We found figuring out which datasets/perspectives to pursue difficult because we wanted to collect data from a large dataset and work from the perspectives of homicide/forensics. We have two big datasets we want to work with, so we couldn't choose one- I guess that's one important concept we understood (using other datasets for supplements and context/comparison is important for conveying a message). That being said, we came to a consensus and decided to review/use both datasets and use them for comparison. We hope to use the U.S. murder dataset and compare it with the country's homicides large dataset (see if there are differences in rates/ features between the two values. For that reason, our team decided to inspect either dataset and we can compare findings and think of a way to integrate it.

After loading the UN's dataset (country homicide), I decided to inspect the data and try to piece together a narrative and questions to address. The column headers in the dataset were as follows: Region, Subregion, Country, iso3_code, Indicator, Disaggregation, Gender, Source, Unit, Year, Value, and Footnote. Iso3_code is just a standardized way to identify countries and Disaggregation is an unspecified variable (but we'd been thinking it was related to the classification of homicides (who was it)- looking at the category (the data was about if it was homicide by family/ partner or someone else. However, that category had an extensive amount of " -" (like it looks like a space).

Moreover, the main components we were focused on were the place/ region where it happened, year (when it happened), Gender, Value (the number of Homicides recorded that year), and Unit (how the value is being measured). Checked for missing values and noticed that besides the issue I started early about the Disaggregation, only the Value and Footnote had missing values (Value- 42, and Footnote- 17018)- out of 24392. Value's missing data could be imputed with median/mean value from homicide. However, footnotes won't be incorporated into the dataset (however, we'll use it for context). Not to mention, there is a variable "Source", but we don't think it'll be relevant to identify patterns.

Looking at the timeframe of this dataset, it starts from 2008 to 2020. It will be interesting to see the progression/ trends of the homicides over time. It also looks like, besides the year variable, the rest of the columns are object data types (which may mean we need compatibility between Year and other regions.

Thinking about how the variables/ data we collected could be used to address a question/ produce a narrative/ help make projections- we were considering using variables like Value, region, year, disaggregation, and gender. Of course, these selection of variables are subject to change, but we were thinking about trying to do a few things: Were there certain years where a particular country/region had an increase in homicides, Which years had high values of homicides (with units to present as metrics), and Possible aggregate/ collect the number of homicides that were related to IPFM (it's not clear what this stands for but we will define it later on the project). Moreover, I think these data might be useful for prediction/ classification purpose models. Maybe using a machine learning model to cluster similar information-based data and make a prediction about future projections.

I think the major challenges with this dataset are imputing the missing values, understanding the purpose/ meaning of some dataset values, and making future predictions. Let's start by addressing/ explaining the missing values problem. Although most data are present the volume footnote, and disaggregation column could pose a problem. The value column has 42 missing values (out of 24392) footnotes have 17018 missing values and disaggregation- although doesn't have missing values- most columns have this symbol "-" which could pose a problem of interpretation (was it supposed to be represented as a space, or does it have an underlying meaning). We'll have to do some more research, but if it represents missing values, then we'd consider accommodating that column by sourcing/ getting context with other sources. The value column only has 42 missing values so performing dropna() wouldn't be a

problem. Footnotes also won't be a problem since we'll be using them as a supplement, but we were curious as to why some of them did have footnotes (possibly from other sources).

Although we haven't cleaned/ performed EDA on the dataset yet (still researching more datasets/ looking into the unclear meaning), we have developed a thorough plan/ timeline to be successful in this project.

Next steps:

Clean both datasets (or other datasets we plan on using)

Perform EDA/ visualizations

Start thinking about machine learning models we could apply to tell a story/ answer underlying questions about trends.