The History of International Soccer Through Data

Benjamin Holderbein

11/30/2021

Preamble

When deciding on what I wanted to do my first data science project on I took into account multiple factors. I want to use data science In ESports, but I figured when making a project that I will use to get my first internship, I would want to base it on something more relatable to a broader audience, so I settled on conventional sports. I'm not super into sports so deciding on a particular one was difficult. But I remembered that I used to play soccer for my hometown's youth league, and thus I picked soccer.

One of the more interesting aspects of soccer is how long it's been around. I thought that using the power of R, finding a data set that captured the history of soccer over thousands of data points could show some very interesting patterns. This lead me to my driving question of "How has soccer changed over time?". Finding the perfect data set to answer this question took a long while but I finally found the perfect one. "International football results from 1872 to 2021" by Mart Jürisoo should be perfect. It has over 40,000 observations that go all the way back to 1872 and features useful variables like where each team was from, where the match was played, what each team's score was and a few more that I will get into later.

Analysis

Loading Libraries:

```
library(tidyverse)
library(ggplot2)
library(lubridate)
library(countrycode)
library(scales)
library(ggpmisc)
```

Loading the data set:

```
results<-read.csv("results.csv")
```

Creating a formula that I will be using for linearization later:

```
my.formula <- y ~ x
```

Creating a column that shows which team won, or if it was a tie.

```
results <- results %>%
  mutate(winner= ifelse(home_score>away_score,home_team,ifelse(away_score>home_score,away_team,"Tie")))
```

Creating a column that has the date as a date object rather than a character string:

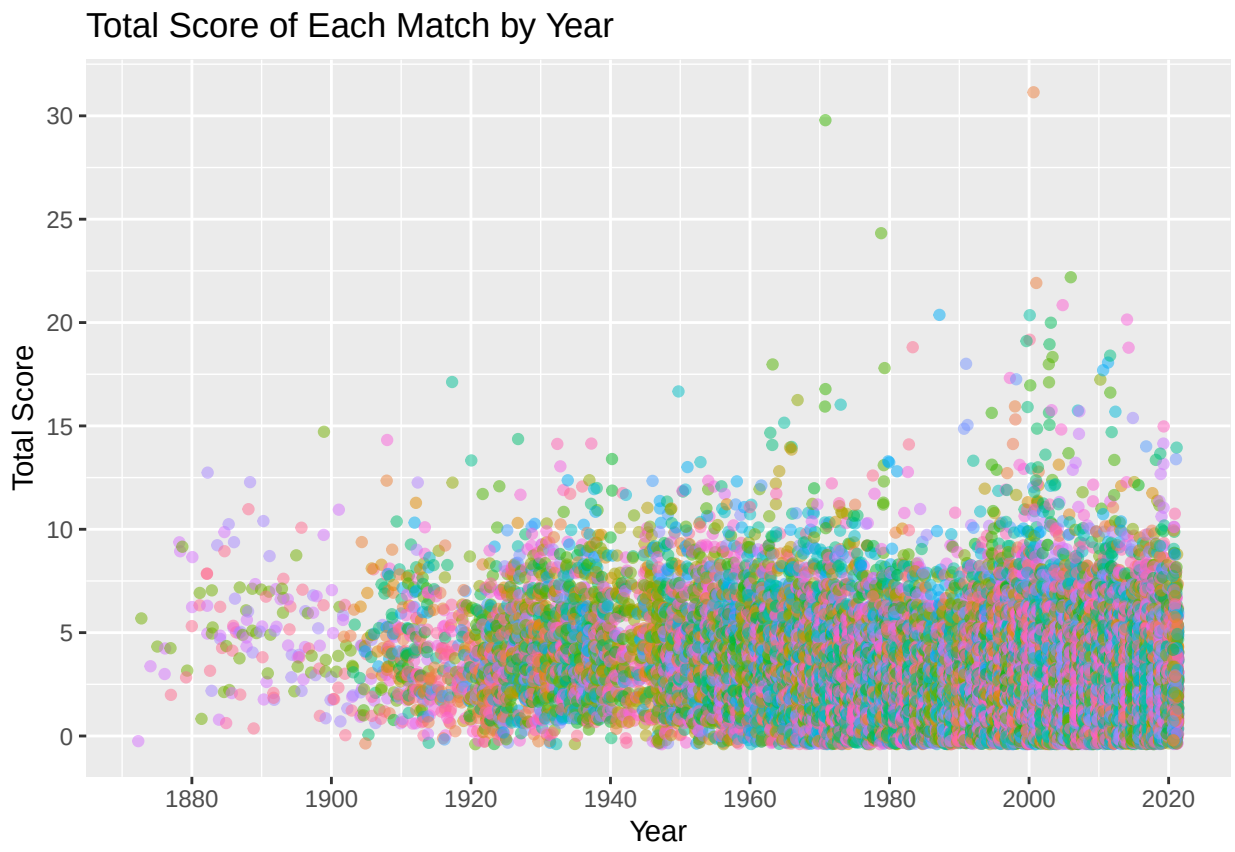
```
results<-results %>%
  mutate(tidy_date=ymd(date))
```

Creating a column that takes the absolute value of the difference between the home and away scores, called "spread":

```
results<-results %>%
  mutate(spread=abs(home_score-away_score))
```

This is where the biggest problem I will encounter started. When plotting this graph I realized that using the country column (what country the match was played in) would never be useful as it has over 200 different countries in it, making it unusable in any graph. So I decided to categorize the countries by continent.

```
ggplot(results,aes(x=year(tidy_date),y=home_score+away_score,color=country))+
  geom_point(alpha=.5,position="jitter",show.legend = FALSE)+
  labs(title="Total Score of Each Match by Year",x="Year",y="Total Score",color="Continent")+
  scale_x_continuous(breaks=seq(1860,2020,by=20))+
  scale_y_continuous(breaks=seq(0,35,by=5))
```



After a lot of searching and looking at what other data scientists do, I landed on the "countrycode" library, making it possible to input countries and output continents.

```
results$continent <- countrycode(sourcevar = results[, "country"],
                            origin = "country.name",
                            destination = "continent")
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
ncipe, Saarland, Saint Martin, Scotland, Serbia and Montenegro, Tanganyika, Wales, Yugoslavia, ZaÃ¯re, Z
```

```
results$w_continent <- countrycode(sourcevar = results[, "winner"],
                            origin = "country.name",
                            destination = "continent")
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some val
ncipe, Saare County, Saarland, Saint Martin, Scotland, Shetland, Silesia, South Ossetia, Surrey, SzÃ©kel

## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some str
```

