Code

* [Show All Code](#gjdgxs)
* [Hide All Code](#gjdgxs)

# Individual Project: Modeling with NHL Power Play Data

#### ***Arusyak Hakobyan***

#### ***December 5, 2018***

Imagine watching NHL team and they draw a powerplay, we think ice is free with the missing player and it will be easier to score. But how important is this seemingly advantageous situation in the NHL?

Setup

Loading some useful packages for analysis and plotting:

library(tidyverse)  
library(RCurl)  
library(broom)  
library(gridExtra)  
library(knitr)  
library(kableExtra)  
library(RColorBrewer)

Firstly, let’s analyze the NHL regular-season over the past eleven years (2007-2018) on the standard 4 vs. 5 power play. The data was retrieved from [Corsica](http://www.corsica.hockey/team_stats/).

data4v5 <- read.csv("nhl\_data.csv")

Let’s load the data and make sure that entries of the Atlanta Thrashers and the Winnipeg Jets are combined, as the Thrashers moved to Winnipeg after the 2011 season. Thus, the Thrashers and the Jets are considered the same team here.

# Combine ATL and WPG as ATL.WPG  
data4v5$Team <- plyr::revalue(data4v5$Team, c(ATL = 'ATL.WPG', WPG = 'ATL.WPG'))

For more interpretable regression model coefficents, let us represent the year as a continuous variable. This way an increase of one year is equivalent to one year of hockey played.

data4v5 <- data4v5 %>%   
 separate(Season, c('StartYear', 'EndYear'), sep = '-', remove = F) %>%  
 mutate\_at(c('StartYear', 'EndYear'), funs(as.numeric))

Next, we sort teams by NHL team division following the 2013-2014 realignment scheme.

# Initialize variables for storing team names  
pacific <- c('S.J', 'CGY', 'L.A', 'ANA', 'EDM', 'VAN', 'ARI')  
central <- c('ATL.WPG', 'NSH', 'STL', 'DAL', 'COL', 'MIN', 'CHI')  
metropolitan <- c('WSH', 'N.J', 'PHI', 'CBJ', 'PIT', 'NYR', 'NYI', 'CAR')  
atlantic <- c('T.B', 'BOS', 'TOR', 'DET', 'MTL', 'FLA', 'OTT', 'BUF')  
  
# Create column for identifying each team's division  
data4v5 <- data4v5 %>%  
 mutate(division = case\_when(data4v5$Team %in% pacific ~ 'Pacific',  
 data4v5$Team %in% central ~ 'Central',  
 data4v5$Team %in% metropolitan ~ 'Metropolitan',  
 data4v5$Team %in% atlantic ~ 'Atlantic'))

See the table detailing the 2013 realignment:

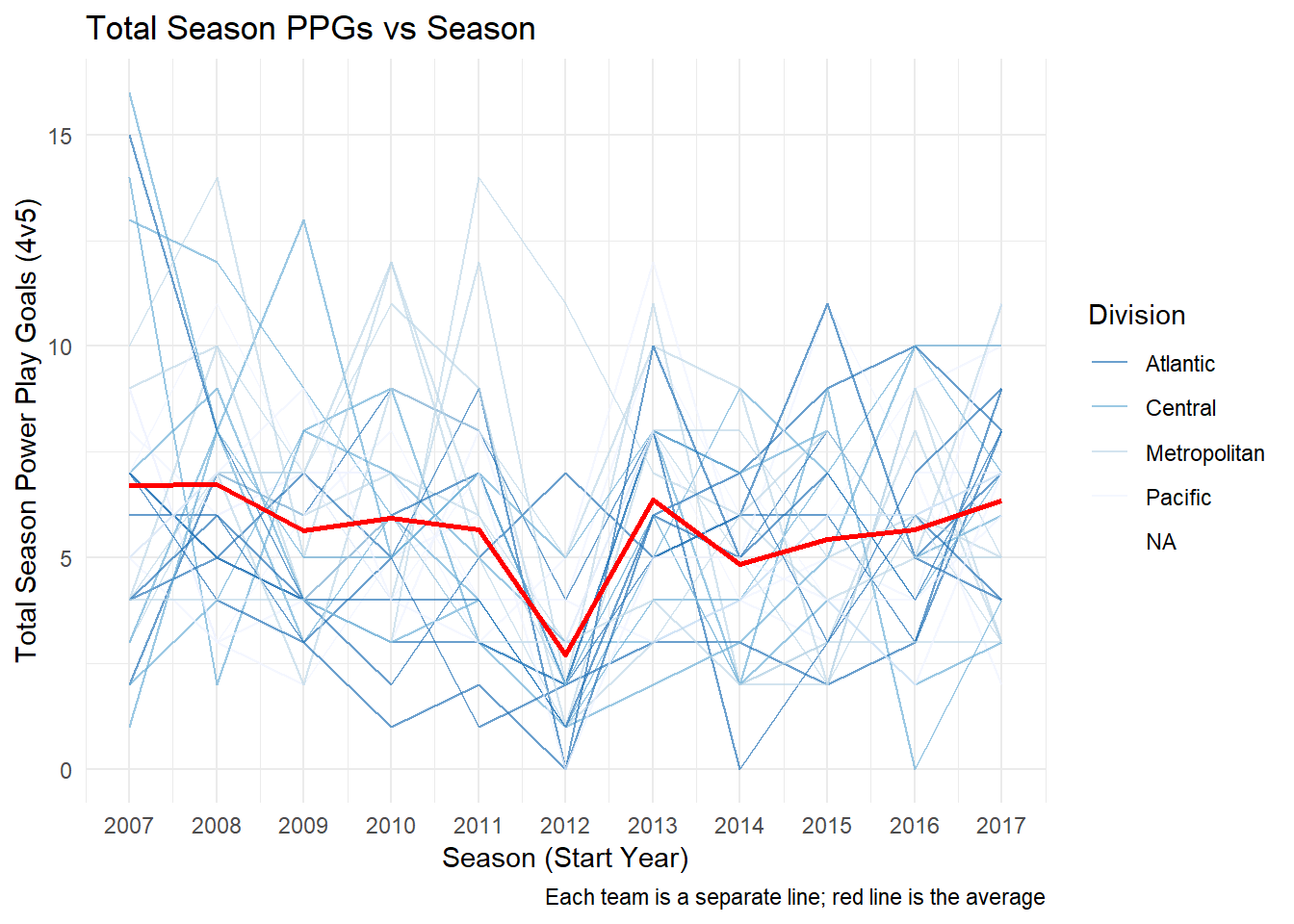
# Creates the data frame for html table  
divisions <- data.frame(Pacific = c(pacific, ''),  
 Central = c(central, ''),  
 Metropolitan = metropolitan,  
 Atlantic = atlantic)  
  
# Prints table  
kable(divisions) %>%  
 kable\_styling(bootstrap\_options = c('striped', 'hover', 'responsive'))

|  |  |  |  |
| --- | --- | --- | --- |
| Pacific | Central | Metropolitan | Atlantic |
| S.J | ATL.WPG | WSH | T.B |
| CGY | NSH | N.J | BOS |
| L.A | STL | PHI | TOR |
| ANA | DAL | CBJ | DET |
| EDM | COL | PIT | MTL |
| VAN | MIN | NYR | FLA |
| ARI | CHI | NYI | OTT |
|  |  | CAR | BUF |

Data Exploration

Before going into analysis, let us first explore the data via various graphs. Try to see if the amount of standard 4 vs. 5 regular season power play goals are changing over time across 11 years in the NHL

