Mathematics Developers Survey 2016

Nejc Ilenic

1 Introduction

Anonymised responses from Stack Overflow Annual Developer Survey are published each year along the results to encourage their further analysis. Being curious about where in the world and in which domain a data scientist should start her/his career, I attempted to answer some of the question by analysing the available dataset.

An important thing to note when interpreting the results however is that this data may not be a represantative sample from the population of mathematics developers. One should keep in mind that these are developers who were aware of the survey and were willing to answer the questions.

2 Data preparation

The dataset was constructed from survey that took place from January 7 to January 25, 2016, with responses originating from Stack Overflow, Stack Exchange technical sites, Facebook and Twitter. Most of the questions were demographic and regarding professional work and technology. Raw data consists of 56030 samples and 66 features, all of which are optional.

In order to obtain an adequately sizable sample, I have decided to include all respondents that belong to the occupation group of mathematics developers, which includes data scientists, machine learning developers and developers with statistics and mathematics backgrounds. After filtering out other occupations and responses with unknown countries we are left with 2132 samples.

3 Exploratory analysis

todo

3.1 Job satisfaction per country

Number of mathematics developers per country can be seen in Figure 1. Minimum number of 40 respondents is required to take the country into account and all others are placed into a single group called *Other*. Note that selected countries and number of people may be different when doing inference of specific features due to missing values (i.e. optional answers in the survey). Majority of respondents are from United States, followed by a combination of countries with less than 40 developers, United Kingdom, Germany and India.
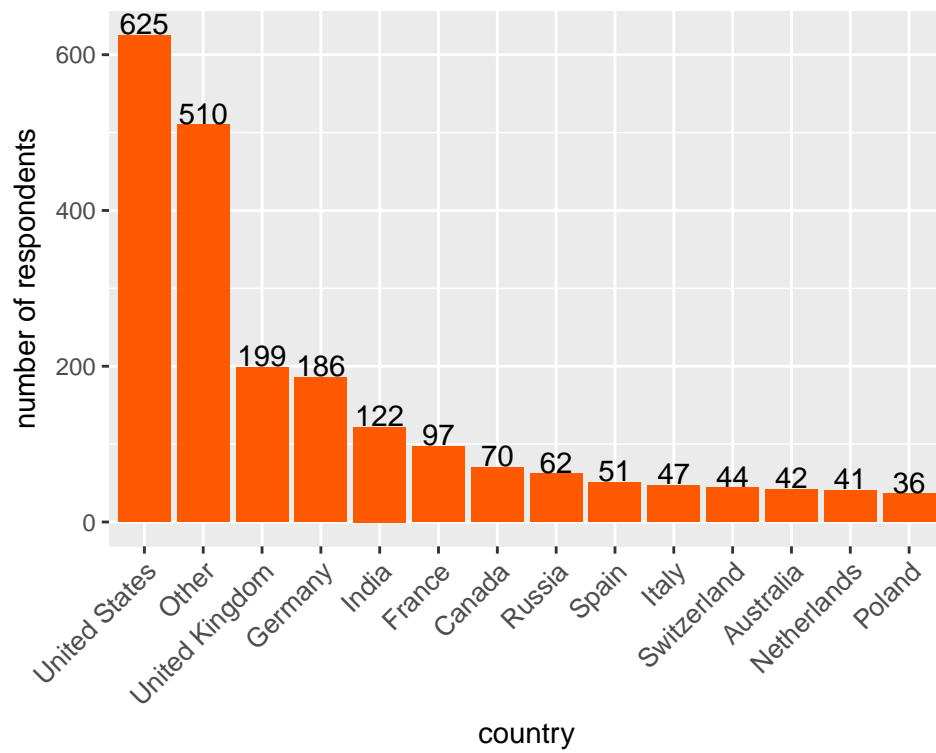
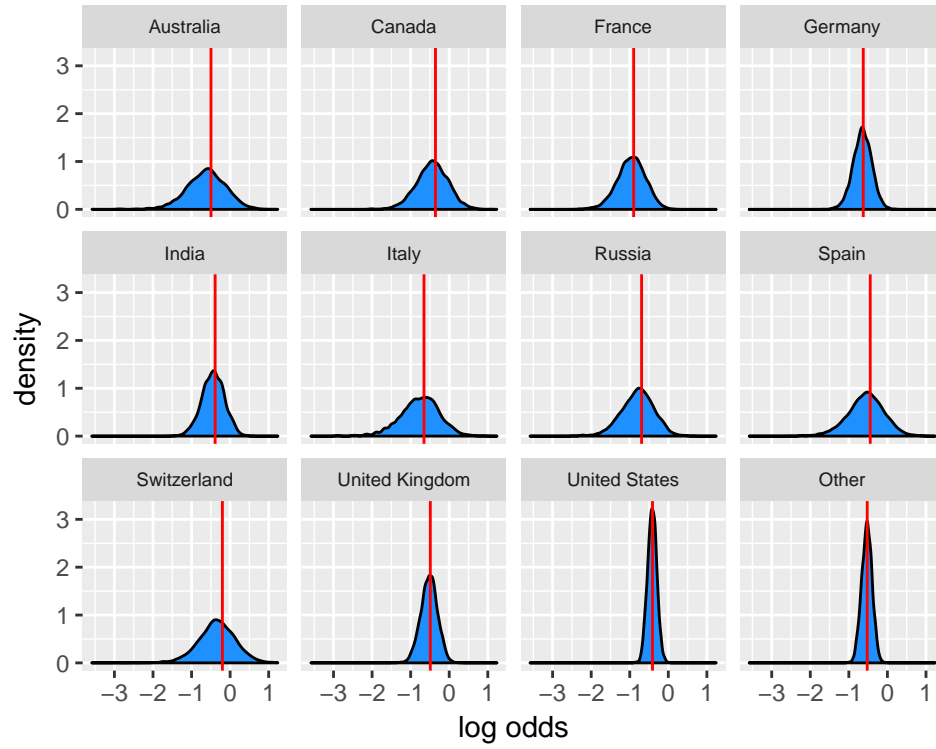Figure 1: Number of mathematics developers per country.

