Presentation Points to hit:

* 10 minutes
* Ask 2 questions on Wednesday
* Use and explain visualizations
* Explain concepts and methods

1. Define the problem. Extra effort for defining the problem and milestones. Problem difficulty
   1. Required by the Bank Secrecy Act of 1970 and within 30 days of initial identification
   2. Purpose: identify individuals or organizations involved in serious financial crimes such as money laundering, fraud, terrorist funding or any other suspicious activity out of the ordinary.
   3. Millions of Suspicious Activity Reports (SARs) are filed a year
   4. While these reports are instrumental in supporting law enforcement, there is a lot of information work through. It can often be difficult to piece the data together to find useful information about criminal behavior
   5. In this project, I set out to identify natural clusters in this public SAR filing data to highlight not only criminal patters but also SAR filing behavior patterns.
   6. However, due to the sheer volume of the data and number of potential dimensions, I decided to narrow my focus to depository institutions (banks) and their deposit accounts in 2020. This helped to interpret the results more accurately in the end because I was controlling for some of the variance. With this, my research question ended up being, “What states exhibit similar suspicious activity for depository institutions in deposit accounts for 2020?”
2. Data Collection Efforts (multiple sources, tools etc) Data difficulty (joins, unions)
   1. Pulling the data was relatively painless. However, the website permitted only a certain amount of records per extract.
   2. I had to pull each suspicious activity types separately.
   3. I pulled the data into R and realized not all of the data types match in each file. So, I had to individually go through each file and update the data types for the continuous variable.
   4. Then I unioned the files together.
   5. Finally, I brought in a single file from the US Census statistics for each state’s population so I can calculate the per capita statistic. I then joined the population data to the unioned sar data
3. Data cleaning effort and tools used (with packages)
   1. I used R for all of my data cleaning efforts, however, I did not need to filter the data to depository institutions, deposit accounts or year to 2020 because the website allowed me to filter upon extract.
   2. The main packages I used in R for cleaning are the dplyr and tidyr.
   3. The biggest thing I had to do was remove irrelevant and pre-aggregated columns in the data, and the Guam territory since I was focusing my efforts on states only.
   4. The data came with Industry, product, and year columns which all could be removed since we already controlled for those columns.
   5. The data also came with pre-aggregated Total
   6. Once I stripped the data of the unneeded information, I could re-aggregate my data to the level of detail in the data set which was State and activity type
   7. Then I calculated the per capita statistic. I decided to use this instead of normalizing the raw data because after initial data inspection, I could see it would compare the states and highlight behavior much better. I then removed the unneeded population and raw data count columns.
   8. Lastly, I had to pivot the data so it could properly be used in the k-means algorithm and replace the nulls produced by the pivot with 0
   9. So the final data looks like this…… new slide
4. Data inspection effort - EDA
   1. A lot of the feature section came from stumbling through the data until I realized there was too much data to work with all of the variables at once. This is why I chose the suspicious activity group and State dimensions to work with to answer the research question as concisely as I could. So a lot of feature selection was done prior to the formal EDA phase through trial and error and a lot of revisiting my research question to make sure I was answering it with the data I chose.
   2. I did find that I ended up with about 3,460,668 suspicious activity observations and 8 columns.
   3. I ran the summary() function to evaluate the spread of each variable. The Money Laundering variable had a massive spread, but we will see why when we get to the conclusion section
   4. I also ran a correlation matrix to verify the was not any major covariance. The relationship between Identification documentation and money laundering was somewhat strong, but I decided to proceed as is.
5. Machine learning and business intelligence methods used. Efforts using analysis methods (optimization)
   1. I started the machine learning phase by identifying the optimal number of clusters for analysis. I used the elbow method as they call it. It searches for the optimal number of clusters by minimizing the total within-cluster sum of squares or compactness.
   2. Basically, here you are looking for the minimum number of clusters that if you add one more cluster, it doesn’t improve the within-cluster sum of squares by much.
   3. Taking what I learned from the WSS method, I wanted to compare the chosen 3 clusters with 2 and 4 clusters to be sure I was selecting the best number of clusters.
   4. Two clusters had an excellent avg silhouette width of .91 which means each observation were well clustered, however, the compactness or WSS was about 81%
   5. Looking at 3 clusters, we can see that the average silhouette width value dropped to .63, but the compactness was higher at 92%
   6. Finally, 4 clusters had great compactness of 97% but its silhouette width dropped to .54.
   7. With all of this considered, the best choice did come out to be 3 clusters which was recommended by the elbow method.
6. Evaluating Results:
   1. Delaware is the only state that makes up cluster 3
   2. Cluster 1 comprises of states that file a higher number of SARs per capita
   3. Cluster 2 are states that generally file a lower number of SARs per Capita
   4. We can also see in cluster 1 that there are more states with spikes for the different suspicious activity types
      1. For example, Ohio in cluster 1 spikes significantly in cyber events and identification documentation
      2. North Carolina – Mortgage Fraud
      3. Hawaii – Structuring
      4. South Dakota – Terrorist Financing
   5. Cluster 2 contains the majority of the states
      1. Huge spike in Alaska for gaming activities and some small spikes in mortgage fraud and structuring
      2. Overall these states have expected criminal activity and sar filing behavior.
   6. Cluster 3: What is happening in Delaware???
      1. Major deviation in money laundering, identification documentation, gaming activities and cyber event
      2. known for their business or corporate conducive environment.
      3. attractive to business owners due to its business-friendly usury laws and light taxation
      4. In fact, before 2017, there were more business registered in Delaware than the total population
      5. Additionally, and more importantly, this state has local laws in place that allow private businesses to protect the identity of their owners
      6. In other words, businesses in Delaware are able to hide the true beneficiary owner for the company
      7. This is a major issue when it comes to anti-money laundering efforts. For example, if a financial institution does not know who the beneficiary owner is of the company, it makes it harder to determine legitimate operations from shell companies. This allows criminals easier access to launder money in the financial system.
7. Problem solving and learning new things independently - What did I learn