TP4

```
#install.packages("lubridate")
```

```
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.2
```

```
Warning: package 'tidyr' was built under R version 4.3.2
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.0     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ---------------------------------------- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
library(lubridate)
library(rvest)
```

```
Attaching package: 'rvest'
```

```
The following object is masked from 'package:readr':

    guess_encoding
```

```
library(dplyr)
```

1)

Afin de faire ce travail, j'ai choisi de prendre une page Wikipédia sur la saison 2022-2023 de Ligue Nationale de Hockey, car j'aimerai regarder si le fait d'avoir des joueurs de son équipe dans les classements de meilleurs scoreurs et meilleurs gardiens influence le classement final lorsque les conférences ouest et est sont regroupées. Je voulais initialement aller sur la page LNH, afin d'avoir aussi les meilleurs défenseurs, mais je n'arrivais pas à scrapper les données.

Je pourrai ainsi comparer qui entre les scoreurs ou les gardiens influencent le plus le résultat final de l'équipe. Les données sont appropriées, car elles correspondent aux données que l'on peut trouver sur le site de la LNH qui regroupe toutes les statistiques officielles.

Le classement des scoreurs qui m'intéresse est le classement des points et pour les gardiens celui des moyennes de buts encaissés par match.

Cette étude permettra de comprendre si la position d'une franchise dans le classement final dépend de certains joueurs ou alors d'une équipe homogène. Si cela dépend de quelques joueurs seulement, est-ce que c'est plus les attaquants ou les gardiens qui font la différence.

```
NHL_class_W_Est <- data_frame(
  Team = c("y - Carolina Hurricanes", "x- New Jersey Devils", "x- New York Rangers"),
  Pts = c(113, 112, 107))
```

```
Warning: `data_frame()` was deprecated in tibble 1.1.0.
i Please use `tibble()` instead.
```

```
NHL_class_W_Est1 <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Stati
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(6)
```

```
NHL_class_W_Est2 <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Stati
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(7)
```

```
NHL_class_W_Ouest <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Stat
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(8)
```

```
NHL_class_W_Ouest1 <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Sta
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(9)
```

```
NHL_class_W_Ouest2 <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Sta
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(10)
```

```
NHL_class_scoreur <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Stat
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(12)
```

```
NHL_class_goal <- read_html("https://en.wikipedia.org/wiki/2022%E2%80%9323_NHL_season#Statist
  html_elements("table") |>
  html_table(fill = T) |>
  pluck(13)
```

2)

Au sein de ma première base de données, s'appelant CLASS_NHL_tri, j'ai seulement 2 variables, la première est le nom de l'équipe et le nombre de points que l'équipe a accumulé, en fonction des victoires hors prolongation, des victoires en prolongation ou des défaites. Cette base de données ne me sert pas pour les analyses, mais me permet de visualiser le classement à titre informatif pour les futures analyses.

Au sein de ma deuxième base de données, s'appelant CLASS_SCOR, j'ai le nom des 10 meilleurs joueurs offensifs comme première variable (Player), en deuxième c'est le nom de l'équipe dans laquelle ils jouent (Team) et en troisième c'est leur nombre de point en fonction du nombre de buts marqués et le nombre d'assists réalisés (Pts). La dernière variable (ClassTeam), correspond à la position de l'équipe du joueur dans le classement final de la ligue avant les playoffs.

