Week 2 Assignment

01 August 2020

Data Coding and Merging

1.Import the packages we will need into R

```
library("tidyverse",quietly = TRUE)
```

2.Import NHL_team data, name it "NHL_Team" in Jupyter Notebook.

```
NHL_Team <- read.csv("NHL_team.csv")
```

3.Display Data to take a quick look of the data

```
head(NHL_Team)
```

```
##   X tid                name       tname      lname tricode abbr      sname
## 1 1   1 Toronto Maple Leafs Maple Leafs    Toronto     TOR  TOR    Toronto
## 2 2   2 MontrÃ©al Canadiens   Canadiens  MontrÃ©al     MTL  MTL  MontrÃ©al
## 3 3   4       Winnipeg Jets        Jets   Winnipeg     WPG  WPG   Winnipeg
## 4 4   5 Washington Capitals    Capitals Washington     WSH  WSH Washington
## 5 5   6  Chicago Blackhawks  Blackhawks    Chicago     CHI  CHI    Chicago
## 6 6   7     St. Louis Blues       Blues  St. Louis     STL  STL   St Louis
```

```
tail(NHL_Team)
```

```
##     X   tid                  name          tname     lname tricode abbr
## 28 28    51    New York Islanders      Islanders  New York     NYI  NYI
## 29 29    52 Columbus Blue Jackets   Blue Jackets  Columbus     CBJ  CBJ
## 30 30    53     Vancouver Canucks        Canucks Vancouver     VAN  VAN
## 31 31    59  Vegas Golden Knights Golden Knights     Vegas     VGK  VGK
## 32 32    66   Carolina Hurricanes     Hurricanes  Carolina     CAR  CAR
## 33 33 11366     Atlanta Thrashers      Thrashers   Atlanta     ATL  ATL
##           sname
## 28 NY Islanders
## 29     Columbus
## 30    Vancouver
## 31        Vegas
## 32     Carolina
## 33      Atlanta
```

4.Delete the "X", "abbr", "tname", "lname", and "sname" variables.

```
NHL_Team <- NHL_Team %>% dplyr::select(-X,-abbr,-tname,-lname,-sname)
head(NHL_Team)
```

```
##   tid                name tricode
## 1   1 Toronto Maple Leafs     TOR
## 2   2 MontrÃ©al Canadiens     MTL
## 3   4       Winnipeg Jets     WPG
## 4   5 Washington Capitals     WSH
## 5   6  Chicago Blackhawks     CHI
## 6   7     St. Louis Blues     STL
```

5.Rename the variable "name" to "team_name" in the NHL_Team dataframe.

```
NHL_Team <- NHL_Team %>% dplyr::rename(team_name = name)
head(NHL_Team)
```

```
##   tid           team_name tricode
## 1   1 Toronto Maple Leafs     TOR
## 2   2 MontrÃ©al Canadiens     MTL
## 3   4       Winnipeg Jets     WPG
## 4   5 Washington Capitals     WSH
## 5   6  Chicago Blackhawks     CHI
## 6   7     St. Louis Blues     STL
```

6.Import NHL_competition data, name it "NHL_Competition" in Jupyter notebook.

```
NHL_Competition <- read.csv("NHL_competition.csv")
head(NHL_Competition)
```

```
##   X comp_id year type                    name tz start end
## 1 1       1 2013    2 2013 NHL Regular Season ET    NA  NA
## 2 2       2 2017    2 2017 NHL Regular Season ET    NA  NA
## 3 3    2453 2013    3        2013 NHL Playoff ET    NA  NA
## 4 4    2541 2017    3        2017 NHL Playoff ET    NA  NA
## 5 5    2661 2012    2 2012 NHL Regular Season ET    NA  NA
## 6 6    2734 2016    2 2016 NHL Regular Season ET    NA  NA
```

7.Delete the "X", "tz","start" and "end" variables.

```
NHL_Competition <- NHL_Competition %>% dplyr::select(-X,-tz,-start,-end)
head(NHL_Competition)
```

```
##   comp_id year type                    name
## 1       1 2013    2 2013 NHL Regular Season
## 2       2 2017    2 2017 NHL Regular Season
## 3    2453 2013    3        2013 NHL Playoff
## 4    2541 2017    3        2017 NHL Playoff
## 5    2661 2012    2 2012 NHL Regular Season
## 6    2734 2016    2 2016 NHL Regular Season
```

8.Rename the variable "name" to "competition_name" in the NHL_Competition dataframe.

```
NHL_Competition <- NHL_Competition %>% dplyr::rename(competition_name = name)
head(NHL_Competition)
```

```
##   comp_id year type        competition_name
## 1       1 2013    2 2013 NHL Regular Season
## 2       2 2017    2 2017 NHL Regular Season
## 3    2453 2013    3        2013 NHL Playoff
## 4    2541 2017    3        2017 NHL Playoff
## 5    2661 2012    2 2012 NHL Regular Season
## 6    2734 2016    2 2016 NHL Regular Season
```

9.Import NHL_game data, name it "NHL_Game" in Jupyter notebook.

```
NHL_Game <- read.csv("NHL_game.csv")
head(NHL_Game)
```

```
##   X gid comp_id       date ascore hscore period status home_away tid
## 1 1  37       2  10/7/2017     NA     NA     NA     NA      away  25
## 2 2  67       2  10/9/2017     NA     NA     NA     NA      away  29
## 3 3 154       1 10/14/2013     NA     NA     NA     NA      away  29
## 4 4 278       1 10/24/2013     NA     NA     NA     NA      away  53
## 5 5 291       1 10/25/2013     NA     NA     NA     NA      away   5
## 6 6 294       1 10/25/2013     NA     NA     NA     NA      away  51
```

10.Delete the "X","period", and "status" variables.

```
NHL_Game <- NHL_Game %>% dplyr::select(-X,-period,-status)
head(NHL_Game)
```

```
##   gid comp_id       date ascore hscore home_away tid
## 1  37       2  10/7/2017     NA     NA      away  25
## 2  67       2  10/9/2017     NA     NA      away  29
## 3 154       1 10/14/2013     NA     NA      away  29
## 4 278       1 10/24/2013     NA     NA      away  53
## 5 291       1 10/25/2013     NA     NA      away   5
## 6 294       1 10/25/2013     NA     NA      away  51
```

