Player tags and traits

Naga Santhosh Kartheek Karnati

3/26/2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Misc. EDA

```r
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(tidyr)
library(stringr)
library(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```r
library(rio)
```

```
## Warning: package 'rio' was built under R version 3.6.2
```

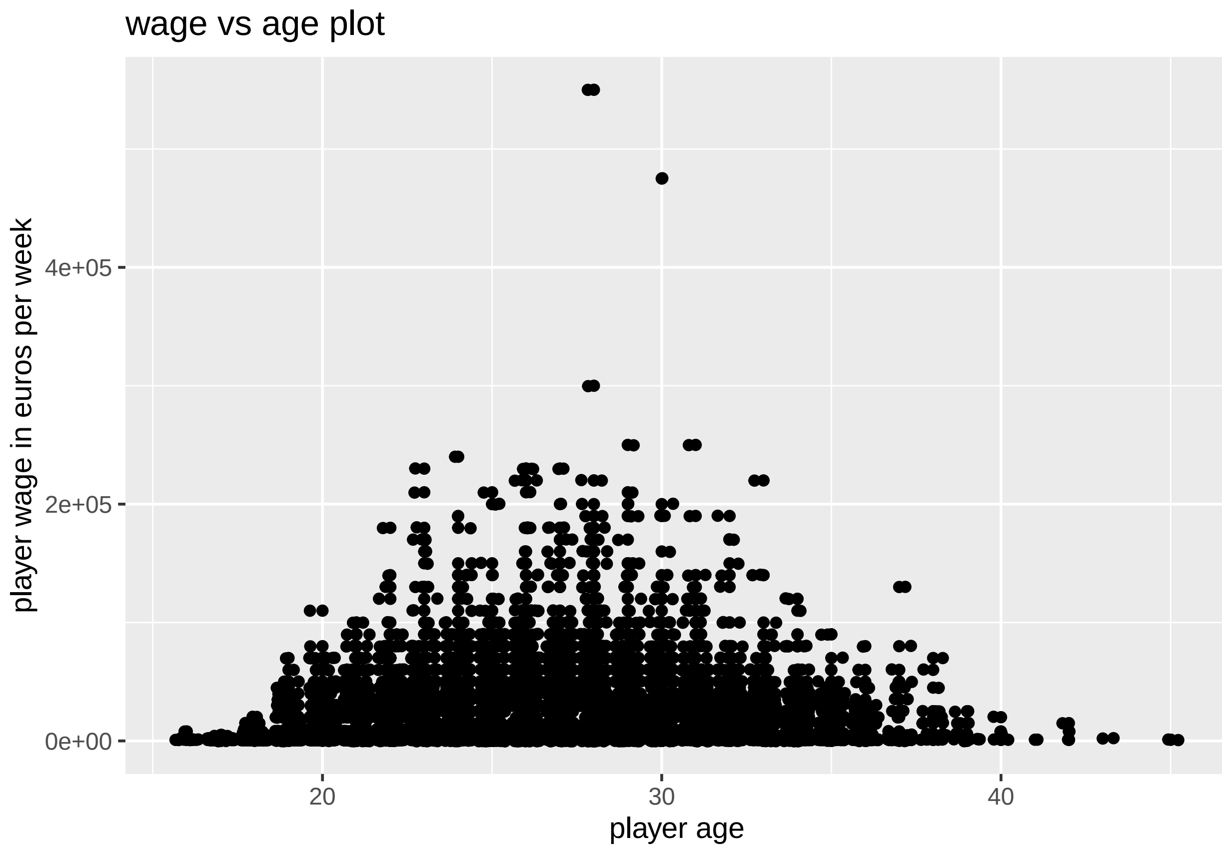
```r
library(modelr)
library(purrr)
```

```
##
## Attaching package: 'purrr'

## The following object is masked from 'package:data.table':
##
##     transpose
```

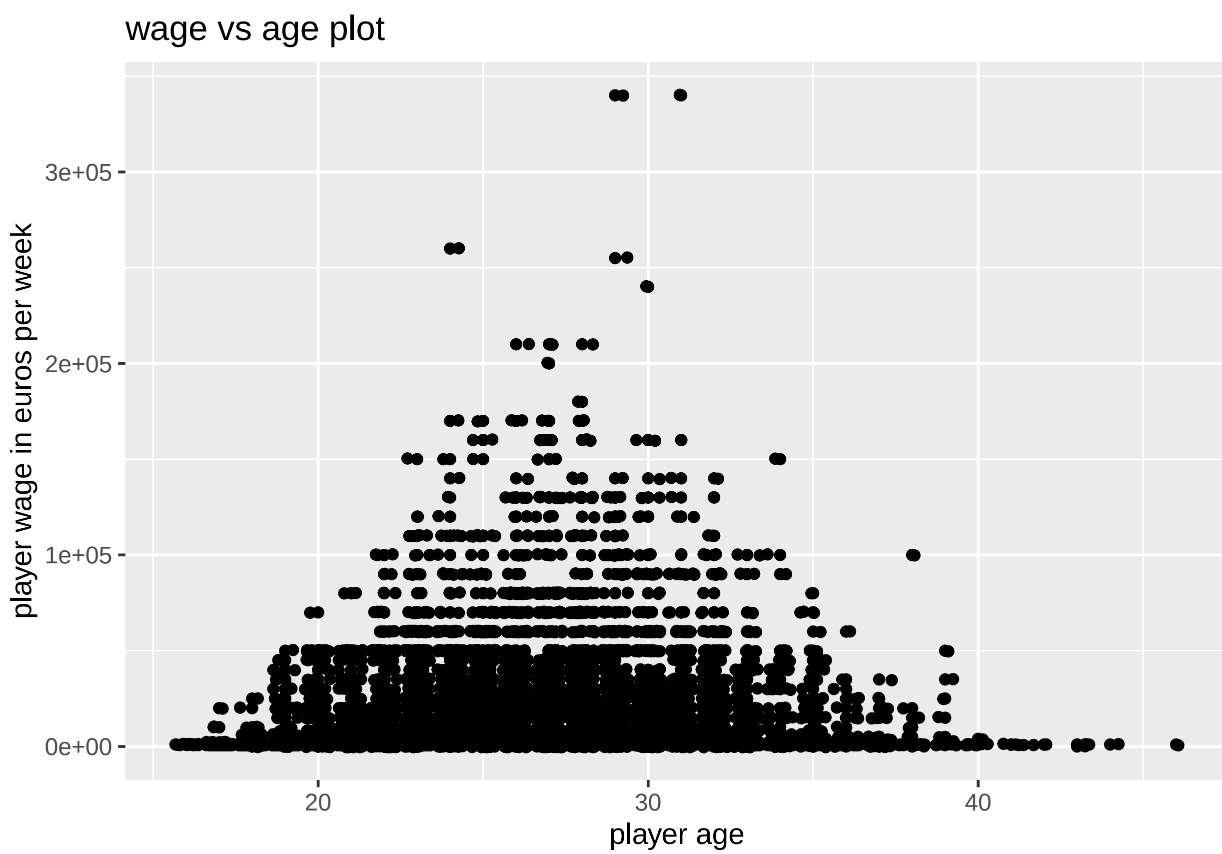
```r
#loading the datasets:
#FIFA20 players dataset:

fi20 <- fread('D:/NEU/Spring 2020/SML/Project/Datasets/players_20.csv')
fi19 <- fread('D:/NEU/Spring 2020/SML/Project/Datasets/players_19.csv')
fi18 <- fread('D:/NEU/Spring 2020/SML/Project/Datasets/players_18.csv')
fi17 <- fread('D:/NEU/Spring 2020/SML/Project/Datasets/players_17.csv')
fi16 <- fread('D:/NEU/Spring 2020/SML/Project/Datasets/players_16.csv')

fi20 <- as_tibble(fi20)
fi19 <- as_tibble(fi19)
fi18 <- as_tibble(fi18)
fi17 <- as_tibble(fi17)
fi16 <- as_tibble(fi16)

fifa_datasets_list = list(fi16, fi17, fi18, fi19, fi20)
years = list("2016", "2017", "2018", "2019", "2020")
```

