2022 FIFA World Cup

By: Chase Webber

Introduction

Motivation

- Main goal is to predict the winner of the 2022 FIFA World Cup.
- It is a relevant topic as the World Cup is going on right now (Happens every 4 years).
- Many people try to predict the winner to win bets, but what if we used statistics to help make that prediction?

- Will be using past data from previous World Cups and international games to make our predictions.
- The main variables we will use to predict this is team_win, team, home_away, and positional rankings.

Raw Data

https://www.kaggle.com/datasets/brenda89/fifa-world-cup-2022

```
fifa <- read.csv("international_matches.csv")
```

```
> colnames(fifa)
 [1] "date"                       "home_team"                  "away_team"
 [4] "home_team_continent"        "away_team_continent"        "home_team_fifa_rank"
 [7] "away_team_fifa_rank"        "home_team_total_fifa_points" "away_team_total_fifa_points"
[10] "home_team_score"            "away_team_score"            "tournament"
[13] "city"                       "country"                    "neutral_location"
[16] "shoot_out"                  "home_team_result"           "home_team_goalkeeper_score"
[19] "away_team_goalkeeper_score" "home_team_mean_defense_score" "home_team_mean_offense_score"
[22] "home_team_mean_midfield_score" "away_team_mean_defense_score" "away_team_mean_offense_score"
[25] "away_team_mean_midfield_score"
>
```

```
> dim(fifa)
[1] 23921    25
>
```

Summary of data cleaning/feature engineering

- Combine all home and away statistics into one table
  - Using **rbind**
-
- Mutate some variables
-
- Split into test and train sets

```r
fifa <- read.csv("international_matches.csv")

# Make variable for total goals scored in a game
fifa_df <- fifa %>%
  mutate(total_goals = (fifa_df$home_team_score + fifa_df$away_team_score))

# Mutated date variable, arranged by current
fifa_current <- fifa_df %>%
  mutate(date = ymd(date), neutral_location = as.logical(neutral_location)) %>%
  arrange(desc(date))

# Creating home and away team tables
home_rank <- fifa_current %>%
  select(date, home_team, home_team_score, away_team_score, home_team_continent, home_team_fifa_rank, home_team_total_fifa_points, home_team_score,
tournament, neutral_location, home_team_result, home_team_goalkeeper_score, home_team_mean_defense_score, home_team_mean_midfield_score,
home_team_mean_offense_score) %>%
  rename(country = home_team, team_score = home_team_score, opponent_score = away_team_score, country_rank = home_team_fifa_rank, continent =
home_team_continent, fifa_points = home_team_total_fifa_points, score = home_team_score, tournament = tournament, neutral_location = neutral_location,
home_team_result = home_team_result, goalkeeper_score = home_team_goalkeeper_score, defense_score = home_team_mean_defense_score, midfield_score =
home_team_mean_midfield_score, offense_score = home_team_mean_offense_score)

away_rank <- fifa_current %>%
  select(date, away_team, away_team_score, home_team_score, away_team_continent, away_team_fifa_rank, away_team_total_fifa_points, away_team_score,
tournament, neutral_location, home_team_result, away_team_goalkeeper_score, away_team_mean_defense_score, away_team_mean_midfield_score,
away_team_mean_offense_score) %>%
  rename(country = away_team, team_score = away_team_score, opponent_score = home_team_score, country_rank = away_team_fifa_rank, continent =
away_team_continent, fifa_points = away_team_total_fifa_points, score = away_team_score, tournament = tournament, neutral_location = neutral_location,
home_team_result = home_team_result, goalkeeper_score = away_team_goalkeeper_score, defense_score = away_team_mean_defense_score, midfield_score =
away_team_mean_midfield_score, offense_score = away_team_mean_offense_score)

home_rank <- home_rank %>%
  mutate(home_away = 'H')
away_rank <- away_rank %>%
  mutate(home_away = 'A')

# Combining home and away tables
ranking_all <- drop_na(rbind(home_rank, away_rank))

# Mutating variables
ranking_all <- ranking_all %>%
  mutate(team= as.factor(country),
         tournament = as.factor(tournament),
         neutral_location = as.logical(neutral_location),
         home_away = as.factor(home_away),
         continent=as.factor(continent),
         country_rank = as.factor(country_rank),
         home_team = as.logical(ifelse(home_away=='H','TRUE', 'FALSE')),
         offense_score = as.numeric(offense_score),
         team_win = as.logical(ifelse(score>opponent_score, 'TRUE', 'FALSE')))

# Create train and test sets
ranking_all_split <- initial_split(ranking_all, prop = 0.75)

ranking_train <- training(ranking_all_split)
ranking_test <- testing(ranking_all_split)

# Creating table w/ home and away that also includes null values for summaries
ranking_all_na <- rbind(home_rank, away_rank)
```

```
> dim(ranking_all)
[1] 13403    18
```

