Synopsis

For our project, we are interested in aggregating data for NFL and NBA teams across states to analyze trends by a specific state.

Extract

Our data sources include:

1. CSV: NFL Kaggle competition with team records since 1966
2. JSON: NFL teams with corresponding city and state locations
3. Web Scraping: convert HTML to table using pandas read_html

Transform

We cleaned all data sources prior to uploading into our database. In summary:

- NFL data: we calculated all win percentages per team and year between 2000-2018. We also controlled for teams who were created in this time frame and changed names/locations.
- NBA data: all team statistics are reported per conference. The data was cleaned so that we can look at the entire league at once.

Load

We combined all cleaned data into a PostgreSQL database. All table schemata were consistent prior to uploading to the database.

NFL

Data Acquisition - Extract

We acquired the NFL data from a Kaggle competition [1]. There are three CSV files describing the win/loss/tie records for each team since 1966. Sample

	schedule_date	schedule_season	schedule_week	schedule_playoff	team_home	score_home	score_away	team_away	team_favorite_id	spread_favorite
0	09/02/1966	1966	1	False	Miami Dolphins	14.0	23.0	Oakland Raiders	NaN	NaN
1	09/03/1966	1966	1	False	Houston Oilers	45.0	7.0	Denver Broncos	NaN	NaN
2	09/04/1966	1966	1	False	San Diego Chargers	27.0	7.0	Buffalo Bills	NaN	NaN
3	09/09/1966	1966	2	False	Miami Dolphins	14.0	19.0	New York Jets	NaN	NaN
4	09/10/1966	1966	1	False	Green Bay Packers	24.0	3.0	Baltimore Colts	NaN	NaN

Figure #: Sample raw data from the NFL Kaggle Competition. Two main features that we focused on for cleaning are retaining accurate team names (i.e. some team names changed over time) and calculating win percentage per season (i.e. this format only shows individual game scores)

Data Cleaning - Transform

Each game is reported as a row, with information about the home team, away team, season, stadium played, and overall score (Figure #). For consistent formatting, we would like the win percentage per year for each team. We would also like to append this dataset with the team city and state information. Overall, the transformation for the NFL data requires the following key steps:

1. Calculate win percentage per year per team
2. Correct for team name/location changes
3. Combine team city/state location

1. Calculate Win Percentage Per Team

We created a dictionary containing keys for each year since 2000 and counter variables like win, loss, and tie.

We then looped through the entire data frame and compared the home team score versus the away team score. We then incremented the win/loss/tie counter for each team in a game to keep count of their overall record throughout the season. Finally, we calculated the win percentage for a season, and added it to the corresponding key.

2. Correct Team Name/Location

Two sets of teams changed names/location between 2000-2019: the San Diego Chargers became the Los Angeles Chargers and the St. Louis Rams became the Los Angeles Rams. We controlled for this change to keep accurate team statistics over the whole time period by renaming the respective teams to a standard key (e.g. “LA/SD Chargers”) and performed a group by.

Finally, any NaNs within our dataset represent teams who were not founded by that point in time (e.g. Houston Texans became a team in 2002) were filled with zeros.

3. Combine Team Location

The state locations were downloaded as a JSON file [2]. Before we could merge this data,

NBA

Data Acquisition – Extract

We acquired all data from sources [3,4] via web scraping with the pandas read_html functions.

Data Cleaning - Transform

One of the challenges with the website's organization was that all data was organized by division per year. We merged all divisions together to get league data. Next we looped through all years to get the full dataset from 2000-2018. Finally, the website marks specific teams with an asterisk if they made it to the playoffs. This formatting was stripped so that we could perform aggregations.

Similar to the NFL data, several teams also changed their name/location throughout the years. This was processed similarly to maintain consistency over the team's history.

References

1. https://www.kaggle.com/tobycrabtree/nfl-scores-and-betting-data#nfl_stadiums.csv
2. Hold for moustafa's json
3. https://www.basketball-reference.com/friv/standings.fcgi?month=2&day=13&year=1992&lg_id=NBA
4. https://geojango.com/pages/list-of-nba-teams