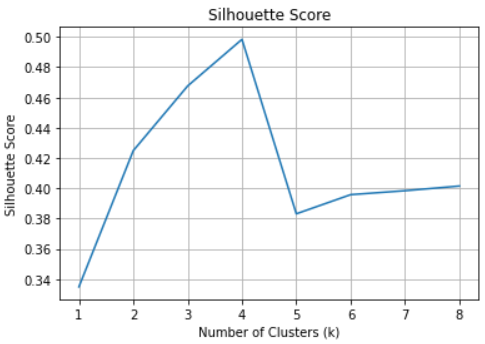
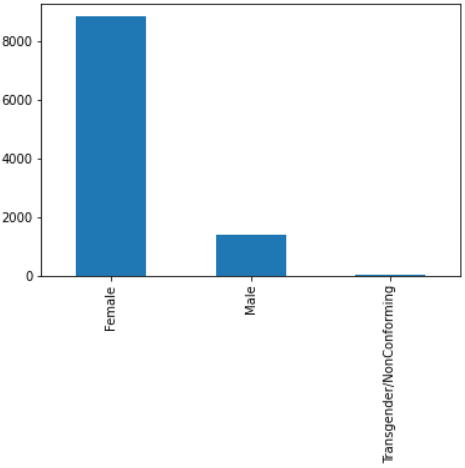
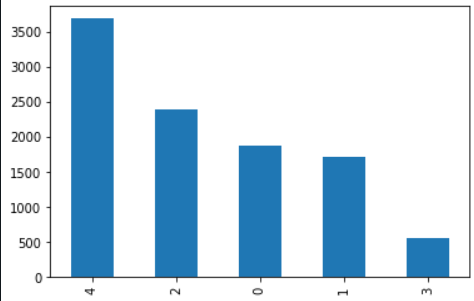
Clustering with PySpark

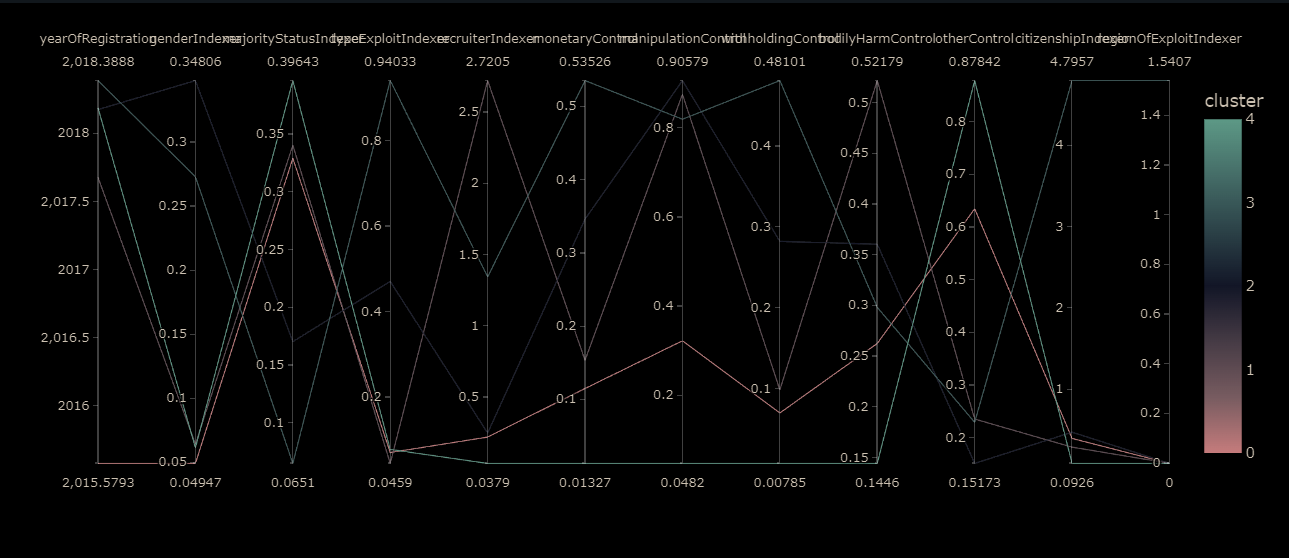
The dataset that I chose to work with was obtained from [The Counter Trafficking Data Collaborative](https://www.ctdatacollaborative.org/page/global-k-anonymized-dataset). This is an anonymized dataset that represents over 156,000 victims of human trafficking from across 189 different countries. It contains tabular data with a mixture of string and integer data types. The CTDC collects various information from victims and survivors of human trafficking such as age (at various times of exploit), gender, type of exploit, citizenship, type of control used, etc. The biggest challenge with this dataset were the large amounts of nulls across all columns. While I started off with 97,750 records, by the end of the cleaning I was only able to analyze 34,612 of them.

This was approached as a clustering problem, to try and identify areas of commonality across victims. The KMeans clustering algorithm was used as it has a straightforward and interpretable application. An optimal cluster number (k) was identified using the silhouette score and visualizing the scores via the plot below. The optimal number of clusters was chosen to be 5. It should be noted that the silhouette scores are all fairly weak, thus it is likely that either the chosen clustering algorithm was not appropriate or there is an underlying issue with the data. Due to the large amount of data loss during clean up and manipulation of the data prior to modeling, I think it is most likely due to an underlying issue with the data.

Streaming was incorporated into the pipeline after the optimal number of clusters was identified. The training set was fit to the pipeline, then the test set was repartitioned and saved to individual files to simulate streaming data. The test data was then read in one file at a time through the stream, transformed via the pipeline, and a query was performed to show the input features and output cluster prediction to test the streaming capability.

An interesting observation from the analyses were that there were a much higher number of females identified as victims of human trafficking (as seen in the plot above and to the left). According to the prediction distribution graph on the right above, there were more data points assigned to cluster 4. The graph below helps explain the different features associated with each cluster.

As mentioned earlier, one of the biggest challenges with this dataset was the gaps in available responses. The 0 for citizenshipRegionIndexer and regionOfExploitIndexer correspond to the region of the United States. Cluster 4 is identified as having the highest number of assigned points. We can see from the above plot though, that several of the other clusters map to 0 or close to 0 for both features as well. Is this because human trafficking is a greater problem in the United States than elsewhere? Is this because the information is being obtained better/easier? Or is there something else that came into play while cleaning and prepping the dataset that affected this outcome? These questions would all need to be more deeply investigated before taking any of these results too seriously.

Ultimately, I believe this calls for an evaluation of how the data is gathered and reported. This is a precarious area to delve into. More complete data, and thus more complete responses, would give more insight to underlying patterns across victims and survivors. However, these data points that are being looked at are the lives of humans that have been tremendously, and often horrifically, affected by this situation. It is important that the utmost care be taken while aiming to obtain more complete records. Survivors can play a key role in helping solve an ongoing problem, but it is not their responsibility alone to solve it. We must remember to treat them with care and respect, that they are not just numbers to be crunched in an algorithm. If victims and supportive organizations work together, insights and benefits can be made for both parties.