11.Merge dataframe NHL_Team into dataframe NHL_Game by 'tid'.

```
NHL_Game <- merge(NHL_Game, NHL_Team, by='tid')
NHL_Game %>% head()
```

```
##   tid  gid comp_id       date ascore hscore home_away           team_name
## 1   1 6836    5662   1/4/2011      2      1      home Toronto Maple Leafs
## 2   1 1103       1 12/24/2013      2      3      away Toronto Maple Leafs
## 3   1 9100    8011   3/4/2015      3      2      away Toronto Maple Leafs
## 4   1 5385    5385  10/7/2015      3      1      home Toronto Maple Leafs
## 5   1   69       2  10/9/2017      3      4      home Toronto Maple Leafs
```

```
## 6   1   74       1  10/8/2013      2      1      home Toronto Maple Leafs
##   tricode
## 1     TOR
## 2     TOR
## 3     TOR
## 4     TOR
## 5     TOR
## 6     TOR
```

12.Merge dataframe NHL_Competition into dataframe NHL_Game by 'comp_id'.

```
NHL_Game <- merge(NHL_Game, NHL_Competition, by='comp_id')
NHL_Game %>% head()
```

```
##   comp_id tid  gid       date ascore hscore home_away           team_name
## 1       1  10  957 12/13/2013      4      2      home     Edmonton Oilers
## 2       1  17 1013 12/18/2013      2      4      home      Buffalo Sabres
## 3       1  16 1248   1/4/2014      2      4      away     San Jose Sharks
## 4       1  46 1699   2/5/2014      3      1      away        Dallas Stars
## 5       1  11 1286   1/7/2014      4      3      away      Calgary Flames
## 6       1  25 1902   3/9/2014      4      3      home Tampa Bay Lightning
##   tricode year type        competition_name
## 1     EDM 2013    2 2013 NHL Regular Season
## 2     BUF 2013    2 2013 NHL Regular Season
## 3     SJS 2013    2 2013 NHL Regular Season
## 4     DAL 2013    2 2013 NHL Regular Season
## 5     CGY 2013    2 2013 NHL Regular Season
## 6     TBL 2013    2 2013 NHL Regular Season
```

13.Create a variable to indicate the goal difference between home and away score (hscore-ascore), name this varible "hgd".

```
NHL_Game['hgd']=NHL_Game['hscore']-NHL_Game['ascore']
NHL_Game %>% head()
```

```
##   comp_id tid  gid       date ascore hscore home_away           team_name
## 1       1  10  957 12/13/2013      4      2      home     Edmonton Oilers
## 2       1  17 1013 12/18/2013      2      4      home      Buffalo Sabres
## 3       1  16 1248   1/4/2014      2      4      away     San Jose Sharks
## 4       1  46 1699   2/5/2014      3      1      away        Dallas Stars
## 5       1  11 1286   1/7/2014      4      3      away      Calgary Flames
## 6       1  25 1902   3/9/2014      4      3      home Tampa Bay Lightning
##   tricode year type        competition_name hgd
## 1     EDM 2013    2 2013 NHL Regular Season  -2
## 2     BUF 2013    2 2013 NHL Regular Season   2
## 3     SJS 2013    2 2013 NHL Regular Season   2
## 4     DAL 2013    2 2013 NHL Regular Season  -2
```

```
## 5     CGY 2013    2 2013 NHL Regular Season  -1
## 6     TBL 2013    2 2013 NHL Regular Season  -1
```

14.Delete observations with missing value in the variable "hgd".

```
NHL_Game <- na.omit(NHL_Game, cols="hgd")
```

15.Drop observations with missing values in the NHL_Game.

```
NHL_Game <- NHL_Game %>% na.omit()
NHL_Game %>% head()
```

```
##   comp_id tid  gid       date ascore hscore home_away           team_name
## 1       1  10  957 12/13/2013      4      2      home     Edmonton Oilers
## 2       1  17 1013 12/18/2013      2      4      home      Buffalo Sabres
## 3       1  16 1248   1/4/2014      2      4      away     San Jose Sharks
## 4       1  46 1699   2/5/2014      3      1      away        Dallas Stars
## 5       1  11 1286   1/7/2014      4      3      away      Calgary Flames
## 6       1  25 1902   3/9/2014      4      3      home Tampa Bay Lightning
##   tricode year type        competition_name hgd
## 1     EDM 2013    2 2013 NHL Regular Season  -2
## 2     BUF 2013    2 2013 NHL Regular Season   2
## 3     SJS 2013    2 2013 NHL Regular Season   2
## 4     DAL 2013    2 2013 NHL Regular Season  -2
## 5     CGY 2013    2 2013 NHL Regular Season  -1
## 6     TBL 2013    2 2013 NHL Regular Season  -1
```

16.What are the number of observations and the number of variables in the NHL_Game data?

```
dim(NHL_Game)
```

```
## [1] 18506    13
```

17.Convert type of "date" variable from object to datetime.

```
NHL_Game['date']=as.Date(NHL_Game['date'],format = "%m/%d/%y")
NHL_Game['date'] %>% head()
```

```
## [1] "2020-12-13" "2020-12-18" "2020-01-04" "2020-02-05" "2020-01-07"
## [6] "2020-03-09"
```

18.What is the time range of the NHL_Game dataframe?

```
NHL_Game['date']%>% summary()
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.
## "2020-01-01" "2020-02-26" "2020-04-15" "2020-06-21" "2020-11-11"
##         Max.
## "2020-12-31"
```

