hw4

Leo

2024-09-26

1 - load the Lahman dataset into R

```
# load libraries
suppressMessages(library(dplyr))
library(ggplot2)
library(Lahman)
```

1a - use left join to filter the data set to display the playerID, yearID, teamID, stint, G (games played), HR (home runs), and salary for all players who hit more than 30 home runs in a single season and played for a team in New York (teamID “NYA” or “NYN”) between 2010 and 2020. How many players match these criteria?

```
# join the batting and salaries tables based on playerID
player_data <- Batting %>% left_join(Salaries, by = c("playerID", "yearID",
"teamID"))

# grab all players who hit more than 300 home runs in a season and played for
a new york team, between 2010 and 2020
player_data_filtered <- player_data %>% filter(HR > 30, yearID >= 2010,
yearID <= 2020, teamID %in% c("NYA", "NYN"))

# display info on them
player_data_filtered <- player_data_filtered %>% select(playerID, yearID,
teamID, stint, G, HR, salary)

# how many players match the condition?
numPlayers <- nrow(player_data_filtered)
numPlayers

## [1] 16
```

How many players match these criteria?

- 16

1b - What is the difference between the following two joins? What is the difference between semi_join and anti_join? Provide an example using the Salaries and Batting tables.

- anti_join(Salaries, Batting, by = c("playerID" = "playerID")) will return rows in Salaries where there is no corresponding row in Batting with the same playerID.
- anti_join(Batting, Salaries, by = c("playerID" = "playerID")) will do the same, but with Salaries and Batting switched.
- semi_join() returns the rows in the first listed dataset that have a corresponding row in the second listed dataset. It is the inverse of anti_join()

```
head(semi_join(Salaries, Batting, by = "playerID"))

##   yearID teamID lgID  playerID salary
## 1   1985    ATL   NL barkele01 870000
## 2   1985    ATL   NL bedrost01 550000
## 3   1985    ATL   NL benedbr01 545000
## 4   1985    ATL   NL  campri01 633333
## 5   1985    ATL   NL ceronri01 625000
## 6   1985    ATL   NL chambch01 800000

anti_join(Salaries, Batting, by = "playerID")

## [1] yearID   teamID   lgID     playerID salary
## <0 rows> (or 0-length row.names)
```

1c - Select the teamID, yearID, and the total number of runs batted in (RBI) for each team in the American League (AL) for the year 2015 (using one or more inner joins with the Teams and Batting tables). How many total home runs were hit by American League teams in 2015?

```
# join the batting table with teams
teams_data <- Batting %>% inner_join(Teams, by = c("teamID", "yearID",
"lgID"))

# filter for the teams in the AL in 2015
teams_data_filtered <- teams_data %>% filter(lgID == "AL", yearID == 2015)

# group by team, compute, and select the required fields
teams_data_filtered <- teams_data_filtered %>% group_by(teamID, yearID)
teams_data_filtered_RBI <- teams_data_filtered %>% summarise(totalRBI =
sum(RBI))

## `summarise()` has grouped output by 'teamID'. You can override using the
## `.groups` argument.
```

```
# how many total home runs were hit by American League teams in 2015?
teams_data_filtered_HR <- teams_data_filtered %>% summarise(totalHR =
sum(HR.x))

## `summarise()` has grouped output by 'teamID'. You can override using the
## `.groups` argument.

totalHRin2015 <- sum(teams_data_filtered_HR$totalHR)
totalHRin2015

## [1] 2634
```

How many total home runs were hit by American League teams in 2015?

- 2634

1d - Using the Managers and Teams tables, determine the number of seasons each manager managed a team. Use group_by and count to get the number of unique managerID and teamID combinations. How many unique combinations of managerID and teamID are present? Are there any players with unusually high number of years as a manager?

```
# use group_by and count to get a list of all manager/team combos, with the
count of how many seasons the combo lasted for
manager_team_data <- Managers %>% group_by(playerID, teamID) %>% count()

# how many unique manager/team combos are there?
uniqueCombos <- nrow(manager_team_data)
uniqueCombos

## [1] 1295

# get players with unusually high years as manager (need total amount of
years as manager across any teams, since some players managed multiple teams)
managersTotalSeasons <- Managers %>% group_by(playerID) %>% count()
UnusuallyHigh_TotalSeasons <- managersTotalSeasons %>% filter(n > 30)
UnusuallyHigh_TotalSeasons

## # A tibble: 3 × 2
## # Groups:   playerID [3]
##   playerID      n
##   <chr>     <int>
## 1 larusto01    36
## 2 mackco01     53
## 3 mcgrajo01    36
```

How many unique combinations of managerID and teamID are present?

