Are we ready for a job in data?

Data Visualisation of the

Stack Overflow 2019 Developer Survey

Student Name: Barry Sheppard

Student Number: 10387786

Module: B8IT107 Data Visualization & Communications

Introduction

When selecting the dataset for this project, my initial goal was to find a dataset of interest to the audience. As the audience is a class of Data Analytics students, some insights into the data job market seemed appropriate, which lead me to the Stack Overflow 2019 Developer Survey. If we compare the skill sets of actual developers in data fields with the subjects covered in the course, we will be able to see if the course fulfils the standards expected by the industry.

Stack Overflow is a question and answer website used by programmers since 2008 (Jeff Atwood, 2019). It has become so popular that people have joked that the field of programming should be renamed ‘searching stack overflow’. Each year the website hosts a developer survey asking about programmers’ job and their favourite technologies. With nearly 90,000 respondents in their 2019 survey, this is a robust dataset with great potential for insight.

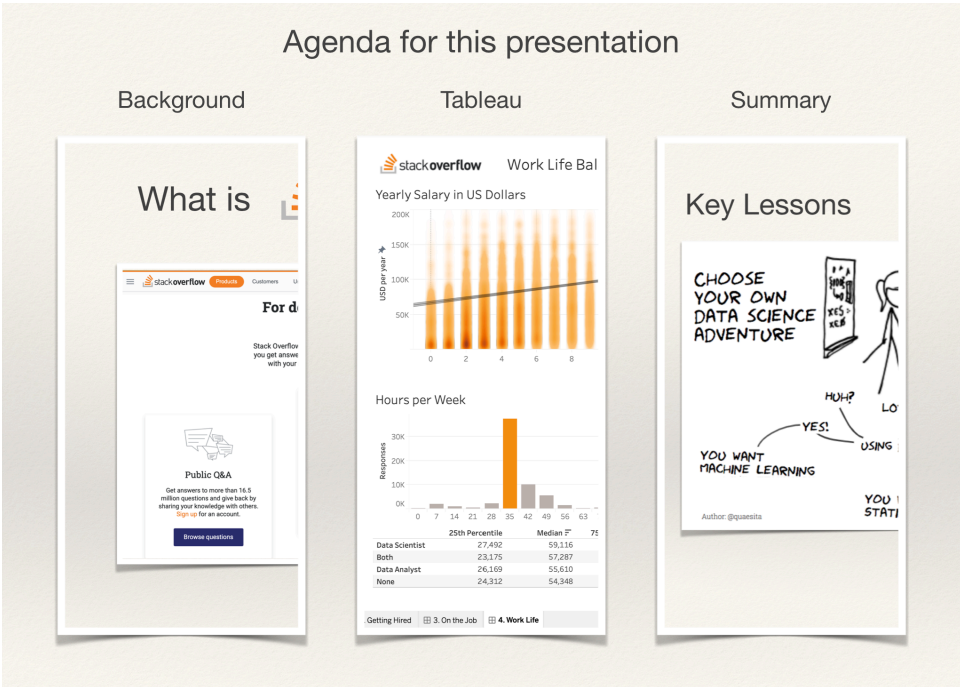
Overview of the dataset

The dataset (Stackoverflow, 2019) contains the survey responses to a total of 85 different questions. The survey itself is broken up into 6 sections covering 1) Basic Information, 2) Education, Work, and Career, 3) Technology and Tech Culture, 4) Stack Overflow Usage + Community, 5) Demographic Information, and 6) a review of the survey. Respondents are users of the Stack Overflow website, so there a selection bias, but usage of the site is so prevalent (Fullerton, 2019) that it should give decent representation. As we are interested in data roles, there will be a focus on respondents who indicated they were Data Analysts or Data Scientists which have around 10 thousand responses. For Irish respondents, there are about 500, with approximately 50 of those in data roles.

Objectives of the visualisation

The core message, and the central question that each aspect of the visualisations will build up to, is whether the Data Analytics course has prepared us for the job market. After an initial demographic overview, this will focus on 3 aspects. The ‘Getting a Job’ visualisations will focus on key elements you need on your resume and what the interview process will look like. The ‘Life on the Job’ visualisations will try to get some insight into what the actual work is like. Finally, the ‘Work life Balance’ visualisations will consider whether the job is a fulfilling one, looking at salary, hours, and job satisfaction.