19.Sort the games by "date" and show the first 15 observations.

```
NHL_Game %>% arrange(date) %>% head(15)
```

```
##    comp_id tid  gid       date ascore hscore home_away
## 1        1   1 1215 2020-01-01      4      3      away
## 2        1  20 1199 2020-01-01      5      3      home
## 3        1   2 1201 2020-01-01      4      5      away
## 4        1  51 1199 2020-01-01      5      3      away
## 5        1  41 1205 2020-01-01      2      3      away
## 6        1  46 1205 2020-01-01      2      3      home
## 7        1  66 1201 2020-01-01      4      5      home
## 8        1  10 1213 2020-01-01      3      4      away
## 9        1   4 1203 2020-01-01      0      3      home
## 10       1  16 1209 2020-01-01      3      6      away
## 11       1  52 1207 2020-01-01      3      5      away
## 12       1  50 1213 2020-01-01      3      4      home
## 13       1  22 1207 2020-01-01      3      5      home
## 14       1  18 1215 2020-01-01      4      3      home
## 15       1  21 1209 2020-01-01      3      6      home
##                team_name tricode year type        competition_name hgd
## 1    Toronto Maple Leafs     TOR 2013    2 2013 NHL Regular Season  -1
## 2          Boston Bruins     BOS 2013    2 2013 NHL Regular Season  -2
## 3    MontrÃ©al Canadiens     MTL 2013    2 2013 NHL Regular Season   1
## 4     New York Islanders     NYI 2013    2 2013 NHL Regular Season  -2
## 5      Los Angeles Kings     LAK 2013    2 2013 NHL Regular Season   1
## 6           Dallas Stars     DAL 2013    2 2013 NHL Regular Season   1
## 7    Carolina Hurricanes     CAR 2013    2 2013 NHL Regular Season   1
## 8        Edmonton Oilers     EDM 2013    2 2013 NHL Regular Season   1
## 9          Winnipeg Jets     WPG 2013    2 2013 NHL Regular Season   3
## 10       San Jose Sharks     SJS 2013    2 2013 NHL Regular Season   3
## 11 Columbus Blue Jackets     CBJ 2013    2 2013 NHL Regular Season   2
## 12       Phoenix Coyotes     PHX 2013    2 2013 NHL Regular Season   1
## 13    Colorado Avalanche     COL 2013    2 2013 NHL Regular Season   2
## 14     Detroit Red Wings     DET 2013    2 2013 NHL Regular Season  -1
## 15         Anaheim Ducks     ANA 2013    2 2013 NHL Regular Season   3
```

20.Seperate the NHL_Game dataframe for home games and away games. Name them "NHL_Home" and "NHL_Away", respectively.

```
NHL_Home <- NHL_Game %>% filter(home_away == 'home')
NHL_Home %>% head()
```

```
##   comp_id tid  gid       date ascore hscore home_away           team_name
## 1       1  10  957 2020-12-13      4      2      home     Edmonton Oilers
## 2       1  17 1013 2020-12-18      2      4      home      Buffalo Sabres
## 3       1  25 1902 2020-03-09      4      3      home Tampa Bay Lightning
```

```
## 4       1   2  532 2020-11-13      2      1      home MontrÃ©al Canadiens
## 5       1  10  719 2020-11-26      5      1      home     Edmonton Oilers
## 6       1  32  943 2020-12-13      1      2      home     Ottawa Senators
##   tricode year type        competition_name hgd
## 1     EDM 2013    2 2013 NHL Regular Season  -2
## 2     BUF 2013    2 2013 NHL Regular Season   2
## 3     TBL 2013    2 2013 NHL Regular Season  -1
## 4     MTL 2013    2 2013 NHL Regular Season  -1
## 5     EDM 2013    2 2013 NHL Regular Season  -4
## 6     OTT 2013    2 2013 NHL Regular Season   1
```

```
NHL_Away <- NHL_Game %>% filter(home_away == 'away')
NHL_Away %>% head()
```

```
##   comp_id tid  gid       date ascore hscore home_away           team_name
## 1       1  16 1248 2020-01-04      2      4      away     San Jose Sharks
## 2       1  46 1699 2020-02-05      3      1      away        Dallas Stars
## 3       1  11 1286 2020-01-07      4      3      away      Calgary Flames
## 4       1  28  837 2020-12-06      3      1      away    New York Rangers
## 5       1  14  714 2020-11-26      1      3      away Philadelphia Flyers
## 6       1  35 2290 2020-04-04      2      3      away      Minnesota Wild
##   tricode year type        competition_name hgd
## 1     SJS 2013    2 2013 NHL Regular Season   2
## 2     DAL 2013    2 2013 NHL Regular Season  -2
## 3     CGY 2013    2 2013 NHL Regular Season  -1
## 4     NYR 2013    2 2013 NHL Regular Season  -2
## 5     PHI 2013    2 2013 NHL Regular Season   2
## 6     MIN 2013    2 2013 NHL Regular Season   1
```

21.Rename variables

- For away games, rename “ascore” to “goals_for” and “hscore” to “goals_against”
- For home games,rename “hscore” to “goals_for” and “ascore” to “goals_against”

```
NHL_Away <- NHL_Away %>% dplyr::rename(goals_for = ascore, goals_against = hscore)

NHL_Home <- NHL_Home %>% dplyr::rename(goals_for = hscore, goals_against = ascore)

NHL_Away %>% head()
```

```
##   comp_id tid  gid       date goals_for goals_against home_away
## 1       1  16 1248 2020-01-04         2             4      away
## 2       1  46 1699 2020-02-05         3             1      away
## 3       1  11 1286 2020-01-07         4             3      away
## 4       1  28  837 2020-12-06         3             1      away
## 5       1  14  714 2020-11-26         1             3      away
## 6       1  35 2290 2020-04-04         2             3      away
```

```
##             team_name tricode year type        competition_name hgd
## 1     San Jose Sharks     SJS 2013    2 2013 NHL Regular Season   2
## 2        Dallas Stars     DAL 2013    2 2013 NHL Regular Season  -2
## 3      Calgary Flames     CGY 2013    2 2013 NHL Regular Season  -1
## 4    New York Rangers     NYR 2013    2 2013 NHL Regular Season  -2
## 5 Philadelphia Flyers     PHI 2013    2 2013 NHL Regular Season   2
## 6      Minnesota Wild     MIN 2013    2 2013 NHL Regular Season   1
```

```
NHL_Home %>% head()
```

```
##   comp_id tid  gid       date goals_against goals_for home_away
## 1       1  10  957 2020-12-13             4         2      home
## 2       1  17 1013 2020-12-18             2         4      home
## 3       1  25 1902 2020-03-09             4         3      home
## 4       1   2  532 2020-11-13             2         1      home
## 5       1  10  719 2020-11-26             5         1      home
## 6       1  32  943 2020-12-13             1         2      home
##             team_name tricode year type        competition_name hgd
## 1     Edmonton Oilers     EDM 2013    2 2013 NHL Regular Season  -2
## 2      Buffalo Sabres     BUF 2013    2 2013 NHL Regular Season   2
## 3 Tampa Bay Lightning     TBL 2013    2 2013 NHL Regular Season  -1
## 4 MontrÃ©al Canadiens     MTL 2013    2 2013 NHL Regular Season  -1
## 5     Edmonton Oilers     EDM 2013    2 2013 NHL Regular Season  -4
## 6     Ottawa Senators     OTT 2013    2 2013 NHL Regular Season   1
```