todo

Figure 2: Job satisfaction between countries posterior predictive check. Densities are sampling distributions of log odds of the answer 'I love my job' calculated from posterior predictive samples of the same sizes as the observed samples. Red lines indicate the observed log odds.
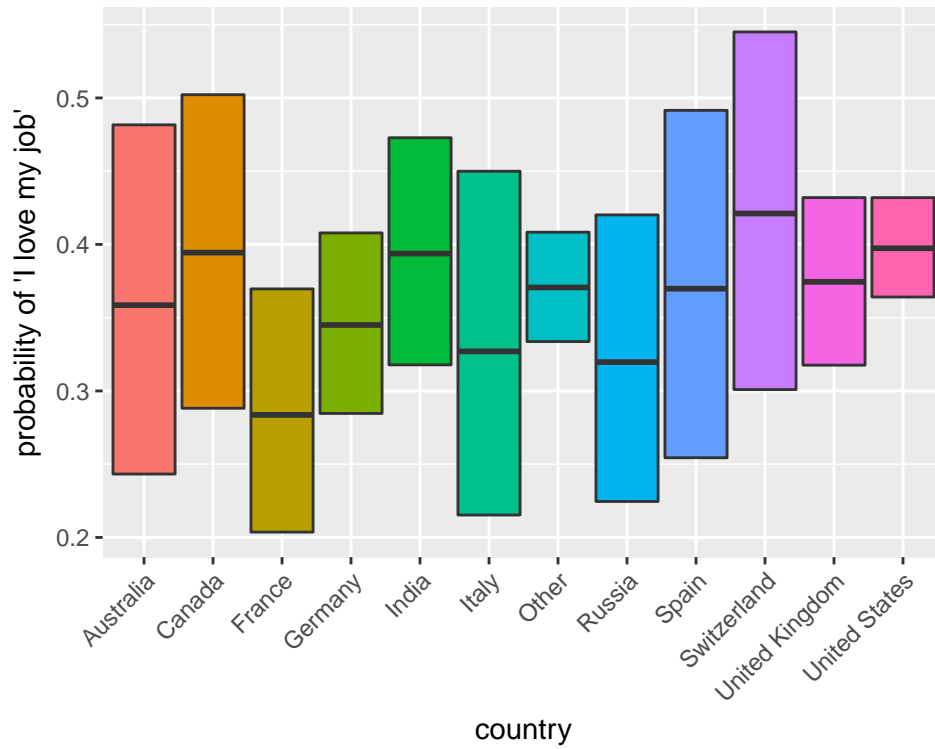
Figure 3: 90% confidence intervals for posterior probabilities of the answer 'I love my job' for all countries.

