**Data Notes for Presentation**

Overview:

* Goal was to find datasets including victims from around the world
* Datasets consisting of historical tabular data ranging from 2002-2018
* This led us to the decision to use a SQL-based database solution for our project **(NEXT SECTION)**

Data Sources:

* Polaris – main source of data containing victim’s demographics and how they were controlled/exploited
* World Bank and World Economic Outlook – wanting to explore how economic factors could have impact on human trafficking
* Data Hub – geojson file with all the countries
* Nations Online Project – country codes with ISO codes and numeric codes **(NEXT SECTION)**

Database:

* Used PostgresSQL hosted via Heroku
* Connections were made using SQLAlchemy
* Our database consists of 6 primary tables **(SCROLL TO ERD)**
* Country table was used to join other tables due to varying country codes **(NEXT SECTION)**

ETL Process:

* Our Polaris data set initially had 63 columns and around 48,801 rows of data
* Cleaned and prepared the data by reviewing missing data and consolidated columns to prevent data loss
* Many columns that contained binary values we condensed into a single categorical feature
  + Use labor type example – 17 columns condensed into single column
* Feature engineering: categorical data to numeric values for models

Santiago will now be presenting how our data was used to visualize our findings on human trafficking **(VISUALIZATIONS TAB)**