Exploratory Analysis of NHL Team Data Extracted from the Web

SurvMeth 727 Final Project

Chia Wen Cheng

The GitHub link for the repository containing these files is linked here.

Questions of Interest

The National Hockey League (NHL) stands as a prominent North American professional ice hockey league, comprising 32 teams—25 in the United States and 7 in Canada. While not holding the title of the most financially rewarding professional sports league in North America, the NHL recorded a substantial total league revenue of 5.93 billion U.S. dollars in the 2021/2022 season. As an ardent ice hockey enthusiast, my curiosity is piqued: Are top-performing teams duly compensated in correlation with their on-ice achievements? This leads to my high-level research question: **Do Top Performing Teams Earn Top Dollars in the NHL?** To unravel this overarching inquiry, I've crafted specific breakdown questions, serving as informative indicators, outlined below.

1. What is the trajectory of average salaries for each NHL team from the 2010 season to the 2023 season?
2. Over the years spanning from 2010 to the present, which NHL teams provide the highest player salaries?
3. What is the annual performance of each NHL team from the 2010 season to the 2023 season?
    1. maximum and minimum shooting percentages by team
    2. average points per game per player by team
    3. average saving percentages by team
4. What is the mean age at signing (the latest contract) for each drafted year?

Methodologies

1. For data collection, I will use web scraping technique with the `rvest` package in R.
2. For data management in exploratory analysis, I will use SQL queries by `sqldf` and `dplyr` in R.
3. For data visualization, I will use `ggplot2` in R to generate line plots and box plots as appropriate.
4. I will be utilizing the `shiny` in R to create an interactive graphic showcasing a series of player salaries organized by teams.
5. I will also use other R packages as needed to manipulate data.

Description of Analysis Plan

With the substantial dataset, I plan to conduct exploratory analysis, examining yearly team average salaries, identifying teams with the highest payments, and aggregating team seasonal performance across multiple years using SQL queries in R. Additionally, I aim to visualize the data through line charts, where the x-axis represents years and the y-axis depicts salaries and aggregated points. I will also use box plots to show the collective maximum and minimum shooting percentages by teams. Given the sequential and categorical nature of the data, employing shiny R might be ideal for illustrating changes across teams and players.

Web scraping raw data

```
# scrabing data from the web from the 2010 season to the 2023 season
salary_df <- data.frame()
page <- c(1:31)
year <- c(2011:2024)
for (x in year){
  for(i in page) {
    salary_url <- paste0("https://www.capfriendly.com/browse/active/",
                         x,
                         "?stats-season=",
                         x,
                         "&age-calculation-date=today&display=draft,signing-age&hide=claus
                         i)
    # check if the url is valid
    #if (url.exists(salary_url)) {
```

```
    #  salary_url <- salary_url
    #}
    url_html <- read_html(salary_url)
    url_nodes <- html_nodes(url_html, "table")
    salary_table <- html_table(url_nodes)[[1]]
    salary <- salary_url %>%
      read_html() %>%
      html_nodes("table") %>%
      html_table()
    salary_df_1 <- salary[[1]] %>%
      mutate(year = x - 1)
    salary_df <- rbind.fill(salary_df, salary_df_1)
  }
}
head(salary_df)
```

```
            PLAYER TEAM                    DRAFTED GP  G  A  P P/GP +/- Sh  Sh%
1 1. Alex Ovechkin  WSH  1 - Round 1 - 2004 (WSH) 18  5  7 12 0.67   1 70 0.07
2 2. Evgeni Malkin  PIT  2 - Round 1 - 2004 (PIT) 21 10  9 19 0.90  -2 53 0.19
3 3. Sidney Crosby  PIT  1 - Round 1 - 2005 (PIT) 21 13 11 24 1.14  11 80 0.16
4    4. Eric Staal    -  2 - Round 1 - 2003 (CAR)  0  0  0  0 0.00   0  0 0.00
5 5. Brad Richards    - 64 - Round 3 - 1998 (TBL)  0  0  0  0 0.00   0  0 0.00
6     6. Rick Nash    -  1 - Round 1 - 2002 (CBJ)  0  0  0  0 0.00   0  0 0.00
    TOI W L SO GAA Sv% SIGNING AGE     SALARY year
1 20:24 - -  -   -   -          22 $9,000,000 2010
2 18:33 - -  -   -   -          21 $9,000,000 2010
3 19:04 - -  -   -   -          19 $9,000,000 2010
4       - -  -   -   -          23 $7,500,000 2010
5       - -  -   -   -          26 $7,800,000 2010
6       - -  -   -   -          26 $7,500,000 2010
```

I initiated the data collection process by scraping table data from this website using the "rvest" package. The provided URL was accessible solely for extracting data of the 2023 season on the first page of the table, which contained 50 rows. However, by modifying the year and page parameters within the URL, I obtained data spanning from the 2010 season to the 2023 season. I converted all HTML tables into data frames and combined them into one data frame. This data encompassesed demographic details, seasonal performance statistics, and yearly salaries for each player.

The raw dataset comprised 20 columns and 20,019 observations. Repeated rows of the same player were seen because they may be actively playing in the NHL for more than one season.

The dataset features various types of information, such as categorical variables including "player name," "drafted detail," "team," "average time on ice," and "year," alongside numerical variables like "number of games played," "number of goals," "number of assists," "points scored," "points earned per game," "number of shots on goal," "amount of salary," etc. In my analysis, I included undrafted players and players playing 0 game in that game season as long as they were recorded to be paid greater than $0.

Data cleaning for analysis

```
# change column names
colnames(salary_df) <- c("player_name", "team", "drafted",
                         "games_played", "goals", "assists",
                         "points", "points_per_game",
                         "plus/minus", "shots_on_goal",
                         "shooting_percentage",
                         "average_time_on_ice", "wins",
                         "loses", "shootouts",
                         "goals against average",
                         "saving_percentage", "signing_age",
                         "salary", "year")

# data management
# cut "$" and "," in the salary column,
# convert the salary values to numeric,
# drop rows that have 0 for their salaries,
# and split the player_name column so that
# the numbers leading the names do not matter
salary_df_2 <- salary_df %>%
  mutate(salary = str_replace_all(salary, "\\$", "")) %>%
  mutate(salary = str_replace_all(salary, ",", "")) %>%
  filter(as.numeric(unlist(salary)) != 0) %>%
  separate(player_name,
           into = c("number", "first_name",
                    "last_name", "more_name",
                    "more_name_2"),
           sep = " ",
           remove = TRUE) %>%
  mutate(player_name = paste(ifelse(!is.na(first_name),
                                    first_name, ""),
                             ifelse(!is.na(last_name),
                                    last_name, ""),
```

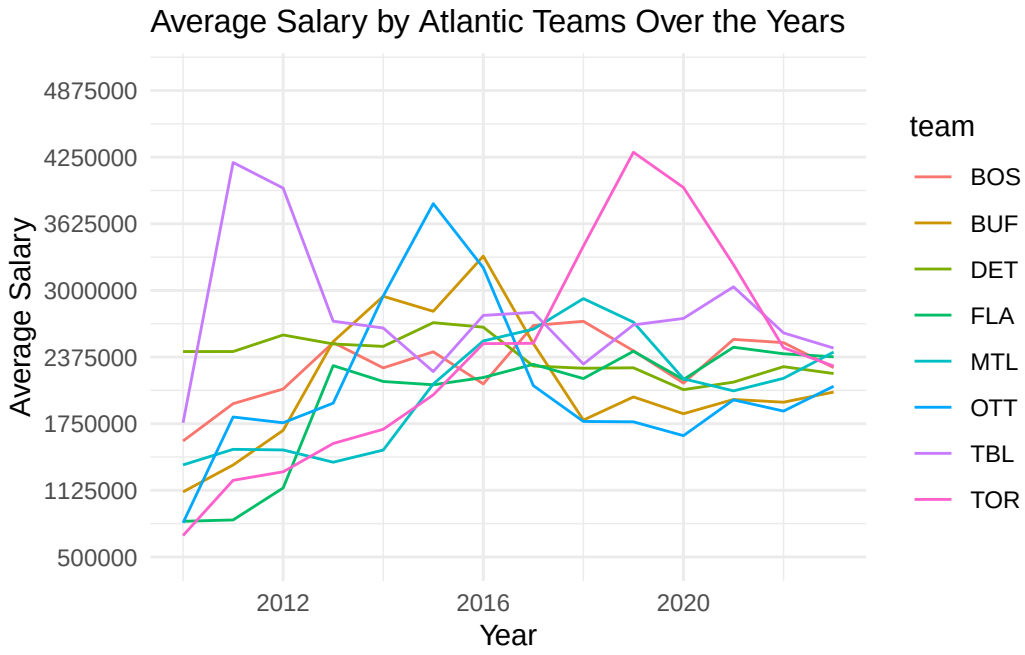
```
                              ifelse(!is.na(more_name),
                                     more_name, ""),
                              ifelse(!is.na(more_name_2),
                                     more_name_2, ""),
                              sep = " ")) %>%
  relocate(player_name, .before = team) %>%
  mutate(drafted_year = substr(drafted,
                               nchar(drafted) - 9,
                               nchar(drafted) - 6)) %>%
  relocate(drafted_year, .after = drafted)
salary_df_2 <- salary_df_2[, -c(1:5, 8)]
salary_df_2$salary <- as.numeric(salary_df_2$salary)
```

