Applied Statistical Programming - Relational Databases

3/7/2022

Write R code to answer the following questions. Write the code, and then show what the computer returns when that code is run. Thoroughly comment your solutions.

You have until the beginning of class 3/9 at 10:00am to complete the assignment below. You may use R, but not any online R documentation. Submit the Rmarkdown and the knitted PDF to Canvas. Have one group member submit the activity with all group members listed at the top.

Relational Databases

Data rarely come in nicely combined CSV files. This exercise gives you practice combining data sources. You are given three sets of Twitter data that need to be combined to answer a set of questions below. The data can be found at the following URLs.

- https://github.com/jmontgomery/jmontgomery.github.io/blob/master/PDS/Datasets/Tweets.csv.zip
- https://raw.githubusercontent.com/jmontgomery/jmontgomery.github.io/master/PDS/Datasets/Mayors.csv
- https://raw.githubusercontent.com/jmontgomery/jmontgomery.github.io/master/PDS/Datasets/TwitterMentions.csv

Once you have imported the data, use relational database commands join data as necessary in order to answer the following questions.

1. For each mayor, calculate the number of times they were mentioned
2. Add to the mentions datset the number of times each mayor tweeted.
3. Create a combined dataset of all tweets from the tweets and mentions data. Subset down to overlapping columns (and rename where needed) to make this easy.
4. Are there any tweets in the mentions dataset from mayors?

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Load datasets
mayors <- read_csv(file = "https://raw.githubusercontent.com/jmontgomery/jmontgomery.github.io/master/P
```

```
## New names:
## * ‘‘ -> ...1
```

```
## Rows: 1473 Columns: 51
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (15): FullName, LastName, FirstName, LastElectionDate, Title, CityName, ...
## dbl (36): ...1, MayorID, GenderMale, GenderFemale, RaceWhite, RaceBlack, Rac...
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
mentions <- read_csv(file = "https://raw.githubusercontent.com/jmontgomery/jmontgomery.github.io/master/
```

```
## New names:
## * ‘‘ -> ...1
## Rows: 61570 Columns: 18-- Column specification ------------------------------------------------
## Delimiter: ","
## chr   (5): ScreenName, Text, MayorHandle, ReplyToSN, StatusSource
## dbl  (12): ...1, TweetID, Favorited, FavoritesCount, IsRetweet, RetweetCount...
## dttm  (1): CreatedTime
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
tweets <- read_csv("Tweets.csv")
```

```
## New names:
## * ‘‘ -> ...1
## Rows: 604818 Columns: 17-- Column specification -----------------------------------------------
## Delimiter: ","
## chr   (4): ScreenName, Text, ReplyToSN, StatusSource
## dbl  (12): ...1, TweetID, Favorited, FavoritesCount, IsRetweet, RetweetCount...
## dttm  (1): CreatedTime
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# In-class drill 1 & 2 ---------------------------------------------------------
tweets <- rename(tweets, TwitterHandle = ScreenName)
left <- tweets %>%
  left_join(select(mayors, TwitterHandle, FacebookLink, TwitterLink, GenderFemale),
    by = "TwitterHandle"
  )

right <- tweets %>%
```

```r
  right_join(select(mayors, TwitterHandle, FacebookLink, TwitterLink, GenderFemale),
    by = "TwitterHandle"
  )

inner <- tweets %>%
  inner_join(select(mayors, TwitterHandle, FacebookLink, TwitterLink, GenderFemale),
    by = "TwitterHandle"
  )

full <- tweets %>%
  full_join(select(mayors, TwitterHandle, FacebookLink, TwitterLink, GenderFemale),
    by = "TwitterHandle"
  )

# Drill 3 ----------------------------------------------------------------------
subsetMayors <- mayors %>%
  filter(TwitterHandle %in% c("robertgarcialb", "rodhiggins2017"))

subsetTweets <- tweets %>%
  filter(TwitterHandle %in% c("robertgarcialb", "rodhiggins2017"))

subsetInner <- subsetTweets %>%
  inner_join(subsetMayors,
    by = "TwitterHandle"
  )

# Item 1 -----------------------------------------------------------------------
## Create count for times mayors were mentioned
mentions_count <- mentions %>%
  group_by(MayorHandle) %>%
  mutate(TimesMentioned = n()) %>%
  rename(TwitterHandle = MayorHandle) %>%
  select(TwitterHandle, TimesMentioned) %>%
  unique()

mentions <- mentions %>% rename(TwitterHandle = MayorHandle)

mayors_mentions_count <- mayors %>%
  left_join(select(mentions_count, TwitterHandle, TimesMentioned),
    by = "TwitterHandle"
  )

# Item 2 -----------------------------------------------------------------------
mayors_tweets <- tweets %>%
  group_by(TwitterHandle) %>%
  mutate(times_tweeted = n()) %>%
  select(TwitterHandle, times_tweeted) %>%
  unique()

mentions <- mentions %>% left_join(mayors_tweets, by = "TwitterHandle")

# Item 3 -----------------------------------------------------------------------
## Columns to remove:
```

```
## - ScreenName
## - Column 1 from each one?

tweets_subset <- tweets %>%
  select(-1)

mentions_subset <- mentions %>%
  select(-c(1, 3, 17))

## DON'T RUN. It's currently not working.
# tweets_mentions <- tweets_subset %>%
#   left_join(mentions_subset, by = "TwitterHandle")

# Item 4-------------------------------------------------------------------------
## Is this an inner join?
inner_tweets <- tweets_subset %>%
  inner_join(mentions_subset)
```

```
## Joining, by = c("TweetID", "TwitterHandle", "Text", "CreatedTime", "Favorited",
## "FavoritesCount", "IsRetweet", "RetweetCount", "Retweeted", "ReplyToSN",
## "ReplyToSID", "ReplyToUID", "Truncated", "StatusSource", "Latitude")
```

```
nrow(inner_tweets)
```

```
## [1] 5824
```