Assignment Nine

BSDS Spring 2021

(due 4/2, 11:59pm PST)

Note: You must submit this assignment as BOTH a Jupyter Notebook file (.ipynb) and an (.html) file on Canvas. All of the following must be satisfied.

- The filenames must be of the form

  `[last_name]_HW9.ipynb` and `[last_name]_HW9.html`

- You must include your name and the assignment number in a Markdown cell at the beginning of the notebook
- You must separate questions using Markdown cells
- If a question requires a short answer rather than code, use a markdown cell.

For this assignment you will need to download and import the following files from the NBA dataset on Canvas (originally found here https://www.kaggle.com/nathanlauga/nba-games):

- `games.csv`
- `teams.csv`
- `ranking.csv` (Hint: you might need to tweak the `guess_max` parameter here!)

1. In each of the three datasets, identify a primary key and decide whether it is a foreign key for another dataset.

2. Consider the `teams` and the `games` datasets.

(a) Use inner join to add home team information to the `games` dataset.

(b) Count the number of columns before and after the join. Explain the result.

(c) Could we use another join to achieve the same dataset? If so, which? If not, why?

3. Consider the `ranking` and the `games` datasets. Our goal will be to see how different game statistics affect home win percentage of the Golden State Warriors over time.

(a) From these two datasets, create a dataset of games where Golden State is the home team and a dataset of Golden State rankings over time. (You may need to use the `teams` dataset to accomplish this!)

(b) Create a column in your new rankings dataset representing Golden State's **home** win percentage on a given date. (Hint: You may need to use the `separate` command)

(c) Use an outer join to join the two datasets in such a way that you complete part (d) and that *minimizes* the amount of rows containing missing values. **Explain** why the other outer joins will have more missing values.

(d) Use a visualization to explore whether there are any correlations in your new dataset.

4. Consider the `flights` and `planes` datasets.

(a) Use the `semi_join` command to create a dataset of flights with just the top 5 carriers who have the most flights.

(b) Use the `anti_join` command to create a dataset of flights whose tail-number does not have a match in the `planes` dataset.

5. Consider the `flights` and `airports` datasets.

(a) Add the origin and destination latitude and longitude for each flight to the `flights` dataset.

(b) BONUS: Pick a particular tail number and a particular date. Find a way to graph the flight path of the plane. Hint: Adding the `borders("state")` layer to a ggplot will add the continental united states to a lat/long plot.