Pre-Analysis Plan
Angie Vasquez
DS 3001

Research Question

I will be using Border Crossing Entry Data provided by the Bureau of Transportation Statistics (BTS) to analyze if there are seasonal patterns to predict border/tourism crossings from the US-Canada and US-Mexico border. The research question guiding this analysis is "Can seasonal patterns in border crossings predict monthly cross-border traffic at major US borders of entry." The data set from data.gov provides reliable and historical data that can be used to explain why traffic patterns may vary seasonally. It will also be able to go beyond seasonal patterns, exploring the relationship of other predictive factors.

Outline

An observation in this study is monthly entry count at a specific location, in this case port. This is categorized by vehicle type (truck, bus, personal vehicle, etc).

This would be an unsupervised learning analysis as I will not be using labeled trained data. It will also focus on regression as the goal of the analysis is to predict crossing volume.

Linear regression will be used as a baseline for the prediction model, as well as Lasso regression as it can help determine which factors contribute the most to crossing volume (while reducing overfitting).

To measure success, it will be determined if a model can predict crossing volume and go beyon by exploring other factors that may be impactful. Certain metrics can measure and answer the research question. For example, Can R^2 measure the variance in crossing volume as shown by seasonal factors, and can RMSE help assess prediction accuracy.

To prepare this data set for the analysis, I can clean and filter the data by changing some to dummy variables. This way, I will be able to use categorical data in the model so that I can have qualitative factors, whether that be location or vehicle, represented.

An anticipated weakness is that seasonal factors within the data will not be enough to influence cross-border traffic as there may be other external influences, such as economic trends or policy/administration changes, that could also create variations. As a result, LASSO may be useful in helping irrelevant features. Another weakness is that focusing on seasonal factors may lead to overfitting, so utilizing cross-validation will be important so that it is not centered on short-term patterns.

If this approach fails, an insight to be gained is "what other methods and metrics may help reach a more accurate prediction of the research question." More specifically, it will be able to demonstrate if border crossing patters are not as driven by seasonal factors as originally expected, showing the need to consider additional sources and models that may explain crossing volume.

To present results, I will create visualizations, such as bar charts for monthly crossing volumes and tables for comparing model metrics to show if there is predictive success and seasonal factors.