Analysis - Best NHL Ice Hockey Team

This analysis wants to find out the best team in NHL (because this is the league with most information: 87% of the data given) since 1918. "Best" is a hard analysis to make; there is no definite answer, but using some metrics we hope to get some interesting results.

First, the basic functions to be used.

```
library(stats)
library(base)
library(magrittr)
library(dplyr, warn.conflicts = FALSE)
library(ggplot2)
library(ggthemes)



teamName <- function(ID) {
  df <- read.csv("professional-hockey-database/Teams.csv", colClasses = character());
  return (df %>%
          filter(tmID==ID) %>%
          select(name) %>%
          apply(1,paste,collapse=" ") %>% head(1));
}
```

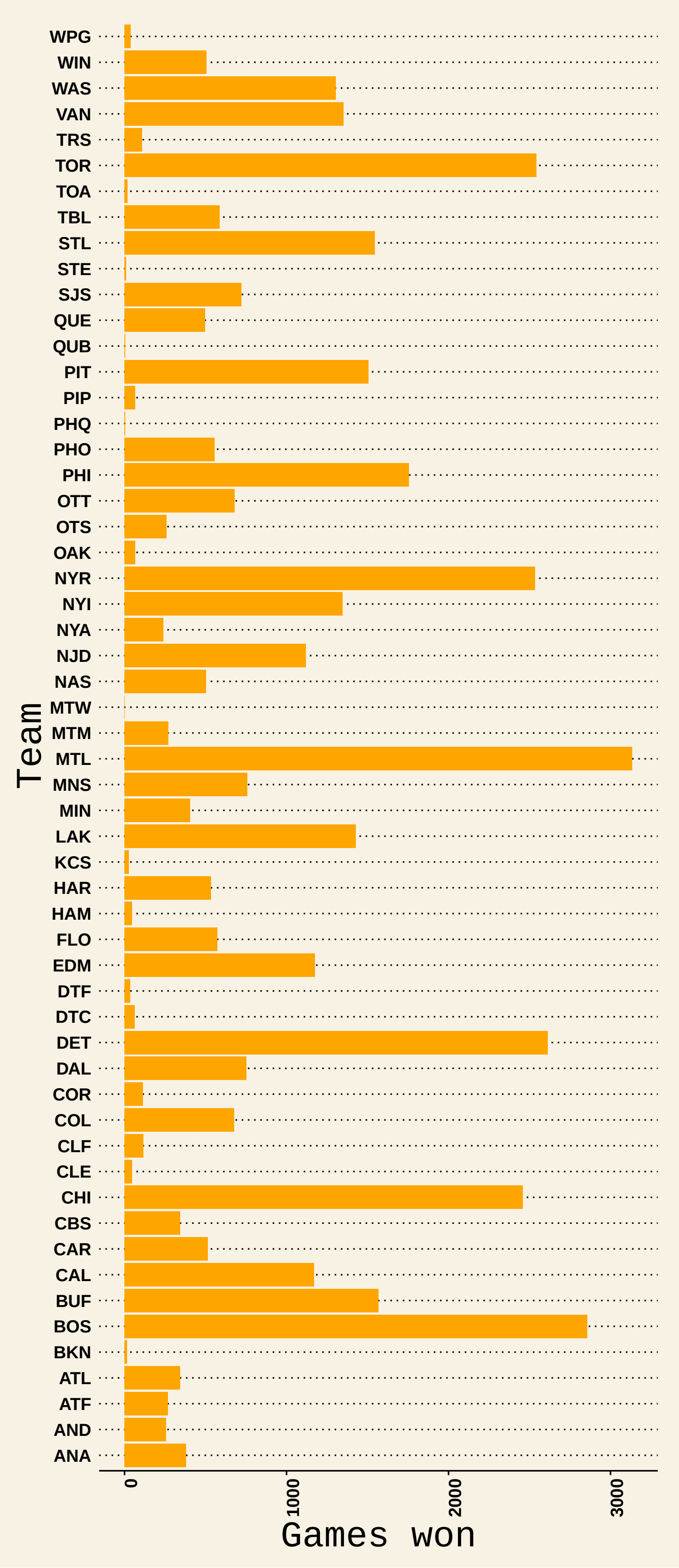
Here we load the dataframe to be used and some values that will be constants through the analysis.

```
  teamsDf <- read.csv("professional-hockey-database/Teams.csv");
  teamsDf <- teamsDf %>%
    filter(lgID=="NHL") %>%
    group_by(tmID) %>%
    summarise(W = sum(W), L = sum(L), G=sum(G), GF=sum(GF), GA=sum(GA))

  count <- nrow(teamsDf)
  subsetSize <- count*0.25
```

