To find an organized human trafficking network, to uncover the truth: *follow the money*.
A tale of great intentions, and great woe.

1) Problem Statement

Human traffickers use the public banking system to store and move monetary value, undetected. In organized networks of traffickers, this effect is heavily exasperated. The banking system allows them to operate far more efficiently as well as mix in with 21st century society, which is increasingly becoming cashless and would otherwise be abrasive to their very existence.

2) Objective

Utilizing python, pandas and any other programming tools necessary, data analysis will be performed on banking transaction records to identify potential activity related to an organized human sex trafficking network.

3) Solution/ Data Use Case Description

The Hypothesis:

Parts in italics marks how we changed our mission once it was obvious we weren't acquiring real banking data in time.
Name) The belief held by a member of Amazing Grace.

Initially we would look for three broad categories of accounts and related transactions.

First, lower level accounts. In Canada and similar first world circumstances, these would be the direct handlers of trafficked sex workers. The pimps, the supervisory muscle, the street boss; individuals who were responsible for ensuring the trafficked victims brought in the necessary income for as long as possible.
These accounts would have large cash deposits, likely at ATM's. These cash deposits would often be similar to each other and follow a semi-regular time schedule, as the pimp made his rounds on his cattle and deposited what he'd taken from them.
These accounts would often not have the pattern of life related activity one would expect from any normal individual earning this level of income, due to the pimp dealing mostly in the safety of cash, or having a different personal account.
These accounts would send much of the deposited cash onto another account.*
*Claude) In days before machine thinking was prevalent, these low level accounts would then send a large percentage of what was deposited onto intermediary accounts. Which would then forward this money onto other intermediary accounts, until it made its way to a higher up account, often in a country with strong banking client privacy laws such as Switzerland. Very possibly smaller organizations still operate in this manner, especially in less developed parts of the world.
*Bitia) In Mexico, the cartels never allowed the street thugs to use banks. They were too stupid to be trusted. Organized human trafficking networks coming from the same peoples also likely don't operate this way.

*Jesse) I have never directly dealt with an organized crime syndicate, but my friends in the military have often told me that since 9/11 the drug syndicates have become significantly more sophisticated in avoiding and countering detection. Looking for transferred funds between accounts might not be an available means of detection anymore. e.g. One possible alternative that doesn't include crypto currency or bogus civil cases where the defendant purposely loses to legally transfer and thus launder money to the extra national claimant, is the ability to withdraw funds from an account from any ATM nearly anywhere. The street level boss could deposit 1000$ cash in Calgary Alberta, and someone further up the organization could withdraw from an ATM on the outskirts of London Ontario, with total anonymity.

Second, the intermediary accounts. In Canada and similar first world circumstances, these would be accounts that were created with either fake identification or with handlers who were bribed or coerced to be an intermediary. As laws change in traditionally safe harbours for illicit income, for instance Switzerland, and as laws about moving money over national borders becomes far stricter; movement of funds to a relatively safe deposit within a country becomes more appealing. Intermediary accounts would break up the trail of money, ensuring far greater difficulty for a higher up to be apprehended by authorities, once a lower level account had been compromised.

These accounts would be the recipients of the transfers from lower level accounts, soon thereafter transferring the money onto other accounts. These accounts would not hold onto the transferred income for any significant length of time. Through the internet it would be possible to make these transfers without a human ever needed to be present at a banking institution.

Alternatively these accounts could closely resemble lower level accounts, being made up of large cash deposits and withdrawals at ATM's, as an intermediary player withdraws from a lower level account and deposits in an intermediate account, using different cards. And another player somewhere else withdraws from this account, to deposit in a different account.

In both situations, these accounts would be highly irregular as other than these transfers to and from, there would be no activity.*

*Jesse) Again giving faith to the potential of driven criminal minds; using intermediaries who also use the accounts for life expenses is likely for sophisticated networks. Finding people, for instance ex-cons, who would be willing to be the consciously ignorant body for an intermediary account would be easy enough. These intermediary bodies would then otherwise use their bank accounts as a normal person would. The intelligence in these organized networks are likely aware of the capabilities of machine learning and wouldn't purposely engage in activity or lack of activity that alert authorities to their presence.

