



Scope of Work for Data Science Final Project

S-109A Introduction to Data Science

Harvard University

Summer 2018

Instructors: Pavlos Protopapas, Kevin Rader

Prepared by Group #28

Mark Dinneen

Mohammad Karim

Ramandeep Harjai

Supplemental Notebook for Web Scraping

This notebook documents all the data which has been collected via web scraping technique

```
In [8]: # import the necessary libraries
%matplotlib inline
import numpy as np
import scipy as sp
import matplotlib as mpl
import matplotlib.cm as cm
import matplotlib.pyplot as plt
import pandas as pd
from pandas import Series
import random

# libraries for scraping data from web
import re
from bs4 import BeautifulSoup
from sys import argv
import requests
from urllib.request import urlopen
from urllib.error import HTTPError
```

Historical FIFA Ranking Data

The latest men's FIFA ranking data is available at: <https://www.fifa.com/fifa-world-ranking/ranking-table/men/index.html> (<https://www.fifa.com/fifa-world-ranking/ranking-table/men/index.html>), which is published on 07-June-2018. The oldest men's FIFA ranking data is available at: <https://www.fifa.com/fifa-world-ranking/ranking-table/men/rank=2/index.html> (<https://www.fifa.com/fifa-world-ranking/ranking-table/men/rank=2/index.html>), which was published on 08-August-1993. We plan to parse this data from 288+ pages, which will provide us the FIFA ranking for various men's teams, for time period between 08-August-1993, and 07-June-2018.

FIFA.com

EN

Men's Ranking

08 August 1993

<

>

All confederations / CAF / CONCACAF / CONMEBOL / OFC / AFC / UEFA

POSITIONS

1-50

51-100

101-150

151-200

201-211

Worst mover

Best mover

Rank	Team	Total Points	Previous Points	+/-	Positions	1993 (100 %)		1992 (50 %)		1991 (30 %)		1990 (20 %)		
						Avg.	AVG WGT	Avg.	AVG WGT	Avg.	AVG WGT	Avg.	AVG WGT	
1	Germany	0 (0)	57	0	◀▶	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
2	Italy	0 (0)	57	0	◀▶	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
3	Switzerland	0 (0)	50	9	▲	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
4	Sweden	0 (0)	55	0	◀▶	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
5	Argentina	0 (0)	51	5	▲	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
6	Republic of Ireland	0 (0)	54	0	◀▶	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
7	Russia	0 (0)	52	1	▲	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
8	Brazil	0 (0)	55	-5	▼	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕
9	Norway	0 (0)	49	5	▲	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	↕

FIFA ranking page for 08-August-1993

```
In [6]: def get_month_number(month_name: str) -> str:
        month_number = 0
        if month_name.lower() == "january":
            month_number = "01"
        elif month_name.lower() == "february":
            month_number = "02"
```

```

    elif month_name.lower() == "march":
        month_number = "03"
    elif month_name.lower() == "april":
        month_number = "04"
    elif month_name.lower() == "may":
        month_number = "05"
    elif month_name.lower() == "june":
        month_number = "06"
    elif month_name.lower() == "july":
        month_number = "07"
    elif month_name.lower() == "august":
        month_number = "08"
    elif month_name.lower() == "september":
        month_number = "09"
    elif month_name.lower() == "october":
        month_number = "10"
    elif month_name.lower() == "november":
        month_number = "11"
    elif month_name.lower() == "december":
        month_number = "12"
    return month_number

def get_year(rank_date: str) -> str:
    rank_date_split = rank_date.split(" ")
    return rank_date_split[2]

def format_date(rank_date: str) -> str:
    rank_date_split = rank_date.split(" ")
    rank_date_formatted = rank_date_split[2] + "-" + \
        get_month_number(rank_date_split[1]) + "-" + \
        rank_date_split[0]
    return rank_date_formatted

def parse_rank_data() -> []:
    ranking = []
    page_url = "https://www.fifa.com/fifa-world-ranking/ranking-table/men/rank={}/index.html"
    page_ids = range(2, 288)

    for page_id in page_ids:
        progress = round( (page_id / max(page_ids))*100 )
        print("Parsing FIFA ranking data: {}%\r".format(progress), end=" ", flush=True)
        rank_page = requests.get(page_url.format(page_id))
        page_soup = BeautifulSoup(rank_page.text, 'html.parser')
        rank_date = page_soup.findAll("ul", {"class":["slider-list","items-1"]})[0].text
        rank_soup = page_soup.findAll("tr", {"class":"anchor"})
        for rank_tag in rank_soup:
            date = format_date(rank_date)
            year = get_year(rank_date)
            rank = rank_tag.find("td", {"class":"tbl-rank"}).text

