# Data Tools 1

# Final Project

Sex Trafficking (FBI Data)

Code Version 1.0

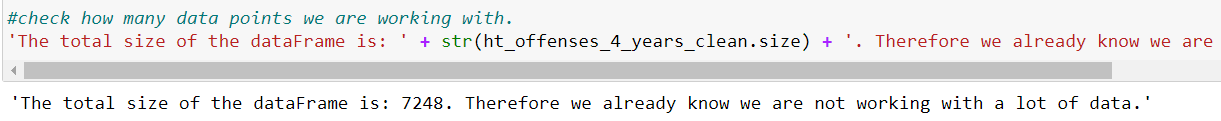
Overview

A review was conducted of the Sex Trafficking Data that was made public by the FBI. This data includes, but is not limited to, the following characteristics:

|  |
| --- |
| * Population Data by State |
| * State Regions |
| * Sex Trafficking Offenses |
| * Sex Trafficking Clearances |
| * Arrests by age (Binary: Adult, Juvenile) |

Human trafficking is the business of stealing freedom for profit. This “business” tricks and/or forces victims into providing commercial sex under cruel or illegal conditions. It is a billion-dollar criminal industry that imprisons 25 million people around the world. With high level description of the problem, the goals of the review is 1. confirm there is still a problem today, and 2. find if there is a predictive pattern in the FBI data that I could extrapolate from the data in regards to how the sex trafficking was being conducted and/or operated by using some of the tools that were taught in this class. Currently, all analysis characteristics are imported as raw data therefore I had to clean the data prior to performing any exploratory analysis.

*Research conducted concluded that although it is easy to see which regions are more involved in the sex trafficking, there is not enough data history to develop a sophisticated enough model to produce consistence predictive output that would prove beneficial for professional use. The total size of the “ht\_offenses\_4\_years\_clean” dataframe is row x column = 7248. Below shows the output that was used for to gather the size.*



## Scale of problem

I believe this would require a Large-scale engineering task develop a predictive model, for the following reasons:

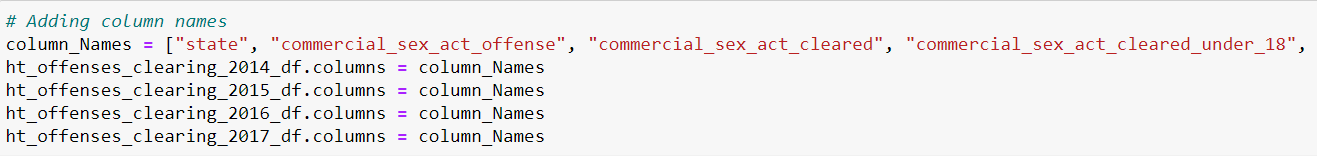
* 1There are roughly 4 million people around the exploited by sex trafficking, annually.
* Its not just young women that are the victims of Sex Trafficking. Victims of human trafficking can be children or adults, male or female, come from all backgrounds, and economic levels. Children as young as 9 can be targeted for exploitation. Given that the selectivity among who gets forced into trafficking is not unique. A model must have lots of other information to be useful.
* Its about power and control in most cases. And unfortunately for the good guys, the “bad guys” are smart too and some research suggests that the “bad guys” use data science to fight the “good guys”. Also, the “bad guys” use predictive modeling to try and predict their own moves, keeping things as random as humanly possible.

## 1. Research Stages

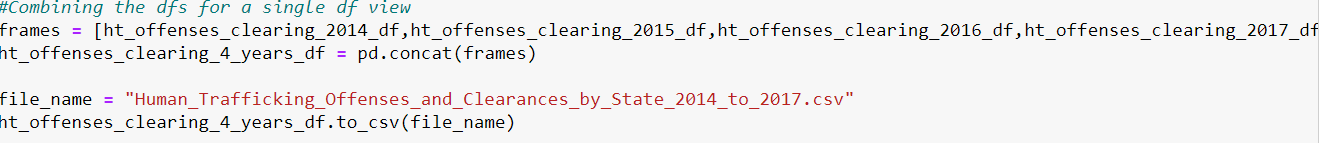
The first step of my process was to clean my data. Given that the accumulated data is in a raw form, meaning, it was not organized in such a manner that is caters to analysis. Therefore, I had to spend a significant amount of time and percent of development work trying to figure out the best way to organize my data. Given that I was not starting with a large amount of data the amount of cleaning was not substantial. However, the amount of time spent on trying to organize the data was because it is hard to extract information from a smaller set of data thus relying on its organization.

