Midterm Project

Christopher Flores and Zack Chipperfield

#Library that we used for creating some of the histograms  
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5 v purrr 0.3.4  
## v tibble 3.1.4 v dplyr 1.0.7  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 2.0.1 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

Reading in the data set from Github:

data\_url <- 'https://raw.githubusercontent.com/florescss14/STAT1341\_Midterm\_Project/main/Bundesliga.csv'  
bund <- readr::read\_csv(data\_url, col\_names = TRUE)

## Rows: 72 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): Season, Club  
## dbl (27): MP, W, D, L, GF, GA, GD, Pts, Total\_Comp, Total\_Att, Total\_Cmp\_Pct...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data\_url <- 'https://raw.githubusercontent.com/florescss14/STAT1341\_Midterm\_Project/main/La\_Liga.csv'  
liga <- readr::read\_csv(data\_url, col\_names = TRUE)

## Rows: 80 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): Season, Club  
## dbl (27): MP, W, D, L, GF, GA, GD, Pts, Total\_Comp, Total\_Att, Total\_Cmp\_Pct...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data\_url <- 'https://raw.githubusercontent.com/florescss14/STAT1341\_Midterm\_Project/main/Ligue\_1.csv'  
ligue <- readr::read\_csv(data\_url, col\_names = TRUE)

## Rows: 80 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): Season, Club  
## dbl (27): MP, W, D, L, GF, GA, GD, Pts, Total\_Comp, Total\_Att, Total\_Cmp\_Pct...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data\_url <- 'https://raw.githubusercontent.com/florescss14/STAT1341\_Midterm\_Project/main/Premier\_League.csv'  
epl <- readr::read\_csv(data\_url, col\_names = TRUE)

## Rows: 80 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): Season, Club  
## dbl (27): MP, W, D, L, GF, GA, GD, Pts, Total\_Comp, Total\_Att, Total\_Cmp\_Pct...

##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

data\_url <- 'https://raw.githubusercontent.com/florescss14/STAT1341\_Midterm\_Project/main/Serie\_A.csv'  
serie <- readr::read\_csv(data\_url, col\_names = TRUE)

## Rows: 80 Columns: 29

## -- Column specification --------------------------------------------------------  
## Delimiter: ","  
## chr (2): Season, Club  
## dbl (27): MP, W, D, L, GF, GA, GD, Pts, Total\_Comp, Total\_Att, Total\_Cmp\_Pct...

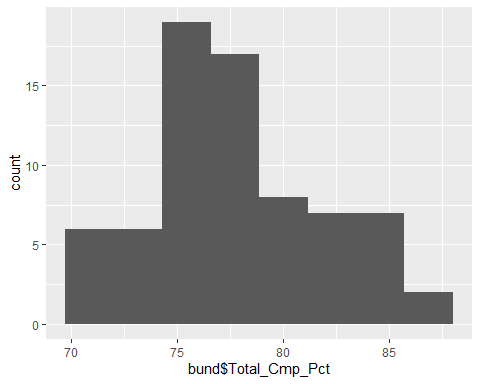
##   
## i Use `spec()` to retrieve the full column specification for this data.  
## i Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

As some leagues have a different number of matches played we want to normalize it to number of points scored per match in each league.

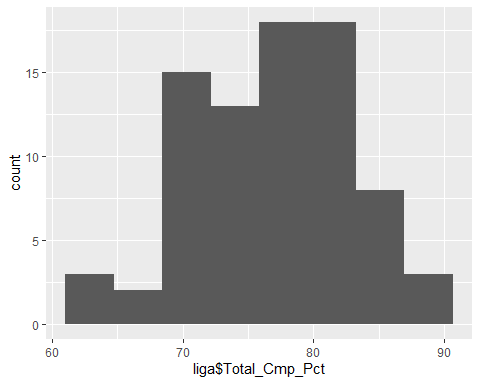
#Creating points per match from each league dataset  
bund$Pts.MP <- bund$Pts / bund$MP  
liga$Pts.MP <- liga$Pts / liga$MP  
ligue$Pts.MP <- ligue$Pts / ligue$MP  
epl$Pts.MP <- epl$Pts / epl$MP  
serie$Pts.MP <- serie$Pts / serie$MP  
  
#Binding all of the leagues into one dataset called all for creating the model later.  
all <- rbind(bund, liga, ligue, epl, serie)

