Bofan Chen QBS181 Project Code

Bofan Chen

2022-11-16

Description of Code

Thie file contains the detailed steps of tidyverse data cleaning and word cloud ploting for the variables team_home, team_away, statdium and stadium_address.

Load the library

```
if (!require("RColorBrewer")) {
  install.packages("RColorBrewer")
  library(RColorBrewer)
}
library(RColorBrewer)
# install.packages("wordcloud")
library(wordcloud)
library(SnowballC)
#install.packages("tm")
library(tm)
```

Tidyverse Data Cleaning

```
spreads <- read.csv("/Users/william/Library/Mobile Documents/com~apple~CloudDocs/QBS 181/spreadspoke_sc
stadiums <- read.csv("/Users/william/Library/Mobile Documents/com~apple~CloudDocs/QBS 181/nfl_stadiums.
teams <- read.csv("/Users/william/Library/Mobile Documents/com~apple~CloudDocs/QBS 181/nfl_teams.csv")

summary(spreads)
```

```
##   schedule_date      schedule_season schedule_week      schedule_playoff
##   Length:13504       Min.   :1966    Length:13504       Mode :logical
##   Class :character   1st Qu.:1983    Class :character   FALSE:12957
##   Mode  :character   Median :1997    Mode  :character   TRUE :547
##                      Mean   :1996
##                      3rd Qu.:2010
##                      Max.   :2022
##
##    team_home          score_home      score_away      team_away
##   Length:13504       Min.   : 0.00   Min.   : 0.00   Length:13504
##   Class :character   1st Qu.:14.75   1st Qu.:13.00   Class :character
```

```
##   Mode  :character    Median :22.00   Median :20.00   Mode  :character
##                       Mean   :22.45   Mean   :19.76
##                       3rd Qu.:29.00   3rd Qu.:27.00
##                       Max.   :72.00   Max.   :62.00
##                       NA's   :224     NA's   :224
##   team_favorite_id   spread_favorite   over_under_line   stadium
##   Length:13504       Min.   :-26.500   Min.   :28.00     Length:13504
##   Class :character   1st Qu.: -7.000   1st Qu.:38.50     Class :character
##   Mode  :character   Median : -4.500   Median :42.00     Mode  :character
##                      Mean   : -5.391   Mean   :42.22
##                      3rd Qu.: -3.000   3rd Qu.:45.00
##                      Max.   :  0.000   Max.   :63.50
##                      NA's   :2703      NA's   :2775
##   stadium_neutral weather_temperature weather_wind_mph weather_humidity
##   Mode :logical   Min.   :-6.00       Min.   : 0.000   Min.   :  4.00
##   FALSE:13396     1st Qu.:48.00       1st Qu.: 3.000   1st Qu.: 57.00
##   TRUE :108       Median :62.00       Median : 8.000   Median : 69.00
##                   Mean   :58.87       Mean   : 7.689   Mean   : 67.22
##                   3rd Qu.:72.00       3rd Qu.:11.000   3rd Qu.: 79.00
##                   Max.   :97.00       Max.   :40.000   Max.   :100.00
##                   NA's   :1224        NA's   :1240     NA's   :5063
##   weather_detail        X              X.1
##   Length:13504     Mode:logical   Mode:logical
##   Class :character NA's:13504     NA's:13504
##   Mode  :character
##
##
##
##
```

