Name: Henok Getachew Mulugeta

Class: Data Science

To: Professor Jordan

Data Manifesto

“Data is just a description of a thing’s qualities: a summary of a book’s plot, the price of a carton of milk, customer reviews of a neighborhood restaurant, the breed of a dog. What’s your name? That's data. Where do you live? That’s data too. What’s your favorite memory from childhood? Also data. Sometimes data is structured in a certain way—not just freeform like a story—but that’s familiar also.”-Melanie Feinberg, Everyday Adventures with Unruly Data. I totally agree with the idea that the quality of a particular object, animal, human, asteroid, president, water, element, enzyme could be data depending on different ideas. Any qualities or characteristics can be used as data that can be organized to answer a specific question and get us tangible information.

I personally think data is a collection of names, numbers, figures, shapes, faces, or anything that we use to analyze and find information. Any qualities, characteristics, trends, behaviors are data. For instance, as of now I am at Reid campus caffe 5/5/2023, 10:32 working on my CS project I am listening to music from Reid radio station, sitting at table with Tenzin and doing my CS project all this is happening at the same time could be data depending on question we ask or analysis we are making. We can find a lot of data from the previous example by asking different questions.

For instance, by just using the raw data in the previous example we can extract relevant data.

Questions: What do people do at Reid caffe?

data: Henok was studying.

Question: Who is Henok’s friend

Data: Tenzin

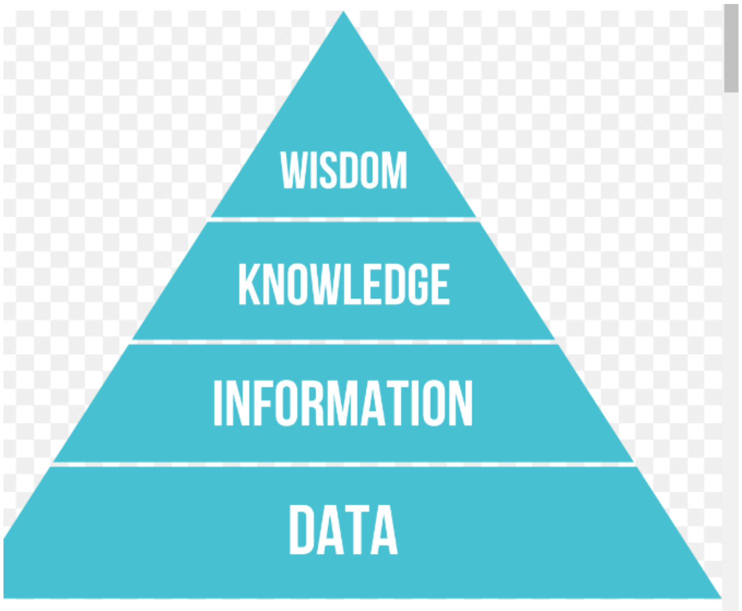
Question: are both Henok and Tezin international students?

Data: Yes

Question: Do international students tend to get along easily with each other at Whitman college?
Data: We can say there might be correlation between international students getting along with one another, it applies for Tenzin and Henok.

As we see above by asking different questions, we can arrive at different conclusions and data points. Of course, to make constructive conclusions we need to analyze more tables at Reid caffe. Moreover, data collection depends upon questions we want to answer, goals we want to reach. A person trying to lose weight may consider food, workouts, sleep hours as data points to reach a goal the person wants to accomplish. Generally, data needs to have a specific purpose, aim or questions it needs to answer and also may change depending on the purpose.