22.Create a "win" variable equals to 1 if the team won the game; 0 if the team lost the game; and 0.5 if it was a draw.

- For home team, win=1 if the variable "hgd" is positive
- For away team, win=1 if the variable "hgd" is negative

```
NHL_Home <- NHL_Home %>% mutate(win = case_when(
  hgd < 0 ~ 0,
  hgd == 0 ~ 0.5,
  hgd > 0 ~ 1))

NHL_Away <- NHL_Away %>% mutate(win = case_when(
  hgd < 0 ~ 1,
  hgd == 0 ~ 0.5,
  hgd > 0 ~ 0))
```

23.Append the NHL_Home and NHL_Away dataframes as the new NHL_Game dataframe.

```
NHL_Game <- rbind(NHL_Home, NHL_Away)
head(NHL_Game)
```

```
##   comp_id tid  gid       date goals_against goals_for home_away
## 1       1  10  957 2020-12-13             4         2      home
```

```
## 2       1  17 1013 2020-12-18             2         4      home
## 3       1  25 1902 2020-03-09             4         3      home
## 4       1   2  532 2020-11-13             2         1      home
## 5       1  10  719 2020-11-26             5         1      home
## 6       1  32  943 2020-12-13             1         2      home
##              team_name tricode year type        competition_name hgd win
## 1     Edmonton Oilers     EDM 2013    2 2013 NHL Regular Season  -2   0
## 2      Buffalo Sabres     BUF 2013    2 2013 NHL Regular Season   2   1
## 3 Tampa Bay Lightning     TBL 2013    2 2013 NHL Regular Season  -1   0
## 4 MontrÃ©al Canadiens     MTL 2013    2 2013 NHL Regular Season  -1   0
## 5     Edmonton Oilers     EDM 2013    2 2013 NHL Regular Season  -4   0
## 6     Ottawa Senators     OTT 2013    2 2013 NHL Regular Season   1   1
```

```
tail(NHL_Game)
```

```
##       comp_id tid  gid       date goals_against goals_for home_away
## 18501    9389  19 9419 2020-04-22             3         2      away
## 18502    9389   2 9435 2020-05-07             2         6      away
## 18503    9389   6 9468 2020-06-04             1         2      away
## 18504    9389  32 9386 2020-04-17             3         2      away
## 18505    9389  18 9395 2020-04-25             0         4      away
## 18506    9389  21 9424 2020-04-21             4         5      away
##                 team_name tricode year type competition_name hgd win
## 18501 Nashville Predators     NSH 2014    3 2014 NHL Playoff   1   0
## 18502 MontrÃ©al Canadiens     MTL 2014    3 2014 NHL Playoff  -4   1
## 18503  Chicago Blackhawks     CHI 2014    3 2014 NHL Playoff  -1   1
## 18504     Ottawa Senators     OTT 2014    3 2014 NHL Playoff   1   0
## 18505   Detroit Red Wings     DET 2014    3 2014 NHL Playoff  -4   1
## 18506       Anaheim Ducks     ANA 2014    3 2014 NHL Playoff  -1   1
```

24.Double check the number of observations in the new NHL_Game dataframe to make sure that we have successfully combined the two dataframes.

```
dim(NHL_Game)
```

```
## [1] 18506    14
```

25.Generate team level variable

Create a team level dataframe called "NHL_Team_Stats" that aggregates the total number of games won, the total number of goals_for and goals_against for each team in each competition.

```
NHL_Team_Stats <- NHL_Game %>%
  group_by(tid,competition_name,type) %>% dplyr::summarise(win = sum(win),
                goals_for = sum(goals_for),
                goals_against = sum(goals_against))
```

```r
head(NHL_Team_Stats)
```

```
## # A tibble: 6 x 6
## # Groups:   tid, competition_name [6]
##     tid competition_name         type   win goals_for goals_against
##   <int> <fct>                   <int> <dbl>     <int>         <int>
## 1     1 2010 NHL Regular Season     2    36       223           259
## 2     1 2011 NHL Regular Season     2    20       129           129
## 3     1 2012 NHL Playoff            3     3        18            22
## 4     1 2012 NHL Regular Season     2    25       144           129
## 5     1 2013 NHL Regular Season     2    38       231           250
## 6     1 2014 NHL Regular Season     2    29       209           258
```

26.Convert the indexes back as variables in the NHL_Team_Stats dataframe.

** There is no such step in R.**

27.Create a dataframe "NHL_Game_Count" that include the total number of games played by each team in competition.

