CS 215
Nov 28 2023
Lucas Zheng

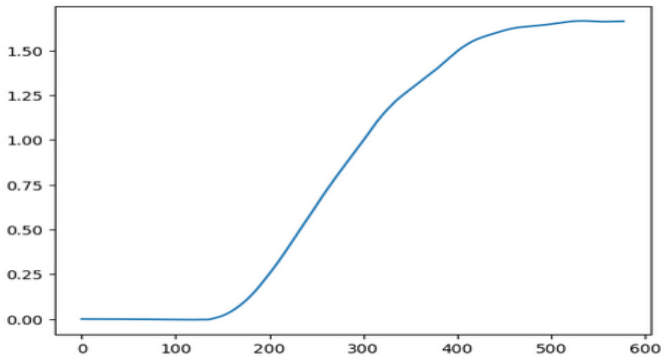
Data Manifesto

Throughout the semester, I have learned a lot of skills about data science, and I am more than excited to continue my future career with a data science major. And there is my thinking of 4 important principles when dealing with data and learn how to be a good data scientist.

Principle 1 is to understand the Data. Not just the general definition of data and what can be a data in data science world. In “Raw Data Is an Oxymoron” by Lisa Gitelman [2], she states her opinion that “data are always already “cooked” and never entirely “raw.” To me, Data is not just raw information, but it is the foundation of building a huge dataset with information, knowledge, and insights. Drawing inspiration from the DIKW pyramid, I recognize that data transforms into information in different ways. And then the information becomes human-readable and digested into knowledge through interpretation. Finally, with enough knowledge, they will evolve into wisdom through application and experience. Through Jill Lepore's article “The Data Delusion” [3], I realized the importance of organizing and categorizing data to unlock its true potential. The article reinstated the need for a systematic approach to handling and structuring data and that helps to make data easily accessible and meaningful. In my opinion, the data are raw information that is ready to be interpreted and developed for human use. Usually, data is quantitative so it can be easier to interpret and analyze. Understanding what Data means is very important for being a data scientist.

Principle 2 will be Understanding the Data Collection and Data usage, which will involve the process of collecting and searching for data in a variety of ways. Reflecting on our class projects, where the sourcing and gathering of data shaped our analyses, this principle underscores the significance of precision in data collection. Just like Project Linear Motion, which involves a self-measure dataset. Knowing how to measure the data is very important so that I can make the data I extract to be accurate. The detail such as moving slowly with the motion sensor at constant speed. Also, keep a steady hand. Throughout the interception process by coding, it also turns out that knowing the data-collecting process was important. The data set was based on time and only the last few seconds were my actual height. So, we did the average height throughout the timeline which made my height to be 60cm. After I revised the data collecting process, I only took the last few seconds and got an average of 166 cm. Which is way more accurate than 60cm. And the picture below would be the actual code from this example.

```
In [15]: df["z position(m)"].plot()
         # this graph make sense because it shows the phone movement by z vs time
Out[15]: <Axes: >
```




```
In [16]: # What is the final height? How would you find it?

         ending_z_positions = df["z position(m)"].tail(577)
         ending_z_positions
Out[16]: 1      -0.000007
         2      -0.000009
         3      -0.000013
         4      -0.000015
         5      -0.000018
                   ...
         573     1.663299
         574     1.663299
         575     1.663316
         576     1.663340
         577     1.663349
         Name: z position(m), Length: 577, dtype: float64

In [17]: last_20_z_positions = df["z position(m)"].tail(20)
         average_final_height = last_20_z_positions.mean()

         average_final_height

         #taking the last 20 positions average to get the average height to make it more accurate.
Out[17]: 1.6625066234516575
```

This principle involves understanding the data landscape, making choices about the data sources and avoiding potential distractions or biases inherent in the collection process. Another example from the homework is Project 4 Critically Examining a Dataset. The project required us to dive into a specific dataset and understand how it has been collected and where this set can be used. That reemphasizes how understanding the collecting process is important. Also, the data usage. When I was dealing with the government dataset with electric cars in WA, I knew how to generalize the result from the dataset, and which group am I looking at by analyzing the data. Examining a Dataset before using code to interpret it is very important. In that way, we can avoid potential bias and distraction from the crowded original dataset and make our next step more efficient.

Principle 3 would be how to Interpret the Dataset by selecting the right tool. This process is the most important part of data processing. Reflecting on our class projects, where we explored datasets using various tools like Seaborn, Pandas, and visualization techniques, this principle highlights the importance of adopting a multifaceted approach to data interpretation. When starting to interpret the dataset, the best approach is to interpret with questions. And then figure out which tool can help us to solve this question the best. That means finding the best tool for each specific data type. First, we need to set up what questions we want to investigate. From Project 6 APIs and Web Scraping, Geospatial data. I was looking for what star has the longest distance from Earth or other planets. With that question, I looked at the data source, which is an API from NASA. To analyze the API, the best tool is to import requests and use the API link. And then using Panda and Jupiter Notebook to manipulate the data obtained from the API and get the desired outcome. From the same project, I also did web scraping. The best tool will be ‘Beautiful-soup’. Other tools like Panda or SQL may not be helpful. Selecting the correct tool is

important for data scientists to analyze the dataset efficiently. Also for project 7, SQL and Geospatial data, with a different purpose, a different tool is needed. To plot the Professor's location on the US map, we used the tool called Observable. That helps me with visualization. I can perform different data analysis tasks with different tools, sometimes data works better on Jupiter Notebook and sometimes it may be more efficient in online Google SQL. Understanding different data analytics requests and being able to perform the most efficient analysis method is important to a data scientist.

Principle 4 will be Human-centered Data. Able to distinguish biased data and protect the data. Data analysis and visualization can solve a variety of problems. However, the data is not neutral. Data science is deeply intertwined with human experiences, biases, and perspectives. The reading "Data Science to Human-Centered Data Science"[1], states that" algorithms reflect the choices made by their human developers, including conscious and unconscious biases. What's worse, algorithms may amplify these biases, make them less transparent to other people, or make it harder to mitigate them." This quote underscores the data is not neutral. I think it is important to emphasize the responsibility of data scientists to approach their work with empathy and an awareness of the potential impact on individuals and communities when dealing with data and try their best to keep the data unbiased. The reading also says the self-reinforcement of biased data. Some data might come in with bias and have been thought of as correct, unbiased data and processed by data scientists repeatedly to make it standard data. As data scientists we cannot change how the data is processed before we get it, so we need to filter out the biased data and get the data as clear as possible. From the recent projects 7 and 8. I analyzed personal data from Professor Wirfs-Brock and my classmates. It is important to keep their data secure and not share it with others. Also, from the earlier project analyzing Whitman student GPA and class, we

need to have our moral standards and use the data in the right way. A moral standard can be different for different people. For myself, the most important part is to keep the rule of using the data. Not sharing private data on purpose and using them in legal and proper ways. Also getting the consent from data owner is important.

In conclusion, being a data scientist, it is important to dive into the data collection process and understand the use for each code package in order to use them more efficiently on the dataset that needs to be analyze.

Works of Citation:

1. Aragon, C. R., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). *Human-centered data science: An introduction*. The MIT Press.
2. Gitelman, L. (2013). *"Raw data" is an oxymoron*. MIT Press.
3. Lepore, J. (2023, March 27). The data delusion. *The New Yorker*. Retrieved November 30, 2023, from https://www.newyorker.com/magazine/2023/04/03/the-data-delusion.