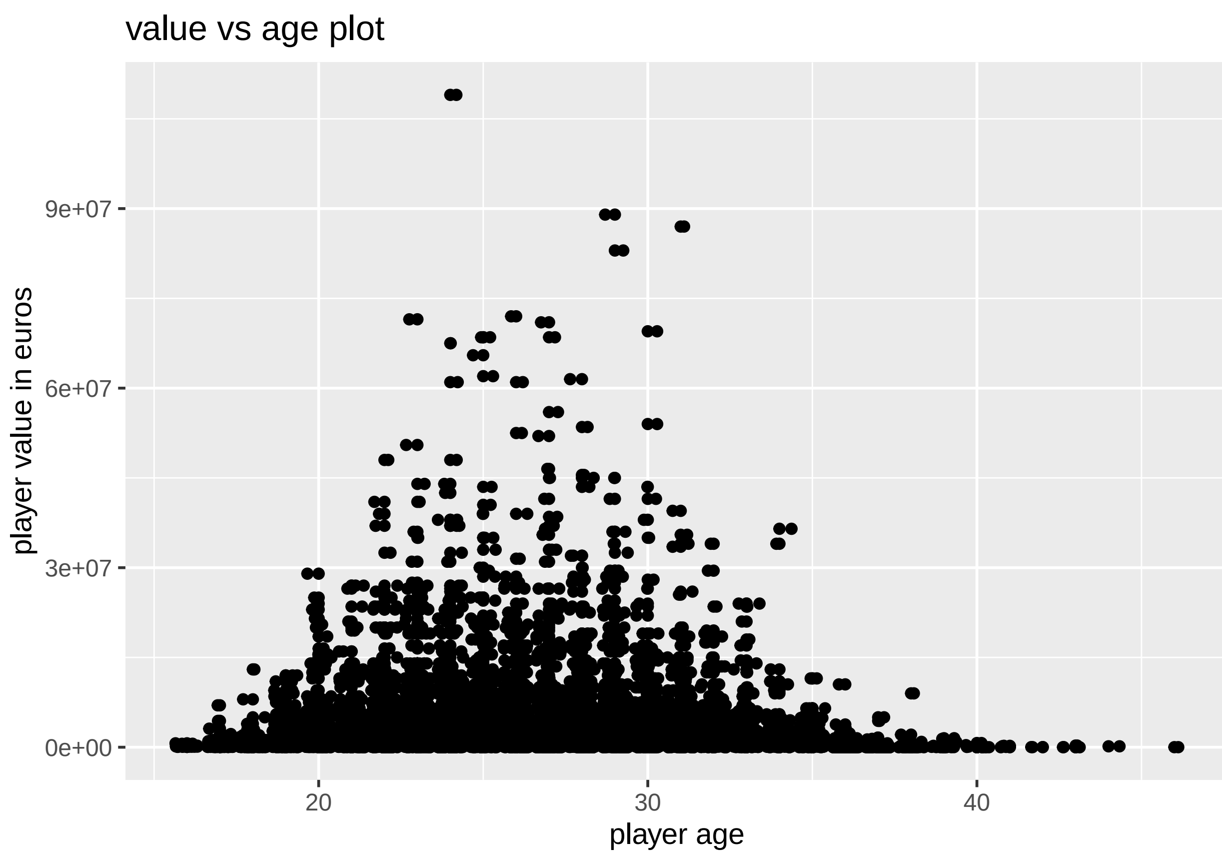
EDA on player age, wage and value

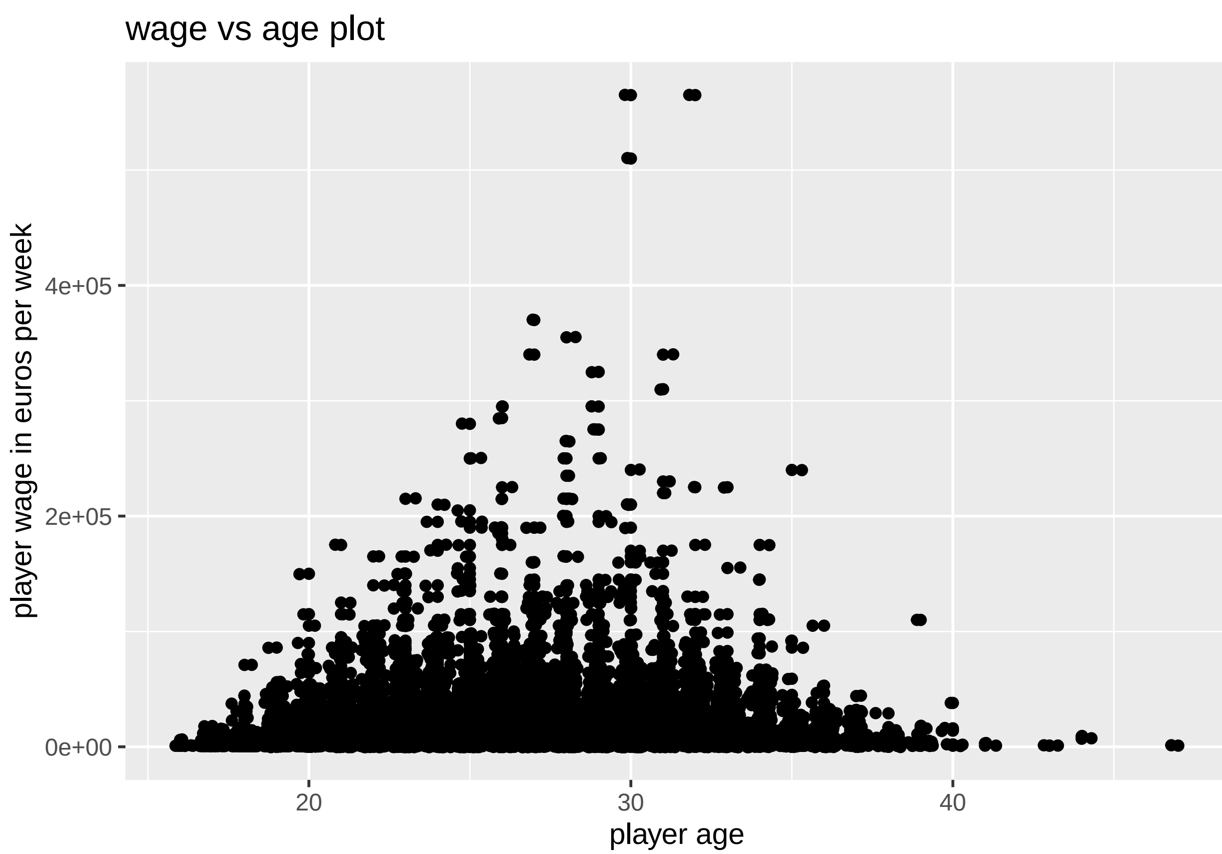
```r
for (i in seq_along(fifa_datasets_list))
{
  wage_plot <- fifa_datasets_list[[i]] %>% ggplot(aes(age, wage_eur))+geom_point()+
    geom_jitter()+labs(x="player age", y="player wage in euros per week",
                       title = "wage vs age plot")
  print(wage_plot)

  value_plot <- fifa_datasets_list[[i]] %>% ggplot(aes(age, value_eur))+geom_point()+
    geom_jitter()+labs(x="player age", y="player value in euros",
                       title = "value vs age plot")
  print(value_plot)

}
```

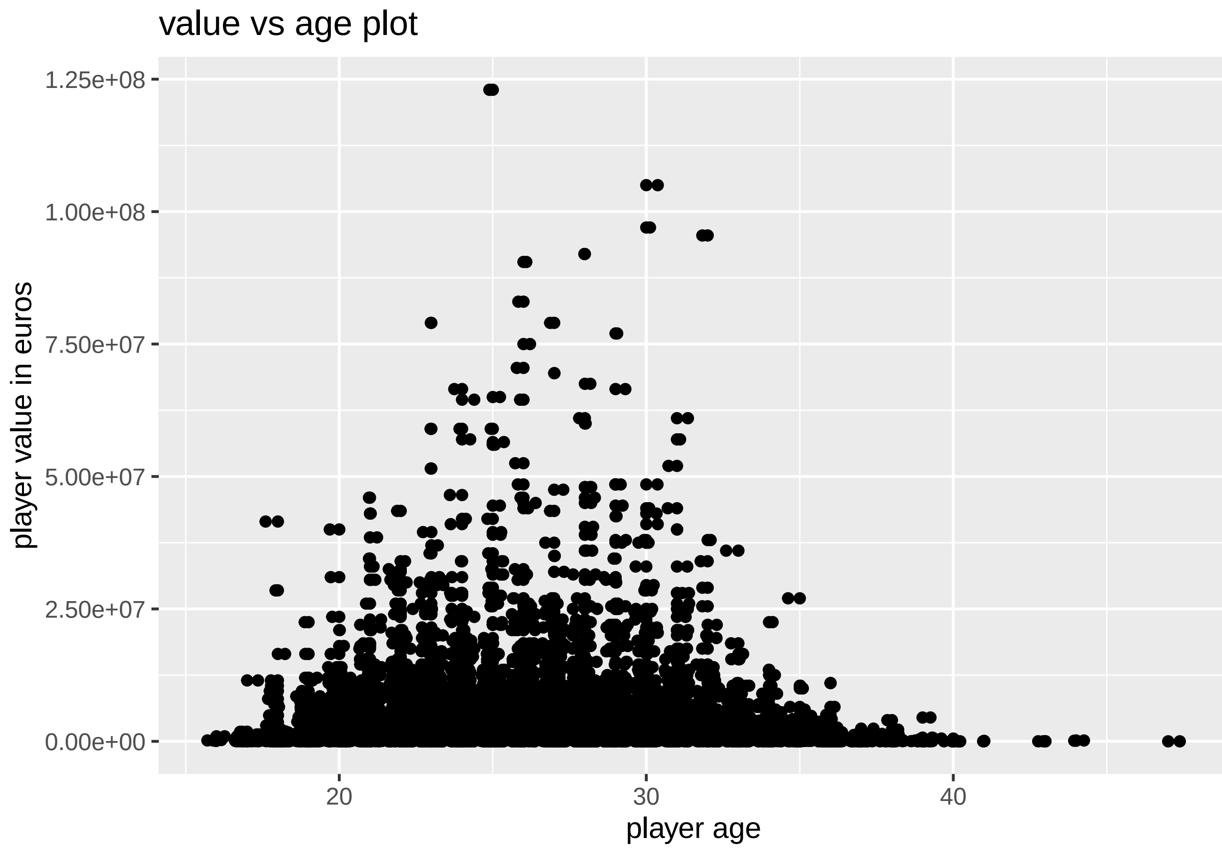

wage vs age plot
player wage in euros per week
4e+05
2e+05
0e+00
20
30
40
player age


value vs age plot
player value in euros
9e+07
6e+07
3e+07
0e+00
20
30
40
player age

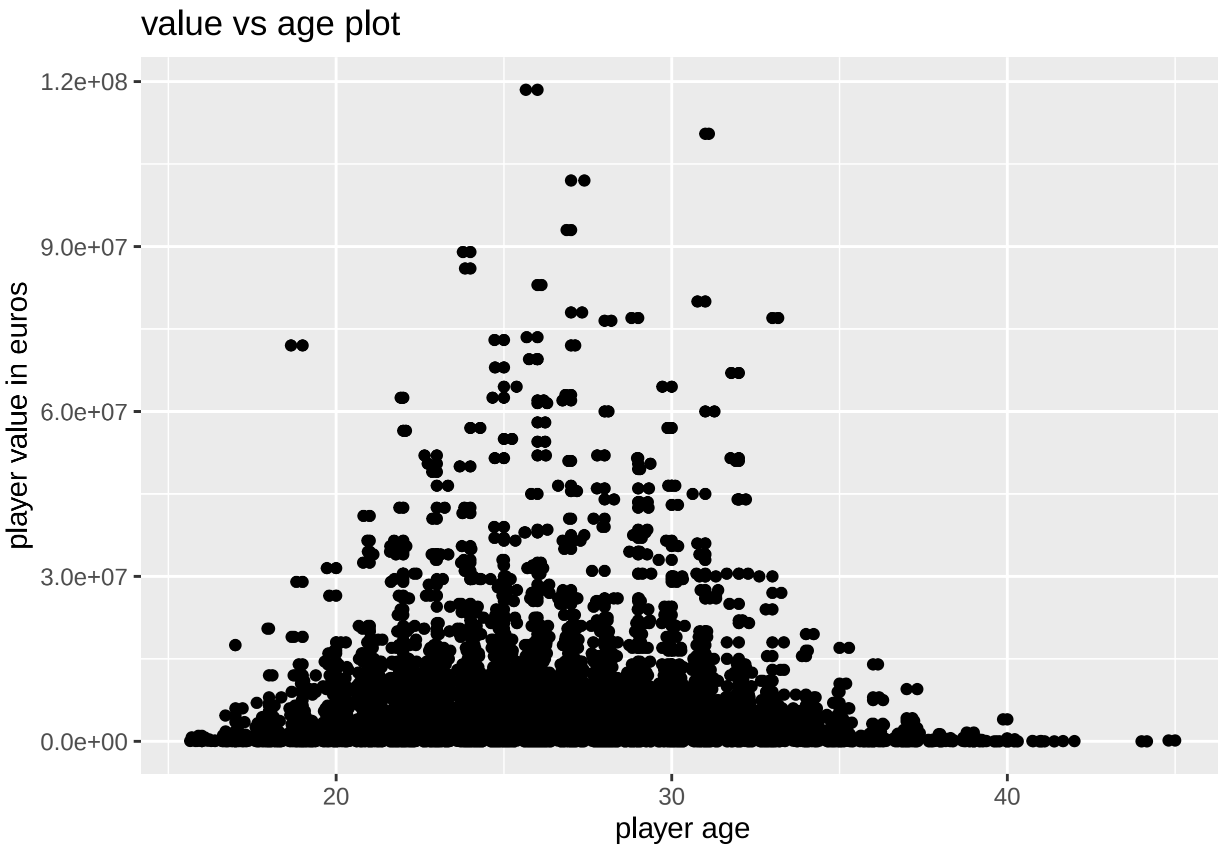

wage vs age plot
player wage in euros per week
3e+05
2e+05
1e+05
0e+00
20
30
40
player age

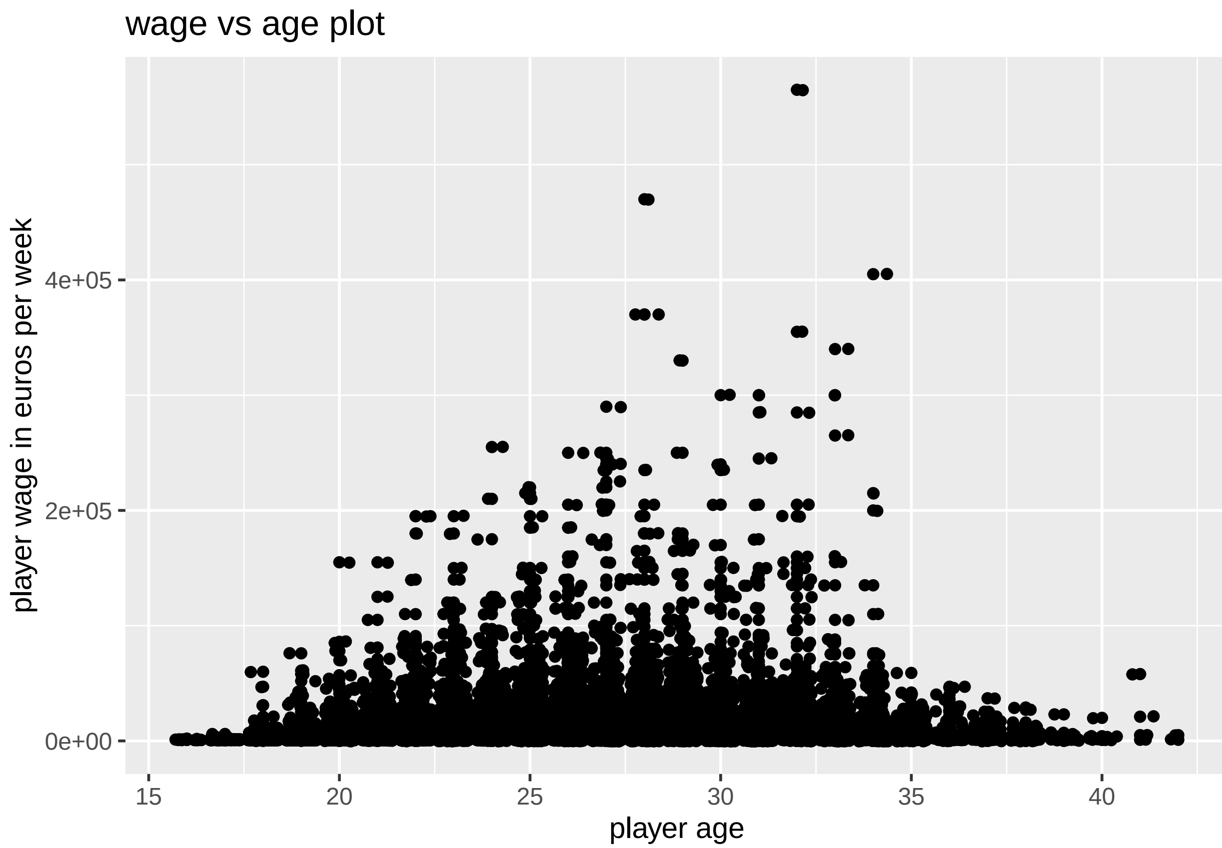

value vs age plot
9e+07
6e+07
3e+07
0e+00
player value in euros
20
30
40
player age


wage vs age plot
player wage in euros per week
4e+05
2e+05
0e+00
20
30
40
player age

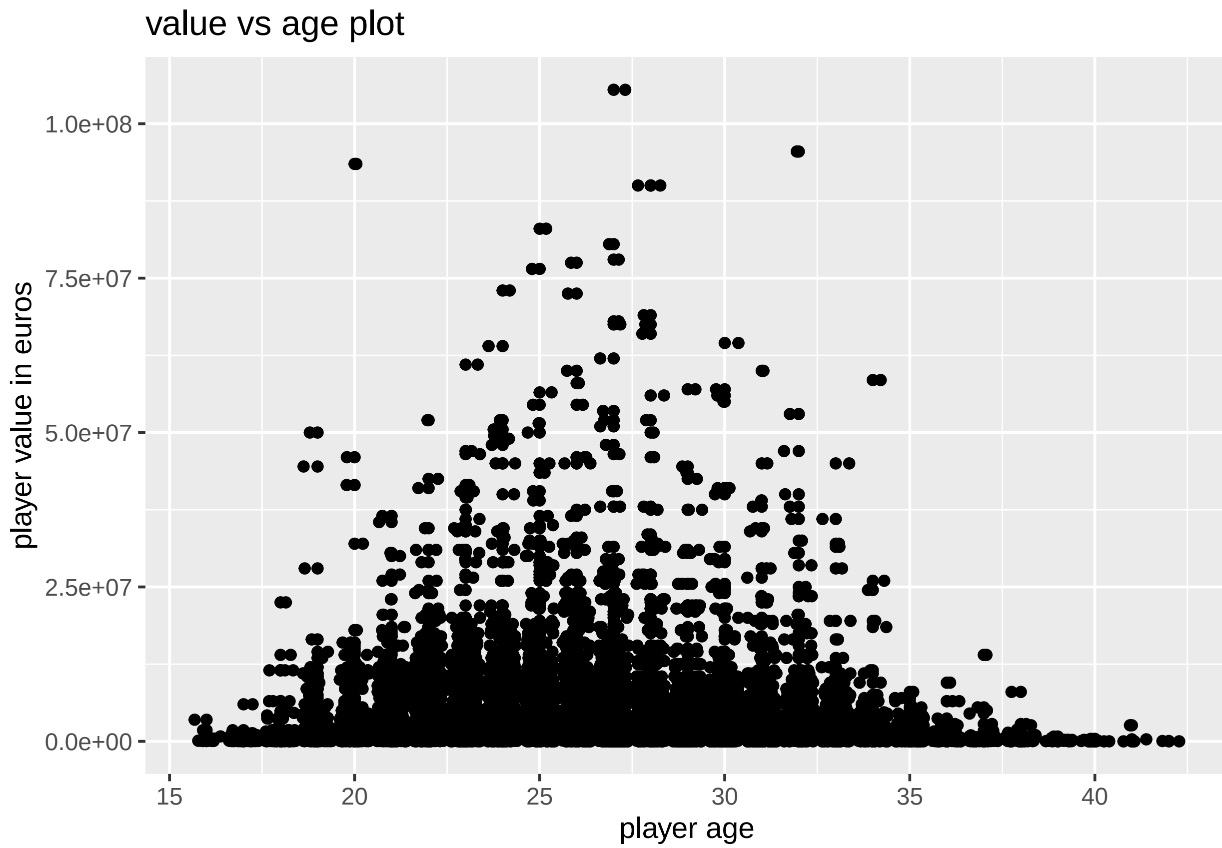

value vs age plot
1.25e+08
1.00e+08
7.50e+07
5.00e+07
2.50e+07
0.00e+00
player value in euros
20
30
40
player age


wage vs age plot
player wage in euros per week
4e+05
2e+05
0e+00
20
30
40
player age


value vs age plot
player value in euros
1.2e+08
9.0e+07
6.0e+07
3.0e+07
0.0e+00
20
30
40
player age



wage vs age plot
player wage in euros per week
4e+05
2e+05
0e+00
15
20
25
30
35
40
player age



Correlation amongst age, wage and value with age threshold 28:

```
for (i in seq_along(fifa_datasets_list))
{
  players28_andless <- fifa_datasets_list[[i]] %>%
    filter(age <= 28)%>% select(age, wage_eur, value_eur)
  players_over28 <- fifa_datasets_list[[i]] %>%
    filter(age > 28)%>% select(age, wage_eur, value_eur)

  cor_28andless <- cor(players28_andless)
  print(paste("Year", years[[i]], ":", "Player age <= 28"))
  print(round(cor_28andless, 2))
  #age is positively correlated with both wage and value.
  #Also wage and value are highly +vely correlated.

  cor_over28 <- cor(players_over28)
  print(paste("Year", years[[i]], ":", "Player age > 28"))
  print(round(cor_over28, 2))
  #age is negatively correlated with both wage and value but the correlation is
  #almost negligable b/w the age and wage pair over the years.
  #Also wage and value are highly +vely correlated.


  }
```

```
## [1] "Year 2016 : Player age <= 28"
##            age wage_eur value_eur
## age       1.00     0.32      0.18
## wage_eur  0.32     1.00      0.91
## value_eur 0.18     0.91      1.00
## [1] "Year 2016 : Player age > 28"
##             age wage_eur value_eur
## age        1.00    -0.10     -0.14
## wage_eur  -0.10     1.00      0.91
## value_eur -0.14     0.91      1.00
## [1] "Year 2017 : Player age <= 28"
##            age wage_eur value_eur
## age       1.00     0.29      0.17
## wage_eur  0.29     1.00      0.87
## value_eur 0.17     0.87      1.00
## [1] "Year 2017 : Player age > 28"
##             age wage_eur value_eur
## age        1.00    -0.12     -0.15
## wage_eur  -0.12     1.00      0.89
## value_eur -0.15     0.89      1.00
## [1] "Year 2018 : Player age <= 28"
##            age wage_eur value_eur
## age       1.00     0.22      0.18
## wage_eur  0.22     1.00      0.85
## value_eur 0.18     0.85      1.00
## [1] "Year 2018 : Player age > 28"
##             age wage_eur value_eur
## age        1.00    -0.09     -0.16
## wage_eur  -0.09     1.00      0.88
## value_eur -0.16     0.88      1.00
## [1] "Year 2019 : Player age <= 28"
##            age wage_eur value_eur
## age       1.00     0.20      0.18
## wage_eur  0.20     1.00      0.87
## value_eur 0.18     0.87      1.00
## [1] "Year 2019 : Player age > 28"
##             age wage_eur value_eur
## age        1.00    -0.08     -0.15
## wage_eur  -0.08     1.00      0.88
## value_eur -0.15     0.88      1.00
## [1] "Year 2020 : Player age <= 28"
##            age wage_eur value_eur
## age       1.00     0.21      0.18
## wage_eur  0.21     1.00      0.88
## value_eur 0.18     0.88      1.00
## [1] "Year 2020 : Player age > 28"
##             age wage_eur value_eur
## age        1.00    -0.06     -0.14
## wage_eur  -0.06     1.00      0.87
## value_eur -0.14     0.87      1.00
```

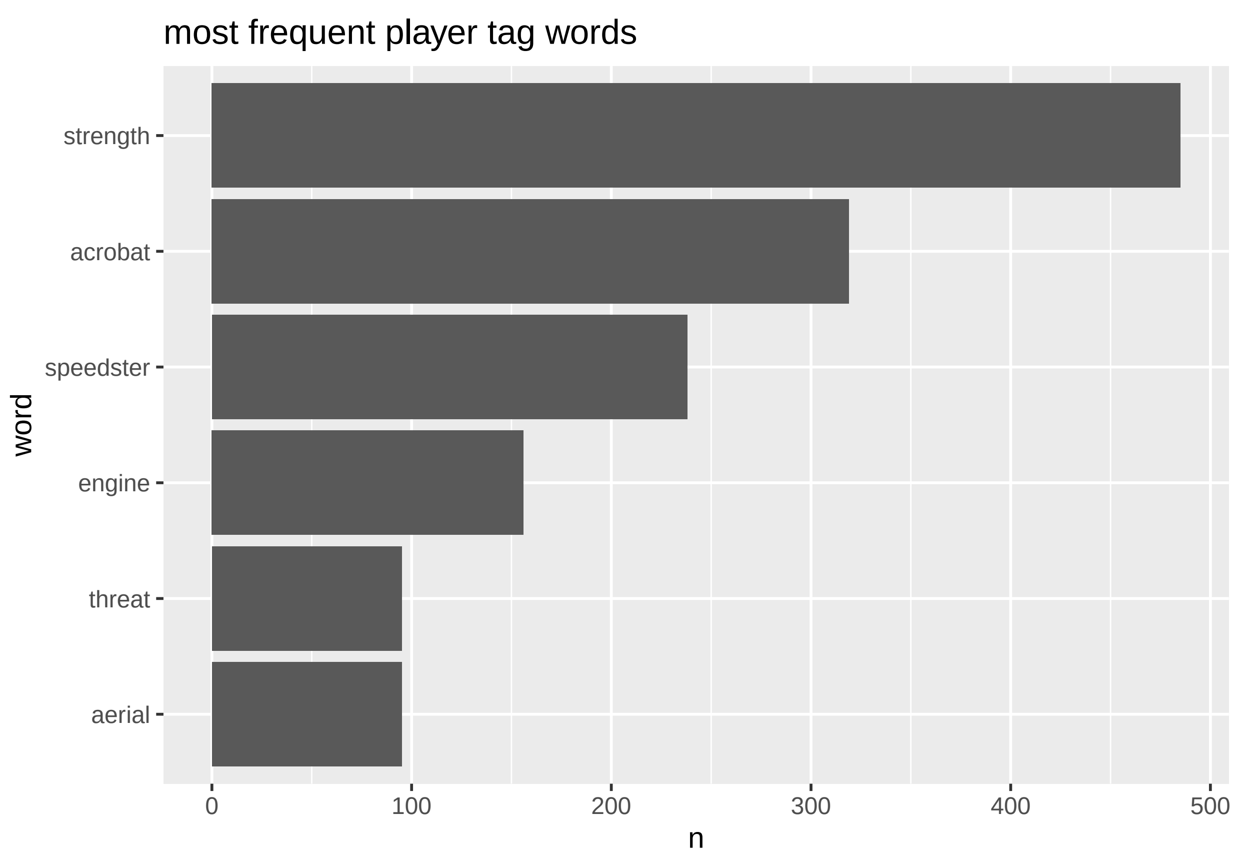
Player tags analysis:

```
library(tokenizers)
library(tidytext)
library(stringr)


#######SINGLE WORDS
for (i in seq_along(fifa_datasets_list))
{
tidy_fifa20 <- fifa_datasets_list[[i]] %>%
  group_by(sofifa_id) %>%
  mutate(row_num = row_number())%>%
  ungroup() %>%
  unnest_tokens(word, player_tags) %>%
  select(sofifa_id, short_name, age,club, nationality, team_position, team_jersey_number,
         overall, potential, weak_foot, skill_moves, work_rate, pace, shooting, passing,
         dribbling, defending, physic, word)

#View(tidy_fifa20)


#Most common player tags (single word):
tags <- as_tibble(tidy_fifa20 %>% group_by(word) %>% count() %>% arrange(desc(n)))
print(paste("Most common single word player tags in", years[[i]],":"))
print(tags)
#Strength seems to be the most common player tag followed by acrobat and speedster.
#Visualization:
tagswordplot <- tidy_fifa20 %>% count(word, sort = TRUE)%>%top_n(5)%>% mutate(word=reorder(word, n))%>%
  ggplot(aes(word,n))+geom_col()+coord_flip()+labs(x="word","frequency",
                                                    title = "most frequent player tag words")
print(tagswordplot)

}
```

```
## [1] "Most common single word player tags in 2016 :"
## # A tibble: 22 x 2
##    word           n
##    <chr>      <int>
##  1 strength     485
##  2 acrobat      319
##  3 speedster    238
##  4 engine       156
##  5 aerial        95
##  6 threat        95
##  7 dribbler      78
##  8 tacklingâ     34
##  9 tacticianâ    33
## 10 crosser       24
## # ... with 12 more rows
```