Figure 1: Histograms of Total Completion Percentage of All Passes for Clubs from all five major leagues from 2017/18 to 2020/21 season

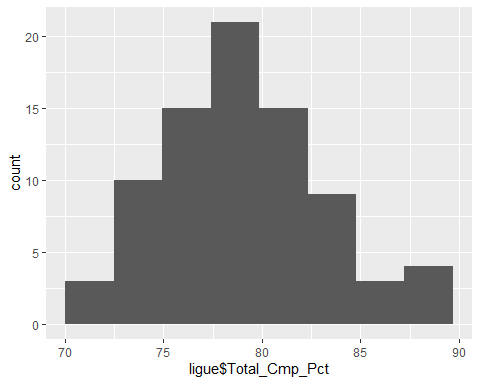
#ggplot is from the tidyverse package and makes a plot  
#geom\_histogram makes a histogram with the number of designated bins  
ggplot() +  
geom\_histogram(mapping = aes(x = bund$Total\_Cmp\_Pct), bins = 8)



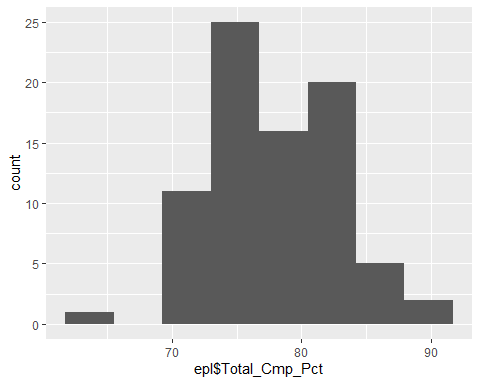
ggplot() +  
geom\_histogram(mapping = aes(x = liga$Total\_Cmp\_Pct), bins = 8)



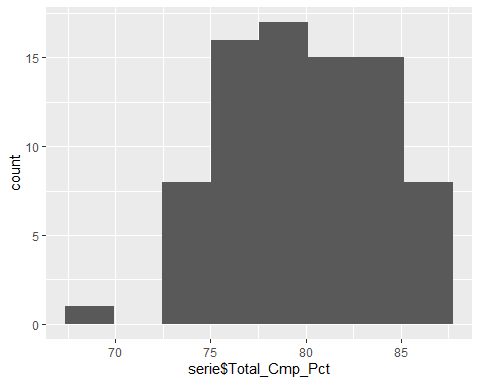
ggplot() +  
 geom\_histogram(mapping = aes(x = ligue$Total\_Cmp\_Pct), bins = 8)



ggplot() +  
 geom\_histogram(mapping = aes(x = epl$Total\_Cmp\_Pct), bins = 8)



ggplot() +  
 geom\_histogram(mapping = aes(x = serie$Total\_Cmp\_Pct), bins = 8)

 Figure 2. Mean and Standard Deviation of Total Pass Completion Percentage of Clubs in Major European Leagues

#Calculating the mean and standard deviation for the total completion percentage  
mean(bund$Total\_Cmp\_Pct)

## [1] 77.82361

sd(bund$Total\_Cmp\_Pct)

## [1] 4.05822

mean(liga$Total\_Cmp\_Pct)

## [1] 77.00625

sd(liga$Total\_Cmp\_Pct)

## [1] 5.876715

mean(ligue$Total\_Cmp\_Pct)

## [1] 79.0875

sd(ligue$Total\_Cmp\_Pct)

## [1] 3.996561

mean(epl$Total\_Cmp\_Pct)

## [1] 77.95625

sd(epl$Total\_Cmp\_Pct)

## [1] 5.148746

mean(serie$Total\_Cmp\_Pct)

## [1] 79.765

sd(serie$Total\_Cmp\_Pct)

## [1] 3.940542

Figure 3: Correlation Values between Total Pass Completion Percentage and Points per Match for Clubs in Major European Leagues from 2017/18 to 2020/21 seasons

cor(bund$Total\_Cmp\_Pct, bund$Pts.MP)

## [1] 0.6918816

cor(liga$Total\_Cmp\_Pct, liga$Pts.MP)

## [1] 0.5465845

cor(ligue$Total\_Cmp\_Pct, ligue$Pts.MP)

## [1] 0.6535265

cor(epl$Total\_Cmp\_Pct, epl$Pts.MP)