Ultimately, I decided to import the four years’ worth of trafficking and abuse data into 4 separate pandas dataframes and clean them separately. At this stage I had the idea I would be combining them at some point, but combining dirty data proved difficult to do because errors would arise from inconsistencies in between the dataframes. Finally, after the data was organized and combined (initially) I learned the process at which the FBI collects their data. Generally, the data is separated by age and is categorized as adult and juvenile. If an individual is over 18 they are adults and under 18 they are juveniles. Further, there are two types of offence classifications, the first being Commercial Sex Acts and the second being Involuntary Servitude. After learning this information, I then regrouped the dataframes as shown in code block 1.4C. Below are images of how I completed the initial cleaning of the separate dataframes and of what the final, joined, dataframe looks like.

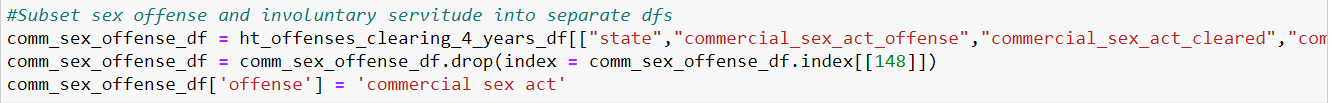
*Code 1.1: Unification of Columns*



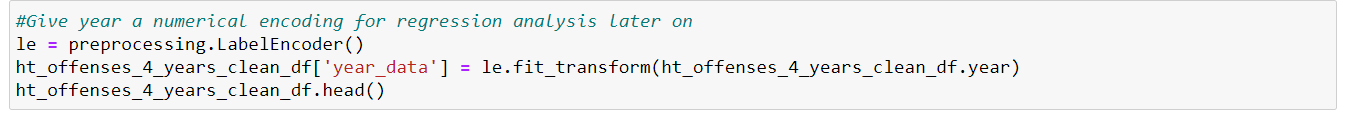
*Code 1.2: Initial joining of dataframes*



*Code 1.3: Splitting dataframe into Sex Offense and Involuntary Servitude*



*Code 1.4: Encoding year\_data*



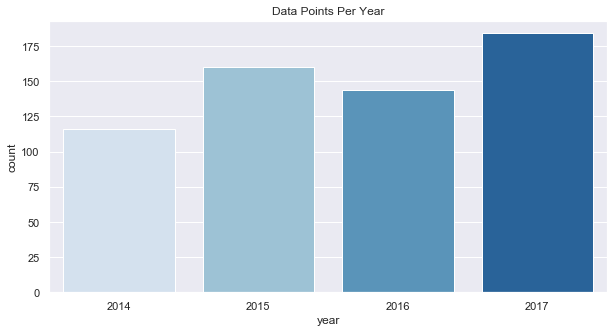
*Code 1.4: Final dataframe*

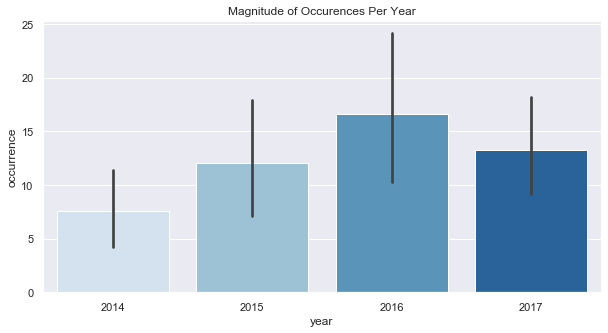


The second step of my process was to visualize my data to see if there is a trend that answers the question of “is there still a problem today?” and insight that will guide us into answering the second questions which of “can a predictive model be built from this data?”.