But what I didn't foresee is the above warning message. This is because the data set, being historical, contains regions and countries not recognized in modern times, making R spit out warnings for every country name it doesn't recognize. To fix this I decided I would make two functions to rename the country from its old name to it's modern day name, one for the winner column, one for the country column:

```
country_rename <- function(old,new){
  results$country[results$country==old] <- new
  return(results)
}
winner_rename <- function(old,new){
  results$winner[results$winner==old] <- new
  return(results)
}
```

Below is 99 calls of this function, each of which I had to manually google and enter the modern name, this took longer than one might expect:

```
results<-country_rename("Ã‰ire","Ireland")
results<-country_rename("CuraÃ§ao","Curaçao")
results<-country_rename("Czechoslovakia","Czechia")
results<-country_rename("England","United Kingdom")
results<-country_rename("German DR","Germany")
results<-country_rename("Irish Free State","Ireland")
results<-country_rename("Kosovo","Serbia")
results<-country_rename("Lautoka","Fiji")
results<-country_rename("Malaya","Malaysia")
results<-country_rename("Manchuria","China")
results<-country_rename("Micronesia","Micronesia (Federated States of)")
results<-country_rename("Northern Ireland","Ireland")
results<-country_rename("RÃ©union","Réunion")
results<-country_rename("SÃ£o TomÃ© and PrÃncipe","São Tomé & Príncipe")
results<-country_rename("Saarland","Germany")
results<-country_rename("Saint Martin","Saint Martin (French part)")
results<-country_rename("Scotland","United Kingdom")
results<-country_rename("Serbia and Montenegro","Montenegro")
```

```
results<-country_rename("Tanganyika","Tanzania")
results<-country_rename("Wales","United Kingdom")
results<-country_rename("Yugoslavia","Bosnia and Herzegovina")
results<-country_rename("ZaÃ¯re","Democratic Republic of the Congo")
results<-country_rename("Zanzibar","Tanzanian")
```

```
results<-winner_rename("Ã‰ire","Ireland")
results<-winner_rename("CuraÃ§ao","Curaçao")
results<-winner_rename("Czechoslovakia","Czechia")
results<-winner_rename("England","United Kingdom")
results<-winner_rename("German DR","Germany")
results<-winner_rename("Irish Free State","Ireland")
results<-winner_rename("Kosovo","Serbia")
results<-winner_rename("Lautoka","Fiji")
results<-winner_rename("Malaya","Malaysia")
results<-winner_rename("Manchuria","China")
results<-winner_rename("Micronesia","Micronesia (Federated States of)")
results<-winner_rename("Northern Ireland","Ireland")
results<-winner_rename("RÃ©union","Réunion")
results<-winner_rename("SÃ£o TomÃ© and PrÃncipe","São Tomé & Príncipe")
results<-winner_rename("Saarland","Germany")
results<-winner_rename("Saint Martin","Saint Martin (French part)")
results<-winner_rename("Scotland","United Kingdom")
results<-winner_rename("Serbia and Montenegro","Montenegro")
results<-winner_rename("Tanganyika","Tanzania")
results<-winner_rename("Wales","United Kingdom")
results<-winner_rename("Yugoslavia","Bosnia and Herzegovina")
results<-winner_rename("ZaÃ¯re","Democratic Republic of the Congo")
results<-winner_rename("Zanzibar","Tanzanian")
results<-winner_rename("Ã…land Islands","Finland")
results<-winner_rename("Abkhazia","Georgia")
results<-winner_rename("Alderney","United Kingdom")
results<-winner_rename("Andalusia","Spain")
results<-winner_rename("Arameans Suryoye","Syria")
results<-winner_rename("Artsakh","Azerbaijan")
results<-winner_rename("Asturias","Spain")
results<-winner_rename("Barawa","Somalia")
results<-winner_rename("Basque Country","Spain")
results<-winner_rename("Bonaire","Netherlands")
results<-winner_rename("Brittany","France")
results<-winner_rename("Canary Islands","Spain")
results<-winner_rename("Cascadia","Canada")
results<-winner_rename("Catalonia","Spain")
results<-winner_rename("Chagos Islands","United Kingdom")
results<-winner_rename("Chameria","Albania")
results<-winner_rename("Corsica","France")
results<-winner_rename("County of Nice","France")
results<-winner_rename("Crimea","Ukraine")
results<-winner_rename("Ellan Vannin","United Kingdom")
results<-winner_rename("Galicia","Spain")
results<-winner_rename("Gotland","Sweden")
results<-winner_rename("Gozo","Malta")
results<-winner_rename("Hitra","Norway")
```

```
results<-winner_rename("Isle of Wight","United Kingdom")
results<-winner_rename("KÃ¡rpÃ¡talja","")
results<-winner_rename("Kabylia","Algeria")
results<-winner_rename("Kernow","United Kingdom")
results<-winner_rename("Matabeleland","Zimbabwe")
results<-winner_rename("Menorca","Spain")
results<-winner_rename("Occitania","France")
results<-winner_rename("Orkney","United Kingdom")
results<-winner_rename("Padania","Italy")
results<-winner_rename("Panjab","India")
results<-winner_rename("Provence","France")
results<-winner_rename("Raetia","Italy")
results<-winner_rename("Republic of St. Pauli","Germany")
results<-winner_rename("Rhodes","Greece")
results<-winner_rename("Romani people","Romania")
results<-winner_rename("SÃ¡pmi","South Africa")
results<-winner_rename("Saare County","Estonia")
results<-winner_rename("Shetland","United Kingdom")
results<-winner_rename("Silesia","Poland")
results<-winner_rename("South Ossetia","Iran")
results<-winner_rename("Surrey","United Kingdom")
results<-winner_rename("SzÃ©kely Land","Hungary")
results<-winner_rename("Tamil Eelam","Sri Lanka")
results<-winner_rename("Two Sicilies","Italy")
results<-winner_rename("United Koreans in Japan","Japan")
results<-winner_rename("Western Isles","United Kingdom")
results<-winner_rename("Ynys MÃ´n","United Kingdom")
results<-winner_rename("Yorkshire","United Kingdom")
```

After that migraine I can finally run the countrycode function again, this time with only a few errors, for countries I simply could not find.

```
results$continent <- countrycode(sourcevar = results[, "country"],
                            origin = "country.name",
                            destination = "continent")
results$w_continent <- countrycode(sourcevar = results[, "winner"],
                            origin = "country.name",
                            destination = "continent")
```

```
## Warning in countrycode_convert(sourcevar = sourcevar, origin = origin, destination = dest, : Some va
```

```
results$w_continent[is.na(results$w_continent)] <- "Tie"
```

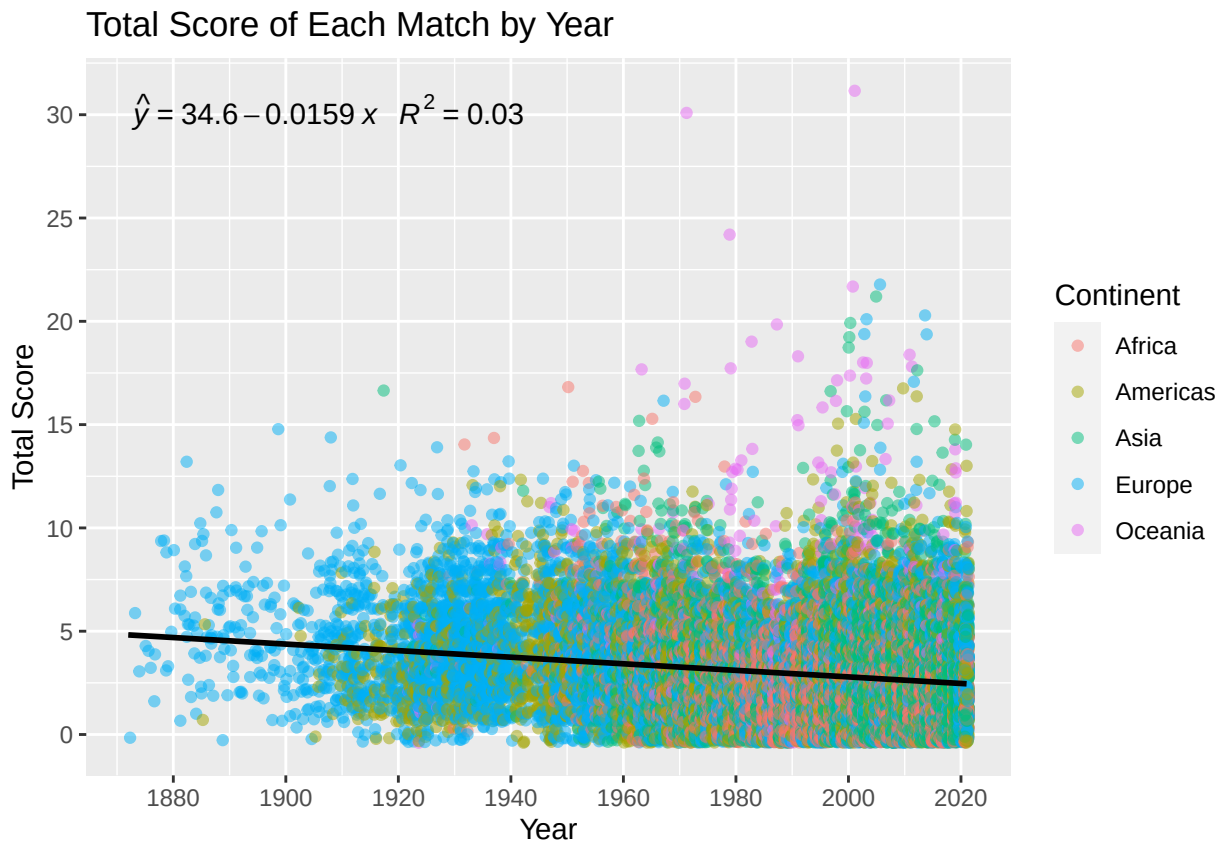
Plotting

This is my first real exploratory plot, a lot going on but it hints at a few interesting patterns. The first thing I notice is how much blue is on the left. The second is that negative slope lsrl hinting at a decrease in score over time, but with coefficient of determination of only 0.03, I can't make conclusions from it.

```
ggplot(results,aes(x=year(tidy_date),y=home_score+away_score,color=continent))+
  geom_point(alpha=.5,position="jitter")+
```

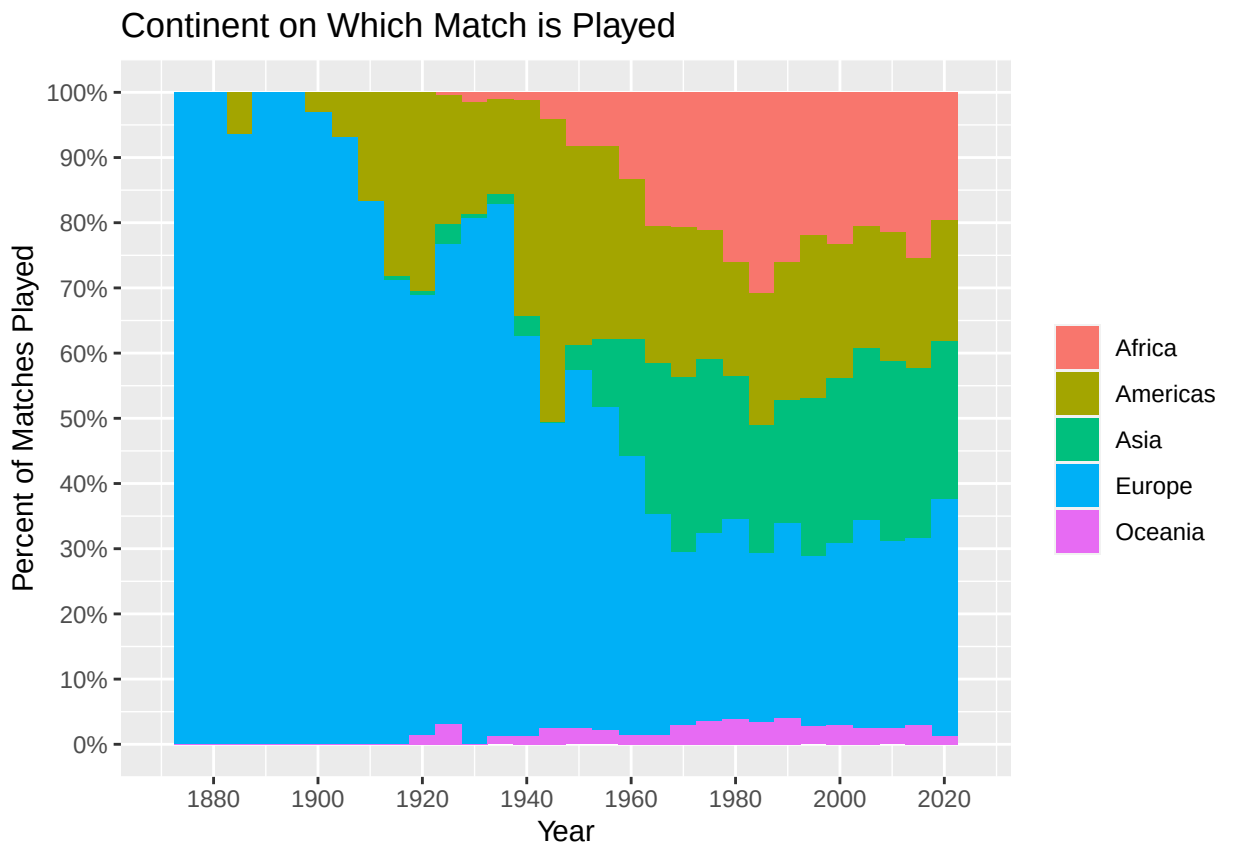
```
  labs(title="Total Score of Each Match by Year",x="Year",y="Total Score",color="Continent")+
  geom_smooth(color="black",method="lm",alpha=1)+
  scale_x_continuous(breaks=seq(1860,2020,by=20))+
  scale_y_continuous(breaks=seq(0,35,by=5))+
  stat_poly_eq(formula = my.formula,
                eq.with.lhs = "italic(hat(y))~`=`~",
                aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
                parse = TRUE,color="black")
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat_smooth).

## Warning: Removed 1 rows containing non-finite values (stat_poly_eq).

## Warning: Removed 1 rows containing missing values (geom_point).
```



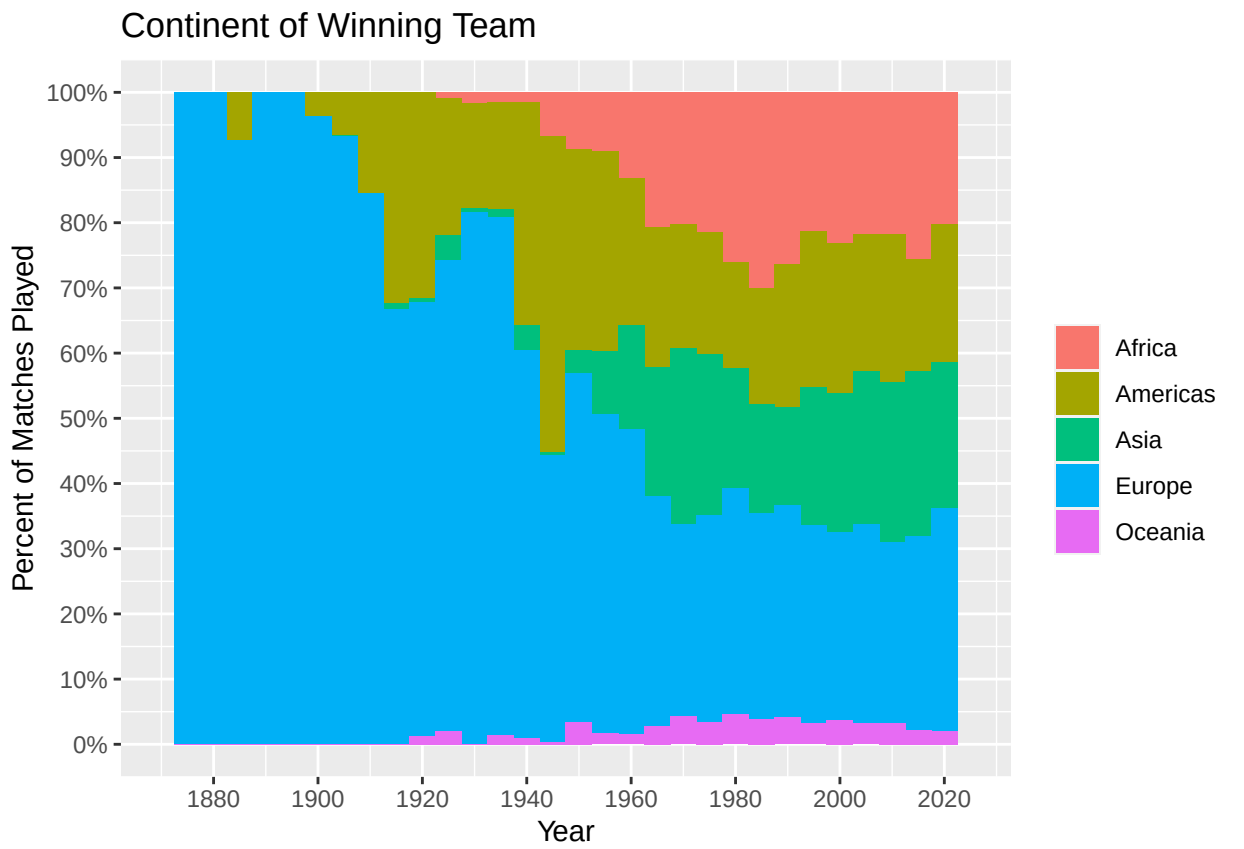
This plot shows where soccer was played over time. As you can see it started in Europe and is fairly well distributed across the continents now.

```
ggplot(results,aes(x=year(tidy_date)))+
  geom_bar(aes(fill=continent),position = "fill",stat = "bin",binwidth=5)+
  labs(title="Continent on Which Match is Played",x="Year",y="Percent of Matches Played",fill="")+
  scale_y_continuous(breaks=seq(0,1,by=.1),limits=c(0,1),labels=label_percent(accuracy = 1))+
  scale_x_continuous(breaks=seq(1860,2020,by=20),limits=c(1870,2025))
```



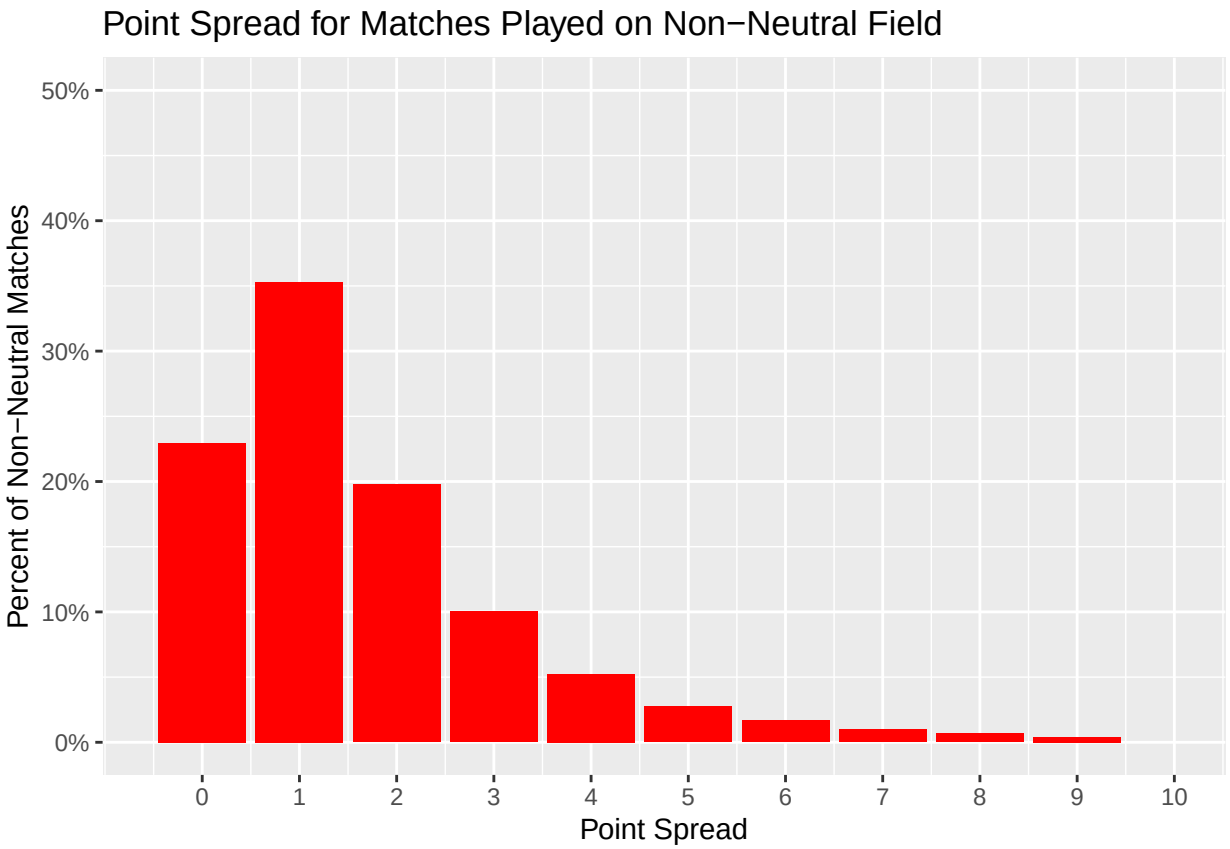
This plot is very simular to the above one, except it shows which continent wins rather than where it was played, but it is nearly identical to the above plot.

```
ggplot(data=subset(results,!(is.na(w_continent)|w_continent=="Tie")),aes(x=year(tidy_date)))+
  geom_bar(aes(fill=w_continent),position = "fill",stat = "bin",binwidth=5)+
  labs(title="Continent of Winning Team",x="Year",y="Percent of Matches Played",fill="")+
  scale_y_continuous(breaks=seq(0,1,by=.1),limits=c(0,1),labels=label_percent(accuracy = 1))+
  scale_x_continuous(breaks=seq(1860,2020,by=20),limits=c(1870,2025))
```

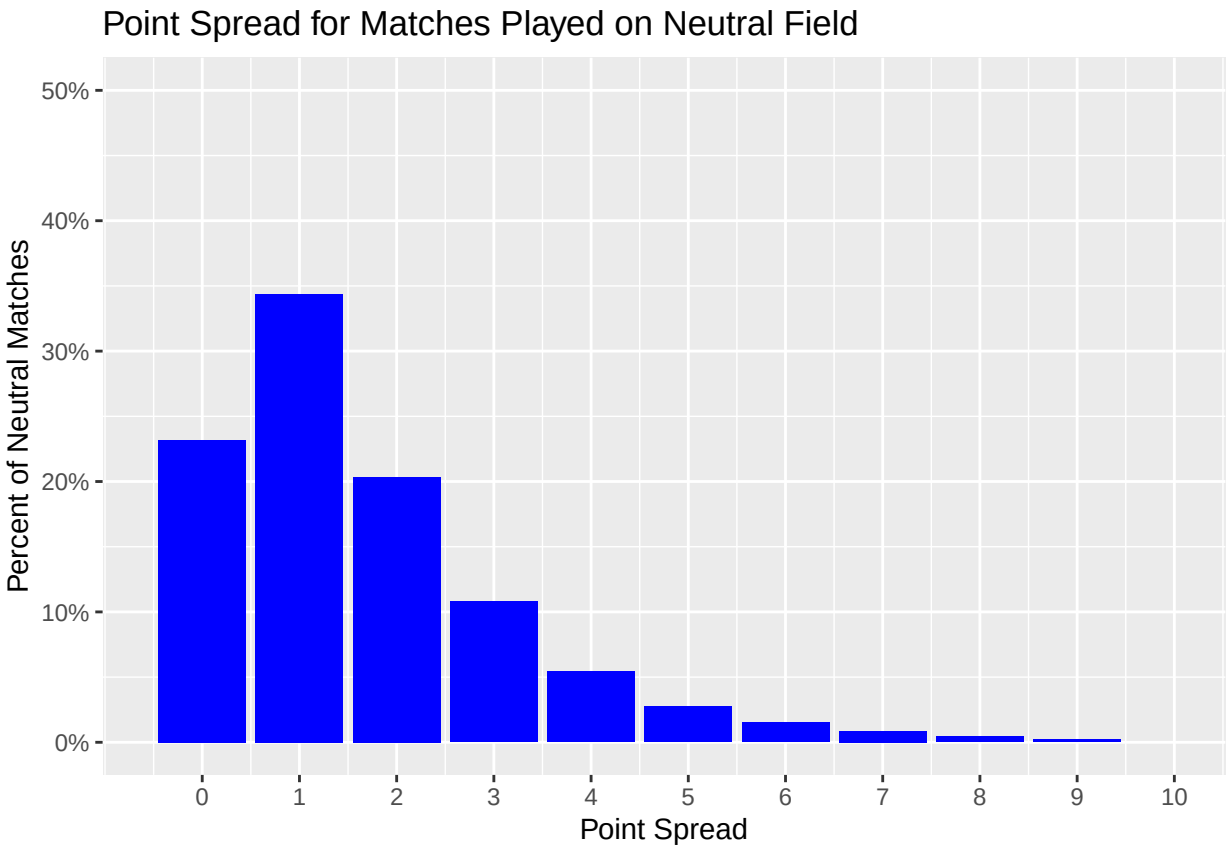


In these two plots I explore what the Neutral variable means. The Neutral variable means that home and away was determined by something arbitrary rather than if one team was on their home field.As you can see the point spread is nearly identical for the graphs.

```
ggplot(subset(results,results$neutral),aes(x=spread,y=(..count..)/sum(..count..)))+
  geom_bar(fill="red")+
  scale_y_continuous(breaks=seq(0,1,by=.1),limits=c(0,.5),labels=label_percent(accuracy = 1))+
  scale_x_continuous(breaks=seq(0,10,by=1),limits=c(-.5,10))+
  labs(title="Point Spread for Matches Played on Non-Neutral Field",y="Percent of Non-Neutral Matches",
```