3.2 Job satisfaction explanatory variables

todo
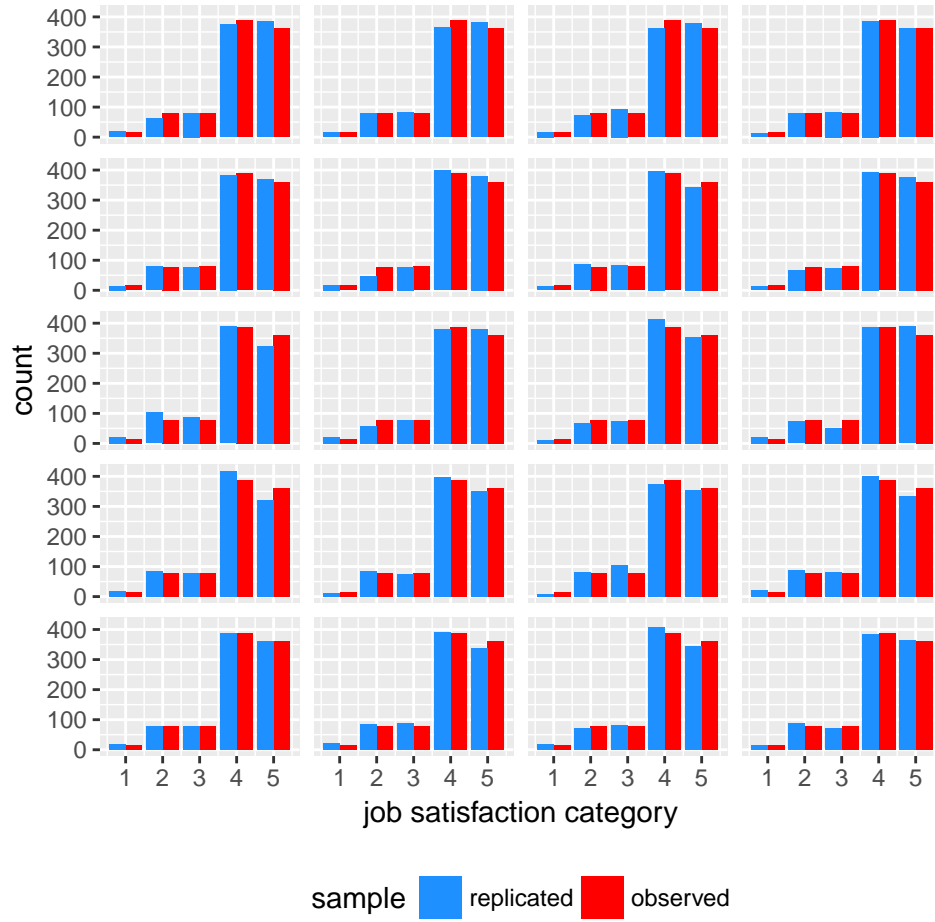
Figure 4: Posterior predictive check for conditional distribution of job satisfaction given explanatory variables. Plotted are histograms of 20 replicated samples along with a histogram of the observed sample.
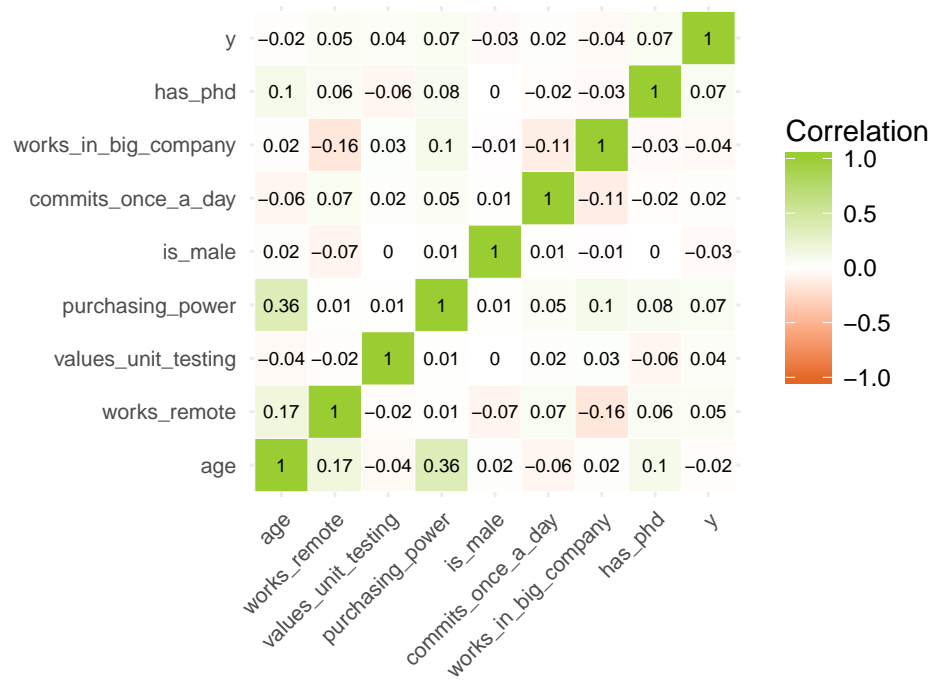
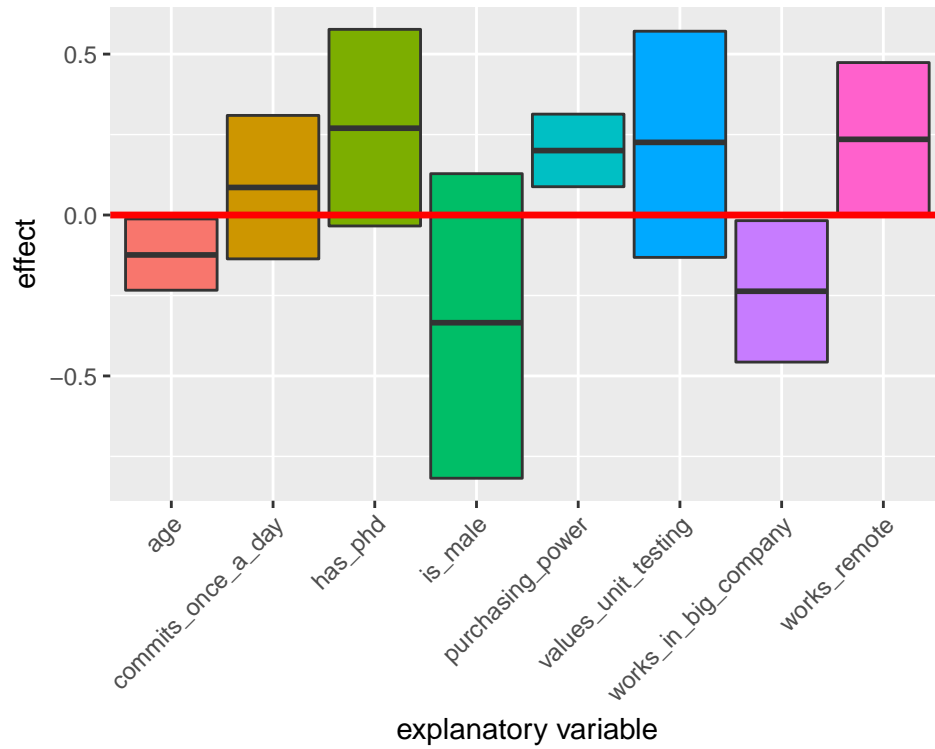Figure 5: Pearson correlation coefficients of selected explanatory variables and the dependent variable y (i.e. job satisfaction).