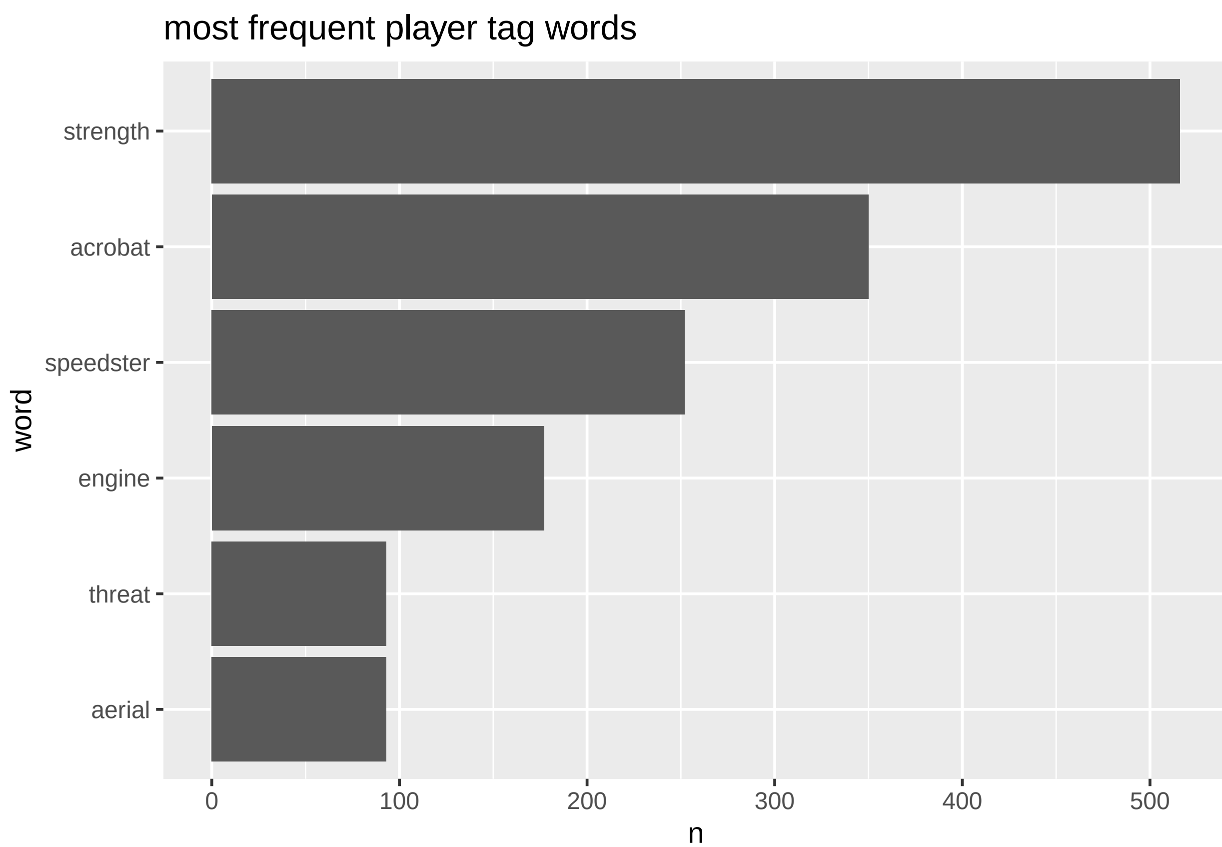
```
## Selecting by n
```

```
## [1] "Most common single word player tags in 2017 :"
## # A tibble: 22 x 2
##    word           n
##    <chr>      <int>
##  1 strength     516
##  2 acrobat      350
##  3 speedster    252
##  4 engine       177
##  5 aerial        93
##  6 threat        93
##  7 dribbler      75
##  8 tacticianâ    33
##  9 crosser       28
## 10 tacklingâ     26
## # ... with 12 more rows

## Selecting by n
```



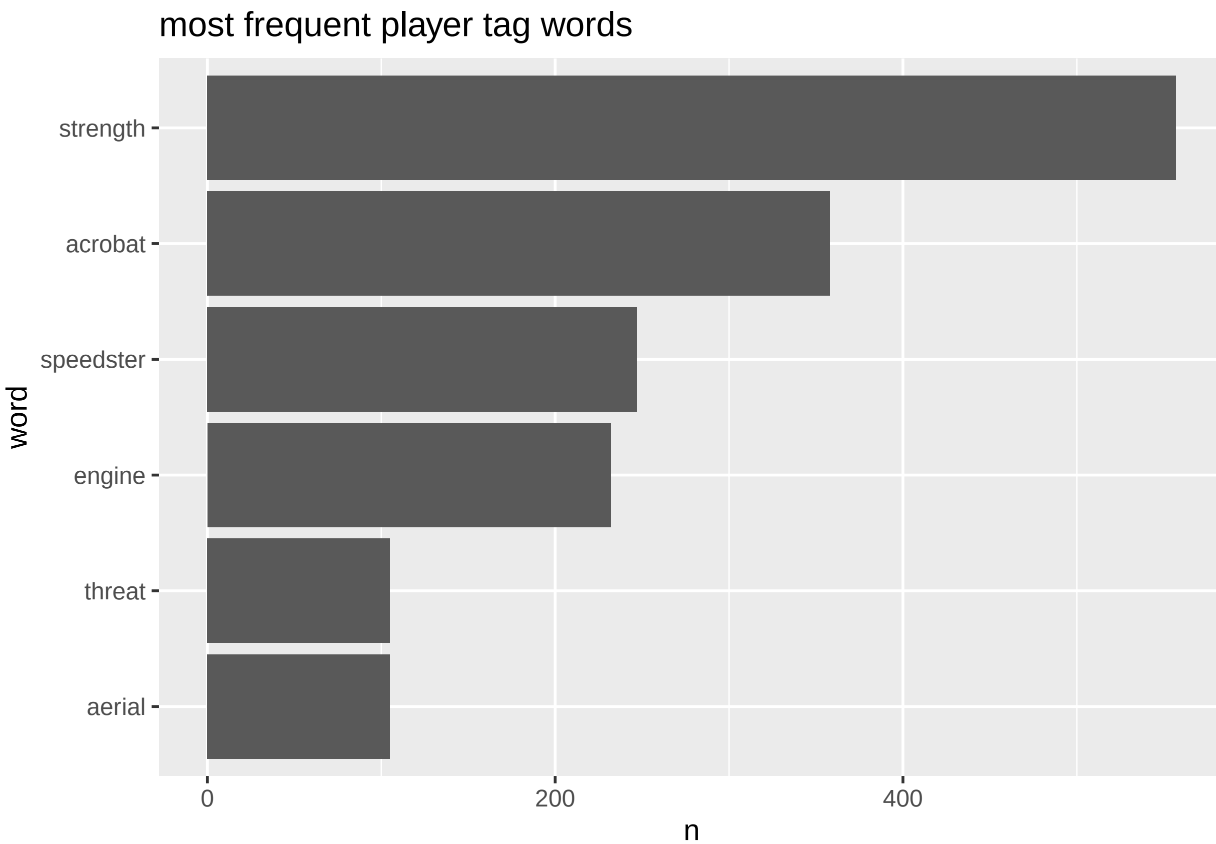

```
## [1] "Most common single word player tags in 2018 :"
## # A tibble: 22 x 2
##    word           n
##    <chr>      <int>
##  1 strength     557
##  2 acrobat      358
##  3 speedster    247
```

```
##   4 engine       232
##   5 aerial       105
##   6 threat       105
##   7 dribbler      93
##   8 tacticianâ    31
##   9 tacklingâ     30
##  10 crosser       27
## # ... with 12 more rows

## Selecting by n
```



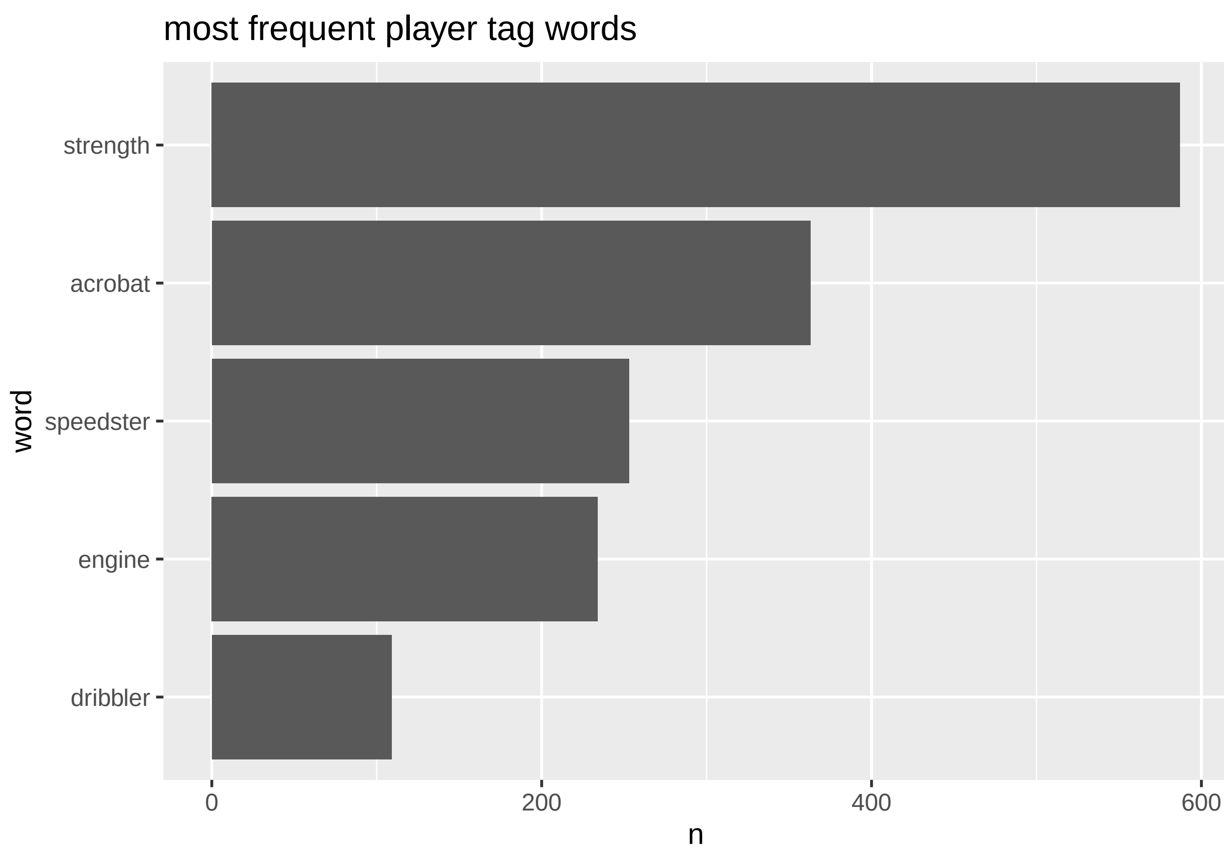

```
## [1] "Most common single word player tags in 2019 :"
## # A tibble: 22 x 2
##    word           n
##    <chr>      <int>
##  1 strength     587
##  2 acrobat      363
##  3 speedster    253
##  4 engine       234
##  5 dribbler     109
##  6 aerial        99
##  7 threat        99
##  8 tacticianâ    37
##  9 crosser       35
## 10 tacklingâ     34
## # ... with 12 more rows
```

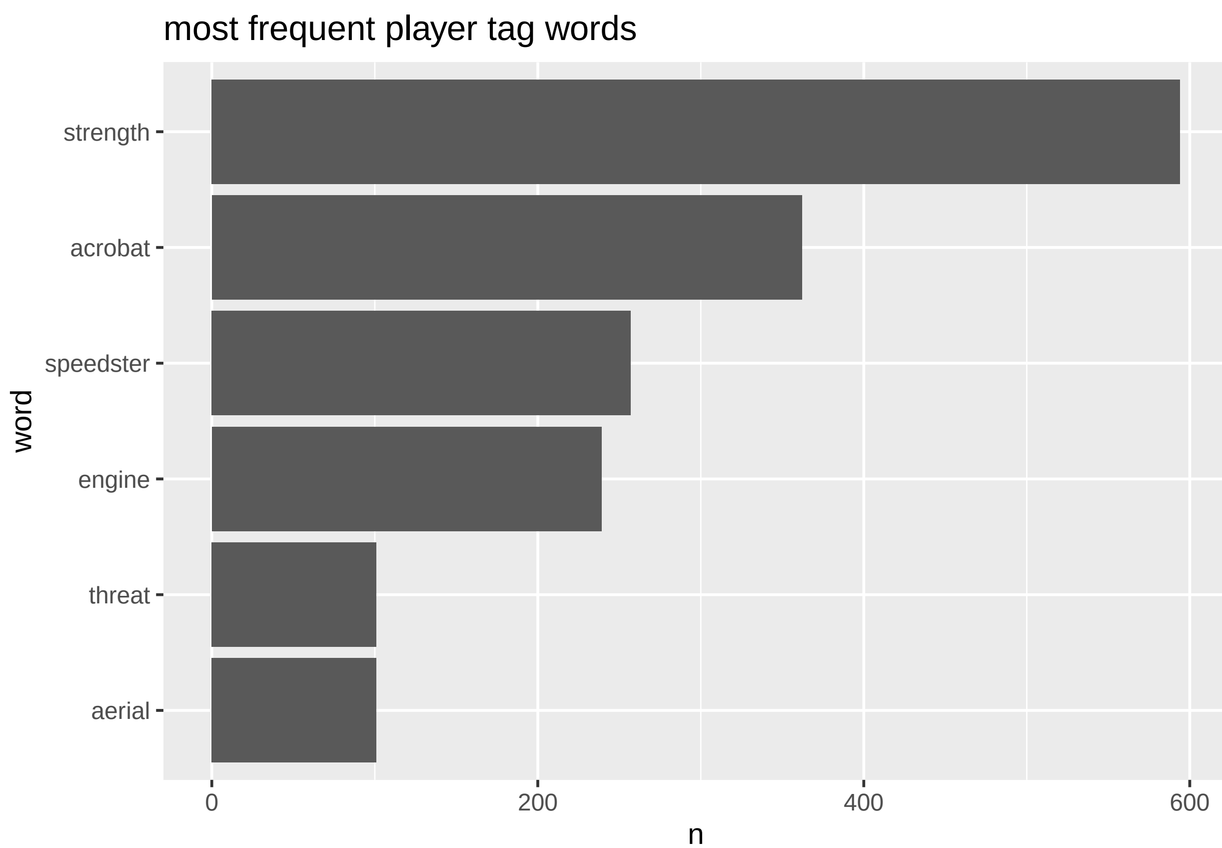
```
## Selecting by n
```




```
## [1] "Most common single word player tags in 2020 :"
## # A tibble: 22 x 2
##    word           n
##    <chr>      <int>
##  1 strength     594
##  2 acrobat      362
##  3 speedster    257
##  4 engine       239
##  5 aerial       101
##  6 threat       101
##  7 dribbler      95
##  8 complete      40
##  9 tacticianâ    36
## 10 crosser       31
## # ... with 12 more rows
```

```
## Selecting by n
```

