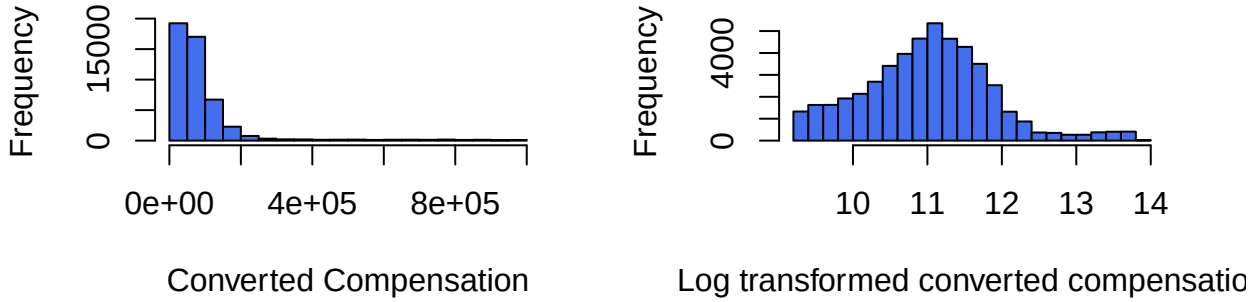
Apart from resolving the above issues, we removed data for respondents in the top 3%. This was done to remove outliers which might hamper our analysis.

Apart from this we verified the converted currency to USD provided by StackOverflow using an external package with latest conversion rates. The difference between the data didn't seem significant and we would be using the converted compensation USD(annually) provided by StackOverdlow.

We also notice major difference in salaried from different countries. There we would be examining the impact of salary on 10 high income countries

We started EDA by looking at the distribution of the response variable (ConvertedComp) and found that they are not normally distributed. They seem to be heavily left skewed; therefore, we apply log transformation

and the results are shown below. ConvertedComp looks normal after the log transformation, and we would be using the log transformation for the rest of our analysis.

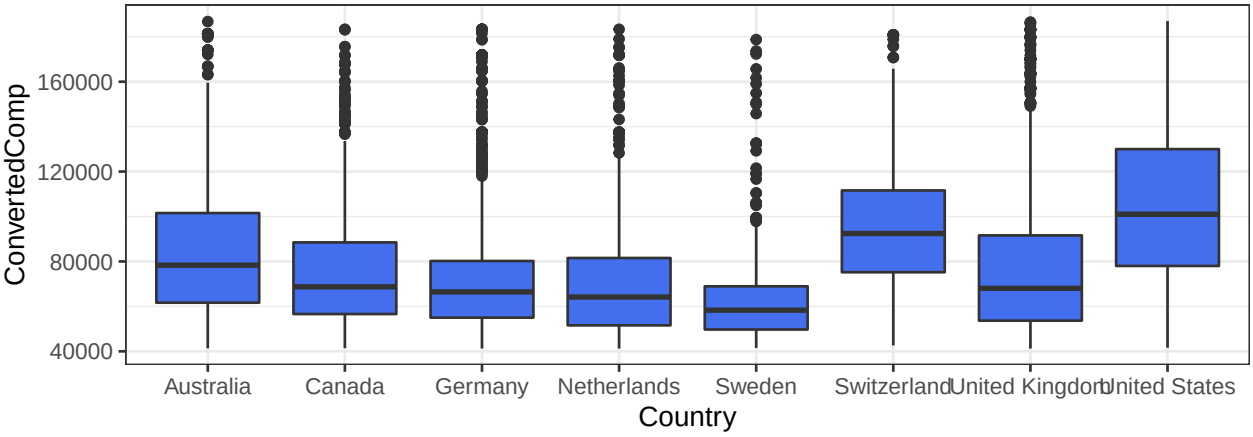


Then we examine different boxplots for different variables “Desgination”, “Primary Technology”, “Education qualification”, “Country”.

Comparing salaries across different countries does not make sense since the buying power of different countres might be very different. Eg. USA and India. Therefore, we only consider countries with high median incomes and high frequency of data.