If we look at *Figure 1.1* we observe that 2017, out of the 4 years of data, has the highest count. This was due to a couple states having data available for collection. In the past states were not in a hurry to publish this data and as we all know the gears of the government work slowly. Nonetheless based on the initial bar chart my thoughts led me to believe that 2017 would have a higher trend of Sex Trafficking cases when compared to prior years. However, what we observe when we view *Figure 1.2* is that although 2017 may have the most data, the chart shows a downtrend in sex trafficking occurrences in 2017.Note that there is some significant uncertainty around the number of occurrences as shown by the black “error bars”. Due to the lack of quantity of data and consistency across years this magnitude of error does not prove to be a big issue. Based on these first two figures we can answer the first question we prompted ourselves in the beginning which was “Is sex trafficking still a problem today?”. Because both of those figures are above 0, the answer to that somewhat naïve question is an emphatic yes.

*Figure 1.1: Quantity of Data per Year*



*Figure 1.2: Sex Trafficking Annual Trends (Commercial and Involuntary)* 

So if it is still indeed a problem today then another question arose for me, can we take the exploratory analysis a step further and determine where the hot spots are based on visual data? Referencing *figure 1.3* we can see that each region is made up of a different number of states and that is why the “count” on the y-axis is higher for some than others. So, initially this made me think the data could be bias towards these regions that have more state representation. However, when we look at *figure 1.4* the chart tells a different story. It is obvious that the number of actual offiences occurrs in the Gulf Coast West. Even though this is faily obvious by the chart, I felt that peeling back another layer and looking at offences by region by year would give me a detailed view of consistency or inconsistencies across regions. Finally, by reviewing *figure 1.5* it is overwheliming how often the Gulf Coast West is the biggest perpatrator of sex trafficking occurrences. Thus we are able to answer where the hot spots for trafficking occur, and the first answer is in the Gulf Coast West. This is quite the indictment considering *figure 1.3* illustrates the Gulf Coast West as one of the smaller categories based on count of state representation.