Figure 6: 90% confidence intervals for posterior regression coefficients of selected explanatory variables for job satisfaction.
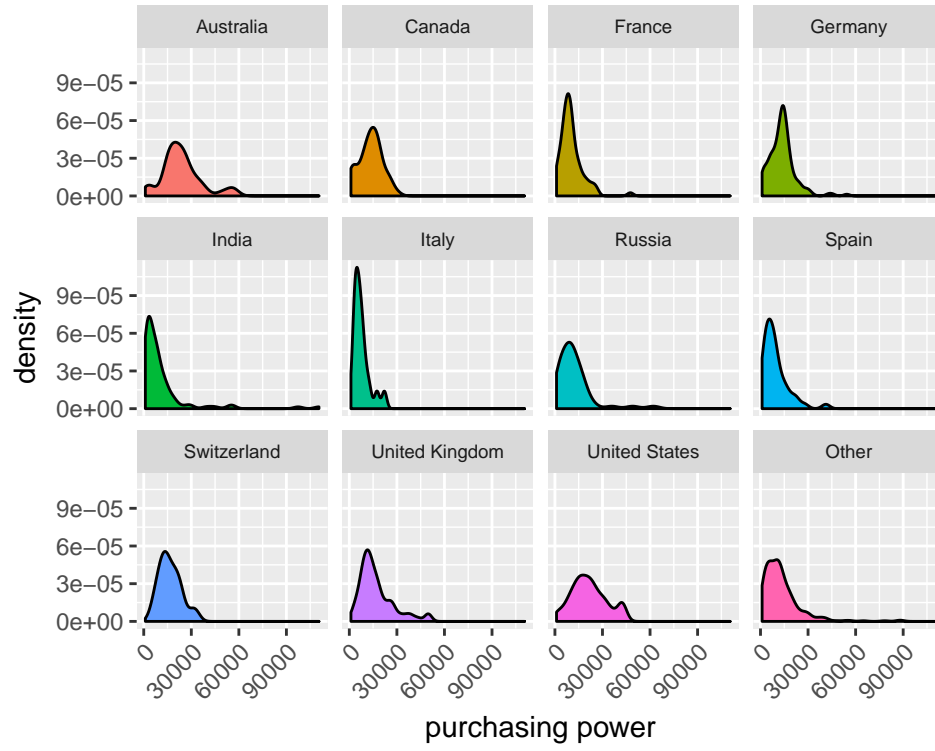
3.3 Purchasing power per country

