Project W25: Stackoverflow survey - differences by region

Members: Rando Tõnso

Please understand, that I am doing this alone and I do not want to waste a lot of time on this report. Because of this the word counts of the sections are too low.

This CRISP-DM is some kind of buzzword for managers and business people. I simply cannot understand why one should standardize an approach of solving a business problem. Standards are for things that can be precisely defined like programming languages, APIs and protocols. Instead of this CRISP-DM thing, one should learn about scientific method.

Business understanding

(This is not really a bussiness project so instead of business coals there are science coals. However I am not going to pretend that this little project is really a true science.) The StackOverflow survey is a self administered yearly survey that everybody can take part of. StackOverflow publishes their own report that illustrates some of the survey results, but there are still lots of undiscovered ways, to view this data. This project uses data from StackOverflow developer survey to compare and visualize the differences in technologies that developers use in different parts of the world. The coal is to discover some new and interesting aspects of the data (like correlations and patterns) and display them in interesting ways. Due to the nature of this task, there is a possibility that nothing interesting gets found and there are no particular differences between different parts of the world. In this unlikely case the project will document the different approaches taken to find something interesting and the results of these approaches.

The data necessary for the survey is freely available and to my knowledge there are no legal issues with using it.

This report will not cover the terminology here, because terminology will naturally develop and will be refined and redefined during the programming phase of the project.

More specific project coals are:

- Working with the data in a reproducible manner so that the entire process can be easily repeated on different computers.
- Not fixing erroneous data by hand or writing special case code that fixes single values.
- Documenting most parts of the code. I think that Jupyter notebooks are good for literate programming.
- Visualizing the discoveries within Jupyter. This may include sliders and animations. Visualizations should be interesting, but not fancy and confusing. It helps to think of different types of plots as functions.
- Including the interesting visualizations in the poster.

Data understanding

The dataset of 2019 survey is given in CSV format and contains about 90 000 rows and 85 columns. The columns can be categorized by topic:

1. General information (7 columns)
2. Education, work, career (36 columns)
3. Technology and culture (21 columns)
4. Stack Overflow usage and community (13 columns)
5. Demographics (6 columns)
6. Survey feedback (2 columns)

The focus of this project is not on StackOverflow usage and community or career and salary.

The columns themselves contain nominal, ordinal and numeric attributes. There are also some columns that I will call composite, for example the LanguageWorkedWith column contains a list of semicolon separated programming language names. These columns need futher processing to be easily usable with pandas.

The survey webpage that the respondents saw is available as a PDF document and can be used to gain some understanding about how the wording of the questions might have affected the answers. Some questions with checkboxes or radio buttons had an extra field to add an answer that was not given by the makers of the survey.

According to Stack Overflow the data is cleaned but I discovered that some developers are one year old. Therefore I should be skeptical and run some basic sanity checks on the data before starting discovery. In general the data seems to be good enough for the project.

There are lots of columns and because of that I have a Python file in which I can “tag” the columns as categorical, numeric, composite and so on.

I had a plan to include data from surveys of the previous years, but “tagging” the columns takes a lot of time. I consider using the previous surveys only if corresponding columns that I am interested in have the same format in all surveys.

I may also use some other data sources to get common statistical data about countries like population or human development index.

Planning

Since I am doing the project alone I have to do 60 hours of work. I will try to divide this into smaller parts and estimate time for each of them. This list is not in strict order and the progress in different areas can happen at the same time as they are overlapping.

- Familiarizing myself with the survey webpage and reading the results found by Stack Overflow. (2h) DONE
- Writing this report. (3h) DONE
- Familiarizing myself with the CSV file and programming code to read it and convert it to suitable pandas DataFrame(s). (10h) PARTLY DONE

- Doing sanity checks on the data. (4h) PARTLY DONE
- Discovering interesting things. (8h)
- Formatting notebooks, adding comments and cleaning up code. (10h)
- Making plots not ugly. (5h)
- Exporting plots to the poster. I have never done that before so hard to predict. (3h)
- Writing the poster. (7h)
- Poster layout and final tweaks. (3h)
- Looking for stupid mistakes. (5h)

This schedule is approximate and may wildly change, because estimating time in any kind of software project is very hard. In practice I will have to divide the time based on the need and not this plan. After all the coal is to make a sufficiently good project in time and all areas of the project should be balanced.

I plan to use following tools:

- Pandas
- Matplotlib
- Something for poster making. My ideas are Adobe Illustrator and LaTeX. The problem is that Adobe illustrator is too low level (I have to drag rectangles, draw boxes, play with colors and borders, align edges) and LaTeX is also too low level (lots of syntax and I have not used it for posters).