# Pinelli Gisler

## Gianni Pinelli and Marius Gisler

2 10 2020

#### **Install Packages**

```
library(dplyr)
library(tinytex)
library(ggplot2)
library(ggrepel)
library(gridExtra)
library(gganimate)
library(gifski)
library(av)
```

#### **Data Cleaning**

```
names(df)[names(df) == "NSA.Code"] <- "Nation"
names(df)[names(df) == "i..Rank"] <- "Rank"
names(df)[names(df) == "Year.of.Birth"] <- "Year_of_birth"
names(df)[names(df) == "Distance..m."] <- "Distance_in_meters"
names(df)[names(df) == "Race.Points"] <- "Race_points"
names(df)[names(df) == "Race.Day"] <- "Race_day"
names(df)[names(df) == "FIS.Code"] <- "FIS_code"
names(df)[names(df) == "Last.Name"] <- "Last_name"
names(df)[names(df) == "First.Name"] <- "First_name"
df$FIS_code <- as.factor(df$FIS_code)
df$Race_day <- as.factor(df$Race_day)
df$Race_day <- as.Date(df$Race_day, "%d.%m.%Y")
df$Race_year <- format(df$Race_day, "%y')
df$Race_year <- as.numeric(df$Race_year)</pre>
```

## Add Additional Columns

```
###Number of participations
df_part <- table(df$FIS_code)
df_part <- as.data.frame(df_part)
names(df_part)[names(df_part) == "Var1"] <- "FIS_code"
names(df_part)[names(df_part) == "Freq"] <- "Number_of_participations"
df <- left_join(df, df_part,by='FIS_code', keep=FALSE)</pre>
```

```
###Age of Rider at Race
df$year_of_race <- format(as.Date(df$Race_day, format="%d.%m.%Y"),"%Y")
df$year_of_race <- as.integer(df$year_of_race)</pre>
df$Age_of_rider <- df$year_of_race - df$Year_of_birth</pre>
###Number of wins
df_winner <- filter(df, df$Rank == 1)</pre>
df winner count <- table(df winner$FIS code)</pre>
df winner count <- as.data.frame(df winner count)</pre>
names(df_winner_count)[names(df_winner_count) == "Var1"] <- "FIS_code"</pre>
names(df_winner_count)[names(df_winner_count) == "Freq"] <- "Number_of_wins"</pre>
df <- left_join(df, df_winner_count,by='FIS_code', keep=FALSE)</pre>
###Number of wins/nation
df_winner_count_nation <- table(df_winner$Nation)</pre>
df_winner_count_nation <- as.data.frame(df_winner_count_nation)</pre>
names(df_winner_count_nation) [names(df_winner_count_nation) == "Var1"] <- "Nation"</pre>
names(df_winner_count_nation) [names(df_winner_count_nation) == "Freq"] <- "Number_of_wins_nation"</pre>
df <- left_join(df, df_winner_count_nation,by='Nation', keep=FALSE)</pre>
df$Number_of_wins_nation[is.na(df$Number_of_wins_nation)] <- 0</pre>
### Number of Participations / nation
nations <- table(df$Nation)</pre>
nations <- as.data.frame(nations)</pre>
names(nations)[names(nations) == "Var1"] <- "Nation"</pre>
names(nations) [names(nations) == "Freq"] <- "Number_of_participations_Nation"</pre>
df <- left_join(df, nations,by='Nation', keep=FALSE)</pre>
###Number of wins/ski
df_winner_ski <- table(df_winner$Ski)</pre>
df_winner_ski <- as.data.frame(df_winner_ski)</pre>
names(df_winner_ski)[names(df_winner_ski) == "Var1"] <- "Ski"</pre>
names(df_winner_ski)[names(df_winner_ski) == "Freq"] <- "Number_of_wins_ski"</pre>
df <- left_join(df, df_winner_ski,by='Ski', keep=FALSE)</pre>
df$Number_of_wins_ski[is.na(df$Number_of_wins_ski)] <- 0</pre>
### Number of Participations / ski
Ski <- table(df$Ski)
Ski <- as.data.frame(Ski)</pre>
names(Ski) [names(Ski) == "Var1"] <- "Ski"</pre>
names(Ski)[names(Ski) == "Freq"] <- "Number_of_participations_Ski"</pre>
df <- left_join(df, Ski,by='Ski', keep=FALSE)</pre>
###Number of podiums
df_podium <- filter(df, df$Rank <4)</pre>
df_podium <- table(df_podium$FIS_code)</pre>
df_podium <- as.data.frame(df_podium)</pre>
names(df_podium)[names(df_podium) == "Var1"] <- "FIS_code"</pre>
names(df_podium)[names(df_podium) == "Freq"] <- "Number_of_podiums"</pre>
df <- left_join(df, df_podium,by='FIS_code', keep=FALSE)</pre>
```

```
###Best Rank

df_rank = df %>%
    select(FIS_code, Rank, Last_name)

df_rank <- df_rank[order(df_rank$Rank),]

df_rank <- df_rank[!duplicated(df_rank$FIS_code), ]

names(df_rank)[names(df_rank) == "Rank"] <- "Best_Rank"

df_rank$Last_name <- NULL

df <- left_join(df, df_rank,by='FIS_code', keep=FALSE)

head(df)</pre>
```

```
##
     Rank Bib FIS code
                          Last_name First_name Year_of_birth Nation
                                                                           Time
                 510993
                                                           1983
                                                                   SUI 2:33.80
       13
           30
                           ALBRECHT
                                         Daniel
## 2
       27
           24
                 510993
                            ALBRECHT
                                         Daniel
                                                           1983
                                                                   SUI 2:33.74
## 3
       45
                 370031 ALESSANDRIA
                                                           1993
           51
                                         Arnaud
                                                                   MON 2:34.06
## 4
       64
          71
                 370031 ALESSANDRIA
                                                           1993
                                                                   MON 1:37.48
                                         Arnaud
## 5
       19
           42
                 104537
                          ALEXANDER
                                         Cameron
