Data Wrangling

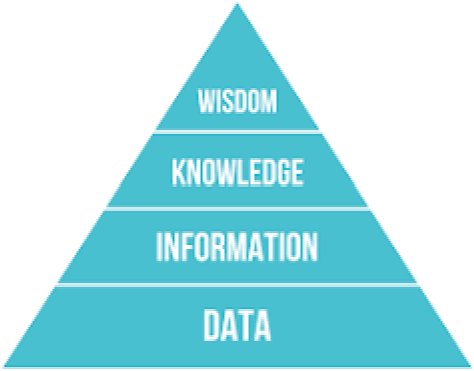
By: Luca Viarengo

Introduction

In the spring of 2023, I signed up for my first ever data science course at Whitman College. I didn't really know what to expect from the course, as I had worked with data briefly in the past, but I knew that there was an entire world of data wrangling that I had yet to discover. I could not have been more correct in that thinking, as over the course of the semester, I learned what working with data is truly like. It is not a very easy process to wrangle data, and answer questions using data, as you will find out in reading this data manifesto. There are some moments where you need to scour the web for answers on how to use certain python tools, like pandas, matplotlib, etc. There are some moments where you will spend hours creating a certain function to make sense of your data that you expect should work, but then you get so many unexpected errors. Data wrangling and visualization can be a long process to get the results that you want, and you might even find some completely different trends that you weren't expecting to see. I'm writing this data manifesto to show how I define data, what principles I follow when working with data, some important tools I've learned over the course of the semester, and examples of functions, data manipulations, and visualizations I have made.

Data

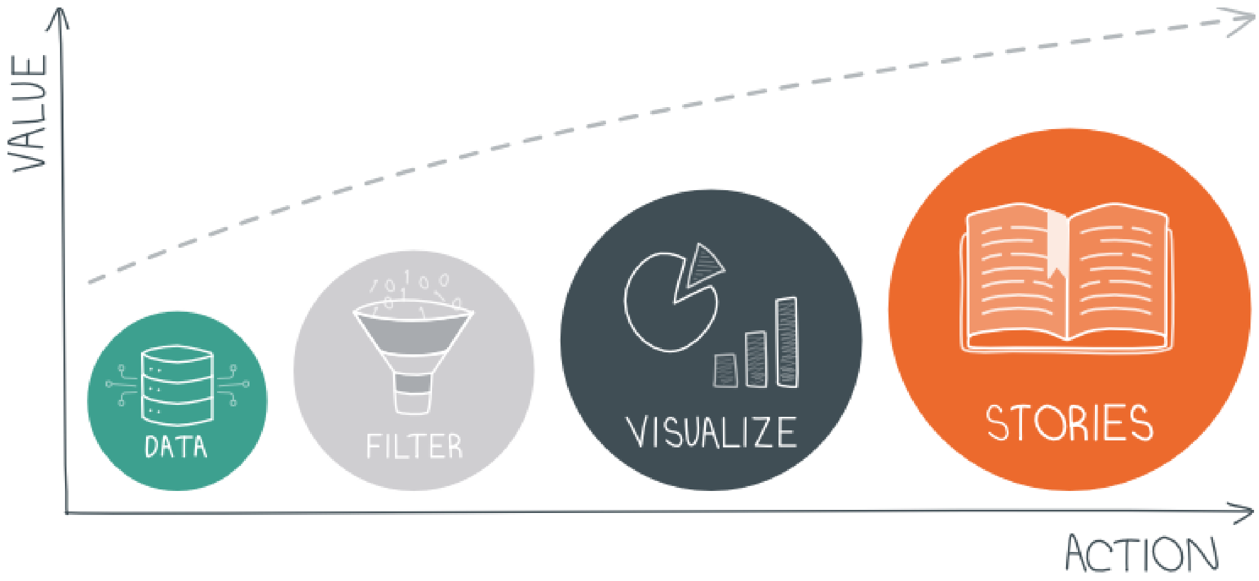
To begin, let's define what data means to me, because contrary to popular belief, data is not a term with a static definition. In general, whenever people see the phrase "data concludes", or a similar phrase to that, they think that it means that whatever is being claimed by the "data" is a proven fact at that point. That is not true in the slightest. The definition that I've personally drawn from my semester working with data is this: Data is a collection of facts and information that, with scrutiny, proper manipulation, and more, can be used to answer questions or inform decisions. Data is NOT fact. Data is NOT knowledge. Data is NOT information. Data is a collection of those three concepts, but it cannot be treated as if it is a cold hard fact. This is reflected in what I think is the best example of what data really is, the DIKW Pyramid (Data, Information, Knowledge, Wisdom), where data is the lowest and biggest tier, while wisdom is the highest and smallest tier.