sumtable {vtable} Summary Statistics

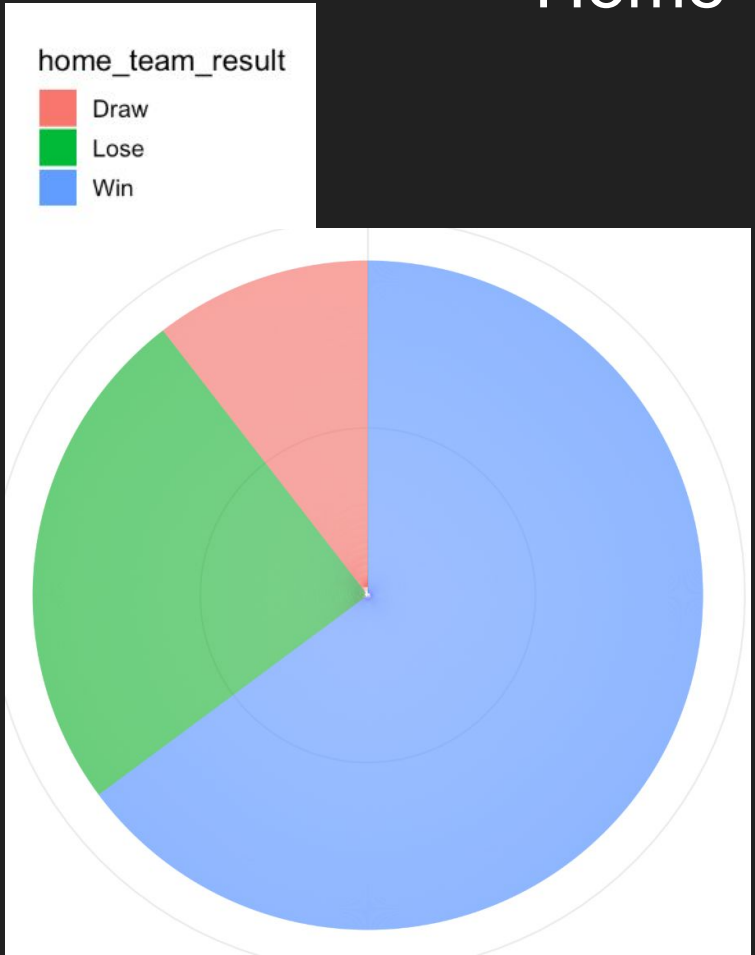
Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
score	47842	1.339	1.483	0	0	2	31
opponent_score	47842	1.339	1.483	0	0	2	31
continent	47842						
... Africa	12191	25.5%					
... Asia	10119	21.2%					
... Europe	14952	31.3%					
... North America	5475	11.4%					
... Oceania	1105	2.3%					
... South America	4000	8.4%					
country_rank	47842	79.326	52.816	1	35	117	211
fifa_points	47842	319.428	495.92	0	0	537	2164
neutral_location	47842						
... FALSE	35894	75%					
... TRUE	11948	25%					
home_team_result	47842						
... Draw	10778	22.5%					
... Lose	13542	28.3%					
... Win	23522	49.2%					
goalkeeper_score	16474	74.595	8.227	47	69	80	97
defense_score	15351	74.667	5.975	52.8	70.8	78.2	91.8
midfield_score	16141	75.578	6.096	54.2	72	79.2	93.2
offense_score	16822	75.622	6.239	53.3	71.3	79.7	93
home_away	47842						
... A	23921	50%					
... H	23921	50%					

```
> dim(ranking_all_na)
[1] 47842    16
```

