Alunos: Iago Leobas e Brayan Mota

Disciplina: Mineração de Dados

Professora: Tamirys Virgulino Ribeiro Prado

Data: 25/11/2022

Análise de Modelo Estatístico

A partir de conceitos adquiridos sobre modelo estatístico:

- Defina um DATASET
- Definir população e AMOSTRA PELO MÉTODO CIENTÍFICO
- Documentação DESCRITIVA e REPRESENTATIVA

Tema Dataset: FIFA World Cup

Link do kaggle: https://www.kaggle.com/datasets/abecklas/fifa-world-cup

Resolução:

Inicialmente, é realizada a importação do dataset e apresentação dos dados:

O dataset escolhido tem como foco apresentar dados das copas do mundo de 1930 até 2014. Ele contém no total 10 **colunas** e 20 **linhas.**

Dentre essas 10 colunas, são elas:

Nome	Descrição
Year	Ano da copa
Country	País que sediou a copa
Winner	Ganhador da copa
Runners-Up	Segundo colocado
Third	Terceiro colocado
Fourth	Quarto colocado
GoalsScored	Total de gols feitos na copa
QualifiedTeams	Total de times que participaram
MatchesPlayed	Partidas realizadas
Attendance	Público total
10 Colunas	**20 Linhas**

	Year	Country	Winner	Runners-Up	Third	Fourth	GoalsScored	QualifiedTeams	MatchesPlayed	Attendance
0	1930	Uruguay	Uruguay	Argentina	USA	Yugoslavia	70	13	18	590.549
1	1934	Italy	Italy	Czechoslovakia	Germany	Austria	70	16	17	363.000
2	1938	France	Italy	Hungary	Brazil	Sweden	84	15	18	375.700
3	1950	Brazil	Uruguay	Brazil	Sweden	Spain	88	13	22	1.045.246
4	1954	Switzerland	Germany FR	Hungary	Austria	Uruguay	140	16	26	768.607

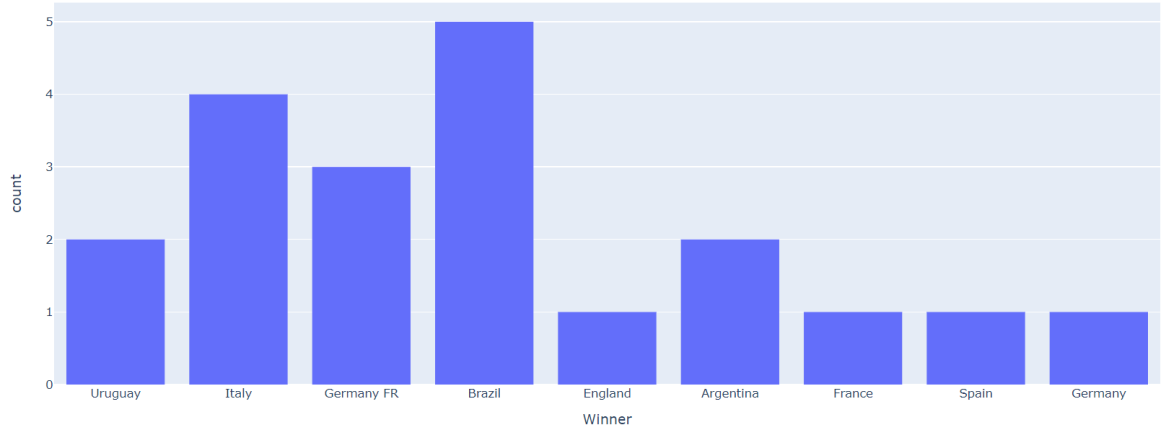
Sendo o dataset bastante curto, não seria interessante realizarmos a geração de uma amostra, pois seria reduzido mais ainda o novo valor. Porém podemos realizar uma geração para podermos ver o que poderá ser feito dela.

