**DAAN 888 – Design and Implementation of Analytics System**

**Group: 7**

**Topic: Predicting climate change using El Nino dataset**

**Jeff Ericson, Anna Ivashko, Eric Lin, Tayeb Ahmadi**

**Data Set Exploration:**

We started with the El Nino Data Set from the UCI Machine Learning Repository (ref 1). This dataset was created from the NOAA Tropical Atmosphere Ocean array using moored buoys to collect the following variables of interest: Zonal Winds, Meridional Winds, Humidity, Air Temp, and Sea Surface Temp. The Data Set also contains the Year, Month, Day, Date, Latitude, and Longitude, for each observation. There were over 70 buoys collecting data from March 1980 to June 1998. Not all of the buoys were operating at the same time. The Gantt chart below (Figure 1) shows the dates of operation for each buoy.

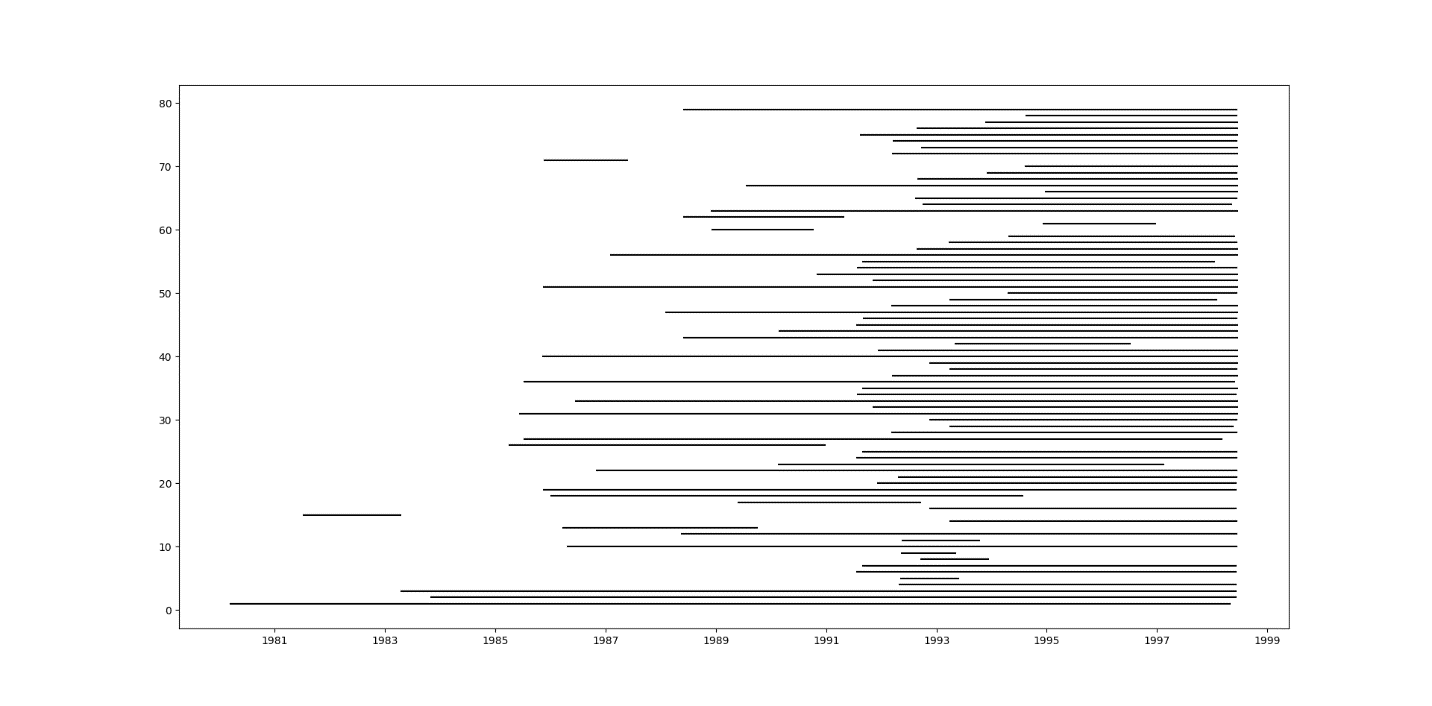


Figure 1: Gantt Chart of Buoy Operational Dates

The significance of this is that there are far fewer observations in the first 9 years than in the last 9 years of the 18-year dataset. In fact, the first 9 years contain only 10.6% of the 178,080 observations.