todo

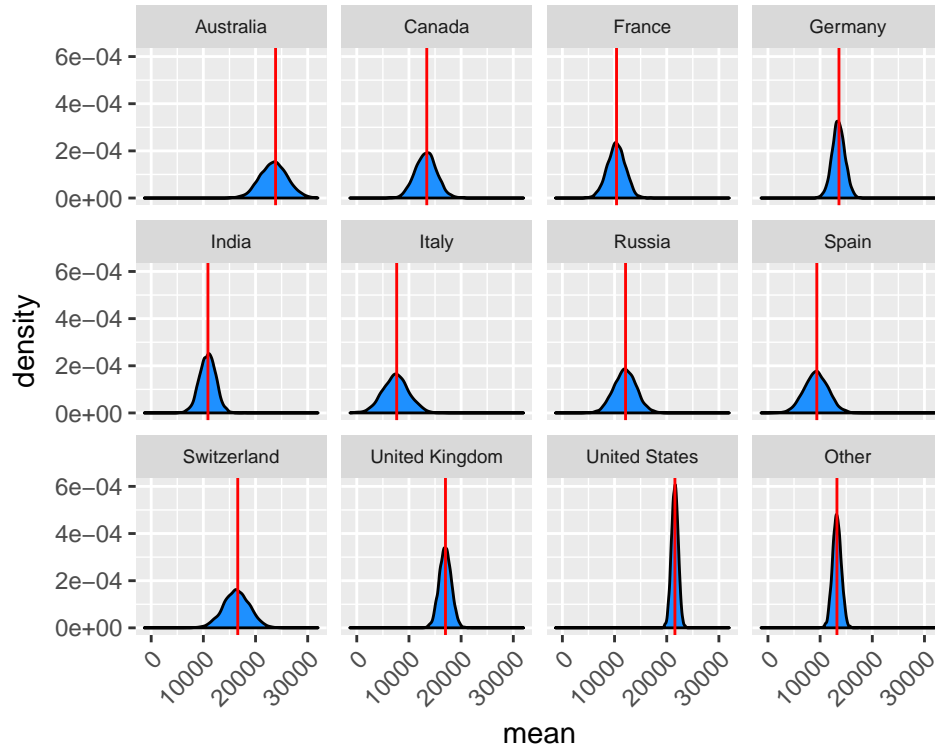Figure 7: Observed densities of purchasing power for all countries.

Figure 8: Purchasing power between countries posterior predictive check. Densities are sampling distributions of means from posterior predictive samples of the same sizes as the observed samples. Red lines indicate the observed mean.
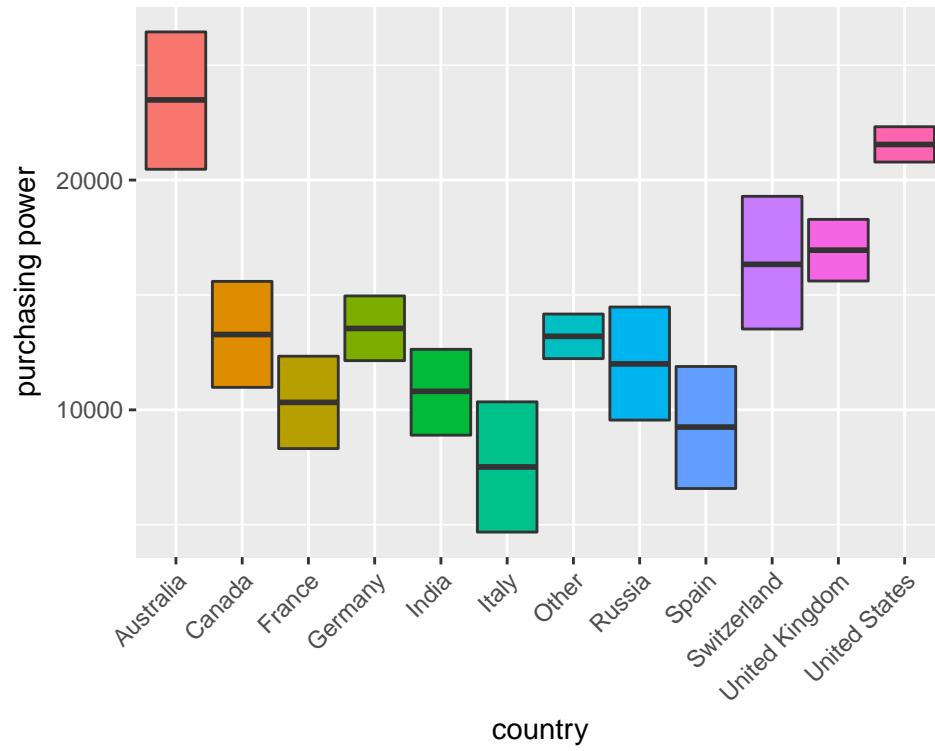
Figure 9: 90% confidence intervals for purchasing power posterior means for all countries.
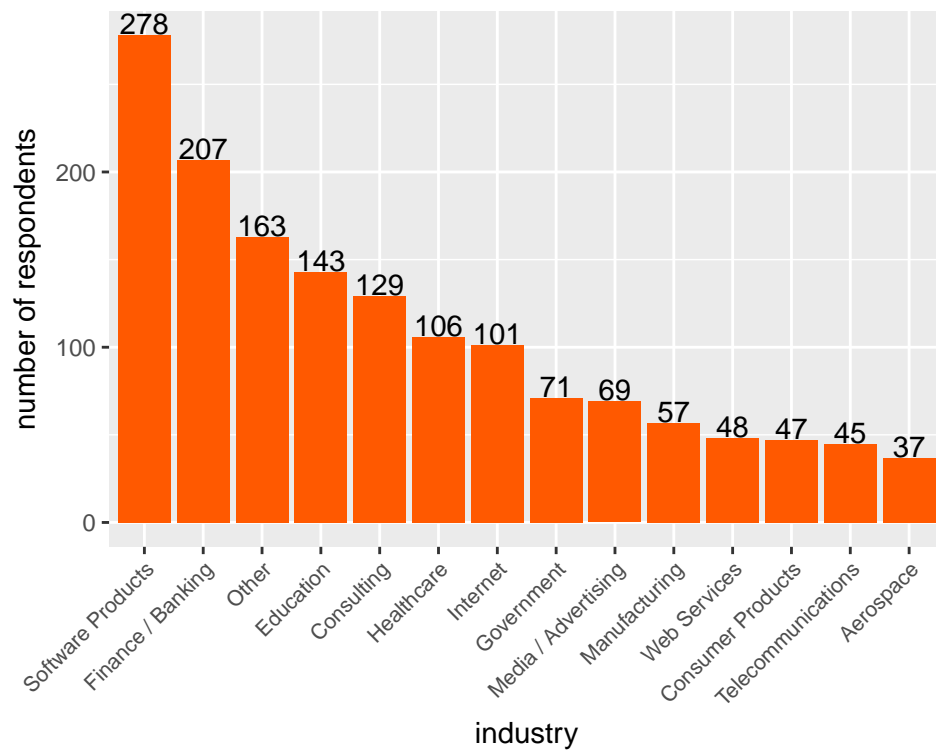
3.4 Job satisfaction per industry

todo

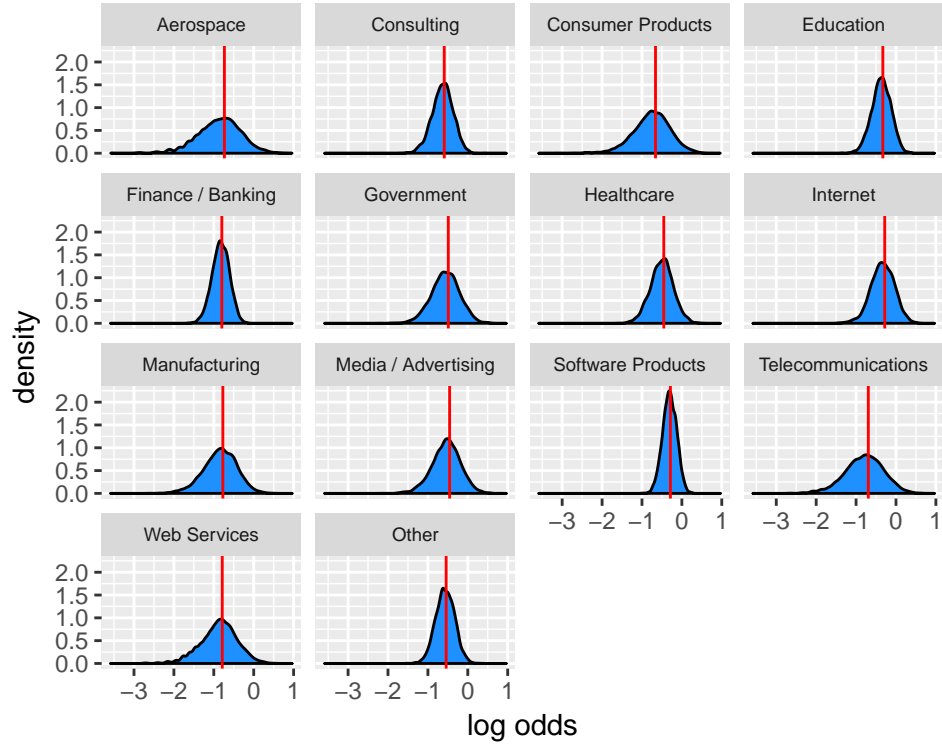Figure 10: Number of mathematics developers per industry.

