***Project 2 Walk through***

For this project I found a competition by IBM. <https://community.ibm.com/community/user/datascience/blogs/susan-malaika/2020/11/10/call-for-code-spot-challenge-for-wildfires>

The point of this dataset/competition is to predict wildfires for Australia for 2021. To achieve that goal 5 datasets were given with different variables such as weather forecasts, fires, land vegetation, sea/water distance to fire area. In this first walk-through of my project, I will detail some of my code and explain my thoughts (please follow along with my code).

I first imported all need packages for data cleaning and then the machine learning ones.

I used pandas to import data sets, and then converted the date column to be able to use data as data frames. The original date columns did not have the correct python type to work with.

I first started the data cleaning by first finding empty data values and then removing the corresponding rows. I removed the rows because all the missing data came from std/Variance confidence. With out these confidence measures I didn’t really want to trust the data rows (line 10). So, after removing the empty rows I then decided to look for outliers, since those can also impact our data in unwanted ways. I did this using z score. The z score measures how ‘far away from the mean’ the value is, so I decided that any z score over 3 would be too high. After removing empty values and outliers I now have clean data comfortable enough to use (for the wildfires dataset). I then repeated these steps for weather data set, and plan to use same steps for the forecast weather data set. Both the land data set and vegetation data are descriptive, so no need for cleaning. The real problem will be on how to find out on what variables to concentrate on.

On the other hand, my clean data without the dates can be see as non-relational, but if we have the dates on the data then it becomes relational. With the addition of dates we can get keys to match the variables using dates as keys, Furthermore, we can potentially use the different regions as well as keys between the datasets.

The last step I want to accomplish is to get the corresponding dates for the clean data and see what dates were the most valuable.

During this first step of my journey, I have accomplished cleaning 2 of the 3 datasets. Next step is to then focus on the visualization of the data.