Então fazemos assim:

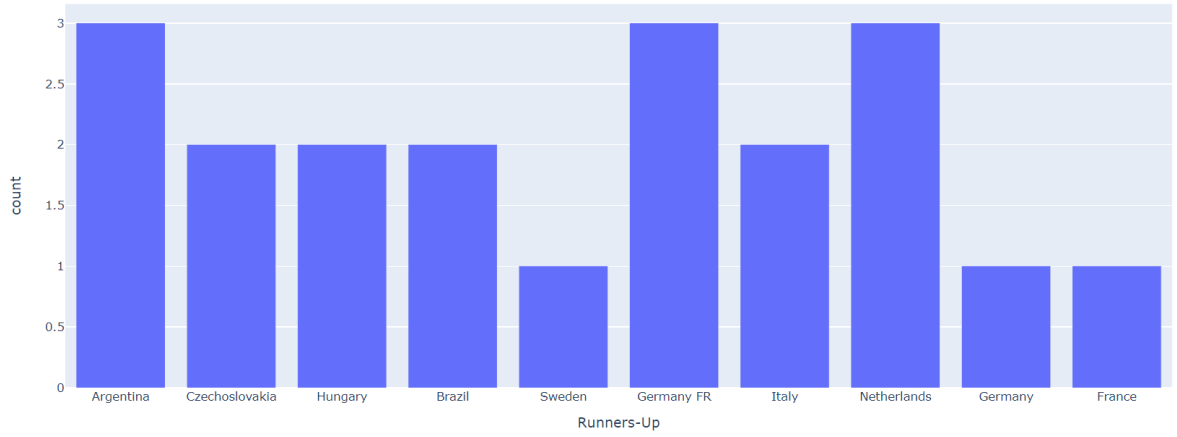
```
colunas = [
  'Year',
  'Country',
  'Winner',
  'Runners-Up',
  'Third',
  'Fourth',
  # 'GoalsScored',
  'QualifiedTeams',
  'MatchesPlayed',
  'Attendance',
]

x = data[colunas]
y = data['GoalsScored']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Nessa divisão foi passado um valor de 25% do dataset original para a criação do dataset de teste, então o dataset de treino ficou com 15 linhas e o de teste ficou com 5. São pouquíssimos dados para serem utilizados, que nesse caso só servirão para demonstração.

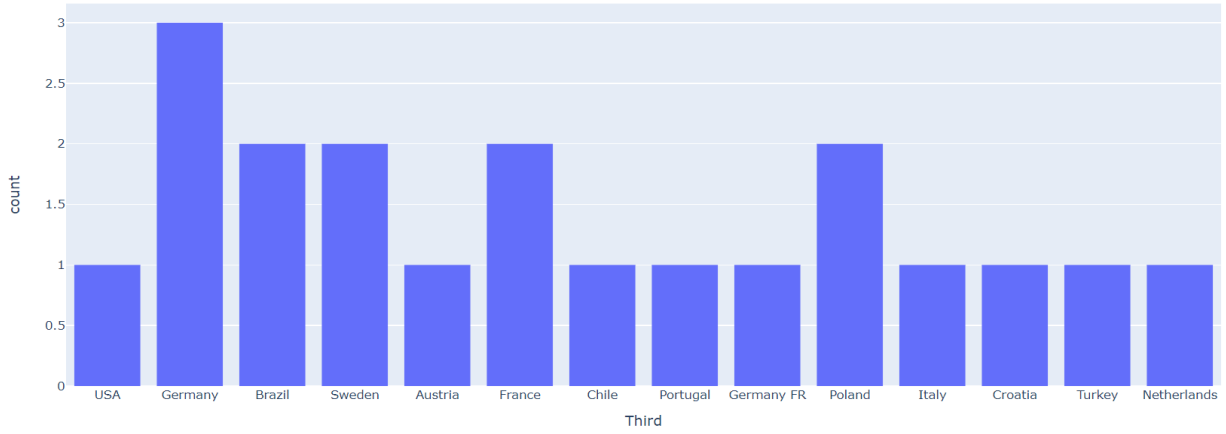