Best team overall

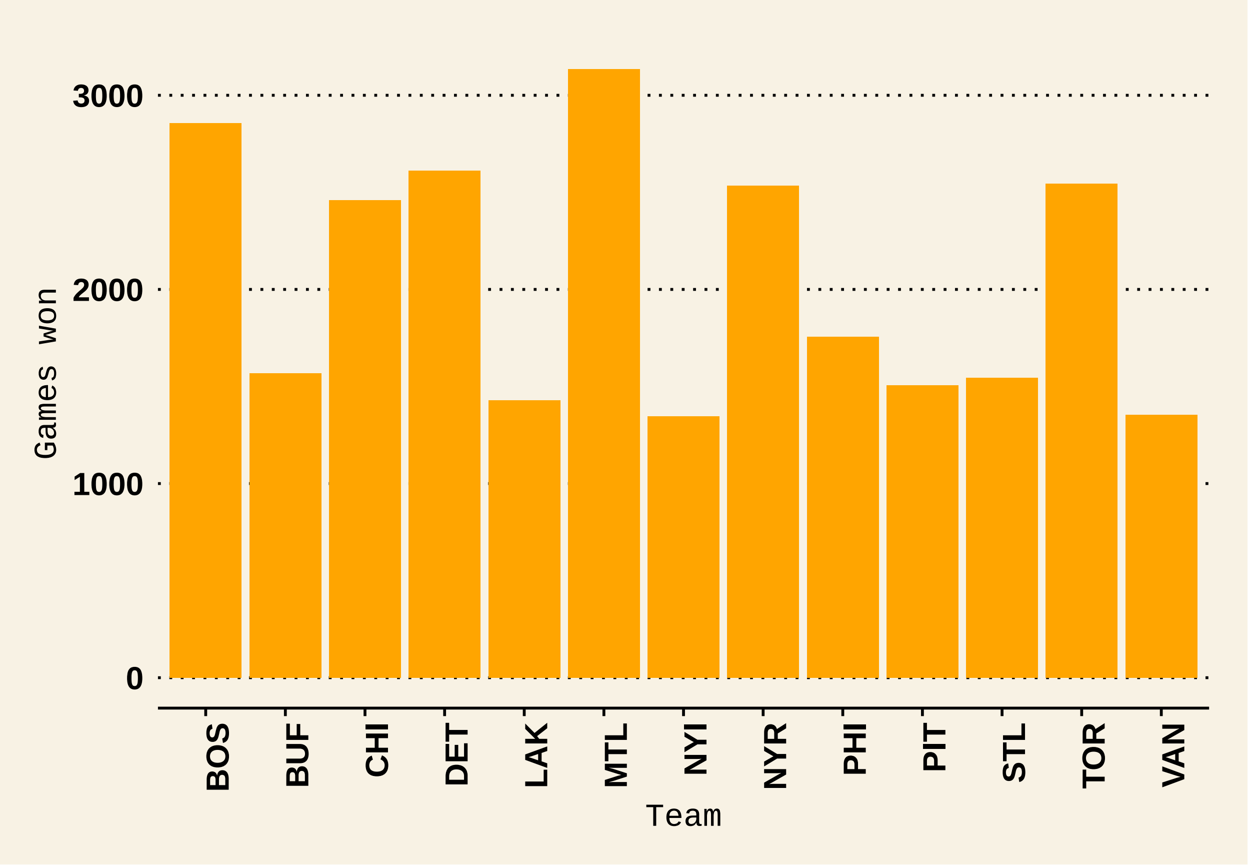
The first and easier analysis to do is check the team who won the most games.

```
  teamsDf %>%
    ggplot(aes(tmID, W)) +
    geom_bar(stat="identity", position="dodge", fill="orange") +
    theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
    coord_flip() +
    theme(axis.title=element_text(size=25)) +
    xlab("Team") +
    ylab("Games won")
```