Figure 2 shows the value counts of observations by year.

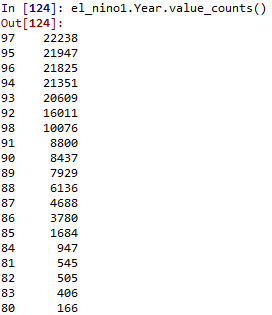


Figure 2: Value Counts by Year

Another issue we encountered with this dataset is that not all buoys were collecting data on all 5 variables of interest. This led to a considerable amount of missing values. The table below (Figure 3) shows the number and percent of missing variables by year for the 5 variables of interest. Of note is the number of missing values for the variable Humidity. This variable was not collected at all for the years 1980 through 1988, and only a small percentage 1989 and 1990.



Figure 3: Missing Values by Variable by Year

We also noticed that the El Nino Data Set did not contain a response variable. We decided to use the Oceanic Nino Index (ONI) from the NOAA Climate Prediction Center (ref 2). This index is based on a 3-month running average of sea surface temperature, calculated monthly from 1950 through 2019. The index contains 2 variables of interest, Total (the 3-month running average temperature) and Anom (the degree anomaly from normal temperature). The Anom variable is used to determine La Nino (less than -0.5 degree anomaly, or El Nino (greater than +0.5 degree anomaly). We decided to use Anom as our response variable.

**Data Preprocessing:**

**El Nino Data Set:** The following steps were taken to preprocess the El Nino Data Set:

1. The variables Zonal Winds, Meridional Winds, Humidity, Air Temp, and Sea Surface Temp we converted from categorical to numeric.
2. The Date variable was converted from an object type to a date type variable.
3. The blank spaces (missing observations) were replaced by NaNs.
4. In excel, a column was added called Buoy to denote which Buoy (1 – 79) recorded each observation.

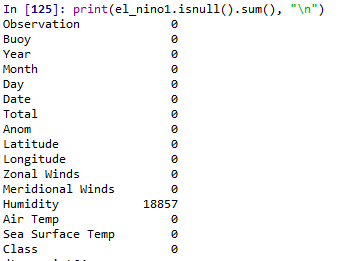
**Oceanic Nino Index Data Set:** These steps were taken to preprocess the ONI Data Set:

1. Rename the column SEAS (for Season) to Month.
2. Recode the SEAS data to reflect the numerical month (1-12).
3. Drop the years (by row) that did not overlap with the El Nino Data Set.
4. Replace the 4-digit values in the Year column with their 2-digit equivalent.
5. Convert the variable Month from categorical to numeric.

**Merging the 2 Data Sets and Preprocessing:**

1. Create a left join of the ONI Data Set with the El Nino Dataset (called oni\_1 then later changed to oni in our code) on the variables Year and Month.
2. Create a new variable called Class based on the Anom value where Anom <= -0.5 is coded as -1 (La Nina), Anom > 0.5 is coded as 1 (El Nino) and coded 0 (Normal) for Anom values in-between.
3. Convert the Class variable from integer to categorical.
4. Reorder the columns to put the Class variable last and rename the merged dataset el\_nino.
5. Subset the el\_nino dataset by year and fill each NaN with the yearly median value. This was done to use annual medians to fill in NaNs rather than using the overall column medians.
6. Recombine the yearly Data Frames into 1 Data Frame called el\_nino1.

