25 Years of Tornado Data **March 20, 2021**

ETL Solvers: Paige Singleton, Sherali Yadav, Tara Polli, Varnika Rachupally, Daniel Ramirez

Tornado data seems readily available dating back to 1950's. After reviewing the data, we learned that data was collected techniques improved so much around 1990 that is appeared as if the prevalence of tornados doubled. Therefore, our group decided to collect data for the twenty-five year span of consistently-collected data from 1990-2015. The primary collector of weather data is National Oceanic and Atmospheric Administration (NOAA).

Extract:

Data sources (CSV):

- Storm Prediction Center (Kaggle) https://www.kaggle.com/jtennis/spctornado
  - Tornadoes_SPC_1950to2015.csv
- Climate at a Glance (NOAA) https://www.ncdc.noaa.gov/cag/national/time-series/110/tmax/all/1/1990-2015
  - 110-tmax-all-1-1990-2015 (1).csv

We read two CSV files into a jupyter notebook. The first dataset was obtained from Kaggle's "Storm Prediction Center" which allowed us to download to CSV file. It was created by the National Weather Service intended to enhance our understanding of where tornados happen, indicators of damage, and weather conditions associated with tornados.

Transform:

In jupyter notebook, we imported Pandas and created tornado_1.ipynb. Using pd_read method, we imported the first dataset Tornadoes_SPC_1950to2015.csv. We then performed the following actions:

- determined which columns to keep using and created a dataframe.
- renamed the columns so the headers are intuitive. (refer to Tornadoes_SPC_Col_Name_Def.pdf for column descriptions).
- This dataset includes data beginning in 1950, so we used the .loc feature to select data between years 1990-2015.
- reset the index. The output from running this notebook was a 'tornadoes_1.csv' file, and is stored in the "Transform" folder on Github.
- There was an unwanted column internal indexing, so we used "Reset_index" method to reset index, making the database (load) cleaner.
- We created and exported tornadoes_1.csv using the "to_csv" method.

Using read_pd method, we imported another dataset for temperature 110-tmax-all-1-1990-2015 (1).csv. We then performed the following actions to transform the data:

- Determined which columns to keep using and created a dataframe.
- Renamed the columns so the headers are intuitive
- Reset the index. The output from running this notebook was a 'temps.csv' file, and is stored in the "Transform" folder on Github.
- We created and exported temps.csv using the "to_csv" method.

Load

We created a SQL database in PgAdmin named 'etl_solvers. We created the tables, *tornadoes* and *temperature*, with columns matching their respective csv files and using index values as primary keys. We then imported the csv files (tornadoes_1.csv and temps.csv).