```

```

        team = rank_tag.find("td", {"class":"tbl-teamname"}).text
        ranking.append([date, year, team, rank])

    return ranking

# un-comment the lines below run data scraping
# fifa_ranking = parse_rank_data() # get FIFA ranking data from fifa.
# com website
# df_ranking = pd.DataFrame(fifa_ranking, columns=["date", "year", "te
# am", "rank"]) # create dataframe
# df_ranking.to_pickle("datasets/ranking_data.pkl") # store dataframe
# to local disk

# read back web scraped dataset
df_ranking = pd.read_pickle("datasets/ranking_data.pkl")
print("\ndf_ranking = ", df_ranking.shape, "\n")
df_ranking.sample(10)

df_ranking = (57793, 4)

```

Out[6]:

	date	year	team	rank
29624	2007-03-14	2007	Romania	14
2466	1994-12-20	1994	Malawi	82
13262	2000-04-12	2000	Samoa	183
14642	2000-11-15	2000	St Kitts and Nevis	146
53579	2016-11-24	2016	France	7
2533	1994-12-20	1994	Guatemala	149
22041	2004-01-14	2004	Croatia	20
32770	2008-06-04	2008	Tunisia	50
41035	2011-11-23	2011	Hungary	37
15411	2001-03-14	2001	Syria	103

Historical FIFA World Cup Winners

The historical FIFA world-cup finals data is available at:

https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_finals

(https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_finals), which is last published (updated) on 25-July-2018. We'll scrape this data to build a dataset of historical FIFA world cup winners.

https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_finals							
List of finals matches, their venues and locations, the finalists, and final scores							
Year	Winners	Final score ^[2]	Runners-up	Venue	Location	Attendance	References
1930	 Uruguay	4–2	 Argentina	Estadio Centenario	Montevideo, Uruguay	80,000	[7][8]
1934	 Italy	2–1 [†] [n 3]	 Czechoslovakia	Stadio Nazionale PNF	Rome, Italy	50,000	[9][10]
1938	 Italy	4–2	 Hungary	Stade Olympique de Colombes	Paris, France	45,000	[11][12]
1950 ^[n 4]	 Uruguay	2–1 [n 5]	 Brazil	Estádio do Maracanã	Rio de Janeiro, Brazil	199,854 ^[13]	[14][15]
1954	 West Germany	3–2	 Hungary	Wankdorf Stadium	Bern, Switzerland	60,000	[16][17]
1958	 Brazil	5–2	 Sweden	Råsunda Stadium	Solna, Sweden	51,800	[18][19]
1962	 Brazil	3–1	 Czechoslovakia	Estadio Nacional	Santiago, Chile	69,000	[20][21]
1966	 England	4–2 [†] [n 6]	 West Germany	Wembley Stadium	London, England	93,000	[22][23]
1970	 Brazil	4–1	 Italy	Estadio Azteca	Mexico City, Mexico	107,412	[24][25]
1974	 West Germany	2–1	 Netherlands	Olympiastadion	Munich, West Germany	75,200	[26][27]
1978	 Argentina	3–1 [†] [n 7]	 Netherlands	Estadio Monumental	Buenos Aires, Argentina	71,483	[28][29]
1982	 Italy	3–1	 West Germany	Santiago Bernabéu	Madrid, Spain	90,000	[30][31]
1986	 Argentina	3–2	 West Germany	Estadio Azteca	Mexico City, Mexico	114,600	[32][33]
1990	 West Germany	1–0	 Argentina	Stadio Olimpico	Rome, Italy	73,603	[34][35]
1994	 Brazil	0–0 [†] [n 8]	 Italy	Rose Bowl	Pasadena, United States	94,194	[36][37]
1998	 France	3–0	 Brazil	Stade de France	Saint-Denis, France	80,000	[38][39]
2002	 Brazil	2–0	 Germany	International Stadium	Yokohama, Japan	69,029	[40][41]
2006	 Italy	1–1 [†] [n 9]	 France	Olympiastadion	Berlin, Germany	69,000	[42][43]
2010	 Spain	1–0 [†] [n 10]	 Netherlands	Soccer City	Johannesburg, South Africa	84,490	[44][45]
2014	 Germany	1–0 [†] [n 11]	 Argentina	Estádio do Maracanã	Rio de Janeiro, Brazil	74,738	[46][47]
2018	 France	4–2	 Croatia	Luzhniki Stadium	Moscow, Russia	78,011	
Upcoming finals							
Year	Team 1	v	Team 2	Venue	Location	Attendance	References
2022				Lusail Iconic Stadium	Lusail, Qatar		
2026				TBD	North America		