                                                           1997
                                                                   CAN 1:44.10
           21
                 194858
                                                           1994
                                                                   FRA 1:44.07
## 6
       15
                             ALLEGRE
                                            Nils
                                                        Ski
     Difference Distance_in_meters Race_points
                                                               Race_day Race_year
## 1
           1.82
                               53.01
                                            15.81
                                                     Atomic 2009-01-17
                                                                              2009
## 2
           3.34
                               97.33
                                            29.98
                                                     Atomic 2008-01-13
                                                                              2008
## 3
           5.70
                              157.98
                                            48.03
                                                    Salomon 2019-01-19
                                                                              2019
## 4
           4.82
                              132.61
                                           71.26
                                                    Unknown 2014-01-18
                                                                              2014
## 5
           1.57
                               44.49
                                            19.14 Rossignol 2020-01-18
                                                                              2020
## 6
           1.54
                               43.65
                                            18.77
                                                    Salomon 2020-01-18
                                                                              2020
     Number_of_participations year_of_race Age_of_rider Number_of_wins
## 1
                              2
                                        2009
                                                        26
                                                                          0
## 2
                                        2008
                                                         25
                                                                          0
## 3
                              2
                                        2019
                                                        26
                                                                          0
## 4
                              2
                                        2014
                                                         21
                                                                          0
## 5
                              1
                                        2020
                                                         23
                                                                          0
                              2
                                                         26
                                         2020
     Number_of_wins_nation Number_of_participations_Nation Number_of_wins_ski
## 1
                          6
                                                            88
                                                                                 1
## 2
                                                            88
                          6
                                                                                 1
## 3
                          0
                                                                                 4
                                                             2
                          0
                                                             2
                                                                                 0
## 4
                                                                                 2
## 5
                          0
                                                            45
## 6
                          0
                                                            76
     Number_of_participations_Ski Number_of_podiums Best_Rank
## 1
                                144
## 2
                                144
                                                     0
                                                               13
## 3
                                 68
                                                     0
                                                               45
## 4
                                                     0
                                 26
                                                               45
## 5
                                 91
                                                     0
                                                               19
## 6
                                 68
                                                     0
                                                               15
```