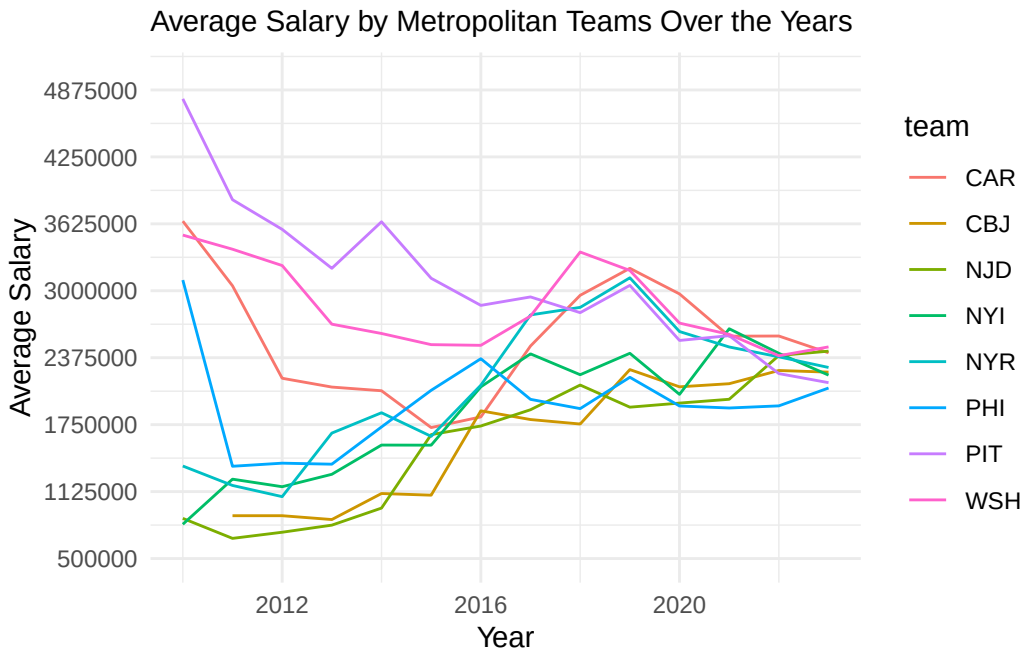
Upon data cleaning by filtering out rows that had salary at the amount of 0, there were 18,989 observations left. The `year` column I mutated represented the starting year of the season, i.e., 2010 indicated the 2010 season, following the fact that one game season crossed two years.

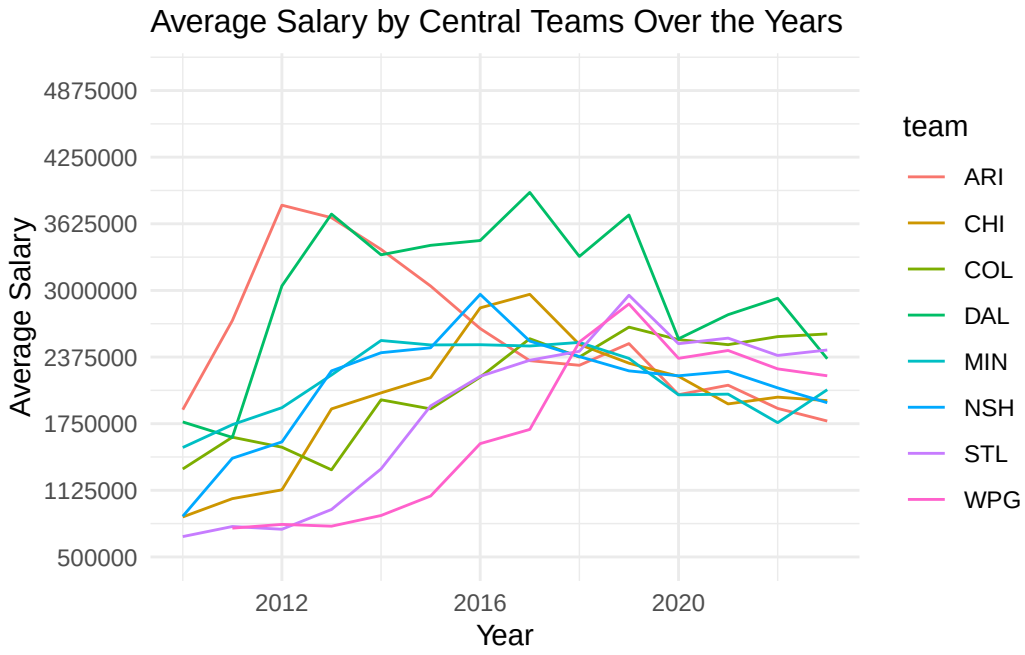
Exploratory Analysis & Data Visualization

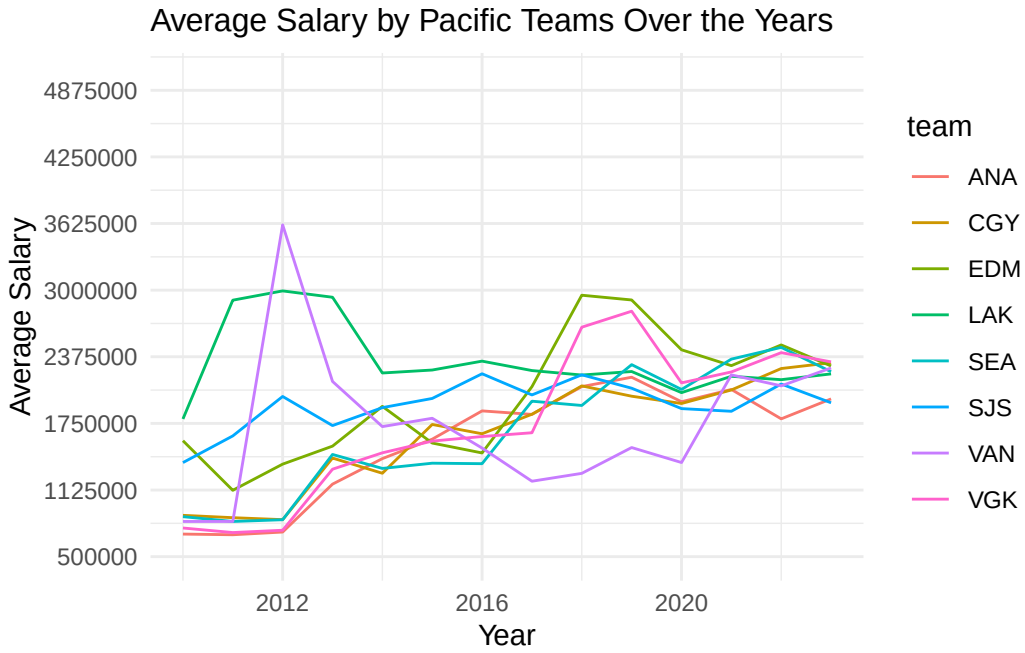
What is the trajectory of average salaries for each NHL team from the 2010 season to the 2023 season?

```
  average_salary team year
1       711875.0  ANA 2010
2       705500.0  ANA 2011
3       731071.4  ANA 2012
4      1180250.0  ANA 2013
5      1419791.7  ANA 2014
6      1602187.5  ANA 2015
```


Average Salary by Atlantic Teams Over the Years
4875000
4250000
3625000
3000000
2375000
1750000
1125000
500000
Average Salary
2012
2016
2020
Year
team
BOS
BUF
DET
FLA
MTL
OTT
TBL
TOR


Average Salary by Metropolitan Teams Over the Years
4875000
4250000
3625000
3000000
2375000
1750000
1125000
500000
Average Salary
2012
2016
2020
Year
team
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH





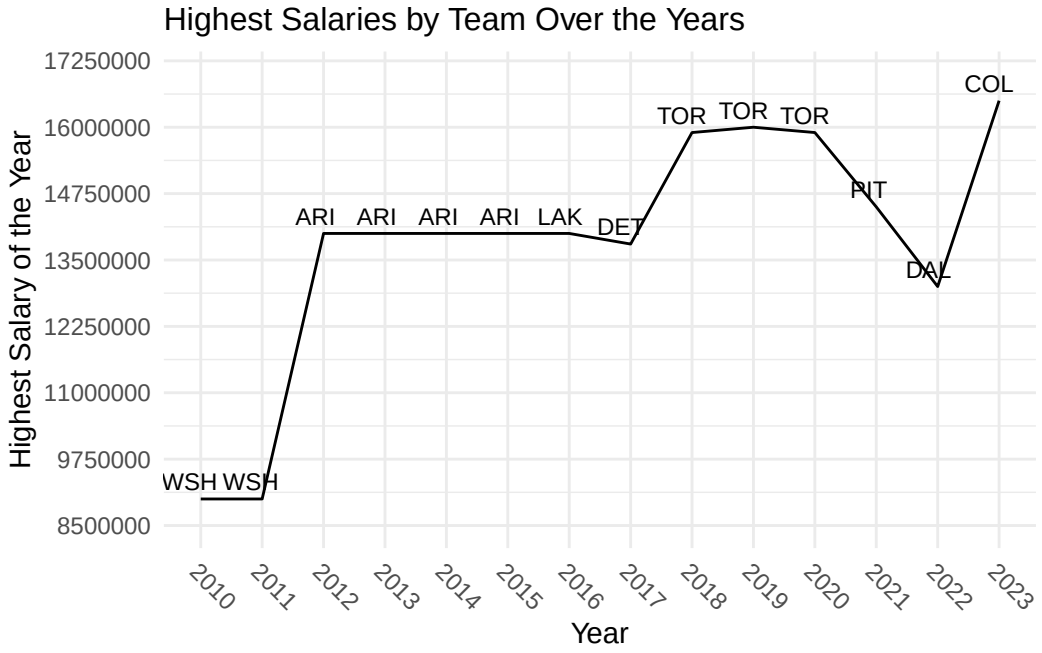
Due to the large number of teams, displaying all teams in the same chart would result in confusing outcomes. As a result, I have divided the teams into four charts based on their division affiliations.

Pittsburgh Penguins, representing the Metropolitan Division, held the highest average salary

over the years of data collection, peaking at approximately $4,875,000. Despite maintaining the highest average salary in the 2014 season among Metropolitan teams, there was a significant drop to $3,625,000, a figure shared by the Carolina Hurricanes in the 2010 season. Teams from the Atlantic Division, namely Toronto Maple Leafs and Tampa Bay Lightning, secured the second-highest average salaries, averaging around $4,250,000. The Ottawa Senators claimed the third-highest salary overall, ranging between $3,625,000 and $3,937,500. The Pacific Division consistently reported the lowest average salaries among all NHL teams. Interestingly, starting from 2019, the average salaries of each team began to narrow the gaps between them. However, it's worth noting that the NHL had implemented a salary cap, restricting the maximum salaries players could receive, long before the period covered by the data collection.

Over the years spanning from the 2010 season to the current season, which NHL teams provide the highest player salaries?

```
  team year  salary
1  WSH 2010 9.0e+06
2  WSH 2011 9.0e+06
3  ARI 2012 1.4e+07
4  ARI 2013 1.4e+07
5  ARI 2014 1.4e+07
6  ARI 2015 1.4e+07
```



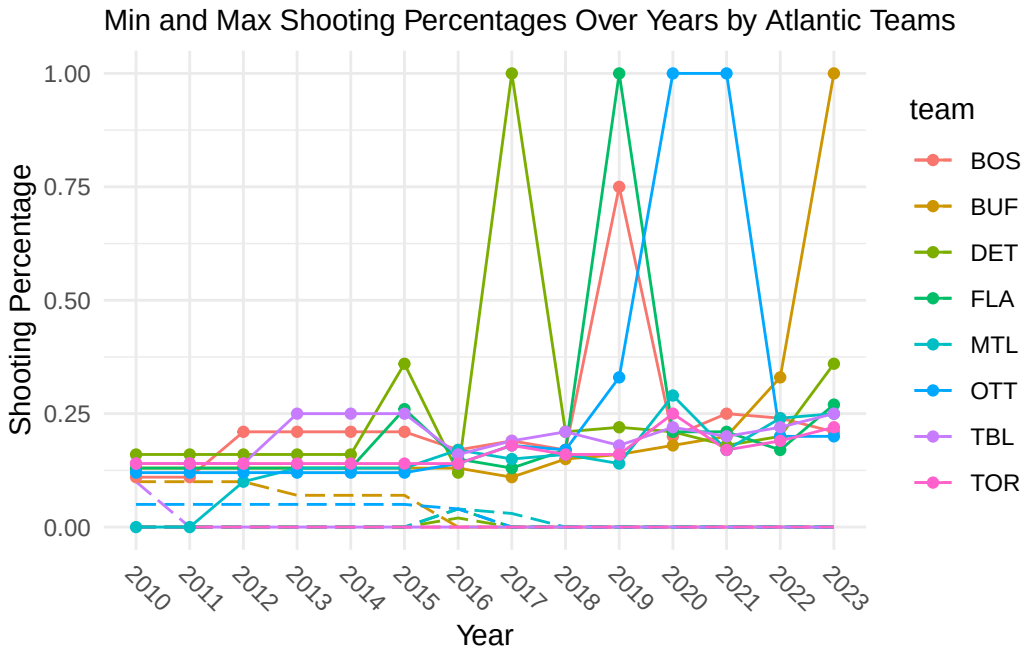
Players without team affiliations were excluded because I would like "team" to be the unit of the following analysis.

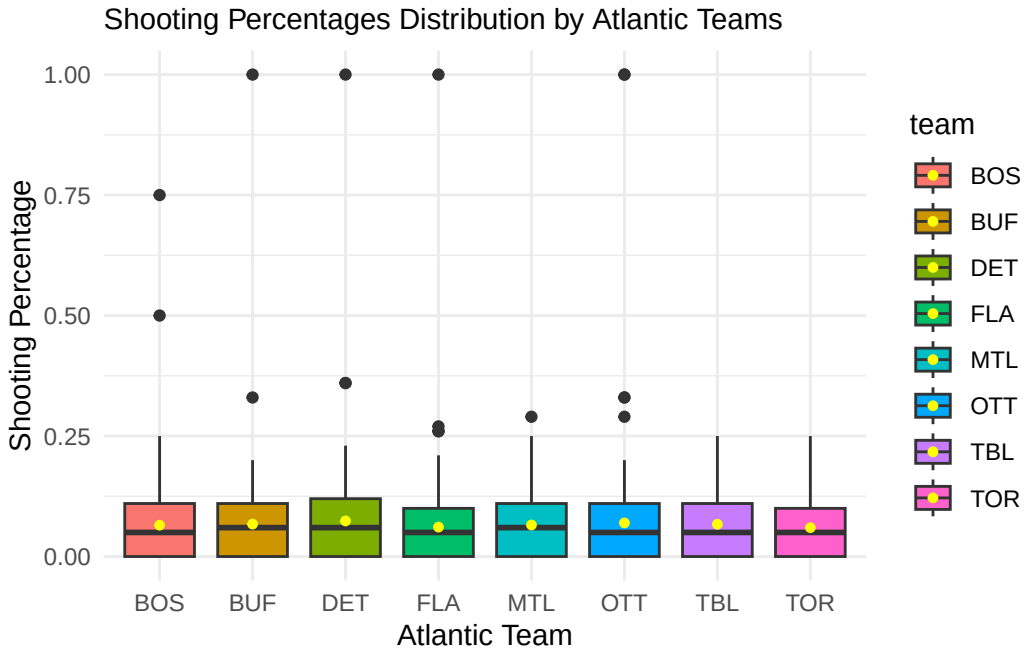
During the 2010 and 2011 seasons, the Washington Capitals held the record for paying the highest salary to at least one of their players, equivalent to approximately $13,336,025.54 in today's value. From the 2012 season to the 2015 season, the Arizona Coyotes maintained this trend, paying around $18,335,616.75 in today's value to at least one player. Following closely were the Los Angeles Kings, Detroit Red Wings, Toronto Maple Leafs, Pittsburgh Penguins, Dallas Stars, and Colorado Avalanche in paying the highest salaries. Two declines were observed in the 2017 and 2022 seasons, and the reasons for these drops are not immediately discernible from the current dataset.