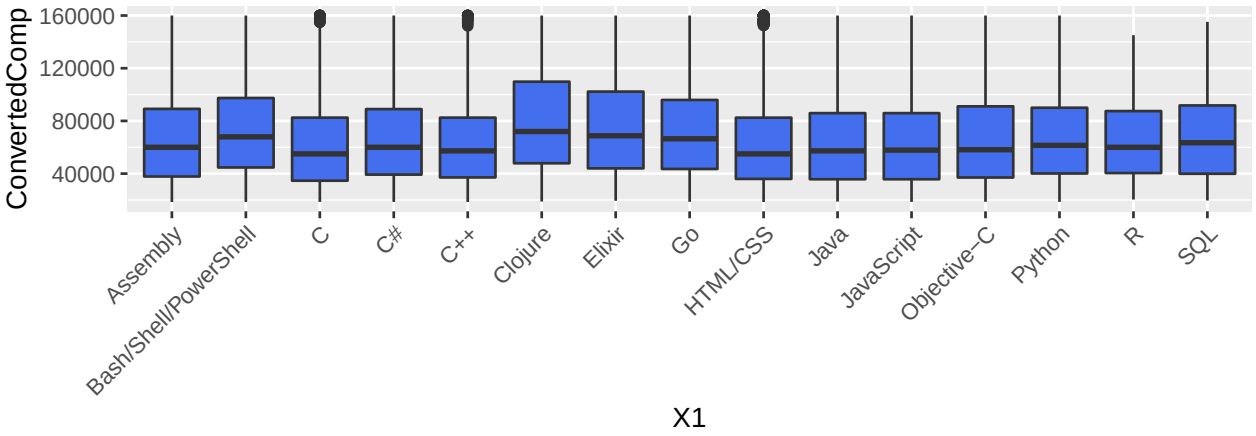
The final set of countries chosen for our analysis are United States, United Kingdom, Germany, Canada, Netherlands, Australia, Sweden, and Switzerland.

Let’s start off with examning it for different countries. We take the top 8 countries by the frequency of data and plot the boxplot.



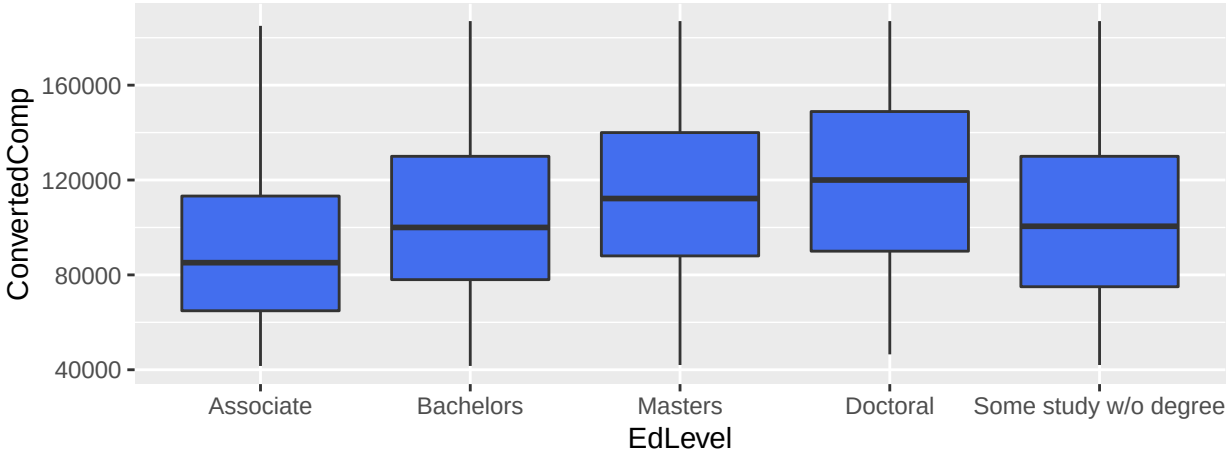
We observe some significant different in median salries amongst the high income countries as well

Next up, we start with boxplots for different technologies. Let’s start with a basic boxplot for top 12 languages as per frequency.



We observe the median salaries of developers in languages "Clojure" and "Elixir" is relatively high. And, HTML/CSS is relatively lower (basic web developers).

Lets plot the same for different education qualifications for the USA.