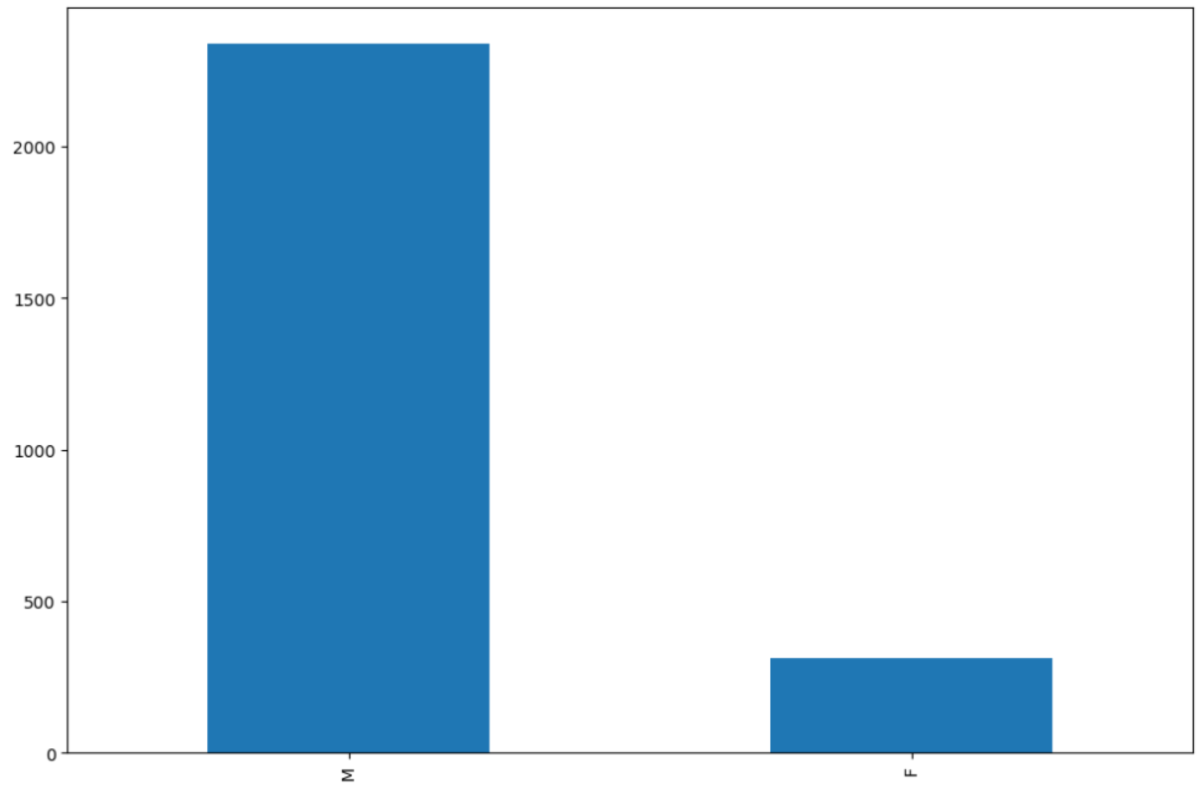
Jill Lepore explained mysteries, something we can't explain and facts as things humans can prove by way of observation, detection, and experiment such as I have one heart and two kidneys etc "Numbers", holds censuses, polls, tallies, national average such as 8 billion people live in the modern world today. Information is an organized or a collection of data or to knowledge and facts that have been communicated, received, or discovered. Knowledge is an organized form of information that we can use to tackle a problem and also solve a particular problem. Data is raw and unorganized making it unique from any facts, knowledge, or Numbers which are far more organized than data. Despite seeming similar, data, knowledge, information are different in terms of complexity and also applicable to solve particular problems or help us in decision making.



. As we see in the diagram above, Data is a basement for information, knowledge and also wisdom. With data there could be no information, knowledge or wisdom. As cells in biology are the building blocks of life, data is the building blocks of decisions we make, information, knowledge, and also wisdom.

