Lab 3: Data Wrangling on Soccer Tournament Data

July 11, 2021

Soccer tournament data wrangling

Read the dataset of football games.

```
d <- read_csv("data/results.csv")
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   date = col_date(format = ""),
##   home_team = col_character(),
##   away_team = col_character(),
##   home_score = col_double(),
##   away_score = col_double(),
##   tournament = col_character(),
##   city = col_character(),
##   country = col_character(),
##   neutral = col_logical()
## )
```

1. Select variables `date`, `home_team` and `away_team`.

```
d %>% select(date, home_team, away_team)
```

```
## # A tibble: 39,669 x 3
##    date       home_team away_team
##    <date>     <chr>     <chr>
##  1 1872-11-30 Scotland  England
##  2 1873-03-08 England   Scotland
##  3 1874-03-07 Scotland  England
##  4 1875-03-06 England   Scotland
##  5 1876-03-04 Scotland  England
##  6 1876-03-25 Scotland  Wales
##  7 1877-03-03 England   Scotland
##  8 1877-03-05 Wales     Scotland
##  9 1878-03-02 Scotland  England
## 10 1878-03-23 Scotland  Wales
## # ... with 39,659 more rows
```

2. Subset games with **Brazil** as the home team.

```
d %>% filter(home_team == "Brazil")
```

```
## # A tibble: 552 x 9
##    date       home_team away_team home_score away_score tournament city  country
##    <date>     <chr>     <chr>          <dbl>      <dbl> <chr>      <chr> <chr>
##  1 1916-07-08 Brazil    Chile              1          1 Copa Amé~  Buen~ Argent~
##  2 1916-07-12 Brazil    Uruguay            1          2 Copa Amé~  Buen~ Argent~
```

```
##  3 1917-10-12 Brazil    Chile              5          0 Copa Amé~  Mont~ Uruguay
##  4 1919-05-11 Brazil    Chile              6          0 Copa Amé~  Rio ~ Brazil
##  5 1919-05-18 Brazil    Argentina          3          1 Copa Amé~  Rio ~ Brazil
##  6 1919-05-26 Brazil    Uruguay            2          2 Copa Amé~  Rio ~ Brazil
##  7 1919-05-29 Brazil    Uruguay            1          0 Copa Amé~  Rio ~ Brazil
##  8 1919-06-01 Brazil    Argentina          3          3 Friendly   Rio ~ Brazil
##  9 1920-09-18 Brazil    Uruguay            0          6 Copa Amé~  Viña~ Chile
## 10 1921-10-12 Brazil    Paraguay           3          0 Copa Amé~  Buen~ Argent~
## # ... with 542 more rows, and 1 more variable: neutral <lgl>
```

3. Choose the games that Brazil won as the home team, and select variables `date`, `away_team` and `tournament`.

```
d %>% filter(home_team=="Brazil", home_score - away_score > 0) %>%
  select(date, away_team, tournament)
```

```
## # A tibble: 395 x 3
##    date       away_team tournament
##    <date>     <chr>     <chr>
##  1 1917-10-12 Chile     Copa América
##  2 1919-05-11 Chile     Copa América
##  3 1919-05-18 Argentina Copa América
##  4 1919-05-29 Uruguay   Copa América
##  5 1921-10-12 Paraguay  Copa América
##  6 1922-10-15 Argentina Copa América
##  7 1922-10-22 Paraguay  Copa América
##  8 1922-10-22 Argentina Copa Roca
##  9 1922-10-29 Paraguay  Friendly
## 10 1923-11-22 Paraguay  Friendly
## # ... with 385 more rows
```

4. Add the difference of goals, and an indicator variable called `goleada` for when the difference of goals is large, and select what we did only for Brazil. **Hint: use `ifelse`**.

```
d %>% mutate(dif = abs(away_score - home_score),
             goleada = ifelse(dif > 5, "Goleada", "Normal  Result")) %>%
  filter(home_team =="Brazil" | away_team=="Brazil", goleada=="Goleada")
```

```
## # A tibble: 41 x 11
##    date       home_team away_team home_score away_score tournament city  country
##    <date>     <chr>     <chr>          <dbl>      <dbl> <chr>      <chr> <chr>
##  1 1919-05-11 Brazil    Chile              6          0 Copa Amé~  Rio ~ Brazil
##  2 1920-09-18 Brazil    Uruguay            0          6 Copa Amé~  Viña~ Chile
##  3 1945-02-21 Brazil    Ecuador            9          2 Copa Amé~  Sant~ Chile
##  4 1949-04-03 Brazil    Ecuador            9          1 Copa Amé~  Rio ~ Brazil
##  5 1949-04-10 Brazil    Bolivia           10          1 Copa Amé~  São ~ Brazil
##  6 1949-04-24 Brazil    Peru               7          1 Copa Amé~  Rio ~ Brazil
##  7 1949-05-11 Brazil    Paraguay           7          0 Copa Amé~  Rio ~ Brazil
##  8 1950-07-09 Brazil    Sweden             7          1 FIFA Worl~ Rio ~ Brazil
##  9 1953-03-01 Bolivia   Brazil             1          8 Copa Amé~  Lima  Peru
## 10 1956-03-13 Brazil    Costa Ri~          7          1 Pan Ameri~ Mexi~ Mexico
## # ... with 31 more rows, and 3 more variables: neutral <lgl>, dif <dbl>,
## #   goleada <chr>
```

