**Executive Summary**

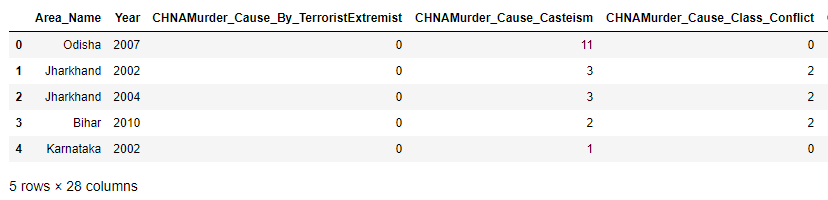
The objective of our project is to create a database to query air quality data against murders/homicides in India. This was done with the purpose of examining if there was a relationship between air quality and murders/homicides in the different states of India. The data is during the period of 2001 through 2010. We gathered different air quality measurement data by state and murder/homicide data by state. The crime data is from the National Crime Records Bureau (NCRB), Govt of India. The air quality data is from the Ministry of Environment and Forests and Central Pollution Control Board of India. By combining this data into a database, we were able to conduct our analysis.

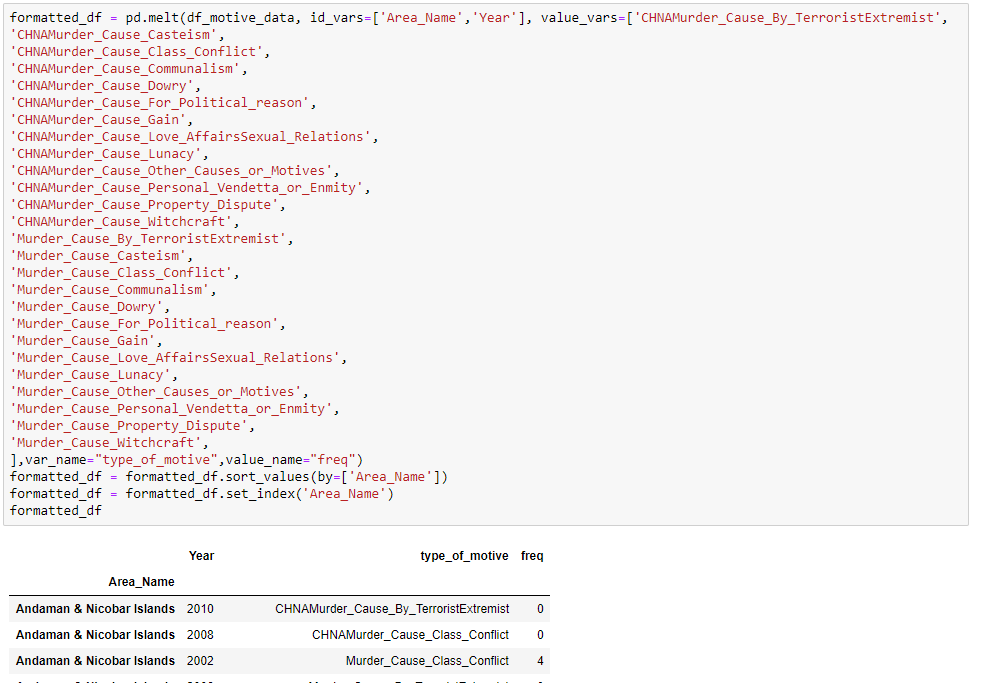
**Data Cleanup**

The goal of this project was to explore potential relationship between air quality of an area and crime rates in that area. To get this done, 4 CSV datasets from Kaggle were obtained- one contains detailed info on air pollution in all states in India, while the others had a breakdown of crimes committed in the said states, in a way that each additional spreadsheet provided granular look into the type of crime - broken down by tools used to commit it, demographics and motive.