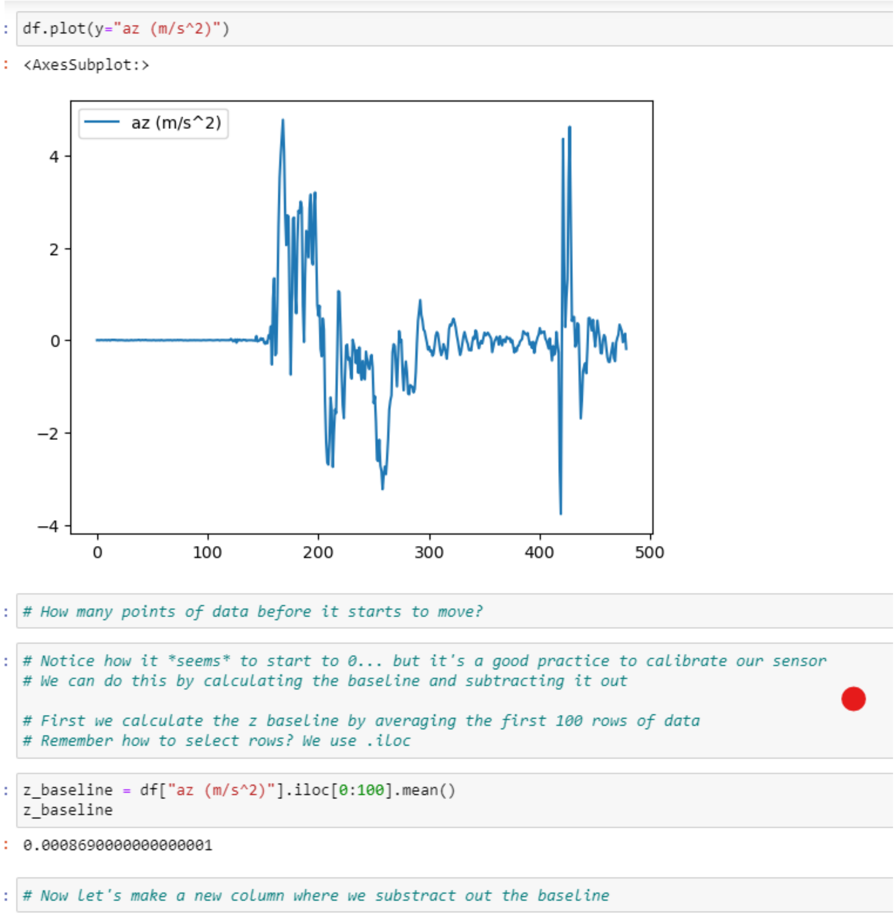
A data scientist finds appropriate or legitimate dataset then tries to find possible limitations of data might or could have by looking at how the dataset was gathered or tries to see any biases a dataset has. Then, answers specific questions out of a dataset or extracts knowledge and facts from a data set on my last project. I did an analysis on billionaires dataset from forbes in which I tried to see the age of those billionaires in which most of them are found to be older than fifty years old. I have also tried to see the number of billionaires in terms of gender. There are many male billionaires which shows the patriarchy in the world we live in right now.

```
<AxesSubplot:>
```



The graph above shows the number of female and male billionaires in the world that can reflect the inequality women face in the modern world we are living in today. Moreover, from the data I found out that some billionaires from my country wasn't included on the Forbes billionaires list. This mainly correlates with the Data feminism book and the story of Cristine Darden. Cristine Darden wasn't given the chance to work in engineering until she took a risk to talk to her boss due to the stereotypical ideology of what women were supposed to be doing. We could see what this stereotype does to the society reflected on the graph above creating wealth inequality in the society. Data feminism also highlights how data could perpetuate stereotypes about marginalized or low income individuals. I totally agree with the idea and I found out that Forbes didn't mention billionaires from my country, Ethiopia, despite Ethiopia having wealthy billionaires. According to Sofia Noble, "Intro to algorithms of operation" highlights problems in relation to algorithms or machine learning that can perpetuate the current biases and discrimination that the world has based on race, gender, ethnicity etc. I believe the root cause for this particular problem is the inaccurate dataset we have, such as the Forbes billionaires dataset that we discussed earlier, that can be used to build algorithms and AI which will be based on inaccurate dataset leading to inequality and injustice. Therefore, an accurate dataset will solve this ripple effect problem. Moreover, we can use data nutrition to highlight the flaws that our data has so that people using the dataset can know the flaws of the dataset.

