Background: Mexican drug cartels face the challenge of moving their ill-gotten US Dollars into Mexico and converting them into Mexican Pesos without alerting banks and authorities. Trade-based money laundering is one example of how they do this. Federal authorities have alerted US banks to an emerging trend where money mules deposit US Dollars in ATMs near the border, and then cross the border into Mexico where they withdraw the funds they just deposited in Mexican Pesos.

Goal: Your manager has asked your team to collect data on ATM transactions and design a model to monitor this type of transaction activity. Your team has also been tasked with generating an alert that will be routed to the Investigations Department of Pig E. Bank for further analysis. This project will be challenging since many Mexicans who work legally in the United States get paid in US Dollars but need Pesos to spend while they are home. The model will require human input and strong data analysis to distinguish normal transaction activity from suspicious transaction activity with an acceptable degree of accuracy.

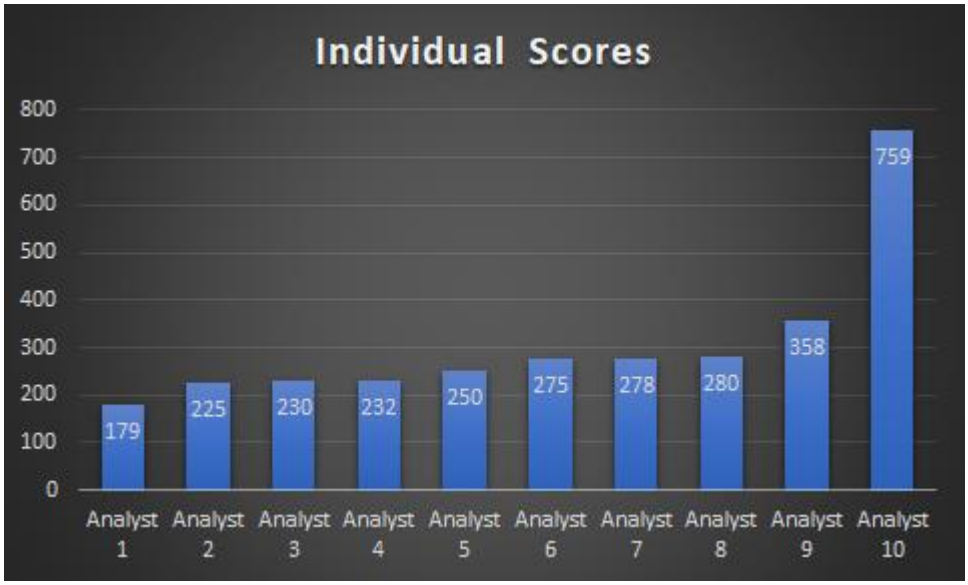
Collection: As a junior analyst on this project, you'll mostly be concerned with pulling and cleaning data to prepare it for testing. You begin writing SQL queries to return a population of ATM and bank transactions patterns on both sides of the US-Mexico border. More specifically, you decide to pull data on cash deposits made at banks in the U.S. that are within 100 miles of the border, as well as subsequent ATM withdrawals in Mexico. Fortunately, your team leader already has some data on ATM transactions from a project they led a few years ago, and it's already been cleaned! This will save a lot of time and effort and move the project along to the more interesting phases. Like many data analysts, you don't particularly enjoy cleaning data and assume that your team lead knows best, so you go along with their plan.