WPG
WIN
WAS
VAN
TRS
TOR
TOA
TBL
STL
STE
SJS
QUE
QUB
PIT
PIP
PHQ
PHO
PHI
OTT
OTS
OAK
NYR
NYI
NYA
NJD
NAS
MTW
MTM
MTL
MNS
MIN
LAK
KCS
HAR
HAM
FLO
EDM
DTF
DTC
DET
DAL
COR
COL
CLF
CLE
CHI
CBS
CAR
CAL
BUF
BOS
BKN
ATL
ATF
AND
ANA
Team
0
1000
2000
3000
Games won

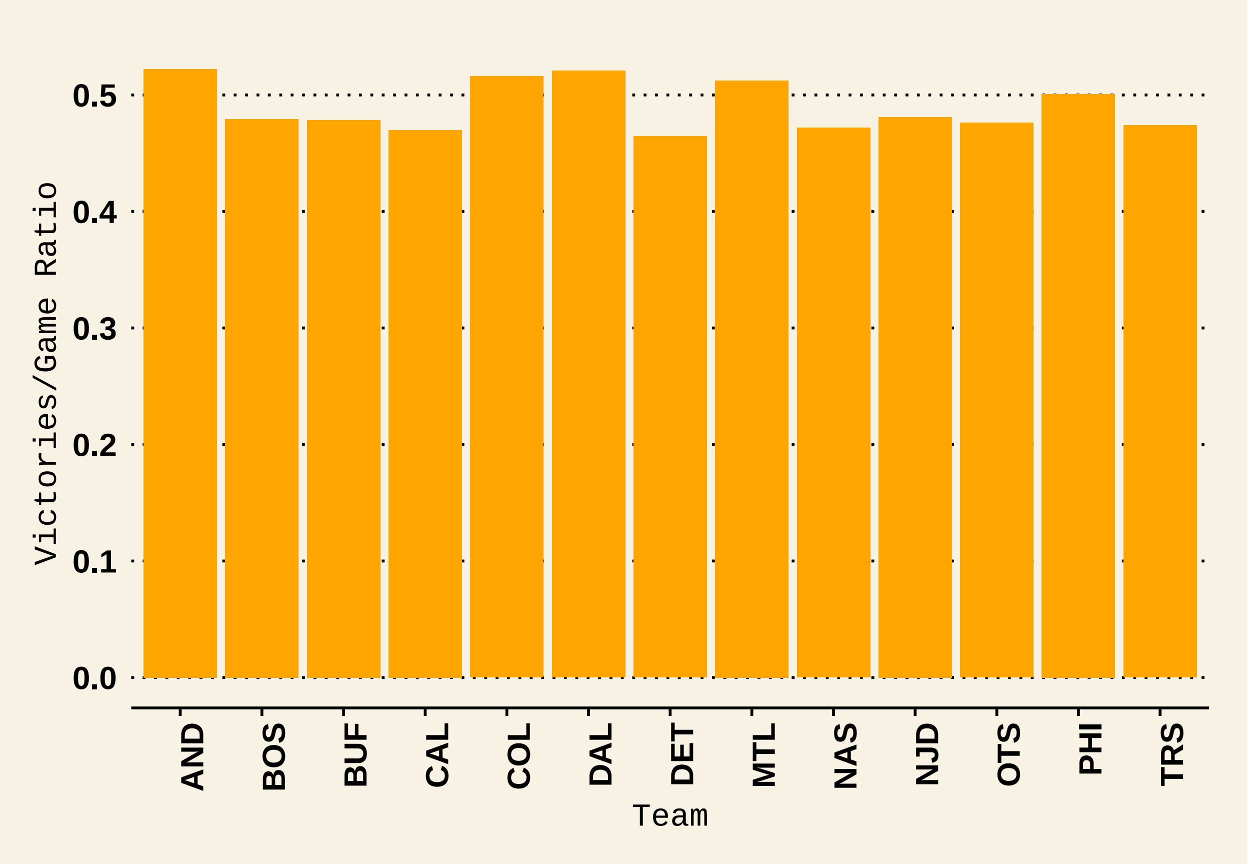
Plotting all the teams makes the graph too big and hard to read, so we might just get around 25% of the teams who won the most.

```
mostWins <- teamsDf %>%
              arrange(-W) %>%
              tibble::rowid_to_column("rowNumber") %>%
              filter(rowNumber<subsetSize)

mostWins %>%
  ggplot(aes(tmID, W)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=12)) +
  xlab("Team") +
  ylab("Games won")
```



