Data Science Assignment 1a

Timothy Fischer

Task Zero: Data collection

Data collection is the process of gathering relevant information specific to some population. (pragmaticinstitute.com) Unfortunately, there a few potential biases that might negatively affect the data collection process. These biases can skew they way we analyse data leading to misinterpretation or misunderstanding of the population.

- Confirmation Bias: Occurs when we only collect data that leans towards or confirms some hypothesis or assumption. For example, if I am collecting data to identify whether cats or dogs make better house pets. I might collect more positive data on dogs if I like dogs more than I like cats.

- Historical Bias: Occurs when our culture or beliefs affect the data collection process. For example, some cultures might not treats cats and dogs the same way as other cultures. Or some culture might view cats differently to dogs.

- Selection Bias: Occurs where the sample created during data collection does not accurately represent the broader population. We select data subjectively rather than objectively. For example, during data collection, we might collect data on only specic breeds of dogs. Or some breeds might be represented disproportionately in the sample.

- Survivorship Bias: Occurs where we only focus on successful cases during data collection. We fail to see that failures can also give us useful information. This can give a skewed view of the data. For example, we might not take into account that certain dog breeds are quite aggressive compared to others and we might miss that.

- Availability Bias: Occurs when current events affect the data collection process. Some events might change the data we collect. For example, if there is a news article about a person who was attacked by a dog, people might shy away from adopting or taking in dogs as pets for a time.

Task 1: Data Cleaning

See attached Assignment_1.ipynb file.

Task 2: Data Integration

Data integration involved combining data from different sources (Google Cloud). Weather its merging multiple datasets or combining data from different formats.

Challenges associated with data integration:

- Scale: The amount of data we process is constantly growing. Finding ways of process what seems to be an infinite amount of data is a very complex issue. For example, trying to store information coming in from social media posts can be very difficult as people are constantly uploading new content.

- Semantics: Data comes in various formats. This requires data to be reformatted and modified in order for it to be combined or integrated. For example, trying to combine and store data in text format with data we get from audio requires the audio to be reformatted or converted to text in some way.

Data quality is vastly improved when integrating data (Jeffrey Richman). This can lead to better decision making since we have access to more data that is highly integrated giving a better view of what is happening. Integrated data is also more central and has a fixed format, meaning we have better access to the data, meaing it is easier and faster to process.