## [1] 0.7229535

cor(serie$Total\_Cmp\_Pct, serie$Pts.MP)

## [1] 0.6701023

Looking at the correlations for short, medium and long completion percentages versus points per match. Although not explicitly used in the paper it is mentioned so it is calculated here.

cor(bund$Short\_Cmp\_Pct, bund$Pts.MP)

## [1] 0.6448393

cor(liga$Short\_Cmp\_Pct, liga$Pts.MP)

## [1] 0.4686661

cor(ligue$Short\_Cmp\_Pct, ligue$Pts.MP)

## [1] 0.5645258

cor(epl$Short\_Cmp\_Pct, epl$Pts.MP)

## [1] 0.6687712

cor(serie$Short\_Cmp\_Pct, serie$Pts.MP)

## [1] 0.5148248

cor(bund$Medium\_Cmp\_Pct, bund$Pts.MP)

## [1] 0.5414421

cor(liga$Medium\_Cmp\_Pct, liga$Pts.MP)

## [1] 0.4526962

cor(ligue$Medium\_Cmp\_Pct, ligue$Pts.MP)

## [1] 0.5863414

cor(epl$Medium\_Cmp\_Pct, epl$Pts.MP)

## [1] 0.6799676

cor(serie$Medium\_Cmp\_Pct, serie$Pts.MP)

## [1] 0.605784

cor(bund$Long\_Cmp\_Pct, bund$Pts.MP)

## [1] 0.6724197

cor(liga$Long\_Cmp\_Pct, liga$Pts.MP)

## [1] 0.6191942

cor(ligue$Long\_Cmp\_Pct, ligue$Pts.MP)

## [1] 0.6646574

cor(epl$Long\_Cmp\_Pct, epl$Pts.MP)

## [1] 0.75147

cor(serie$Long\_Cmp\_Pct, serie$Pts.MP)

## [1] 0.6802777

Figure 4. All Five Leagues Summary off Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression Points per Match, from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for all leagues  
intercept\_only <- lm(Pts.MP ~ 1, data=all)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = all)  
step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ Long\_Cmp\_Pct + Ent\_Opposing\_PA + PrgDist +   
## Ent\_Final\_Third + Prog\_10\_plus\_yards, data = all)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.03153 -0.17631 0.00429 0.17558 0.87569   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.430e-01 1.765e-01 -1.377 0.1695   
## Long\_Cmp\_Pct 2.495e-02 2.790e-03 8.943 < 2e-16 \*\*\*  
## Ent\_Opposing\_PA 1.058e-03 4.348e-04 2.434 0.0154 \*   
## PrgDist -1.480e-05 2.132e-06 -6.944 1.63e-11 \*\*\*  
## Ent\_Final\_Third 8.432e-04 1.760e-04 4.792 2.36e-06 \*\*\*  
## Prog\_10\_plus\_yards 2.965e-04 2.073e-04 1.430 0.1535   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2917 on 386 degrees of freedom  
## Multiple R-squared: 0.6005, Adjusted R-squared: 0.5953   
## F-statistic: 116 on 5 and 386 DF, p-value: < 2.2e-16

Calculating predicted Points per match and the error

all$pred\_Pts.MP <- predict(step\_model, all)  
all$error <- all$Pts.MP - all$pred\_Pts.MP

Figure 5: Plot of Actual Points per Match vs. Predicted Points per Match (using the model in Figure 4) of All Clubs in Major Five Leagues from 2017/18 to 2020/21 season

plot(all$Pts.MP, all$pred\_Pts.MP)

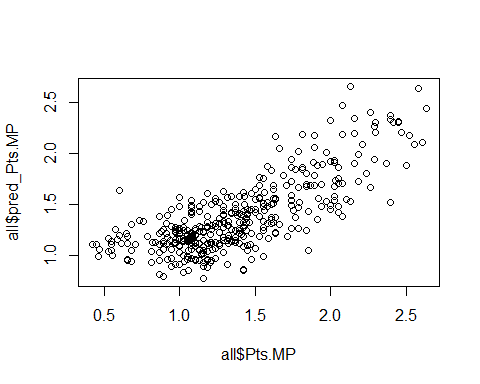


Figure 6: Res Plot of Overall Regression

plot(all$pred\_Pts.MP, all$error)