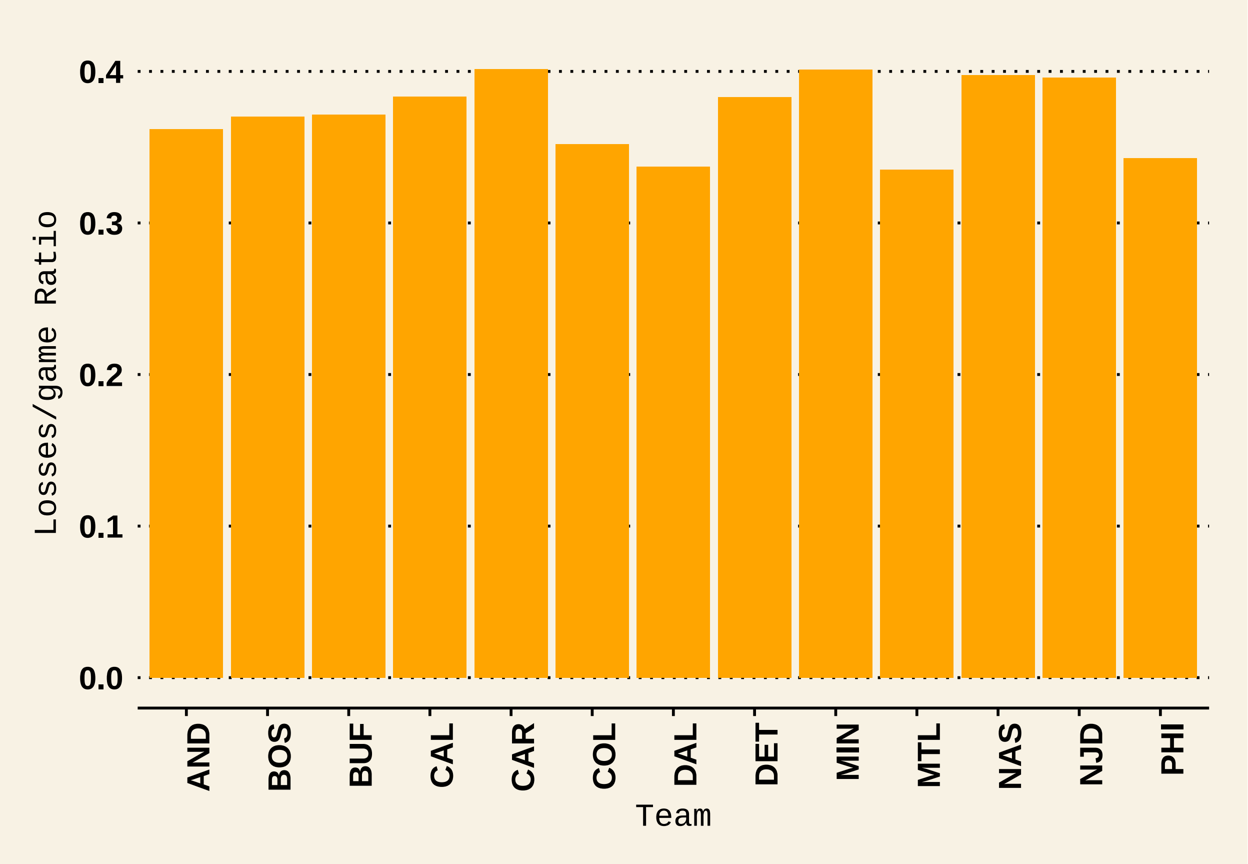
Having the teams with most wins/losses is cool with our objective, but maybe it's a bit unfair. It seems to be a good idea to get the ratio between victories and losses in order to estimate the better team.

```
bestWinRatio <- teamsDf %>%
  mutate(WinRatio = W/G) %>%
  arrange(-WinRatio) %>%
  tibble::rowid_to_column("rowNumber") %>%
  filter(rowNumber<subsetSize)

bestWinRatio %>%
  ggplot(aes(tmID, WinRatio)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=12)) +
  xlab("Team") +
  ylab("Victories/Game Ratio")
```