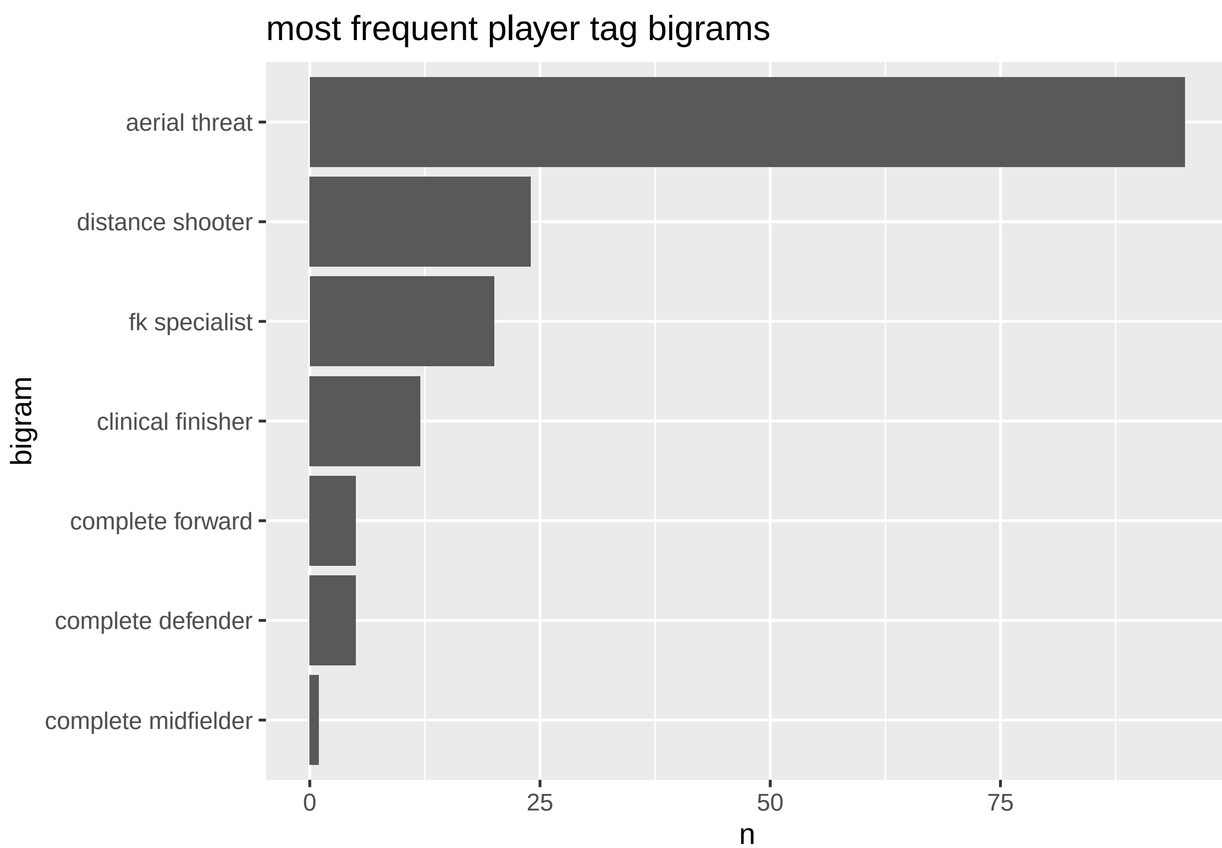

most frequent player tag words
word
strength
acrobat
speedster
engine
dribbler
0
200
400
600
n



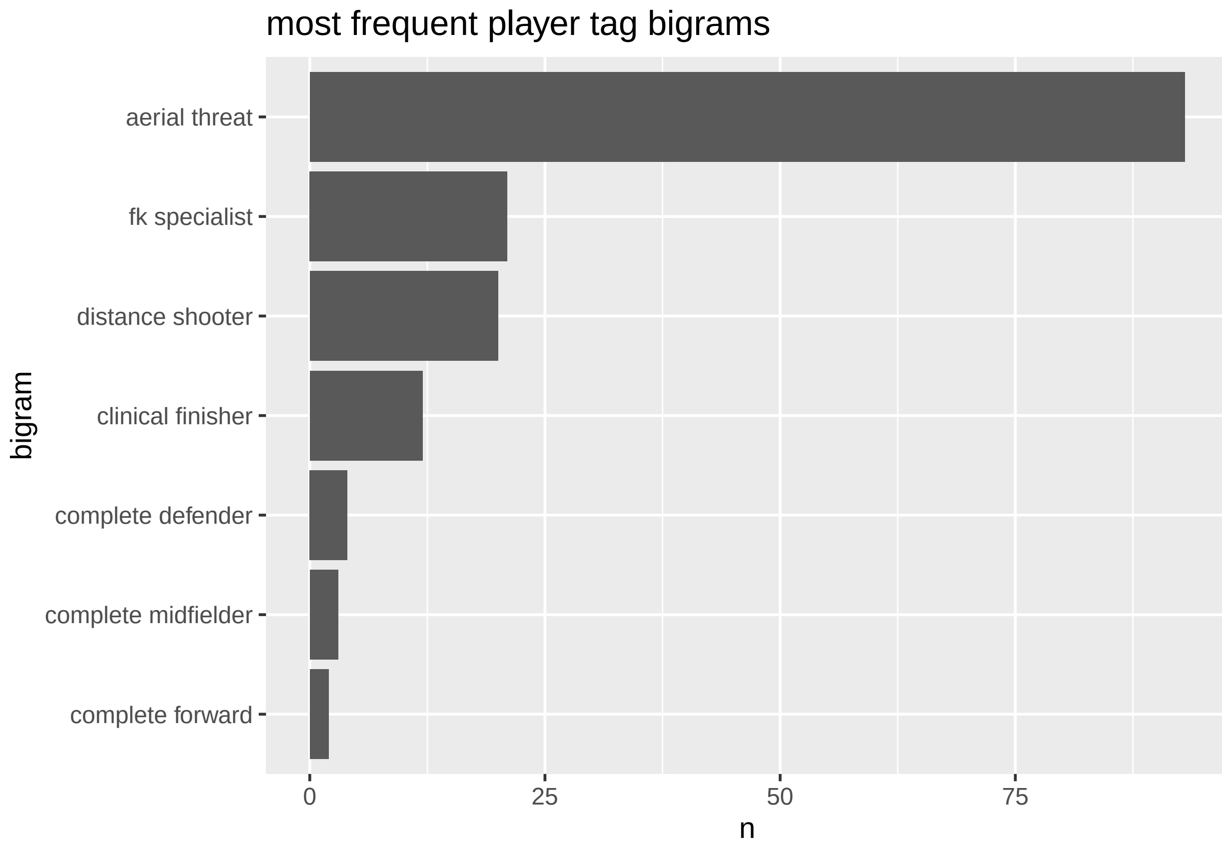

```
#####BIGRAMS
for (i in seq_along(fifa_datasets_list))
{
tidy_fifa20_bigrams <- fifa_datasets_list[[i]] %>%
  group_by(sofifa_id) %>%
  mutate(row_num = row_number())%>%
  ungroup() %>%
  unnest_tokens(bigram, player_tags, token = "ngrams", n=2) %>%
  select(sofifa_id, short_name, age,club, nationality, team_position, team_jersey_number,
         overall, potential, weak_foot, skill_moves, work_rate, pace, shooting, passing,
         dribbling, defending, physic, bigram, attacking_heading_accuracy,
         attacking_finishing)

#View(tidy_fifa20_bigrams)


#actual player tag bigrams:
actual_bigrams <- tibble(bigram = c('aerial threat','fk specialist','clinical finisher',
                                    'distance shooter','complete midfielder','complete forward','comple
tidy_fifa20_bigrams <- tidy_fifa20_bigrams %>%
  semi_join(actual_bigrams, by="bigram")


#Most common bigrams in player tags:
tags_bigrams <- tidy_fifa20_bigrams %>% count(bigram, sort = TRUE)
print(paste("Most common player tag bigrams in", years[[i]],":"))
```

```r
print(tags_bigrams)
#aerial threat is the most common bigram followed by fk specialist and distance shooter.
#Visualization:
tags_bigrams_plot <- tidy_fifa20_bigrams %>% count(bigram, sort = TRUE)%>%top_n(7)%>%
  mutate(bigram=reorder(bigram, n))%>%
  ggplot(aes(bigram,n))+geom_col()+coord_flip()+labs(x="bigram","frequency",
                                                     title = "most frequent player tag bigrams")
print(tags_bigrams_plot)

}
```

```
## [1] "Most common player tag bigrams in 2016 :"
## # A tibble: 7 x 2
##   bigram                  n
##   <chr>               <int>
## 1 aerial threat          95
## 2 distance shooter       24
## 3 fk specialist          20
## 4 clinical finisher      12
## 5 complete defender       5
## 6 complete forward        5
## 7 complete midfielder     1

## Selecting by n

## [1] "Most common player tag bigrams in 2017 :"
## # A tibble: 7 x 2
##   bigram                  n
##   <chr>               <int>
## 1 aerial threat          93
## 2 fk specialist          21
## 3 distance shooter       20
## 4 clinical finisher      12
## 5 complete defender       4
## 6 complete midfielder     3
## 7 complete forward        2

## Selecting by n
```

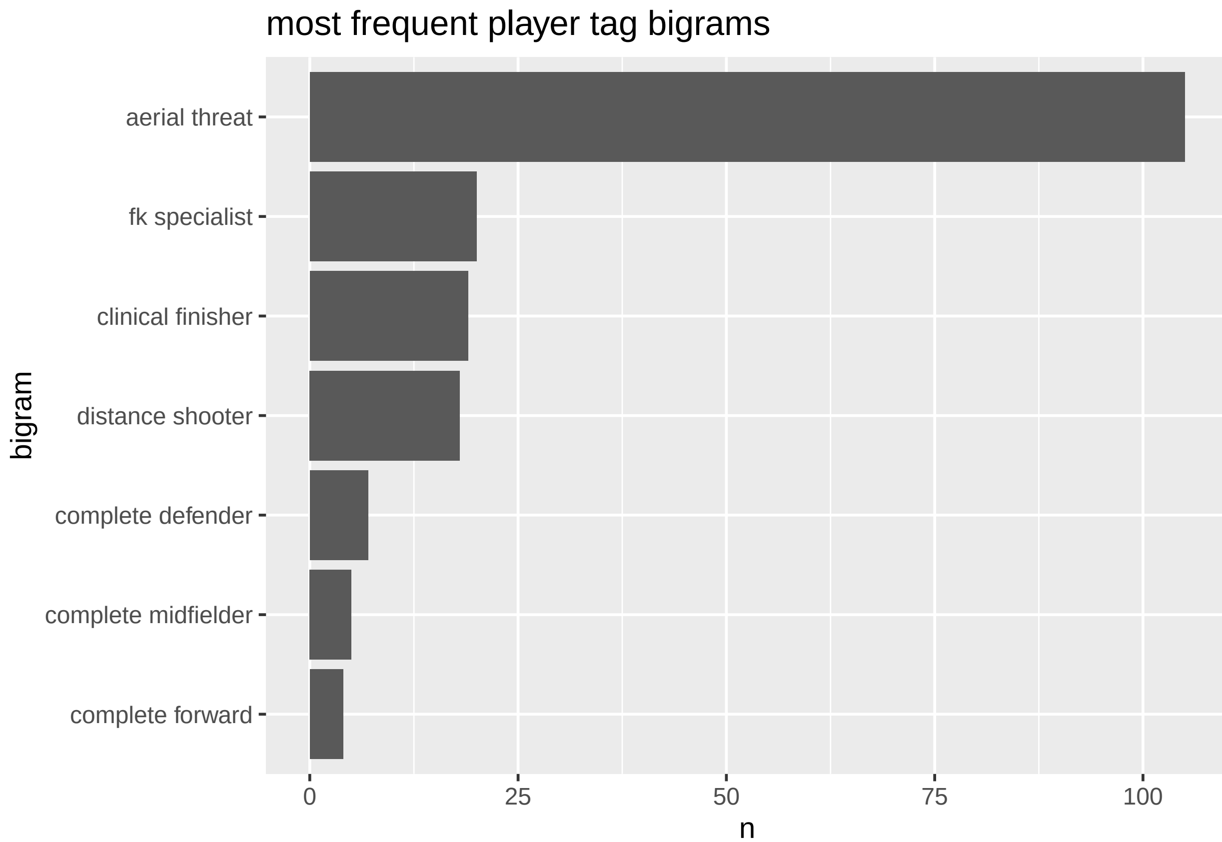



```
## [1] "Most common player tag bigrams in 2018 :"
## # A tibble: 7 x 2
##   bigram                  n
##   <chr>               <int>
## 1 aerial threat         105
## 2 fk specialist          20
## 3 clinical finisher      19
## 4 distance shooter       18
## 5 complete defender       7
## 6 complete midfielder     5
## 7 complete forward        4

## Selecting by n
```



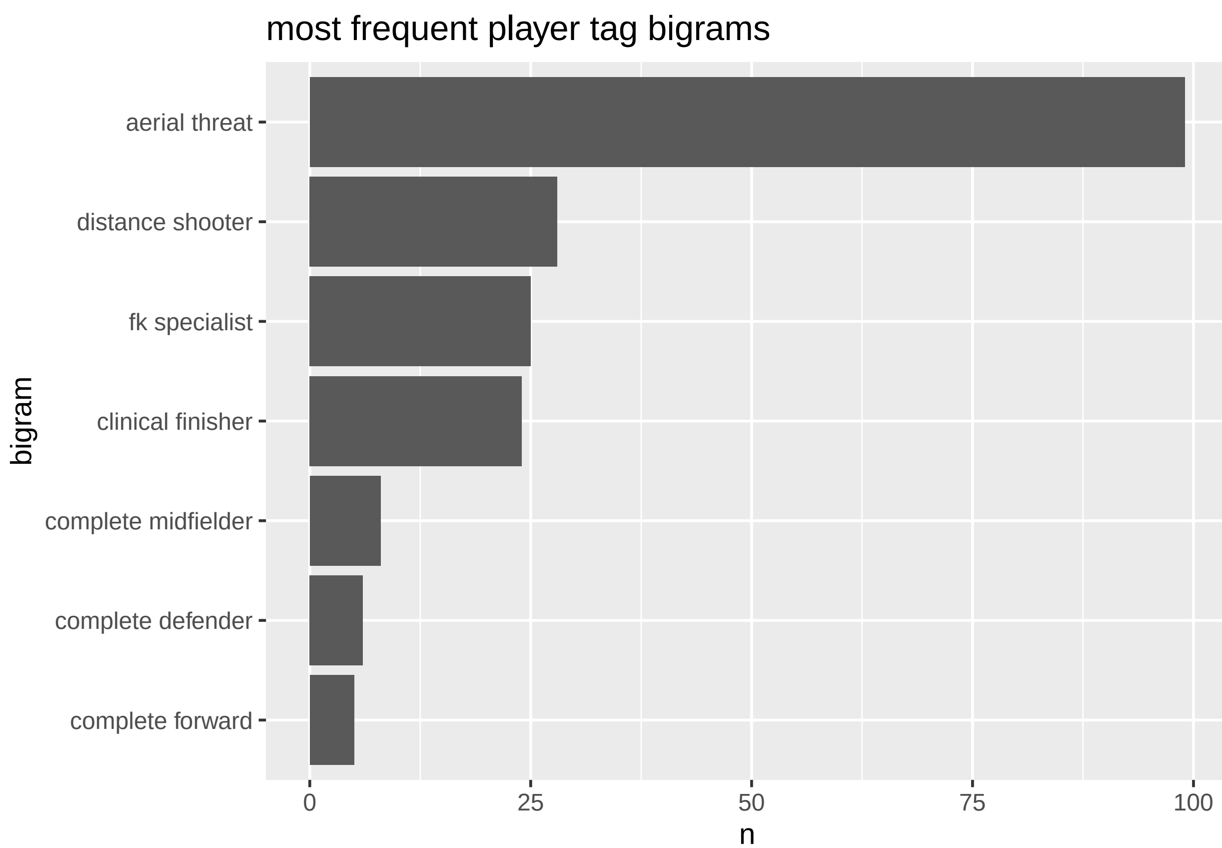

```
## [1] "Most common player tag bigrams in 2019 :"
## # A tibble: 7 x 2
##   bigram                  n
##   <chr>               <int>
## 1 aerial threat          99
## 2 distance shooter       28
## 3 fk specialist          25
## 4 clinical finisher      24
## 5 complete midfielder     8
## 6 complete defender       6
## 7 complete forward        5
```

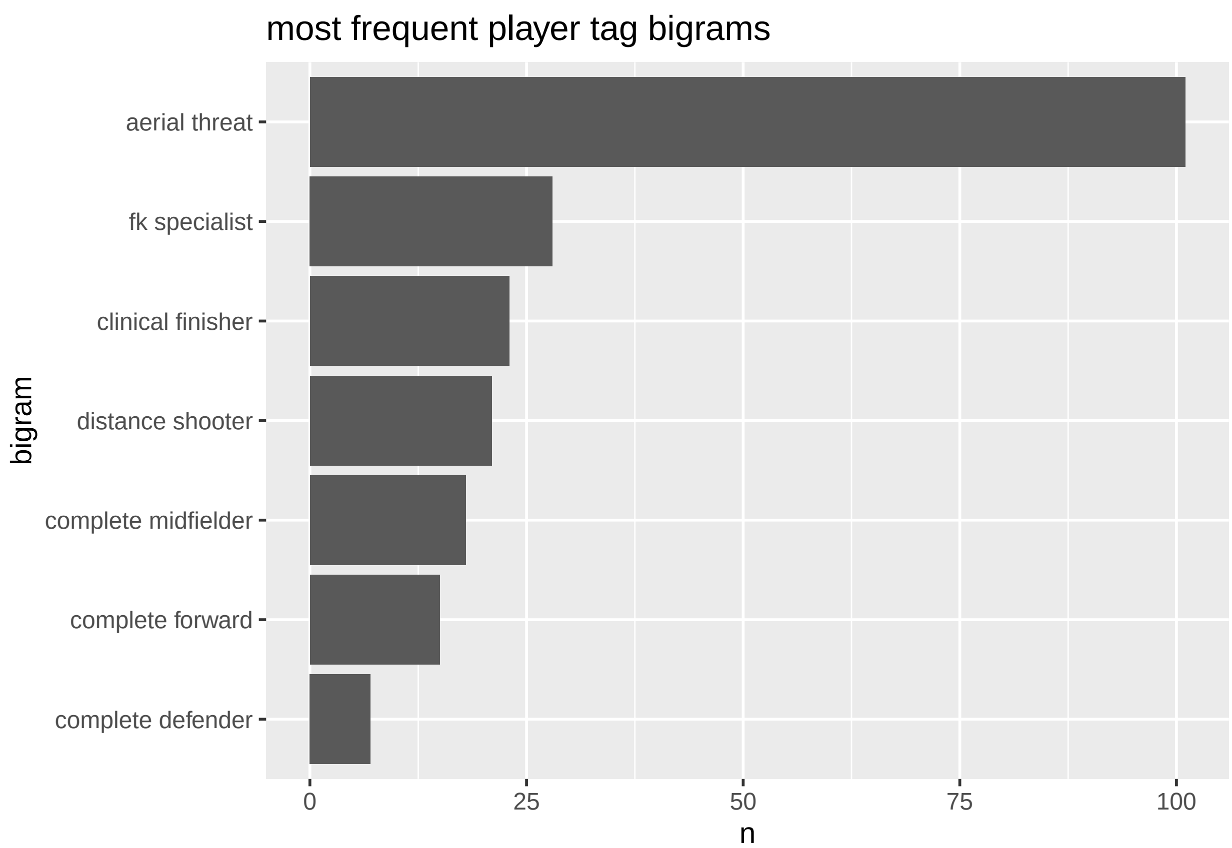
```
## Selecting by n
```




```
## [1] "Most common player tag bigrams in 2020 :"
## # A tibble: 7 x 2
##   bigram                  n
##   <chr>               <int>
## 1 aerial threat         101
## 2 fk specialist          28
## 3 clinical finisher      23
## 4 distance shooter       21
## 5 complete midfielder    18
## 6 complete forward       15
## 7 complete defender       7
```

```
## Selecting by n
```


most frequent player tag bigrams
aerial threat
distance shooter
fk specialist
clinical finisher
complete midfielder
complete defender
complete forward
bigram
0
25
50
75
100
n



Difference in heading abilities for pllayers based on aerial threat tag:

```
aerial_threats <- tidy_fifa20_bigrams %>% filter(bigram == "aerial threat")
not_aerial_threats <- tidy_fifa20_bigrams %>% filter(bigram != "aerial threat")

aerial_threats %>% summarise(avg_heading = mean(attacking_heading_accuracy))
```

```
## # A tibble: 1 x 1
##   avg_heading
##         <dbl>
## 1        80.3
```

```
#avg heading of 80
not_aerial_threats %>% summarise(avg_heading = mean(attacking_heading_accuracy))
```

```
## # A tibble: 1 x 1
##   avg_heading
##         <dbl>
## 1        67.1
```

```
#avg heading of 67.1
```

```r
#Clubs with the most aerial threats:
tidy_fifa20_bigrams %>% filter(bigram=="aerial threat") %>%
  count(club) %>% arrange(desc(n))
```

```
## # A tibble: 92 x 2
##    club                          n
##    <chr>                     <int>
##  1 Cagliari                      2
##  2 Cardiff City                  2
##  3 Celtic                        2
##  4 Chelsea                       2
##  5 Manchester United             2
##  6 Parma                         2
##  7 Shandong Luneng TaiShan FC    2
##  8 Sporting CP                   2
##  9 Viktoria PlzeÅˆ               2
## 10 1. FC KÃ¶ln                    1
## # ... with 82 more rows
```

Difference in shooting ability for pllayers based on fk specialist tag:

```r
tidy_fifa20_bigrams %>% filter(bigram=="fk specialist") %>%
  summarise(avg_shooting = mean(shooting))
```

```
## # A tibble: 1 x 1
##   avg_shooting
##          <dbl>
## 1         78.3
```

```r
#avg shooting 78
tidy_fifa20_bigrams %>% filter(bigram!="fk specialist") %>%
  summarise(avg_shooting = mean(shooting))
```

```
## # A tibble: 1 x 1
##   avg_shooting
##          <dbl>
## 1         68.2
```

```r
#avg shooting 68


##Clubs with the most fk specialists or clubs mostly likely to get a goal via a fk:
tidy_fifa20_bigrams %>% filter(bigram=="fk specialist") %>%
  count(club) %>% arrange(desc(n))
```

```
## # A tibble: 25 x 2
##    club                   n
##    <chr>              <int>
##  1 Juventus               2
```

```
##   2 Paris Saint-Germain     2
##   3 Real Madrid             2
##   4 Al Hilal                1
##   5 Atalanta                1
##   6 Borussia Dortmund       1
##   7 Chelsea                 1
##   8 Ecuador                 1
##   9 Everton                 1
##  10 FC Barcelona            1
## # ... with 15 more rows
```

```
#Juventus, PSG and Real Madrid.
```

Difference in finishing ability for pllayers based on clinical finisher tag:

```
tidy_fifa20_bigrams %>% filter(bigram=="clinical finisher") %>%
  summarise(avg_finishing = mean(attacking_finishing))
```

```
## # A tibble: 1 x 1
##   avg_finishing
##           <dbl>
## 1          88.6
```

```
#avg finishing 88.5
tidy_fifa20_bigrams %>% filter(bigram!="clinical finisher") %>%
  summarise(avg_finishing = mean(attacking_finishing))
```

```
## # A tibble: 1 x 1
##   avg_finishing
##           <dbl>
## 1          65.2
```

```
#avg finishing 65

#Clubs with the most clinical finishers:
tidy_fifa20_bigrams %>% filter(bigram=="clinical finisher") %>%
  count(club) %>% arrange(desc(n))
```

```
## # A tibble: 16 x 2
##    club                    n
##    <chr>               <int>
##  1 FC Barcelona            3
##  2 Borussia Dortmund       2
##  3 FC Bayern MÃ¼nchen      2
##  4 Juventus                2
##  5 Napoli                  2
##  6 Tottenham Hotspur       2
##  7 LA Galaxy               1
##  8 Lazio                   1
##  9 Liverpool               1
```

```
## 10 Manchester City         1
## 11 Paris Saint-Germain     1
## 12 RC Celta                1
## 13 Real Betis              1
## 14 Real Sociedad           1
## 15 Sampdoria               1
## 16 TSG 1899 Hoffenheim     1
```

```
#FC Barcelona
```

Player traits analysis:

```
player_traits <- fi20 %>% separate_rows(player_traits, convert = TRUE, sep = ", ")
View(player_traits)

#Most common player traits:
player_traits %>% filter(player_traits != "") %>%
  count(player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 27 x 2
##    player_traits                 n
##    <chr>                     <int>
##  1 Crowd Favourite            1336
##  2 Argues with Officials      1222
##  3 Early Crosser              1168
##  4 Selfish                    1106
##  5 Power Free-Kick             946
##  6 Leadership                  731
##  7 Long Passer (CPU AI Only)   710
##  8 Finesse Shot                684
##  9 Injury Prone                682
## 10 Giant Throw-in              552
## # ... with 17 more rows
```

```
#clubs with most crowd favs:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Crowd Favourite")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 545 x 3
##    club                player_traits       n
##    <chr>               <chr>           <int>
##  1 Manchester City     Crowd Favourite     9
##  2 Manchester United   Crowd Favourite     9
##  3 Real Madrid         Crowd Favourite     9
##  4 BeÅŸiktaÅŸ JK       Crowd Favourite     8
##  5 Celtic              Crowd Favourite     8
##  6 Independiente       Crowd Favourite     8
##  7 Paris Saint-Germain Crowd Favourite     8
##  8 RSC Anderlecht      Crowd Favourite     8
##  9 Sporting CP         Crowd Favourite     8
```

```
## 10 Ajax                Crowd Favourite     7
## # ... with 535 more rows
```

```
#Man Utd and Man City.

#clubs with players who argue with officials the most:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Argues with Officials")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 523 x 3
##    club                     player_traits             n
##    <chr>                    <chr>                 <int>
##  1 Juventus                 Argues with Officials    10
##  2 Napoli                   Argues with Officials     9
##  3 Borussia Dortmund        Argues with Officials     7
##  4 FC Bayern MÃ¼nchen       Argues with Officials     7
##  5 Fiorentina               Argues with Officials     7
##  6 Manchester United        Argues with Officials     7
##  7 Medipol BaÅŸakÅŸehir FK Argues with Officials     7
##  8 Real Madrid              Argues with Officials     7
##  9 Stade Brestois 29        Argues with Officials     7
## 10 Trabzonspor              Argues with Officials     7
## # ... with 513 more rows
```

```
#Juventus, Napoli, Dortmund and Bayern Munich.


#Clubs with the most selfish players:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Selfish")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 481 x 3
##    club                     player_traits     n
##    <chr>                    <chr>         <int>
##  1 Ajax                     Selfish          11
##  2 Manchester City          Selfish          10
##  3 Manchester United        Selfish          10
##  4 Paris Saint-Germain      Selfish          10
##  5 Juventus                 Selfish           9
##  6 Liverpool                Selfish           9
##  7 Leicester City           Selfish           8
##  8 Medipol BaÅŸakÅŸehir FK Selfish           8
##  9 Real Madrid              Selfish           8
## 10 Borussia Dortmund        Selfish           7
## # ... with 471 more rows
```

```
#Ajax, man utd and city, PSG.
#Do clubs with most selfish players have high wage bills?
fi20 %>% group_by(club)%>%
  summarise(sum_wage = sum(wage_eur)) %>% arrange(desc(sum_wage))
```

```
## # A tibble: 698 x 2
##    club                 sum_wage
##    <chr>                   <int>
##  1 Real Madrid           5354000
##  2 FC Barcelona          4950000
##  3 Manchester City       3984000
##  4 Juventus              3750000
##  5 Manchester United     2874000
##  6 Chelsea               2806000
##  7 Liverpool             2667000
##  8 Tottenham Hotspur     2603000
##  9 FC Bayern MÃ¼nchen    2516000
## 10 Paris Saint-Germain   2396000
## # ... with 688 more rows
```

```
#Man City, Man Utd and PSG are among the top 10 clubs who have the highest wage bill
#per week


#Clubs with the most injury prone players:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Injury Prone")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 400 x 3
##    club                   player_traits     n
##    <chr>                  <chr>         <int>
##  1 Manchester United      Injury Prone      7
##  2 Bayer 04 Leverkusen    Injury Prone      5
##  3 Eintracht Frankfurt    Injury Prone      5
##  4 Fortuna DÃ¼sseldorf    Injury Prone      5
##  5 Getafe CF              Injury Prone      5
##  6 Sheffield United       Injury Prone      5
##  7 1. FC Heidenheim 1846 Injury Prone      4
##  8 Ã-stersunds FK         Injury Prone      4
##  9 Arsenal                Injury Prone      4
## 10 AtlÃ©tico Madrid       Injury Prone      4
## # ... with 390 more rows
```

```
#Man Utd
#Since Man Utd is in the top 10 clubs with highest weekly wage bills, they tend to
#lose alot of money to players who dont play much due to injuries.

#Clubs with the most leaders:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Leadership")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 404 x 3
##    club                         player_traits     n
##    <chr>                        <chr>         <int>
##  1 KAA Gent                     Leadership        7
##  2 Manchester United            Leadership        7
```

```
##   3 SD Eibar                    Leadership        7
##   4 Getafe CF                   Leadership        6
##   5 Royal Antwerp FC            Leadership        6
##   6 San Jose Earthquakes        Leadership        6
##   7 Cardiff City                Leadership        5
##   8 Nottingham Forest           Leadership        5
##   9 Portland Timbers            Leadership        5
##  10 Shandong Luneng TaiShan FC Leadership        5
## # ... with 394 more rows
```

```
#KAA Gent, Man Utd and Eibar.

#Clubs with the most divers:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Diver")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 303 x 3
##    club                     player_traits     n
##    <chr>                    <chr>         <int>
##  1 Sheffield United        Diver             8
##  2 Vitesse                 Diver             8
##  3 Manchester United       Diver             7
##  4 Wolverhampton Wanderers Diver             6
##  5 Millwall                Diver             5
##  6 Birmingham City         Diver             4
##  7 Celtic                  Diver             4
##  8 Juventus                Diver             4
##  9 Olympique Lyonnais      Diver             4
## 10 Rangers FC              Diver             4
## # ... with 293 more rows
```

```
#Sheffield United and Vitesse

#clubs with most dribblers:Speed Dribbler (CPU AI Only)
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Speed Dribbler (CPU AI Only)")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 210 x 3
##    club                 player_traits                     n
##    <chr>                <chr>                         <int>
##  1 PSV                  Speed Dribbler (CPU AI Only)    11
##  2 FC Twente            Speed Dribbler (CPU AI Only)    10
##  3 Real Madrid          Speed Dribbler (CPU AI Only)     9
##  4 Fortuna Sittard      Speed Dribbler (CPU AI Only)     8
##  5 BeÅŸiktaÅŸ JK        Speed Dribbler (CPU AI Only)     7
##  6 Willem II            Speed Dribbler (CPU AI Only)     7
##  7 Bayer 04 Leverkusen  Speed Dribbler (CPU AI Only)     6
##  8 Borussia Dortmund    Speed Dribbler (CPU AI Only)     6
##  9 Monterrey            Speed Dribbler (CPU AI Only)     6
## 10 Tigres U.A.N.L.      Speed Dribbler (CPU AI Only)     6
## # ... with 200 more rows
```

```r
#PSV and FC Twente

#clubs with most injury free players:
player_traits %>% filter(player_traits != "") %>%
  filter(player_traits == "Injury Free")%>%
  count(club, player_traits) %>% arrange(desc(n))
```

```
## # A tibble: 286 x 3
##    club              player_traits     n
##    <chr>             <chr>         <int>
##  1 FC ZÃ¼rich        Injury Free       4
##  2 Levante UD        Injury Free       4
##  3 QuerÃ©taro        Injury Free       4
##  4 Sunderland        Injury Free       4
##  5 Alanyaspor        Injury Free       3
##  6 AS Nancy Lorraine Injury Free       3
##  7 BeÅŸiktaÅŸ JK     Injury Free       3
##  8 Crystal Palace    Injury Free       3
##  9 FC Porto          Injury Free       3
## 10 Huddersfield Town Injury Free       3
## # ... with 276 more rows
```

```r
#FC Zurich, Levante, Querataro and Sunderland
```