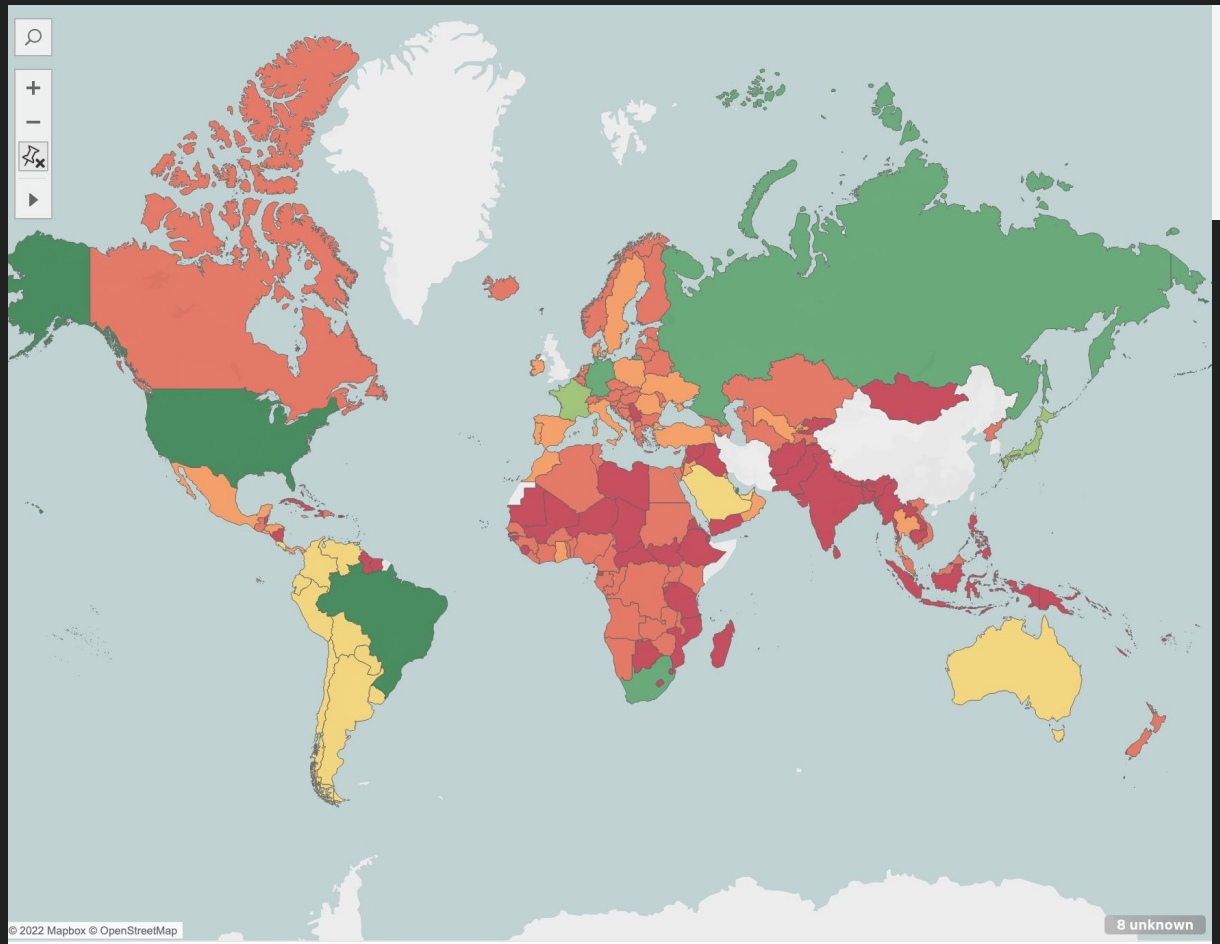
tournament	47842	
... ABCS Tournament	16	0%
... AFC Asian Cup	462	1%
... AFC Asian Cup qualification	1082	2.3%
... AFC Challenge Cup	200	0.4%
... AFC Challenge Cup qualification	178	0.4%
... AFF Championship	572	1.2%
... AFF Championship qualification	4	0%
... African Cup of Nations	980	2%
... African Cup of Nations qualification	2548	5.3%
... African Nations Championship	450	0.9%
... African Nations Championship qualification	128	0.3%
... Afro-Asian Games	8	0%
... AmÃlcar Cabral Cup	146	0.3%
... Arab Cup	142	0.3%
... Arab Cup qualification	32	0.1%
... Baltic Cup	92	0.2%
... CECAFA Cup	616	1.3%
... CFU Caribbean Cup	216	0.5%
... CFU Caribbean Cup qualification	512	1.1%
... CONCACAF Nations League	252	0.5%
... CONCACAF Nations League qualification	94	0.2%
... Confederations Cup	272	0.6%
... CONMEBOLâ€“UEFA Cup of Champions	2	0%
... Copa AmÃ©rica	588	1.2%
... Copa AmÃ©rica qualification	4	0%
Copa Confraternida	2	0%
Copa del PacÃfico	8	0%
Copa Paz del Chaco	10	0%
COSAFA Cup	618	1.3%
COSAFA Cup qualification	54	0.1%
Cup of Ancient Civilations	4	0%
Cyprus International Tournament	140	0.3%
Dragon Cup	8	0%
Dunhill Cup	30	0.1%
Dynasty Cup	28	0.1%
EAFF Championship	214	0.4%
FIFA World Cup	864	1.8%
FIFA World Cup qualification	11056	23.1%
Friendly	17116	35.8%
Gold Cup	582	1.2%
Gold Cup qualificatio	74	0.2%
Gulf Cup	408	0.9%
Intercontinental Cup	28	0.1%
King Hassan II Tournament	24	0.1%
King's Cup	152	0.3%
Kirin Challenge Cup	40	0.1%
Kirin Cup	98	0.2%
Korea Cup	28	0.1%
Lunar New Year Cu	42	0.1%
Mahinda Rajapaksa Cup	14	0%
Malta International Tournament	74	0.2%
Merdeka Tournamen	60	0.1%
...ournament		
... Merdeka Tournamen t	60	0.1%
... Millennium Cup	20	0%
... Nations Cup	12	0%
... Navruz Cup	8	0%
... Nehru Cup	92	0.2%
... Nile Basin Tourname nt	28	0.1%
... Nordic Championshi p	20	0%
... Oceania Nations Cu p	144	0.3%
... Oceania Nations Cu p qualification	82	0.2%
... OSN Cup	8	0%
... Pacific Games	74	0.2%
... Pacific Mini Games	20	0%
... Prime Minister's Cup	44	0.1%
... SAFF Cup	288	0.6%
... Simba Tournament	34	0.1%
... SKN Football Festival	18	0%
... South Pacific Game s	68	0.1%
... SuperclÃ¡sico de las AmÃ©ricas	16	0%
... Three Nations Cup	6	0%
... TIFOCO Tournamen t	2	0%
... Tournoi de France	12	0%
... UEFA Euro	506	1.1%
... UEFA Euro qualification	3446	7.2%
... UEFA Nations League	830	1.7%
... UNCAF Cup	290	0.6%
... UNIFFAC Cup	30	0.1%