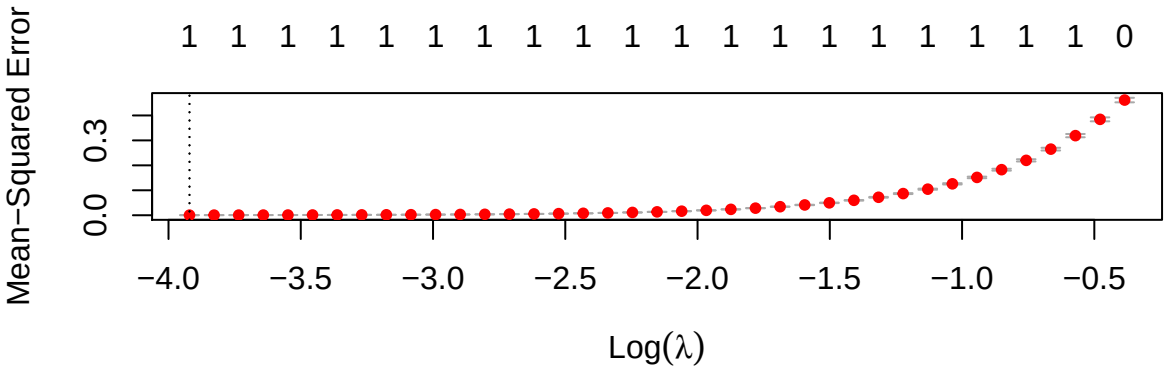


As expected masters and doctorals earn more than any other degrees. Surprisingly, the trend is not the acorss the countries we chose for our analysis, in some cases there is no difference between masters and bachelors salary at all. Please refer to the code in Appendix for further details.

Model

For training our model, we first randomly divide the training and test set in 70-30 ratio.

For the selection of our model we use domain knowledge, coupled with stepwise BIC and Lasso regularization. While using Lasso Regularization we don't see promising results for features and we obtain the value of lambda which doesn't remove any features.

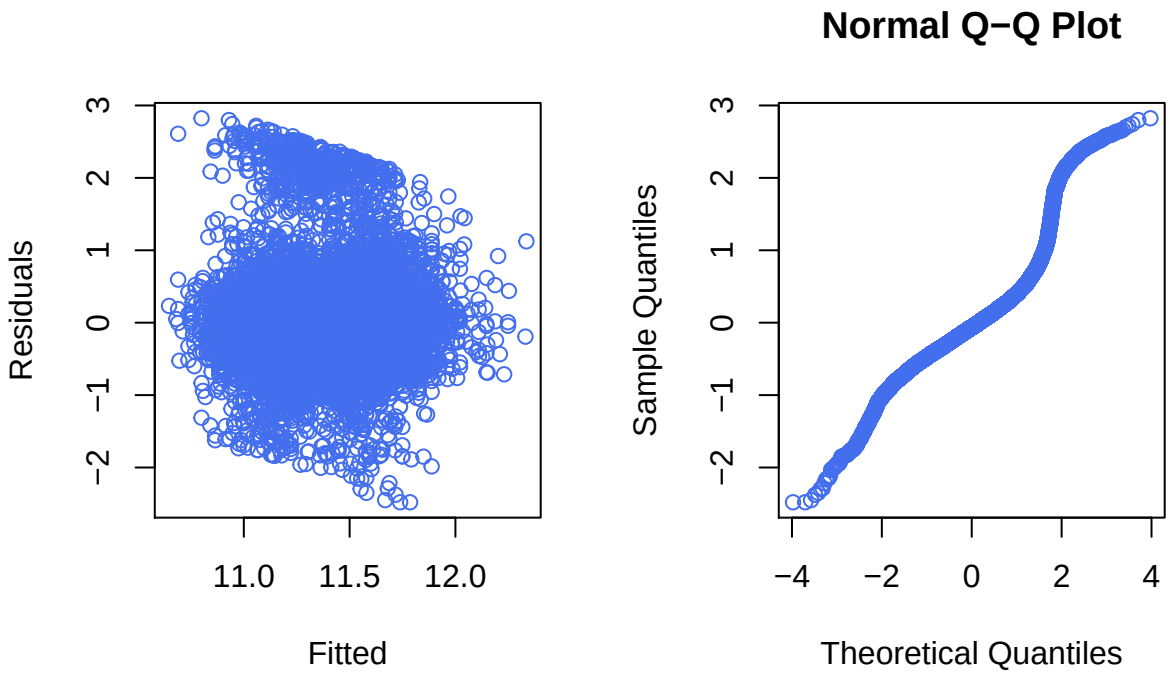


Since the number of levels in data is very high, we tried BIC for feature selection. Removing some features like gender manually (since these were not a part of this research question) . We add random intercept for technology and country to answer our research questions.

```
final_model <- lmer(ConvertedComp~(1|Country)+EdLevel+(1|Tech)+Designation+OpenSourcer,data = train)
```

Model Assessment and Validation

We start off with seeing the residual normality plot and fitted vs. residual plots



