# import python modules for stats and graphical output

import os

import statistics

import scipy as sp

import math

import pandas as pd

import numpy as np

import matplotlib as mpl

import matplotlib.pyplot as plt

import matplotlib.ticker as tick

import seaborn as sns

import statsmodels.api as sm

os.chdir('c:\cs project\data')

data = pd.read\_csv('consolidated\_data\_2021.csv',sep=';',engine='python')

#the three data tables are created from each of the three seasons csv files

data05 =  pd.read\_csv('consolidated\_data\_2019.csv',sep=';',engine='python')

data1 = pd.read\_csv('consolidated\_data\_2020.csv',sep=';',engine='python')

data1=pd.DataFrame.append(data,data1)

data1=pd.DataFrame.append(data1,data05,ignore\_index=True)

data1.sort\_values('value', ascending=False)

#log values in millions\*\*\*

def millions(x, pos):

    'The two args are the value and tick position'

    return '%1.1fM' % (x \* 1e-6)

def ln(x):

    return np.log(x)

#getting values from 2021 consolidated file

values=data["value"]

#finding the average value

avgvalue=statistics.mean(values)

#converting a column to integer

values=values.astype(np.int64)

#performing logarithms on values

lnvalues=ln(values)

#setting label

formatter = mpl.ticker.FuncFormatter(millions)

#sorting by value descending and keeping the first 10

dataplot=data.sort\_values('value',ascending=False)[0:10]

#sorting by ascending the players selected above

dataplot=dataplot.sort\_values('value',ascending=True)

dataplot['player']=['Trent Alexander-Arnold','Lionel Messi','Jadon Sancho', 'Kevin De Bruyne', 'Mohamed Salah', 'Sadio Mané', 'Harry Kane','Neymar Jr.','Raheem Sterling','Kylian Mbappé']

#setting up the graphs

fig, ax = plt.subplots(2, 2, figsize=(16, 12))

fig.suptitle('Data Analysis', fontsize=20)

#scatter graph for distribution of players vs value

ax[0,0].scatter(data.index,data['value'], color='lightsteelblue')

ax[0,0].yaxis.set\_major\_formatter(formatter)

ax[0,0].set\_xlabel('Number of Players')

ax[0,0].set\_ylabel('Values')

ax[0,0].set\_title('Distribution')

#histogram of player values with the mean value of players value

ax[0,1].hist(values, bins=40, rwidth=2, color='lightsteelblue')

ax[0,1].xaxis.set\_major\_formatter(formatter)

ax[0,1].set\_xlabel('Values')

ax[0,1].set\_ylabel('Count')

ax[0,1].set\_title('Mean Value of Players')

ax[0,1].axvline(avgvalue, color='k', linestyle='dashed', linewidth=1)

min\_ylim, max\_ylim = plt.ylim()

ax[0,1].text(1400, 700, '       Mean: %1.2fM' % (avgvalue \* 1e-6))

#bar chart of top 10 most valuable players

ax[1,0].barh(dataplot['player'],dataplot['value'],color='lightsteelblue')

ax[1,0].axis([0, 220000000,-1,10])

for i, v in enumerate(dataplot['value']):

    ax[1,0].text(v-100, i-0.3, '%1.2fM'% (v \* 1e-6), color='black',fontweight='medium')

ax[1,0].xaxis.set\_major\_formatter(formatter)

ax[1,0].set\_xlabel('Values')

ax[1,0].set\_title('Top 10 most valuable players')

#histogram the number of players per logarithm of value

ax[1,1].hist(lnvalues, bins=40, rwidth=2, color='lightsteelblue')

ax[1,1].set\_xlabel('Logarithms of values')

ax[1,1].set\_ylabel('Count')

ax[1,1].set\_title('Histogram of Values')

plt.show()