Along with the Tableau visualisations, some slides were prepared, these were intentionally kept very simple, typically having only a title and an image. The overall theme for both the slides and the Tableau visualisations was kept similar to the Stack Overflow website. Primary colours were a background of white, with a palette of black, grey, and oranges. In several cases, titles used a similar format to the Stack Overflow bolding the last word. These kept the whole presentation in a unified theme and retained the link to the website.



Data cleaning in R and Tableau

Data cleaning was completed with R code producing a .csv file with new columns. Although the process outlined below follows a straight data cleaning, data import, and data visualisation process, in reality, the process was iterative, returning back to data cleaning as new visualisation requirements appeared.

For several questions, respondents could select from multiple options. In the dataset, the answers to questions like this appear in a single column that contains semi-colon separated lists. To visualise this, extra columns were added for each of the available options with a True or False response. Using R, a Grep text search was completed on the arrays to populate the new columns.

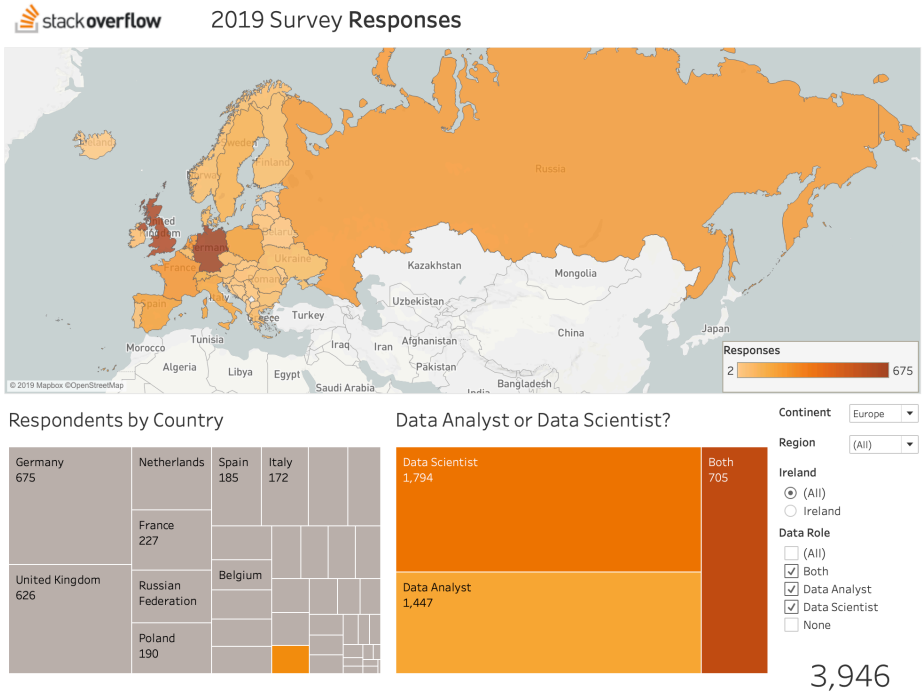
Some filter variables were also generated, such as ‘Region’, ‘Continent’, and ‘Ireland’ for geographical data or ‘Data Role’ to focus on specific aspects. Similarly, many plot specific variables were generated to aggregate data or complete other conversions, such as converting the currency to the Euro.

For some visualisations, such as the languages histogram, a calculated total was created for the Boolean values. For example, the Language_Python column contained 88883 responses of True or False generated a ‘Count of Python’ value of 36443, for total Python users.

All files were saved to a GitHub repository (Barry Sheppard, 2019).

