*Capstone 1 Observations*

**On use of pyspark vs simple pandas**

First of all, it was extremely helpful to have a nicely scaled-down version of the dataset. Much of my code was drafted and debugged using this smaller data and it also helped point me in the right direction for data exploration, engineering/processing and analyses without getting bogged down with execution time.

When I did flip the settings to use the entire dataset, I was surprised and how well it the standard pandas/numpy functionality still worked without much consternation. As long as I was careful with my garbage collecting (i.e. deleting any variables/references) and regularly uploaded the (cleaned) dataset with targeted features for each section, I was able to run the code without interruption. Even the clustering steps I took with the weather data did not take too long to complete.

While this remains a work in progress, I do not believe the final notebook would be exclusively in pyspark as there are some steps that I think are better done with the familiar pandas functionality.

**On working with this dataset**

This project highlights a key insight that I often have when working with data analysis: there is so much insight to be had from creative (but not necessarily complex) visualizations and inventive feature engineering. Building fancy machine learning models can be very useful and necessary for practical purposes but, often, my time as a data analyst is spent working with and manipulating raw data and visuals to inform model choices. That said, this project really highlighted how uncomfortable I am with Tableau. I feel much more comfortable with some of the visualization tools from standard python libraries including plotly for some more sophisticated illustrations.

It also highlighted how helpful it was to take the time to understand all the features—although, I do not think this was the raw data from the APIs in the academic research papers. I also enjoyed the opportunity to research other dataset options—like the census info on commutes—as a way of supplementing the original data.

The biggest frustration with this project was how little information was contained in the statistical tests that I performed across all these features. With such a large dataset—even with the scaled-down version—even the smallest of absolute differences in the target features were considered statistically significant. Especially with severity, a discrete variable with limited range, using seven million observations to conduct a statistical analysis was not going to give much by way of actional recommendations.

The biggest limitation of this dataset was the inability (at least without doing extensive additional data scrapping/wrangling) was the lack of “control” observations, for lack of a better term. For the weather and infrastructure-related features, it would have been useful to develop a way of tracking how often accidents did not occur in order to analyze how much these features contributed to the rate or probability of an accident(s).