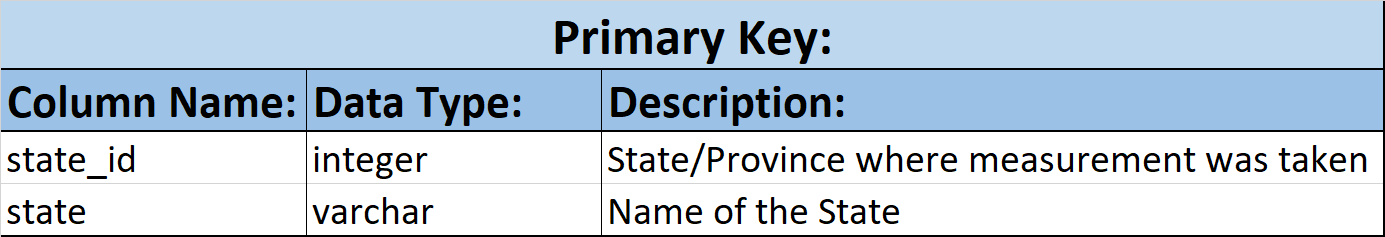
CSV files for both datasets were read, and converted into Pandas Dataframes. Dataframe containing pollution information required rather basic transformations - several irrelevant columns were dropped, “NaN” values in remaining columns were replaced with “No Data” as the value of ‘0’ indicated infinite absorption of such particles. See more below on the explanation of air quality coefficients used in our analysis. Additionally, index was switched to State, and a relevant time frame was selected using pd.datetime.

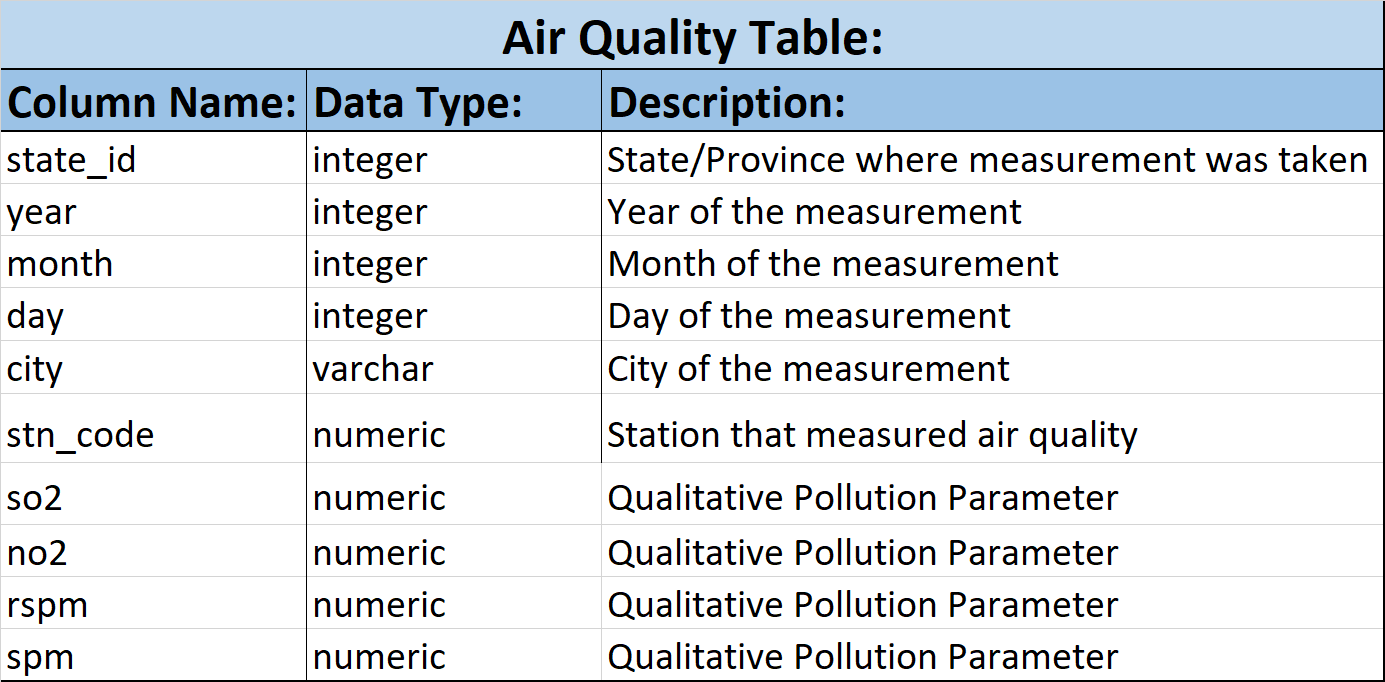
Same cannot be said of the 3 supporting spreadsheets. One of the biggest challenges we’ve faced was with the format of the spreadsheet in its’ original state. But one of the instructors who shall remain nameless directed our team to tidy-date-in-python document. the below transformation is what came of the new skill. Transposing the dataframe using pd.melt was challenging but also educational. Besides that we’ve also removed useless columns containing total values and replaced “null” with 0. (in this case it actually meant zero observations)

