Lab 3: Data Wrangling on Soccer Tournament Data

July 9, 2020

Soccer tournament data wrangling

Read the dataset of football games.

```
setwd("D:/github/Rlab/")
d <- read_csv("data/results.csv")
```

```
## Parsed with 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`.

```
d1 = d %>% dplyr::select(date, home_team, away_team)
```

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

```
d2 = d1 %>% dplyr::filter(home_team == "Brazil")
```

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

```
d3 = d %>% dplyr::filter(home_team == "Brazil") %>%
  dplyr::filter(home_score>away_score) %>%
  dplyr::select(date, home_team, tournament)
```

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`**.

```
d4 = d %>% dplyr::mutate(diff_goal = home_score-away_score)%>%
  dplyr::mutate(goleada = ifelse(diff_goal >= 5|diff_goal<=-5, 1, 0)) %>%
  dplyr::filter(home_team == "Brazil"|away_team == "Brazil")
```

5. What was the largest difference in goals within these games?
6. The top 5 goleadas?
7. Summary on goals scored by home teams, such as `mean` of `home_score` and `away_score`, `std`, using `group_by` and `summarise`

```r
d %>% dplyr::group_by(home_team)%>%
  summarize(mean_homescore = mean(home_score))%>%
  ungroup()
```

```
## # A tibble: 291 x 2
##    home_team      mean_homescore
##    <chr>                   <dbl>
##  1 <U+00C5>land Islands         1.74
##  2 Abkhazia                2.21
##  3 Afghanistan             1.36
##  4 Albania                 1.11
##  5 Alderney                0.5
##  6 Algeria                 1.78
##  7 American Samoa          0.75
##  8 Andorra                 0.324
##  9 Angola                  1.42
## 10 Anguilla                0.688
## # ... with 281 more rows
```

```r
d %>% dplyr::group_by(home_team)%>%
  summarize(cont_awayscore = mean(away_score))%>%
  ungroup()
```

```
## # A tibble: 291 x 2
##    home_team      cont_awayscore
##    <chr>                   <dbl>
##  1 <U+00C5>land Islands          1.52
##  2 Abkhazia                0.571
##  3 Afghanistan             1.58
##  4 Albania                 1.14
##  5 Alderney                3.83
##  6 Algeria                 0.865
##  7 American Samoa          6.2
##  8 Andorra                 2.23
##  9 Angola                  0.857
## 10 Anguilla                2.38
## # ... with 281 more rows
```

```r
d %>% dplyr::group_by(home_team)%>%
  summarize(std_homescore = sd(home_score))%>%
  ungroup()
```

```
## # A tibble: 291 x 2
##    home_team      std_homescore
##    <chr>                  <dbl>
##  1 <U+00C5>land Islands         1.35
##  2 Abkhazia               2.49
##  3 Afghanistan            1.40
##  4 Albania                1.14
##  5 Alderney               0.837
##  6 Algeria                1.69
##  7 American Samoa         0.967
##  8 Andorra                0.552
##  9 Angola                 1.34
## 10 Anguilla               1.08
```

```
## # ... with 281 more rows
```

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

```
d1 = d %>% dplyr::filter(home_team == "Brazil"|away_team == "Brazil")%>%
  dplyr::mutate(against = ifelse(home_team == "Brazil",away_team,home_team))%>%
  dplyr::mutate(all = 1)%>%
  dplyr::mutate(win = ifelse(home_team == "Brazil",ifelse(home_score>away_score,1,0),ifelse(home_score<a


d1%>% group_by(tournament,against)%>%
  summarize(win_rate = sum(win)/sum(all))
```

```
## # A tibble: 209 x 3
## # Groups:   tournament [19]
##    tournament              against        win_rate
##    <chr>                   <chr>             <dbl>
##  1 Atlantic Cup            Argentina         0.5
##  2 Atlantic Cup            Paraguay          1
##  3 Atlantic Cup            Uruguay           0.5
##  4 Brazil Independence Cup Czechoslovakia    0
##  5 Brazil Independence Cup Portugal          1
##  6 Brazil Independence Cup Scotland          1
##  7 Brazil Independence Cup Yugoslavia        1
##  8 Confederations Cup      Argentina         1
##  9 Confederations Cup      Australia         0.333
## 10 Confederations Cup      Cameroon          0.5
## # ... with 199 more rows
```