Unit 8 - Data - Study Guide

Getting Started with Data

- **Data** refers to any information about the world that's been placed in a computer-readable form.
    - Computers are only able to read binary, so we need to transform human-readable information like physical pictures, handwriting, etc. into digital bits somehow.
    - Data is not super useful all by itself - it needs to be processed, interpreted, or worked with in some way for us to gain any knowledge from it.
- We use computers to process data because they are much better at processing large quantities of information than humans are - they can do more calculations in less time.
- We can use tools like Databases to make storing large quantities of data simpler, and we can use Database Languages to make retrieving information easier!

Visualizing Data

- Data Visualization is the process of using graphs, charts, or images to display complex data. Visualization allows us to draw conclusions based on the available data much more easily than looking at raw numbers.
- While computers make visualizing data much easier, they aren't strictly necessary!
- There are a variety of different visualizations that we can use to better understand data.
    - Table
        - Tables are most useful when we want to display **precise** data.
    - Bar Chart
        - Bar Charts allow us to **compare different categories** of data against one another.
            - We can stack the bars for a slightly different effect!
    - Pie Chart
        - Pie charts are used when we want to display **percentages** of a whole.
    - Histogram
        - Histograms are used to display the **frequency** of events.
    - Line Chart
        - Line Charts are best suited to showing how numbers change **over time**.
- If we can add some degree of interactivity to our visualizations, we can increase engagement. This can lead people to gain a better understanding of the information being displayed!

Collecting Data

- There are many different ways in which we can collect data!
    - Surveys
        - If we want to get information from people about people, we can send out surveys!
    - Sensors
        - We can use all kinds of different sensors to gather information about the world around us!
            - Thermometers, barometers, etc.
        - Usually these'll interface directly with our computers, which lets us store the data we collect directly!
    - Transactional data from credit cards
        - Anytime you make a transaction using a credit/debit card, it gets tracked!
        - This allows us to see our own spending history
        - This also allows credit card companies to see how we spend our money!
    - Websites storing information about you
        - Websites will often track how long their users spend on any given page, what they click on, etc. in order to improve their services
    - Crowdsourcing data
        - Pose a question to a large number of Internet users, or potentially track the behavior of a large group of people on the Internet
- Once we've collected some data, we will need to store it somewhere.
    - Databases and Data centers are most often used for this!

Data Limitations

- Something that we need to be wary of when looking at visualizations of data are **misleading** visualizations. We learned about a few different ways to create a misleading representation of data.
    - Truncation of Axes
        - If only a small portion of an axis is visible, it can potentially show a much larger change than is accurate or reasonable.
    - Omission of Data
        - A misleading tale can be told if certain data points that don't align with the desired narrative are omitted.
    - Breaking of Convention
        - If graphs are created in a way that doesn't match how we would normally expect to interpret a graph of that type, it can be misleading.
            - Ex: Having the slices of a pie chart not accurately represent the portion of the whole that they should
    - Using Correlation to Imply Causation
        - Just because two events occur at a similar frequency does not imply that they are connected in any way - it's more than likely that one did not cause the other.
- Metadata, or information about the data, can be helpful in determining whether a dataset is viable and trustworthy. Here are some questions to consider about a data source before using to draw any conclusions:
    - Where was this data collected?
    - Who collected this data?
    - How long ago was this data collected?
    - How large is this data set? Is it accurate?