TEAM 21 - VICKY CHERNOKHLEBOVA AND NELSON LEUNG

# ETL on NFL stats and Weather Conditions

Data Sources: Kaggle.com

July 20, 2019

# PROJECT SUMMARY

## Objective

Using two sources of data found on Kaggle, NFL play by play game stats and historical weather data for select cities, we will be building a usable database to be able to correlate offensive NFL stats with different weather patterns.

## Goals

Extract raw data from csv files sourced from Kaggle.
Transform data into usable formats
Load into database to enable analyses

Examples of potential analysis:  Breaking down and aggregating the number of passes or rushes based on different bins in weather patterns.

- Are there more rushing plays when it is snowing hard vs light?
- Do those plays end up in more total yardage?
- Are there more passing plays when it is sunny out?

## Solution

Through an ETL process on data sourced from Kaggle, we will be creating aggregate tables that will enable analyses on our goal questions accomplished by joining game stats with the weather conditions in the home stadium city.  This will be loaded into a relational database

## Project Outline

- Sourced raw data from CSV's found on kaggle.com:
    1. https://www.kaggle.com/maxhorowitz/nflplaybyplay2009to2016
    2. https://www.kaggle.com/selfishgene/historical-hourly-weather-data
- ETL Process:
    A. Extracting initial data:
    - When trying to load NFL play by play csv directly to a Postgres database via Tableplus, errors were being generated that prevented the upload.  Attempt to use PGAdmin to upload also failed to create the correct tables
    - Hypothesis is that the number of columns with in the CSV (255) caused the database to incorrectly try to populate the tables

    - Solution - after creating the table schema first based on the columns headers, we read the data directly into a dataframe with pandas read_csv function and then called the to_sql function to load the raw data into the database

  B. Transforming Data:

    - NFL data: Filtering, merging and aggregating data based on GameID and play types
        - We selected the key metrics needed to perform offensive stats analyses and created an aggregate table based on grouping stats by game_id and play types
    - Weather data: filtering, sorting, renaming, transforming columns into rows
        - Challenge - team did not know how to transpose data from columns into rows. Through research, determined that a melt function applied to the dataframe would accomplish this.

  C. Loading Data: Using pandas to_sql function, we uploaded the data into our Postgres database

- Created relational database with tables for raw NFL data, aggregated offensive stats, and weather conditions for select cities. We chose this structure because of the ability to relate games with cities and subsequently the weather conditions in each of those cities during game day.

- The final tables in our production database listed below
    - NFL2009-2018: raw NFL data
    - offense_stats_by_game: aggregated table with stats per game by play type
    - game_conditions: lookup table matching home team to city conditions during the game date
    - city_attributes: key for city locations
    - weather_conditions: (ie. Light snow, cloudy, rain, etc)
    - Temperature: temperatures by city

Limitations in project:

- Due to the need to pay for historical weather data coupled with time constraints, our weather data set was only limited to the cities that were available.
- Additional translations and date cleanup not completed due to resource and time constraints and incomplete datasets.