Historical FIFA World Cup Final Match Results

```
In [103]: def parse_past_winners() -> []:
            url = "https://en.wikipedia.org/wiki/List_of_FIFA_World_Cup_finals"
            "
            page = requests.get(url)
            page_soup = BeautifulSoup(page.text, 'html.parser')
            table_soup = page_soup.findAll("tbody")[2]
```

```

row_soup = table_soup.findAll("tr")

results = []
for row in range(1,len(row_soup)-4):
    result = []
    result.append(re.findall(r"[0-9]+", row_soup[row].find("th").text)[0])
    result.append(row_soup[row].findAll("td")[0].find("a").text)
    result.append(row_soup[row].findAll("td")[2].find("a").text)
    result.append(re.findall(r"[0-9]+", row_soup[row].findAll("td")[1].find("a").text)[0])
    result.append(re.findall(r"[0-9]+", row_soup[row].findAll("td")[1].find("a").text)[1])
    result.append(row_soup[row].findAll("td")[0].find("a").text)
    result.append(row_soup[row].findAll("td")[3].findAll("a")[0].text)
    result.append(row_soup[row].findAll("td")[4].findAll("a")[0].text)
    result.append(row_soup[row].findAll("td")[4].findAll("a")[1].text)
    result.append(re.findall(r"[0-9,]+", row_soup[row].findAll("td")[5].text)[0].replace(',',''))
    results.append(result)

return results

# un-comment the lines below run data scraping
# colnames = ['year','team1','team2','team1_score','team2_score',
#             'winner','venue','city','country','attendance']
# df_fifa_finals = pd.DataFrame.from_records(parse_past_winners(), columns=colnames)
# df_fifa_finals["year"] = df_fifa_finals["year"].astype(int)
# df_fifa_finals["team1"] = df_fifa_finals["team1"].astype(str)
# df_fifa_finals["team2"] = df_fifa_finals["team2"].astype(str)
# df_fifa_finals["team1_score"] = df_fifa_finals["team1_score"].astype(int)
# df_fifa_finals["team2_score"] = df_fifa_finals["team2_score"].astype(int)
# df_fifa_finals["winner"] = df_fifa_finals["winner"].astype(str)
# df_fifa_finals["venue"] = df_fifa_finals["venue"].astype(str)
# df_fifa_finals["city"] = df_fifa_finals["city"].astype(str)
# df_fifa_finals["country"] = df_fifa_finals["country"].astype(str)
# df_fifa_finals["attendance"] = df_fifa_finals["attendance"].astype(int)
# df_fifa_finals = df_fifa_finals.replace("West Germany", "Germany")
# df_fifa_finals.to_pickle("datasets/fifa_finals_data.pkl") # store dataframe to local disk

# read back web scraped dataset
df_fifa_finals = pd.read_pickle("datasets/fifa_finals_data.pkl")
print("\ndf_fifa_finals = ", df_fifa_finals.shape, "\n")
df_fifa_finals

```

```
df_fifa_finals = (21, 10)
```

```
Out[103]:
```

	year	team1	team2	team1_score	team2_score	winner	venue
0	1930	Uruguay	Argentina	4	2	Uruguay	Estadio Centenario
1	1934	Italy	Czechoslovakia	2	1	Italy	Stadio Nazionale PM
2	1938	Italy	Hungary	4	2	Italy	Stade Olympique de Colombes
3	1950	Uruguay	Brazil	2	1	Uruguay	Estádio do Maracanã
4	1954	Germany	Hungary	3	2	Germany	Wankdorf Stadium
5	1958	Brazil	Sweden	5	2	Brazil	Råsunda Stadium
6	1962	Brazil	Czechoslovakia	3	1	Brazil	Estadio Nacional
7	1966	England	Germany	4	2	England	Wembley Stadium
8	1970	Brazil	Italy	4	1	Brazil	Estadio Azteca
9	1974	Germany	Netherlands	2	1	Germany	Olympiastadion
10	1978	Argentina	Netherlands	3	1	Argentina	Estadio Monumental
11	1982	Italy	Germany	3	1	Italy	Santiago Bernabéu
12	1986	Argentina	Germany	3	2	Argentina	Estadio Azteca
13	1990	Germany	Argentina	1	0	Germany	Stadio Olimpico
14	1994	Brazil	Italy	0	0	Brazil	Rose Bowl
15	1998	France	Brazil	3	0	France	Stade de France
16	2002	Brazil	Germany	2	0	Brazil	International Stadium
17	2006	Italy	France	1	1	Italy	Olympiastadion