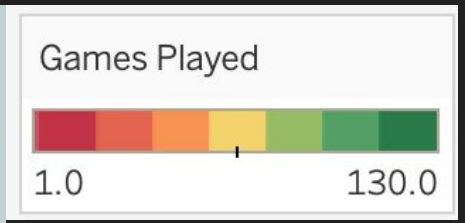
Home Team Advantage?



```{r}

ranking_all_nn <- ranking_all %>% filter(neutral_location == FALSE)

ggplot(data = ranking_all_nn, aes(x="", y=home_team_result, fill=home_team_result)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)

```

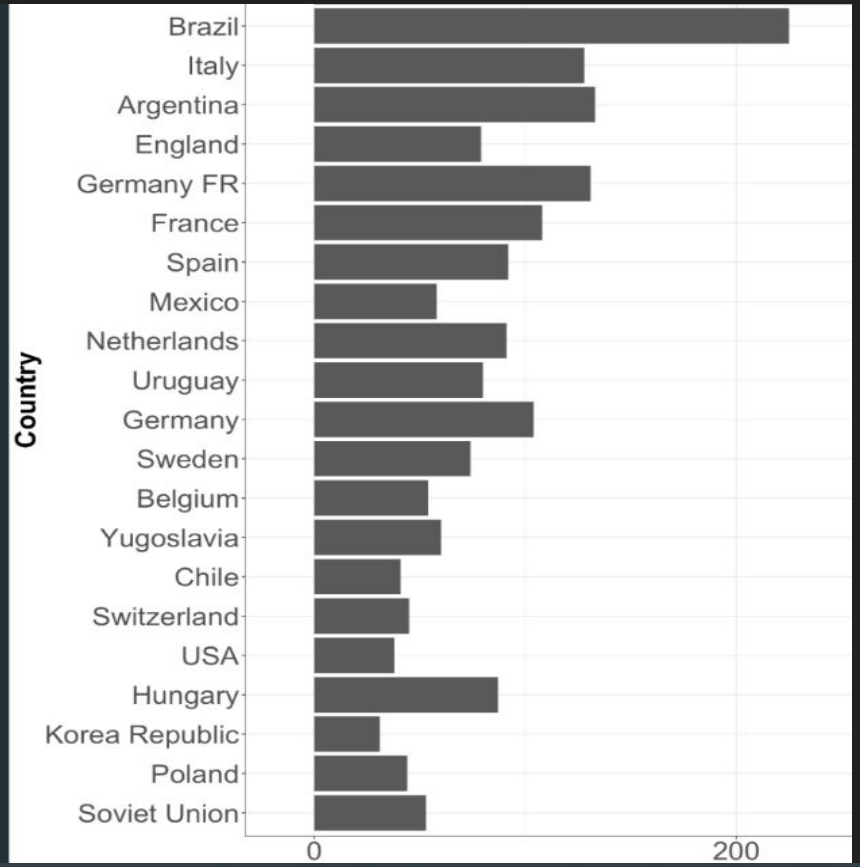
Amount of World Cup + qualifier games played per country