Third, the upper level accounts. The proverbial kingpin, or more reasonably the local kingpin. Crime lords are likely many more layers removed and undetectable from top down data analysis.

These accounts would be very difficult to discover without the trail of other accounts leading to it, as activity other than cash deposits/ transfers would be extremely varied*.

*Claude) In a smaller network where this account belongs to a top level kingpin, if the network isn't that sophisticated then there would be largess in life expenses. In a more sophisticated network the kingpin likely deals almost exclusively in cash, so the only activity would be withdrawing cash from ATM's for later use, and purchases that can't use cash (e.g. Amazon).

In a larger network, the only difference between this upper level account and an intermediary account is the level of personal involvement by the account holder. An intermediary account belongs to someone with fake credentials, or someone beholden to the network who gains enough benefit for being an intermediary in return for possibly being a 'fall guy' (e.g. soldiers in the Italian mob). An upper level account belongs to a real person with a lot of power and control in the network. The upper level account would send onto another layer of intermediary accounts a required amount, possibly after laundering it, while keeping the rest.

Our intention would be to use machine learning to exhaustively feature select and uncover accounts with characteristics of these three categories. Once identifying suspicious transactions, and thus apprehending these possibly illicit accounts, we would investigate further for patterns suggesting either a cash deposit/withdrawal criminal network, and/or a network that uses transfers to move money around. Our collective intuition was that if we can find just one player in a network, the whole network would be compromised.

Our hope was that since banks have available the credit reports and transaction files for Kaggle competitions, with all identifying characteristics on given files removed, that we could acquire similar reports for banking transactions. Even if we were given excessively old banking transaction reports we could try and find patterns resembling our hypothesized three categories.

Scores of banking institutions from several continents, large and small and everything in between, gave firm no's to our requests, or asked us to solicit records through certain channels that would have taken far too long to pursue; in the remote chance our request for data was eventually approved.

Unfortunately many banks did not respond formally until later in the data jam timeline, and we were so blind to the possibility of not acquiring any data, that we didn't even think of a plan B until 48 hours before the submission deadline.

This hard reality became apparent. Our first recourse was to convert Kaggle fraud detection credit card files to banking transaction files, but it with a mere few hours of exploratory analysis it became apparent that even if members of criminal networks had credit card transactions in these files; none of the aforementioned patterns would be prevalent. No criminal would use credit cards to move money around in the manner mentioned above. We would need to add cooked transactions that met the pattern Claude and all of us laid out.

As modifying .csv files was something only I, of Team Amazing Grace, could do, and as I would also be the person using data science to uncover the mentioned patterns; we collectively decided this was too unavoidably bias prone a path to follow.

Zach discovered a website called '<u>mockaroo.com</u>', which was a very easy to use source of .csv (and other) files one could craft. He suggested that a completely fabricated dataset was made and that we undertake a triple blind study.

With 36 hours to go, we split into three mini-teams to try and put our hypothesis to test.

Team 1: Claude and Zach. They created the dataset from scratch.

Team 2: Myself, Jesse. I would use computer programming and data science to parse the dataset into observed patterns. I would forward on my findings without giving my own deduction; giving only a technical readout of uncovered patterns. I would aim to be as verbose as possible in what I uncover, to eliminate my own bias.

Team 3: Britt and Bitia. They would go over my findings and draw their own conclusions, and deliver a report based on what they deduced.

1) Neither Claude nor Zach had any machine learning or coding experience, at all. To understand how machine learning on banking datasets work they watched several videos on data science fraud detection. In addition, with the intention of creating a criminal network whose account activity was difficult for me to uncover, I did a machine learning demo for them; making survivor predictions on the famous Titanic dataset.

They elected to replicate a file released by a bank that had very little data on it, representing the banking industry's adherence to privacy laws, that we had found insurmountable. Only transaction amount, time, date and an index number (called client_id, but wasn't the actual account number, but could easily be associated to an account number by the bank if deemed necessary).

They purposely elected to leave out source information for deposits, so I couldn't see if it was cash, nor if it came through an ATM. The also purposely left out if withdrawals were transfers, or cash.