18	2010	Spain	Netherlands	1	0	Spain	Soccer City
19	2014	Germany	Argentina	1	0	Germany	Estádio do Maracanã
20	2018	France	Croatia	4	2	France	Luzhniki Stadium

```
In [119]: df_fifa_finals.groupby("winner").agg({
            'year': 'count'
        }).sort_values(by='year', ascending=False)
```

Out[119]:

	year
winner	
Brazil	5
Germany	4
Italy	4
Argentina	2
France	2
Uruguay	2
England	1
Spain	1

FIFA World Cup — All Time Team Rankings

FIFA all time team rankings, and associated team statistics is available at <https://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html> (<https://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html>). We'll scrape this data to build a dataset of FIFA team rankings.

FIFA World Cup™ Teams Statistics										
FIFA World Cup™ - All-time rankings										
RANK	TEAM	PTS	MP	W	D	L	GS	GA	AV. PTS	APPS.
1	 BRAZIL	227	104	70	17	17	221	102	2.2	20
2	 GERMANY	218	106	66	20	20	224	121	2.1	18
3	 ITALY	156	83	45	21	17	128	77	1.9	18
4	 ARGENTINA	140	77	42	14	21	131	84	1.8	16
5	 SPAIN	99	59	29	12	18	92	66	1.7	14
6	 ENGLAND	98	62	26	20	16	79	56	1.6	14

FIFA World Cup — All Time Team Rankings

```
In [141]: def parse_team_rankings() -> []:
            url = "https://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html"
            page = requests.get(url)
            page_soup = BeautifulSoup(page.text, 'html.parser')
            table_soup = page_soup.findAll("table", {"class":["table","tbl-all-timeranking"]})[1].find("tbody")
            row_soup = table_soup.findAll("tr")

            results = []
            for row in row_soup:
                result = []
                result.append(row.findAll("td")[0].text)
                result.append(row.findAll("td")[1].find("span", {"class":"t-nT
```

```

ext")).text)
        result.append(row.findAll("td")[3].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[4].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[5].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[6].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[7].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[8].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[9].find("span", {"class":"text
    })).text)
        result.append(row.findAll("td")[10].text)
        result.append(row.findAll("td")[11].find("span", {"class":"tex
t"})).text)
        results.append(result)

    return results

# un-comment the lines below run data scraping
# colnames = ['rank', 'team', 'points', 'matches', 'win', 'draw', 'lost',
#             'goal_for', 'goal_against', 'points_avg', 'appearances']
# df_fifa_wc_ranking = pd.DataFrame.from_records(parse_team_rankings()
, columns=colnames)
# df_fifa_wc_ranking["rank"] = df_fifa_wc_ranking["rank"].astype(int)
# df_fifa_wc_ranking["points"] = df_fifa_wc_ranking["points"].astype(i
nt)
# df_fifa_wc_ranking["matches"] = df_fifa_wc_ranking["matches"].astype
(int)
# df_fifa_wc_ranking["win"] = df_fifa_wc_ranking["win"].astype(int)
# df_fifa_wc_ranking["draw"] = df_fifa_wc_ranking["draw"].astype(int)
# df_fifa_wc_ranking["lost"] = df_fifa_wc_ranking["lost"].astype(int)
# df_fifa_wc_ranking["goal_for"] = df_fifa_wc_ranking["goal_for"].asty
pe(int)
# df_fifa_wc_ranking["goal_against"] = df_fifa_wc_ranking["goal_against"]
.astype(int)
# df_fifa_wc_ranking["points_avg"] = df_fifa_wc_ranking["points_avg"].
astype(float)
# df_fifa_wc_ranking["appearances"] = df_fifa_wc_ranking["appearances"
].astype(int)
# df_fifa_wc_ranking.to_pickle("datasets/fifa_wc_ranking.pkl") # store
dataframe to local disk

# read back web scraped dataset
df_fifa_wc_ranking = pd.read_pickle("datasets/fifa_wc_ranking.pkl")
print("\ndf_fifa_wc_ranking = ", df_fifa_wc_ranking.shape, "\n")
df_fifa_wc_ranking.head(10)