More EDA:

```r
fifa_top20_nations <- c("Belgium","France","Brazil","England","Uruguay","Croatia",
                        "Portugal","Spain","Argentina","Colombia","Mexico","Switzerland",
                        "Italy","Netherlands","Germany","Sweden","Chile",
                        "Poland","Senegal")

for (i in seq_along(fifa_datasets_list))
{

#Hypothesis: Of the top 20 ranked national teams, Spain has the best passers
passing_hyp <- fifa_datasets_list[[i]] %>% filter(nationality %in% fifa_top20_nations)%>%
  filter(!is.na(passing))%>%
  group_by(nationality) %>% summarise(avg_passing = mean(passing)) %>%
  arrange(desc(avg_passing))
print(paste("best passing country in",years[[i]],"is :"))
print(passing_hyp)
#Our hypothesis is true


#Hypothesis: of the top 20 ranked national teams, Brazil has the best dribblers
dribbling_hyp <- fifa_datasets_list[[i]] %>% filter(nationality %in% fifa_top20_nations)%>%
  filter(!is.na(dribbling))%>%
  group_by(nationality) %>% summarise(avg_dribbling = mean(dribbling)) %>%
  arrange(desc(avg_dribbling))
print(paste("best dribbling country in",years[[i]],"is :"))
print(dribbling_hyp)
```

```r
#Portugal has the best dribblers and Brazil is second. So, our hypothesis is
#not true.


#Hypothesis: of the top 20 ranked national teams, England has the fastest players:
fastest_hyp <- fifa_datasets_list[[i]] %>% filter(nationality %in% fifa_top20_nations)%>%
  filter(!is.na(pace))%>%
  group_by(nationality) %>% summarise(avg_pace = mean(pace)) %>%
  arrange(desc(avg_pace))
print(paste("fastest country in",years[[i]],"is :"))
print(fastest_hyp)
#Senegal has the fastest players, England is 6th so our hypothesis is not true.


#Hypothesis: of the top 20 ranked national teams, Italy has the best defenders
defen_hyp <- fifa_datasets_list[[i]] %>% filter(nationality %in% fifa_top20_nations)%>%
  filter(team_position %in% c("LB","RB","LWB","RWB","LCB","RCB","CDM","LDM","CDM"))%>%
  filter(!is.na(defending))%>%
  group_by(nationality) %>% summarise(avg_defending = mean(defending)) %>%
  arrange(desc(avg_defending))
print(paste("best defending country in",years[[i]],"is :"))
print(defen_hyp)
#Brazil has the best defenders, Italy is 7th, so our hypothesis is not true.

}
```

```
## [1] "best passing country in 2016 is :"
## # A tibble: 19 x 2
##    nationality avg_passing
##    <chr>             <dbl>
##  1 Brazil             61.7
##  2 Netherlands        61.0
##  3 Spain              60.9
##  4 Belgium            60.8
##  5 Portugal           60.6
##  6 Uruguay            59.1
##  7 Germany            59.1
##  8 France             58.7
##  9 Mexico             58.6
## 10 Senegal            58.0
## 11 Switzerland        57.6
## 12 Argentina          57.4
## 13 Croatia            57.4
## 14 Chile              56.7
## 15 Italy              56.6
## 16 Sweden             56.3
## 17 Poland             54.8
## 18 Colombia           53.2
## 19 England            53.1
## [1] "best dribbling country in 2016 is :"
## # A tibble: 19 x 2
##    nationality avg_dribbling
##    <chr>               <dbl>
```

```
##  1 Brazil               67.4
##  2 Portugal             66.2
##  3 Netherlands          65.8
##  4 Uruguay              64.9
##  5 Belgium              64.9
##  6 Spain                63.8
##  7 Germany              63.3
##  8 Senegal              63.2
##  9 Mexico               62.8
## 10 Argentina            62.5
## 11 France               62.4
## 12 Italy                62.1
## 13 Chile                62.0
## 14 Croatia              61.9
## 15 Switzerland          61.5
## 16 Sweden               60.5
## 17 Colombia             60.1
## 18 Poland               59.0
## 19 England              58.4
## [1] "fastest country in 2016 is :"
## # A tibble: 19 x 2
##    nationality avg_pace
##    <chr>          <dbl>
##  1 Brazil          70.3
##  2 Senegal         69.8
##  3 Switzerland     69.3
##  4 Uruguay         69.2
##  5 Portugal        69.0
##  6 Colombia        68.9
##  7 Belgium         68.5
##  8 Netherlands     68.4
##  9 Chile           68.4
## 10 Germany         68.3
## 11 England         68.2
## 12 Argentina       67.9
## 13 France          67.8
## 14 Spain           67.5
## 15 Italy           67.5
## 16 Mexico          67.2
## 17 Poland          67.0
## 18 Sweden          66.9
## 19 Croatia         66.6
## [1] "best defending country in 2016 is :"
## # A tibble: 19 x 2
##    nationality avg_defending
##    <chr>               <dbl>
##  1 Brazil               73.6
##  2 Uruguay              72.7
##  3 Portugal             71.2
##  4 Netherlands          69.8
##  5 Senegal              69.6
##  6 Germany              69.5
##  7 Italy                69.4
##  8 Spain                69.4
```

```
##  9 France                69.2
## 10 Belgium               69.1
## 11 Argentina             68.7
## 12 Mexico                67.2
## 13 Colombia              65.9
## 14 Croatia               65.9
## 15 Chile                 65.7
## 16 Switzerland           65.2
## 17 Sweden                65.2
## 18 Poland                64.5
## 19 England               63.4
## [1] "best passing country in 2017 is :"
## # A tibble: 19 x 2
##    nationality avg_passing
##    <chr>             <dbl>
##  1 Spain              61.9
##  2 Belgium            60.9
##  3 Portugal           60.6
##  4 Brazil             60.3
##  5 Netherlands        59.4
##  6 Germany            59.3
##  7 France             59.1
##  8 Mexico             58.7
##  9 Uruguay            58.4
## 10 Argentina          57.9
## 11 Croatia            57.8
## 12 Senegal            56.4
## 13 Italy              56.3
## 14 Switzerland        56.1
## 15 Sweden             56.1
## 16 Poland             55.8
## 17 Chile              55.6
## 18 Colombia           54.1
## 19 England            53.1
## [1] "best dribbling country in 2017 is :"
## # A tibble: 19 x 2
##    nationality avg_dribbling
##    <chr>               <dbl>
##  1 Portugal             66.8
##  2 Spain                64.9
##  3 Belgium              64.9
##  4 Brazil               64.6
##  5 Uruguay              64.4
##  6 Netherlands          64.2
##  7 Germany              63.6
##  8 France               63.2
##  9 Mexico               63.2
## 10 Argentina            63.1
## 11 Senegal              62.5
## 12 Croatia              62.3
## 13 Italy                62.0
## 14 Chile                61.1
## 15 Switzerland          60.3
## 16 Sweden               60.2
```

```
## 17 Colombia             59.9
## 18 Poland               59.8
## 19 England              58.8
## [1] "fastest country in 2017 is :"
## # A tibble: 19 x 2
##    nationality avg_pace
##    <chr>          <dbl>
##  1 Senegal         70.9
##  2 Portugal        69.6
##  3 Colombia        69.4
##  4 Belgium         68.9
##  5 Uruguay         68.9
##  6 England         68.7
##  7 France          68.6
##  8 Chile           68.3
##  9 Mexico          68.2
## 10 Brazil          68.2
## 11 Germany         68.1
## 12 Switzerland     67.8
## 13 Argentina       67.8
## 14 Netherlands     67.6
## 15 Italy           67.6
## 16 Poland          67.3
## 17 Sweden          66.7
## 18 Spain           66.5
## 19 Croatia         66.1
## [1] "best defending country in 2017 is :"
## # A tibble: 19 x 2
##    nationality avg_defending
##    <chr>               <dbl>
##  1 Portugal             70.9
##  2 Brazil               70.7
##  3 Netherlands          70.4
##  4 Spain                70.3
##  5 Germany              69.6
##  6 Senegal              69.3
##  7 Argentina            69.3
##  8 Italy                69.0
##  9 France               69.0
## 10 Uruguay              68.4
## 11 Croatia              68.4
## 12 Belgium              68.3
## 13 Mexico               66.9
## 14 Poland               65.9
## 15 Switzerland          65.6
## 16 Colombia             65.4
## 17 Chile                65.4
## 18 Sweden               64.7
## 19 England              64.2
## [1] "best passing country in 2018 is :"
## # A tibble: 19 x 2
##    nationality avg_passing
##    <chr>             <dbl>
##  1 Chile              62.5
```

```
##  2 Spain              61.6
##  3 Portugal           61.3
##  4 Brazil             61.2
##  5 Belgium            61.1
##  6 Netherlands        60.1
##  7 Uruguay            59.0
##  8 France             58.7
##  9 Argentina          58.2
## 10 Mexico             58.2
## 11 Croatia            58.2
## 12 Sweden             56.8
## 13 Germany            56.1
## 14 Italy              56.1
## 15 Poland             55.9
## 16 Senegal            55.4
## 17 Switzerland        55.1
## 18 Colombia           54.9
## 19 England            53.5
## [1] "best dribbling country in 2018 is :"
## # A tibble: 19 x 2
##    nationality avg_dribbling
##    <chr>               <dbl>
##  1 Portugal             67.6
##  2 Brazil               66.0
##  3 Chile                65.5
##  4 Spain                65.0
##  5 Netherlands          65.0
##  6 Belgium              64.8
##  7 Uruguay              64.1
##  8 Argentina            63.8
##  9 France               63.0
## 10 Croatia              63.0
## 11 Mexico               62.7
## 12 Senegal              62.5
## 13 Italy                62.0
## 14 Sweden               61.2
## 15 Germany              61.0
## 16 Colombia             60.7
## 17 Poland               60.1
## 18 Switzerland          59.6
## 19 England              59.4
## [1] "fastest country in 2018 is :"
## # A tibble: 19 x 2
##    nationality avg_pace
##    <chr>          <dbl>
##  1 Senegal         71.4
##  2 Portugal        69.5
##  3 Colombia        69.5
##  4 Brazil          68.9
##  5 Chile           68.8
##  6 Uruguay         68.5
##  7 England         68.5
##  8 Belgium         68.2
##  9 Argentina       68
```

```
## 10 Mexico          67.7
## 11 France          67.5
## 12 Netherlands     67.4
## 13 Germany         67.3
## 14 Sweden          66.9
## 15 Italy           66.7
## 16 Spain           66.7
## 17 Croatia         66.4
## 18 Poland          66.2
## 19 Switzerland     66.0
## [1] "best defending country in 2018 is :"
## # A tibble: 19 x 2
##    nationality avg_defending
##    <chr>               <dbl>
##  1 Brazil               71.4
##  2 Chile                70.6
##  3 Uruguay              70.5
##  4 Spain                70.1
##  5 Belgium              69.4
##  6 Netherlands          69.1
##  7 France               69.0
##  8 Portugal             68.9
##  9 Argentina            68.6
## 10 Italy                68.0
## 11 Mexico               66.7
## 12 Germany              66.6
## 13 Switzerland          66.5
## 14 Senegal              66.2
## 15 Croatia              66.2
## 16 Colombia             65.6
## 17 Sweden               64.6
## 18 England              64.0
## 19 Poland               62.7
## [1] "best passing country in 2019 is :"
## # A tibble: 19 x 2
##    nationality avg_passing
##    <chr>             <dbl>
##  1 Portugal           62.8
##  2 Belgium            62.1
##  3 Spain              61.9
##  4 Brazil             61.5
##  5 Netherlands        60.4
##  6 Uruguay            60.2
##  7 France             59.2
##  8 Argentina          58.6
##  9 Croatia            58.1
## 10 Mexico             57.7
## 11 Italy              57.5
## 12 Switzerland        57.3
## 13 Senegal            56.3
## 14 Sweden             56.2
## 15 Germany            56.2
## 16 Chile              56.1
## 17 Poland             54.9
```

```
## 18 Colombia           54.8
## 19 England            53.8
## [1] "best dribbling country in 2019 is :"
## # A tibble: 19 x 2
##    nationality avg_dribbling
##    <chr>               <dbl>
##  1 Portugal             68.9
##  2 Brazil               66.2
##  3 Belgium              65.9
##  4 Spain                65.6
##  5 Uruguay              65.5
##  6 Netherlands          65.1
##  7 Argentina            64.3
##  8 Senegal              64.3
##  9 France               63.7
## 10 Italy                63.3
## 11 Croatia              62.7
## 12 Mexico               62.1
## 13 Switzerland          61.4
## 14 Germany              61.4
## 15 Chile                61.0
## 16 Colombia             60.8
## 17 Sweden               60.5
## 18 England              59.5
## 19 Poland               59.3
## [1] "fastest country in 2019 is :"
## # A tibble: 19 x 2
##    nationality avg_pace
##    <chr>          <dbl>
##  1 Senegal         71.1
##  2 Portugal        69.9
##  3 Colombia        69.6
##  4 Belgium         68.8
##  5 Brazil          68.7
##  6 Switzerland     68.1
##  7 Uruguay         67.8
##  8 Netherlands     67.8
##  9 France          67.7
## 10 Germany         67.6
## 11 England         67.6
## 12 Chile           67.4
## 13 Poland          67.2
## 14 Argentina       67.2
## 15 Sweden          67.0
## 16 Italy           66.9
## 17 Mexico          66.6
## 18 Croatia         66.3
## 19 Spain           66.0
## [1] "best defending country in 2019 is :"
## # A tibble: 19 x 2
##    nationality avg_defending
##    <chr>               <dbl>
##  1 Brazil               72.2
##  2 Senegal              71.2
```

```
##  3 Belgium             70.6
##  4 Uruguay             70.2
##  5 Netherlands         69.8
##  6 Spain               69.7
##  7 Portugal            69.6
##  8 Croatia             69.5
##  9 France              68.7
## 10 Argentina           68.5
## 11 Italy               68.4
## 12 Germany             66.8
## 13 Mexico              65.7
## 14 Chile               65.3
## 15 Colombia            65.0
## 16 England             64.6
## 17 Sweden              64.4
## 18 Switzerland         64.2
## 19 Poland              64.0
## [1] "best passing country in 2020 is :"
## # A tibble: 19 x 2
##    nationality avg_passing
##    <chr>             <dbl>
##  1 Spain              62.4
##  2 Belgium            62.3
##  3 Portugal           61.9
##  4 Uruguay            61.7
##  5 Brazil             61.6
##  6 Netherlands        60.3
##  7 Argentina          59.9
##  8 France             59.0
##  9 Croatia            58.8
## 10 Mexico             58.8
## 11 Italy              57.7
## 12 Chile              56.9
## 13 Switzerland        56.6
## 14 Senegal            56.5
## 15 Germany            56.3
## 16 Sweden             55.9
## 17 Colombia           55.3
## 18 Poland             54.3
## 19 England            53.7
## [1] "best dribbling country in 2020 is :"
## # A tibble: 19 x 2
##    nationality avg_dribbling
##    <chr>               <dbl>
##  1 Portugal             68.0
##  2 Brazil               66.4
##  3 Belgium              66.3
##  4 Spain                66.2
##  5 Uruguay              66.0
##  6 Argentina            65.5
##  7 Netherlands          65.1
##  8 Senegal              64.2
##  9 Italy                63.9
## 10 France               63.7
```

```
## 11 Croatia              63.5
## 12 Mexico               63.3
## 13 Germany              61.8
## 14 Chile                61.7
## 15 Colombia             61.4
## 16 Switzerland          61.2
## 17 Sweden               60.3
## 18 England              59.5
## 19 Poland               59.0
## [1] "fastest country in 2020 is :"
## # A tibble: 19 x 2
##    nationality avg_pace
##    <chr>          <dbl>
##  1 Senegal         70.1
##  2 Colombia        68.9
##  3 Brazil          68.9
##  4 Portugal        68.8
##  5 Uruguay         68.0
##  6 England         67.8
##  7 Netherlands     67.7
##  8 Argentina       67.6
##  9 Switzerland     67.5
## 10 Germany         67.5
## 11 France          67.5
## 12 Belgium         67.4
## 13 Chile           67.2
## 14 Mexico          67.1
## 15 Italy           67.1
## 16 Poland          66.9
## 17 Sweden          66.4
## 18 Spain           66.3
## 19 Croatia         65.5
## [1] "best defending country in 2020 is :"
## # A tibble: 19 x 2
##    nationality avg_defending
##    <chr>               <dbl>
##  1 Brazil               71.9
##  2 Uruguay              71.2
##  3 Portugal             70.5
##  4 Belgium              70.4
##  5 Spain                69.7
##  6 Argentina            69.2
##  7 Italy                69.0
##  8 Senegal              68.6
##  9 Croatia              68.3
## 10 Netherlands          68.2
## 11 France               68.2
## 12 Switzerland          66.1
## 13 Germany              65.7
## 14 Chile                65.0
## 15 Mexico               64.7
## 16 Colombia             64.7
## 17 Poland               64.1
## 18 England              63.9
```

```
## 19 Sweden                  63.1
```