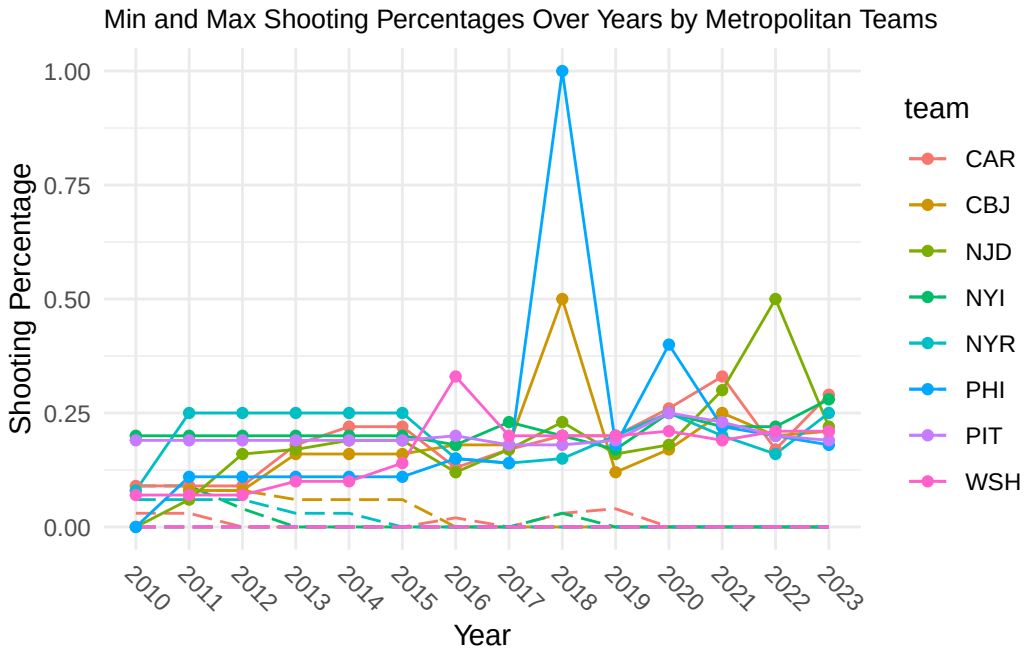
What is the annual performance of each NHL team from the 2010 season to the 2023 season?

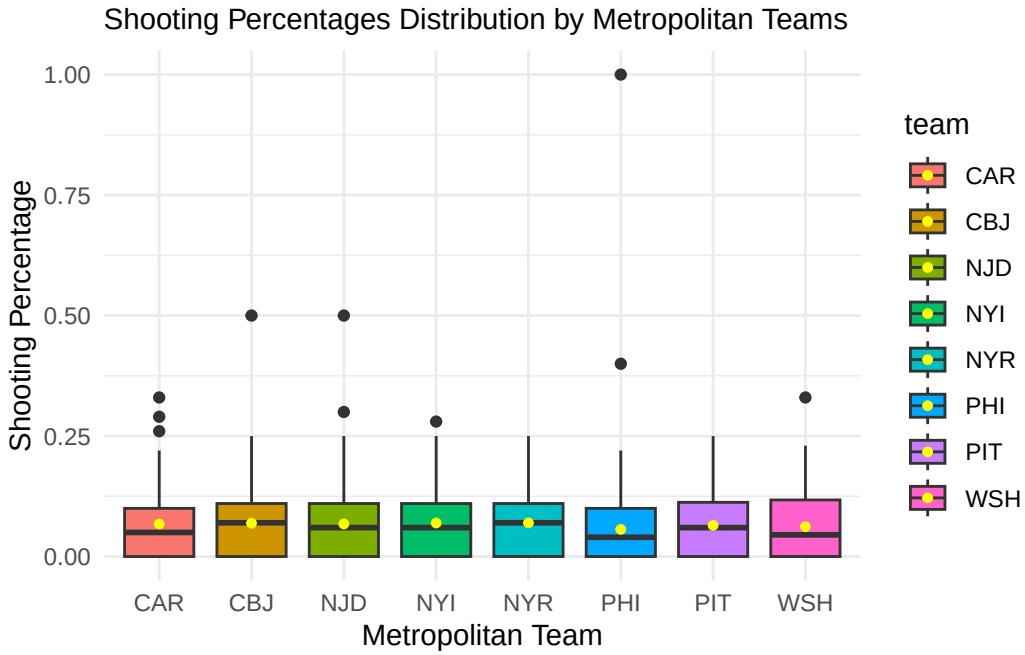
Maximum and Minimum Shooting Percentages by Team

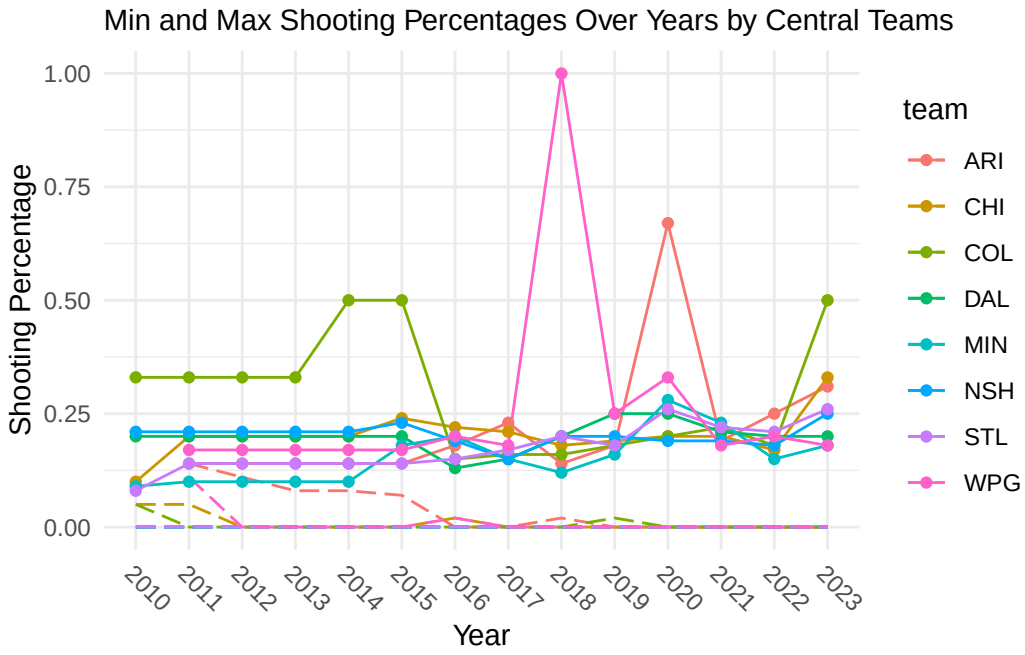
	max_shooting_percentage	min_shooting_percentage	year	team
1	0.15	0.06	2010	ANA
2	0.15	0.03	2011	ANA
3	0.16	0.03	2012	ANA
4	0.16	0.03	2013	ANA
5	0.16	0.00	2014	ANA
6	0.16	0.00	2015	ANA

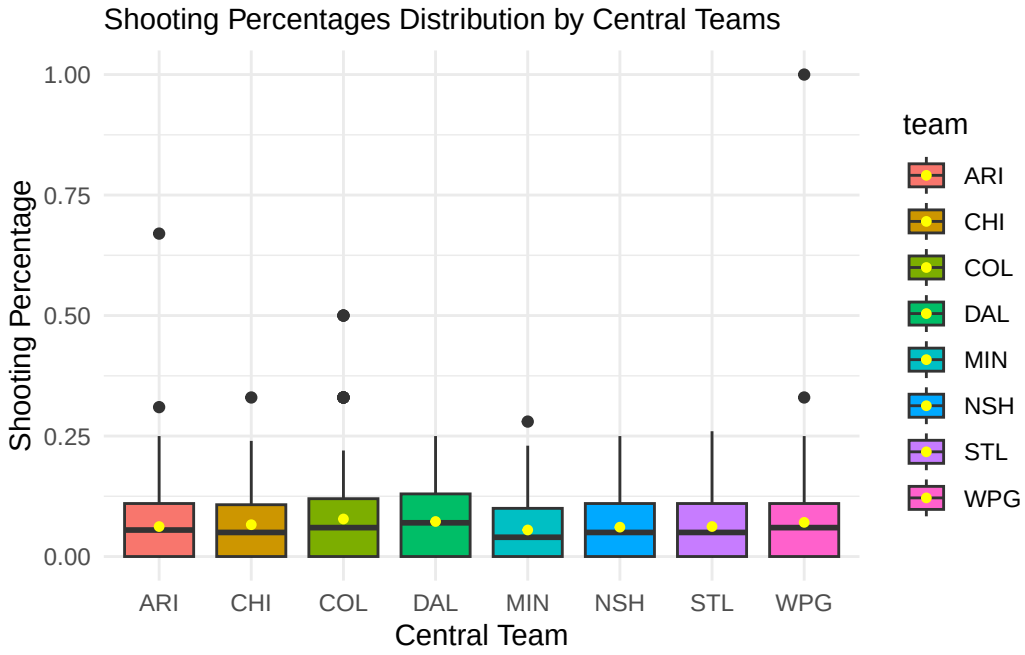

Min and Max Shooting Percentages Over Years by Atlantic Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
team
BOS
BUF
DET
FLA
MTL
OTT
TBL
TOR