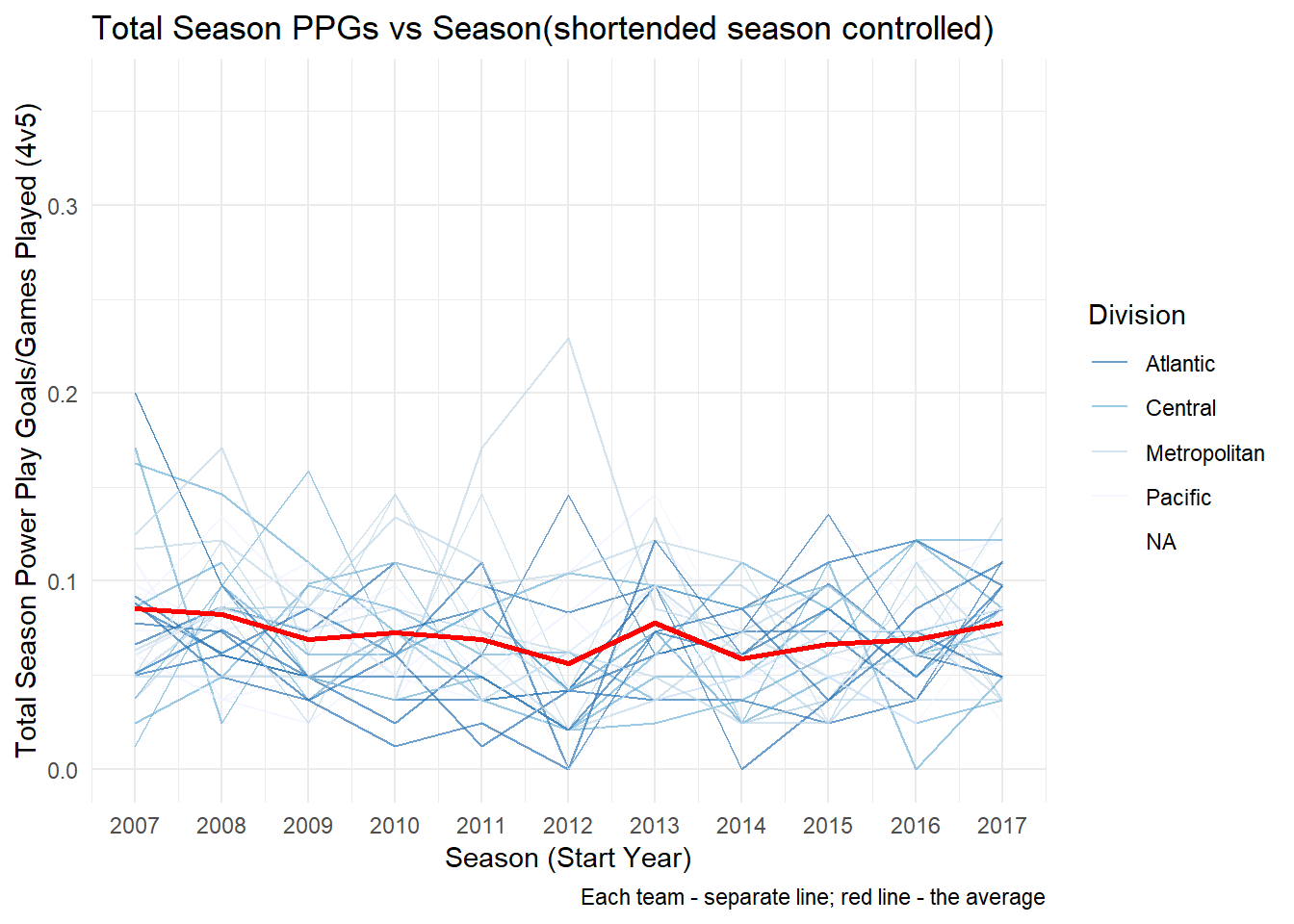
ggplot(data4v5, aes(StartYear, GF, group = 1)) +  
 geom\_line(aes(group = Team, color = division), alpha = 2/3) +   
 geom\_line(stat = 'summary', fun.y = 'mean', size = .9, color = 'red') +  
 scale\_color\_brewer(palette = 'Blues', direction = -1) +  
 labs(x = 'Season (Start Year)',  
 y = 'Total Season Power Play Goals (4v5)',  
 caption = 'Each team is a separate line; red line is the average') +  
 ggtitle("Total Season PPGs vs Season") +  
 guides(color = guide\_legend(title = 'Division')) +  
 scale\_x\_continuous(limits = c(2007, 2017), breaks = c(2007:2017)) +  
 theme\_minimal()



The plot shows a dramatic decrease in power play goals during the 2012-2013 season. This is a result of the shortened season, which would greatly reduce the amount of powerplay goal attempts, due to the 2012-2013 NHL lockout.

For controling this shortended season, let us divide the powerplay goals by the amount of games played per season:

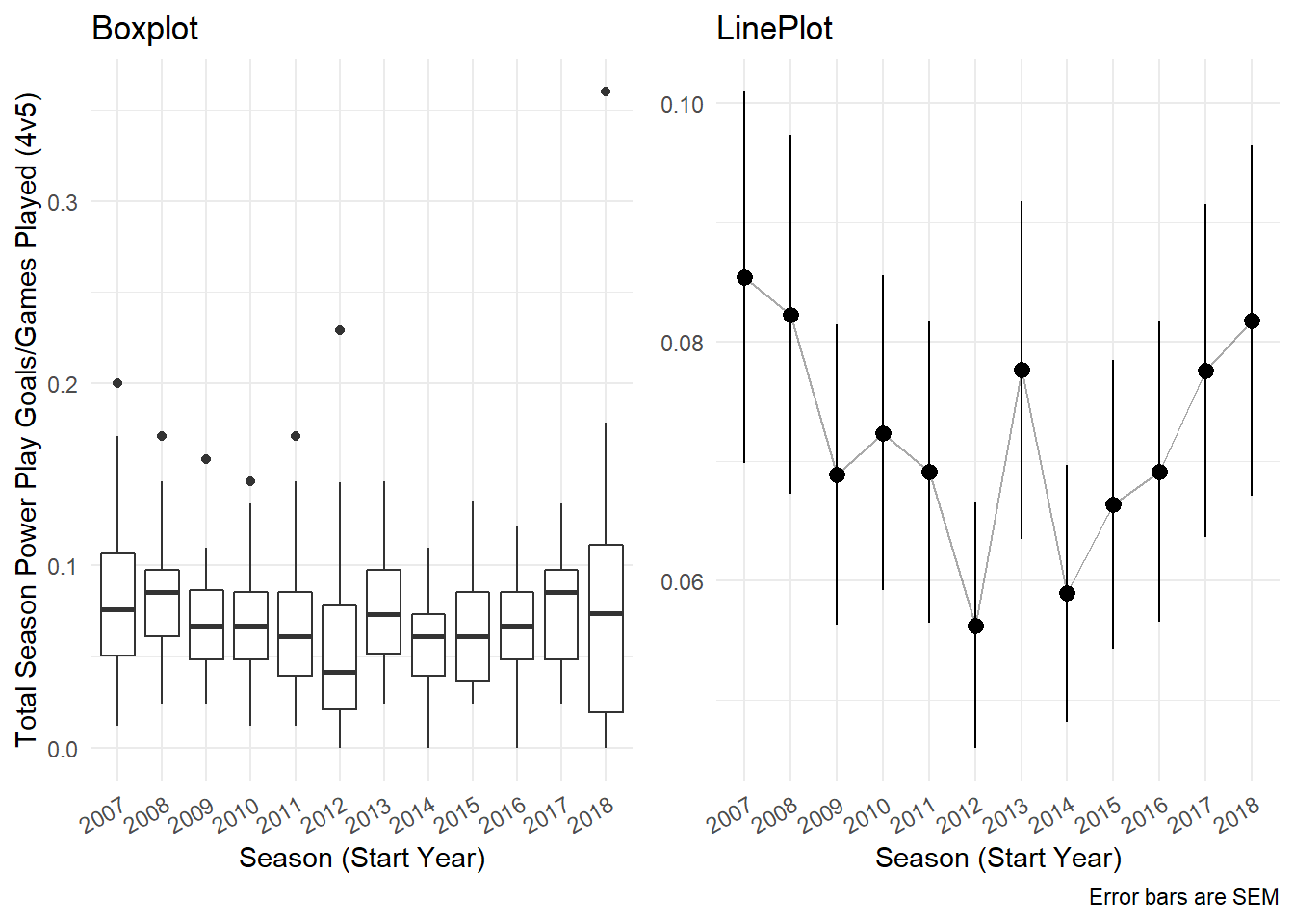
ggplot(data4v5, aes(StartYear, GF/GP, group = 1)) +  
 geom\_line(aes(group = Team, color = division), alpha = 2/3) +   
 geom\_line(stat = 'summary', fun.y = 'mean', size = .9, color = 'red') +  
 scale\_color\_brewer(palette = 'Blues', direction = -1) +  
 labs(x = 'Season (Start Year)',  
 y = 'Total Season Power Play Goals/Games Played (4v5)',  
 caption = 'Each team - separate line; red line - the average') +  
 ggtitle("Total Season PPGs vs Season(shortended season controlled)") +   
 guides(color = guide\_legend(title = 'Division')) +  
 scale\_x\_continuous(limits = c(2007, 2017), breaks = c(2007:2017)) +  
 theme\_minimal()



Now having controlled the number of games played per season, we can see that there seems to be a stable average powerplay goals scored per season for each subsequent year of the play.