Data science requires programming skills in which in the class we used python pandas to work with a dataset. We need to know relevant dataset formats, such as CSV, Excel, or HTML to extract or manipulate the relevant information. After finding the dataset, we then need to clean the data. Data cleaning is the process of identifying and correcting errors that are found in a dataset.



```
: df.plot(y="az (m/s^2)")

: <AxesSubplot:>
```

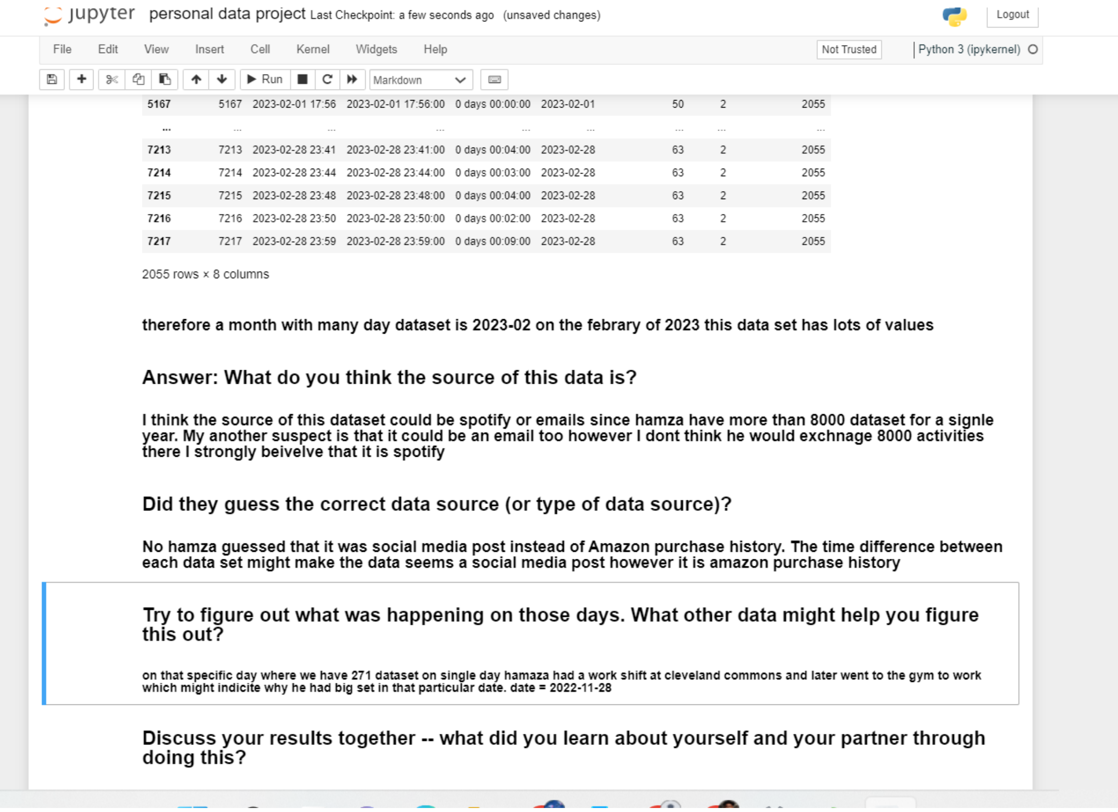
```
: # How many points of data before it starts to move?

: # Notice how it *seems* to start to 0... but it's a good practice to calibrate our sensor
  # We can do this by calculating the baseline and subtracting it out

  # First we calculate the z baseline by averaging the first 100 rows of data
  # Remember how to select rows? We use .iloc

: z_baseline = df["az (m/s^2)"].iloc[0:100].mean()
  z_baseline

: 0.0008690000000000000001

: # Now let's make a new column where we substract out the baseline
```

In the example above in which we calculated the height of the white board integrating acceleration into velocity and then velocity into distance in which here we calculated the baseline for acceleration and substrate the baseline from acceleration as we see in the picture above in which this will help us clean the dataset that we have.

