﻿

"""

Created on Wed Feb 27 14:43:41 2019

@author: roryeggleston

"""

import numpy as np

import xarray as xr

import matplotlib.pyplot as plt

import matplotlib as mpl

import pandas as pd

file = '/Users/roryeggleston/Downloads/movies.xls'

movies = pd.read\_excel(file)

#EXERCISE 1

movies\_sheet1 = pd.read\_excel(file, sheetname=0, index\_col=0)

movies\_sheet1.head()

/anaconda3/lib/python3.7/site-packages/pandas/io/excel.py:329: FutureWarning: The `sheetname` keyword is deprecated, use `sheet\_name` instead

\*\*kwds)

Out[6]:

Year ... IMDB Score

Title ...

Intolerance: Love's Struggle Throughout the Ages 1916 ... 8.0

Over the Hill to the Poorhouse 1920 ... 4.8

The Big Parade 1925 ... 8.3

Metropolis 1927 ... 8.3

Pandora's Box 1929 ... 8.0

[5 rows x 24 columns]

﻿movies\_sheet2 = pd.read\_excel(file, sheetname=1, index\_col=0)

movies\_sheet2.head()

Out[7]:

Year ... IMDB Score

Title ...

102 Dalmatians 2000 ... 4.8

28 Days 2000 ... 6.0

3 Strikes 2000 ... 4.0

Aberdeen 2000 ... 7.3

All the Pretty Horses 2000 ... 5.8

[5 rows x 24 columns]

﻿movies\_sheet3 = pd.read\_excel(file, sheetname=2, index\_col=0)

movies\_sheet3.head()

Out[8]:

Year ... IMDB Score

Title ...

127 Hours 2010.0 ... 7.6

3 Backyards 2010.0 ... 5.2

3 2010.0 ... 6.8

8: The Mormon Proposition 2010.0 ... 7.1

A Turtle's Tale: Sammy's Adventures 2010.0 ... 6.1

[5 rows x 24 columns]

﻿movies = pd.concat([movies\_sheet1, movies\_sheet2, movies\_sheet3])

﻿movies.shape

Out[10]: (5042, 24)

﻿movies["Year"].plot(kind="hist", bins = 10, color = "purple")

plt.xlabel('Year')

plt.ylabel('Number of films')

Out[11]: Text(0, 0.5, 'Number of films')



﻿

movies["Year"].plot(kind="hist", bins = 5, color = "black")

plt.xlabel('Year')

plt.ylabel('Number of films')

Out[12]: Text(0, 0.5, 'Number of films')



﻿#It appears that the number of films being produced over the past century has increased exponentially.

#EXERCISE 2

﻿plt.scatter(movies["Year"], movies["Budget"], color = "purple")

plt.xlim(1914,2017)

plt.ylim(0,450000000)

plt.xlabel("Year")

plt.ylabel("Budget")

plt.suptitle("Film Year by Budget")

Out[13]: Text(0.5, 0.98, 'Film Year by Budget')



﻿ #The overall budgets for films also appears to increase exponentially with the years.

#EXERCISE 3

﻿plt.scatter(movies["Gross Earnings"], movies["Budget"], color = "purple")

plt.xlim(0, 450000000)

plt.ylim(0,450000000)

plt.xlabel("Gross Earnings")

plt.ylabel("Budget")

plt.suptitle("Gross Earnings by Budget")

Out[14]: Text(0.5, 0.98, 'Gross Earnings by Budget')



#Although it is not possible to calculcate the correlation because of the large number of NAN values, it appears as though these variables are at least weakly positively correlated

#EXERCISE 4

﻿sorted\_by\_IMDB = movies.sort\_values(['IMDB Score'], ascending=True)

sorted\_by\_IMDB["IMDB Score"].head(10)

Out[15]:

Title

Justin Bieber: Never Say Never 1.6

Foodfight! 1.7

The Helix... Loaded 1.9

Superbabies: Baby Geniuses 2 1.9

Disaster Movie 1.9

Who's Your Caddy? 2.0

United Passions 2.0

Glitter 2.1

From Justin to Kelly 2.1

Crossover 2.1

Name: IMDB Score, dtype: float64

#"Justin Bieber: Never Say Never" is the film with the lowest IMDB score, with a score of 1.6.

#EXERCISE 5

﻿﻿BadMovies = movies[movies["IMDB Score"] <= 5]

BadMovies["IMDB Score"].head(10)

Out[16]:

Title

Over the Hill to the Poorhouse 4.8

The Brain That Wouldn't Die 4.1

The Stewardesses 4.0

Exorcist II: The Heretic 3.7

Alien Zone 4.1

Meteor 4.9

Can't Stop the Music 4.5

Galaxina 3.4

Raise the Titanic 4.7

Graduation Day 4.5

Name: IMDB Score, dtype: float64

#Output Excel file uploaded to GitHub