Data Processing

Process Book

May 16, 2024

Full Name	Student ID	University
William Bærenholdt	15348857	UvA

Amsterdam

Date of entry: May 07

What I've worked on:

Initially, my plan for the project was to track 'the development of life expectancy and fertility for different countries'. Therefore, I began following the 'Bokeh visualization library' module to understand the package. However, I ended up changing my subject to consider expected ages of deaths between men and women, as this theme is more aligned with my background studies (Actuarial Mathematics). Therefore, I also spent quite some time finding data.

What problems I encountered:

I really wanted to learn the Bokeh library, which is why it took me some time to get through the notebook. I found it pretty difficult to understand some of the themes, especially because I was not even sure if I was going to use the specific type of effect or visualization yet. Also, with that in mind, my initial idea was to make a project about something else. Additionally, my Dutch skills were insufficient to find the right data, so I used some Danish (my home country) data because I knew where to look for it. I am still not sure what my project should be completely about yet.

What I learned:

I almost finished the notebook and got a broad idea about what to expect from the library. I made a lot of good notes for my future graphs so I can style them from the library and know where to find the specific style.

Which resources did I use:

I used the provided notebook https://sp.proglab.nl/project/bokeh. Furthermore, I briefly read the recommended further readings: Visual encoding, Michael Dubakov; Learning to see, Oliver Reichenstein; Data visualization: Clarity of Aesthetics, Ben Jones (however, I found sources quite overlapping).

Date of entry: May 08

What I've worked on:

Today, I began to explore my dataset to get a sense of the type of information I could use. I created some test graphs and functions as well. I delved into making a plot where you can toggle the graph on and off using the Bokeh library. Today was primarily focused on data processing, and due to the time spent on it, I've decided to incorporate some of my data handling thoughts into the project.

What problems I encountered:

I always find it daunting to download and read a dataframe for some reason. There were many mistakes in my data which I initially handled manually but later attempted to rectify using Python code. I know you're not supposed to manually alter your data files, but I couldn't find a solution otherwise. Apart from that, I began to delve into some Python code. Not because I encountered specific problems, but I mention it here since the first time you code something is never correct.

What I learned:

Nothing specific comes to mind other than some Bokeh library functions. And yes, that display() function is much prettier than print()... There was some missing data in my data, so I revisited the section on how to handle missing data. However, I began to implement linear regression to address it, but then realized my project was becoming overly mathematical. Therefore, I opted for a simpler approach without compromising the quality of the code.

Which resources did I use:

https://docs.bokeh.org/en/latest/index.html

Date of entry: May 09

What I've worked on:

I'm not entirely settled on the formulation of my project, but I aim to derive some results on life expectancy for men and women and how it changes over the years, as well as the differences between them. This is highly relevant in my field since we often discuss people's longevity (which is significant for a pension company). My main goal for today was to derive a function to calculate the expected age of death for different years, and I am quite satisfied with the result. I also realized that there was a GitHub part of the module, which I spent some time understanding.

What problems I encountered:

I needed to find a way to calculate the expected age of death for an arbitrary person in my dataframe. This required some mathematics, and I revisited some courses I had taken in Denmark. Thereby, I derived a function to calculate this (which took me longer than expected, despite the formula not being that complicated). Although one might think using a nested loop (a for within another) would be appropriate for calculating since it's a sum of a sum, I wanted to avoid it due to computation time. I am quite satisfied with the result, and the function works completely as well as when I calculate by hand. Regarding the GitHub repository, I don't quite understand why it is necessary to use it, so for now, I will leave it and ask in the upcoming exercise class (which, sadly, is on Monday due to a public holiday - today is Thursday).

What I learned:

Well, never to give up, and every time you have some duplicating code, there is always a way to shorten it, which is quite useful when you find a mistake and only have to redo the problem once. This gave me the eagerness to really set a nice standard for the coding in this project, which I am quite satisfied with so far.

Which resources did I use:

Lecture notes from a course in "Mathematics in Life Insurance" (University of Copenhagen)

Date of entry: May 10-12

Note: This notebook encompasses everything I've worked on, the problems I encountered, and what I learned during the weekend. Since I've been working on the project sporadically, it's better to compile everything for the entire weekend.

What I've worked on:

I focused on my scientific question, 'Is there a gender that we expect to live longer than the other in Denmark?' This required a lot of time using different graphs and styles from the Bokeh package. I am pretty satisfied with the outcome even though it took longer than expected. I made sure that my graphs were nicely presented before proceeding, as I wanted to demonstrate the skills I've acquired using the Bokeh package.

What problems I encountered:

This part has been more about project writing and technical issues. I needed to determine which results were useful and presentable. I delved into some rabbit holes, misusing some data and deriving incorrect results, which led me to delete a lot of work. I wanted to create an interactive graph where I could toggle lines on and off in the notebook, but I couldn't implement this functionality within the notebook itself (it would only be possible in a web browser). This was quite frustrating, although my results remained unchanged.

What I learned:

I learned how to create interactive graphs using the Bokeh package. I also discovered some other Bokeh features, such as 'hover' and other styling options, which are showcased in the project.

Which resources did I use:

Nothing new worth mentioning.

Date of entry: May 14

What I've worked on:

I've dedicated a lot of time to the project, and without recapping the project itself, I derived some results that led to new scientific questions and significantly shaped the project. Since I quickly concluded that women live longer than men, I decided to explore another question: "Do people in general live longer and longer?". I intended to attend the Lab class to learn about GitHub, but the school has been shut down due to demonstrations. Although I'm satisfied with the effort I've put into the project, I was concerned that it didn't showcase enough of my new skills, as I primarily used Bokeh as a styling package. Despite having some interactive elements in my project where Bokeh was helpful, I decided to incorporate one more skill into my project to further support my results. Therefore, I explored a statistics package to assess the statistical support for my findings.

What problems I encountered:

Learning a new package is still time-consuming, especially finding the right one. Despite knowing that I was going to explore a statistics package, I still needed to find one that I could understand and that would be useful for my project.

What I learned:

I learned how to use the stats package from SciPy.

Which resources did I use:

https://docs.scipy.org/doc/scipy/reference/stats.html

Date of entry: May 15

I just want to wrap up my project with a final note for today. I completed my project, including writing the text, refining code to ensure it looks polished and aligns with what we've learned in the course. I wrote a conclusion and the project description. Likewise, I corrected this logbook for grammar mistakes.