Au sein de ma troisième base de données, s'appelant CLASS_SCOR, j'ai le nom des 10 meilleurs gardiens de la ligue comme première variable (Player), en deuxième c'est le nom de la franhise pour laquelle il joue (Team) et en troisième, c'est sa statistique de buts encaissés par match (plus elle est basse, meilleur le gardien est) (GAA). La dernière variable (ClassTeam), correspond à la position de l'équipe du joueur dans le classement final de la ligue avant les playoffs.

```
NHL_class_W_Ouest3 <- subset(NHL_class_W_Ouest2, select = -Div)
NHL_class_Ouest <- rbind(NHL_class_W_Ouest, NHL_class_W_Ouest1, NHL_class_W_Ouest3)
NHL_class_Ouest1 <- subset(NHL_class_Ouest, select = - c(GP, W, L, OTL, RW, GF, GA,GD, Pos))
  na.omit(NHL_class_Ouest)
NHL_Ouest <- NHL_class_Ouest1 |> rename(Team = `Team v  t  e`)
```

```
A <- subset(NHL_class_W_Est1, select = - c(GP, W, L, OTL, RW, GF, GA,GD, Pos))
B <- subset(NHL_class_W_Est2, select = - c(Div, GP, W, L, OTL, RW, GF, GA,GD, Pos)) |>
  na.omit(NHL_class_W_Est2)

A1 <-A |> rename(Team = `Team v  t  e`)
B1 <- B |> rename(Team = `Team v  t  e`)
NHL_Est <- rbind(NHL_class_W_Est, A1, B1)
```

```
CLASS_NHL <- rbind(NHL_Est, NHL_Ouest)
CLASS_NHL_tri <- CLASS_NHL |>
  arrange(desc(Pts))

CLASS_S <- subset(NHL_class_scoreur, select = -c(GP, G, A, PIM))
CLASS_S1 <- subset(CLASS_S, select = -ncol(CLASS_S))
```

```r
CLASS_G <- subset(NHL_class_goal, select = -c(GP, TOI, W, L, OTL, GA, SO))
CLASS_G1 <- subset(CLASS_G, select = c("Player", "Team", "GAA"))

Class_G <- c(1, 11, 1, 4, 15, 8, 3, 9, 2, 14)
CLASS_GOAL <- CLASS_G1 |> mutate(ClassTeam = Class_G)

Class_S <- c(7, 7, 1, 13, 6, 8, 17, 6, 7, 22)
CLASS_SCOR <- CLASS_S1 |> mutate(ClassTeam = Class_S)
```

```r
glimpse(CLASS_NHL_tri)
```

```
Rows: 32
Columns: 2
$ Team <chr> "p - Boston Bruins", "y - Carolina Hurricanes", "x- New Jersey De~
$ Pts  <dbl> 135, 113, 112, 111, 111, 109, 109, 108, 107, 104, 103, 100, 98, 9~
```

```r
glimpse(CLASS_SCOR)
```

```
Rows: 10
Columns: 4
$ Player    <chr> "Connor McDavid", "Leon Draisaitl", "David Pastrnak", "Nikit~
$ Team      <chr> "Edmonton Oilers", "Edmonton Oilers", "Boston Bruins", "Tamp~
$ Pts       <int> 153, 128, 113, 113, 111, 109, 109, 105, 104, 102
$ ClassTeam <dbl> 7, 7, 1, 13, 6, 8, 17, 6, 7, 22
```

```r
glimpse(CLASS_GOAL)
```

```
Rows: 10
Columns: 4
$ Player    <chr> "Linus Ullmark", "Filip Gustavsson", "Jeremy Swayman", "Ilya~
$ Team      <chr> "Boston Bruins", "Minnesota Wild", "Boston Bruins", "Toronto~
$ GAA       <dbl> 1.89, 2.10, 2.27, 2.33, 2.34, 2.37, 2.45, 2.48, 2.48, 2.49
$ ClassTeam <dbl> 1, 11, 1, 4, 15, 8, 3, 9, 2, 14
```

3)

L'objectif de l'analyse est d'évaluer l'impact de l'appartenance à une équipe de meilleurs scoreurs ou de meilleurs gardiens sur sa position dans le classement final de la Ligue nationale de hockey. Afin de vérifier s'il existe un lien statistique entre le nombre de points offensifs d'un joueur et le classement final de son équipe, j'ai décidé de faire une corrélation bivariée. J'ai effectué la même analyse, afin de vérifier le potentiel lien statistique entre la moyenne de buts encaissés par match par un gardien et le classement final de son équipe dans la ligue de hockey.

Lors du test de la première relation, le coefficient rho de Spearman est de -0.31, ce qui signifie qu'il y a une corrélation moyenne monotone décroissante entre les points offensifs marqués par un joueur et le classement de sa franchise. Cependant, cette corrélation n'est pas statistiquement significative à un niveau de confiance de 95% (p=0.8). De plus, on remarque visuellement, l'absence de lien entre ces deux variables.