5. What was the largest difference in goals within these games?

```
d %>% mutate(dif = abs(away_score - home_score)) %>%
  arrange(desc(dif)) %>%
  slice(1)
```

```
## # A tibble: 1 x 10
##   date       home_team away_team  home_score away_score tournament city  country
##   <date>     <chr>     <chr>           <dbl>      <dbl> <chr>      <chr> <chr>
## 1 2001-04-11 Australia American ~         31          0 FIFA Worl~ Coff~ Austra~
## # ... with 2 more variables: neutral <lgl>, dif <dbl>
```

6. The top 5 goleadas?

```
d %>% mutate(dif = abs(away_score - home_score)) %>%
  arrange(desc(dif)) %>%
  slice(1:5) # top_n(5) here would also do the trick
```

```
## # A tibble: 5 x 10
##   date       home_team  away_team home_score away_score tournament city  country
##   <date>     <chr>      <chr>          <dbl>      <dbl> <chr>      <chr> <chr>
## 1 2001-04-11 Australia  American~         31          0 FIFA Worl~ Coff~ Austra~
## 2 1979-08-30 Fiji       Kiribati          24          0 South Pac~ Naus~ Fiji
## 3 2001-04-09 Australia  Tonga             22          0 FIFA Worl~ Coff~ Austra~
## 4 2005-03-11 Guam       Korea DPR          0         21 EAFF Cham~ Taip~ Taiwan
## 5 1987-12-15 American ~ Papua Ne~          0         20 South Pac~ Noum~ New Ca~
## # ... with 2 more variables: neutral <lgl>, dif <dbl>
```

7. Summary on goals scored by home teams, such as `mean` of `home_score` and `away_score`, `std`, using `group_by` and `summarise`

```
d %>% group_by(home_team) %>%
  summarise(mean_home_gols= mean(home_score, na.rm = TRUE),
            sd_home_gols= sd(home_score),
            mean_home_gols_op = mean(away_score), count=n()) %>%
  ungroup() %>%
  top_n(., 10, wt=mean_home_gols)  # note here, it does not give you in order
```

```
## # A tibble: 10 x 5
##    home_team       mean_home_gols sd_home_gols mean_home_gols_op count
##    <chr>                    <dbl>        <dbl>             <dbl> <int>
##  1 Cascadia                  4           NA                0         1
##  2 Gotland                   3.35         2.96             2.1      20
##  3 Guersney                  3.25         2.51             1.04     24
##  4 Kárpátalja                3.4          1.14             1.8       5
##  5 Micronesia FS             3.5          4.95             9         2
##  6 North Vietnam             3.6          3.13             1.8       5
##  7 Northern Cyprus           3.70         3.97             0.609    23
##  8 Sápmi                     4.12         5.75             1.5      16
##  9 Somaliland                3.5          4.95             2.5       2
## 10 Western Sahara            3.33         1.53             2.33      3
```

```
                                # you can add arrange
```

8. Proportion of victories of **Brazil** on different tournaments against each opponent, for instance, **Argentina**.

```
d %>% filter(home_team=="Brazil") %>%
  mutate(dif = home_score - away_score, victory = ifelse(dif>0, 1, 0)) %>%
```

```r
  group_by(away_team, tournament) %>%
  summarise(mean_victory_brazil= mean(victory, na.rm = TRUE), number_games=n()) %>%
  filter(away_team=="Argentina")
```

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

```
## # A tibble: 7 x 4
## # Groups:   away_team [1]
##   away_team tournament                    mean_victory_brazil number_games
##   <chr>     <chr>                                       <dbl>        <int>
## 1 Argentina Atlantic Cup                                    1            1
## 2 Argentina Confederations Cup                              1            1
## 3 Argentina Copa América                                    0.7         10
## 4 Argentina Copa Roca                                       0.538       13
## 5 Argentina FIFA World Cup                                  0            1
## 6 Argentina FIFA World Cup qualification                    1            3
## 7 Argentina Friendly                                        0.5         16
```