summary(stadiums)

```
##   stadium_name       stadium_location    stadium_open   stadium_close
##   Length:118         Length:118          Min.   :1909   Min.   :1970
##   Class :character   Class :character    1st Qu.:1970   1st Qu.:1995
##   Mode  :character   Mode  :character    Median :1992   Median :2000
##                                          Mean   :1985   Mean   :2000
##                                          3rd Qu.:2002   3rd Qu.:2009
##                                          Max.   :2020   Max.   :2016
##                                          NA's   :29     NA's   :77
##   stadium_type       stadium_address     stadium_weather_station_code
##   Length:118         Length:118          Length:118
##   Class :character   Class :character    Class :character
##   Mode  :character   Mode  :character    Mode  :character
##
##
##
##
##   stadium_weather_type stadium_capacity   stadium_surface      STATION
##   Length:118           Length:118         Length:118         Length:118
##   Class :character     Class :character   Class :character   Class :character
##   Mode  :character     Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##        NAME              LATITUDE         LONGITUDE          ELEVATION
##   Length:118         Min.   :25.79    Min.   :-122.41   Min.   :   1.8
##   Class :character   1st Qu.:33.46    1st Qu.: -97.07   1st Qu.:  13.4
##   Mode  :character   Median :39.10    Median : -86.12   Median : 163.4
##                      Mean   :37.68    Mean   : -91.47   Mean   : 208.3
##                      3rd Qu.:40.68    3rd Qu.: -80.21   3rd Qu.: 227.1
##                      Max.   :47.65    Max.   : -71.00   Max.   :1611.2
##                      NA's   :57       NA's   :57        NA's   :57
```

```
summary(teams)
```

```
##    team_name         team_name_short       team_id          team_id_pfr
##  Length:44          Length:44          Length:44          Length:44
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##  team_conference    team_division      team_conference_pre2002
##  Length:44          Length:44          Length:44
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##  team_division_pre2002
##  Length:44
##  Class :character
##  Mode  :character
```

```r
library(tidyverse)
```

```r
# Adding Abbreviations
df1 <- teams[,c("team_name", "team_id")]
df1 <- df1 %>% rename(team_home=team_name)
spread1 <- left_join(spreads, df1, by="team_home")

# Difference Favored
spread1$difference_favored_minus_notfavored <- ifelse(spread1$team_id == spread1$team_favorite_id, sprea

# Comparing Spread
spread1$abs_spread <- abs(spread1$spread_favorite)
spread1$actual_diff_spread <- spread1$difference_favored_minus_notfavored - spread1$abs_spread

# Comparing Over Under
spread1$actual_total_over_under <- (spread1$score_home + spread1$score_away) - spread1$over_under_line

# Merge Stadiums
stadiums<- stadiums %>% rename(stadium=stadium_name)
merged_stadiums <- left_join(spread1, stadiums, by="stadium")

# Merge Teams
df2 <- teams[,c("team_name", "team_conference", "team_division", "team_conference_pre2002", "team_divis
df2 <- df2 %>% rename(team_home=team_name)
merged_teams <- left_join(merged_stadiums, df2, by="team_home")
```

Visualize team_home

Visualize team_stadium

```
stadium <- read.csv("/Users/william/Library/Mobile Documents/com~apple~CloudDocs/QBS 181/stadium.csv")
# remove Stadium, Filed, Bowl, County, Memorial, Municipal, Coliseum, Dome, Superdome, Metrodome, Silve
stadium <- Corpus(VectorSource(stadium$stadium))


#dev.new(width = 2000, height = 2000, unit = "px")
wordcloud(stadium
          , scale=c(3.5,0.2)     # Set min and max scale
          , max.words=100        # Set top n words
          , random.order=FALSE   # Words in decreasing freq
          , rot.per=0.15         # % of vertical words
          , use.r.layout=FALSE   # Use C++ collision detection
          , colors=brewer.pal(8, "Dark2"))
```

Visualize team_away

```r
team_away <- merged_teams$team_away
team_away <- Corpus(VectorSource(team_away))

#dev.new(width = 2000, height = 2000, unit = "px")
wordcloud(team_away
          , scale=c(5,0.5)       # Set min and max scale
          , max.words=100        # Set top n words
          , random.order=FALSE   # Words in decreasing freq
          , rot.per=0.35         # % of vertical words
          , use.r.layout=FALSE   # Use C++ collision detection
          , colors=brewer.pal(8, "Dark2"))
```

Visualize stadium_address

```r
stadium_address <-  merged_teams$stadium_address
stadium_address <- Corpus(VectorSource(stadium_address))
# Convert to lowercase
# stadium_address <- tm_map(stadium_address, tolower)
# Remove conjunctions etc.
stadium_address <- tm_map(stadium_address, removeWords, stopwords("english"))
# Remove suffixes to the common 'stem'
stadium_address <- tm_map(stadium_address, stemDocument)
# Remove commas etc.
stadium_address <- tm_map(stadium_address, removePunctuation)

#dev.new(width = 2000, height = 2000, unit = "px")
wordcloud(stadium_address
          , scale=c(5,0.5)      # Set min and max scale
          , max.words=100       # Set top n words
          , random.order=FALSE  # Words in decreasing freq
          , rot.per=0.35        # % of vertical words
          , use.r.layout=FALSE  # Use C++ collision detection
          , colors=brewer.pal(8, "Dark2"))
```