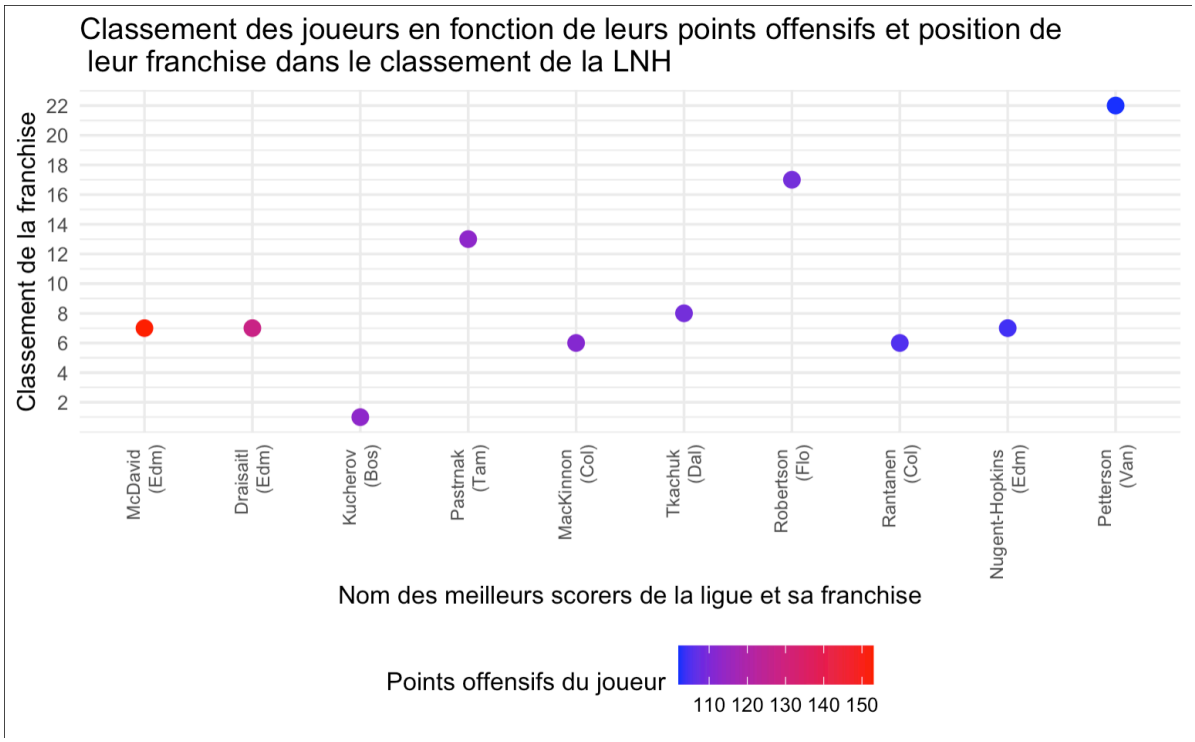


Figure 1: Graphique présentant la relation entre points offensifs d'un joueur et classement de sa franchise au sein de la LNH

Concernant la vérification d'un lien statistique potentiel entre le score moyen de buts encaissés par un gardien et le classement final de la franchise, il existe une relation monotone moyenne et croissante. Cependant, selon le p-value (p=0.16), cette corrélation n'est pas statistiquement

significative à un niveau de confiance de 95%. Visuellement, une certaine corrélation entre ces deux variables est visible, cependant, le peu de valeurs à l'étude peut expliquer la valeur de p élevée.