A seguir serão apresentados alguns gráficos obtidos com o dataset com informações interessantes sobre as copas do mundo.



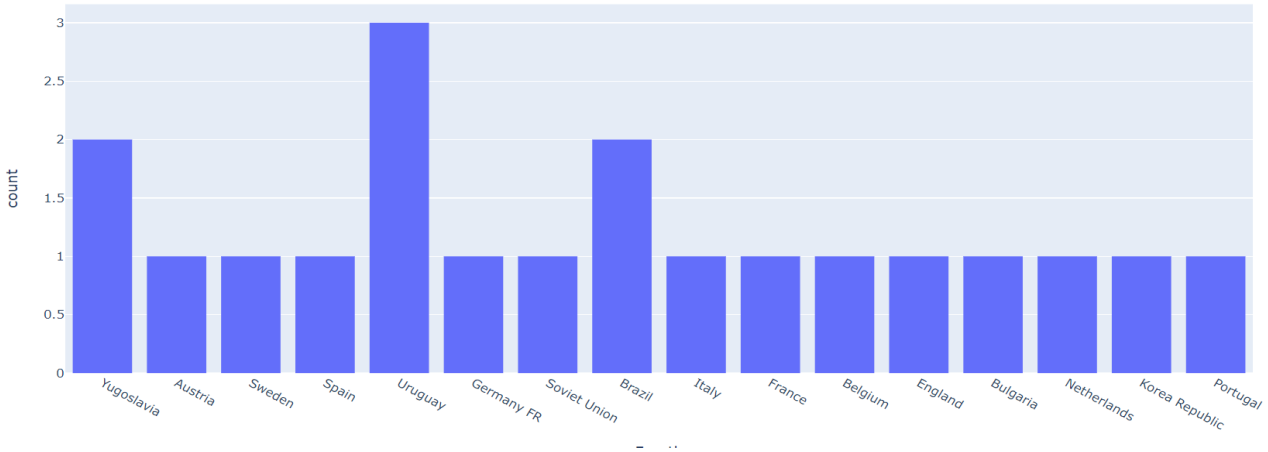
Inicialmente podemos apresentar os países que mais venceram a copa do mundo, tendo o Brasil (5 vitórias) como aquele que tem o maior número de vitórias na competição, seguido da Itália (4 vitórias) e Alemanha (3 vitórias).



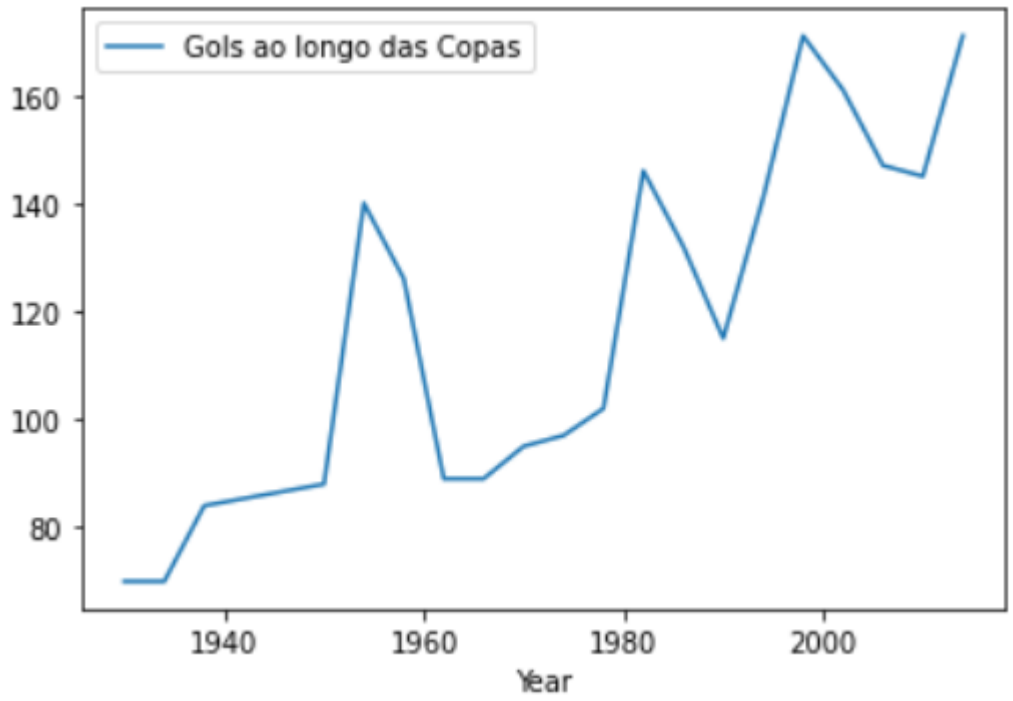
Seguindo temos os países que mais ficaram em segundo colocado nas copas, temos em primeiro lugar empatadas as seleções da Argentina (3 vezes segundo lugar), Alemanha (3 vezes segundo lugar) e Holanda (3 vezes segundo lugar). Seguidos por