Add additional Dataframes

```
### Dataframe Podium
df_podium <- filter(df, df$Rank <4)

### Dataframe Winner

df_winner <- filter(df, df$Rank == 1)

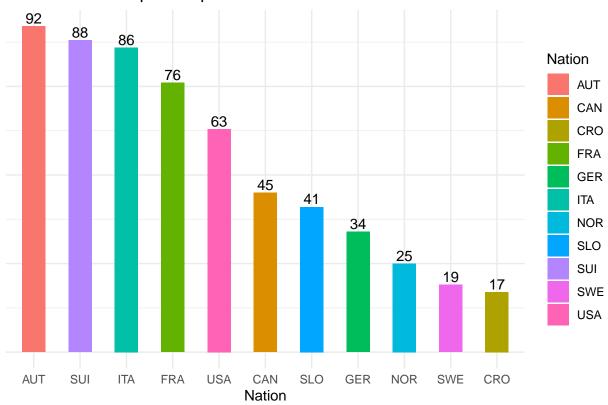
### Dataframe Top 10
df_top_10 <- filter(df, df$Rank <11)

### Riders with more than 5 Participations
df_participations <- filter(df, df$Number_of_participations >6)
df_participations <- df_participations[order(df_participations$Rank),]
df_participations <- df_participations[!duplicated(df_participations$FIS_code), ]</pre>
```

## Data Visualisation: Nations most participated

```
### Nations number of Participants
nations <- table(df$Nation)</pre>
nations <- as.data.frame(nations)</pre>
names(nations)[names(nations) == "Var1"] <- "Nation"</pre>
names(nations) [names(nations) == "Freq"] <- "Number_of_participations_Nation"</pre>
nations <- filter(nations, nations$Number_of_participations_Nation >9)
ggplot(data = nations, aes(x=reorder(Nation, -Number_of_participations_Nation), y=Number_of_participati
  theme_minimal()+
  geom_bar(stat="identity", width=0.5, aes(fill = Nation))+
  ggtitle("Number of Participations per Nation")+
 xlab("Nation")+
  ylab("Number of Participations")+
  geom_text(aes(label=Number_of_participations_Nation), position=position_dodge(width=0.9), vjust=-0.25
  theme(axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank())
```

## Number of Participations per Nation



Data Visualisation: Comparison participations, wins and wins per number of participations

```
df$Wins_per_Start_nation <- df$Number_of_wins_nation/df$Number_of_participations_Nation
nations <- filter(df, df$Number_of_participations_Nation >24)
plot_nations_1 <- ggplot(data = nations, aes(x=reorder(Nation, -Number_of_participations_Nation), y=Num</pre>
  theme_minimal()+
  geom_bar(stat="identity", width=0.5, position = "dodge", aes(fill = Nation), show.legend = FALSE)+
  ylab("Participations")+
  xlab("Nation")
plot_nations_2 <- ggplot(data = nations, aes(x=reorder(Nation, -Number_of_wins_nation), y=Number_of_win</pre>
  theme minimal()+
  geom_bar(stat="identity", width=0.5, position = "dodge", aes(fill = Nation), show.legend = FALSE)+
  ylab("Number of Wins")+
  xlab("Nation")
plot_nations_3 <- ggplot(data = nations, aes(x=reorder(Nation, -Wins_per_Start_nation), y=Wins_per_Start</pre>
  theme minimal()+
  geom_bar(stat="identity", width=0.5, position = "dodge", aes(fill = Nation), show.legend = FALSE)+
  ylab("Wins per Start")+
  xlab("Nation")+
  scale_y_continuous(labels = scales::percent)
```