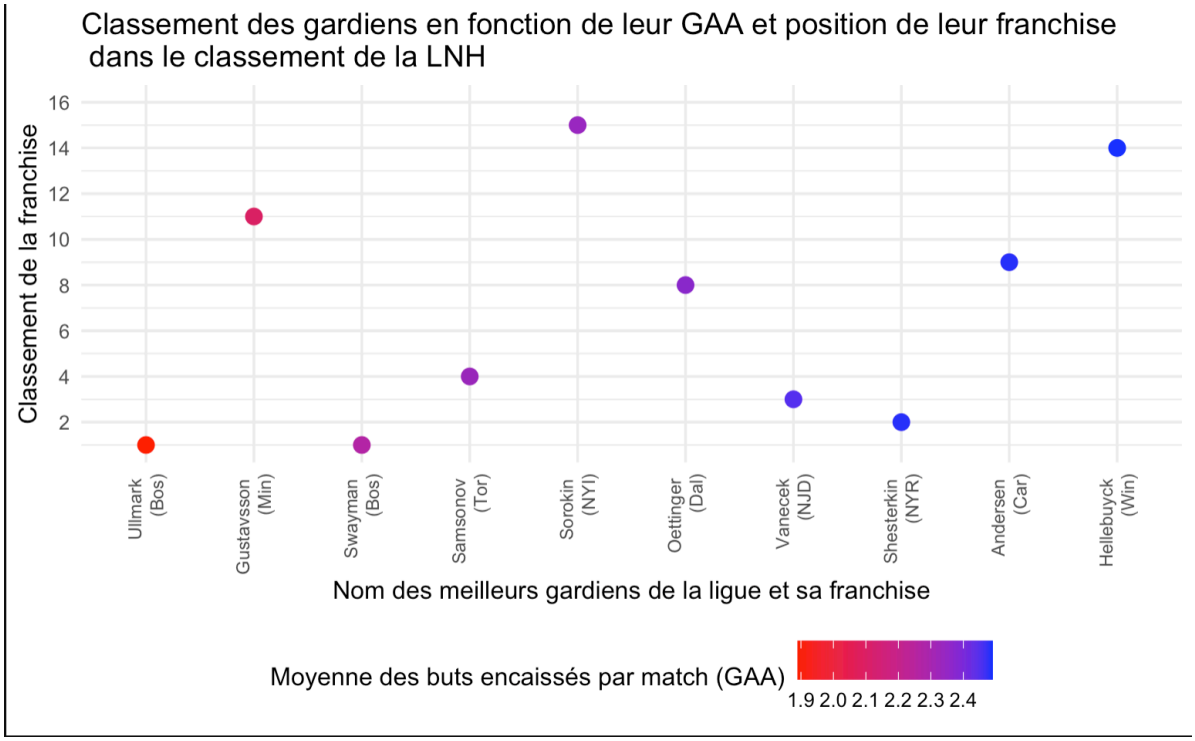


Figure 2: Graphique présentant la relation entre moyenne des buts encaissés par match d'un gardien et classement de sa franchise au sein de la LNH

L'absence de relation significative entre le nombre de points offensifs ou la moyenne des buts encaissés par match et la position finale de la franchise du joueur dans la ligue nationale de hockey peut être compris comme le fait que pour réussir dans la ligue, la franchise doit se doter de joueur ayant tous un bon niveau et non forcément de « supers-joueurs ». Cette absence de relation met aussi de l'avant qu'il existe également d'autres facteurs à prendre en compte dans la réussite de la franchise, comme les tactiques de jeu.

Finalement, le manque de significativité des résultats est dû au peu de valeurs que comportent les bases de données. Cette étude pourrait être réalisée avec des données de plusieurs saisons ou alors avec le classement complet des joueurs offensifs et des gardiens. Cependant, cela n'était pas disponible sur la page Wikipédia. On aurait pu également évaluer la relation avec les joueurs défensifs, mais le classement de ces meilleurs joueurs n'était pas disponible sur Wikipédia.

4)

Le scraping de données pour l'analyse comporte des responsabilités éthiques importantes, notamment en matière de consentement, de respect de la vie privée ou encore de la transparence.

Premièrement, concernant le consentement des données, dans cette situation, les joueurs en jouant dans la LNH acceptent implicitement ou explicitement de voir leurs statistiques officielles, car la LNH elle-même les diffuse. Cependant, afin de respecter la véracité des données, je suis allée vérifier si les données de la page Wikipédia correspondaient à celles du site de la LNH.

Deuxièmement, concernant le respect de la vie et les données que j'ai collecté, je ne pense pas qu'il peut y avoir de grands dommages avec ce type de données.

Finalement, il faut être transparent quant à la collecte des données, ce qui a été fait, car on retrouve l'ensemble de mon code à la suite du travail.

Annexe code

```
cor.test(x = CLASS_SCOR$Pts, y= CLASS_SCOR$ClassTeam, alternative = 'greater', method = 'spea
```

```
Warning in cor.test.default(x = CLASS_SCOR$Pts, y = CLASS_SCOR$ClassTeam, :
Cannot compute exact p-value with ties
```

```

    Spearman's rank correlation rho

data:  CLASS_SCOR$Pts and CLASS_SCOR$ClassTeam
S = 217.11, p-value = 0.813
alternative hypothesis: true rho is greater than 0
sample estimates:
       rho
-0.3158031
```

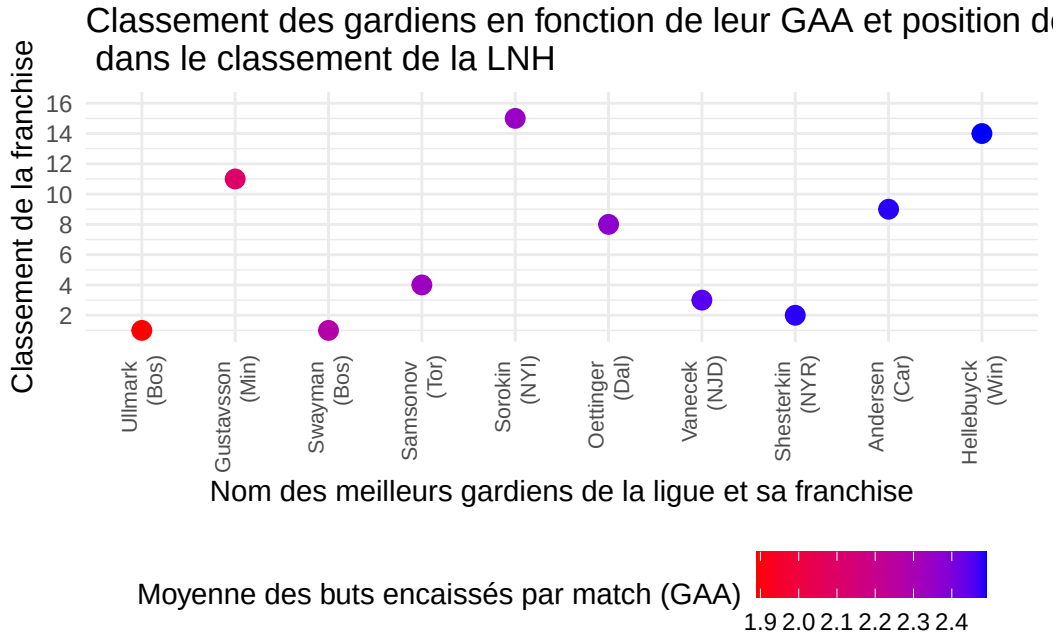
```
cor.test(x = CLASS_GOAL$GAA, y= CLASS_GOAL$ClassTeam, alternative = 'greater', method = 'spea
```

```
Warning in cor.test.default(x = CLASS_GOAL$GAA, y = CLASS_GOAL$ClassTeam, :
Cannot compute exact p-value with ties
```

```

    Spearman's rank correlation rho

data:  CLASS_GOAL$GAA and CLASS_GOAL$ClassTeam
S = 107.65, p-value = 0.1625
alternative hypothesis: true rho is greater than 0
sample estimates:
     rho
0.347561
```

