

5.2: Data Ethics: Data Bias

1. What types of potential bias exist in your team lead's collection plan? Why was it biased? Please explain your answer. You may also think of biases that go beyond this reading (e.g., cultural bias).

Due to exclusion bias, sampling bias, cultural bias, and a lack of knowledge about the data collection process, the results may be biased. The information is based on US institutions that are less than 100 miles from the border, thus it could not be typical of all instances of money laundering. Due to the data being from a few years ago, the sample might perhaps not be typical of the population. The inspection of ATMs within 100 miles of the border results in an exclusion bias because other potential sources of suspicious behaviour outside of that radius are not considered. Furthermore, the team leader's assumption of appropriate cleansing has resulted in out-of-date information.

2. How might these biases distort the results? What could you do to avoid these biases?

Recognising the potential impact of missing data in pre-cleaned data as well as the dynamic nature of money mule operations is crucial for ensuring reliable analytical results. It is essential to comprehend data collection techniques, probe biases, and deal with sample, measurement, and exclusion biases. It is possible to reduce collecting bias while still obtaining a representative and up-to-date sample by extending the data range beyond 100 miles and incorporating transaction data from Mexican banks. To combat measurement bias, equal training and experience should be guaranteed. To avoid exclusion bias, the justification for a 100-mile distance must be carefully examined. Furthermore, seeking recent data takes into account any prospective changes in the way money mules transport money across borders. Overall, these steps are intended to prevent skewed results and present suspicious and regular activities more realistically.

3. If you know that there is bias in the collection method, what could you do to communicate your concerns to your team lead? Please be as specific as possible.

I would begin by conducting tests to validate my assumptions and address any biases of concern before presenting compelling evidence and alternative data collection approaches to my team lead in a private setting. Before sharing results with decision-makers, I would emphasise the importance of discussing biases, expressing my concerns, and seeking clarification. Alongside playing the role of "Devil's Advocate" to identify errors, I would propose alternative ideas and request that my team lead carefully consider them. By adopting a constructive and collaborative approach, I aim to ensure the integrity and quality of the data.

4. Read through the details of testing. How might the lack of transparency around the experience and training of the investigators allow for bias?

Significant measurement bias may result from the lack of transparency around the background, education, and training of the 10 distinct analysts. When reporting a

favourable or bad outcome, a person's objectivity may be impacted by variables like age, occupation, gender, and nationality. The Banks use threshold- and logic-based models, however, what is considered "logical" by one person may not be the same for another. Measurement bias is possible due to the investigators' lack of transparency on their background and training because this can influence the results of the analysis. To complete the analysis for this project, each analyst has to have received fair and sufficient training.

5. Analyse the bar chart showing the scores of individual analysts and see where their scores fall on the distribution curve. If the mean of the scores was 307 and the standard deviation is 166, which score or scores might you eliminate to control for bias? Why?

Data with a normal distribution are those that are within one standard deviation of the 307 mean. The top score of Analyst 10 (759), which skews the results, should be removed. Understanding the analyst's biases and whether they were brought on by a lack of training or experience will help you avoid getting results that are distorted. All values would be within two standard deviations of the new mean, which would be 256 with a standard deviation of 47. This choice needs to be carefully thought through since, before completely deleting Analyst 10's suspicious transactions, it is necessary to comprehend their context.