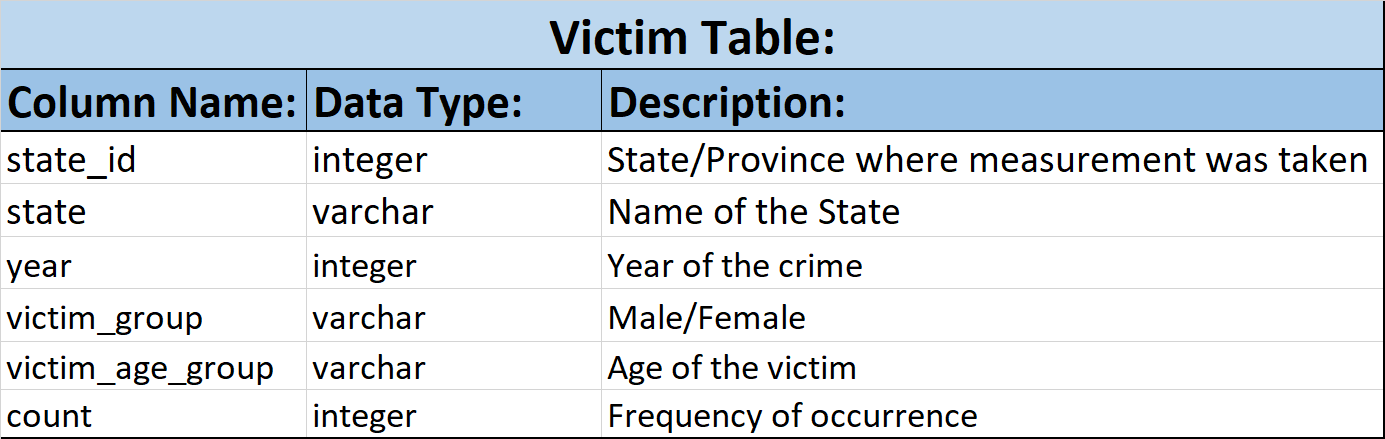


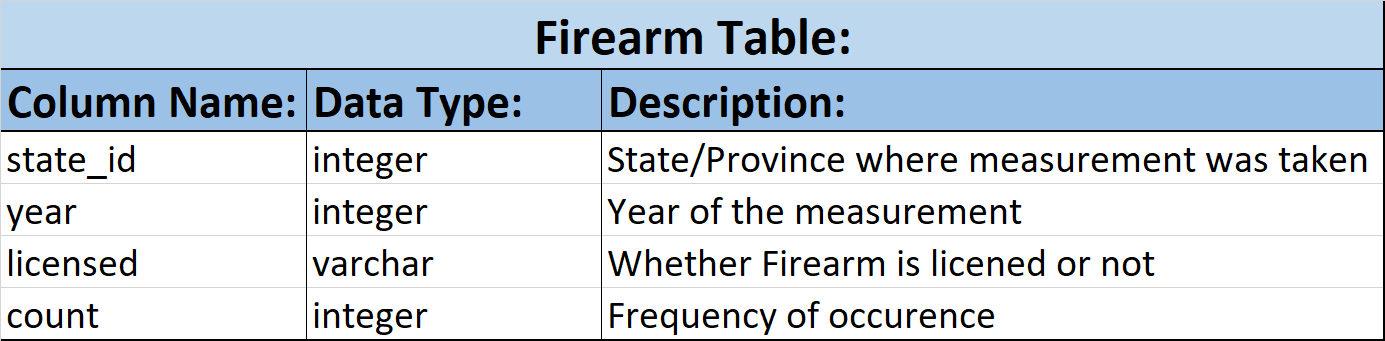


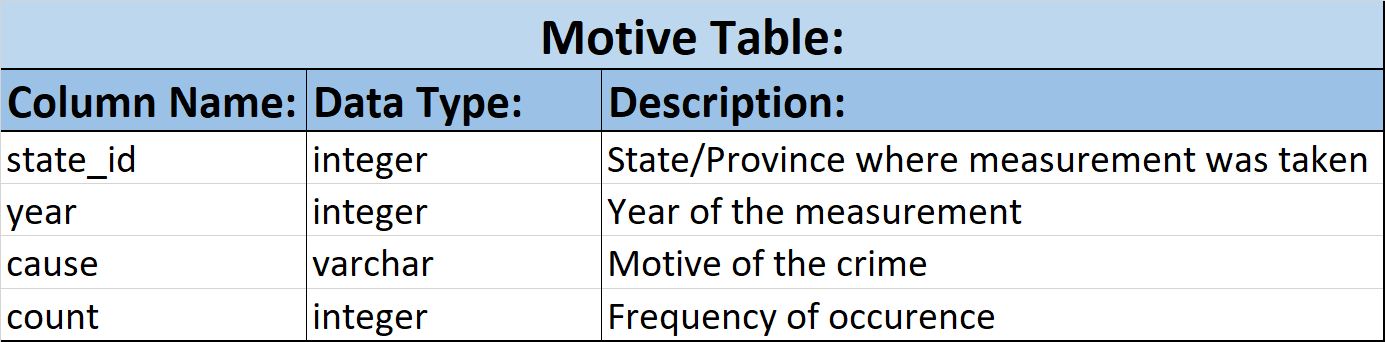