If we consider that the only negative result a team might have is a loss, we could also get the losses ratio. In this case, we are taking ties as results as good as victories, but it's still an interesting metric.

```
bestLossRatio <- teamsDf %>%
  mutate(LossRatio = L/G) %>%
  arrange(LossRatio) %>%
  tibble::rowid_to_column("rowNumber") %>%
  filter(rowNumber<subsetSize)

bestLossRatio %>%
  ggplot(aes(tmID, LossRatio)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=12)) +
  ylab("Losses/game Ratio") +
  xlab("Team")
```



Now we have three tables with meaningful results: 25% of the teams that won the most, that won the most considering the amount of games and that loss the least considering the amount of games. We can combine them and find which teams are in all of them.

```
mostWinners <- suppressMessages(mostWins %>% select(tmID) %>%
              semi_join(bestWinRatio %>% select(tmID)) %>%
              semi_join(bestLossRatio %>% select(tmID)))
mostWinners <- as.list(mostWinners)
mostWinners <- lapply(mostWinners[[1]], teamName) %>% unlist() %>% as.data.frame()
mostWinners
```

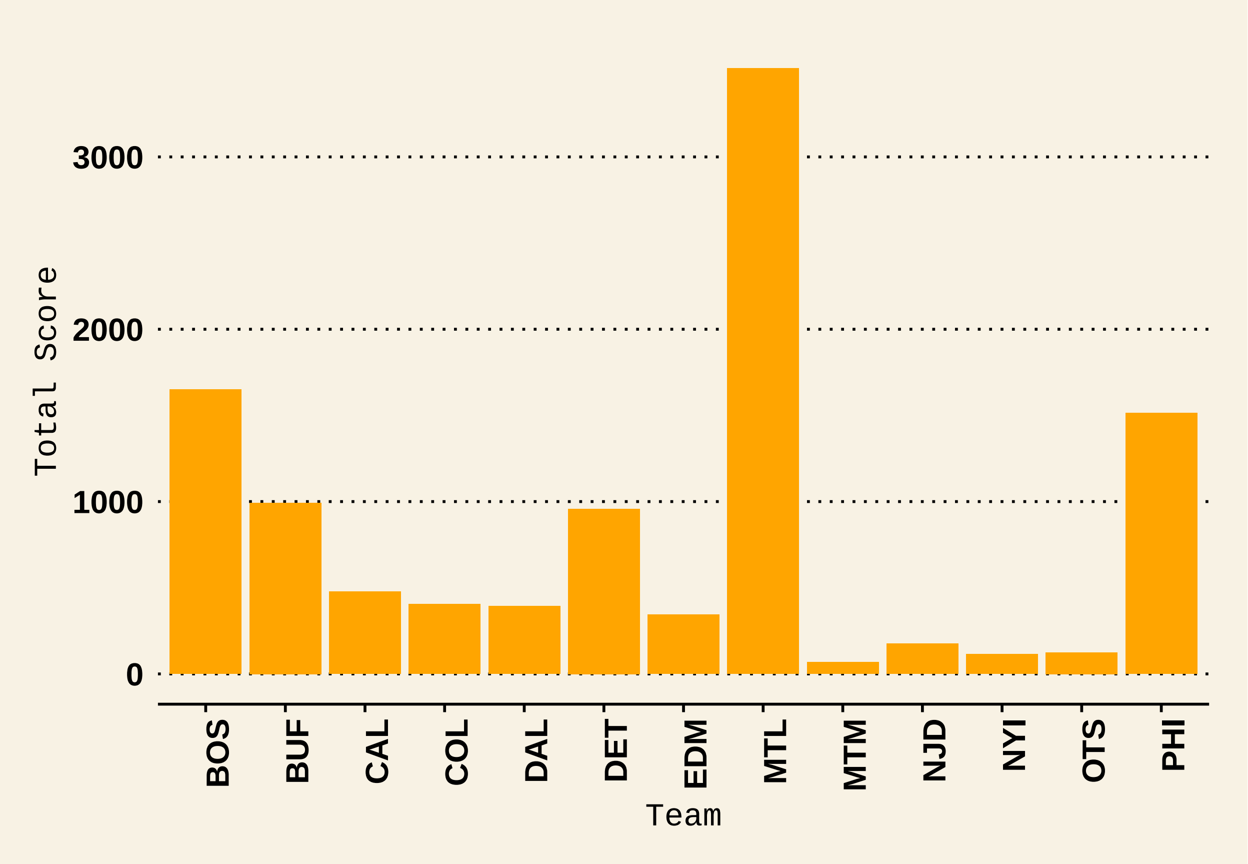
```
##                       .
## 1  Montreal Canadiens
## 2        Boston Bruins
## 3   Detroit Red Wings
## 4 Philadelphia Flyers
## 5      Buffalo Sabres
```

There we have it. Considering these three parameters, those five teams had achieved the best score based on their victories and losses.

