Name: Andrea Fox

Date: June 20, 2021

Course: DSC680 – T301 Applied Data Science

Assignment: Project 1 – Project Check-in/Milestone 2

1. Any surprises from your domain from these data?

No surprises from the data necessarily. It was a much larger file than expected, so it does take some time to process it when I start. It also included some columns I had not thought to use or even think about so that was a bonus. There are 47 columns total included in the file, but I chose to remove several of them that I did not think would have any influence on my overall analysis and project.

1. The dataset is what you thought it was?

I mentioned this above, but there were a few things included that I had thought about so I may be able to use that as well in analysis. Mostly a few extra weather-related variables along with the different sunsets, which I am going to have to look up the difference on. There were several nan values that came as a surprise especially with weather related topics, so that will be something I am thinking about when I start working with the data and the different methods. Thankfully, Sam Loyd posted an article that I remember seeing often that talks about the different ways to handle missing data because there are reasons to keep or not keep.

1. Have you had to adjust your approach or research questions?

I will have to adjust some of my questions. One, because as Professor Williams called out, they were more EDA related questions instead of data science project whole related. Two, I was missing quite a bit of weather data. I know how bad the Midwest can be when we get precipitation, and we are used to having crazy weather. You add snow to say Atlanta, and it shuts down the city completely and traffic gets crazy. Granted that is more of an outlier situation, it still plays into my thought process for this project.

1. Is your method working?

Sadly, I have not reached my method stage yet as I spent way too much time trying to clean up my data and update values to be int instead of Boolean or strings. I was having issues replacing some of the values into integers, but Sam Loyd saved the day again and led me in the right direction. Unfortunately, it meant replacing values column by column but it worked so I am happy. Once the cleaning is finished, I will load into PowerBI for some visualizations on the current data, then also create a separate notebook for splitting into train and test then working through testing it on different method types. I stated in my proposal that I plan on using Naïve Bayes, but again I think it does not hurt for me to look at multiple methods for the practice if nothing else.

1. What challenges are you having?

Time has been a big challenge for me this past week. I spent quite a bit of time cleaning and organizing my GitHub that I did not give myself a ton of time to work on this milestone. I also managed to get bronchitis which had me down and out for a few days. The other challenge I am having is the sheer number of records I am working with and the fact I could not get my values replaced with 0’s or 1’s. I had to save off a csv that I had cleaned up and then change those values manually. Overall, I am not pleased with my progress and understand if you take points off for not being farther along on this. I have not managed my time as well as I would like, and of course outside influences did not help the situation. I will say this week I will be focusing completely on the project work to be ready for milestone 3.