```

```
df_fifa_wc_ranking = (77, 11)
```

Out[141]:




	rank	team	points	matches	win	draw	lost	goal_for	goal_against	points_a
0	1	Brazil	227	104	70	17	17	221	102	2.2
1	2	Germany	218	106	66	20	20	224	121	2.1
2	3	Italy	156	83	45	21	17	128	77	1.9
3	4	Argentina	140	77	42	14	21	131	84	1.8
4	5	Spain	99	59	29	12	18	92	66	1.7
5	6	England	98	62	26	20	16	79	56	1.6
6	7	France	96	59	28	12	19	106	71	1.6
7	8	Netherlands	93	50	27	12	11	86	48	1.9
8	9	Uruguay	72	51	20	12	19	80	71	1.4
9	10	Sweden	61	46	16	13	17	74	69	1.3





FIFA World Cup — Participations

FIFA World Cup - team participations, and associated team statistics is available at <https://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html> (<https://www.fifa.com/fifa-tournaments/statistics-and-records/worldcup/teams/index.html>). We'll scrape this data to build a dataset of FIFA team participations in world cup tournament.

FIFA.com



EN



FIFA World Cup™ Teams Statistics

Victories







All-time rankings

Participations

Matches

Cards

Teams with the most tournament participations

TEAM	PARTICIPATIONS	EDITIONS
 BRAZIL	20	1930, 1934, 1938, 1950, 1954, 1958, 1962, 1966, 1970, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
 GERMANY	18	1934, 1938, 1954, 1958, 1962, 1966, 1970, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
 ITALY	18	1934, 1938, 1950, 1954, 1962, 1966, 1970, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
 ARGENTINA	16	1930, 1934, 1958, 1962, 1966, 1974, 1978, 1982, 1986, 1990, 1994, 1998, 2002, 2006, 2010, 2014
 MEXICO	15	1930, 1950, 1954, 1958, 1962, 1966, 1970, 1978, 1986, 1994, 1998, 2002, 2006, 2010, 2014
 ENGLAND	14	1950, 1954, 1958, 1962, 1966, 1970, 1982, 1986, 1990, 1998, 2002, 2006, 2010, 2014

FIFA World Cup — Team Participations

```

In [147]: def parse_team_participations() -> []:
            url = "https://www.fifa.com/fifa-tournaments/statistics-and-record
s/worldcup/teams/index.html"
            page = requests.get(url)
            page_soup = BeautifulSoup(page.text, 'html.parser')
            table_soup = page_soup.findAll("table", {"class":["table","tbl-all
timeranking"]})[2].find("tbody")
            row_soup = table_soup.findAll("tr")

            results = []
            for row in row_soup:
                result = []
                result.append(row.findAll("td")[0].find("span", {"class":"t-nT
ext"}).text)
                result.append(row.findAll("td")[2].find("span", {"class":"text
"}).text)
                result.append(row.findAll("td")[3].find("span", {"class":"text
"}).text)
                results.append(result)

            return results

# un-comment the lines below run data scraping
# colnames = ['team', 'participations', 'years']
# df_fifa_participations = pd.DataFrame.from_records(parse_team_partic
ipations(), columns=colnames)
# df_fifa_participations["participations"] = df_fifa_participations["p
articipations"].astype(int)
# df_fifa_participations.to_pickle("datasets/fifa_participations.pkl")
# store dataframe to local disk

# read back web scraped dataset
df_fifa_participations = pd.read_pickle("datasets/fifa_participations.
pkl")
print("\ndf_fifa_participations = ", df_fifa_participations.shape, "\n
")
df_fifa_participations.head(10)

```

```
df_fifa_participations = (77, 3)
```

Out[147]:

	team	participations	years
0	Brazil	20	1930, 1934, 1938, 1950, 1954, 1958, 1962, 1966...
1	Germany	18	1934, 1938, 1954, 1958, 1962, 1966, 1970, 1974...
2	Italy	18	1934, 1938, 1950, 1954, 1962, 1966, 1970, 1974...
3	Argentina	16	1930, 1934, 1958, 1962, 1966, 1974, 1978, 1982...
4	Mexico	15	1930, 1950, 1954, 1958, 1962, 1966, 1970, 1978...
5	England	14	1950, 1954, 1958, 1962, 1966, 1970, 1982, 1986...
6	France	14	1930, 1934, 1938, 1954, 1958, 1966, 1978, 1982...
7	Spain	14	1934, 1950, 1962, 1966, 1978, 1982, 1986, 1990...
8	Belgium	12	1930, 1934, 1938, 1954, 1970, 1982, 1986, 1990...
9	Uruguay	12	1930, 1950, 1954, 1962, 1966, 1970, 1974, 1986...

```
In [120]: from IPython.core.display import HTML
def css_styling(): styles = open("cs109.css", "r").read(); return HTML(
(styles)
css_styling()
```

Out[120]: