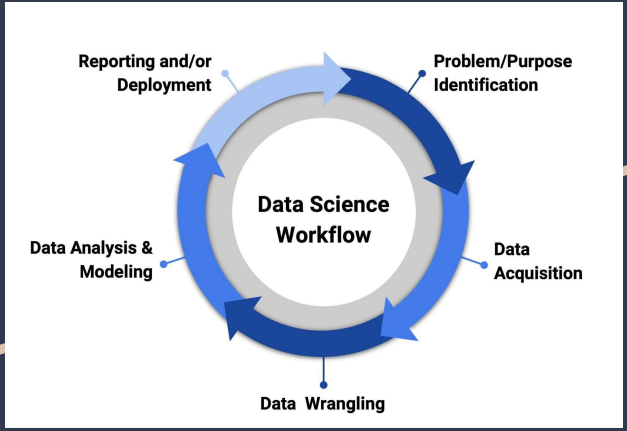
What I learned in Data Science

Aaryan Samanta

Module 1: Introduction to Data Science



- **Data Science Summary:** It's a field that uses scientific methods and systems to get valuable knowledge and insights from data.
- **Real-World Applications:** It's used everywhere from business to medicine. For example, it helps companies recommend products to customers or aids doctors in making a diagnosis.
- We learned the main steps of a data project: asking a question, getting the data, cleaning it, analyzing it, and then sharing the results.

Module 2: Python for Data Science

- We learned the basics of the **Python** programming language.
- **NumPy:** Used for working with numbers and performing fast calculations.
- **Pandas:** A powerful tool used to organize and analyze data in a table format.
- We worked on various coding exercises to practice our skills with these libraries.

Module 3: Data Collection and Ethics

- **Sources of Data:** We discussed where data comes from, like public websites, APIs, and surveys.
  - **APIs** are like digital messengers that allow different programs to send and receive data from each other.
  - **Surveys** are a way to gather information from people by asking them questions.
- **Ethical Considerations:** We learned about the importance of being ethical with data. This includes getting **consent** from people, protecting their **privacy**, and making sure the data doesn't have **bias** that could lead to unfair results.

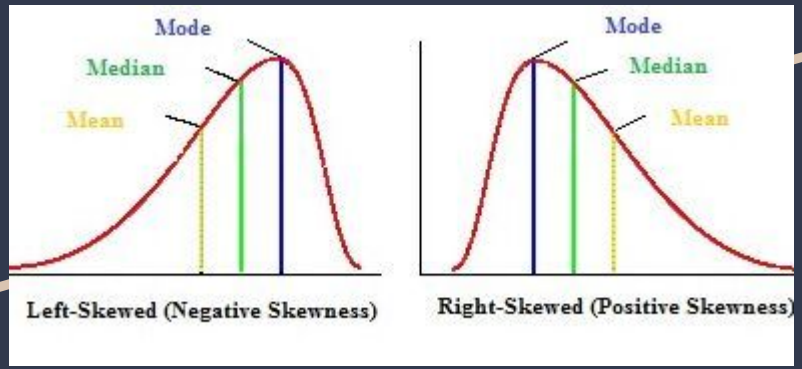
Module 4: Data Cleaning and Preparation

- **Importance:** Real-world data is often messy, with mistakes or missing information, so we have to clean it to get accurate results.
- **Key Steps:**
  - **Handling missing values:** We learned how to fill in or remove missing data.
  - **Removing duplicates:** We made sure each data point was unique.
  - **Transforming datasets:** We prepared the data so it was ready for analysis.

Module 5: Data Visualization Basics

- **Best Practices:** We learned how to create charts that are clear and easy to understand.
- We also learned the importance of choosing the right graph to represent certain data
  - Bar graphs are optimal for representing categorical data
  - Heatmaps are optimal for representing correlations
- **Tools for Visualizations:**
  - **Matplotlib:** A basic tool for creating all kinds of plots.
  - **Seaborn:** A more advanced tool for making beautiful statistical charts.