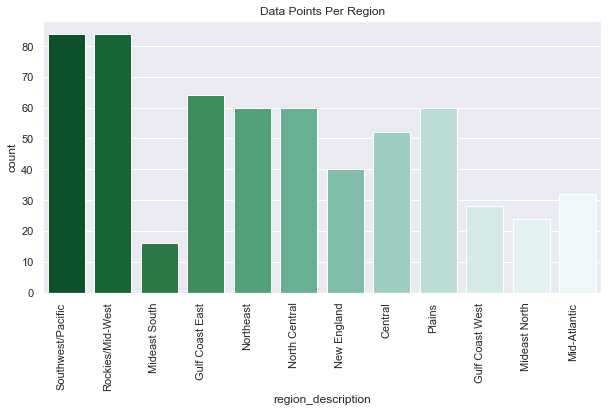
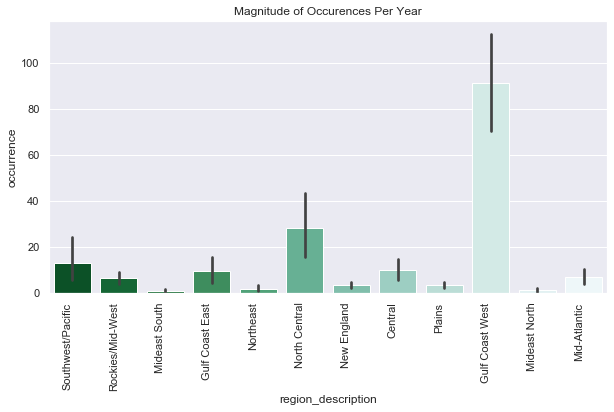
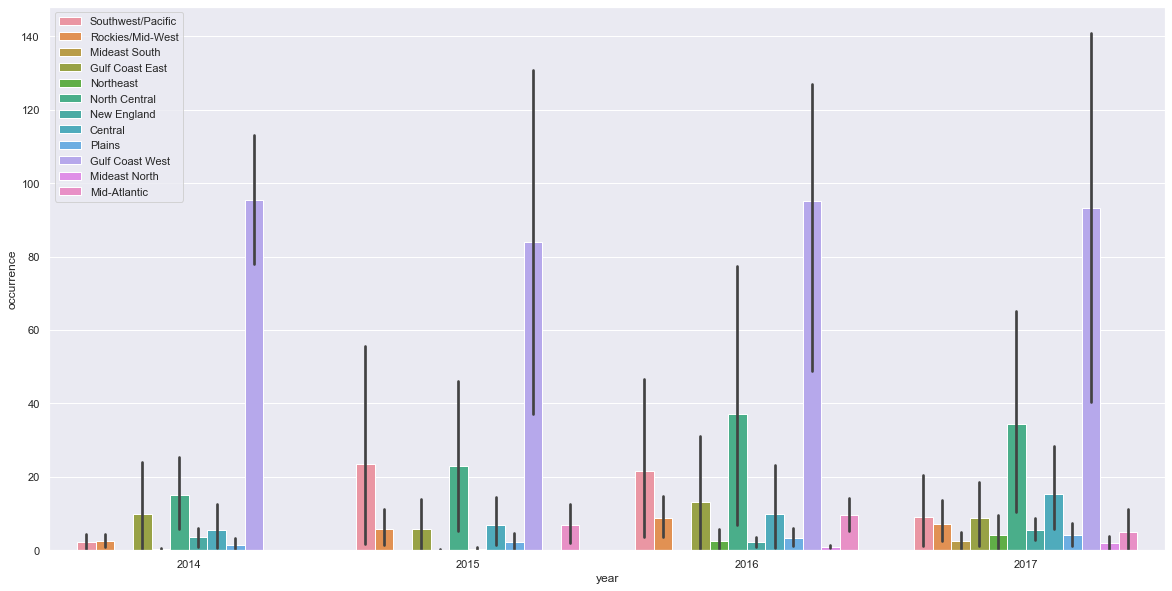
*Figure 1.3: Quantity of Data per Region*

Figure 1.4: *Sex Trafficking Regional Data*

*Figure 1.5 Sex Trafficking Annual Trends per Region by Year (Commercial and Involuntary)*



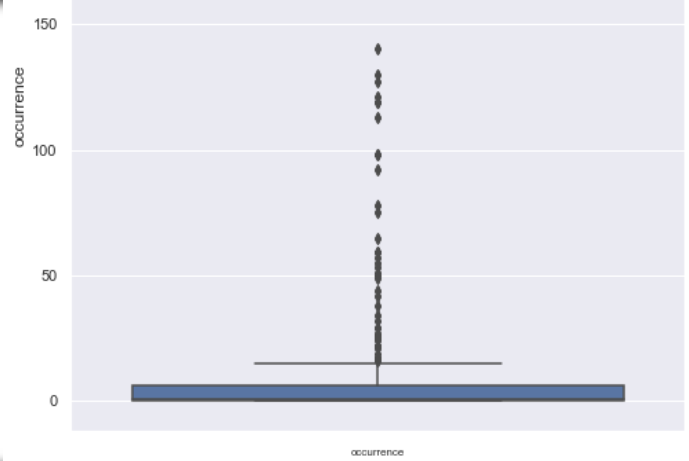
The final step of my process was to determine if I could build a model that was predictive and reliable. One of characteristics of a good model is good input data. Regardless of how good the model is, if you feed it garbage you will get garbage as an output. With that thought in mind, I needed to take a quick look and see if I the data is reliable. *Table 1.1*, below, shows the summary statistics for all four years’ worth of data. At first glance, there is significant variance in each variable of data which does not bode well for output stability. Meaning, in any given period of data production the datapoint could be significantly different then the prior data point thus causing variation in the predictability of the model. For example, after we standardize the number of occurrences by region per 100K of population, the variance, .44, is 2x the arithmetic average of .22.

*Table 1.1: Cumulative Annual Summary Statistics by Column*



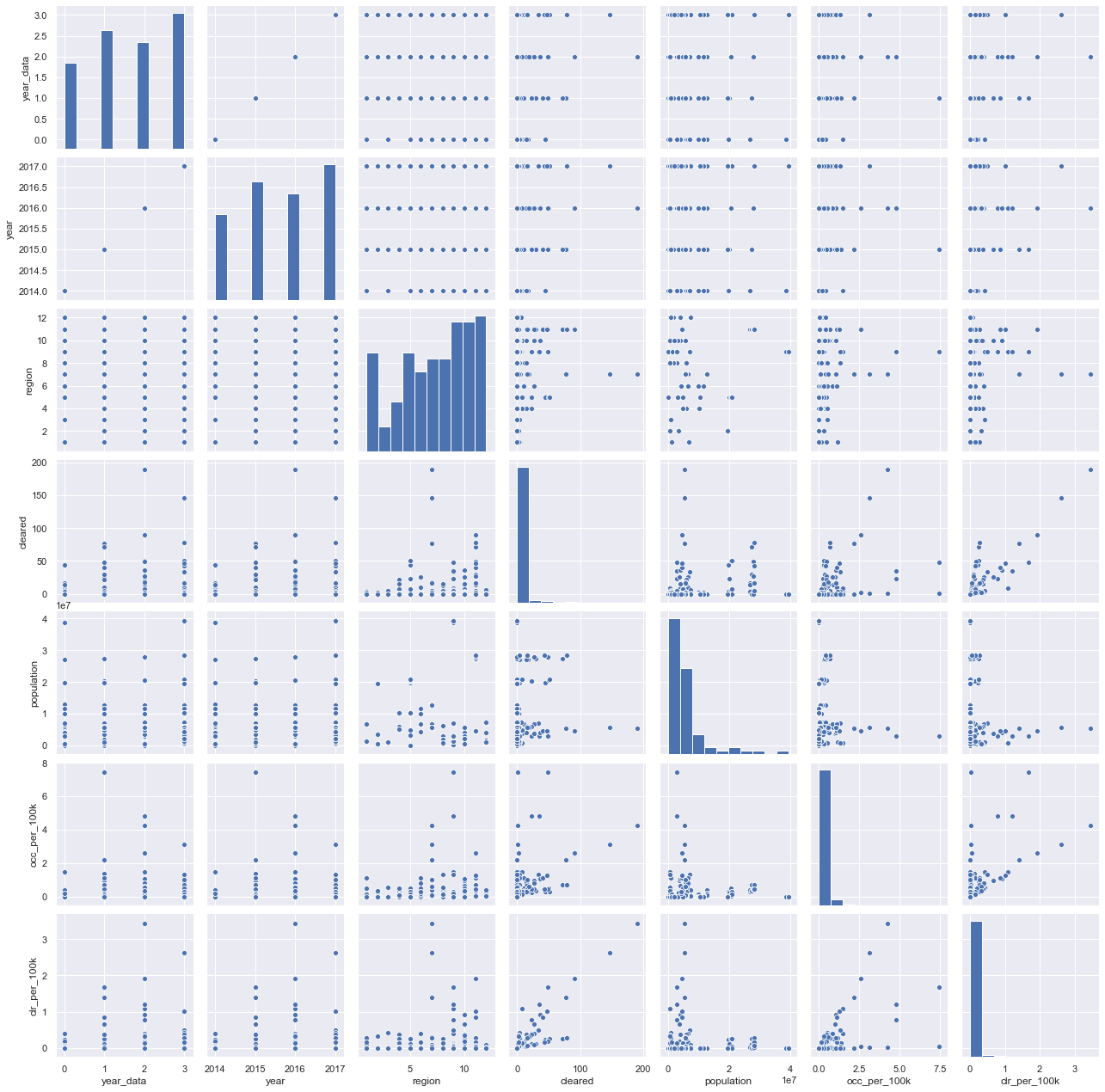
Unfortunately, the significant variance in the data is problematic given the small quantity of total data points that we are working with. As referenced in the Overview section of this paper there a total of 7248 data points (rows times columns). Granted, the large variance may be described by the lack of data, but nonetheless it will be hard to engineer any sort of reliable model when the data is of such low quality. The last real data quality check I conducted was for outliers. I had an idea from the high variance that there would be a large quantity of outliers. I checked all columns but *Figure 1.6* gives the idea for what I was seeing across all variables, which is a massive amount of outliers. Statistically this makes sense because if the data has high variance, then more data is in the “tails” of the distribution, which would result in more outliers. At this point I have decided that a predictive model cannot be built with this data.