Best score

Another metric we might use is the amount of goals scored. More than that, we probably should subtract the amount of goals taken, so we can have the best score overall. Again, we will take only the best 25%.

```
bestScore <- teamsDf %>%
  mutate(TotalScore = GF-GA) %>%
  arrange(-TotalScore) %>%
  tibble::rowid_to_column("rowNumber") %>%
  filter(rowNumber<subsetSize)

bestScore %>%
  ggplot(aes(tmID, TotalScore)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=12)) +
  ylab("Total Score") +
  xlab("Team")
```



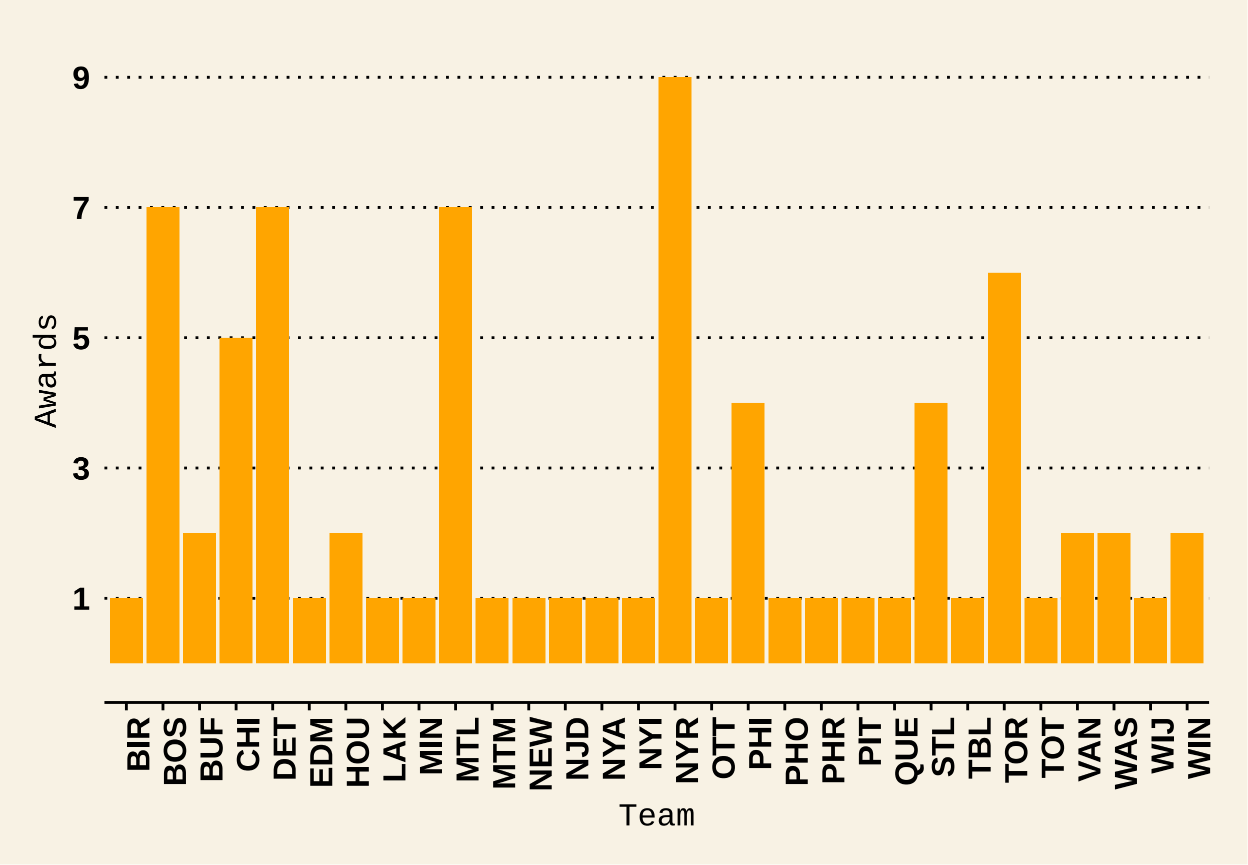
Well, this graph says a lot. The Montreal Canadiens team, which was already on our previous top 5, has more than twice the score of the second place, Boston Bruins. It may look interesting, but actually the entire top 5 from the previous session is here, so we might have not gained much more information.