Choice of Visualisations

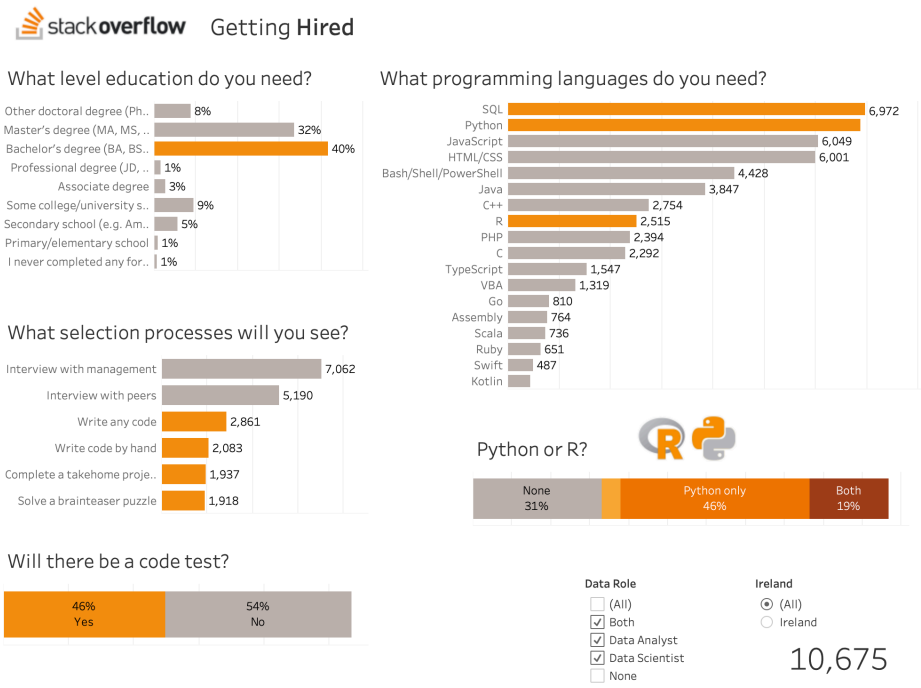
The first dashboard, 1. Overview, gives some context to the data, both geographically and career-wise. It consists of 3 charts. The first is a choropleth map based on responses by country. The number of responses is represented by a diverging pallet going from light orange to dark orange. The first view shows the full world map showing that the dataset is worldwide, but filters can drill this down by continent, region, country, and role to focus on specific areas.



The second plot is a treemap showing the same data, but in a form that makes comparisons more straightforward. Due to the large number of countries, a treemap is the most transparent option. As Ireland is of interest but is a small value, it was highlighted in orange to make sure it can always be seen. The last chart was another treemap which looks at whether respondents have data roles or not. A selected total appears in the bottom right along with filter selections, so we can always see what the total number of responses currently selected. This count was also used across the rest of the dashboards next to a selection of filters.

The next 3 dashboards are designed to tell a story. The first is around getting hired, the second is around life in the job, and the last is about work-life balance.

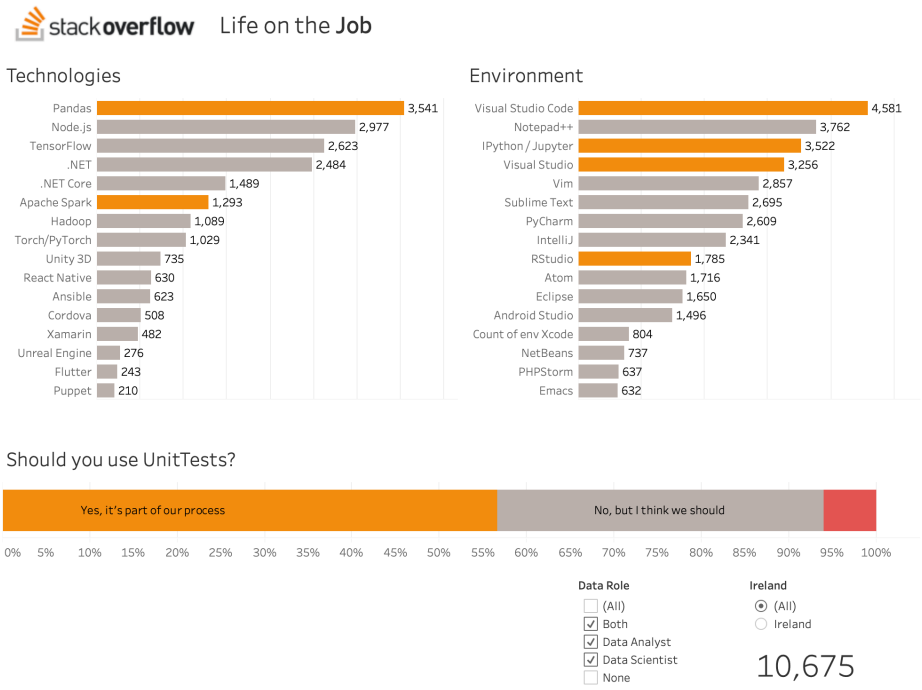
Dashboard, 2. Getting hired gives an overview of what companies are looking for and what the recruitment process looks like. The first chart was a histogram that shows the highest level of education received. The Higher Diploma of the course is the equivalent of a Bachelors degree (QQI, 2019) which was highlighted in orange. The second chart was another histogram looking at the types of interview processes candidates have gone through. One of the critical aspects I wanted to show in this plot is that we are likely to have a code test. Each of the code questions has a response rate of around 20%, but that doesn't tell us how likely respondents are to have at least one of those questions. An additional stack bar plot was added to show this. Combined, they tell a story, initially setting up the audience to think that maybe they don't have to worry about a test before revealing that almost 50% of the time they will have a test.