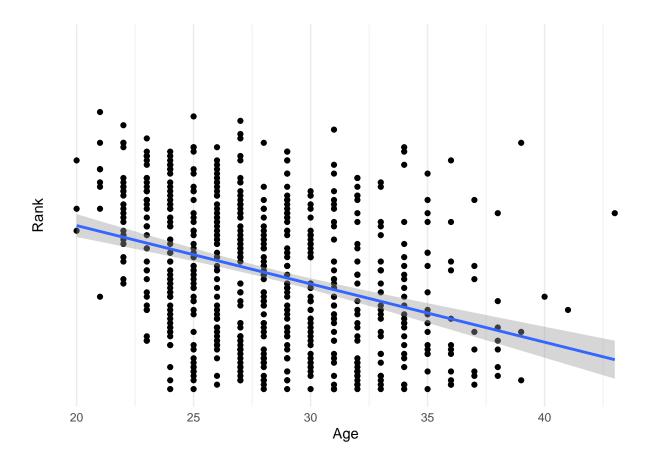


Austria has participated the most, however the Swiss won more races. Interestingly the Norwegians have a better rating in Wins divided by number of participations, this might indicate that the small norwegian team is strong or has had a lucky day by wining a race.

## Data Visualisation: Age and Rank over time

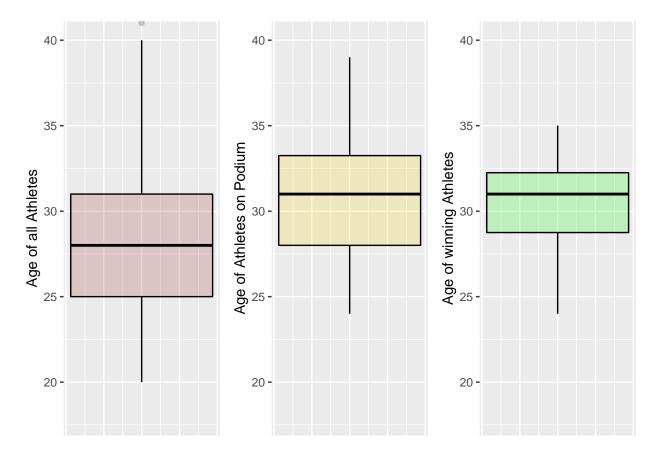
```
ggplot(data = df, aes(x=Age_of_rider, y=Rank)) +
  geom_point()+
  coord_cartesian(ylim = c(1, 80))+
  scale_y_reverse()+
  geom_smooth(method = "lm")+
  theme_minimal()+
  xlab("Age")+
  ylab("Rank")
```

## 'geom\_smooth()' using formula 'y ~ x'



#### Data Visualisation: Age Comparison

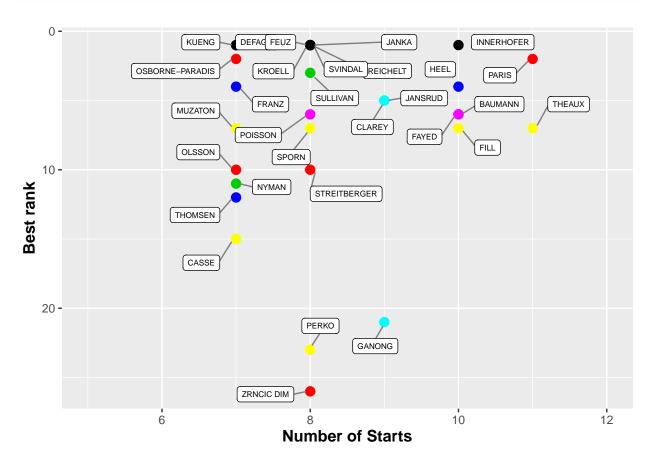
```
age_average <- ggplot(data = df, aes(y=Age_of_rider))+</pre>
 geom_boxplot(color="black", fill="brown", alpha=0.2)+
 ylab("Age of all Athletes")+
  coord_cartesian(ylim = c(18, 40))+
  theme(axis.title.x=element_blank(),
       axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
age_podium_average <- ggplot(data = df_podium, aes(y=Age_of_rider))+</pre>
  geom_boxplot(color="black", fill="gold", alpha=0.2)+
 ylab("Age of Athletes on Podium")+
  coord_cartesian(ylim = c(18, 40))+
  theme(axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
age_winner_average <- ggplot(data = df_winner, aes(y=Age_of_rider, fill = Age_of_rider))+
  geom_boxplot(color="black", fill="green", alpha=0.2)+
 ylab("Age of winning Athletes")+
 coord_cartesian(ylim = c(18, 40))+
```