*Figure 1.6: Cumulative Annual Summary Statistics by Column*

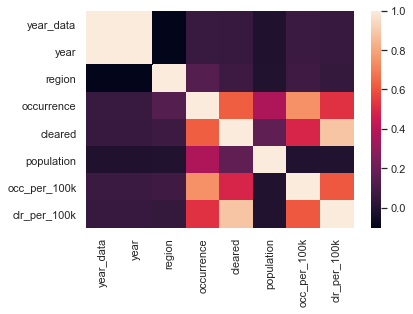


Even though I have chosen not to build a model due to data reliability issues; I could still analyze if there is a significant relationship among characteristics. By using a Seaborn Pairplot I was able to output *Figure 1.7.* This figure provides a high level overview of the relationships between characteristics. Some of the relationships do not make much sense, others do not show much relationship, and the rest have some sort of relationship. For example when referring to the figure we can see a relationship between the number of cases cleared and the population standardized by 100K of population. Again, this makes sense because as long as clearances is above 0 then it is in someway a function of population.

*Figure 1.7: Cumulative Annual Summary Statistics by Column*



Taking the relationship amongst characteristics another step or into another view I created a correlation matrix represented as a heat map. Reviewing Figure 1.8 we can see that a significant portion of variables in the lower left or upper right have little to no relationship with each other. As our viewing moves down and to the right we start to see some sort of positive relationship between variables. As a confirmation for what we observed in Figure 1.7, “occurrence” of sex trafficking cases is related to number of clearances per population, “clr\_per\_100k”. Meaning, as the number of occurrences increases so does the number of clearances. Other relationships are not ones that we could pull significant information from. What I mean is of course the number of “cleared” cases is highly correlated with the number of cleared cases per 100k of population, “clr\_per\_100k” because it’s the numerator of the standardization.



## 2. Conclusion Recommendations

When I started this research project, I had hoped to answer two questions. The first question was “Is sex trafficking still a big problem today?” and the second question was “Am I able to build a reliable model that could find predictive patters in the FBI’s data?”. The good news is I was able to answer both questions during the research. The bad news is that neither of them were answers I had hoped for. Regarding the first question, sex trafficking is still a huge problem today and something that data science could address. Granted it is unrealistic to say the data science community could get the numbers to zero it is not unrealistic to think we could find ways impact the numbers significantly. For the second question, as much as I can hope to find the end all be all model solution, I cannot. Although the reasons I cannot engineer a predictive model may be extended beyond the limitations of the data, the data limitations is the reason that no model is the solution. As mention earlier, regardless of how good the model is constructed if you feed it garbage it will give you garbage as a result. And as can be seen from the correlation analysis there is not much information in the limited amount of data. Further the variance of the input data would cause too much instability in the outputs of the model. Therefore, in conclusion, I would need more data in order to build a proper model to tack the issue of sex trafficking.

## 3. Appendix: Citations and Other Important Code Snippets

1Senechal, Isabelle. “How Some Researchers Are Using Data Science to Fight Sex Trafficking.” *America The Jesuit Review*, 2020 American Press, 12 June 2020, www.americamagazine.org/politics-society/2020/06/12/how-some-researchers-are-using-data-science-fight-sex-trafficking.