The hurdle they encountered was the file size limit placed by mockaroo.com for the free services, which is only 1000 rows. Due to a communication mishap, they also did not understand how easy it is for a machine learning engineer to combine .csv files into a larger file. As a result, the generated a file with a little over 33000 account transactions over the course of a year, belonging to almost 1000 accounts. This in contrast to they initially wanting to create one hundred million account transactions over half a decade, belonging to half a million different accounts, containing three different networks of differing sizes.

The 33000+ row file they did create was crafted in layers.

a) Zach/ Claude created 984 random accounts. These accounts would have a large number of smaller withdrawals, representing life expenses such as groceries, gas and hobbies. They would have a smaller number of larger withdrawals, representing larger purchases and credit card and car payments. They would also have few but regular large payments representing rent. These accounts would have regular somewhat infrequent large deposits, representing their wages earned.

b) Zach/ Claude created 10 low level illicit accounts. These had regular larger deposits and consistent withdrawals, representing the pimp's trade and transfers.

c) Zach/ Claude created 3 intermediary accounts that would take in these consistent transfers, and send them on.

d) Zach/ Claude created 1 upper level account, that would take in these same consistent transfers, and spend much of it. They decided that this account would belong to a local kingpin, so he would not be spending a substantial portion of his illegal earnings.

Note: I was not privy to any of this information until after doing my machine learning analysis.

Due to their inexperience with .csv files and the necessity for research, their responsibilities did not complete until early morning of the last day of the data jam.

2) Less than 12 hours to go.

As neither Zach nor Claude were familiar with statistics, they did not know that any account they fabricate needs to fall somewhere within three standard deviations of a standard distribution. As a result, their relatively wealthy kingpin was an instantly observable outlier. Just as government watch dogs pay closer attention to larger accounts with lots of large activity, so did I. This resulted in my uncovering their network with almost no feature extraction or other EDA.

Their baked criminal network also had other peculiarities that likely only occurred due to their total lack of experience with data, and would not be accurate to real banking data. (Please see associated .ipynb file)

All of the criminals in their lower and mid tiers withdrew exclusively in 500$ amounts, irreducibly uncovering their entire network was unavoidable. In addition, their criminal accounts conveniently had index numbers 985-999, and had no non-illegal activity whatsoever.

3) Bitia and Britt laughed off our collective stress. There wasn't anything to report, other than if client_id's 985-999 do belong to a criminal network, then they are the most ignorant buffoons outside of politics.

At least everyone came away with golden rays of hope; with next to no effort or intrepid machine learning sleuthing, I was able to identify their entire network. The human trafficking syndicates of the world would be far more careful about their banking patterns, but patterns would exist, and these worst forms of human consciousness would be identified, and apprehended.

4) Pitch

If fraudulent transactions, which require no pattern save for their anomalous nature, can be detected.
If malware across terabytes of constantly changing app software can be detected.
If efficiency improvements can be found in the unending gigabytes of generated data for each of the tens of thousands of planes taking off each day.
Then, human trafficking network banking transactions will leave a pattern significant enough to identify these accounts.

Unfortunately Team Amazing Grace was unable to even remotely conclusively prove this over the datajam. Hopefully in the future this will be rectified.

5) Datasets

A fabricated dataset representing what 'X' Bank was willing to release about a year's account activity for nearly 1000 accounts: transactions.csv (Please see associated .csv file).

6) Project Code

Python, through Jupyter Notebook was used, utilizing the Pandas and Matplotlib libraries. (Please see associated .ipynb file)

7) Additional Docs

(Please see associated .pdf file)
In a .pdf file is recorded for posterity, a readout of the motley inimitable crew of Amazing Grace and what we each experienced and contributed on this paradigm altering sojourn that drew so many pain wracked tears and brought five very unlikely people together as a team.

Thank you all organizers and personnel involved in the IBM datajam against human trafficking! Thank you a thousand, thousand times! This global fight means so much to each of us on Amazing Grace, and though our contribution ultimately failed to be anything useful, this datajam has produced five resolute warriors in this heart felt cause.
When next there is a datajam against human trafficking, you can be Amazing Grace will return, and this time with terribly promising results.
Truly, thank you!