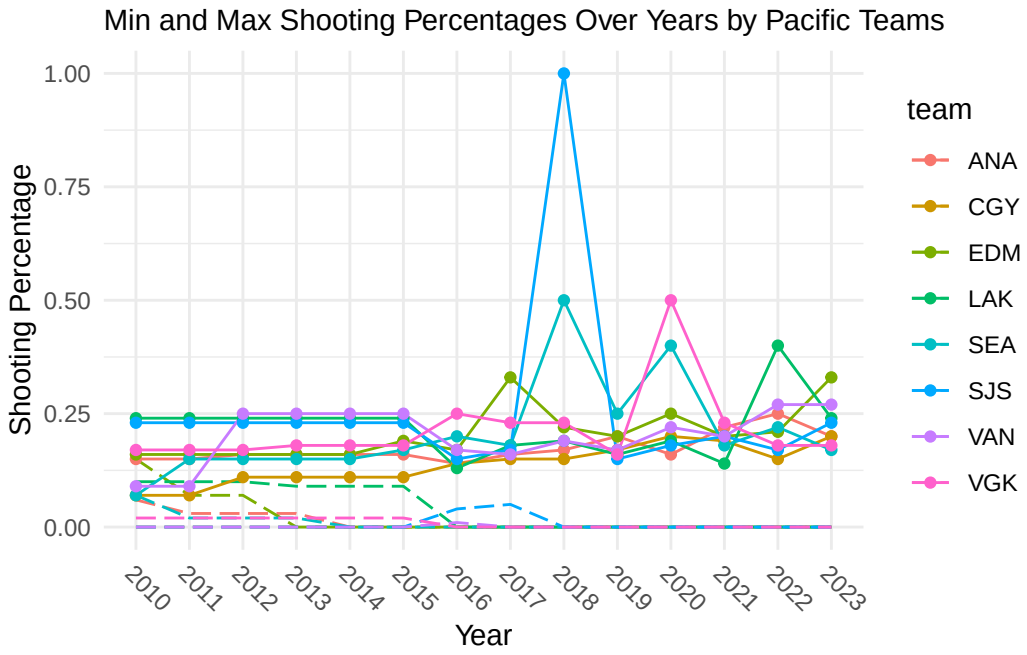

Shooting Percentages Distribution by Atlantic Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
BOS
BUF
DET
FLA
MTL
OTT
TBL
TOR
Atlantic Team
team
BOS
BUF
DET
FLA
MTL
OTT
TBL
TOR

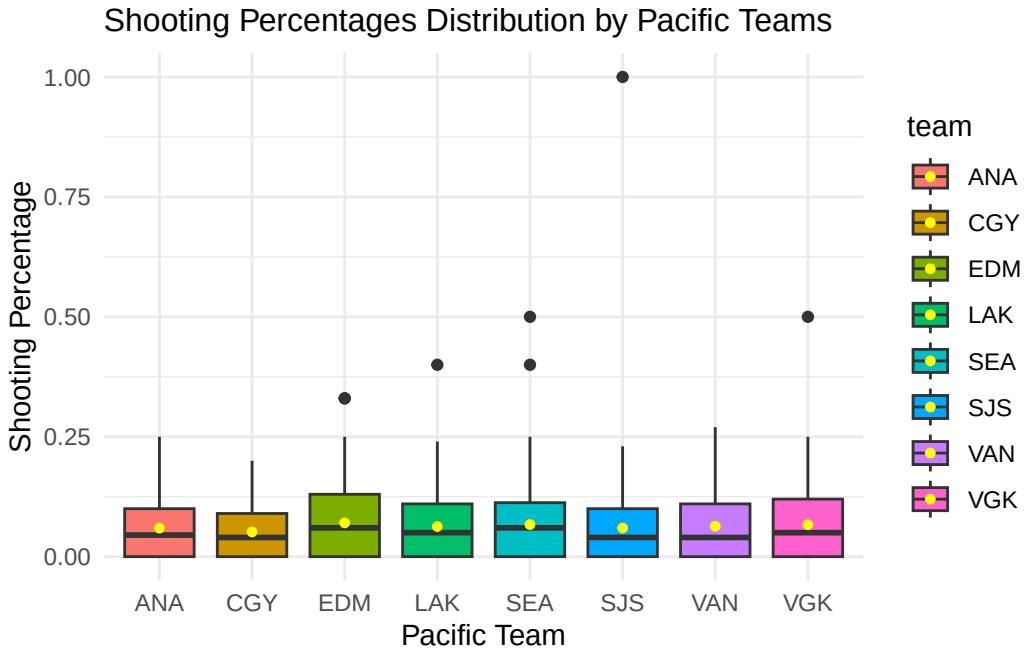

Min and Max Shooting Percentages Over Years by Metropolitan Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
team
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH


Shooting Percentages Distribution by Metropolitan Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH
Metropolitan Team
team
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH


Min and Max Shooting Percentages Over Years by Central Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
team
ARI
CHI
COL
DAL
MIN
NSH
STL
WPG


Shooting Percentages Distribution by Central Teams
Shooting Percentage
1.00
0.75
0.50
0.25
0.00
ARI
CHI
COL
DAL
MIN
NSH
STL
WPG
Central Team
team
ARI
CHI
COL
DAL
MIN
NSH
STL
WPG





A skater's shooting percentage is calculated by dividing the number of goals scored by the number of shots on goal. Therefore, a shooting percentage of 0.00 for a skater makes sense when no goals were scored, regardless of the number of shots on goal made. I excluded players that had no teams in the year, played no games, or were goaltenders (i.e., having "-" in their

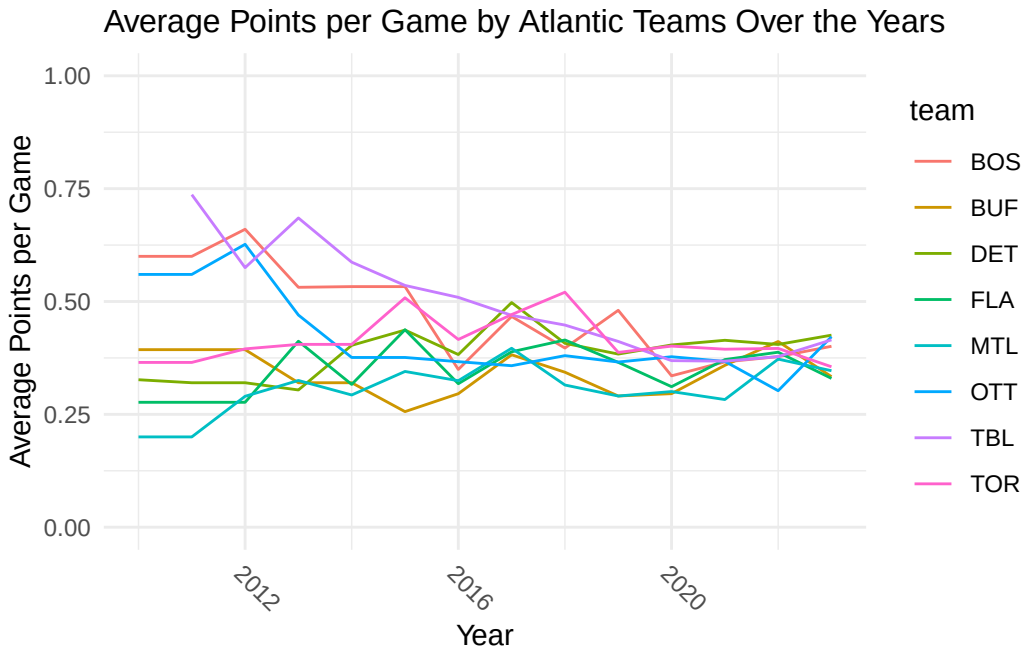
shooting percentage records) from the following analysis.

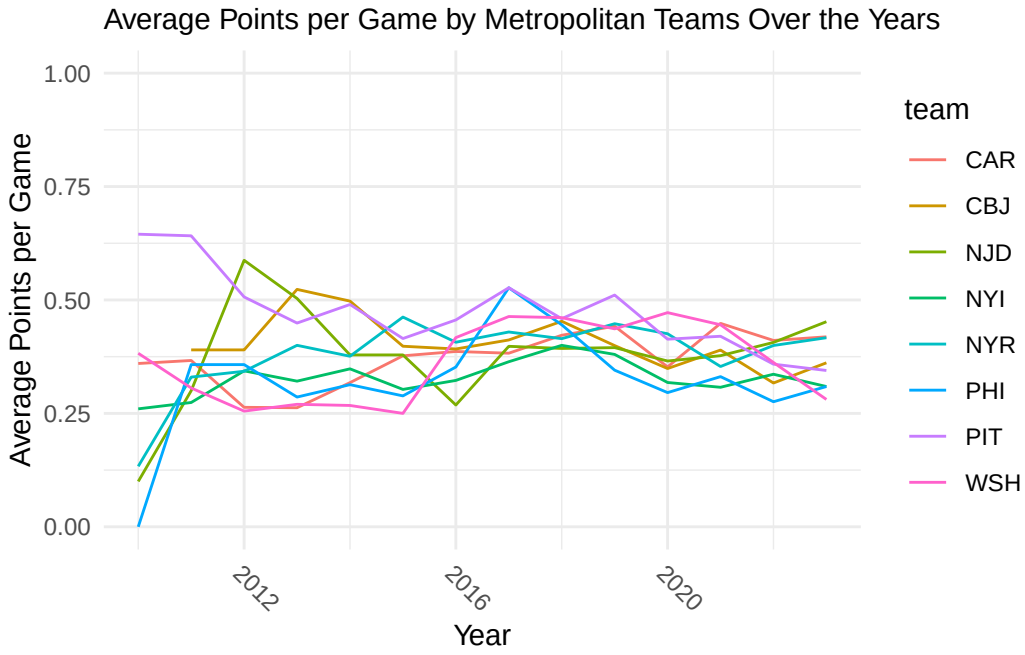
Analyzing the box plots, it appears that the Atlantic Division stands out as the most competitive, as half of its teams achieved an extreme shooting percentage of 1.00 collectively across the years. Furthermore, the Atlantic teams have the highest medians and means of shooting percentages across all teams in the league. In contrast, each of the other three divisions had only one team reaching a 1.00 shooting percentage across the years. Notably, all teams in the Pacific Division exhibited mean shooting percentages greater than their median shooting percentages, indicating positively skewed distributions for their shooting performance.