Moving over to the top right, there was a histogram showing the most popular programming languages. This was sorted by the most popular language for the data roles and when filtered will retain this original order, so changes in values are more evident. The languages covered during the course (SQL, Python, and R) have been

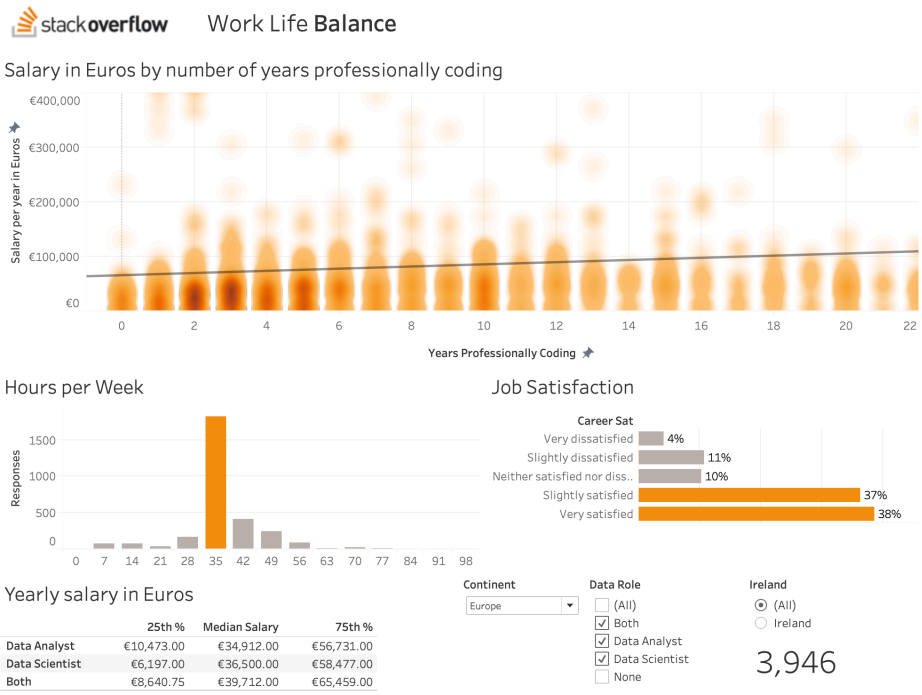
highlighted in orange to draw attention. Beneath that chart, was another stacked bar plot, this shows the number of respondents who know Python, R, neither, or both. Rather than just looking at whether respondents know the language or not, this also looks at the overlap and answers the question of which language we should focus on.

Dashboard 3. Life on the Job, gives some insight into what a typical day would involve. The top two plots are a histogram looking at some of the technologies and environments used, respectively. Again, those we covered on the course are highlighted in orange. The bottom plot was a stacked barplot which showed how popular UnitTests are, as this was a major part of the programming module this was highlighted in orange.



