THE HISTORY OF INTERNATIONAL SOCCER THROUGH DATA
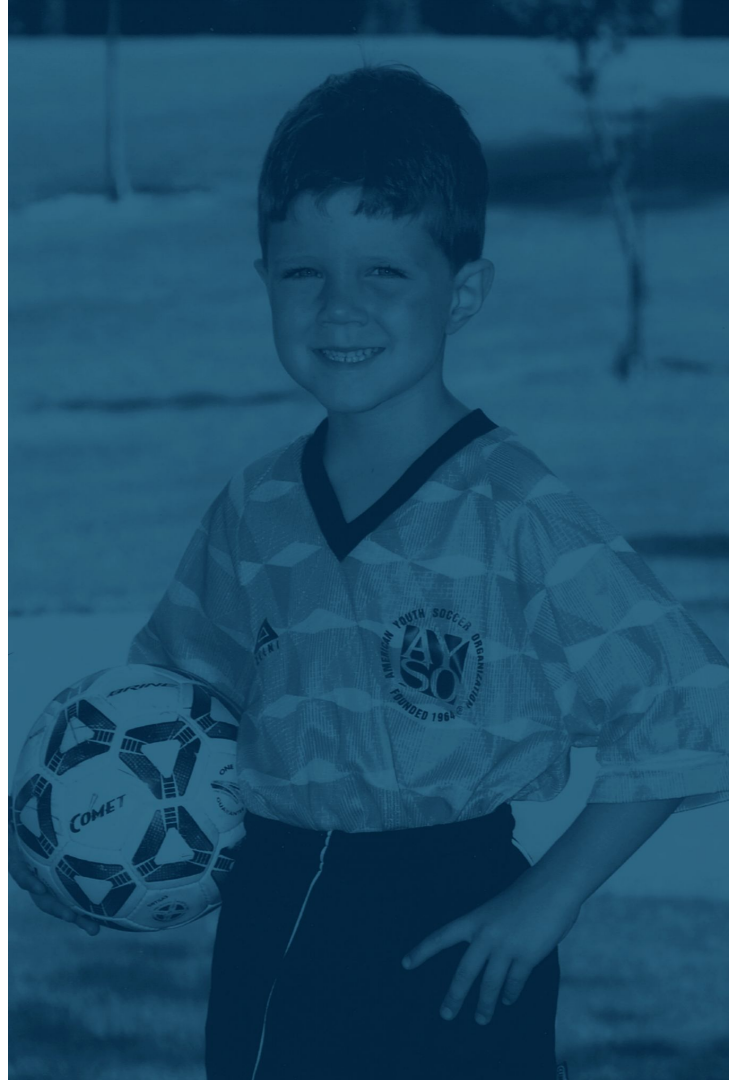
By Benjamin Holderbein

TABLE OF CONTENTS

WHY SOCCER

- I want to use data science in Esports
- I've always loved soccer
- Seems like data science is applicable
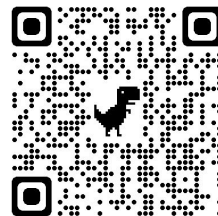
HOW HAS SOCCER CHANGED OVER TIME

THE DATA

International football results from 1872 to 2021:

- By Mart Jürisoo
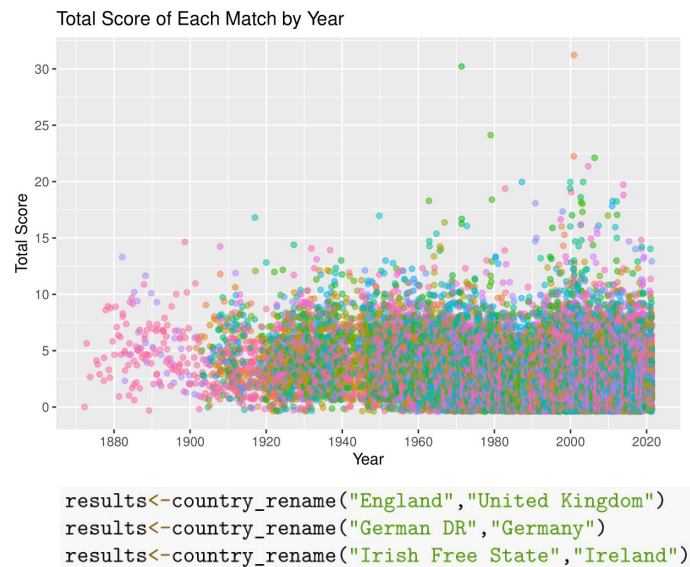- Entire history of International "football" matches
- 42,899 Observations