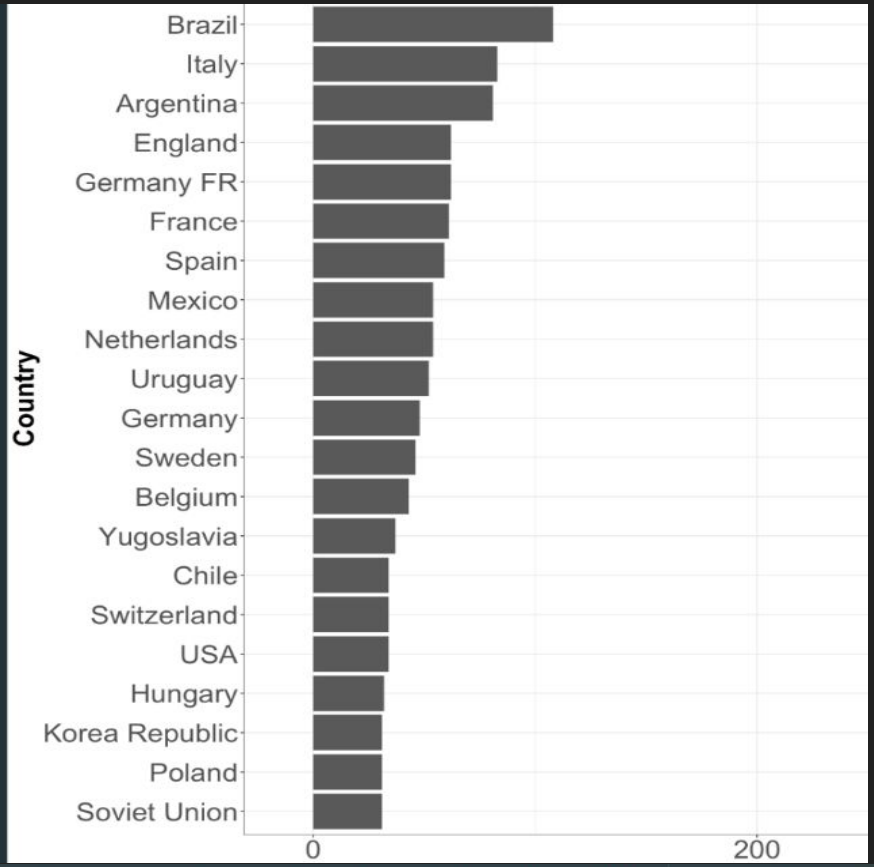
*using Tableau

Total Goals Scored in World Cup




```
goals_hist <- goals_country %>% ggplot() + aes(x = Goals, y = fct_lump_n(fct_rev(fct_infreq(Country)), n = 20)) +
geom_col() + labs(x = "Total Goals Scored in World Cup", y = "Country", title = "Top 20 Countries in World Cup
Goals Scored", caption = "*disregard other")+theme(axis.text=element_text(size=30),
        axis.title=element_text(size=30,face="bold"))
goals_hist
```