Dashboard, 3. Work Life, looks at what we as staff would get out of the role. The first plot considers annual salary. This plot was quite complicated for many reasons but primarily due to the massive number of responses. Plotting 88 thousand dots on a scatterplot gets quite complicated. Initially, a boxplot was considered, but due to the high maximum, these charts were not informative. The option decided upon, was to use a density plot showing annual salary on one axis with the number of years professional coding on the other. The axes were then

artificially limited to focus the attention on the area of highest density. To add context to this, a table was used, which showed the Data role, 25th percentile, median, and 75th percentile. This showed the range of the middle 50% of the data and should give a reasonable perspective.



Two additional histograms were added, displaying the number of hours worked in a week and the overall satisfaction. The number of hours variable was binned into groups of 7 to represent the average working day, and the typical business week of 35 hours was highlighted in orange. For the job satisfaction plot, the positive responses were also highlighted in orange.

Interpretations and Conclusions

There was always an expectation that the data would show the course prepared the class well, but there were a few surprises. On the first dashboard, we could see the number of Data Scientists was roughly equal with the number of Data Analysts, which was unexpected as Data Scientist is supposed to be a more prestigious and hence rarer role.

On the second dashboard, we confirmed the higher diploma level qualification was the education threshold needed and that SQL and Python were important. We also learned that Analysts use SQL more and favour HTML and Javascript, suggesting proficiency with web development. Data Scientists instead focused on Python and had a higher rate of knowing both Python and R.

On the third dashboard, we confirmed that Pandas was a key technology. For Analysts, we had a similar finding as earlier with web technologies such as node.js and .NET. We did see Tensorflow as an essential technology for Data Scientists, suggesting an opportunity for future growth. The environments included several we had already seen during the course. The last section, that was certainly a surprise, was how many developers were using UnitTests with over 50% saying they do and another 45% saying they don't but wanted to. This was a topic that had appeared on the course, but up until now, its importance was not evident.

On the last dashboard, we had some reassurance that the roles are worth taking. Pay is good, hours are reasonable, and most importantly, job satisfaction is high. Based on the data, the conclusion is that we are ready for a job in data. The visualisations clearly showed the importance of the topics we have covered but also gave a few opportunities which could be incorporated into the capstone project.

Ethical challenges

Due to its large number of questions, even though the dataset is anonymised, it may be possible to identify specific individuals and reveal sensitive information including their attitude to their manager, whether they are considering leaving their current employment, and any mental disorders.

The industry is heavily gender-biased, and this dataset is no different. In 2014, Amazon started working on artificial intelligence that could take 100 resumes and would pick out the top 5 candidates. They did this using supervised machine learning, where they took historic applicants and indicated which the successful candidates were. In retrospect, it really shouldn't have been a surprise that the algorithm produced was horribly biased, female applicants were rated lower than men as the data in the training dataset was biased (Dastin, 2018).

Reflections on working on this project

I also found working with such a large dataset frustrating at times and in retrospect would use the data cleaning process to create a trimmed dataset only using the columns I required.

One feature I was not familiar with was Tableau's animation, which is especially impressive when using time series data. The stack overflow website does have data from previous years available and incorporating that could have resulted in some interesting visualisations especially with relation to the growing popularity in some technologies suggesting trends for the future.

References

Sheppard, B., 2019. *barrysheppard/B8IT107DataVis* [WWW Document]. GitHub. URL https://github.com/barrysheppard/B8IT107DataVis (accessed 7.10.19).

Dastin, J., 2018. *Amazon scraps secret AI recruiting tool that showed bias against women - Reuters* [WWW Document]. URL https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G (accessed 7.10.19).

Fullerton, D., 2019. *State of the Stack 2019: A Year in Review* [WWW Document]. Stack Overflow Blog. URL https://stackoverflow.blog/2019/01/18/state-of-the-stack-2019-a-year-in-review/ (accessed 7.10.19).

Atwood, J., 2019. *What does Stack Overflow want to be when it grows up?* [WWW Document]. URL https://blog.codinghorror.com/what-does-stack-overflow-want-to-be-when-it-grows-up/ (accessed 7.10.19).

QQI, 2019. *Qualifications Frameworks - A European View* [WWW Document]. URL http://www.nfq-qqi.com/qualifications-frameworks.html (accessed 7.10.19).

Stackoverflow, 2019. *Stack Overflow Insights - Developer Hiring, Marketing, and User Research* [WWW Document]. URL https://insights.stackoverflow.com/survey/ (accessed 7.10.19).