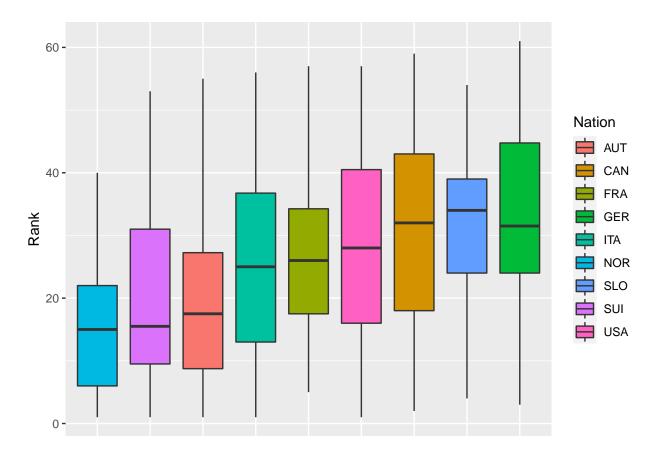
The age of the winning athletes and the athletes on the podium seem to be higher than the overall age of all athletes.

Data Visualisation: Athletes with more than 6 starts, what is the highest rank reached?

```
ylab("Best rank")+
xlab("Number of Starts")
```



Data Analyisis: How do the nations perform on average?



```
lm_nations <- lm(nations$Rank ~ nations$Nation)
summary(lm_nations)</pre>
```

```
##
## Call:
## lm(formula = nations$Rank ~ nations$Nation)
##
## Residuals:
##
             1Q Median
     Min
                           3Q
                                 Max
## -29.85 -11.18 -1.24 10.48 34.53
##
## Coefficients:
##
                    Estimate Std. Error t value Pr(>|t|)
                     20.4674
                                 1.4917 13.721 < 2e-16 ***
## (Intercept)
## nations$NationCAN
                     9.8882
                                 2.6027
                                          3.799 0.000162 ***
## nations$NationFRA
                     6.0458
                                          2.726 0.006618 **
                                 2.2178
## nations$NationGER 12.3855
                                 2.8715
                                          4.313 1.91e-05 ***
## nations$NationITA
                     5.0675
                                 2.1460
                                          2.361 0.018562 *
## nations$NationNOR -4.2274
                                 3.2270 -1.310 0.190744
## nations$NationSLO 11.1911
                                 2.6866
                                          4.166 3.62e-05 ***
## nations$NationSUI -0.8538
                                 2.1334 -0.400 0.689173
## nations$NationUSA
                     7.5485
                                 2.3397 3.226 0.001330 **
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
##
```

```
## Residual standard error: 14.31 on 541 degrees of freedom
## Multiple R-squared: 0.1002, Adjusted R-squared: 0.08693
## F-statistic: 7.534 on 8 and 541 DF, p-value: 1.466e-09
```

Data Analysiis: Are there correlations between Rank and Bib?