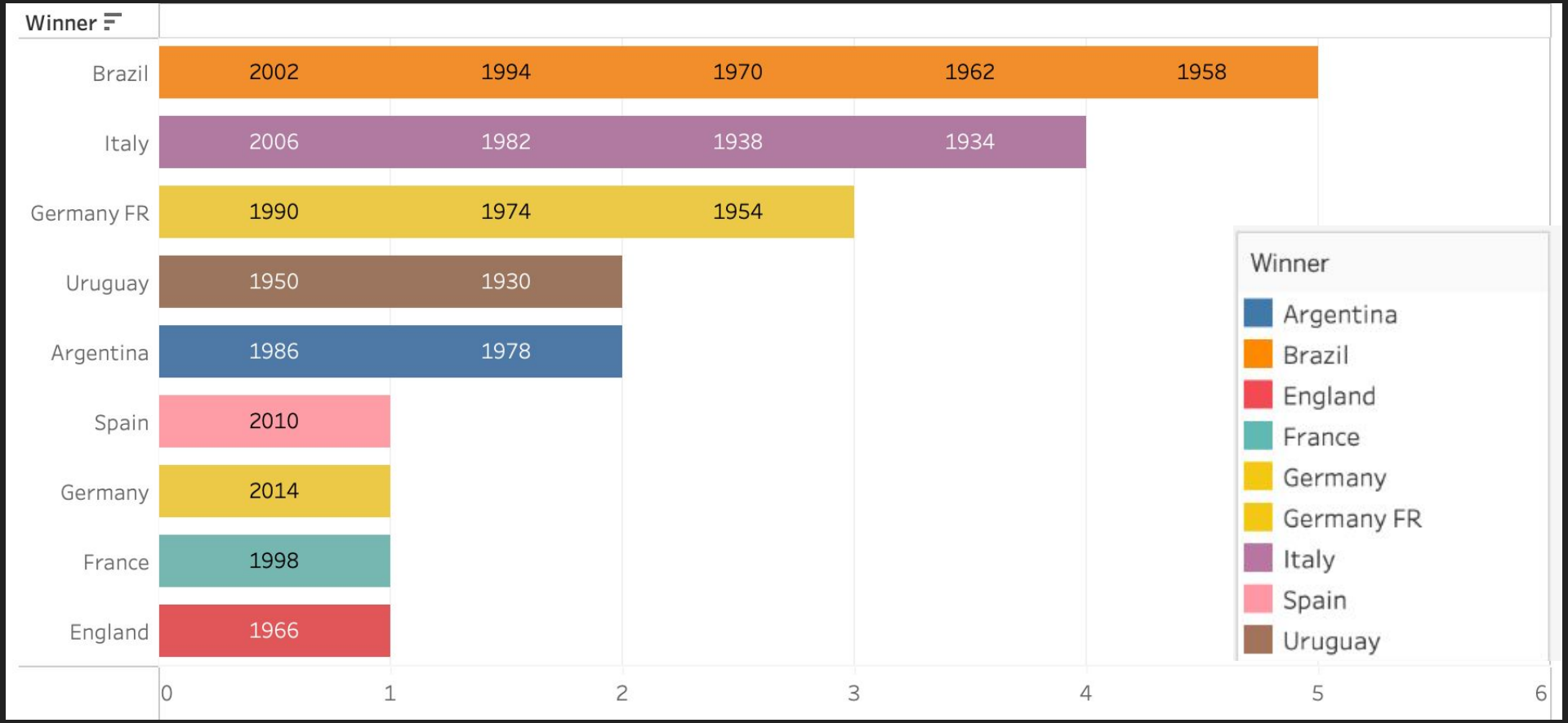
Total Games Played in World Cup




```
ggplot(goals_country, aes(fct_lump_n(fct_rev(fct_infreq(Country)), n = 20))) + geom_histogram(stat = 'count',
binwidth = 50) + coord_flip() + labs(x = "Country", y = "Total Games Played in World Cup", title = "Top 20
Countries in World Cup Games Played", caption = "*disregard other")+theme(axis.text=element_text(size=30),
        axis.title=element_text(size=30,face="bold"))
```

World Cup Winners

*using Tableau



Current Country Rankings

```
head(current_fifa_ranking, 30)
```

Country <chr>	current_country_rank <int>
Brazil	1
Belgium	2
France	3
Argentina	4
England	5
Italy	6
Spain	7
Portugal	8
Mexico	9
Netherlands	10

Country <chr>	current_country_rank <int>
Denmark	11
Germany	12
Uruguay	13
Switzerland	14
USA	15
Croatia	16
Colombia	17
Wales	18
Sweden	19
Senegal	20

Country <chr>	current_country_rank <int>
IR Iran	21
Peru	22
Japan	23
Morocco	24
Serbia	25
Poland	26
Ukraine	27
Chile	28
Korea Republic	29
Nigeria	30

Predictive Model - Classification

- Edited dataset to show:

1. Only show matches played in FIFA World Cup + qualification matches
2. Only show matches from 2018-now to show current ranks
3. Filter teams and only show teams left in Round of 16 World Cup 2022

```
world_cup_t <- ranking_train %>% filter(grepl('FIFA World Cup', tournament))
world_cup_s <- ranking_test %>% filter(grepl('FIFA World Cup', tournament))

world_cup_t <- with(ranking_train, ranking_train[(date >= "2018-01-01"), ])

world_cup_2022_16_t <- world_cup_t[world_cup_t$team %in% c('Ecuador', 'Senegal', 'Netherlands', 'England', 'USA', 'Argentina', 'Poland',
'France', 'Australia', 'Spain', 'Japan', 'Morocco', 'Croatia', 'Brazil', 'Switzerland', 'Portugal'), ]

world_cup_s <- with(ranking_test, ranking_test[(date >= "2018-01-01"), ])

world_cup_2022_16_s <- world_cup_s[world_cup_s$team %in% c('Ecuador', 'Senegal', 'Netherlands', 'England', 'USA', 'Argentina', 'Poland',
'France', 'Australia', 'Spain', 'Japan', 'Morocco', 'Croatia', 'Brazil', 'Switzerland', 'Portugal'), ]

clean_16cc_t <- data.frame(world_cup_2022_16_t)

clean_16cc_s <- data.frame(world_cup_2022_16_s)
```