- Name this variable "game_count"

```r
NHL_Game_Count <- NHL_Game %>%
  group_by(tid,competition_name,type) %>%
  dplyr::summarise(game_count = n())
head(NHL_Game_Count)
```

```
## # A tibble: 6 x 4
## # Groups:   tid, competition_name [6]
##     tid competition_name         type game_count
##   <int> <fct>                   <int>      <int>
## 1     1 2010 NHL Regular Season     2         82
## 2     1 2011 NHL Regular Season     2         40
## 3     1 2012 NHL Playoff            3          7
## 4     1 2012 NHL Regular Season     2         46
## 5     1 2013 NHL Regular Season     2         79
## 6     1 2014 NHL Regular Season     2         78
```

```r
tail(NHL_Game_Count)
```

```
## # A tibble: 6 x 4
## # Groups:   tid, competition_name [6]
##     tid competition_name         type game_count
##   <int> <fct>                   <int>      <int>
## 1    66 2013 NHL Regular Season     2         80
## 2    66 2014 NHL Regular Season     2         82
## 3    66 2015 NHL Regular Season     2         81
## 4    66 2016 NHL Regular Season     2         82
```

```
## 5    66 2017 NHL Regular Season     2         80
## 6 11366 2010 NHL Regular Season     2         79
```

28.Merge the NHL_Game_Count dataframe to the NHL_Team_Stats dataframe.

```
NHL_Team_Stats <- merge(NHL_Team_Stats, NHL_Game_Count, by=c('tid','competition_name','t
head(NHL_Team_Stats)
```

```
##   tid        competition_name type win goals_for goals_against game_count
## 1   1 2010 NHL Regular Season    2  36       223           259         82
## 2   1 2011 NHL Regular Season    2  20       129           129         40
## 3   1        2012 NHL Playoff    3   3        18            22          7
## 4   1 2012 NHL Regular Season    2  25       144           129         46
## 5   1 2013 NHL Regular Season    2  38       231           250         79
## 6   1 2014 NHL Regular Season    2  29       209           258         78
```

29.Merge the NHL_Team dataframe to the NHL_Team_Stats dataframe.

```
NHL_Team_Stats <- merge(NHL_Team_Stats, NHL_Team,by='tid')
head(NHL_Team_Stats)
```

```
##   tid        competition_name type win goals_for goals_against game_count
## 1   1 2010 NHL Regular Season    2  36       223           259         82
## 2   1 2011 NHL Regular Season    2  20       129           129         40
## 3   1        2012 NHL Playoff    3   3        18            22          7
## 4   1 2012 NHL Regular Season    2  25       144           129         46
## 5   1 2013 NHL Regular Season    2  38       231           250         79
## 6   1 2014 NHL Regular Season    2  29       209           258         78
##             team_name tricode
## 1 Toronto Maple Leafs     TOR
## 2 Toronto Maple Leafs     TOR
## 3 Toronto Maple Leafs     TOR
## 4 Toronto Maple Leafs     TOR
## 5 Toronto Maple Leafs     TOR
## 6 Toronto Maple Leafs     TOR
```

30.Import PPPK dataset and name it "NHL_PPPK".

```
NHL_PPPK <- read.csv("PP.PK.PPgf.csv")
head(NHL_PPPK)
```

```
##   tricode  pp  pk ppgf competition_name
## 1     ANA  35  27    9 2010 NHL Playoff
## 2     BOS 126 116   22 2010 NHL Playoff
## 3     BUF  48  46   13 2010 NHL Playoff
## 4     CHI  27  39    6 2010 NHL Playoff
## 5     DET  59  55    6 2010 NHL Playoff
## 6     LAK  24  27    5 2010 NHL Playoff
```

31.Merge the NHL_PPPK dataframe to the NHL_Team_Stats dataframe.

```
NHL_Team_Stats <- merge(NHL_PPPK, NHL_Team_Stats, by =c('tricode','competition_name'))
head(NHL_Team_Stats)
```

```
##   tricode        competition_name  pp  pk ppgf tid type  win goals_for
## 1     ANA        2010 NHL Playoff  35  27    9  21    3  2.0        19
## 2     ANA 2010 NHL Regular Season 401 378   92  21    2 43.5       227
## 3     ANA 2011 NHL Regular Season 203 176   29  21    2 11.0        96
## 4     ANA        2012 NHL Playoff  28  28   10  21    3  3.0        21
## 5     ANA 2012 NHL Regular Season 196 173   38  21    2 30.0       149
## 6     ANA        2013 NHL Playoff  78  69   12  21    3  7.0        35
##   goals_against game_count     team_name
## 1            22          6 Anaheim Ducks
## 2           225         78 Anaheim Ducks
## 3           133         39 Anaheim Ducks
## 4            18          7 Anaheim Ducks
## 5           125         48 Anaheim Ducks
## 6            37         13 Anaheim Ducks
```

32.Create additional variables in the "NHL_Team_Stats" dataframe.

- winning percentage('win_pct')= 'win' / total number of games played
- average goals for per game ('avg_gf')= total number of goals for / total number of games played
- average goals against per game ('avg_ga') = total number of goals against / total number of games played

```
NHL_Team_Stats[,'win_pct']=NHL_Team_Stats[,'win']/NHL_Team_Stats[,'game_count']
NHL_Team_Stats[,'avg_gf']=NHL_Team_Stats[,'goals_for']/NHL_Team_Stats[,'game_count']
NHL_Team_Stats[,'avg_ga']=NHL_Team_Stats[,'goals_against']/NHL_Team_Stats[,'game_count']
```

```
NHL_Team_Stats['win_pct'] %>% summary()
```

```
##     win_pct
##  Min.   :0.0000
##  1st Qu.:0.4067
##  Median :0.5000
##  Mean   :0.4722
##  3rd Qu.:0.5625
##  Max.   :0.8000
```

33.In the NHL_Competition dataframe, the variable type indicates the type of competition. type=2 means it is regular season competition.

Create a dataframe (NHL_Team_R_Stats) that contains team statistics for games only during regular seasons.

```
NHL_Team_R_Stats <- NHL_Team_Stats %>% filter(type == 2)
```

Descriptive and Summary Analyses

Return to the NHL_Game dataframe,

34.Calculate summary statistics for the variable "goals_for".

```
NHL_Game['goals_for'] %>% summary()
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   2.000   3.000   2.826   4.000  10.000
```

35.Caclulate the summary statistics for the "goals_against" for the entire NHL_Game dataframe by whether the game is home or away.

```
NHL_Game %>% group_by(home_away) %>%
  summarise(count = n(),
            mean = mean(goals_against),
            std = sd(goals_against),
            min = min(goals_against),
            Q1 = quantile(goals_against,0.25),
            Median = quantile(goals_against,0.5),
            Q3 = quantile(goals_against,0.75),
            max = max(goals_against))
```

```
## # A tibble: 2 x 9
##   home_away count  mean   std   min    Q1 Median    Q3   max
##   <fct>     <int> <dbl> <dbl> <int> <dbl>  <dbl> <dbl> <int>
## 1 away       9253  2.96  1.69     0     2      3     4    10
## 2 home       9253  2.69  1.61     0     1      3     4    10
```