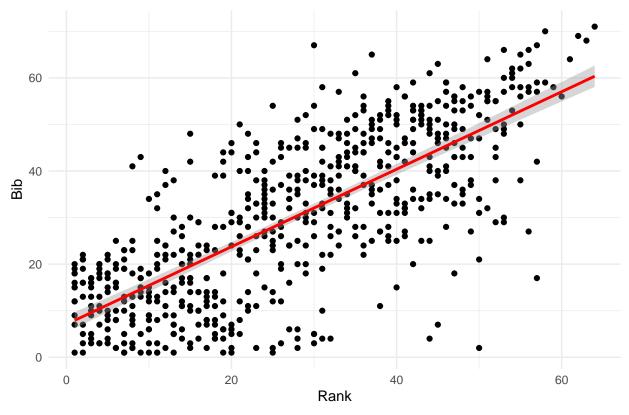
```
cor(df$Rank, df$Bib)
```

## ## [1] 0.7484561

```
ggplot(df, aes(x=Rank, y=Bib))+
  geom_point()+
  stat_smooth(method = 'lm', col = 'red')+
  ylab("Bib")+
  xlab("Rank")+
  ggtitle('Correlation between Rank and Bib')+
  theme_minimal()
```

## 'geom\_smooth()' using formula 'y ~ x'

# Correlation between Rank and Bib



```
lm.rank_1 <- lm(Rank ~ Bib, data = df)
summary(lm.rank_1)</pre>
```

```
##
## Call:
## lm(formula = Rank ~ Bib, data = df)
##
## Residuals:
##
               1Q Median
      Min
                               3Q
                                      Max
## -27.170 -6.914 -0.576
                            6.751 41.433
##
## Coefficients:
##
              Estimate Std. Error t value Pr(>|t|)
                                    8.817
## (Intercept) 7.22097
                          0.81897
                                            <2e-16 ***
               0.67323
                          0.02367 28.439
                                            <2e-16 ***
## Bib
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
## Residual standard error: 10.51 on 635 degrees of freedom
## Multiple R-squared: 0.5602, Adjusted R-squared: 0.5595
## F-statistic: 808.8 on 1 and 635 DF, p-value: < 2.2e-16
```

Data Analysis: Are there correlations between Rank and Number of Participations?

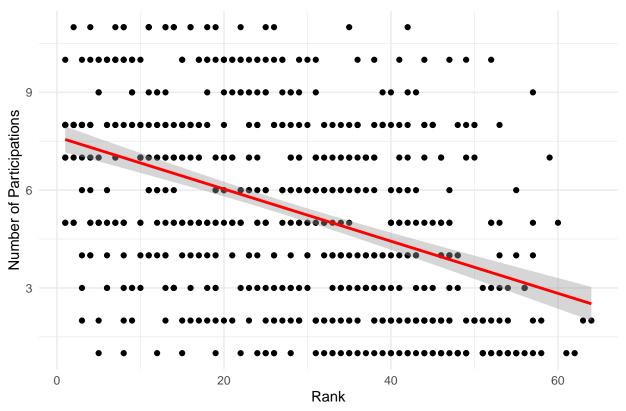
```
cor(df$Rank, df$Number_of_participations)

## [1] -0.4339373

ggplot(df, aes(x=Rank, y=Number_of_participations))+
    geom_point()+
    stat_smooth(method = 'lm', col = 'red')+
    ylab("Number of Participations")+
    xlab("Rank")+
    ggtitle('Correlation between Rank and Bib')+
    theme_minimal()
```