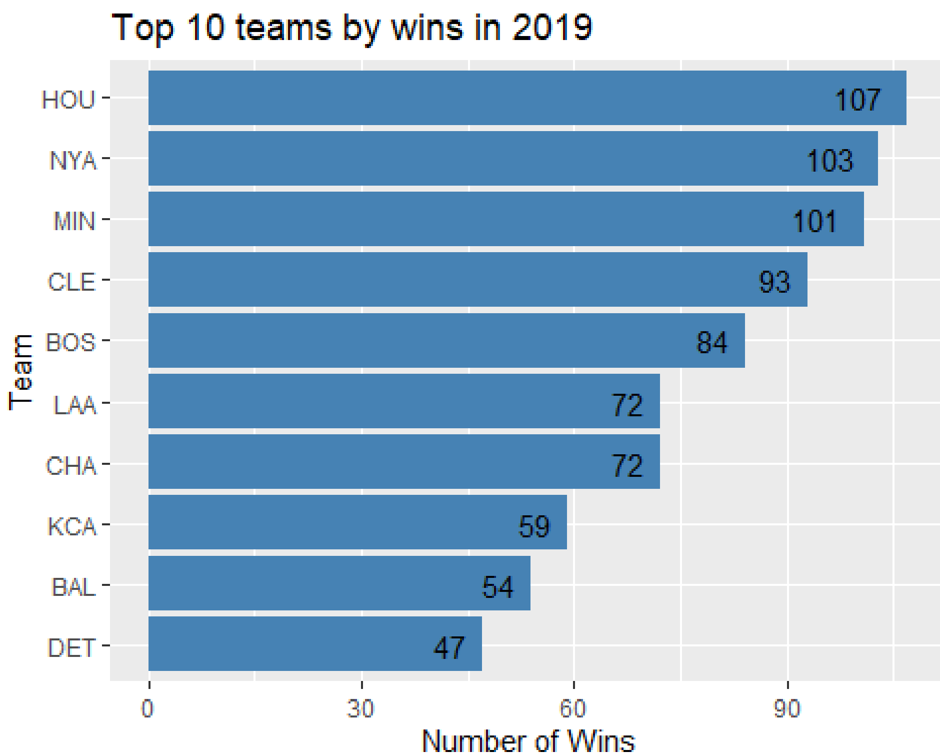
- 1295

Are there any players with unusually high number of years as a manager?

- Yes. I determined that anything longer than 30 years could be considered an unusually long time to serve as a manager. There are 3 players that served over 30 years as a manager.

1e - Using the provided template as a start, produce a horizontal bar plot that shows the number of wins for the top 10 teams in 2019. Adjust the axis labels to clearly represent the teams and the number of wins. Add a meaningful title to the plot, and include the number of wins as text on each bar for clarity.

```
teams_filtered <- Teams %>%
  filter(yearID == 2019) %>%
  select(teamID, W) %>%
  slice(1:10) # grab the top 10

ggplot(teams_filtered, aes(x = reorder(teamID, W), y = W)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    coord_flip() +
    labs(title = "Top 10 teams by wins in 2019", x = "Team", y = "Number of
Wins") +
    geom_text(aes(label = W), hjust = 1.5) # include number for each bar
```

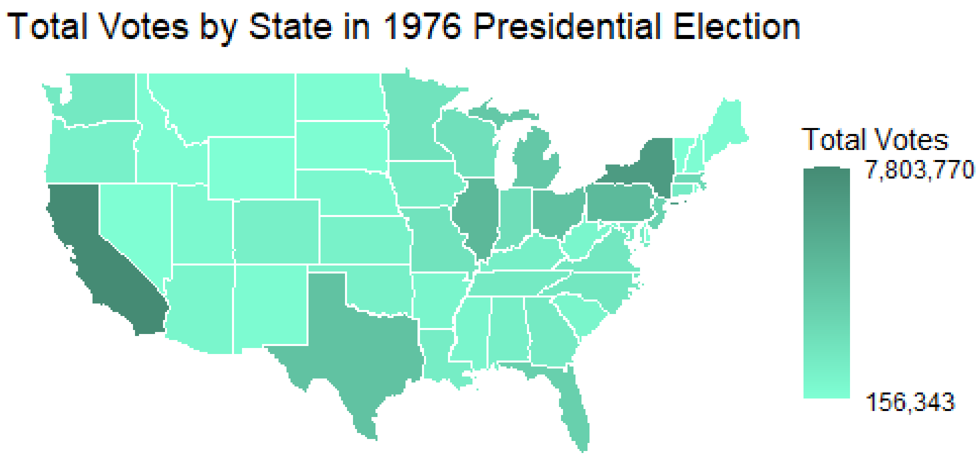


2 - create a visualization of the US map showing the states/territories and the number of presidential votes received during an election year.

