https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

This will be my data for the final project. This data has a massive and extensive amount of data. There are exactly 2,845,342 entries in this data, which is nearly 3 million total entries. There are some ordinal variables such as accident severity on a 1 to 4 scale (Severity), DateTime variables such as (Start_Time and (End_Time). There are locational variables such as (Start_Lng), (End_Lat), and several others. There are categorical variables such as (Street), (Side), (City), and many more detailing the locations of the crash. There are more on the conditions including normal quantitative attributes like (Pressure), (Visibility), and more. In all there are 47 different attributes in this data allowing for a great amount of potential analysis and data processing.

The usage scenario for this data would be a policy maker in the United States. The number of fatalities and injuries from car crashes in the US is very high. Someone seeking to reduce these may want to know what is actually causing these events to happen and see if there are things that can be done to fix them. They will want to see various factors, including those that are controllable and those that are not. For example, the ability of government policy to control accidents correlated with driving at night or in bad weather, may be difficult to control. But if there are an increased number of accidents in certain states controlling for these factors, there may be something that can be done. For example, seeing if crossings cause more accidents or various other factors. These are things that government policy can determine and being able to see the correlation between these things can be helpful to a policy maker. There are various things that a policy maker may want to know, and this data needs to be able to demonstrate these things.