Data Manifesto: A Personal Reflection on Data Science

Introduction

Data is all around us and is an important component of our lives. Taking the introduction to data science class this semester, I have learned that data is much more than just numbers and figures. It is like the raw material that is used to build knowledge and understanding, which is the foundation that helps us to learn and gain insights about the world around us. My experiences in this class have inspired me to create this data manifesto that outlines my philosophy on data, data science, and how I approach working with data.

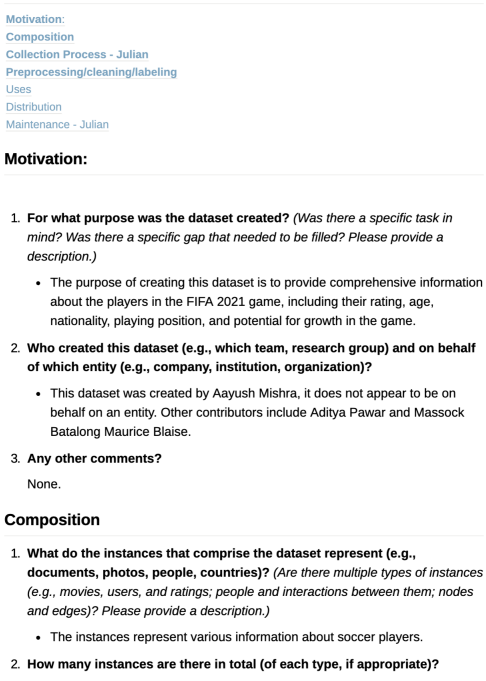
Data is not neutral

The starting point for data feminism, as D'Ignazio and Klein argue, is the recognition that power is not distributed equally in the world. This fundamental insight is also at the core of the principle that "data is not neutral." Data is constructed and shaped by the people who collect it, the questions they ask, and the decisions they make.

As a result, it is important to critically examine the sources of data, the methods used to collect it, and the assumptions and biases that may be present. We cannot assume that data is objective or unbiased.

In this class, I had the opportunity to work on `Project 4: Critically examining a dataset`, which involved critically examining a dataset of FIFA players. Throughout this project, I learned to ask various questions, such as which attributes were measured, the purpose of the dataset, how it was collected, who created it, and whether there were any errors or redundancies in the dataset. This has helped me gain a better understanding of the data, which

helped to avoid any subjectivity or bias when working with it.



Motivation:
Composition
Collection Process - Julian
Preprocessing/cleaning/labeling
Uses
Distribution
Maintenance - Julian

Motivation:

1. **For what purpose was the dataset created?** *(Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.)*
   - The purpose of creating this dataset is to provide comprehensive information about the players in the FIFA 2021 game, including their rating, age, nationality, playing position, and potential for growth in the game.
2. **Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**
   - This dataset was created by Aayush Mishra, it does not appear to be on behalf on an entity. Other contributors include Aditya Pawar and Massock Batalong Maurice Blaise.
3. **Any other comments?**

   None.

Composition

1. **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *(Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.)*
   - The instances represent various information about soccer players.
2. **How many instances are there in total (of each type, if appropriate)?**

A part of Project 4

Ultimately, the principle that "data is not neutral" reminds us that we must always question the sources and assumptions of our data and be mindful of the power dynamics that shape them.

Data visualization is a powerful tool for communication

Data visualization is a helpful way to turn complex data into clear pictures that are easy to understand. Data by itself may not be meaningful until it is arranged in a way that is visually attractive and easy to comprehend.

I worked on a project that used Google Map data, and it gave me access to a few attributes such as longitude, latitude, and timestamp. These numbers by themselves don't give much information and need to be organized to make sense of them.

		Unnamed: 0	lat	lon	timestamp
2	909677	909677	39.723636	-105.2389627	2022-03-09 00:16:38.646000+00:00
3	909678	909678	39.723636	-105.2389627	2022-03-09 00:52:39.544000+00:00
4	909679	909679	39.7236301	-105.2389794	2022-03-09 01:11:38.079000+00:00
5	909680	909680	39.723639	-105.2389794	2022-03-09 01:11:56.416000+00:00
6	909681	909681	39.723639	-105.2389794	2022-03-09 01:47:57.004000+00:00
7	909682	909682	39.723639	-105.2389794	2022-03-09 02:26:31.466000+00:00
8	909683	909683	39.723639	-105.2389794	2022-03-09 03:02:43.510000+00:00
9	909684	909684	39.7236405	-105.2389695	2022-03-09 03:09:34.299000+00:00
10	909685	909685	39.7236304	-105.2389792	2022-03-09 03:11:53.878000+00:00
11	909686	909686	39.7236304	-105.2389792	2022-03-09 03:16:49.414000+00:00
12	909687	909687	39.7236243	-105.2389721	2022-03-09 03:16:50.971000+00:00
13	909688	909688	39.7236506	-105.2389716	2022-03-09 03:17:41.961000+00:00