For observing this in details, plot these data using boxplot and an average trend line:

boxplot <- ggplot(data4v5, aes(factor(StartYear), GF/GP, group = StartYear)) +  
 geom\_boxplot() +  
 labs(x = 'Season (Start Year)',  
 y = 'Total Season Power Play Goals/Games Played (4v5)',  
 caption = ' ') +  
 ggtitle("Boxplot") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 30, hjust = 1))  
  
data4v5\_sum <- data4v5 %>%   
 group\_by(StartYear) %>%   
 summarise(mean = mean(GF/GP),  
 sd = mean(GF/GP),  
 n = n(),  
 sem = sd/sqrt(n))  
  
lineplot <- ggplot(data4v5\_sum, aes(factor(StartYear), mean, group = 1)) +  
 geom\_line(alpha = 1/3) +  
 geom\_pointrange(aes(ymax = mean + sem, ymin = mean - sem)) +  
 labs(x = 'Season (Start Year)', caption = 'Error bars are SEM') +  
 ggtitle("LinePlot") +  
 theme\_minimal() +  
 theme(axis.title.y = element\_blank(), axis.text.x = element\_text(angle = 30, hjust = 1))  
  
   
grid.arrange(boxplot, lineplot, ncol = 2)



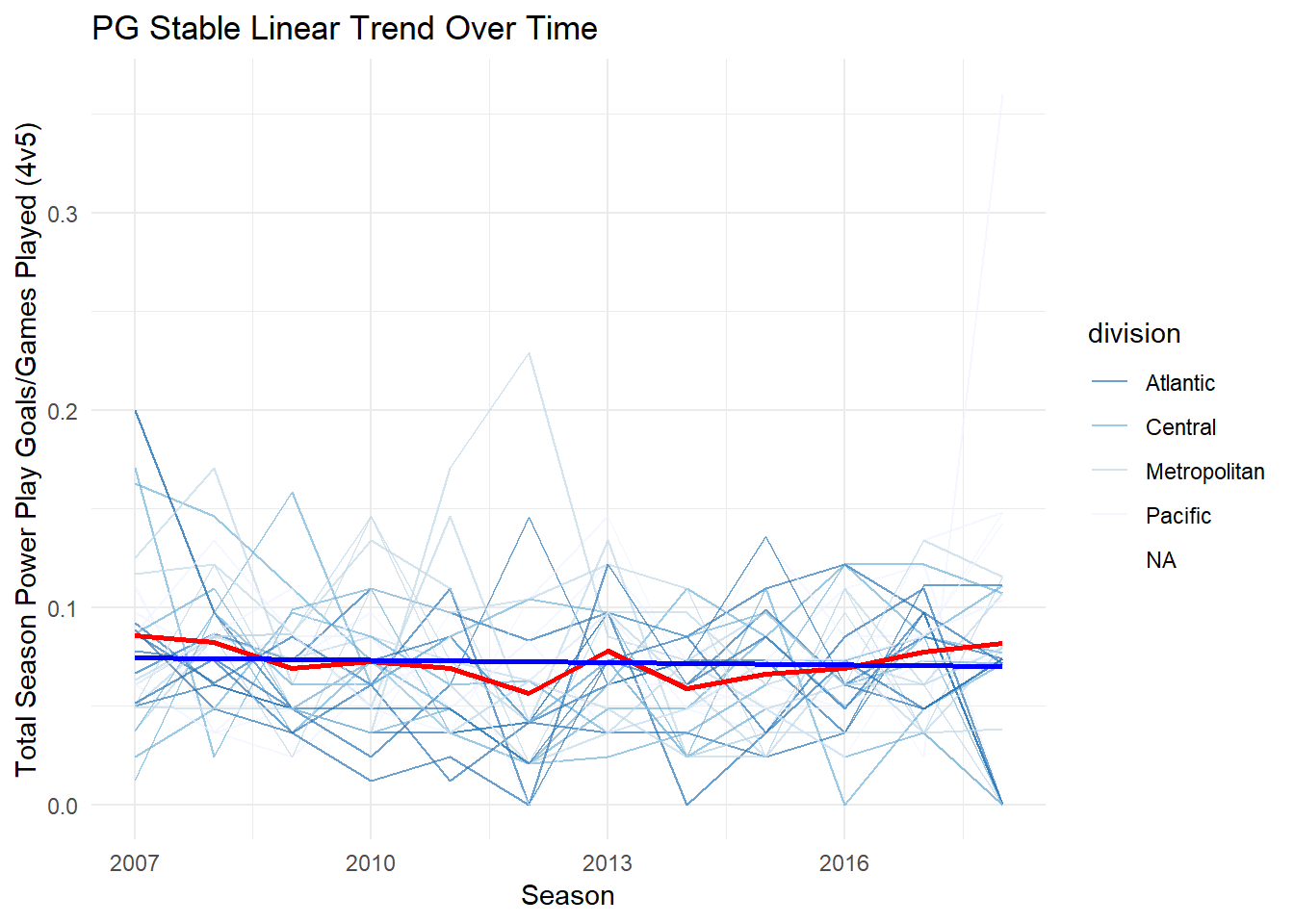
We now see each team’s power play performance controlled for the games played that season. Above plot demnstrates stable rate, that is clode to 0.07 in power play goals across time. Let’s model it with a linear regression, which predicts the total season power play goals controlled for games played as a function of time/season.