Brasil (2 vezes segundo lugar), Hungria (2 vezes segundo lugar) e Tchecoslováquia (2 vezes segundo lugar).



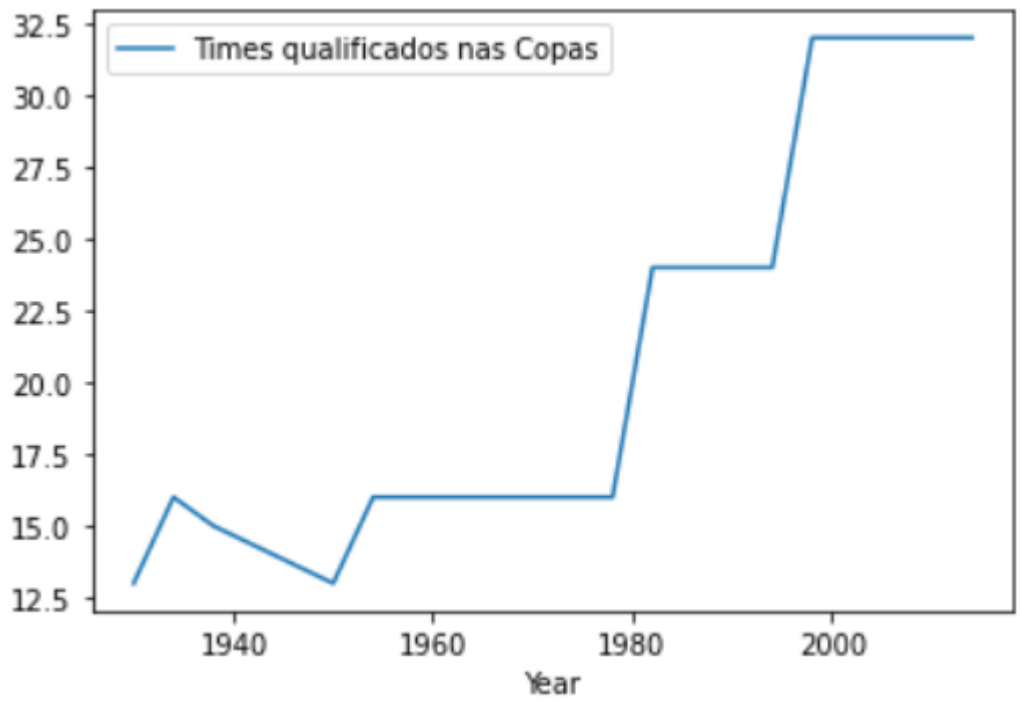
Outro dado interessante são os países que mais ficaram em terceiro lugar na competição, essa conquista foi alcançada mais vezes pela Alemanha (3 vezes terceira colocada).



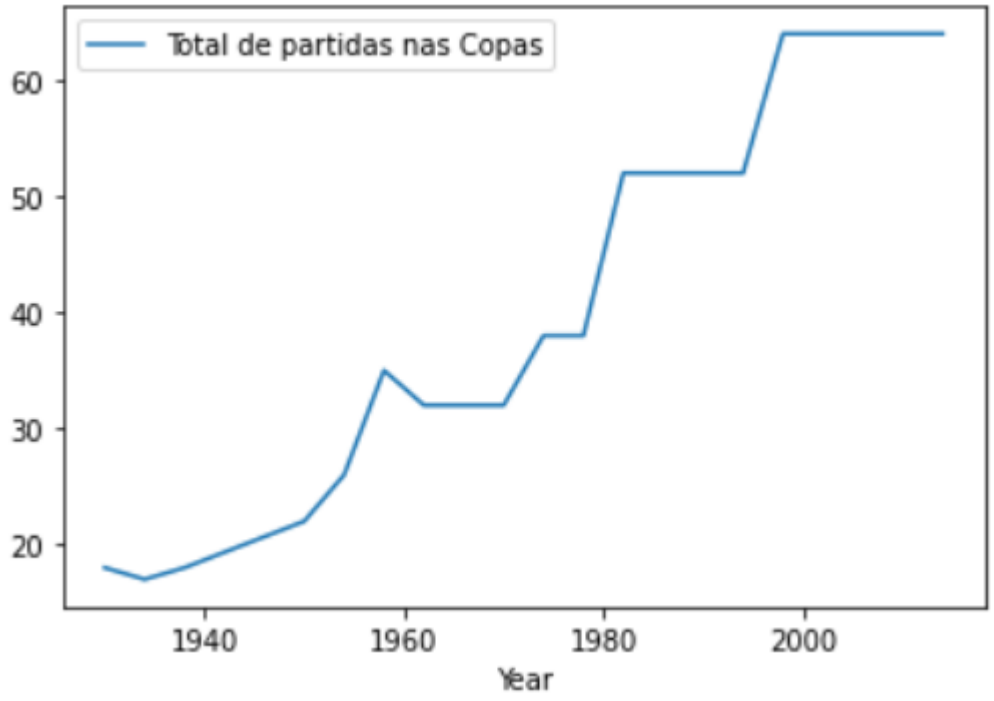
E por fim os dados de maiores quartos colocados na história da competição, com o Uruguai (3 vezes quarto colocado) encabeçando a lista.



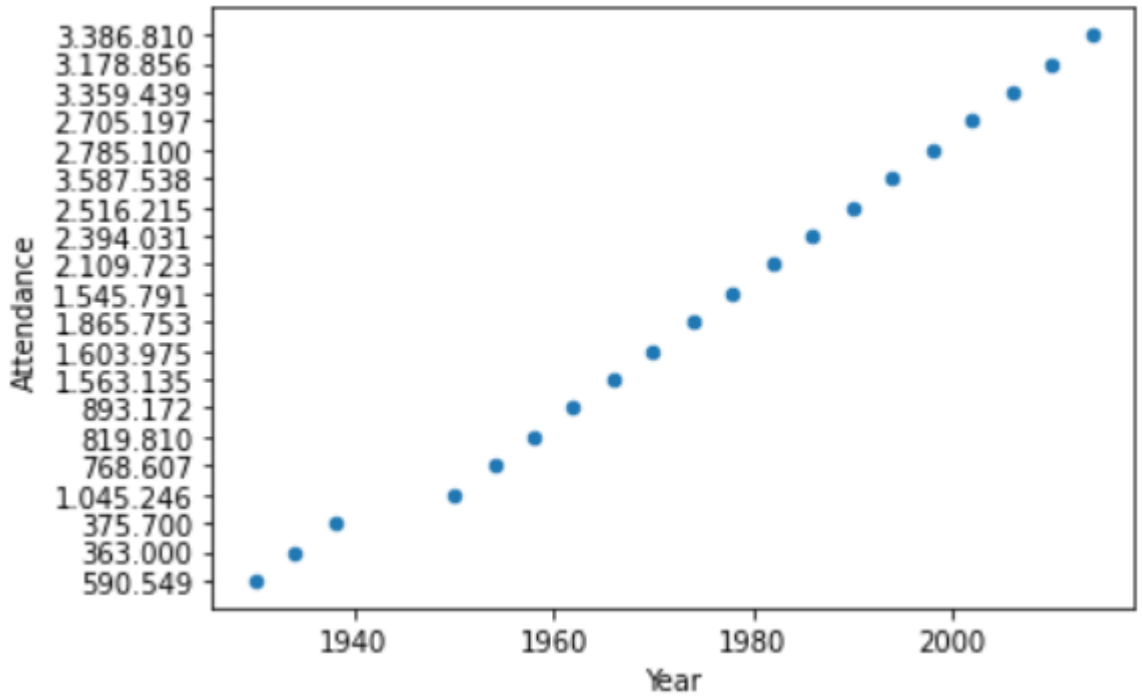
Outro dado relevante é a quantidade de gols feitos ao longo de cada edição da copa, é notório que essa quantidade vai aumentando à medida que os anos se passam. Começando com uma média de 70 gols, hoje ultrapassa os 160, mais que o dobro.



A quantidade de times envolvidos na competição também aumentou ao decorrer dos anos, começando com uma média de 13 times, hoje a média ultrapassa os 30.



A quantidade de partidas nas copas também vem aumentando ao decorrer do tempo, iniciando com menos de 20 partidas, hoje ultrapassa as 60.



Assim como o público total presente em cada copa, todos esses indicadores vêm aumentando ao decorrer dos anos.

Link do Colab:
https://colab.research.google.com/drive/1CoFDkynpEeere8lnngRvNBd5hRK1dgnX?usp=sharing