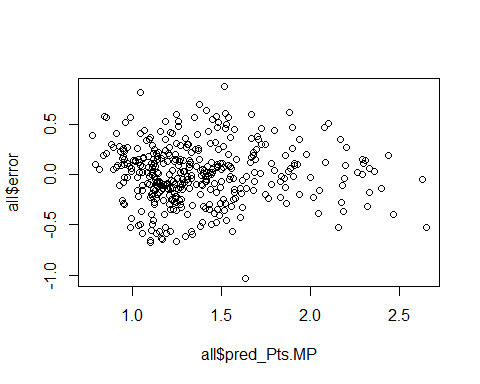


Figure 7: Bundesliga Summary of Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression on Points per Match, accounting for all 5 major leagues from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for Bundesliga  
intercept\_only <- lm(Pts.MP ~ 1, data=bund)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = bund)  
bund\_step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(bund\_step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ PrgDist + Ent\_Opposing\_PA, data = bund)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.65394 -0.14919 -0.00399 0.15507 0.80295   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.207e-01 3.631e-01 -2.536 0.01348 \*   
## PrgDist 1.669e-05 5.730e-06 2.913 0.00481 \*\*  
## Ent\_Opposing\_PA 2.640e-03 8.198e-04 3.221 0.00195 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2591 on 69 degrees of freedom  
## Multiple R-squared: 0.6769, Adjusted R-squared: 0.6676   
## F-statistic: 72.29 on 2 and 69 DF, p-value: < 2.2e-16

bund$pred\_Pts.MP <- predict(bund\_step\_model, bund)  
bund$error <- bund$Pts.MP - bund$pred\_Pts.MP

Figure 8: English Premier League Summary of Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression on Points per Match, from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for EPL  
intercept\_only <- lm(Pts.MP ~ 1, data=epl)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = epl)  
epl\_step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(epl\_step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ Ent\_Final\_Third, data = epl)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.82842 -0.16659 0.02317 0.15018 0.61119   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -5.655e-02 1.141e-01 -0.495 0.622   
## Ent\_Final\_Third 1.278e-03 9.741e-05 13.119 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2725 on 78 degrees of freedom  
## Multiple R-squared: 0.6881, Adjusted R-squared: 0.6841   
## F-statistic: 172.1 on 1 and 78 DF, p-value: < 2.2e-16

epl$pred\_Pts.MP <- predict(epl\_step\_model, epl)  
epl$error <- epl$Pts.MP - epl$pred\_Pts.MP

Figure 9: La Liga Summary of Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression on Points per Match, from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for La Liga  
intercept\_only <- lm(Pts.MP ~ 1, data=liga)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = liga)  
liga\_step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(liga\_step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ Cross\_into\_PA + Ent\_Opposing\_PA + Ent\_Final\_Third,   
## data = liga)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.75315 -0.19345 0.01924 0.18810 0.57510   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.4707648 0.1782245 2.641 0.01002 \*   
## Cross\_into\_PA -0.0040574 0.0014623 -2.775 0.00695 \*\*  
## Ent\_Opposing\_PA 0.0025415 0.0009208 2.760 0.00724 \*\*  
## Ent\_Final\_Third 0.0004500 0.0002899 1.552 0.12474   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2817 on 76 degrees of freedom  
## Multiple R-squared: 0.5851, Adjusted R-squared: 0.5688   
## F-statistic: 35.73 on 3 and 76 DF, p-value: 1.639e-14

liga$pred\_Pts.MP <- predict(liga\_step\_model, liga)  
liga$error <- liga$Pts.MP - liga$pred\_Pts.MP

Figure 10: Ligue 1 Summary of Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression Points per Match, accounting for all 5 major leagues from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for Ligue 1  
intercept\_only <- lm(Pts.MP ~ 1, data=ligue)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = ligue)  
ligue\_step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(ligue\_step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ Ent\_Opposing\_PA + Total\_Cmp\_Pct + PrgDist +   
## KP + Ent\_Final\_Third, data = ligue)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.69615 -0.12899 -0.01421 0.13729 0.54989   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.559e+00 6.731e-01 -3.802 0.000293 \*\*\*  
## Ent\_Opposing\_PA 1.409e-03 8.104e-04 1.738 0.086283 .   
## Total\_Cmp\_Pct 4.929e-02 8.489e-03 5.807 1.49e-07 \*\*\*  
## PrgDist -2.016e-05 4.245e-06 -4.748 9.80e-06 \*\*\*  
## KP 2.690e-03 9.553e-04 2.816 0.006232 \*\*   
## Ent\_Final\_Third 7.029e-04 2.994e-04 2.348 0.021571 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2514 on 74 degrees of freedom  
## Multiple R-squared: 0.6818, Adjusted R-squared: 0.6603   
## F-statistic: 31.72 on 5 and 74 DF, p-value: < 2.2e-16

ligue$pred\_Pts.MP <- predict(ligue\_step\_model, ligue)  
ligue$error <- ligue$Pts.MP - ligue$pred\_Pts.MP