## 'geom\_smooth()' using formula 'y ~ x'

## Correlation between Rank and Bib



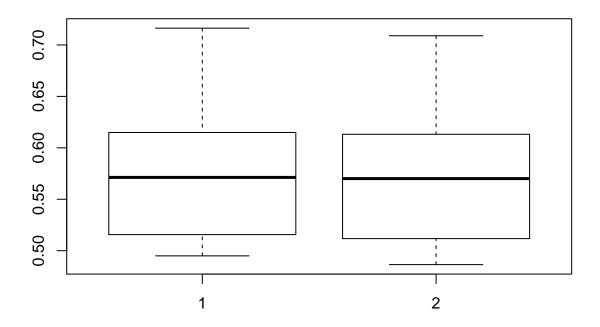
lm.rank\_2 <- lm(Rank ~ Number\_of\_participations, data = df)
summary(lm.rank\_2)</pre>

```
##
## Call:
## lm(formula = Rank ~ Number_of_participations, data = df)
##
## Residuals:
##
       Min
                1Q Median
                               ЗQ
                                      Max
## -32.752 -11.045
                    0.955
                           10.602 38.077
##
## Coefficients:
##
                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)
                            40.1058
                                        1.1993
                                                 33.44
                                                         <2e-16 ***
## Number_of_participations -2.3537
                                        0.1939 -12.14
                                                         <2e-16 ***
## ---
## Signif. codes: 0 '*** 0.001 '** 0.01 '* 0.05 '.' 0.1 ' 1
##
## Residual standard error: 14.28 on 635 degrees of freedom
## Multiple R-squared: 0.1883, Adjusted R-squared: 0.187
## F-statistic: 147.3 on 1 and 635 DF, p-value: < 2.2e-16
```

Comparison of models: Which one is better?

A 10n fold cross validation is applied and the R-squards are compared.

```
r.squared.simple <- c()</pre>
r.squared.few <- c()</pre>
set.seed(12)
for(i in 1:10){
  ## 1) prepare data
  smp_size <- floor(0.9*nrow(df))</pre>
  train <- sample(seq_len(nrow(df)), size = smp_size)</pre>
  df_train <- df[train,]</pre>
  df_test <- df[-train,]</pre>
  ## model 1
  ## 2) fit the model with "train" data
  lm.all <- lm(Rank ~ Bib + Number_of_participations, data = df)</pre>
  ## 3) make prediction on the test data
  predicted_lm_all <- predict(lm.all,</pre>
                                 newdata = df_test)
  ## 4) compute R^2
  r.squared.simple[i] <- cor(predicted_lm_all,</pre>
                                df_test$Rank)^2
  ## model 2
  ## 2) fit the model with "train" data
  lm.few <- lm(Rank ~ Bib, data = df_train)</pre>
  ## 3) make prediction on the test data
  predicted_lm_few <- predict(lm.few,</pre>
                                 newdata = df_test)
  ##
  ## 4) compute R^2
  r.squared.few[i] <- cor(predicted_lm_few,</pre>
                            df_test$Rank)^2
  }
mean_rsquared_lm_all <- mean(r.squared.simple)</pre>
mean_rsquared_lm_few <- mean(r.squared.few)</pre>
boxplot(r.squared.simple,r.squared.few)
```



```
print(mean_rsquared_lm_all)
## [1] 0.5770742
print(mean_rsquared_lm_few)
```

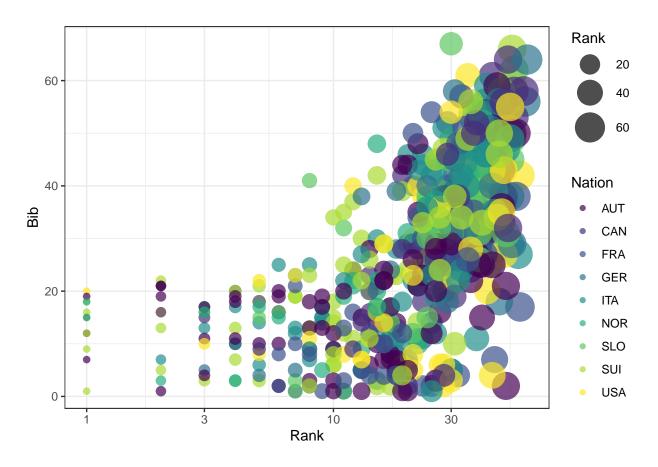
## [1] 0.5724064

The r-squared seem to do not differ significantly.

## Own Chapter

Application of gganimate, gifski, and av.

```
geom_point(show.legend = TRUE, alpha = 0.7) +
scale_color_viridis_d() +
scale_size(range = c(2, 10)) +
scale_x_log10() +
labs(x = "Rank", y = "Bib")
p
```



```
p + transition_time(Race_year) +
labs(title = "Year: {frame_time}")
```