Google Map Dataset used on Project 7: Geospatial Data

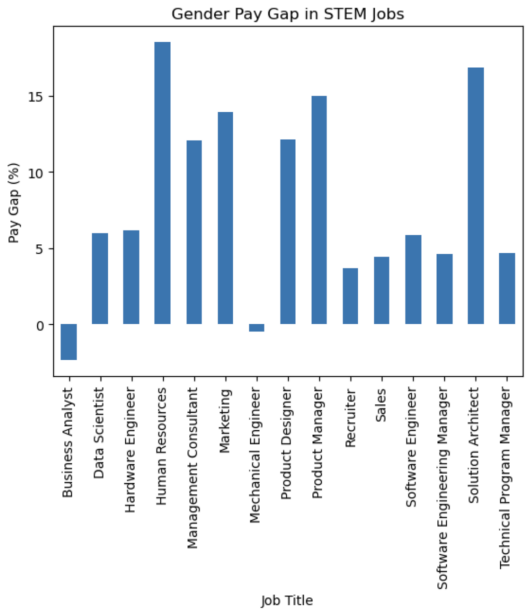
By using the longtitude and latitude data, we can plot each record on a map and see the location of the dots. And if we add the timestamp, we can also see when each dot was recorded. This makes the data much easier to understand.

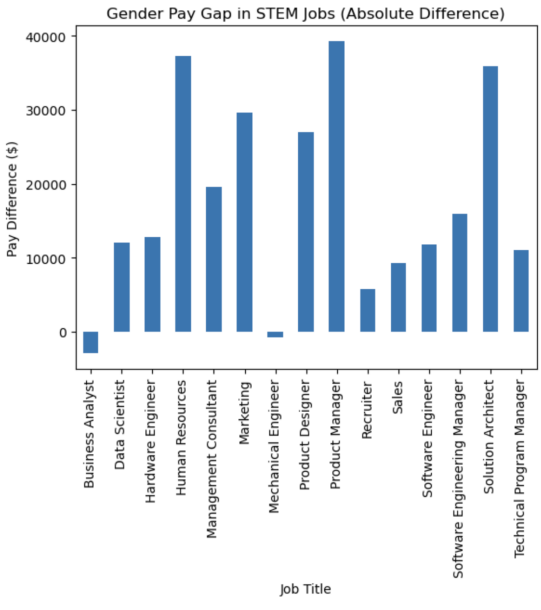
Raw data after being plotted on a map

With data visualization, we can even see patterns and compare different things to show the significance, and even use these visualizations to tell stories that inspire actions.

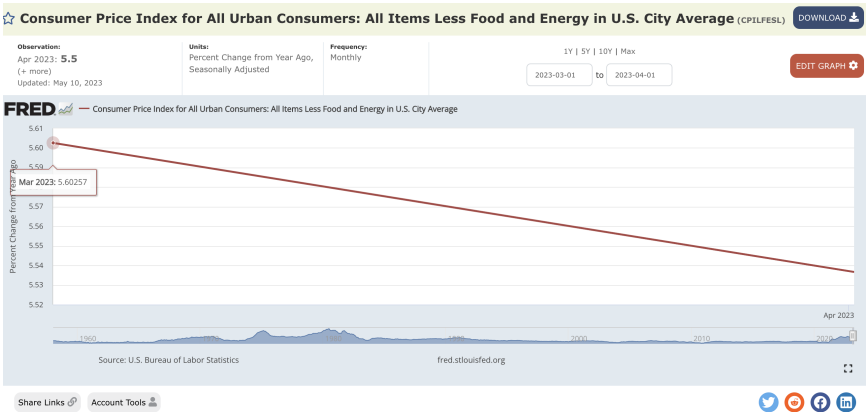
For example, when I worked on the salary dataset as part of `Project 9: A data analysis of my own` to identify if there's a gender pay gap in the STEM field, I used a bar chart to present my findings.

The bar charts clearly displayed that a pay gap does exist, with the difference amounting to as much as $40,000 in absolute terms or 20% in percentage difference. This visualization could be used to inspire action to close the gender pay gap in STEM fields.



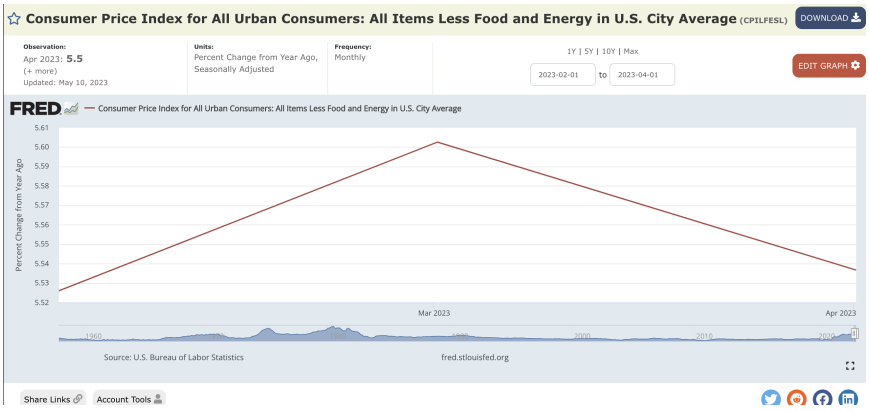


However, we need to be careful when creating visualizations by using appropriate methods to avoid misleading images. One example I have seen is the recent CPI report released a few days ago. The media has shared like a great news about the core CPI going down from 5.6% in March to 5.5% in April. They made it sound like a huge victory.



https://fred.stlouisfed.org/graph/?g=rocU#0

Looking at the overall time frame from February 2023 to April 2023, there has been very little change in the core CPI, which remained at 5.5%. However, some people may only focus on the slight decrease from 5.6% in March to 5.5% in April and overlook the lack of progress over the entire period. This could lead to a misunderstanding of efforts to fight inflation.



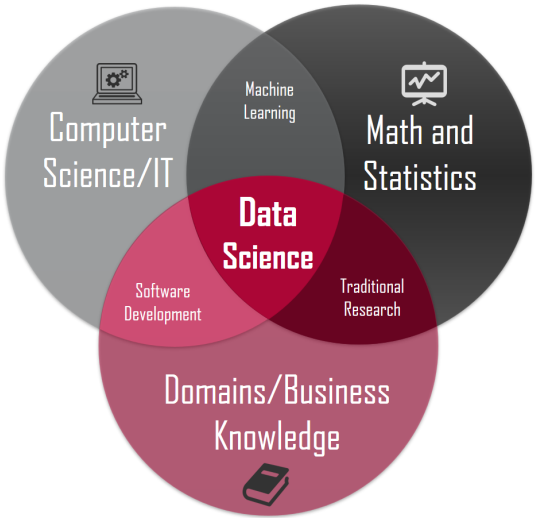
As a result, It's important to be aware of the limitations and biases in data and visualization, and work together to overcome them.

By using data visualization appropriately and responsibly, we can help people understand the world of data and work towards a fair and sustainable world.

Data science is a multidisciplinary field

Data Science is more than just about programming or statistics. It requires a mix of skills from different areas like Computer Science, Math, and specific knowledge in a field.

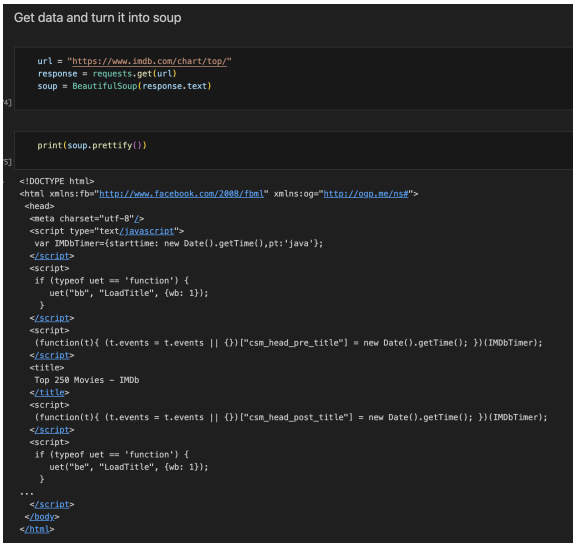
As a student who took the introduction to data science class, I have learned to be able to identify a problem, clean and prepare data, perform data analysis, and communicate my findings effectively.



https://towardsdatascience.com/introduction-to-statistics-e9d72d818745

An example when I applied these skills is when I worked on the `Project 6: API and Web Scraping.`

I would need to identify the problem I want to solve and understand the data I'm working with. In this case, I want to see if there's any correlation between the year the movie produced and its rating by looking into the top 250 movies. In order to do this, I would need to get the data from the IMDB website by using Beautiful soup and understand the structure of the data.