# Linearly model this relationship between power play goals/game and season  
mod <- lm((GF/GP) ~ factor(Season), data = data4v5)  
summary(mod) # Very significant negative trend over time

##   
## Call:  
## lm(formula = (GF/GP) ~ factor(Season), data = data4v5)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.081755 -0.027702 -0.004472 0.019919 0.278245   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.085397 0.007333 11.645 < 2e-16 \*\*\*  
## factor(Season)2008-2009 -0.003112 0.010371 -0.300 0.76427   
## factor(Season)2009-2010 -0.016487 0.010371 -1.590 0.11279   
## factor(Season)2010-2011 -0.013019 0.010371 -1.255 0.21018   
## factor(Season)2011-2012 -0.016291 0.010371 -1.571 0.11711   
## factor(Season)2012-2013 -0.029147 0.010371 -2.810 0.00522 \*\*   
## factor(Season)2013-2014 -0.007755 0.010371 -0.748 0.45512   
## factor(Season)2014-2015 -0.026454 0.010371 -2.551 0.01117 \*   
## factor(Season)2015-2016 -0.019041 0.010371 -1.836 0.06720 .   
## factor(Season)2016-2017 -0.016241 0.010371 -1.566 0.11824   
## factor(Season)2017-2018 -0.007821 0.010287 -0.760 0.44758   
## factor(Season)2018-2019 -0.003641 0.010287 -0.354 0.72356   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04017 on 350 degrees of freedom  
## Multiple R-squared: 0.04678, Adjusted R-squared: 0.01682   
## F-statistic: 1.562 on 11 and 350 DF, p-value: 0.1084

As the regression model and coefficients show, power play goals are changing slightly with each successive season. Plot this stable trend in blue to see more clearly.