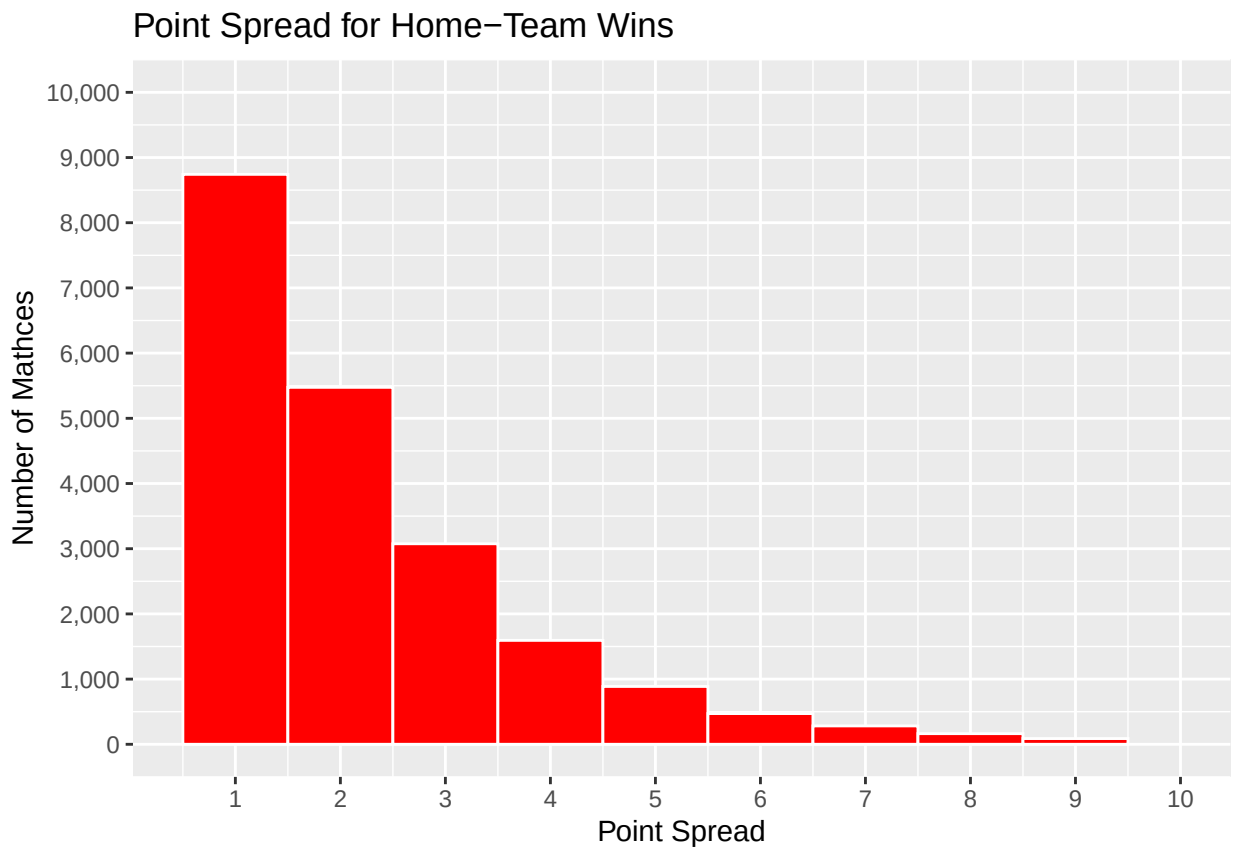



```
ggplot(subset(results,!results$neutral),aes(x=spread,y=(..count..)/sum(..count..)))+
  geom_bar(fill="blue")+
  scale_y_continuous(breaks=seq(0,1,by=.1),limits=c(0,.5),labels=label_percent(accuracy = 1))+
  scale_x_continuous(breaks=seq(0,10,by=1),limits=c(-.5,10))+
  labs(title="Point Spread for Matches Played on Neutral Field",y="Percent of Neutral Matches",x="Point
```