Once I've collected the data, I need to make sure it's correct and ready for analysis. This includes cleaning, organizing, and formatting the data so that it's easy to work with. To do this, I may need to use programming and problem-solving skills.

```
title_and_year_tags = soup.find_all("td", {"class": "titleColumn"})
rating_tags = soup.find_all("td", {"class": "ratingColumn imdbRating"})
```

We will now use a list of json (in python3 json is similar to a dictionary) to store the information of each movie, with the keys are title, year, and rating

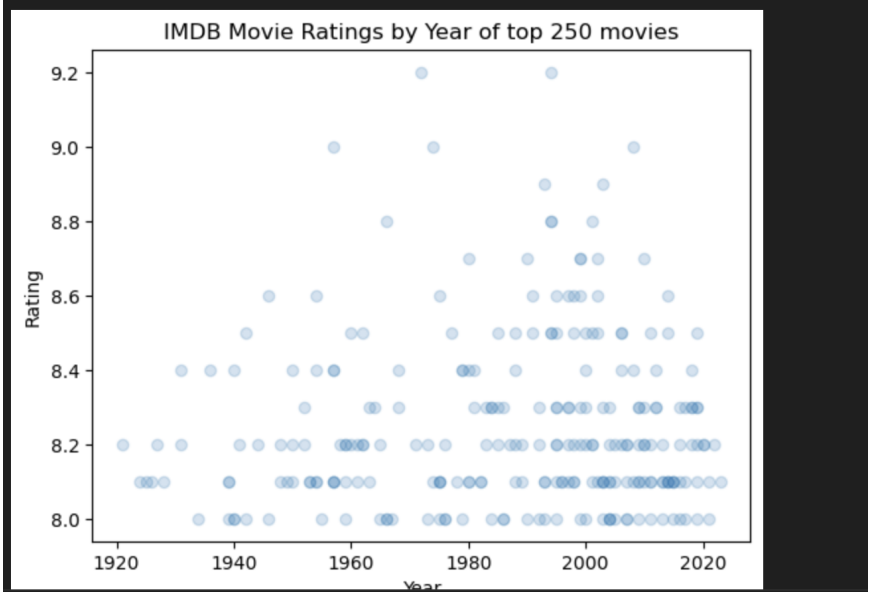
```
movies = []

for title_and_year, rating in zip(title_and_year_tags, rating_tags):
    # get title, year, and rating
    title = title_and_year.find("a").get_text()
    year = int(title_and_year.find("span", {"class": "secondaryInfo"}).get_text().strip("()"))
    rating = float(rating.find("strong").get_text())

    movie = {"title": title, "year": year, "rating": rating}
    movies.append(movie)
```

```
movies
```

```
[{'title': 'The Shawshank Redemption', 'year': 1994, 'rating': 9.2},
 {'title': 'The Godfather', 'year': 1972, 'rating': 9.2},
 {'title': 'The Dark Knight', 'year': 2008, 'rating': 9.0},
 {'title': 'The Godfather Part II', 'year': 1974, 'rating': 9.0},
 {'title': '12 Angry Men', 'year': 1957, 'rating': 9.0},
 {'title': "Schindler's List", 'year': 1993, 'rating': 8.9},
 {'title': 'The Lord of the Rings: The Return of the King',
  'year': 2003,
  'rating': 8.9},
 {'title': 'Pulp Fiction', 'year': 1994, 'rating': 8.8},
 {'title': 'The Lord of the Rings: The Fellowship of the Ring',
  'year': 2001,
  'rating': 8.8},
 {'title': 'The Good, the Bad and the Ugly', 'year': 1966, 'rating': 8.8},
 {'title': 'Forrest Gump', 'year': 1994, 'rating': 8.8},
 {'title': 'Fight Club', 'year': 1999, 'rating': 8.7},
 {'title': 'The Lord of the Rings: The Two Towers',
  'year': 2002,
  'rating': 8.7},
 {'title': 'Inception', 'year': 2010, 'rating': 8.7},
```

Once I have the data in an ideal clean format, I need to analyze it to find patterns and trends. I can use different tools like Google Sheet or Matplotlib to create graphs that show how the movie rating is related to the year the movie was made. This will help me understand if there is any correlation between the two.

```python
import matplotlib.pyplot as plt

# Create a scatter plot of rating vs. year
plt.scatter(df['year'], df['rating'], alpha=0.2)

# Set the x-axis label
plt.xlabel('Year')

# Set the y-axis label
plt.ylabel('Rating')

# Set the title of the plot
plt.title('IMDB Movie Ratings by Year of top 250 movies')

# Display the plot
plt.show()
```



Based on the analysis, I may identify patterns in the movie data such as a correlation between a movie's production year and its rating, showing that movies made later tend to have a slightly higher rating. With this finding, I would need to communicate these insights effectively to others who may not have a background in data science. This may involve creating visualizations or presentations that highlight the findings and recommendations.