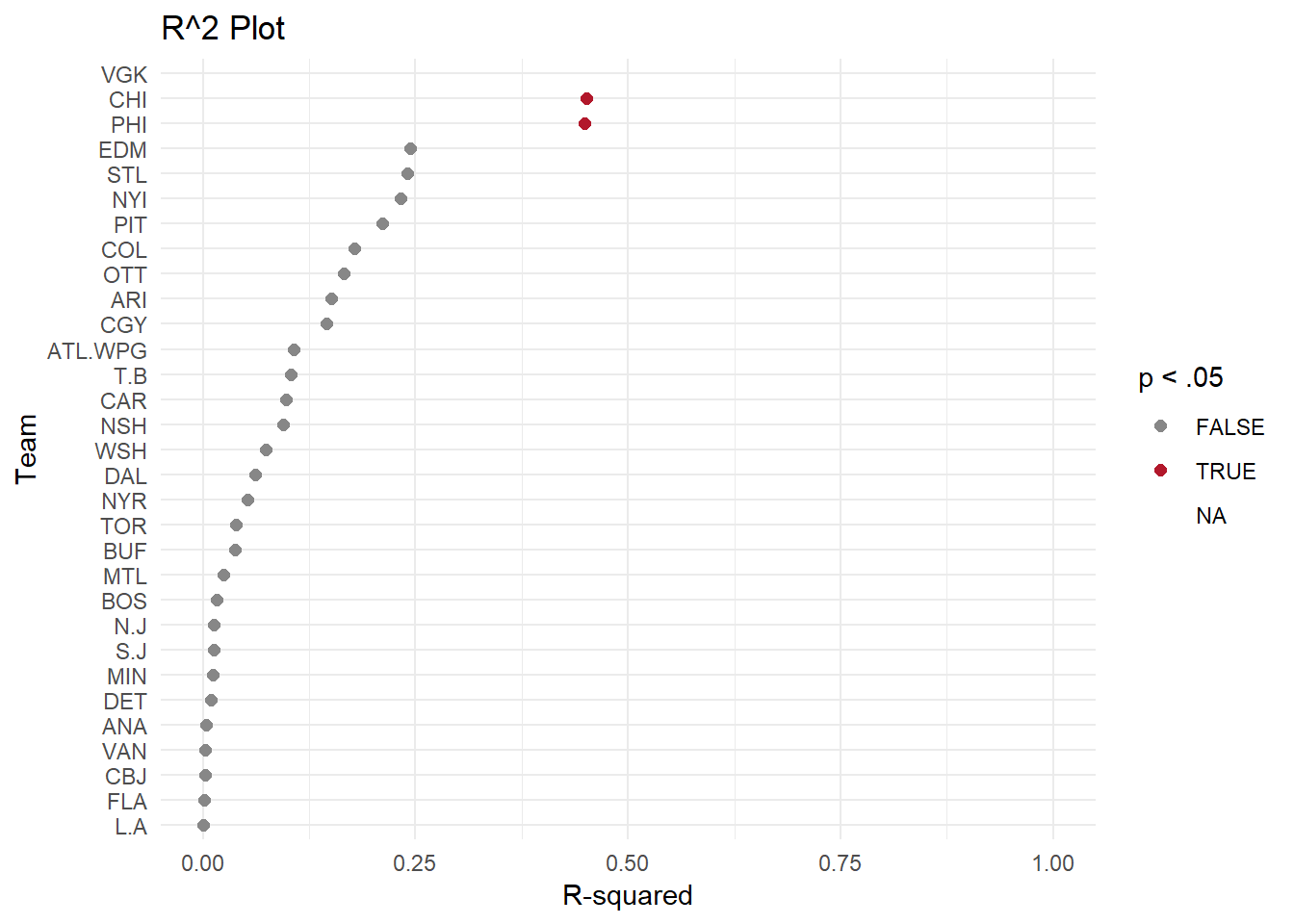
ggplot(data4v5, aes(StartYear, GF/GP, group = 1)) +  
 geom\_line(aes(group = Team, color = division), alpha = 2/3) +   
 geom\_line(stat = 'summary', fun.y = 'mean', size = .9, color = 'red') +  
 scale\_color\_brewer(palette = 'Blues', direction = -1) +  
 xlab('Season') +  
 ylab('Total Season Power Play Goals/Games Played (4v5)') +  
 ggtitle("PG Stable Linear Trend Over Time") +  
 stat\_smooth(method = 'lm', col = 'blue', se = F) +  
 theme(axis.text.x = element\_text(angle = 15, hjust = 1)) +  
 theme\_minimal()



Each team is going to be modeled individually using a linear regression for predicting powerplay goals per games played as a function of time (across 11 years). There are 30 teams and thus 30 distinct linear regressions will be performed, that is one for each team in the NHL.

We’ll use *R2* statistics to examine all 30 regression models simultaneously. These will allow us to account for the variability of power plays within each team.

