Data Extract

1)We extract the data from three different data source:

afltables.com;

austadiums.com;

Kaggle.com

2)Download it as csv file and excel file :

game csv,

status csv,

AFL Stadiums csv (venue sheet, stadium sheet)

3)Extract and convert to four data frame

AFL game results,

AFL players’ performance,

AFL Venue In-use information,

AFL stadium location and capacity information.

4)Draw the ERD diagram after the evaluating the data columns and its correlation

5)We Created 7 tables connecting by the primary keys to each table……

Data Transform

**Manipulate data frame using pandas:**

We transformed the four dataframe to 7 tables and connecting each table with different id :

1)Player table: contains player\_id, first name , last name:

* Create a filtered data frame for specific columns: first and last name
* Split the first and last name from full name column using split function, and give values to the new column
* Rename the column headers to match the ERD table
* Drop the duplicates and setting the index
* Call the data frame to check the header
* Using the same procedure to create the other 6 tables:

2)Team table: contains team\_id, name(team name), stadium\_id

* Besides the same procedure as city table with pandas, the hard part is to create the city\_id
* Using the range and length to look for the length of the team\_transformed data frame’s row, and set the auto-increment id as the value of the team\_id; the length is counting + 1, and the range starts from 1;

3)City table: contains city\_id, name (city name),state

4)Stadium table: contains: stadium\_id.name(stadium name),city\_id, start year, end year, capacity ,active-ind

* Stadium table is hard part as it combines two different source of data, the stadium sheet and venue
* Each source of data gives different name spelling of the same stadium, we replace the other one first using replace function
* Then set the data frame using set function and get unique names with unique .union to join venue name, stadium name , and venue name together and rename the column
* Split the year information from in-use column and merge the stadium\_name , venue together
* Outer merge (inclusively merge) the stadium table with the venue, stadium data frame, and using if function to set boolean value: if the stadium is active or not , as if the end year is over 2022
* We would like to keep the most of the stadium data available first even venue\_df and stadium\_df miss match with each other where some of the city\_id is not given for stadium in the other , and venue in\_use information is not given in this dafa frame.
* And in the end we can dropNa for unwanted information.
* The rest of table is applying the same data cleaning, filtering, aggregating process

5)Player Status table: contains game\_id, player\_id, team\_id, rebond, inside\_50s, clearance, contested possessions, performance

6)Game table: contains : game\_id, year, round, date, start time, stadium id, attendance, home team id, home team score, away team id, away team score, rain fall.

7)TVS table: team id, TVS, yr