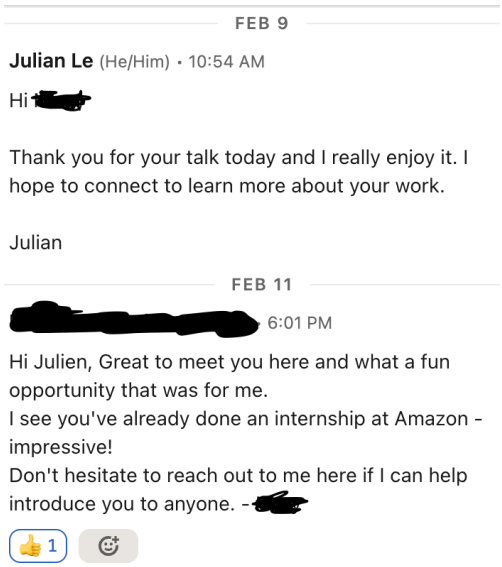
However, solving this problem can vary depending on the data and my objectives. I may need to use different methods or techniques, such as machine learning algorithms or other datasets, to further explore correlations in the top movies.

Continuous learning is essential

To succeed in the field of data science, we need to keep learning and updating our skills because the field is constantly changing. To achieve this, we should be curious, open-minded, and willing to learn. During class, we had the opportunity to use

various tools and techniques such as Python, Jupyter notebooks, API, Web Scraping, Javascript, and machine learning algorithms. Additionally, I expanded my knowledge by using resources like Data Camp, which offers helpful tutorials on topics I am interested in. These resources serve as a foundation for me to keep exploring and learning, and to work on my own data science projects.

We should also be active members of the data science community, attending conferences, participating in online forums, and sharing knowledge with others. In the class, Professor Wirfs-Brock has invited various guests to share their experience with students, giving the opportunity to not only learn from the experts in the field in an interactive way, but also have a chance to ask questions directly. Last but not least, I have also learned to connect with the speakers after the sharing thanks to Professor's suggestions.



Connect with guest speaker on LinkedIn after the talk

Collaboration is key

Data science is a team sport, and collaboration is essential for success.

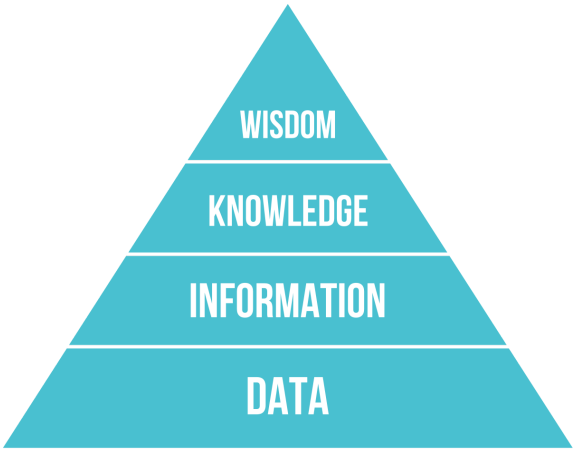
During our class, we did some group projects which helped me to develop skills such as teamwork and learning from my classmates. I realized that working alone on these projects would have been difficult and taken a lot of time.

One example is when I worked on the `Project 2: Linear Motion` in a team of three. We needed to do several tasks in order to accomplish our goals. One of the them requires us to measure how tall a group member is using the phone acceleration sensor. Doing this task alone would have been difficult. Moreover, working in a group allowed us to help each other when we faced difficulties, which helped us move forward. Also, pair programming helped us catch mistakes and bugs, and complete tasks faster.

Data is more than just numbers

Data is more than just a collection of numbers or facts. It carries important information, significance, and knowledge that we can learn from it.

The DIKW pyramid provides a framework to understand how different levels of data are related to each other. It explains that data is the raw material that is transformed into information, which gives it meaning. Knowledge is built on top of information, and it requires understanding and decision-making abilities. Wisdom is the highest level and involves applying knowledge and experience to make wise decisions.



DIKW pyramid

Conclusion

To sum up, data is an important part of our lives, and we have learned in the Introduction to Data Science class that it is not just numbers but also shaped by its collectors. We should be aware of where the data comes from and examine it critically. We have also learned how to use data visualization to communicate complex information in a clear and understandable way, but it is essential to avoid misleading visualizations that can manipulate people's understanding. By using data ethically and keeping these principles in mind, we can gain knowledge and understanding that will help us learn more about the world.