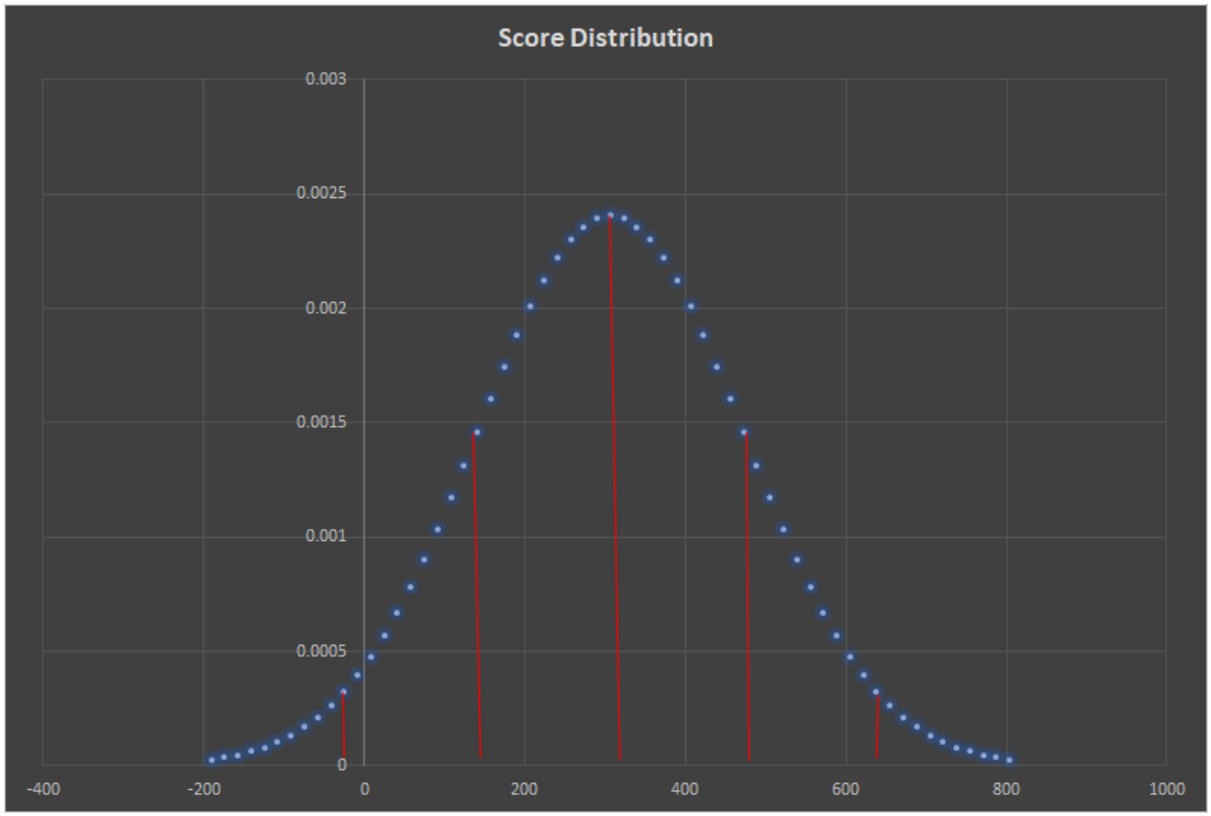
Testing: Your team lead needs human input to train the model to distinguish between normal activity and suspicious transaction activity for this particular data set. They create a test in which humans will analyze the ATM data and score each instance as "positive" if it's suspicious and "negative" if it's not. The Investigations Department loans your team 10 analysts to work on the project.

Tip!
Banks use several different models, also known as "rules," to monitor transaction activity. These models are basically a combination of logic and thresholds that define the transaction pattern and monetary amounts that must be exceeded to create an alert. This alert is the model predicting that investigators will find the alert suspicious.

The analysts must consider variables such as the age of the account, transaction amounts, the time between deposits and withdrawals, and how often they occur. They must also consider customer variables such as age, occupation, gender, nationality, and address. The model will then use the analysts' inputs to identify variables that are common to suspicious and normal activity. Your team lead wants to be sure that the sample data used for the test is large enough, so they give each analyst 1,000 work items to score.

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.





When analyzing the data sample further, you notice that 11 percent of customers whose transaction activity fits the logic are Mexican citizens. But 75 percent of the items flagged as suspicious (i.e., marked positive) involve Mexican Citizens.

Create a new text document and call it “Answers 5.2.” Write your answers to the questions below in this document. Try to limit your answers to 100 words per question:

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).

There is certainly measurement bias at play. Analyst 10 has identified over 2 x the numbers of positive of anyone else – and that will drag the criteria "upwards" i.e. more people will fit the criteria. Perhaps excluding the top and bottom scoring analysts would have been a better idea.

There may be collection bias – as transaction patterns from a few years ago may not fit patterns of transactions today.

We don't know much about the sample of the old data, so there may be sample bias. Likewise, what's been excluded when this was collected – what don't we know about the transactions.

We can't rule out there being other human biases at play e.g. if analysts might tend to view Mexicans as less trust-worthy, they may trust measurements that fit that conclusion.

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

In the case of using Analyst 10s interpretation, then more transactions will get flagged as suspicious – and likely more than actually are suspicious.

Wit the data being old, we don't know that it is truly reliable. Use of cash has changed hugely in recent years (at least in the UK) - there could now be very different types of transactions going on, and indeed we could be missing criminal activity by relying on these results.

If our sample is not representative of the population we are now looking at – it could lead us to entirely the wrong conclusions

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 gather as much information as I could on the method of collection. I would then document an ideal collection method.

This would allow me to present to my team lead, where bias could be allowed to enter the collection.

That then allows a discussion about what bias could be present in the data, whether that can be mitigated, or whether a different type of data collection altogether would be needed.

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

From what we've been told, we don't know anything about the testers. They could all be highly credible and reliable.

We do know that "They must also consider customer variables such as age, occupation, gender, nationality, and address" – but what we do not know is how they are considering these. Their own assumptions, or prejudice, may mean that they give some variable more weight than others. We do not know if this has happened consistently across the testers.

If they haven't been selected in the same way, and briefed in the same method – then there is a a lot of opportunity for bias to enter the system.

5. 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?

If the mean score is 307, and the standard deviation is 166, then we may consider eliminating scores that are less than 141 or greater than 473. These scores will be outside of the main body of the bell curve. We would certainly consider strongly eliminating scores below -25 (but that is not possible) or above 639 as these are tow standard deviations away from the mean.

In this case, I would suggest we remove analyst 10s score of 759. It is so far out of the standard deviation as to be unreliable.