Classification

- Predicting probability of teams in the top 16 winning any given game in the World Cup or qualifier

```
logit_fitc_t <- glm(team_win ~ team + goalkeeper_score +
                      defense_score + midfield_score + offense_score,
                    family = binomial,
                    data = clean_16_cw_t)

summary(logit_fitc_t)
```

- Variables used: team_win, goalkeeper_score, defense_score, midfield_score, offense_score

```
Coefficients:
                  Estimate
(Intercept)      -1.36541
teamAustralia     0.27690
teamBrazil        0.45166
teamCroatia       0.37502
teamEcuador      -0.98693
teamEngland      -0.34946
teamFrance        0.32190
teamJapan         0.47778
teamMorocco       0.04302
teamNetherlands  -0.46415
teamPoland       -1.10609
teamPortugal     -0.00839
teamSenegal       0.12786
teamSpain        -0.72666
teamSwitzerland  -1.11818
teamUSA          -0.22664
goalkeeper_score  0.02402
defense_score     0.15568
midfield_score   -0.15804
offense_score     0.00292
---
```

Classification

- Highest win rate amongst teams in the World Cup Top 16:
  1. Japan - 61.2% higher win rate than average
  2. Brazil - 57.1% higher win rate than average
  3. Croatia - 45.5% higher win rate than average

- Coefficients of defense, offense, etc. carry a rank to how they contribute to a win

```
exp(logit_fitc_t$coefficients)
```

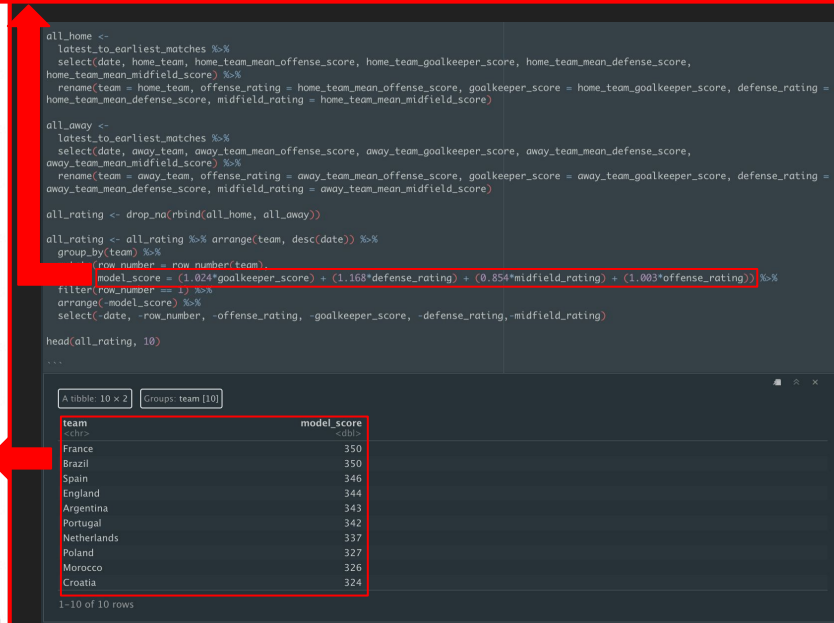
```
    (Intercept)     teamAustralia        teamBrazil       teamCroatia      teamEcuador      teamEngland       teamFrance
          0.255             1.319             1.571             1.455            0.373            0.705            1.380
     teamJapan      teamMorocco   teamNetherlands       teamPoland      teamPortugal      teamSenegal        teamSpain
          1.612             1.044             0.629             0.331            0.992            1.136            0.484
teamSwitzerland         teamUSA goalkeeper_score   defense_score   midfield_score   offense_score
          0.327             0.797             1.024             1.168            0.854            1.003
```

