Week 11

Analyze data for bias and credibility.

Quality of data

Data ethics and privacy

Bias and unbiased data

* Sampling bias: When a sample isn’t representative of the population as whole.
* Observer bias (experimenter bias or research bias): The tendency for people to observe things differently.
* Interpretation bias: The tendency to always interpret ambiguous situations in a positive, or negative way.
* Confirmation bias: The tendency to search for or interpret information in a way that confirms pre-existing beliefs.
* Unbiased sampling: When a sample is representative of the population being measured.

Identifying good data sources (ROCCC):

* Reliable
* Original
* Comprehensive
* Current
* Cited

Data ethics: Well-founded standards of right and wrong that dictate how data is collected, shared, and used.

GDPR: General Data Protection Regulation of the European Union.

Aspects of data ethics

* Ownership: Individuals own the raw data they provide and they have primary control over its usage, how it’s processed, and how it’s shared.
* Transaction transparency: All data-processing activities and algorithms should be completely explainable and understood by the individual who provides their data.
* Consent: An individual’s right to know explicit details about how and why their data will be used before agreeing to provide it.
* Currency: Individuals should be aware of financial transactions resulting from the use of their personal data and the scale of these transactions.
* Privacy: Preserving a data subject’s information and activity any time a data transaction occurs.
* Openness: Free access, usage, and sharing of data.

Open data: free access, usage, and sharing data.

Data interoperability: The ability of data systems and services to openly connect and share data

Luckily for data analysts, there are lots of trustworthy sites and resources available for open data. It is important to remember that even reputable data needs to be constantly evaluated, but these websites are a useful starting point:

U.S. government data site: Data.gov is one of the most comprehensive data sources in the US. This resource gives users the data and tools that they need to do research, and even helps them develop web and mobile applications and design data visualizations.

U.S. Census Bureau: This open data source offers demographic information from federal, state, and local governments, and commercial entities in the U.S. too.

Open Data Network: This data source has a really powerful search engine and advanced filters. Here, you can find data on topics like finance, public safety, infrastructure, and housing and development.

Google Cloud Public Datasets: There are a selection of public datasets available through the Google Cloud Public Dataset Program that you can find already loaded into BigQuery.

Dataset Search: The Dataset Search is a search engine designed specifically for data sets; you can use this to search for specific data sets.