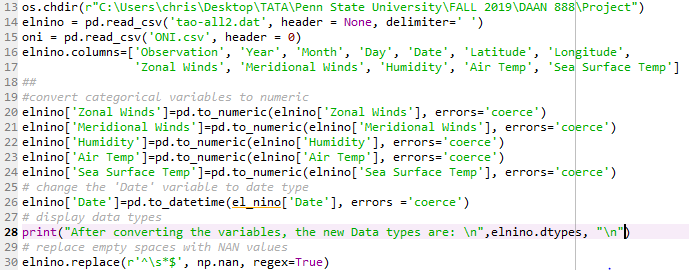
This still left us with a significant number of NaNs for the variable Humidity. This is because there were no observations for Humidity taken from 1980 – 1988, so there is no median value to replace NaNs for Humidity in those years. Figure 4 shows the amount of missing values for each variable in the merged file called el\_nino1.



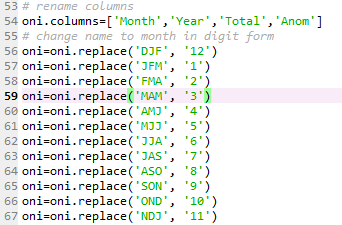
**Figure 4: Missing Values in Merged Data Set (el\_nino1)**

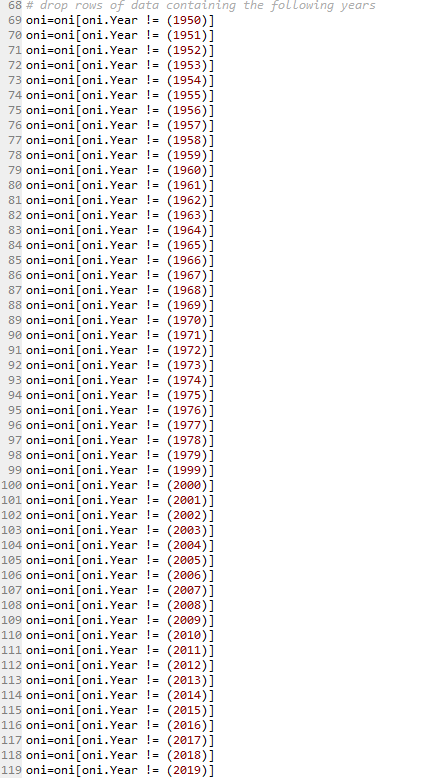
**Code:**

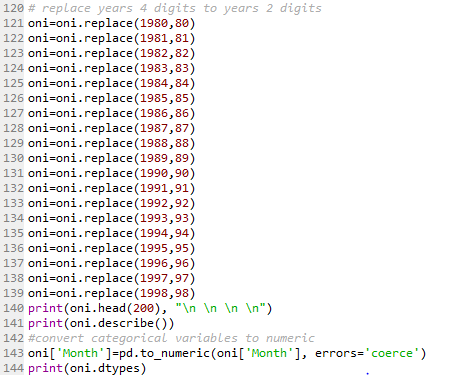
**Code for preprocessing El\_Nino Dataset:**



**Preprocessing the ONI Data Set:**





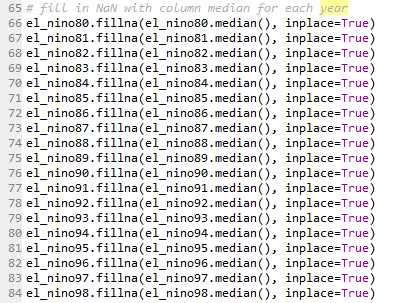


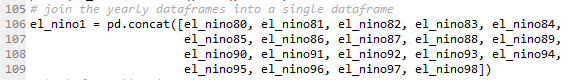
**Preprocessing the Merged Data Set:**











**References:**

1: <https://archive.ics.uci.edu/ml/datasets/El+Nino>

2: https://catalog.data.gov/dataset/climate-prediction-center-cpcoceanic-nino-index