**WEATHER DATA:**

**Matthew:**

1. Add weatherlat and weatherlon for both of the stations and inserted them right after the station column
2. Tavg(M) - replaced missing values with Tmin + (tmax-tmin / 2) Only 9 or so values replaced
3. Dropped 4 columns: DewPoint, Depth, Water1, SnowFall
4. Preciptotal (M and T) plans to put value for station 1 into station 2 or visa versa if 1 of the stations has a value and 0 if both stations have T
5. STNpressure (M) Only 4 missing values so put value from station 1 or 2 to replace the missing value
6. Sealevel (M) Only 9 missing values so put value from station 1 or 2 to replace the missing value
7. Avgspeed (M) Only 3 missing values so put value from station 1 or 2 to replace the missing value

**Amy**:

1. Depart (M) -
   1. 1472 missing values founded.
   2. Changed missing values from ‘M’ to ‘NaN’ .
   3. All 1472 missing values belong to station 2
   4. Replace missing value with the corresponded value in station 1
2. Wetbulb (M) - find the formula of a humidity percentage. Change the column name to Relative Humidity
   1. Relative Humidity Calculation = <https://www.1728.org/relhum.htm>
   2. Since there are only 4 missing variables (station 1 : 3missing variables, station2: 1missing variables), we are going to find replace missing values with the mean of each station)
   3. Change the column name to Relative Humidity
3. Heat (M) - index variable that measures how hot it is as compared to 65 degrees (keeping as it is). Replace the missing value with value for station 1

**Sonal**:

* Cool (M) - index variable that measures how cool it is as compared to 65 degrees (keeping as it is) Replace the missing value with value for station 1
* Sunrise (-) -
  + 1. 1472 rows affected
    2. ‘-’ only for station 2
    3. need to convert it into time (24 hour format) and replace the ‘-’ for station 2 with the time for station 1
    4. New column for Sunrise for Station 1 : sunrise\_new1\_Station1
    5. New column for Sunrise for Station 2 : sunrise\_new1\_Station2
* Sunset (-) -
  + 1. 1472 rows affected
    2. ‘-’ only for station 2
    3. need to convert it into time (24 hour format) and replace the ‘-’ for station 2 with the time for station 1
    4. New column for Sunset for Station 1: sunset\_new1\_Station1
    5. New column for Sunset for Station 2: sunset\_new1\_Station2
* Codesum ( ‘ ‘) -
  + 1. replacing blanks with ‘No event occured’.

**SPRAY DATA**

**Sonal:**

1. Dropped duplicate rows (~540)
2. Dropped the Time column
3. Create an identifier column for Spray = 1
4. Checked for the duplicates at Latitude and Longitude level
5. Exported the sheet

**MASTER DATABASE CREATION (from Mano’s code)**

1. Info:
   1. The purpose is to join train + spray + weather data altogether
   2. Spray had only 10 distinct spray date
2. First, joining the train\_cleaned.csv(14294 rows) and spray\_cleaned.csv(4616 rows) using **many to many** (no foreign key used to join). This resulted in 65,981,104 rows of output.
3. Then from the **many to many** join output, calculate the distance (using haversine) and date diff (spray date – train date) for each row. Resulted in 2 additional columns **dist** & **days\_frm\_spray**.
4. Rank the row ascendingly for **dist** & **days\_frm\_spray**.
5. Then pivot these 2 columns to treat them as separate independent variables.