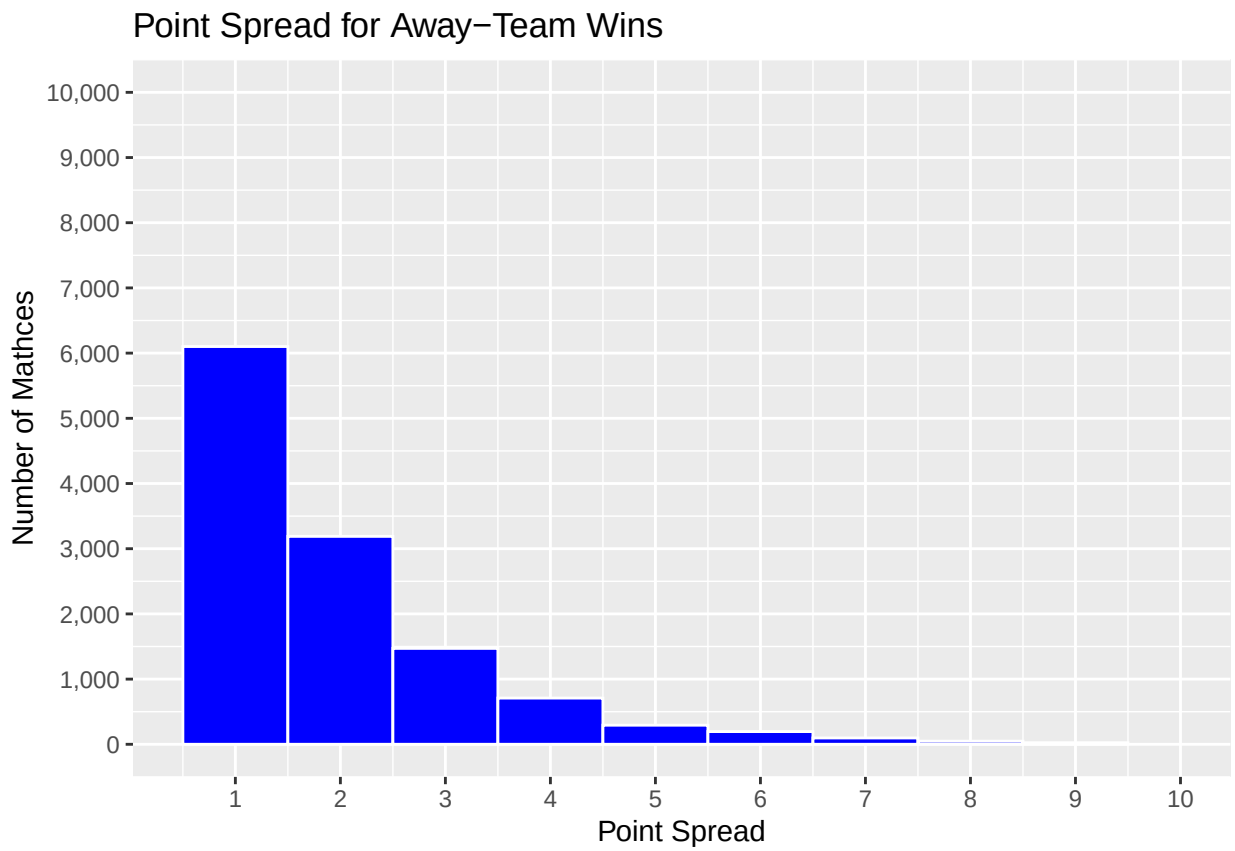


Here I explore whether the home team has an advantage or not. As you can see there are far more home team wins than away team wins. This makes sense as home and away is either determined by one team being on their home field, giving them an advantage, or generally speaking which team is better obviously giving them an advantage.

```
ggplot(subset(results,home_score>away_score),aes(x=spread))+
  geom_histogram(binwidth=1,fill="red",color="white")+
  scale_x_continuous(breaks=seq(1,10,by=1),limits=c(.5,10))+
  scale_y_continuous(breaks=seq(0,10000,by=1000),limits=c(0,10000),labels=comma)+
  labs(title="Point Spread for Home-Team Wins",x="Point Spread",y="Number of Mathces")
```

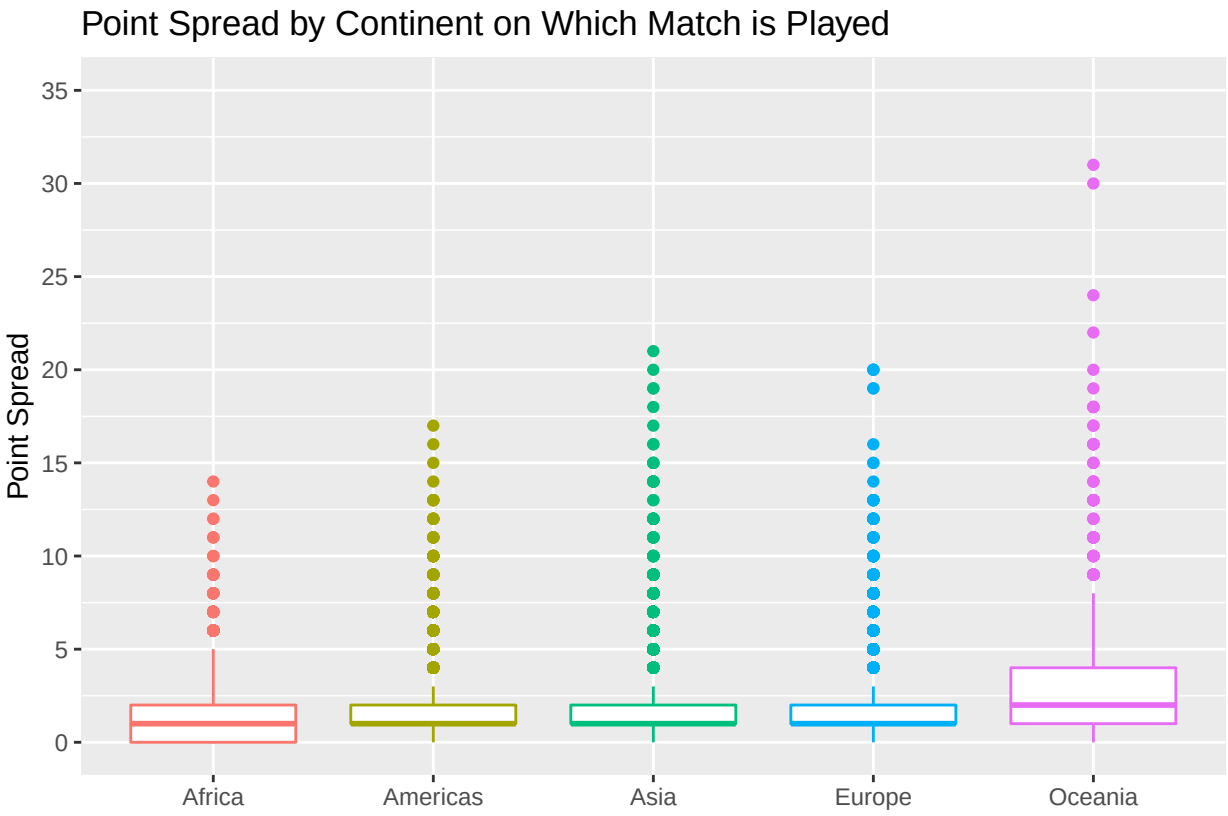



```
ggplot(subset(results,home_score<away_score),aes(x=spread))+
  geom_histogram(binwidth=1,fill="blue",color="white")+
  scale_x_continuous(breaks=seq(1,10,by=1),limits=c(.5,10))+
  scale_y_continuous(breaks=seq(0,10000,by=1000),limits=c(0,10000),labels=comma)+
  labs(title="Point Spread for Away-Team Wins",x="Point Spread",y="Number of Mathces")
```



This plot shows the mean, standard deviation, min/max, and outliers of the difference of scores for each continent. As you can see its pretty similar across the continents except for oceania where it varies a bit more.

```
ggplot(results,aes(x=continent,y=spread,color=continent))+
  geom_boxplot(show.legend = FALSE)+
  labs(title="Point Spread by Continent on Which Match is Played",x="",y="Point Spread")+
  scale_y_continuous(breaks=seq(0,35,by=5),limits=(c(0,35.00001)))
```



Here I generate a table and two statements that describe the relation ship of Neutral, home vs. away and winning. The table shows Home team has a clear advantage, on both Neutral and non-neutral fields, with a even larger disparity for home-field wins, nearly 2:1.

```
All_Matches<-percent(c(sum(results$home_score>results$away_score,na.rm=T)/length(results$home_score),su

Neutral_Field<-percent(c(sum(subset(results,neutral)$home_score>subset(results,neutral)$away_score,na.r

Home_Field<-percent(c(sum(subset(results,!neutral)$home_score>subset(results,!neutral)$away_score,na.rm

df<-data.frame(All_Matches,Neutral_Field,Home_Field)

rownames(df)<-c("Home Wins","Away Wins","Tie")
colnames(df)<-c("All Matches","Neutral Field","Home Field")
df
```

```
##           All Matches Neutral Field Home Field
## Home Wins       48.6%         43.0%      50.5%
## Away Wins       28.3%         34.2%      26.4%
## Tie             23.0%         22.8%      23.1%
```

```
paste(format(100*sum(results$neutral)/length((results$neutral)),digits=3),"% of all games were Neutral"
```

```
## [1] "24.7% of all games were Neutral"
```

```
paste(format(100*sum(!results$neutral)/length((results$neutral)),digits=3),"% of all games were not Neu
```