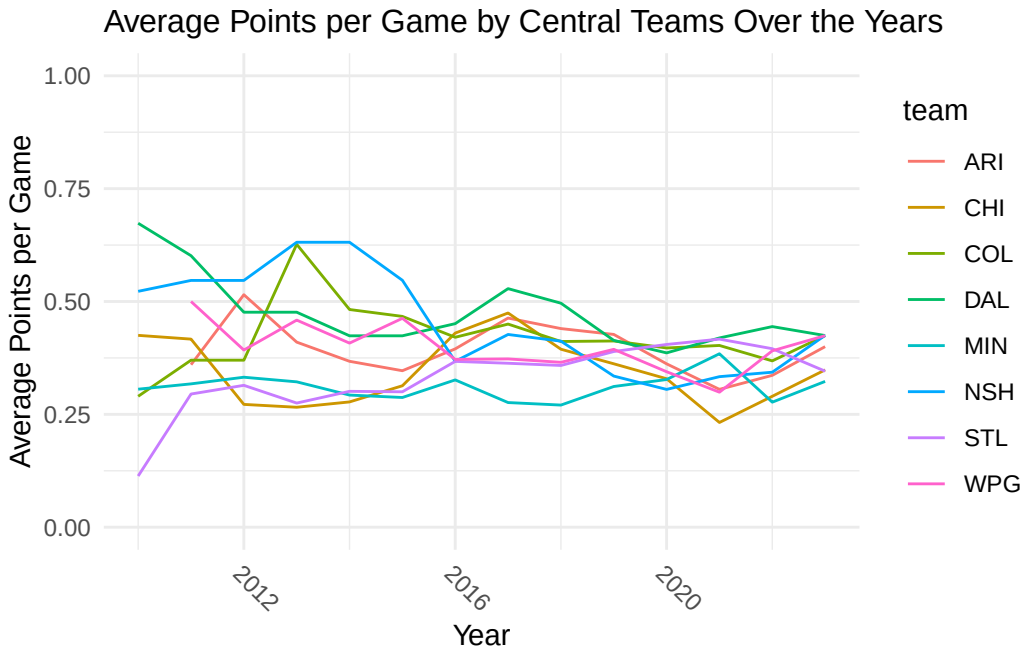
Long-dashed lines in the line charts presents the minimum shooting percentages. No significant patterns were observed from the line charts depicting each team's shooting performance over time. However, it could be worthwhile to examine the years and teams where the minimum shooting percentages exceeded 0.00, suggesting a higher average capability among skaters.

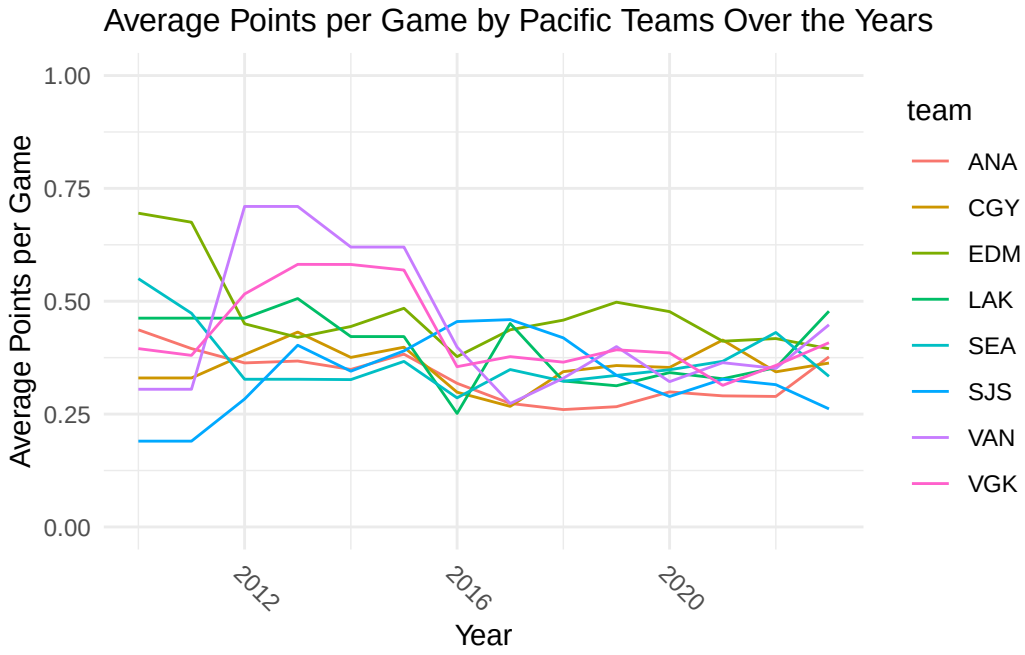
Average Points per Game per Player by Team

	average_points_per_game	team	year
1	0.4366667	ANA	2010
2	0.3950000	ANA	2011
3	0.3633333	ANA	2012
4	0.3675000	ANA	2013
5	0.3488889	ANA	2014
6	0.3827273	ANA	2015


Average Points per Game by Atlantic Teams Over the Years
Average Points per Game
1.00
0.75
0.50
0.25
0.00
2012
2016
2020
Year
team
BOS
BUF
DET
FLA
MTL
OTT
TBL
TOR


Average Points per Game by Metropolitan Teams Over the Years
Average Points per Game
1.00
0.75
0.50
0.25
0.00
2012
2016
2020
Year
team
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH





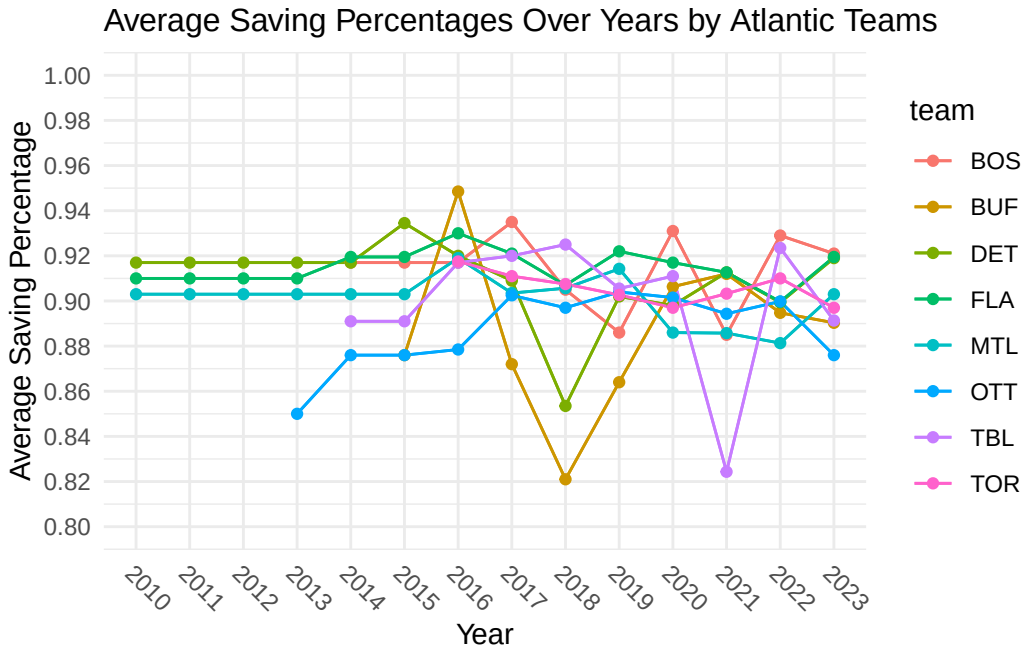
I excluded players that played no games, thus having 0 points per gamem or had no affiliated teams in the year, and applied this subset for the following analysis.