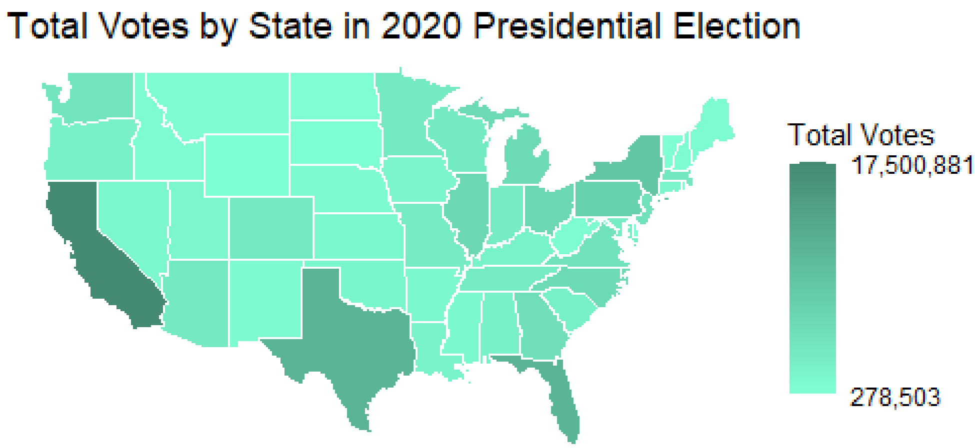
For this question, you will create two visualizations of the US map for two presidential years of your choice coloring the states or sizing the point/marker for the states according to the number of total votes received from that state for the presidential election.

```
# load libraries
library(sf)

## Linking to GEOS 3.12.1, GDAL 3.8.4, PROJ 9.3.1; sf_use_s2() is TRUE

library(maps)
library(scales)

# load dataset
presidents_data <- read.csv("us-presidents.csv")

# filter the data for 2 election years
presidents_data_1976 <- presidents_data %>% filter(year == 1976)
presidents_data_2020 <- presidents_data %>% filter(year == 2020)
```

```
# get map data of the US, using the maps package in R
us_map <- map_data("state") # returns the coords for plotting the boundaries
of the states
#us_map

# get the votes per state for a given year and join it to the states in the
map
presidents_data_1976_byState <- presidents_data_1976 %>% mutate(region =
tolower(state)) # add a region column with the state name in lower case, so
we can join it to the map table
map_1976 <- left_join(us_map, presidents_data_1976_byState, by = "region") #
join the tables
presidents_data_2020_byState <- presidents_data_2020 %>% mutate(region =
tolower(state))
map_2020 <- left_join(us_map, presidents_data_2020_byState, by = "region") #
join the tables

# create the maps
ggplot(map_1976, aes(x = long, y = lat, group = group, fill = totalvotes)) +
# initialize the plot object with map information
  geom_polygon(color = "white") + # draw the shapes of the states, with white
outlines
  coord_fixed(1.3) + # this makes the proportions of the map more standard
  scale_fill_gradient(low = "aquamarine", high = "aquamarine4",
                      breaks = c(min(map_1976$totalvotes),
max(map_1976$totalvotes)),
                      labels = comma) + # create a gradient for smaller to
larger numbers
  labs(title = "Total Votes by State in 1976 Presidential Election", fill =
"Total Votes") +
  theme_void()
```




```
ggplot(map_2020, aes(x = long, y = lat, group = group, fill = totalvotes)) +
# initialize the plot object with map information
  geom_polygon(color = "white") + # draw the shapes of the states, with white
outlines
  coord_fixed(1.3) + # this makes the proportions of the map more standard
  scale_fill_gradient(low = "aquamarine", high = "aquamarine4",
                      breaks = c(min(map_2020$totalvotes),
max(map_2020$totalvotes)),
                      labels = comma) + # create a gradient for smaller to
larger numbers
  labs(title = "Total Votes by State in 2020 Presidential Election", fill =
"Total Votes") +
  theme_void()
```



Compare both maps and comment on any observations.

I Chose to compare the 2 available years that were farthest apart in time so that I could observe a larger change between the 2. I have several observations, the first being that the overall scale for 1976 is much lower than 2020, since the population was significantly less back then. Additionally, in 1978 there was a high concentration of voters in states directly south of the great lakes but by 2020 those states seem to be much more in line with the average. Meanwhile, Florida was only a little bit above average in 1976 and in 2020 it appears to be in the top 3 states for total votes. Finally, California was by far the largest total vote state in both years, meaning that even back then it must have been very densely populated when compared to the rest of the states.

3 - Create a word cloud for an interesting (relatively short, say a couple of pages) document of your own choice.

```
# load libraries
library(wordcloud)
library(tm)

# load in the text for the word cloud
text <- readLines("meta-earnings-transcript.txt")

# Create a corpus. this is basically a collection of text
```

```r
corpus <- Corpus(VectorSource(text))

# Clean the text
corpus <- tm_map(corpus, content_transformer(function(x) gsub("’", "", x))) #
remove apostrophes
corpus <- tm_map(corpus, removePunctuation)  # Remove rest of punctuation
corpus <- tm_map(corpus, removeNumbers)  # Remove numbers
corpus <- tm_map(corpus, removeWords, stopwords("english"))  # Remove common
stopwords (words like "the", "and", etc)

# Create the word cloud
wordcloud(corpus, random.order = FALSE, max.words = 200, colors =
brewer.pal(5, "Dark2"))
```

Caption:

- “Transcript for Meta’s Q2 2024 earnings call”