AMA
LEARNING

Data: Notes

In this lesson:

- Learning from Data/Trends
- Exploring One Column
- Filtering & Cleaning Data
- Big, Open, and Crowdsourced Data
- Machine Learning/Bias

Learning from Data/Trends:

- Data visualizations help us answer:
    - **"I think this visualization tells me this…"**
        - Something is more popular than something else
        - Something is more important than the other
        - Something has become more or less searched over time
    - **"...but I am not sure because…"**
        - I don't know how the data was collected
        - The data doesn't tell the reason for a certain trend/correlation
        - We need more data
- **Visualizations**
    - **Help us look at lots of data at once**
    - **Help see patterns that are "invisible" if you just look at a table**
    - When looking, consider:
        - What does this data show? (FACT)
        - Why might this be the case? (OPINION)
    - **CORRELATION ≠ CAUSATION**
        - Use this when making assumptions about data
    - Programs (data visualizer) help princess data so we can understand it and learn
    - charts/visuals help find and communicate what we've learned from data

Exploring One Column:

- **Metadata: data about data**
    - Can be changed without impacting the primary data
    - Used for finding, organizing, and managing information
    - Increases effective use of data by providing extra information
    - Allows data to be structured and organized
- **Data analysis process:**
    - **1. Collect or choose data**
    - **2. Clean and/or filter**

  - **3. Visualize and find patterns**
  - **4. Generate new information**
- **Bar Chart:** count how many times each value in the column appears and make a bar at that height
  - If column has too many unique values, it gets difficult to make any sense of them or find patterns
- **Histograms:** similar to bar chart but first, all numbers in a range are grouped together
  - Can only be created with numbers but useful when bar chart is hard to read

Filtering & Cleaning Data:

- **Cleaning Data**
  - Why clean data?
    - **Data is incomplete**
    - **Data is messy**
    - **Data is invalid**
  - What is messy data?
    - User enters different types of data (“two”, 2)
    - Users use different abbreviations to represent same information (“Feb”, “Febr”)
    - Data has different spellings (“Color”, “Colour”) or inconsistent spellings (“Spring”, “spring”)
- **Filtering Data: allows the user to look at a subset of the data**
  - Instead of using traversals, software programs with built in tools (data visualizer) can be used to filter data

Big, Open, and Crowdsourced Data:

- **Big Data:** Evaluating the large amount of information that a computer has to go through by using parallel processing (Taking large numbers of data and many computers process at the same time)
- **Open Data:** Data that is available to the public by being published publicly
  - Ex: weather app, GPS app
- **Citizen Science:** effort to get public engaged in science and using knowledge to solve real world problems
  - Example: Migratory patterns, climate change information
  - Members of the public collect, analyze, spread information about data for research
  - Technology allows anyone to be involved in citizen science

Machine Learning/Bias:

- **Machine Learning:** computers recognize patterns and make decisions without being explicitly programmed
    - Ex: text, email, filters
    - Computers learn with trial and error
    - Machine Learning can take any kind of data
    - Training data needs to be of high quality
- **Bias Data:** Favors some things/de-prioritizes others
    - Need a lot of training data to prevent this
- **Algorithmic bias:** exclusionary experiences/discriminated
    - Need full spectrum inclusion
- **Bias warning:** Need to make sure to not use prejudice when training data that perpetuated human bias