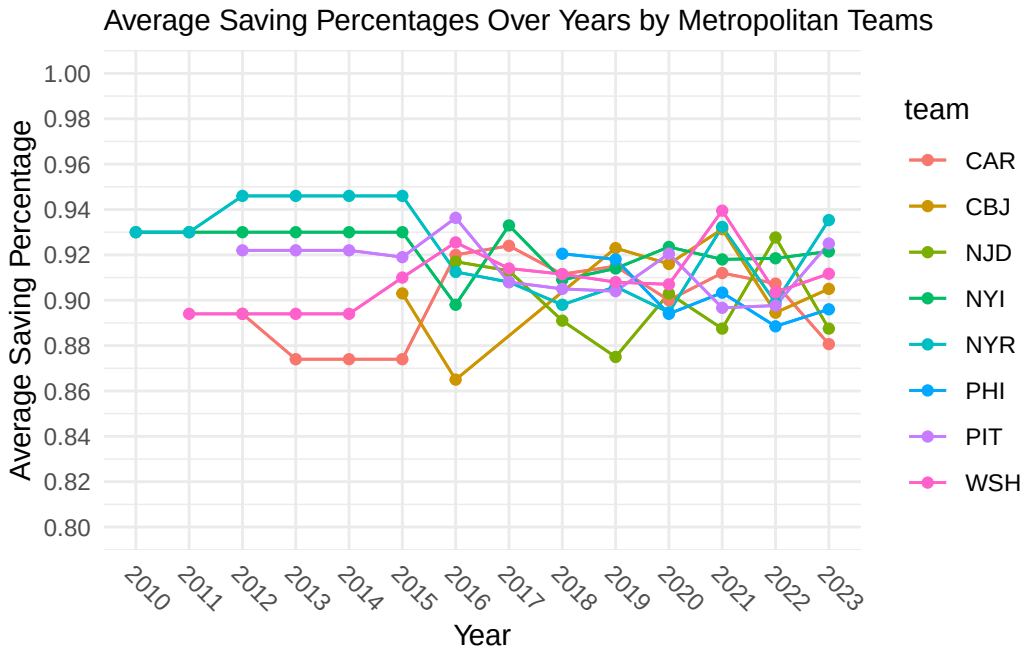
Tampa Bay Lightning, representing the Atlantic Division, had once achieved the highest average points per game across all seasons and teams included. The Atlantic and Pacific Divisions

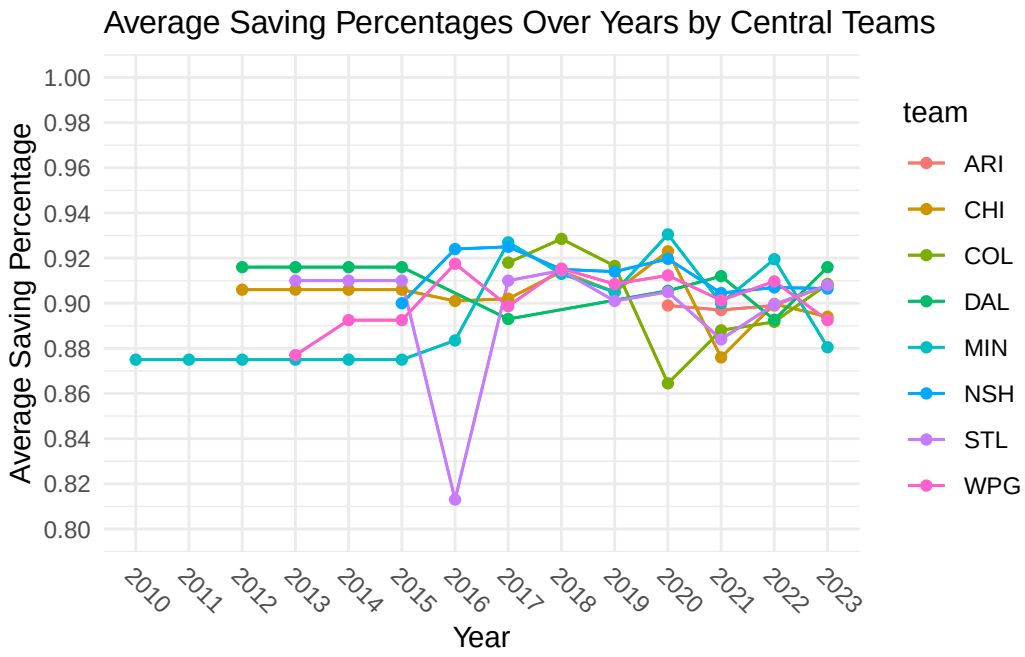
appeared to be relatively evenly matched, showing similar trends in average points scored per game among their respective teams. Similarly, the Metropolitan and Central Divisions displayed comparable levels of competitiveness, with overlapping ranges of maximum and minimum average points earned per game, though the Central Division may exhibit slightly higher competitiveness.

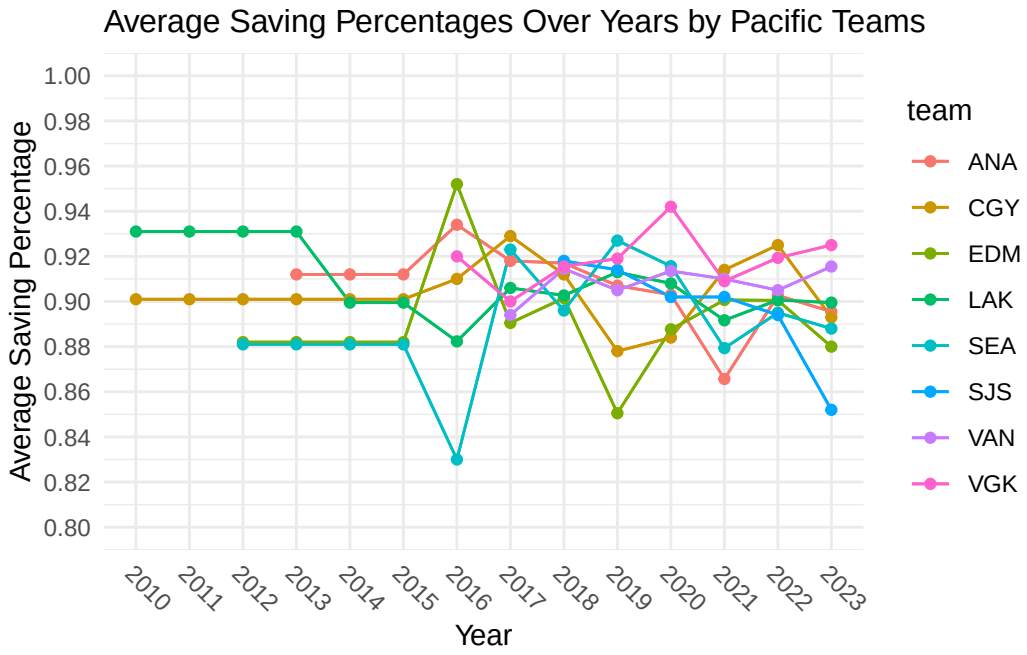
Average Saving Percentage by Team

```
  average_saving_percentage team year
1                     0.912  ANA 2013
2                     0.912  ANA 2014
3                     0.912  ANA 2015
4                     0.934  ANA 2016
5                     0.918  ANA 2017
6                     0.917  ANA 2018
```




Average Saving Percentages Over Years by Metropolitan Teams
Average Saving Percentage
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
team
CAR
CBJ
NJD
NYI
NYR
PHI
PIT
WSH


Average Saving Percentages Over Years by Central Teams
Average Saving Percentage
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
0.84
0.82
0.80
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
Year
team
ARI
CHI
COL
DAL
MIN
NSH
STL
WPG



Similarly, within the analysis of saving percentages, I focused exclusively on goaltenders' saving percentages. Consequently, I omitted individuals without team affiliations for the year, those who didn't participate in any games, and skaters, identifiable by the presence of "-" in their saving percentage records, from the following analysis.

Taking the Atlantic Division as an example, the Buffalo Sabres had both the highest and lowest average saving percentages within the data collection period from the 2010 season to the 2023 season. Tampa Bay Lightning also had a significantly lower saving percentage spotted than other teams. The goaltenders of Detroit Red Wings, Florida Panthers, and Montreal Canadiens showed a more moderate annual average performance across the years.

In the Metropolitan Division, the New York Rangers' goaltenders consistently demonstrated superior performance, holding the highest saving percentages among Metropolitan teams for 7 years. On the other hand, the Carolina Hurricanes, Columbus Blue Jackets, and New Jersey Devils initially recorded the lowest saving percentages. However, each team witnessed significant improvements in the subsequent years, possibly attributed to strategic goaltender trades aimed at enhancing their performance.

