

Mary Kane

Exercise 5.2 Data Ethics: Data Bias

Answers 5.2

1. Carefully read the background and collection plan again. 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).

a. **Sample bias:** There were only collected cases that match "cash deposit within 100 miles from the border and with withdrawal in Mexico"

There are legitimate customers with this cross-border behavior (ex: Mexican citizens working in U.S., families sending money home).

b. **Measurement bias:** The team is using "cleaned data" from years ago and requesting the analyst to mark the scenarios as suspicious for cartel-driven laundering vs not.

The definitions in ATM networks, customer behavior, and reinforcement patterns may have changed. Also, framing the data into a suspicious bucket based on sensitive demographic attributes, such as nationality, already limits the analyst.

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

The sample bias overrepresents people who live/work near the border or travel to Mexico, potentially leading to false positives for legitimate border communities.

To reduce bias: Fix sampling (add payroll deposits), create subgroups that better represent transaction behavior (border/non-border, frequent traveler/non-frequent traveler), justify behavior (e.g., transaction + withdrawal within 2hr), and hide demographic information.

The previous cleaning choices could have removed anomalies that better represent the new laundering tactics. It's important to standardize labeling to avoid overlapping labeling.

Mary Kane

Exercise 5.2 Data Ethics: Data Bias

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.

Explain the disparity observed in a short message, including evidence, offering fixes and a short pilot to validate the concern:

*“While checking the labeled data, I noticed something that worries me. Mexican citizens are about **11%** of the customers in our sample, but they are **about 75%** of the items marked **suspicious (positive)**.*

*This could mean our data or labeling is **skewed**, and the model might learn “nationality/location” instead of the real suspicious behavior. That could create a lot of **false alerts** for normal customers.*

Before we move to the next step, can we do a quick check? I can pull:

- *Positive rate by nationality (Mexican vs not Mexican)*
- *Positive counts by analyst (to see if analysts score very differently)*
- *A small test where analysts label cases **without seeing nationality/gender**”*

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

A lack of transparency about the investigators' **backgrounds, training, and how they were instructed** can create bias because the “ground truth” labels may reflect the investigators—not the actual risk. There is no prove they were trained the same way or if they have hidden conflicts in cultural understanding that affect their evaluation.

Exercise 5.2 Data Ethics: Data Bias

- Analyze 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?

Results: Of the 10,000 work items analyzed, the total aggregate score was 3,066 positives. Your team lead believes that this aggregate score is good enough to move your model to the next iteration. The bar graph below shows the individual analyst scores for the number of items they deemed positive or suspicious.



The mean of the “positive counts” = 307

Standard deviation = 106

Most analysts (179–358) fall within about **-0.77 to +0.31 SD** of the mean → *well within the main part of the curve*.

759 has a z-score of:

$$z = \frac{759 - 307}{166} \approx 2.72$$

That's **~2.7 standard deviations above the mean**, which puts it in the **extreme right tail** (an outlier).

Eliminate the analyst 10, because he is labeling way more items as suspicious than everyone else, reflecting different standards, misunderstanding or personal bias.