#Datascraping teams

import requests

from bs4 import BeautifulSoup

import pandas as pd

import os

os.chdir('/cs project/teams')

#all the teams have a unique url for each season and these are saved as a txt file

teams=pd.read\_csv('teams2021.txt', header=None)

teams=pd.Series((teams[0]))

length1=len(teams)

#the headers user agent is required to ensure that browser doesn’t think it is a bot

headers = {'User-Agent':

           'Mozilla/5.0 (X11; Linux x86\_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/47.0.2526.106 Safari/537.36'}

df=pd.DataFrame()

#scraping for every team

for x in range(length1):

    page = "https://www.transfermarkt.co.uk" + teams[x]

    pageTree = requests.get(page, headers=headers)

    pageSoup = BeautifulSoup(pageTree.content, 'html.parser')

    Players = pageSoup.find\_all("span", {"class": "show-for-small"})

#the class values for this site are in German – the following is for the right hyperlink

    Values = pageSoup.find\_all("td", {"class": "rechts hauptlink"})

    PlayersList = []

    ValuesList = []

length = len(Players)

#adding players to array’s with their values

for i in range(0,length):

    PlayersList.append(Players[i].text)

    ValuesList.append(Values[i].text)

    df1=pd.DataFrame({"Players":PlayersList,"Values":ValuesList})

    df=df.append(df1)

#removing players not valued in millions as these are outliers and won't be considered in modelling

df = df[df['Values'].str.contains('Mio', na=False)]

df['Values']=((df['Values'].str.split(' ').str[0]).replace(',','.',regex=True).astype(float))\*1000000

df