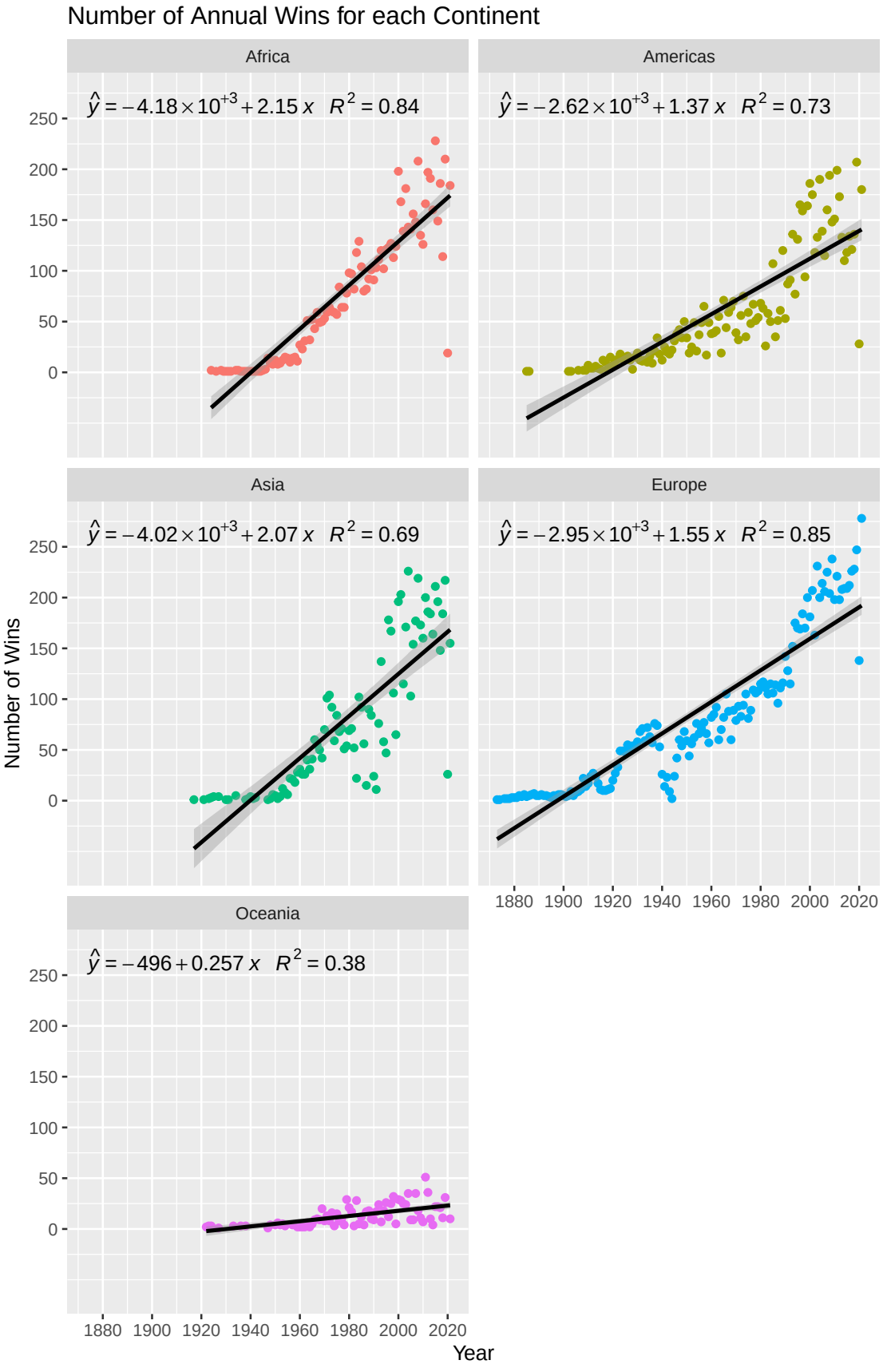
```
## [1] "75.3% of all games were not Neutral"
```

These final plots show annual wins for each continent over time. The linearizations plotted over each shows the growth of both sport and skill for each continent.These all have decent coefficients of determination so we can make some pretty cool conclusions. Africa has by far the highest growth rate, where Oceania has the lowest.

```
win_by_year<-results %>%
  group_by(year(tidy_date),w_continent) %>%
  summarise(n())
```

```
## `summarise()` has grouped output by 'year(tidy_date)'. You can override using the `.groups` argument
```

```
win_by_year<- win_by_year %>%
  filter(w_continent!="Tie")

colnames(win_by_year)<-c("year","w_continent","wins")

ggplot(win_by_year,aes(x=year,y=wins,color=w_continent))+
  geom_point(show.legend = FALSE)+
  geom_smooth(formula=y~x,method="lm",color="black")+
  stat_poly_eq(formula = my.formula,
               eq.with.lhs = "italic(hat(y))~`=`~",
               aes(label = paste(..eq.label.., ..rr.label.., sep = "~~~")),
               parse = TRUE,color="black") +
  facet_wrap(~ w_continent,ncol=2)+
  labs(title="Number of Annual Wins for each Continent",x="Year",y="Number of Wins")+
  scale_x_continuous(breaks=seq(1860,2020,by=20))+
  scale_y_continuous(breaks=seq(0,300,by=50))
```


Number of Annual Wins for each Continent
Africa
$\hat{y} = -4.18 \times 10^{+3} + 2.15\,x \quad R^2 = 0.84$
Americas
$\hat{y} = -2.62 \times 10^{+3} + 1.37\,x \quad R^2 = 0.73$
Asia
$\hat{y} = -4.02 \times 10^{+3} + 2.07\,x \quad R^2 = 0.69$
Europe
$\hat{y} = -2.95 \times 10^{+3} + 1.55\,x \quad R^2 = 0.85$
Oceania
$\hat{y} = -496 + 0.257\,x \quad R^2 = 0.38$
Number of Wins
0
50
100
150
200
250
1880
1900
1920
1940
1960
1980
2000
2020
Year

Conclusions

This project left us with some pretty cool conclusions: -Oceania has more variation than other continents. -Home team has an advantage, on both neutral and non-neutral fields. -Home team has a BIG advantage on non-neutral fields. -Africa has the largest win's growth rate. -Oceania is growing the slowest. -Soccer started in Europe and is now fairly distributed across all continents making it the most international sport ever seen.

In the future I would like to: -Look more closely at outliers -Look more closely at home team advantage on neutral fields -Look more closely at the nonlinear annual growth -Separate home team advantage for neutral and home fields