📅 date	⬆ home_team	⬆ away_team	# home_score	# away_score
1872-11-30	Scotland	England	0	0
1873-03-08	England	Scotland	4	2
1874-03-07	Scotland	England	2	1
1875-03-06	England	Scotland	2	2
1876-03-04	Scotland	England	3	0

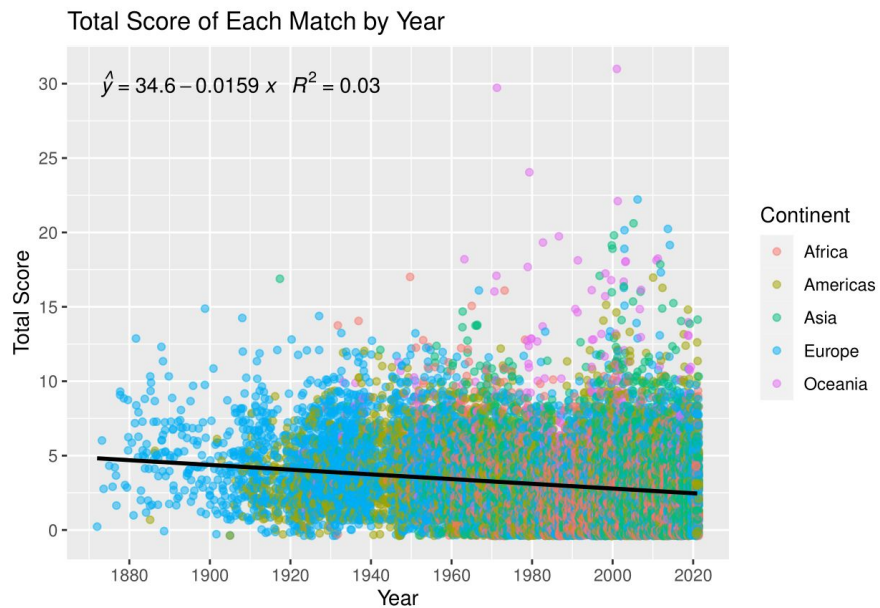| ⬆ tournament	⬆ city	⬆ country	✓ neutral
Friendly	Glasgow	Scotland	FALSE
Friendly	London	England	FALSE
Friendly	Glasgow	Scotland	FALSE
Friendly	London	England	FALSE
Friendly	Glasgow	Scotland	FALSE

CHALLENGES

- Problem: Way too many countries to be meaningful
- Solution: Bin them down to continents
- Problem: How the heck do I do that?
- Solution: library(countrycode)
- Problem: There are countries that don't exist anymore
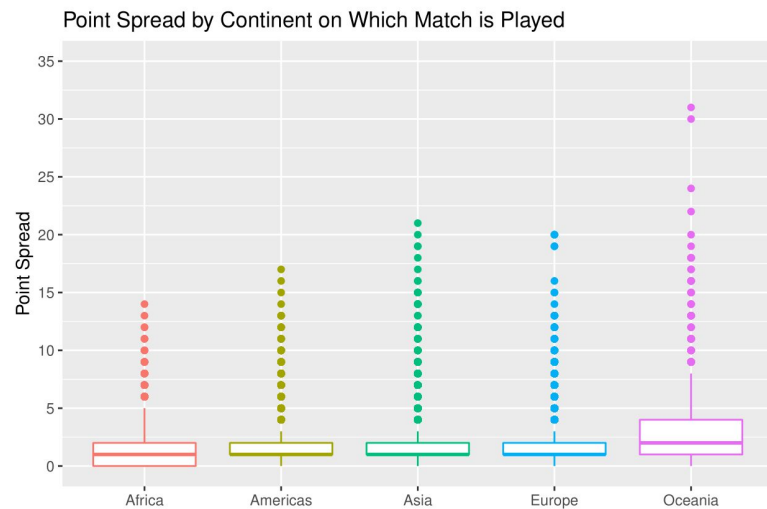- Solution: Create a function that can manually rename 1 country at a time

Total Score of Each Match by Year


```
results<-country_rename("England","United Kingdom")
results<-country_rename("German DR","Germany")
results<-country_rename("Irish Free State","Ireland")
```