Figure 11: Job satisfaction between industries posterior predictive check. Densities are sampling distributions of log odds of the answer 'I love my job' calculated from posterior predictive samples of the same sizes as the observed samples. Red lines indicate the observed log odds.
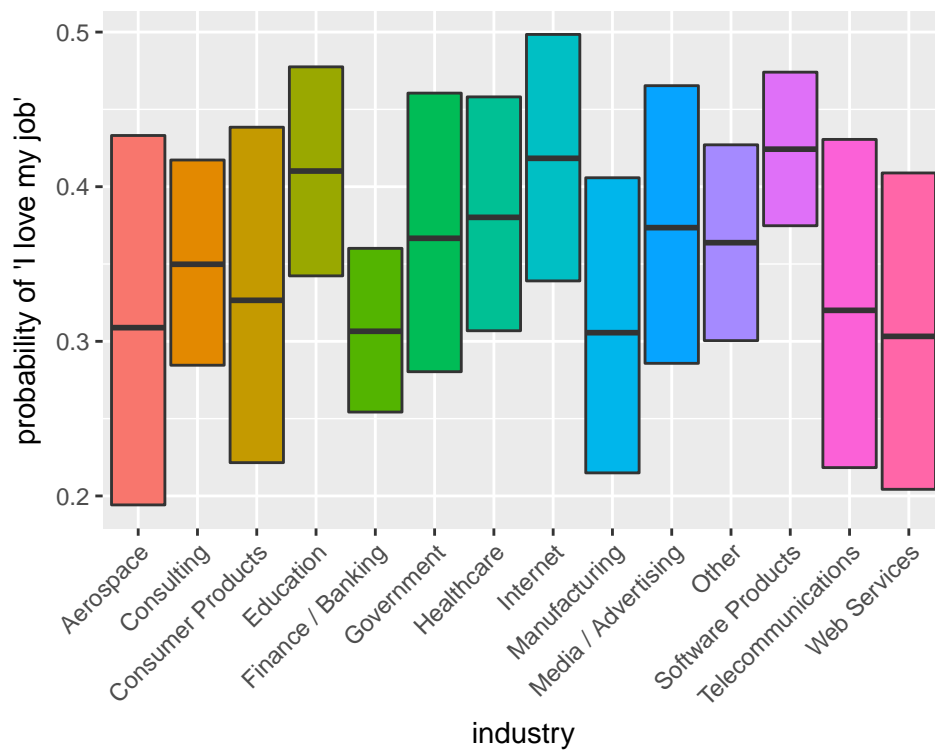
Figure 12: 90% confidence intervals for posterior probabilities of the answer 'I love my job' for all industries.