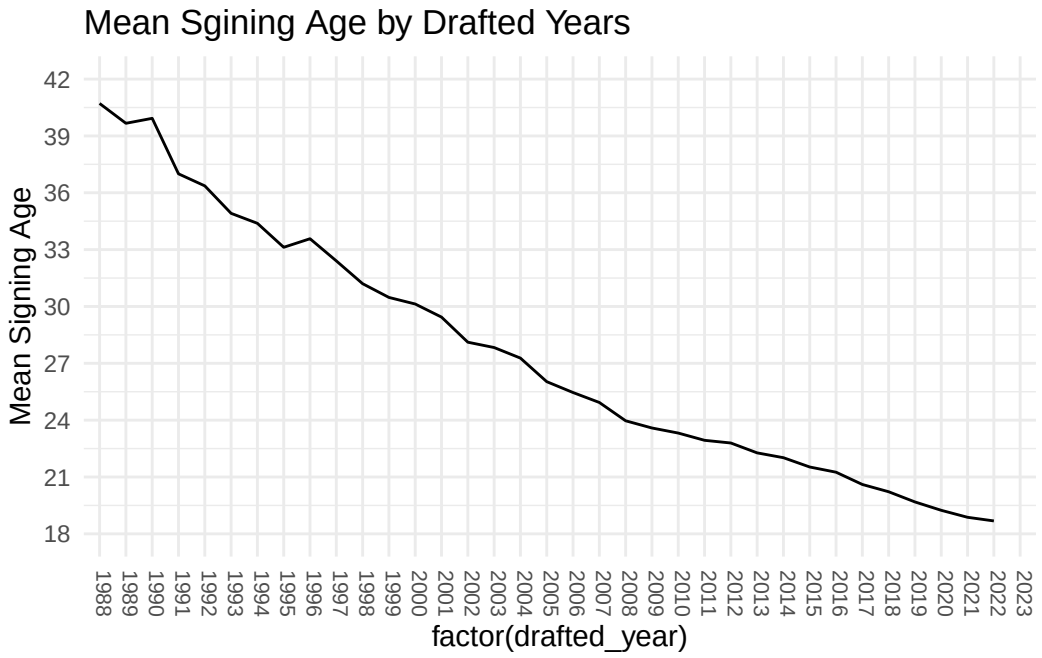
In the Central Division, both the St. Louis Blues and Minnesota Wilds showcased subpar performances, consistently registering some of the lowest saving percentages. The Edmonton Oilers demonstrated a mixed record in the Pacific Division, having the highest saving percentage and the second lowest saving percentage. Conversely, the Seattle Kraken held the lowest saving percentage in the division. While not consistent, the Vegas Golden Knights predominantly maintained a top-ranking position. Notably, the San Jose Sharks experienced

a significantly poorer saving percentage this season compared to other teams within the same division.

In summary, the Atlantic and Pacific Divisions exhibited the highest saving percentages at approximately around 0.95. Conversely, the Central Division recorded the lowest saving percentage, hovering around 0.81. The Metropolitan Division showed the smallest variations in saving performance among all divisions.

What is the mean age at signing (the latest contract) for each drafted year?

```
  mean_signing_age drafted_year
1         40.71429         1988
2         39.66667         1989
3         39.92857         1990
4         37.00000         1991
5         36.36364         1992
6         34.91429         1993
```



This plot illustrates that as the draft year of players approaches the current year, the age at which they sign contracts tends to decrease linearly, with an average annual decline of one year. Some greater drops are also spotted. This trend is quite reasonable because, according to the website I scraped, the age at which players sign contracts reflects the signing age of the most recent contract. Players drafted earlier by professional teams may be older now.

Additionally, the NHL stipulates that drafted players must be between 17 and 20 years old, so the average signing age in 2024 almost equals the draftable age. However, it's interesting to note that there are spikes in the average signing age from 1989 to 1990 and from 1995 to 1996. This may be due to the inclusion of more European players (whose age must be over 20) compared to other drafted years, or it could be a result of a smaller sample size in the data, leading to more pronounced effects of extreme values on the average.

Interactive Graphics

```
# Data
salary_df_3 <- salary_df_2[, c(1:3, 19:20)]
# Define UI # what users are able to see when they look at the screen
ui <- fluidPage(
  titlePanel("Introductory Visualization of NHL Team Salary"),
  # Sidebar layout with input and output definitions
  sidebarLayout(
    # Inputs
    sidebarPanel(position = "left",
      # Select team
      selectInput(inputId = "team",
                  label = "Team",
                  choices = c("ANA", "ARI", "BOS", "BUF",
                              "CAR", "CBJ", "CGY", "CHI",
                              "COL", "DAL", "DET", "EDM",
                              "FLA", "LAK", "MIN", "MTL",
                              "NJD", "NSH", "NYI", "NYR",
                              "OTT", "PHI", "PIT", "SEA",
                              "SJS", "STL", "TBL", "TOR",
                              "VAN", "VGK", "WPG", "WSH",
                              "-"),
                  selected = "-"),
    ),
    # Outputs
    mainPanel(
      plotOutput(outputId = "scatterplot"),
      br(), br(),
      dataTableOutput(outputId = "table")
    )
  )
)
```

```
# Define server function
server <- function(input, output) {
  # Create scatterplot object
  output$scatterplot <- renderPlot({
    salary_df_3 %>%
      filter(team == input$team) %>%
      ggplot(aes(x = factor(year), y = salary,
                 group = team)) +
        geom_point(size = 1.5, shape = 20) +
        geom_smooth(method = "lm", se = TRUE, size = 0.5) +
       labs(title = "Scatter Plot of Salaries by Year with Linear Regression Line",
       x = "Year",
       y = "Salary") +
      scale_y_continuous(breaks = seq(140000,
                                      17000000,
                                      1250000),
                         limits = c(140000, 17000000)) +
      coord_cartesian(ylim = c(140000, 17000000)) +
      theme(axis.text.x = element_text(angle = -45,
                                       hjust = 0,
                                       size = 10))
  })
  # Create data table
  output$table <- renderDataTable({
    datatable(data = salary_df_3 %>% select(player_name,
                                            team,
                                            drafted_year,
                                            salary,
                                            year),
              options = list(pageLength = 10),
              rownames = FALSE)
  })
}

# Create a Shiny app object
shinyApp(ui = ui, server = server)
```

Readers can use this interactive graphic to choose their preferred teams and view the salary distribution, which includes all players associated with that team. Additionally, there is a team category labeled as "-". Consequently, this interactive graphic incorporates all players from the data I extracted from the website, specifically those with a salary exceeding $0.

Given the limited data points obtained from the website scraping, employing the linear regres-

sion function may not be the most ideal approach. Nevertheless, this could serve as a valuable template for future endeavors, particularly when working with a more extensive dataset containing increased data points for each year and team.

Conclusions

Examining the points scored per game and average annual saving percentages, the Atlantic and Pacific Divisions demonstrated a relatively balanced competitiveness, with the former holding a slightly higher rank when considering shooting percentages. Broadly, salaries in the Atlantic Division seemed to correlate with players' capabilities, with teams like Toronto Maple Leafs, Tampa Bay Lightning, and Ottawa Senators securing the second- and third-highest average salaries among all 32 teams. Toronto Maple Leafs also maintained the record of paying the highest player salaries for three consecutive seasons within the 14 collected. In contrast, the Pacific Division consistently reported the lowest average salaries, seemingly not aligning with the highly ranked game performances of its players. Notably, all Pacific Division teams exhibited mean shooting percentages greater than their median shooting percentages, emphasizing concerns about player underpayment.

Similarly, the Metropolitan and Central Divisions showed comparable competitiveness in terms of points scored per game. However, concerning saving percentages, the Central Division reported the lowest saving percentage, while the Metropolitan Division displayed the smallest variations among all divisions. Despite comparatively poorer performances, six out of the sixteen teams in these divisions, including the Pittsburgh Penguins, Detroit Red Wings, Carolina Hurricanes, Washington Capitals, Dallas Stars, and Colorado Avalanche, demonstrated generosity in players' salaries. Four of these teams held records of paying the highest salaries, and two paid the highest average salary.

Limitations

Due to the limited information contained in this web-scraped dataset, certain intriguing phenomena remain unexplained. Examples include the two drops in the top-paid salary by team per game season, the two increases in the mean signing age by drafted year, and the narrowed gaps between average salaries of each team starting from 2019.

Moreover, my exploratory analysis, accompanied by data visualizations, aims to highlight key seasonal team statistics that might interest fans who do not use the teams' mobile applications, which typically provide detailed analyses for each play of the games. The analyses are designed to be basic and accessible for ice hockey fans of any level, refraining from delving into more intricate analyses that could facilitate deeper interactions between fans and their favorite teams or players. Some illustrations include the complexity of the formations of contracts signed and salaries paid to NHL players.

Finally, the raw data limitations also hinder me from conducting more meaningful analyses. For instance, my preference is to focus on determining the average initial signing age of each drafted year rather than the average latest signing age of each drafted year.