Figure 11: Serie A Summary of Final Model of Coefficient Estimates, Standard Errors, Test Statistics, P-Values, and Significance for Linear Regression Points per Match, accounting for all 5 major leagues from 2017/18 to 2020/21 seasons

#Stepwise model for points per match for Serie A  
intercept\_only <- lm(Pts.MP ~ 1, data=serie)  
all\_predictors <- lm(Pts.MP ~ Total\_Cmp\_Pct + Ent\_Final\_Third + Ent\_Opposing\_PA + Cross\_into\_PA + PrgDist + Short\_Cmp\_Pct + Medium\_Cmp\_Pct + Long\_Cmp\_Pct + Prog\_10\_plus\_yards + KP, data = serie)  
serie\_step\_model <- step(intercept\_only, direction = "both", scope=formula(all\_predictors), trace=0)  
summary(serie\_step\_model)

##   
## Call:  
## lm(formula = Pts.MP ~ Long\_Cmp\_Pct + Ent\_Final\_Third + PrgDist +   
## Ent\_Opposing\_PA, data = serie)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.16831 -0.14064 0.02565 0.16763 0.79838   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -9.577e-01 4.864e-01 -1.969 0.05266 .   
## Long\_Cmp\_Pct 3.668e-02 7.484e-03 4.901 5.37e-06 \*\*\*  
## Ent\_Final\_Third 1.105e-03 3.635e-04 3.041 0.00324 \*\*   
## PrgDist -1.669e-05 4.530e-06 -3.685 0.00043 \*\*\*  
## Ent\_Opposing\_PA 1.693e-03 9.977e-04 1.697 0.09384 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3221 on 75 degrees of freedom  
## Multiple R-squared: 0.612, Adjusted R-squared: 0.5913   
## F-statistic: 29.58 on 4 and 75 DF, p-value: 9.108e-15

serie$pred\_Pts.MP <- predict(serie\_step\_model, serie)  
serie$error <- serie$Pts.MP - serie$pred\_Pts.MP

Figure 12: Plot of Actual Points per Match vs. Predicted Points per Match (using the models indicated in Figure 7-11) of All Clubs in Major Five Leagues from 2017/18 to 2020/21 season

par(mfrow=c(2,3)) #Used to bundle the plots into one figure to view easier in the paper  
plot(bund$Pts.MP, bund$pred\_Pts.MP)  
plot(epl$Pts.MP, epl$pred\_Pts.MP)  
plot(liga$Pts.MP, liga$pred\_Pts.MP)  
plot(ligue$Pts.MP, ligue$pred\_Pts.MP)  
plot(serie$Pts.MP, serie$pred\_Pts.MP)

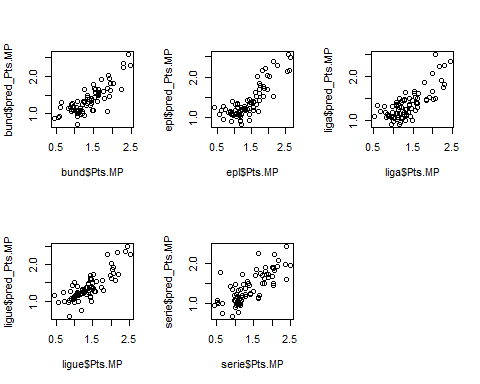


Figure 13: Res Plot of all 5 individual leagues

par(mfrow=c(2,3)) #Used to bundle the plots into one figure to view easier in the paper  
plot(bund$pred\_Pts.MP, bund$error)  
plot(epl$pred\_Pts.MP, epl$error)  
plot(liga$pred\_Pts.MP, liga$error)  
plot(ligue$pred\_Pts.MP, ligue$error)  
plot(serie$pred\_Pts.MP, serie$error)