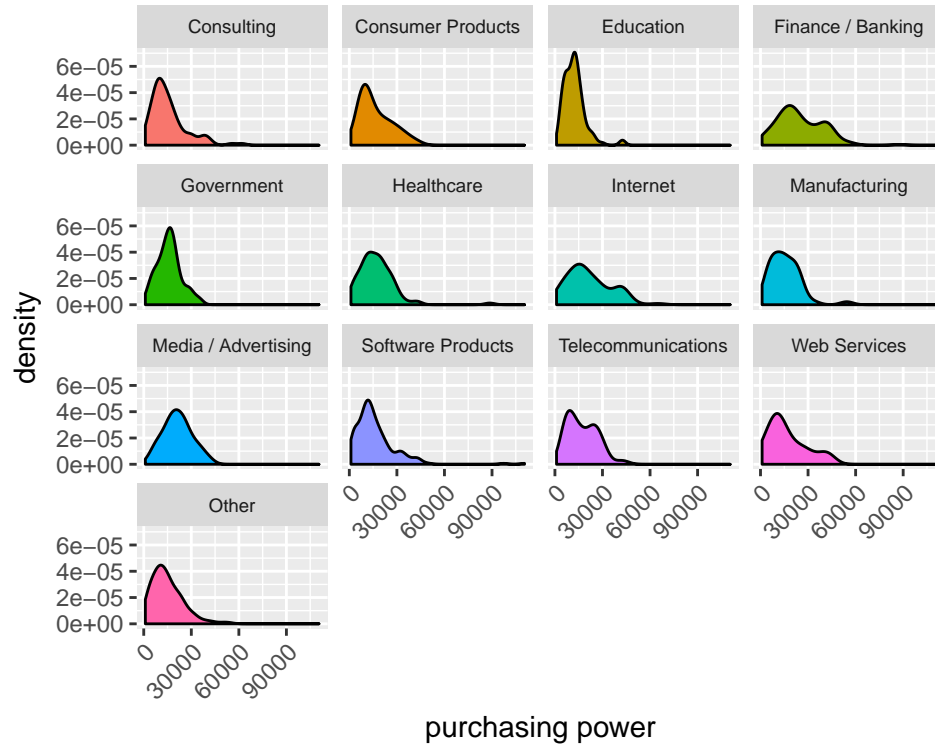## 3.5 Purchasing power per industry

todo

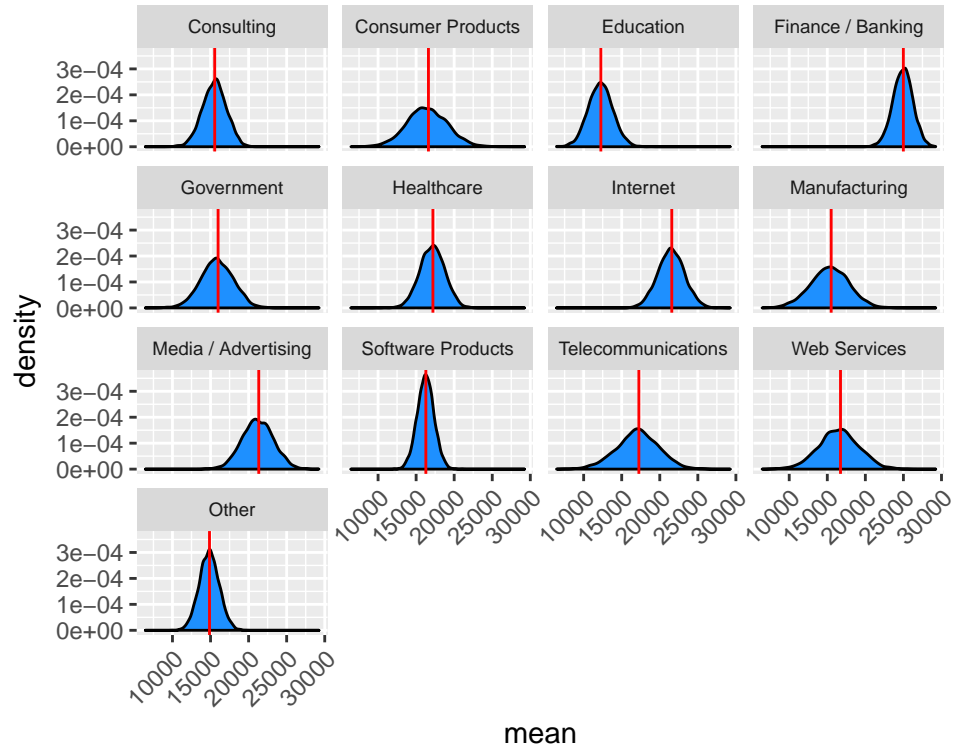Figure 13: Observed densities of purchasing power for all industries.

Figure 14: Purchasing power between industries posterior predictive check. Densities are sampling distributions of means from posterior predictive samples of the same sizes as the observed samples. Red lines indicate the observed mean.
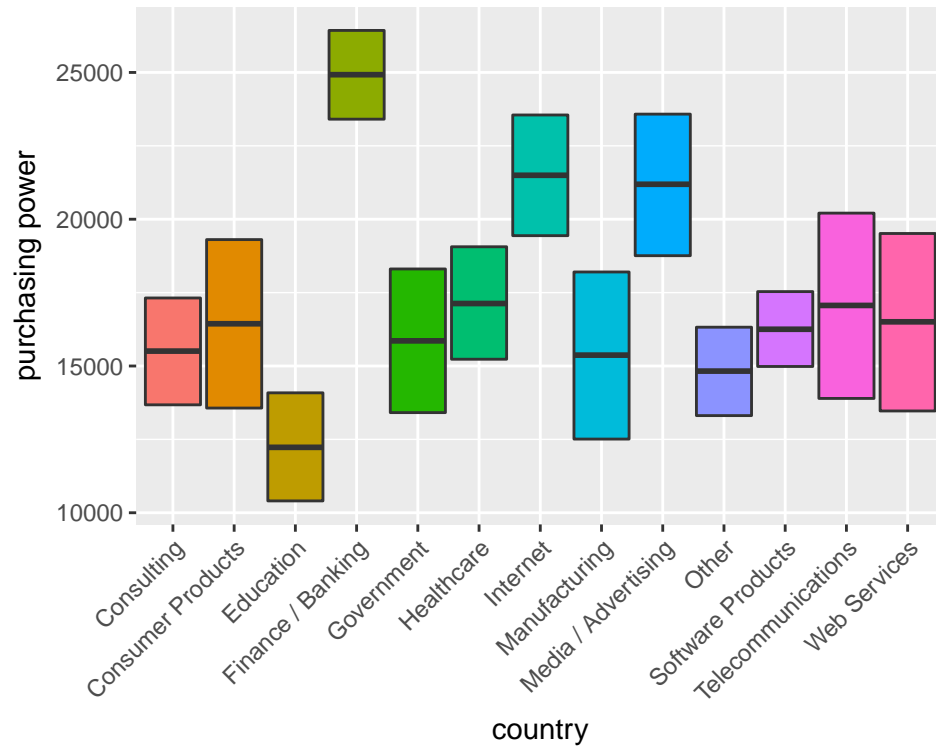
Figure 15: 90% confidence intervals for purchasing power posterior means for all industries.

4 Conclusion

todo