Classification

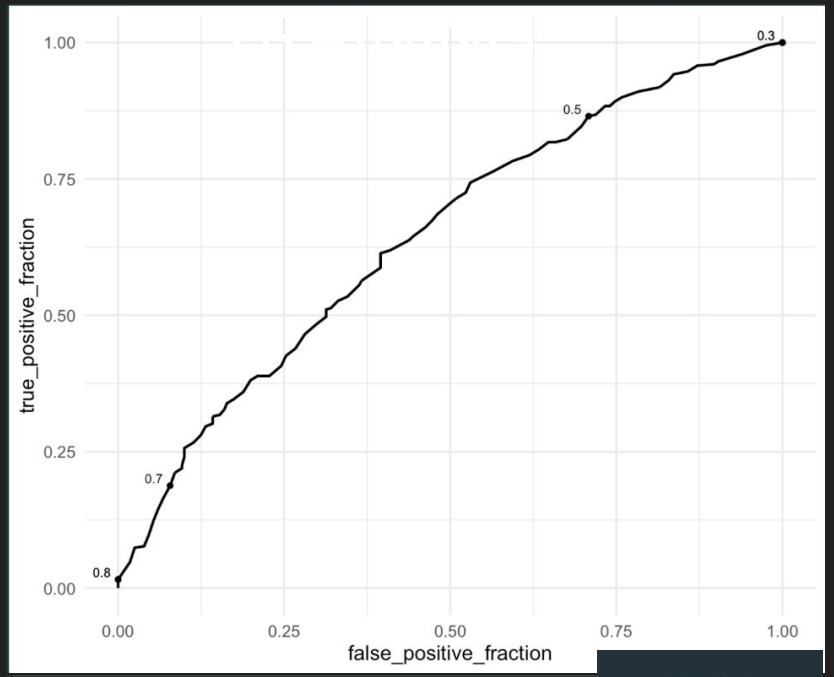
goalkeeper_score	defense_score	midfield_score	offense_score
1.024	1.168	0.854	1.003

```
all_rating <- all_rating %>%
  mutate(model_score = (1.024*goalkeeper_score) + (1.168*defense_rating) + (0.854*midfield_rating) + (1.003*offense_rating))
```

team	model_score
<chr>	<dbl>
France	350.1257
Brazil	349.7583
Spain	345.7470
England	344.2720
Argentina	343.0286
Portugal	341.9026
Netherlands	337.0156
Poland	326.6089
Morocco	325.8295
Croatia	323.6753

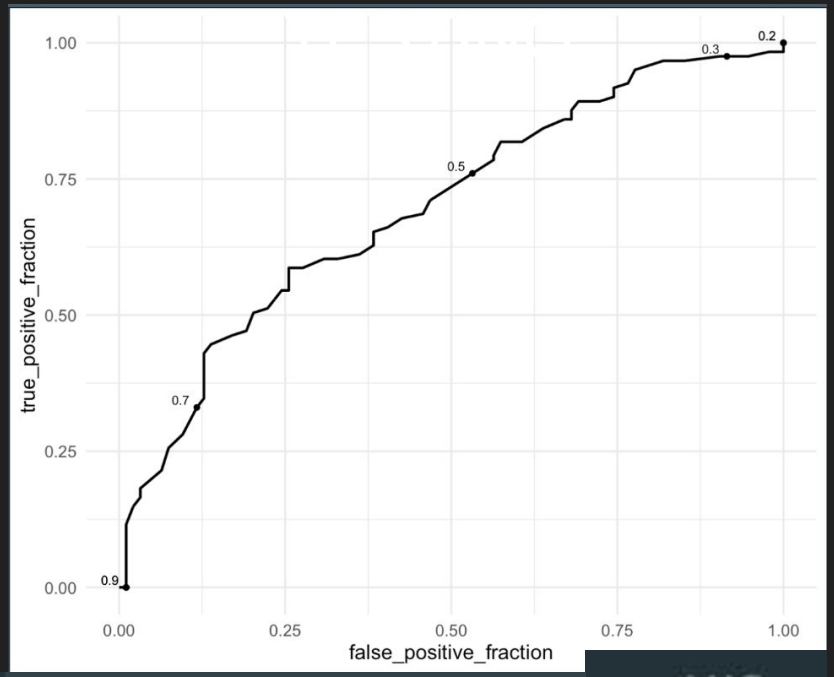


Classification




```
p_t <- ggplot(results_trainc_t,
              aes(m = prob_eventc_t, d = true_classc_t)) +
  geom_roc(labelsize = 3.5,
           cutoffs.at =
             c(0.99,0.9,0.7,0.5,0.3,0.1,0)) +
  theme_minimal(base_size = 16)
print(p_t)
```

AUC
<dbl>
0.644




```
p_s <- ggplot(results_trainc_s,
              aes(m = prob_eventc_s, d = true_classc_s)) +
  geom_roc(labelsize = 3.5,
           cutoffs.at =
             c(0.99,0.9,0.7,0.5,0.3,0.1,0)) +
  theme_minimal(base_size = 16)
print(p_s)
```

AUC
<dbl>
0.697

Performance of Model

- AUC results are considered poor
- Underfit model
  - Due to low AUC scores and
  - Testing score is greater than training score

Conclusion

- Nothing can ever 100% predict who is going to win a match
  - But predictive models give us a good estimate
- The classification model does not have a great performance score, but since our objective is to help gamblers, they may still trust it