We also need to know how to import relevant modules such as pandas, matplotlib, requests etc. We can make different analyses, and finally present our findings in a graph using data visualization. Data visualization is the ability to clearly communicate insights through graphs, charts or anysort of visualization. Data visualization can be helpful in presenting results, correlation, or comparison between data.

I recommend people to start projects and fail as many as possible. We can mainly learn data science through building projects and facing problems then finding solutions to those problems. Currently, chatGPT is the main resource for coding problems. I recommend people to use chatGPT wisely. Never use chatGPT as a way to escape a particular problem or find a shortcut for a problem. Utilize chatGPT as an instrument of learning when facing a problem, try to understand the solution chatGPT comes up with and apply it using your own understanding. I generally recommend using chatGPT for problems similar to stackoverflow websites.

We can make comparisons using matplotlib bar graphs. We can use bar graphs to make comparisons between two quantities. For project nine, I tried to present self made billionaires for a few selected countries in which the stacked bar graph can help us present which country has most self made billionaires. We can also use graphs to show correlation within a dataset. We can summarize results or detect a pattern using a data analysis. Data analysis can also be used to find a solution for a particular problem or draw cause and effect conclusions or relationships. For instance in the medical field during drug trials people use data to determine if a drug works or not. We also do trend analysis and figure out a particular trend and patterns could be identified in a dataset.



5167	5167	2023-02-01 17:56	2023-02-01 17:56:00	0 days 00:00:00	2023-02-01	50	2	2055
...	...	...	...	...	...	...	...	...
7213	7213	2023-02-28 23:41	2023-02-28 23:41:00	0 days 00:04:00	2023-02-28	63	2	2055
7214	7214	2023-02-28 23:44	2023-02-28 23:44:00	0 days 00:03:00	2023-02-28	63	2	2055
7215	7215	2023-02-28 23:48	2023-02-28 23:48:00	0 days 00:04:00	2023-02-28	63	2	2055
7216	7216	2023-02-28 23:50	2023-02-28 23:50:00	0 days 00:02:00	2023-02-28	63	2	2055
7217	7217	2023-02-28 23:59	2023-02-28 23:59:00	0 days 00:09:00	2023-02-28	63	2	2055

2055 rows × 8 columns

therefore a month with many day dataset is 2023-02 on the febrary of 2023 this data set has lots of values

Answer: What do you think the source of this data is?

I think the source of this dataset could be spotify or emails since hamza have more than 8000 dataset for a signle year. My another suspect is that it could be an email too however I dont think he would exchnage 8000 activities there I strongly beivelve that it is spotify

Did they guess the correct data source (or type of data source)?

No hamza guessed that it was social media post instead of Amazon purchase history. The time difference between each data set might make the data seems a social media post however it is amazon purchase history

Try to figure out what was happening on those days. What other data might help you figure this out?

on that specific day where we have 271 dataset on single day hamaza had a work shift at cleveland commons and later went to the gym to work which might indicite why he had big set in that particular date. date = 2022-11-28

Discuss your results together -- what did you learn about yourself and your partner through doing this?

For example, in the personal project we did, I used Hamaz's dataset and noticed a pattern. The pattern was Hamza's dataset varies with 3-5 minutes and also Hamza dataset has many data points which made me come to the conclusion that Hamza dataset is from spotify.

As we have seen throughout the semester, data is not neutral and data science always has people behind it. Thus, you should also reflect on what you, as a person, bring to your work as a data scientist.

Generally, we should have a specific question or goal then finding an accurate dataset should be our first priority when we do any data analysis. Then, we can look at the dataset and try to answer our question in many cases. After looking at the dataset, we come up with further questions and correlations to be made. We can finally show this correlation or result in terms of graphs, charts, or any visualizations