ANALYSIS

ANALYSIS 01

Total Score of Each Match by Year

$\hat{y} = 34.6 - 0.0159\,x$ $R^2 = 0.03$

Continent

- Africa
- Americas
- Asia
- Europe
- Oceania

- First real graph
- Looking for patterns
- Interesting LSRL but R^2 is poor

ANALYSIS 02

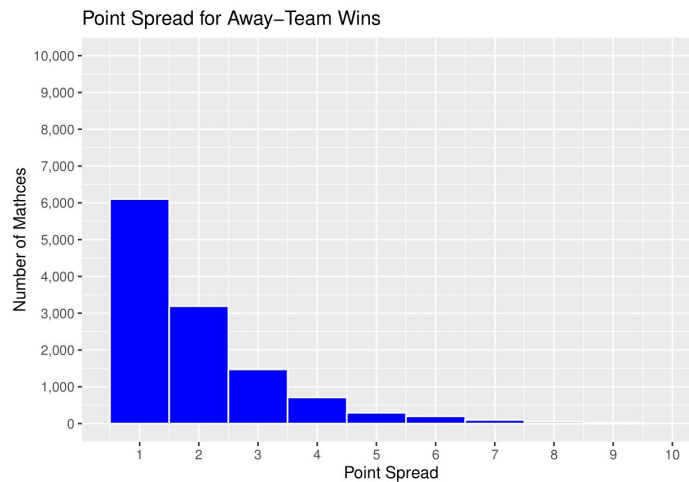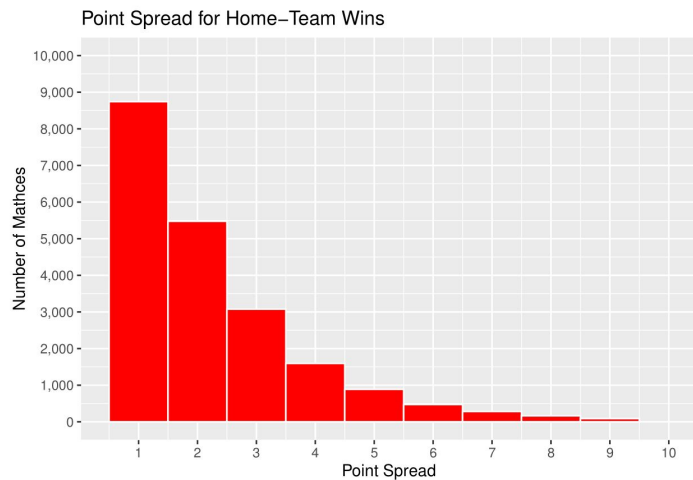Point Spread by Continent on Which Match is Played

- Oceania has much larger SD than other continents
- Otherwise regular between continents

Point Spread = |Home Score-Away Score|

ANALYSIS 03

- Home team has an advantage
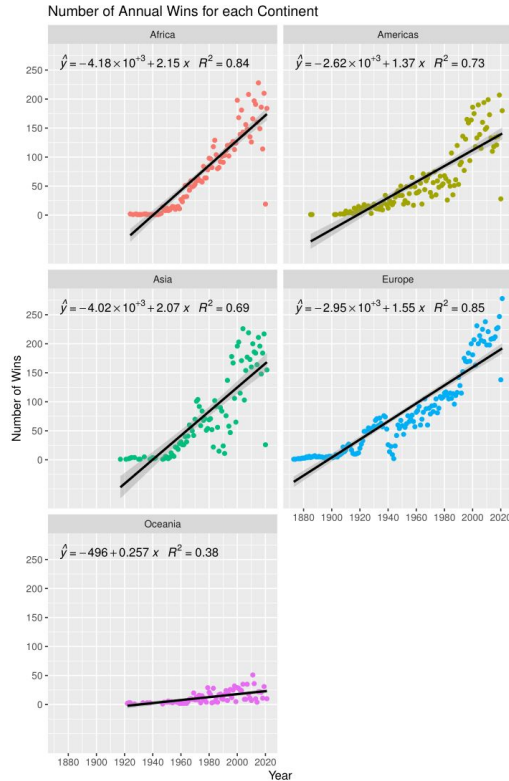- This is because Away/Home is determined by which team is better

Point Spread for Home–Team Wins

Point Spread for Away–Team Wins

ANALYSIS 04

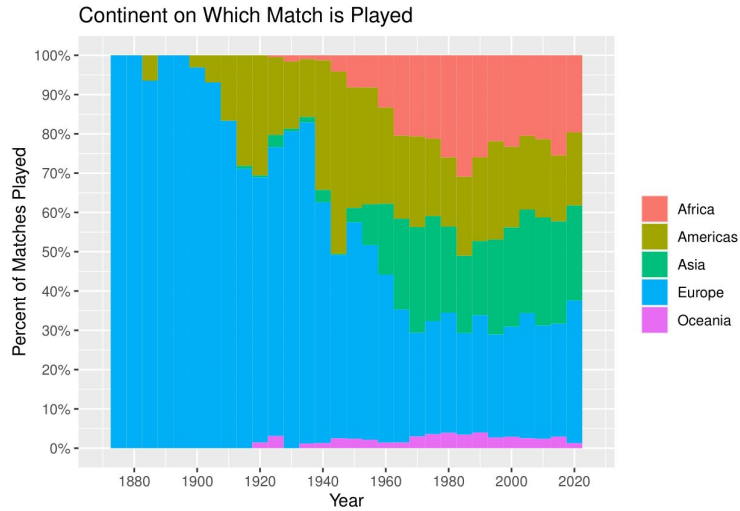|	All Matches	Neutral Field	Home Field
Home Wins	48.6%	43.0%	50.5%
Away Wins	28.3%	34.2%	26.4%
Tie	23.0%	22.8%	23.1%

