DSC680 Project 1 Check-In

# Katie Briggs Winter 2020 <https://github.com/Briggskm9/DSC680/tree/main/Project%201>

# Domain Surprises

There were a few surprises within the domain perspective. This project looks at the sun cycle, monitors solar flares and winds to determine a kp index which alerts and forecasts the chance to see northern light, also known as the Aurora Borealis. Most of the space weather is documented on National Oceanic and Atmospheric Administration and National Aeronautics and Space Administration so finding already created datasets were difficult. This domain will focus on the probability of seeing the northern lights and aid in travel promotions.

# Dataset As Expected

For the most part the dataset is as I would have expected. The major surprise for me was that I will need to do a lot of gathering of data from data.gov and other federal agencies. Additionally, I may need to conduct web scraping with Beautiful Soup to get accurate forecasts for the kp index from archived data. The solar flare and wind information I am not currently familiar with and will need to research what thresholds are likely to create the northern lights. I will have to create a legend to go by and use as informational purposes for the audience in the near future.

# Adjust Approach

My approach has had to change, mostly due to where the archive data is being stored. The amount of data is massive and am creating an approach to sample the data within the last ten years. That way it will be easier to conduct a train, test and split during the predictive analysis phase. When writing the data frame I did select a random number of rows, this may need to be adjusted since the sun does have a season, much like earth. This will need to be taken into consideration to provide accurate predictions. Due to the large amount of data, requesting random sampling of data and removing unnecessary data it may lock up my computer like it did in the previous class project. I may need to find different ways to beef up my CPU RAM/memory.

# Method

My method is working so far. If I need to adjust my methods, I will conduct that as needed. I am planning on following the CRISP-DM and what I have learned over the past year for each process. I have used online resources and will cite the sources throughout my code.

# Challenges

As noted, my major challenges are where the data is located, the events that need to occur to create a kp index, the size of data and the time processing the data. The time that is taking to gather and clean the data has taken up most of my time. Not to mention that I do have COVID19 and it causes confusion, lethargic and making it difficult to stay awake. That being said, I have not looked at relationship of variables, correlation, or regression as of yet and have not focused on a type of data preparation and visualizations. I believe my audience will be anyone that is interested in viewing the northern lights or to plan vacations around them.