Data Manifesto

# Data science is a relatively new field, yet the insights gained from data analysis have been marketed as pillars of truth in today’s digital age. Even with the widespread use of data science, there is no official data scientist code of conduct. As a new data scientist, I must face the following questions: What is data, and how do I work with it? Data Science 215 introduced me to these questions and several approaches to answering them. After being exposed to all these unique ideas, I’m left with the question: which do I incorporate? How do I act going forward as a future data scientist? Going forward, I’ve decided to remember four main principles.

# Data is Human

One of the primary reasons people have come to trust and rely on data so heavily is its perceived inhumanness. Conclusions drawn from numbers can seem like conclusions made without bias and personal agendas, but this assumption is untrue. Humans are involved in every step of the data science process. Humans collect, store, review, and analyze data, and they are affected by the products of data. Therefore, data is influenced by subjective decisions at every step.

During project 5, I worked with hierarchical and k-means clustering to analyze parks and recreation data from U.S. cities. Working with both types of clustering methods helped solidify the large number of subjective decisions data scientists make. I made decisions, which included educated guesses, about which clustering method would be best for the dataset, which normalization method to use, which distance matrix method to use, the number of clusters to make, the visualizations to use, and what further exploratory analysis I should do. A screenshot of a computer

AI-generated content may be incorrect.

Figure My decision to use the ward method to create the distance matrix.

A screen shot of a computer

AI-generated content may be incorrect.

Figure My decision to create 7 clusters after looking at an elbow graph with no distinct "elbow."

Data scientists' decisions around data collection, wrangling, analysis, and visualization affect their final result. Even if data scientists choose what they believe to be the best option rather than the easiest, it's impossible for personal bias not to be incorporated. During Project 7, I worked with a partner to request and analyze our personal data. Although people are allowed to control some settings that restrict data collection and can request to download their personal data from apps and accounts they use, data is still constantly collected about them. For example, I used the IP address data Google collected from my personal account, which recorded "logins" or "logouts" to the account. Google's "Privacy and Terms" section of their policies explains that they use location information for relevant advertising, location-relevant results, and security. Much of companies' collection and use of user data often remains unseen, but that does not mean their users are unaffected. During my project, I was careful when sharing my IP address data. Google, however, likely does not take the same precautions I do.A screenshot of a computer program

AI-generated content may be incorrect.

Figure In project 7, I removed parts of my code including my initial analysis of my IP address data to prevent my IP address from being shared.

In the future, I will strive to provide transparency on everyone involved in the data science process. By acknowledging the people behind the sources I use and my subjective decisions, I can enforce accountability. Furthermore, considering who will be affected by a product I create, regardless of whether they are the intended recipients, will help me recognize and fight the moral or ethical concerns my products may create.

# Data is Messy

Most formats data can be found in (especially the formats non-data scientists see most often) seem organized and logical. Excel spreadsheets and statistical diagrams can feel more consistent than other ways people process, organize, and visualize ideas (such as art or a physical file cabinet). However, these ways of storing and presenting data often hide gaps. Since people are behind data collection and storage, their decisions can make analysis more challenging due to a lack of clarity, missing data, or even mistakes in datasets. For example, I worked with a college scorecard dataset in Project 2. I spent a good chunk of time during the project trying to understand the dataset I was working with, since the dataset's column names were obscure.A table of data with numbers

AI-generated content may be incorrect.

A white sheet with black text

AI-generated content may be incorrect.

Figure The lack of clarity due to column names made it challenging for me to understand how I could explore the college scorecards data set.

Regarding presenting data, the choices someone makes will inevitably highlight some insights while highlighting others. For example, my partner and I created three visualizations to compare our timestamp data in project 7. First, we made a date-time scatter plot. The scatter plot highlighted the differences in the range of days for when our timestamp datasets collected their data. A graph with blue dots

AI-generated content may be incorrect.

Figure A scatterplot visualizing my partner and I's timestamp data sets.

Our second plot was a pie chart that further highlighted the distribution in range and the differences in distributions of timestamp data across the two days.

A screenshot of a graph

AI-generated content may be incorrect.

Figure A pie chart visualizing my partner and I's timestamp data sets.

Finally, our overlapped histogram highlighted the differences in the number of total timestamps taken and the differences in the distribution of our timestamps across the time of day.

A graph of time stamps

AI-generated content may be incorrect.

Figure A histogram visualizing my partner and I's timestamp datasets.

Each visualization captured something new about our combined timestamp datasets. Therefore, choosing the types of visualizations you show an audience will affect their understanding of the data. They can only form conclusions based on the information given to them. Unfortunately, a couple of visualizations can't fully encapsulate most datasets.

In the future, I will clarify what is missing along with the findings I want to highlight. During the data wrangling step, I can explain how my data methods for cleaning and wrangling data altered the dataset and keep the original dataset available. I can also reference the original dataset and its contents in data visualizations so viewers know what data is excluded from the visualization.

# Data Is Fragile

With services such as the Cloud, data can feel permanent and safe. It is hard to imagine my data being stored in large, physical data centers. However, with the current administration deleting data from official government websites, I've realized how impermanent data is and will continue to be. Data storage always comes at a cost. Data requires people's desires and resources to be stored and kept. Therefore, data is vulnerable to monetary, political, societal, and even personal shifts. For example, being able to work with a dataset from the Department of Education during Project 2 was extra significant knowing of Trump's executive order to close the Department of Education, risking the vast amount of data the DOE collected and stored.A screenshot of a white page

AI-generated content may be incorrect.

Figure A press release statement on the DOE's official website (https://www.ed.gov/about/news/press-release/statement-president-trumps-executive-order-return-power-over-education-states-and-local-communities)

The availability of APIs is also subject to change. For example, the Trefle API I worked with in Project 6 is currently being restored by Mashum. Previously, the API’s code had been archived due to the lack of time and means to keep it open.

A screenshot of a computer

AI-generated content may be incorrect.

Figure The announcement from the Trefle webpage that Trefle would be returning (https://trefle.io/#:~:text=Students,is%20now%20backed%20by%20Mashum).

Now that I know how vulnerable data is, I can take extra steps to save data. Saving data sets in multiple places, such as Cloud services, internal and portable hard drives, and USB flash drives, can make data storage more secure. However, I also have to ask myself: What data is most vital to save? Like everyone else, I need time and other resources to secure data.

# Data needs Fluidity

With data being as subjective, messy, and fragile as humans, shouldn't data be considered multifaceted like humans? I've already seen people use data in many ways, from analyzing relationships and trends and creating predictive models to storytelling and art. With the diversity in the ways people can utilize data, the role and conduct of a data scientist shouldn't be limited. In the future, I've decided to tailor my conduct to the goal I wish to accomplish.