**Data Dictionaries for each Table**

****

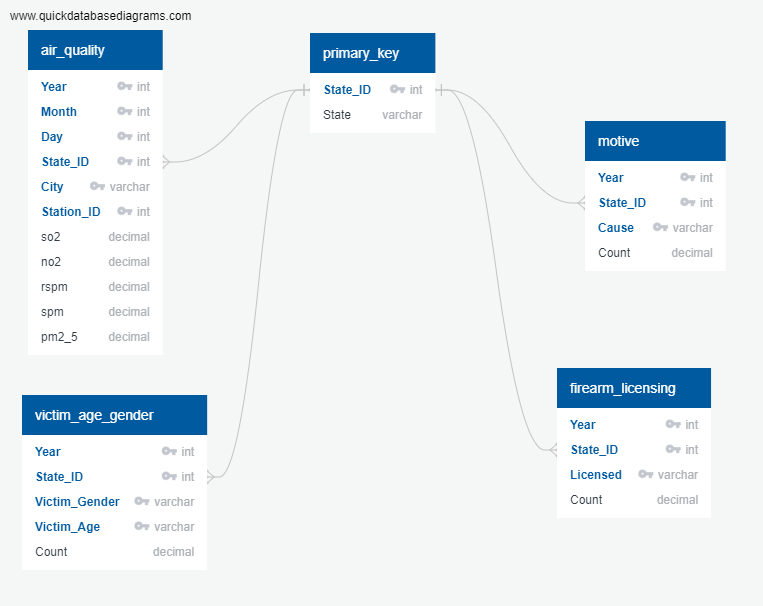
****

****

****

****

**ERD Diagram**



**Queries**

