Assignment Eight

BSDS Spring 2021

(due 3/26, 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]_HW8.ipynb` and `[last_name]_HW8.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 the following files from Canvas:

- `patient_info.csv`
- `pokemon2.csv`
- `us_presidents.csv`
- `mlb_data.csv`

General Hint: Remember you can select all columns between "Col_A" and "Col_B" (including these two columns!) with `Col_A:Col_b`

1. Which command should you use to import `patient_info.csv`? Import and print the dataset.

2. Print the first 10 lines of `pokemon2.csv`. Now import it.

3. Use pivot commands to perform the following.

(a) For the pre-loaded dataset `relig_income`, change the column headers relating to income (and "don't know") into a variable and move the value to a new column.

(b) Import the .csv file titled `us_presidents.csv`, change the values of "age_assumed_office" and "order" to column headers with their values coming from the "value" column.

4. Look at the pre-loaded dataset `billboard`.

(a) What is this dataset displaying? (don't forget about `?` !)

(b) Tidy the dataset

(c) Use `separate` to create a column displaying the week variable as type integer. (Hint: You'll need to separate into two columns, use the `convert` option, and use `select` to get rid of one of them).

(d) Pick a song (or songs!) and graph its ranking over time (starting from when it entered the top 100). Why did tidying the data help you do this?

5. Do the following for the `flights` dataset in the `nycflights13` library.

(a) Unite the year, month, and day (in that order!) columns using "`-`" as the separator. Convert this new column into a column of type `Date`.

(b) Separate the "dep_time" and "arr_time" columns into "dep_time_hour", "dep_time_minute", "arr_time_hour", and "arr_time_minute" columns.

6. Import the dataset in the file `mlb_data.csv`

(a) Graph the win *percentage* over time of all the Major League Baseball Teams in California (SFG, SDP, LAA, LAD, OAK).

(b) Note that some teams do not have number of wins for certain years. Use the `complete` command to make explicit these implicit missing values.