Name: Nigel Tatem

Date: November 3, 2023

Subject: Data Analysis with Python

Source:

freeCodeCamp. "Data Analysis with Python - Full Course for Beginners (Numpy, Pandas, Matplotlib, Seaborn)." Youtube, 15 April 2020, https://youtu.be/r-uOLxNrNk8?si=vLDyg4dFjuujWECe

Annotations

Assessment:

After creating the original work proposal I laid out a path to creating my final product: DEEP AI (Data-Driven Efficiency Enhancement Program & Artificial Intelligence). My passion has become Data Science as stated in my last research assessment, driven by my connections with businessmen like Bradley Shogren at UnitedHealth helping me see the value of information. The first part of my product requires using my knowledge of programming and independent research to create efficient data analysis for large quantities of data. Thus, I found the proper way to approach this is through Python, one of the most simple coding languages with a very high learning curve. I came across its ability to visualize, chart, and sort massive amounts of data with little to no processing time in a 4-hour video.

The video covers reasons why you should use Python over software like Excel, Tableau, and other auto-managed tools. First off it states that these softwares are closed-source: underlying source code is not made available to the public; In this model, the software vendor or developer retains exclusive control over the source code and restricts access to it. Users of closed source software can typically only access the compiled executable version of the software, which can be run on their computers or devices. This is a technical explanation but put simply this form of software is bad for four main reasons. 1. Limited access to source code means users cannot view, modify, or redistribute the source code of the software making your modifications limited, transparency limited, and you are at the will of the company that owns the program. 2. Licensing and usage restrictions in which the users have to pay a license that costs companies a lot of money compared to Python which is free, also the terms of the license have to be followed by the user further restricting them and are subject to change meaning its unknown what price you will be paying two years down the line. 3. Proprietary nature meaning it is only controlled by one entity relying solely on them for updates, bug fixes, and support. 4. Lack of transparency, meaning users have limited insight into how the software operates internally, which can make it difficult to verify security, privacy, or quality aspects of the software.

The videos in depth explanation on the drawbacks of entity owned software opened my eyes to the SPECIFIC reasoning on why learning Python will be so advantageous as a foundation for data analysis. Although it is harder to learn, the best things in life often are harder to achieve. The video talks about how Data Analysts who know how to program in SQL or Python have a 20% salary increase compared to those who do not, a HUGE difference. This is due to the extreme efficiency boost they provide and the fact that most companies utilizing data science use these languages if not entity-owned software. Although Python is harder to learn than a pre-formatted interface, it can be simple and intuitive compared to other languages like Java. It also contains powerful libraries available to all due to its open source nature, also it is free to use and has an amazing community with endless documentation and conferences available. Each of the libraries used to create data analysis charts, and program features to extract data are free and listed in the video: pandas, matplotlib, numpy, seaborn, statsmodel, scipy, scikit-learn.

Finally the source covers the 5-step process which although in order of how it is done, is usually a circular process that involves jumping back and forth between actions in order to perfect the final product. The first step is Data Extraction in which you scrape data, import file formats, deal with API's to have databases merge, purchase data, and use distributed databases. The second step is Data Cleaning where you find missing values, input data, incorrect types or values dealt with, outliers and non relevant data removed, and statistical sanitation. The third step is Data Wrangling in which data is ranked in Hierarchical fashion, categorically, reshaping structures, indexing for quick access, merging and combining data. The fourth step is analysis: This is when the exploration of data occurs, with Python you build quick statistical models of the data to more easily understand, visualize and representations, correlation vs causation, hypothesis testing, statistical analysis, and reporting. Finally is action, this is the Artificial Intelligence part. Build Machine Learning models, feature real Engineering, Live Dashboard and reporting, and decision making and real-life tests.

I consider this to be my most useful source yet guiding me on my journey to create the final product DEEP AI. My research will likely continue in topics discussed in the video. Each day will be taken step by step and the hardest challenge for now is what I do not yet know about Python.