Module 6: Introduction to Statistics

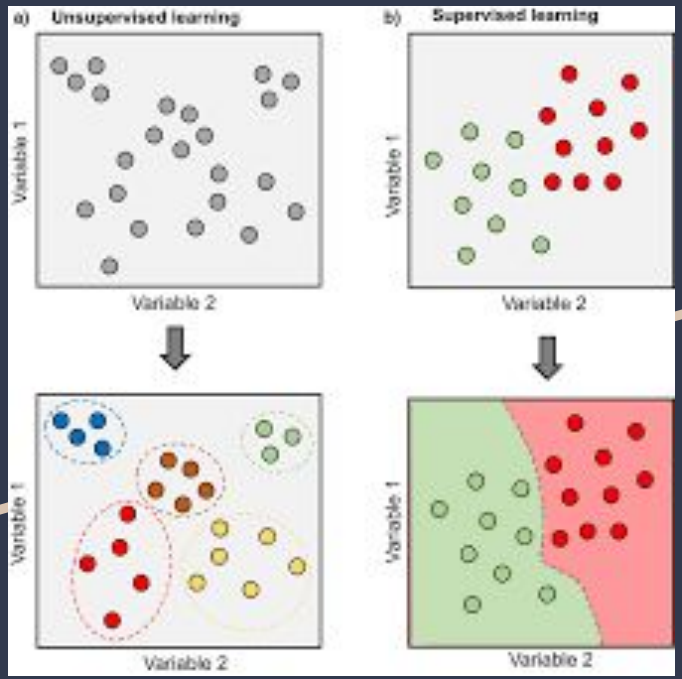


- We learned about the most common concepts used to understand data:
  - **Mean:** The average value.
  - **Median:** The middle value.
  - **Mode:** The most frequent value.
  - **Variance:** How spread out the data is.
- **Correlation vs. Causation:** We learned that just because two things are related (correlation) doesn't mean one causes the other (causation).

Module 7: Exploratory Data Analysis (EDA)

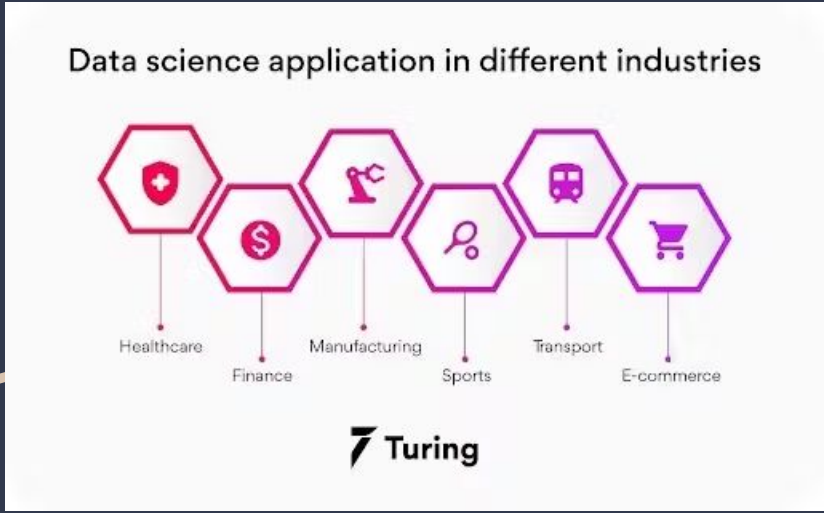
- **Identifying Patterns:** EDA is like detective work. It's the process of looking for trends, patterns, and strange data points.
- **Hypothesis Testing:** We learned how to use EDA to test a guess or idea about the data to see if it's true.
- We used a correlation heatmap in our code to find patterns, and we used a histogram to see a general trend of which category the patients fell into

Module 8: Introduction to Machine Learning



- **Supervised Learning:** This is when a computer learns from data that already has the answers. We use this to make predictions.
- **Unsupervised Learning:** This is when a computer finds hidden patterns or groups in data on its own.
- **Algorithms Explored:** We worked with algorithms like **linear regression** (for predictions) and **k-means clustering** (for grouping).

Module 9: Real-World Applications



- **Case Studies:** We looked at examples of data science in action:
  - **Healthcare:** Predicting diseases.
  - **Sports:** Analyzing player performance.
  - **E-commerce:** Recommending products.
- **Career Paths:** We discussed different jobs in data science, such as a Data Analyst or Machine Learning Engineer.

Module 10: Capstone Project

- **The Problem:** I analyzed patient data to find relationships between symptoms, demographics, and diseases.
- **The Data:** I worked with a dataset that included a patient's age, gender, blood pressure, cholesterol, and symptoms.
- **Key Insights:**
  - I cleaned the data to make it usable.
  - I found that a patient's blood pressure is related to their cholesterol level.

Conclusion

- The most important thing I've learned is that data science is a powerful tool that can be applied to almost any problem.
- I've found that data visualization and machine learning were particularly interesting because they help make sense of complex data.
- I think data science will be essential in my future career, because it will allow me to make data-based decisions and solve real-world challenges