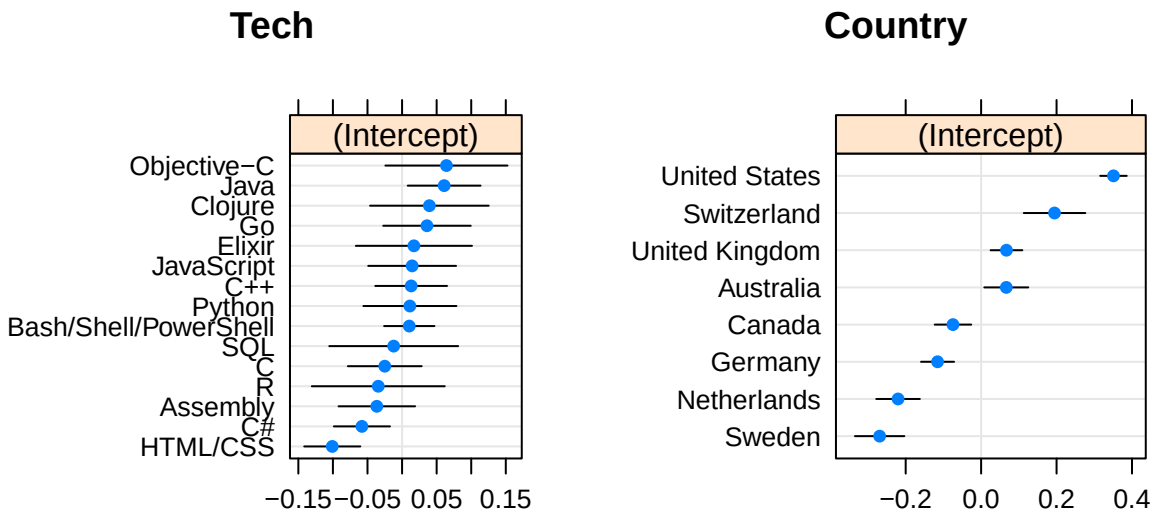
We observe that the plots look almost okay. Some kind of a trend in residuals variance, but nothing standing out.

We then run our model on the test set. We observe an R-squared value of 0.32, which is a bit on the lower side, and this might be one of the limitations of our model.

Results

To answer our research questions, lets start with analyzing random effects.

```
## $Tech
```




```
## $Country
```

1. We notice the baseline difference in salaris across different languages is very small, unless one is a HTML/CSS developer (completely different prgramming). This helps us conclude that since programming fundamentals remain consistent, so if you are good at one language you should be good.
2. There seems to be larger variance amongst countries, and we can see developers in USA get paid the highest compared to any other country.

Let's start analyzing fixed effects now. Some of the key finding from our analysis are which are significant at 0.05 significance leve.

1. Model baseline is a person with an associate degree, business analyst, with more than 1 open source contribution per year has an average salary of 64000$.
2. A DevOps Specialist is expected to earn 15.5% more keeping other factors constant.
3. A person with more frequent open source contributions is expected to earn 6% more keeping other factors constant.
4. Masters students are expected to earn 21% more than the baseline, keeping other factors constant.

Please refer to the Appendix for the complete coeffecients table (non exponentiated).

Appendix

	Estimate	Std. Error	df	t value	Pr(>|t|)
(Intercept)	11.070	0.083	10.468	132.907	0.000
EdLevelBachelor's degree (BA, BS, B.Eng., etc.)	0.199	0.026	14227.137	7.682	0.000
EdLevelMaster's degree (MA, MS, M.Eng., MBA, etc.)	0.337	0.027	14229.540	12.318	0.000
EdLevelOther doctoral degree (Ph.D, Ed.D., etc.)	0.471	0.041	14222.138	11.579	0.000
EdLevelSome college/university study without earning a degree	0.104	0.029	14225.826	3.604	0.000
DesignationData scientist or machine learning specialist	0.020	0.030	14191.831	0.673	0.501
DesignationDesigner	-0.013	0.028	13888.170	-0.451	0.652
DesignationDeveloper	0.058	0.019	12457.766	2.972	0.003
DesignationDevOps specialist	0.145	0.052	14221.463	2.759	0.006
DesignationEngineer, data	0.079	0.070	14226.028	1.122	0.262
DesignationEngineering manager	0.419	0.063	14227.096	6.609	0.000
OpenSourcerLess than once per year	-0.041	0.014	14227.919	-2.880	0.004
OpenSourcerNever	-0.111	0.014	14124.313	-7.939	0.000
OpenSourcerOnce a month or more often	0.061	0.019	14225.987	3.136	0.002