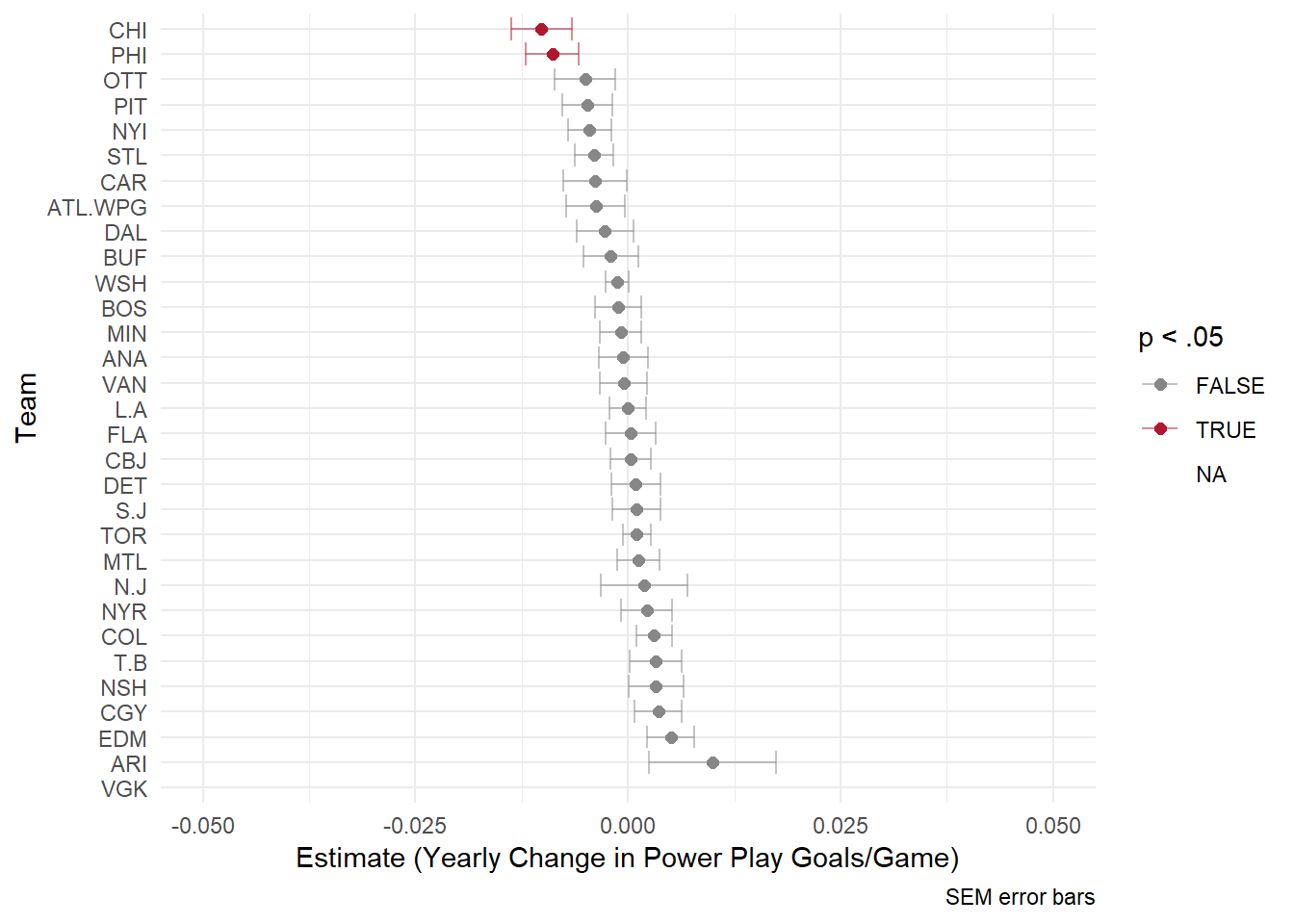
modTeam <- data4v5 %>%   
 group\_by(Team) %>%  
 do(level = lm((GF/GP) ~ StartYear, data = .))  
  
statistics <- modTeam %>% glance(level) %>% mutate(sig = p.value < .05)  
  
sigColorPal <- brewer.pal(11,'RdGy')  
  
ggplot(statistics, aes(r.squared, reorder(Team, r.squared), color = sig)) +  
 geom\_point(size = 2) +  
 scale\_color\_manual(values = sigColorPal[c(9,2)]) +  
 labs(x = 'R-squared', y = 'Team') +  
 ggtitle("R^2 Plot") +  
 guides(color = guide\_legend(title = 'p < .05')) +  
 theme\_minimal()



# Extracting coefficients  
levelCoef <- modTeam %>%   
 tidy(level) %>%   
 filter(term == 'StartYear') %>% # Facilitates plotting  
 mutate(sig = p.value < .05) # For later plotting

Now let’s plot these coefficients ordered by size:

xax.lock <- c(-.05, .05)  
ggplot(levelCoef, aes(estimate, reorder(Team, -1\*estimate), color = sig)) +  
 geom\_point(size = 2) +  
 geom\_errorbarh(aes(xmin = estimate - std.error,   
 xmax = estimate + std.error),  
 alpha = 1/2) +  
 scale\_color\_manual(values = sigColorPal[c(9,2)]) +  
 coord\_cartesian(xlim = xax.lock) +  
 labs(x = 'Estimate (Yearly Change in Power Play Goals/Game)',   
 y = 'Team',  
 caption = 'SEM error bars') +  
 guides(color = guide\_legend(title = 'p < .05')) +  
 theme\_minimal()



As the graph above illustrates each team had a different predicted rate of change in powerplay goals per games. The x-axis represent the regression coefficient of time. For example, the Arizona Coyotes has an estimate close to 0.015. This means that for every increase in 1 year (every season) the model predicted a decrease of -0.015 powerplay goals per games played. This estimate is colored grey, which means it was not found to be significantly different from zero.