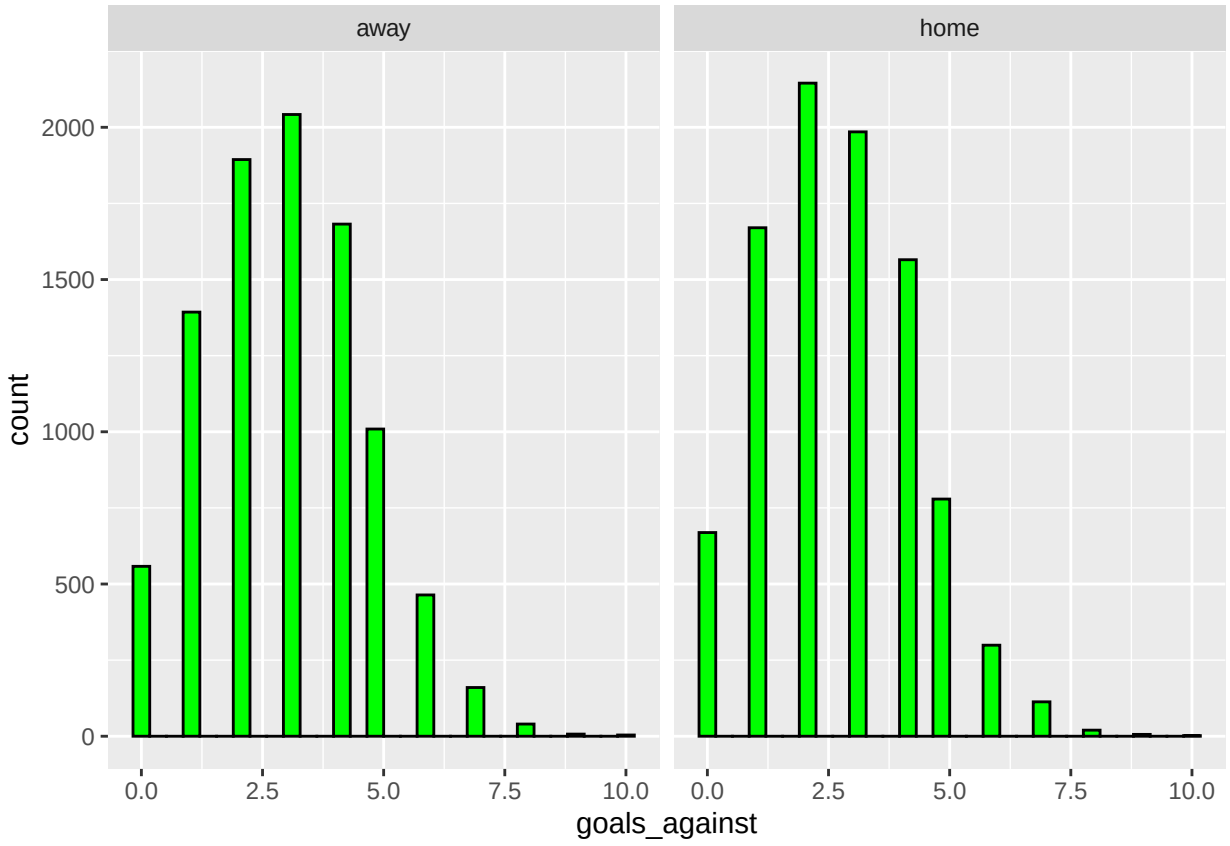
36.Create a histogram of the "goals_against" variable by whether the game is home or away.

- Make the color of the histogram green
- Set the number of bins to be 20
- Make sure that the two sub-histograms share the same ranges for the x-axis and y-axis

Save the histogram as "goals_against.png".

```
p1 <- ggplot(data = NHL_Game, aes(x=goals_against))+
  geom_histogram(color="black", fill="green")+
  facet_grid(.~home_away)
p1
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
ggsave("goals_against.png",p1,width = 4, height = 4.5)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

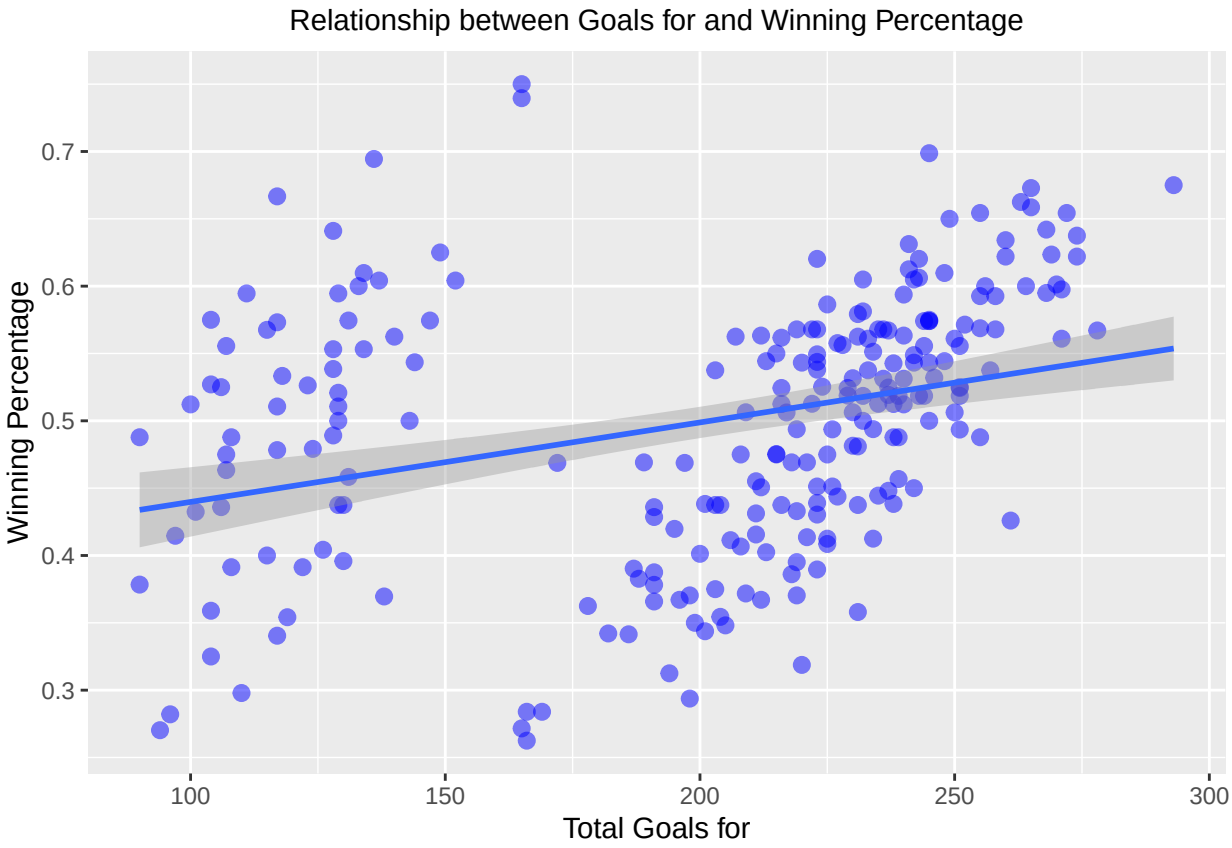
Correlation Analyses

37.In the NHL_Team_R_Stats dataframe, make a scatter plot to depict the relationship between the total number of goals for and the winning percentage.

- Plot the total number of *goals for* on the x-axis and winning percentage on the y-axis
- Add a regression line to the scatter plot
- Make the title of the graph "Relationship between Goals for and Winning Percentage" and make the font size 11
- Label the x-axis as "Total Goals for" and label the y-axis as "Winning Percentage"