Awarded coaches

What is a team without its coach? Maybe we might find some nice results checking the amount of coach awards won by every team.

```
coachesDf <- read.csv("professional-hockey-database/Coaches.csv") %>%
  select(coachID, tmID, year)
```

```
coachesAwardedDf <- read.csv("professional-hockey-database/AwardsCoaches.csv") %>%
  select(coachID, year)

coaches <- merge(coachesDf, coachesAwardedDf, by=c("year", "coachID"))

coaches %>%
  group_by(tmID) %>%
  mutate(Awards = n()) %>%
  ggplot(aes(tmID, Awards)) +
  scale_y_discrete("Awards", limits=c(1, 3, 5, 7, 9)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=12)) +
  ylab("Awarded coaches") +
  xlab("Team")
```

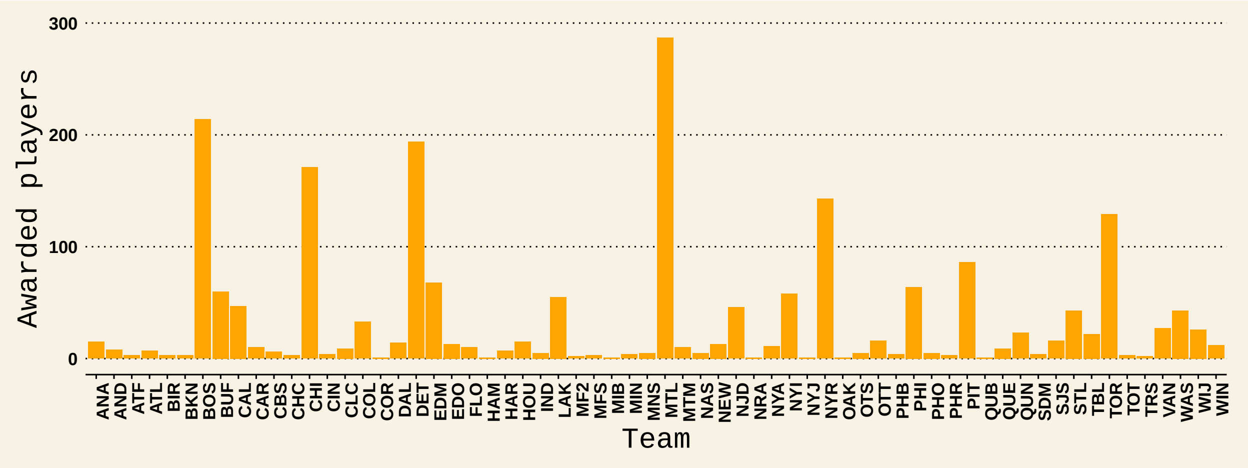


Well, finally a new possibility. The New York Rangers won 2 coach awards more than the invincible Montreal Canadiens. However, it's the only new team on the competition, if we keep analysing only the top 5. From now on, Philadelphia Flyers is out of the game.

Awarded players

If we checked for the number of awards for the coaches, it's unfair to not check the players' as well. So, here it is:

```r
playersDf <- read.csv("professional-hockey-database/Scoring.csv") %>%
  select(playerID, tmID, year)
playersAwardedDf <- read.csv("professional-hockey-database/AwardsPlayers.csv") %>%
  select(playerID, year)

players <- merge(playersDf, playersAwardedDf, by=c("year", "playerID"))

players <- players %>%
  group_by(tmID) %>%
  mutate(Awards = n())

players %>%
  ggplot(aes(tmID, Awards)) +
  geom_bar(stat="identity", position="dodge", fill="orange") +
  theme_wsj() + theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme(axis.title=element_text(size=20)) +
  ylab("Awarded players") +
  xlab("Team")
```



Well, again we have the Canadiens at the top.

Conclusion

We recognize that, in this work, there wasn't much of statistics exactly. There is no confidence interval or something like that, but it seems like the approach used in this scenario was the best possible.

In all of the metrics we used, we had basically the same 4~5 teams. However, it is clear that Montreal Canadiens' results were special. The team has a great difference in comparison with the other considered the bests, so they probably should receive the award of best team from the NHL according to this study.