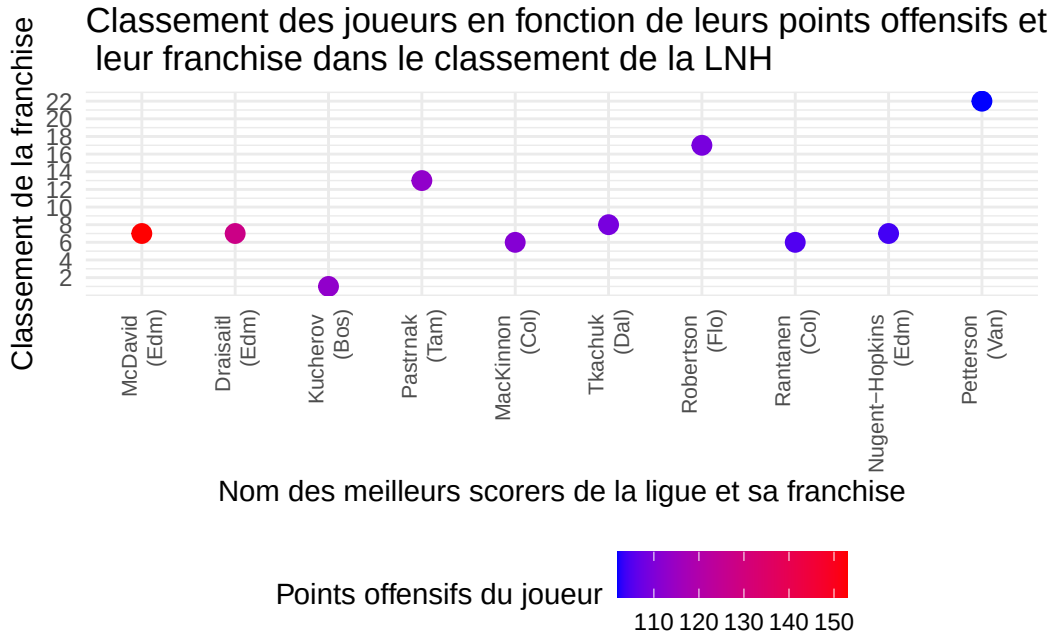
```
a <- ggplot(data = CLASS_GOAL, aes(x = reorder(Player, GAA), y = ClassTeam,  color = GAA)) +
  geom_point(size= 3) +
  scale_y_continuous(limits = c(1, 16),
                     breaks = c(2, 4, 6, 8, 10, 12, 14, 16)) +
  scale_x_discrete(labels = c("Ullmark \n \n (Bos)", "Gustavsson \n \n (Min)", "Swayman \n \
  scale_color_gradientn(colours = c("red", "blue")) +
  xlab("Nom des meilleurs gardiens de la ligue et sa franchise") +
```

```
  ylab("Classement de la franchise") +
  labs(color = "Moyenne des buts encaissés par match (GAA)") +
  ggtitle("Classement des gardiens en fonction de leur GAA et position de leur franchise \n d
  theme_minimal() +
  theme(axis.text.x = element_text(size = 8, angle = 90, hjust = 1, vjust = 0.5, lineheight =
        legend.position = "bottom")

a
```




```
b <- ggplot(data = CLASS_SCOR, aes(x = reorder(Player, - Pts), y = ClassTeam,  color = Pts))
  geom_point(size= 3) +
  scale_y_continuous(limits = c(1, 22),
                     breaks = c(2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22)) +
  xlab("Nom des meilleurs scorers de la ligue et sa franchise") +
  ylab("Classement de la franchise") +
  scale_x_discrete(labels = c("McDavid \n \n (Edm)", "Draisaitl \n \n (Edm)", "Kucherov \n \
  scale_color_gradientn(colours = c("blue", "red")) +
  labs(color = "Points offensifs du joueur") +
  ggtitle("Classement des joueurs en fonction de leurs points offensifs et position de \n leu
  theme_minimal() +
  theme(axis.text.x = element_text(size = 8, angle = 90, hjust = 1, vjust = 0.5, lineheight =
        legend.position = "bottom")
```

```
b
```


Classement des joueurs en fonction de leurs points offensifs et leur franchise dans le classement de la LNH
Classement de la franchise
22
20
16
14
12
10
6
4
2
McDavid (Edm)
Draisaitl (Edm)
Kucherov (Bos)
Pastrnak (Tam)
MacKinnon (Col)
Tkachuk (Dal)
Robertson (Flo)
Rantanen (Col)
Nugent−Hopkins (Edm)
Petterson (Van)
Nom des meilleurs scorers de la ligue et sa franchise
Points offensifs du joueur
110 120 130 140 150