```
p2 <- ggplot(data = NHL_Team_R_Stats,aes(x = goals_for, y = win_pct))+
  geom_point(color = "blue", size = 2.5,alpha = 0.5)+
  geom_smooth(method='lm',formula=y~x)+
  xlab("Total Goals for")+
  ylab("Winning Percentage")+
  ggtitle("Relationship between Goals for and Winning Percentage")+
    theme(
        plot.title = element_text(size = 11,hjust = 0.5))
```

```
p2
```




```
ggsave("scatter_gf_winpct.png",p2,width = 4, height = 4.5)
```

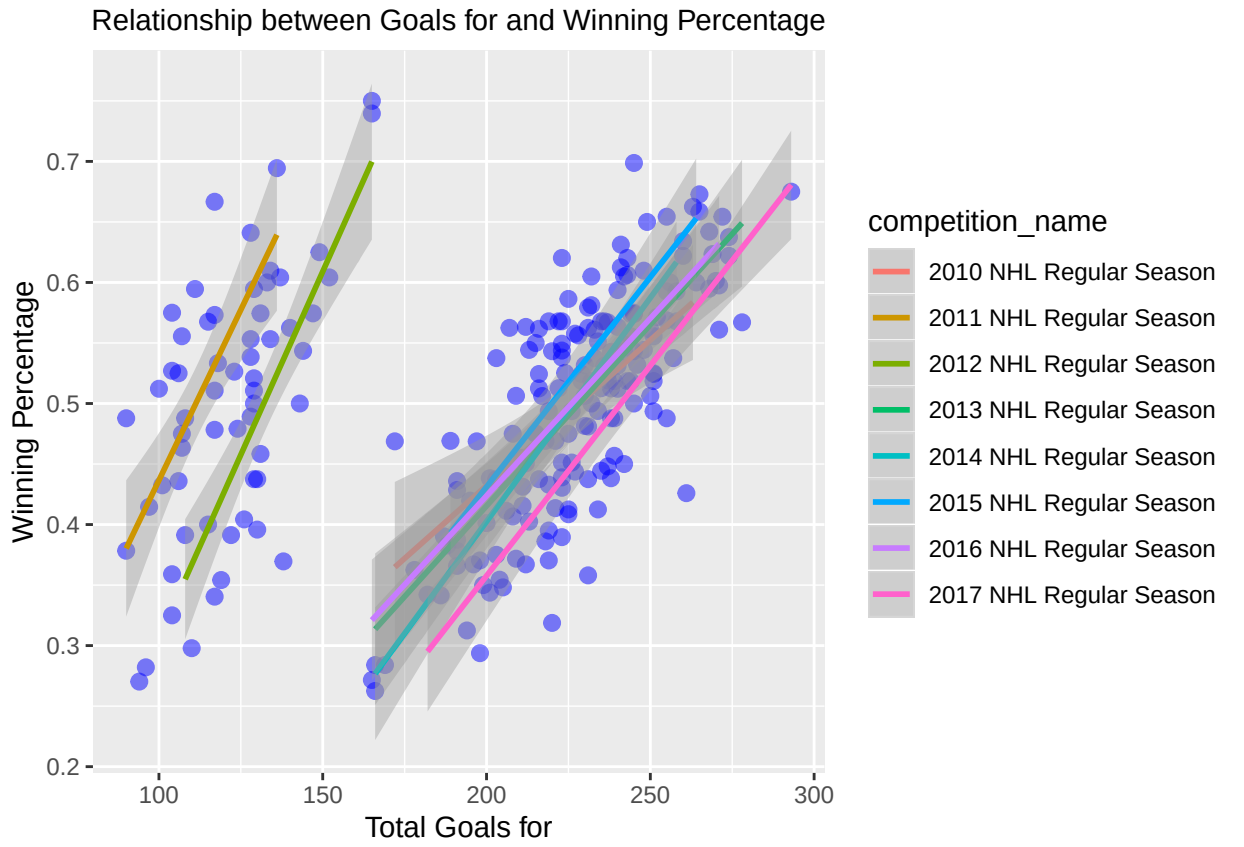
38.Calulate the correlation coefficient between total *goals for* and winning percentage.