As shown here, data is abundant, but doesn't really provide anything on its own. Data that is collected and processed correctly will become information. Learning that information and knowing it turns that information into knowledge, and finally, becoming an expert with practice and persistence turns that knowledge into wisdom. This pyramid best describes my relationship with data, as I feel like data is **everywhere**, but it shouldn't be treated as **fact, information, or knowledge** on its own.

Being a Data Scientist

Data science is a tricky field, meaning that being a data scientist isn't a walk in the park. Being a data scientist means being someone who is willing to look through spreadsheets of numbers and text in order to decipher what is going on. Being a data scientist means you need to be able to manipulate the data you're given. But most of all, Being a data scientist means you need to be a storyteller. That might sound absurd at first, but it makes total sense. A storyteller takes a seemingly random set of words and sentences, and creates a story with meaning and value. A data scientist takes a collection of information and facts and creates meaning for that data in the form of data wrangling, data manipulation, and visualizations. A data scientist turns collections of numbers and words into something that other non-data scientists can understand.



It takes a multitude of skills to be a good data scientist. To start, a basic knowledge of computer science is extremely useful. You need analytical skills to be able to wrangle information out of the data. Furthermore, you need to be able to be persistent, as there are times when the code you worked so hard to analyze the data on doesn't work, and you need to be able to try and figure out what went wrong and how to fix it. A good data scientist is also able to explain the meaning behind the data, and what story that data is capable of telling. This is shown through clear and easy to understand visualizations, and strong explanations of those visualizations. Essentially, a good data scientist is a storyteller with a computer science background and the ability to persevere through debugging and data wrangling.

Personal Principles

When presented with data, how does one go about processing and manipulating it? Do you start with a question, then find/create data that best answers the question? Do you stick to your original hypothesis/question, or can you divert halfway through? These, along with many others, are examples of questions that data scientists ask themselves before deciding to start a data project of any kind. The answers to these questions help shape the data scientists personal principles. Each data scientist will have their own defining principles that guide how they work with data. A great example of this is Giorgia Lupi's Data manifesto called "Data Humanism". In her manifesto, she details what she thinks about when she begins a new project, creating a few key principles that she tends to live by when working with data. With that in mind, here are the main principles that determine what I do when I decide to start a data science project.

Question or Data?

Do I start with a question and find data that can best answer my question? Or do I find a cool dataset online and create a question based on that data? That is the first thing that I think about when deciding my newest data project. Throughout the course of the semester, it seems clear to me what my preference truly is: Find a cool dataset, and then draft a question based on that dataset. One of the most prevalent examples of this was my latest project. In this project, we were tasked with just showcasing our skills as a data scientist. There were no guidelines on what tools you could use, what dataset was needed, or what visualizations you had to do. It was a complete free for all as a data project. So I just went on data.gov to find a dataset that seemed cool to me. I eventually struck gold with a dataset that tracks all of the electric vehicles currently registered in Washington state. Scouring through the CSV file that the dataset was stored in, I saw that cities, makes and models of cars, as well as location data was present. This allowed me to think of both the questions that I would come up with that this data set could answer, as well as the visualization I could create with the location data. I initially drafted two questions that I thought the dataset could answer: “What is the most popular make of electric car registered in the state of Washington?” and “What is the city with the most electric cars registered?”. This is an example of me choosing my dataset first, then drafting a question based on that. I find that I’m able to come up with better questions if I choose based on data I am given, rather than having a question first, then finding data, since if I first draft a question, I might never find a dataset that truly answers the question. Not to mention how difficult it is to draft a question when there are infinite questions that I can come up with. It just makes more sense to me to find an interesting dataset and work from there

Biases

Another important principle that I think about when working with datasets is bias. There are two main aspects to be aware of when you think about inherent biases: your own biases, and the biases of the researchers who collected the dataset that you plan to work with. To begin, let's start with how I try to acknowledge my biases so that they don't seep into the data project that I'm working with. I first try to think about my own personal life, and if my own personal life relates to this dataset in any way. More often than not it does, given as I stated earlier, I tend to work with datasets that I find are cool, which typically is because it is about something that I care about. After I figure out what personal relationship I might have with the dataset, I then make sure to keep that relationship in mind whenever I start wrangling the data. The most common question that I ask myself is: "Is the data truly coming to this conclusion, or do I WANT the data to come to this conclusion?" A great example of this is when I asked for my personal data from Instagram, then needed to extract timestamp data from my personal data.

Given that this was my personal data from Instagram, I knew I couldn't avoid the personal relationship I had with the data, so when I decided to get my timestamps from my comment history, I knew I couldn't really factor in my comments in my data. Just to make sure of that, I deleted the column that held what was in each comment, and focused purely on the timestamp data. This example shows the steps that I try to take to make sure that my personal biases don't affect the work that I'm doing.

Unfortunately, personal bias is not the only bias that needs to be taken into account when working with data, which brings me to the biases inherent in the data itself. This bias is much harder to spot because whenever you come across a dataset you might initially think that the numbers can't have a bias because they are just numbers that are collected and organized. However, how they were collected, how they were organized, and who/what was behind it all can introduce loads of biases into the dataset itself. How does that play out? Well let's look at an example that I found when looking for a dataset for one of the projects during the semester. I found a dataset that tracked both milk production, consumption, and general health of communities. Now I initially was interested in looking into maybe seeing if there was a correlation between health of a community and consumption of milk, but then I thought to myself, why was this data collected? I went to the author's page, and a few hyperlinks later, I found out that the dataset was collected by a pro-milk agency. So my general assumption was that they might be cherry picking where to find consumption data and health, and even skewing what can be considered as "health" to make people think that consuming milk causes better health. The numbers of the data set initially seemed unbiased, but a little bit of extra research shows a hidden agenda behind the numbers.

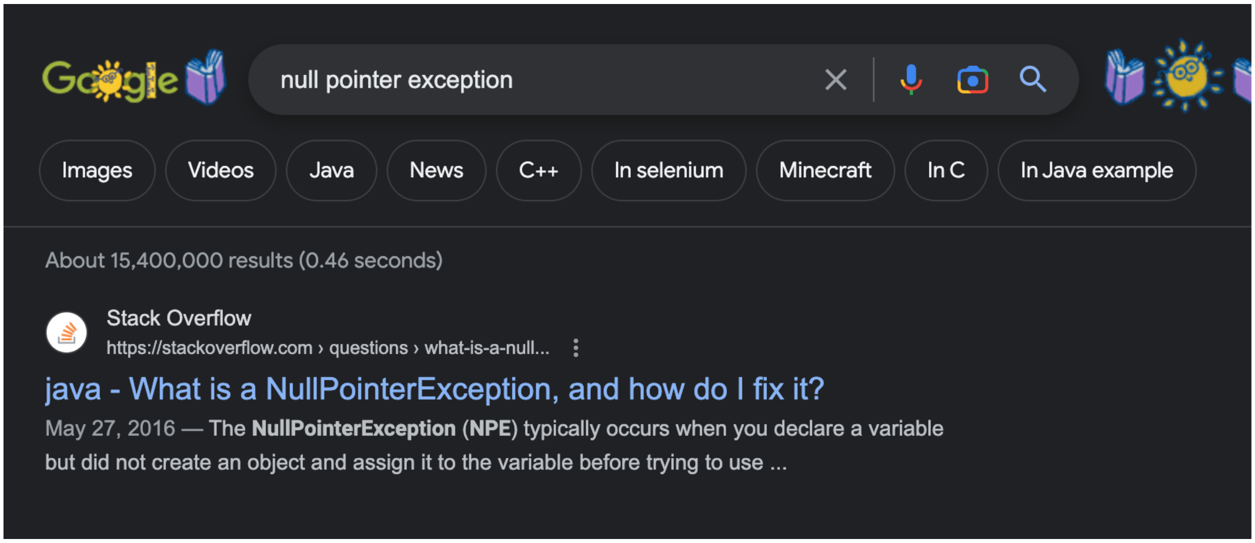
So I decided to move past that dataset and choose one that seems more unbiased. This checking of the dataset is integral to how I work with data, to make sure I'm working with as much of an unbiased source as possible.

Use your resources

Data scientists get stuck… a lot. I've gotten frustrated countless times because some code I wrote wasn't doing what I expected it to do. That's why one of the key abilities of a good data scientist that I mentioned earlier is persistence. Luckily, one of the easiest ways to get yourself out of a whole is to make use of the resources available to you. Resources like data science textbooks, a teacher/mentor who has more experience, or previous projects you have done can certainly help fix some of the problems that you may face when working with data. However, the biggest and most useful resource available to a data scientist is the **internet**.

Google

That may seem obvious but the internet is integral to getting past many of the problems that you face. Most likely, regardless of the problem you might face, someone, somewhere online has had the same if not a similar problem, and has the solution available for you to take and apply to your own project. People online have created multiple packages of code that make complicated, seemingly impossible tasks much easier. All it takes is a Google search and you can finally break through the problem that you were previously stuck on. Not only are you able to search up a problem, you can even search up specific errors that you get. For example, if you get a "Null Pointer Exception", you can search that specific line and see the results.



null pointer exception

About 15,400,000 results (0.46 seconds)

Stack Overflow
https://stackoverflow.com › questions › what-is-a-null...

java - What is a NullPointerException, and how do I fix it?

May 27, 2016 — The **NullPointerException** (**NPE**) typically occurs when you declare a variable but did not create an object and assign it to the variable before trying to use ...

Using the internet as a main resource can help you find solutions to many of the problems that you are stuck with, as long as you are willing to comb through code documentation. It's by far the resource that I have found most useful in my data science career so far.

Be Open Minded

The final main principle that guides my data science work is to be open minded. That principle can come in many different forms. Being open-minded about what dataset to use and not just sticking to the first one you come across is one way to be flexible in your data science learning. Using the internet as a resource when you get stuck, rather than trying to brute force your own solution, is another way to show your open mindedness, as previously mentioned in the earlier section. However the main one I want to discuss is to be open-minded with what you want to get out of a dataset. Many people, myself included during my first few data science projects, tend to find a dataset they like, and draft a bunch of questions from the start, and then lock in on a few questions for the rest of the project. That is a quite limited way to look at things, especially if the initial questions can't be completely answered by the dataset that was given. Instead, what I've learned to do is to be open minded while going through a dataset so that I can be flexible in case my initial questions aren't best suited for the dataset, or in case there's more I want to explore with the dataset.

The best example I can remember where I showed this open-mindedness is from the electric vehicles dataset that I had discussed earlier in this piece. I had talked about how I **initially** drafted two questions: "What is the most popular make of electric car registered in the state of Washington?" and "What is the city with the most electric cars registered?". However, as I continued to work with the dataset, I realized that there was another, even more specific question that this dataset, and my tools, could answer: "What is the most popular model of the most popular make of car in the city with the most electric vehicles registered in the state of Washington?”. I could have easily stopped at the first two questions, but I kept my mind open to new ideas, and found a new, more interesting story that I could tell from this dataset. That is why I believe it is vital as a data scientist to not close yourself off to new ideas and to remain flexible, as there can be even more interesting stories that can be told from data if you keep an open mind.

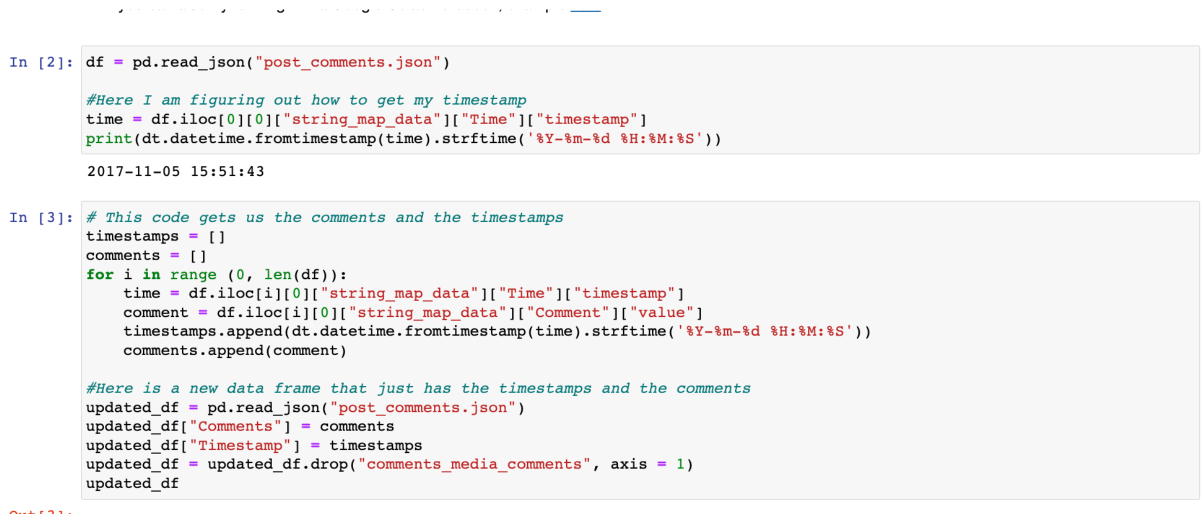
Tools

Data scientists would have quite a difficult time if it weren’t for tools that make wrangling and visualizing data so much easier. These tools increase optimization, efficiency, and productivity for the data scientist, as many of these tools automate what would otherwise be a cumbersome task for the data scientist to do on their own. Through my data science career, I’ve been able to be exposed to many different tools that have helped me process data. These are the few that I want to highlight as having been most useful to my projects.

Jupyter Notebooks

Data scientists need a program to give them a place to wrangle and visualize the code. One of the ones I was introduced to, and the main one used for the entire semester, was Jupyter Notebooks.

Jupyter Notebooks is a web based interactive computing program that allows you to create and share programs. Every class, our professor would send us a link to a jupyter notebook that she created, and once it was on our computer, we could edit it in any way we wanted, without affecting the original link that was sent to us. Jupyter Notebooks uses python as its coding language of choice, and allows you to import python packages directly from the internet. One package that came in quite handy for data wrangling and visualizations was the python pandas package, as it allowed us to read datasets like .CSV and .JSON format into a pandas dataframe, where pandas has easy to use functions to allow us to process the code. It also had rudimentary visualization tools like creating line graphs, which helped us get a taste of what data visualizations could look like. However, the feature that I really enjoyed above all else was the fact that you could run individual cells, and that these cells were self contained.

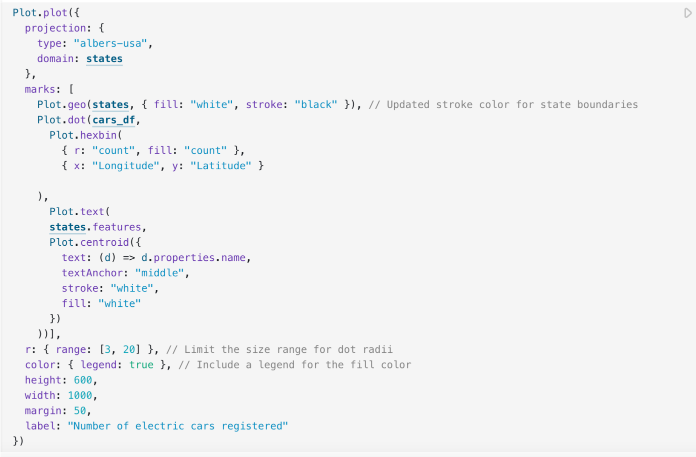


```
In [2]: df = pd.read_json("post_comments.json")

        #Here I am figuring out how to get my timestamp
        time = df.iloc[0][0]["string_map_data"]["Time"]["timestamp"]
        print(dt.datetime.fromtimestamp(time).strftime('%Y-%m-%d %H:%M:%S'))

        2017-11-05 15:51:43

In [3]: # This code gets us the comments and the timestamps
        timestamps = []
        comments = []
        for i in range (0, len(df)):
            time = df.iloc[i][0]["string_map_data"]["Time"]["timestamp"]
            comment = df.iloc[i][0]["string_map_data"]["Comment"]["value"]
            timestamps.append(dt.datetime.fromtimestamp(time).strftime('%Y-%m-%d %H:%M:%S'))
            comments.append(comment)

        #Here is a new data frame that just has the timestamps and the comments
        updated_df = pd.read_json("post_comments.json")
        updated_df["Comments"] = comments
        updated_df["Timestamp"] = timestamps
        updated_df = updated_df.drop("comments_media_comments", axis = 1)
        updated_df
```

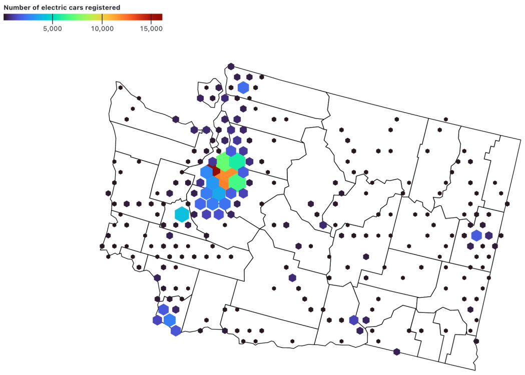
As you can see in this image, there are two separate blocks of code. I can run the first block without running the second block, and it will output whatever I need from the first block without affecting the second block. I find that really interesting because typically, in other programs that allow you to run code, you need everything within a class to work in order to run the program without bugs. In this case, you can run individual cells and fix the bugs self contained in that cell before you move on, which I find really useful in data wrangling. I can also test out certain outputs, and show what my data looks like using these self contained code blocks. I can even change the type of the code block to instead be a text-style block to add some explanation for certain sections of the code. Overall, Jupyter Notebooks has been a really useful tool for data manipulation, and in the future, I will probably choose Jupyter Notebooks anytime I work with data.

Observable

Observable, like Jupyter Notebooks, is another useful tool for data scientists to use. While Jupyter Notebooks focuses more on the data manipulation side of being a data scientist, Observable's main appeal to data scientists is its ability to create beautiful data visualizations. When you first open up Observable, you get the small sense that it's like GitHub and Jupyter Notebooks combined. You have the ability to run code in individual blocks, like Jupyter Notebooks, but you also have a central hub where you can access all your notebooks with a similar feel and style to GitHub. Using JavaScript as its language of choice, Observable allows you to create stunning visualizations, especially when it comes to mapping data onto a place in the world. Specifically, it has a Plot.geo function that allows you to map the world in many different styles, then add data to plot onto the created map. I must admit, the first time I saw and tried to work with Observable, I was a little overwhelmed and didn't really understand how it worked. However, in my final project, I decided to try it to get more familiar with the tool, and test my limits a little bit, which resulted in the data and visualization that you see below.



```
Plot.plot({
  projection: {
    type: "albers-usa",
    domain: states
  },
  marks: [
    Plot.geo(states, { fill: "white", stroke: "black" }), // Updated stroke color for state boundaries
    Plot.dot(cars_df,
      Plot.hexbin(
        { r: "count", fill: "count" },
        { x: "Longitude", y: "Latitude" }

    ),
      Plot.text(
      states.features,
      Plot.centroid({
        text: (d) => d.properties.name,
        textAnchor: "middle",
        stroke: "white",
        fill: "white"
      })
    ))],
  r: { range: [3, 20] }, // Limit the size range for dot radii
  color: { legend: true }, // Include a legend for the fill color
  height: 600,
  width: 1000,
  margin: 50,
  label: "Number of electric cars registered"
})
```



These images show a snapshot of what you can do in Observable. After running this block of code I was able to map on my data from an imported CSV file to the state of Washington, and show density using the hexagonal feature that exists with the Plot library, creating the map you see on the right. It may look really complex at first, but once I spent the time to actually look through each part of the code, it actually made a lot more sense to me. Overall, Observable gives data scientists the liberty to create stunning visualizations that help drive the story that the scientist is trying to tell.

Conclusion

This was quite a long data manifesto, which I honestly didn’t expect going into writing this. However, it just goes to show how much one can learn about being a data scientist in such a short amount of time. You learn what data means to you, how to actually be a data scientist, what principles guide you in your decision making when it comes to visualizing and manipulating data, and what tools are essential to have a successful career as a data scientist. For me, it was clear that data was not fact, information, or knowledge, just a collection of them, and that on its own, it doesn’t mean a whole lot. In fact, it should never be treated as a fact, knowledge, or information, and instead should be carefully combed through to turn the data into useful information. The people who are able to turn data into something actually useful are data scientists, whose job is to wrangle the data and make sense of it. Data scientists end up having to act as storytellers with their data, grabbing somewhat connected pieces of data and telling a palatable story out of it that is presentable through visualizations. Every data scientist has guiding principles that shape their process in how they go about working with data. As a new data scientist myself, the work I did this semester helped shape the key concepts and ideas that I think about whenever I begin a new data project. The 4 main ones that I narrowed it down to were whether to start with the dataset or a question, be aware of the biases inherent in myself and the dataset itself, use your resources, ESPECIALLY, the internet, as much as you can, and be open minded as you go through the project, as you might discover something even more interesting than when you first started. However, these guiding principles would be hard to implement if it weren't for useful data science tools like Jupyter Notebooks and Observable to help make projects go more efficiently and create stunning visualizations. Overall, I learned a lot about how to be a data scientist from this semester of being immersed in data science projects, and I don’t plan on quitting now. After all, what’s the point of acquiring all of these tools and principles if not to actually do something with it down the line?