- Neutral Field means game is played on the field of a third country
- Home team has a big advantage on home fields: almost 2 to 1
- Home team also has an advantage on neutral fields: 1.25 to 1

ANALYSIS 05

Number of Annual Wins for each Continent

- This graph shows each continent's growth in annual wins each year
- Africa has grown the most with a M of 2.15
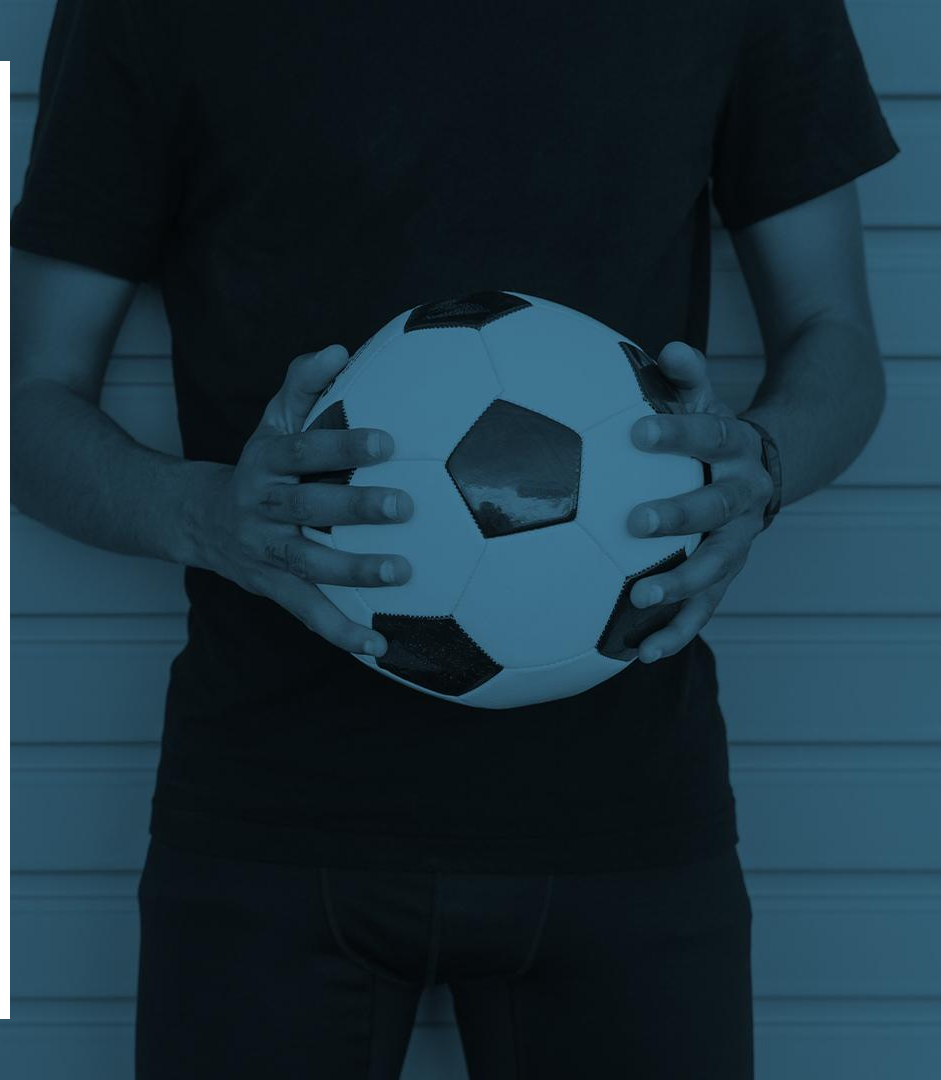- Oceania has grown the least with a M of 0.26

ANALYSIS 06

Continent on Which Match is Played

- Shows change in where soccer is played
- Started in Europe and is now fairly distributed

CONCLUSIONS

CONCLUSIONS

- Oceania has much larger SD 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

FUTURE WORK

- 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

THANK YOU