```
cor(NHL_Team_R_Stats[,'goals_for'],NHL_Team_R_Stats[,'win_pct'])
```

```
## [1] 0.3156646
```

39.Create a scatter plot of total number of *goals for* and winning percentage, grouped by competition.

```
p3 <- ggplot(data = NHL_Team_R_Stats,aes(x = goals_for, y = win_pct),)+
  geom_point(color = "blue", size = 2.5,alpha = 0.5)+
  geom_smooth(aes(color = competition_name), method='lm',formula=y~x)+
  xlab("Total Goals for")+
  ylab("Winning Percentage")+
  ggtitle("Relationship between Goals for and Winning Percentage")+
    theme(
        plot.title = element_text(size = 11,hjust = 0.5))
p3
```



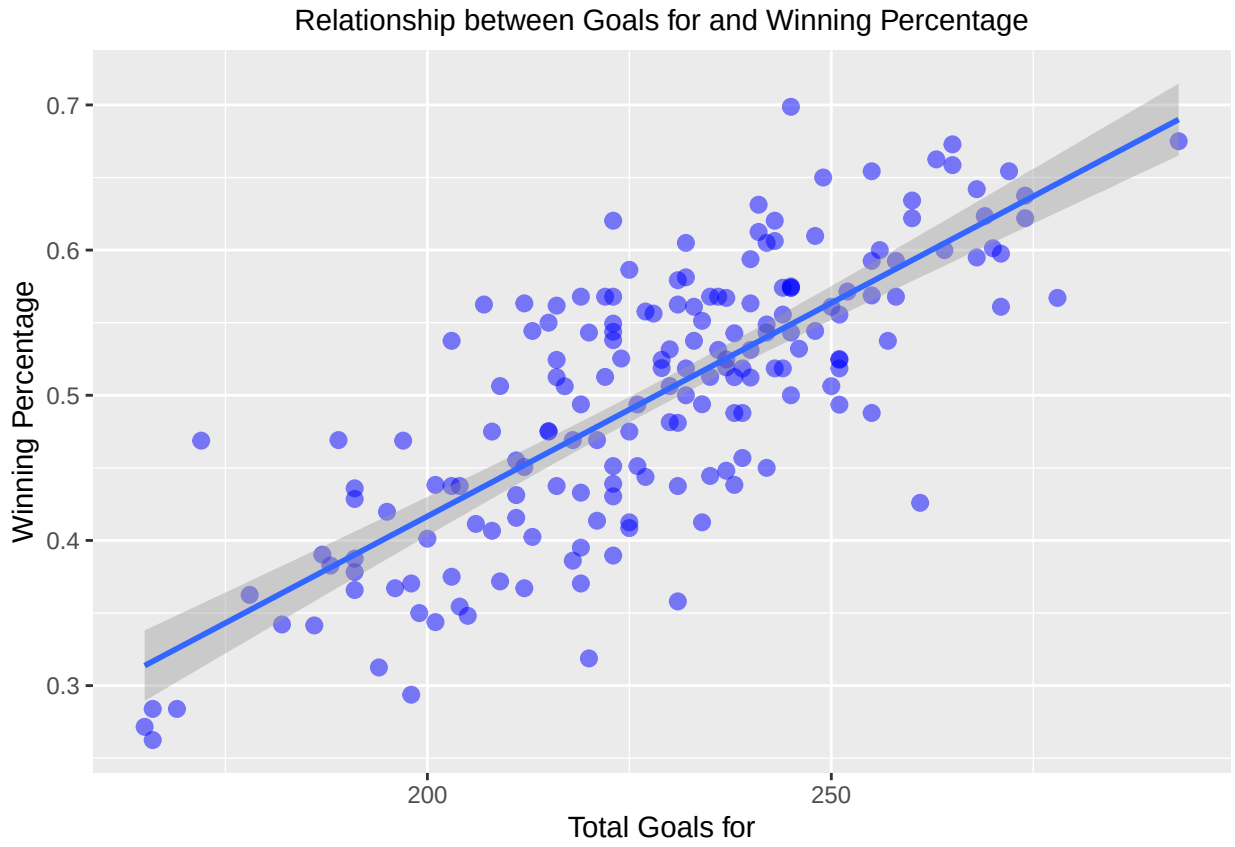

```
ggsave("scatter_gf_winpct_comp.png",p3,width = 4, height = 4.5)
```

40.Filter out the 2011 and 2012 seasons, continue to name this dataframe NHL_Team_R_Stats

```
NHL_Team_R_Stats <- NHL_Team_R_Stats%>%filter((competition_name != '2011 NHL Regular Sea
&(competition_name != '2012 NHL Regular Season'))
```

41.Create a scatter plot to depict the relationship between total *goals for* and winning percentage without seperating the data by competition.

```
p4 <- ggplot(data = NHL_Team_R_Stats,aes(x = goals_for, y = win_pct))+
  geom_point(color = "blue", size = 2.5,alpha = 0.5)+
  geom_smooth(method='lm',formula=y~x)+
  xlab("Total Goals for")+
  ylab("Winning Percentage")+
  ggtitle("Relationship between Goals for and Winning Percentage")+
    theme(
        plot.title = element_text(size = 11,hjust = 0.5))
p4
```




```
ggsave("scatter_gf_winpct_new.png",p4,width = 4, height = 4.5)
```

42.Calculate the correlation coefficient between goals for and winning percentage in the updated "NHL_Team_R_Stats" dataframe.

```
cor(NHL_Team_R_Stats[,'goals_for'],NHL_Team_R_Stats[,'win_pct'])
```

```
## [1] 0.7706255
```

43.Save dataframes as csv files.

- Name the updated NHL_Game as "NHL_Game2"
- Name the NHL_Team_Stats as "NHL_Team_Stats"
- Name the NHL_Team_R_Stats as "NHL_Team_R_Stats" ** - Exclude indexes in the csv files **

```
write.csv(NHL_Game, file = "NHL_Game2.csv")
write.csv(NHL_Team_Stats, file = "NHL_Team_Stats.csv")
write.csv(NHL_Team_R_Stats, file = "NHL_Team_R_Stats.csv")
```