ETL MLB Weather Project - Final Report

Team: Jie Feng, Wei Wu, Nick Ayvazian, Chidozie Nwagbo, Riley Williamson

Datasets: MLB Game Data from Kaggle

Atlanta Weather Data pulled from OpenWeather API

Mock Research Question: Does weather affect the total number of points scored in a baseball game?

Process:

1. **E**xtract: We extracted the MLB Game Data from Kaggle. It lists teams, date, scores, and type of game for every MLB game from 2010 to 2020. We extracted the data as a CSV file and were able to read it into a dataframe. The weather data came from OpenWeather in a JSON format. It has records of Atlanta weather data for every hour going back the last forty years.
2. **T**ransform: We cleaned the MLB data first. We filtered the home team column data to only show games in Atlanta and to remove any data from Preseason games. We also removed all quotes from the data for a clearer presentation.Next, we added a time to every game: 7pm for weekday games and 1pm for weekend games.We created a game\_id column. We combined the separate day, month, year, and time data to one column. We renamed the column names for a clarification. We converted game times to timestamps and added a timestamp column. Next we worked on transforming the weather data. From this JSON, we pulled data from the game timestamps and created a dataframe of the following columns: “Timestamp”, “Temperature”, “Humidity”, “Cloudiness”, “Weather”, and “Wind Speed.” The weather data did not require anymore cleaning; therefore, the final step in transforming was to combine the two dataframes based on the shared timestamps.
3. **L**oad: After that we used sqlalchemy to create an engine. We then loaded the data to a SQL final production database.One combined table went into the database with the following columns: Away\_Team, Away\_Team\_Score, Home\_Team, Home\_Team\_Score, Game\_ID, Time, Timestamp, Temperature, Humidity, Cloudiness, Weather, Wind\_Speed. We chose to use a regional database because it was more organized and easier to edit.

Challenges: The biggest challenge for this project was deciding how to combine the data based on time when both used different measurements of time. We decided to combine the separate date, month, and year columns. We also decided to assign a specific time to every game. This way the data will be more accurate than the 12am data that would have been generated. Finally, we converted the data into a timestamp that could be used to retrieve OpenWeather data.

Final DataFrame:

