Đề thi:

PYTHON FOR MACHINE LEARNING, DATA SCIENCE AND VISUALIZATION

Thời gian: 120 phút

Ngày thi : __/__/____

**** Học viên tạo 1 thư mục là **LDS2_HoVaTen**, lưu tất cả bài làm vào để nộp chấm điểm ****

**** Học viên được sử dụng tài liệu ****

Chú ý, với mỗi câu:

- Cần hiển thị thông tin chung của dữ liệu bằng cách dùng shape, dtypes, columns, head(), tail(), info()… để có cái nhìn ban đầu về dữ liệu.
- Mỗi câu là 1 file viết trên Jupyter Notebook, các yêu cầu nhận xét kết quả trong từng câu được viết trong cell dưới định dạng Markdown.

1. Udemy Courses: (7.5 điểm)

Cho dữ liệu **courses.csv, subjects.csv**, thực hiện các yêu cầu sau :

a) Đọc dữ liệu, hiển thị thông tin chung của dữ liệu : shape, columns, head, tail, info (0.5 điểm)

b) Xử lý dữ liệu (1.25 điểm):

- Kết dữ liệu bảng courses và subjects, sau đó:

- Xóa các cột không cần thiết : course_id,url,id,id_sub,total

- Tạo thêm cột 'revenue'= num_subscribers * price

- Hiển thị thông tin các cột như sau :

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3678 entries, 0 to 3677
Data columns (total 12 columns):
course_title           3678 non-null object
is_paid                3678 non-null bool
price                  3678 non-null int64
num_subscribers        3678 non-null int64
num_reviews            3678 non-null int64
num_lectures           3678 non-null int64
level                  3678 non-null object
content_duration       3678 non-null float64
published_timestamp    3678 non-null object
subject                3678 non-null object
dean                   3678 non-null object
revenue                3678 non-null int64
dtypes: bool(1), float64(1), int64(5), object(5)
memory usage: 348.4+ KB
```

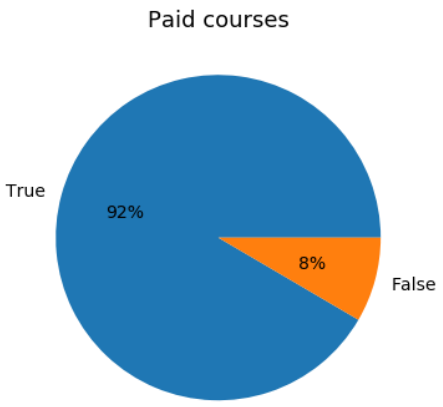
c) Lọc ra các khóa học thuộc nhóm **Musical Instruments** mà tiêu đề (course_title) có chứa "**Guitarist**" như hình sau: (0.5 điểm)

	course_title	is_paid	price	num_subscribers	num_reviews	num_lectures	level	content_duration	published_timestamp	subject	dean
1798	7 Ways A Beginner Guitarist Can Sound Better, ...	False	0	4529	193	7	Beginner Level	0.6	2015-12-21T18:50:50Z	Musical Instruments	Joseph Miller
1858	Guitar Fretboard Mastery for the Blues Guitarist	True	20	257	6	5	All Levels	1.0	2014-01-16T11:30:07Z	Musical Instruments	Joseph Miller
2010	The Circle of Fifths for Guitarists	True	50	1018	2	5	Intermediate Level	1.5	2016-12-15T21:57:42Z	Musical Instruments	Joseph Miller
2113	The Art of the Blues Shuffle - A Guitarist's G...	True	20	8	0	11	Intermediate Level	1.5	2016-06-26T16:57:13Z	Musical Instruments	Joseph Miller

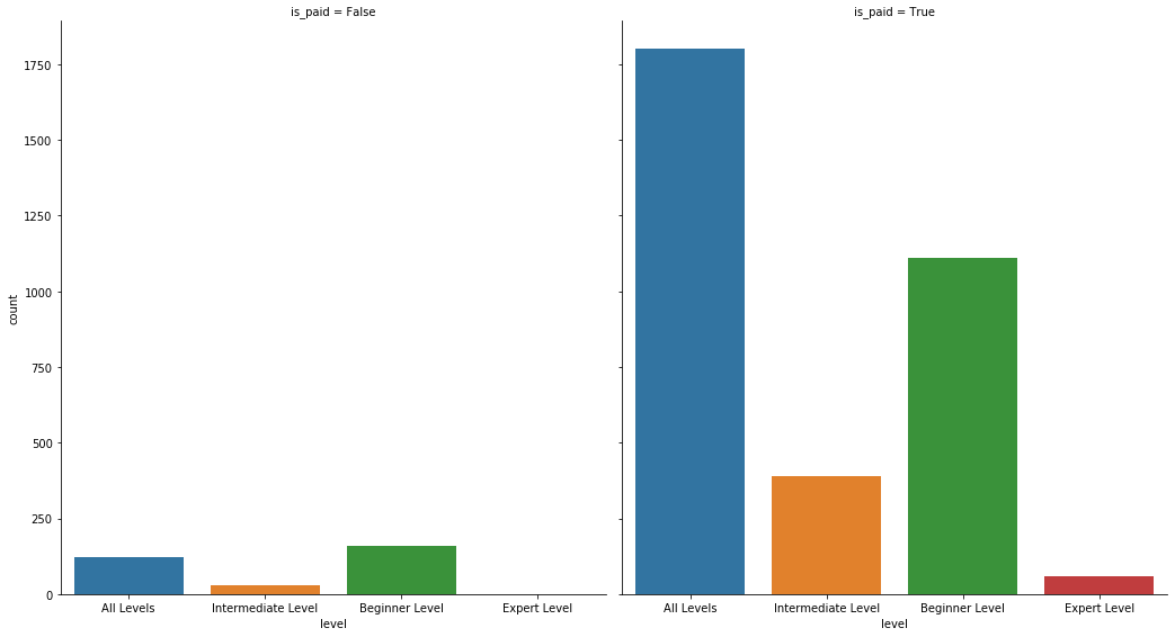
d) Thống kê giá trung bình và doanh thu trung bình của các khóa học theo subject và level, kết quả gợi ý như hình sau, bạn có nhận xét gì với kết quả có được (0.75 điểm)

subject	level	Avg Price	Avg Revenue
Business Finance	All Levels	75.20	120811.44
	Beginner Level	53.38	80147.49
	Expert Level	95.97	129994.68
	Intermediate Level	66.05	65395.08
Graphic Design	All Levels	62.70	159184.80
	Beginner Level	53.48	109278.72
	Expert Level	70.00	5067.00
	Intermediate Level	49.82	52035.70
Musical Instruments	All Levels	55.98	139663.73
	Beginner Level	43.38	35228.34
	Expert Level	36.43	37739.29
	Intermediate Level	51.04	40793.07
Web Development	All Levels	82.53	667633.91
	Beginner Level	68.36	397559.78
	Expert Level	113.67	76648.33
	Intermediate Level	71.26	253906.89

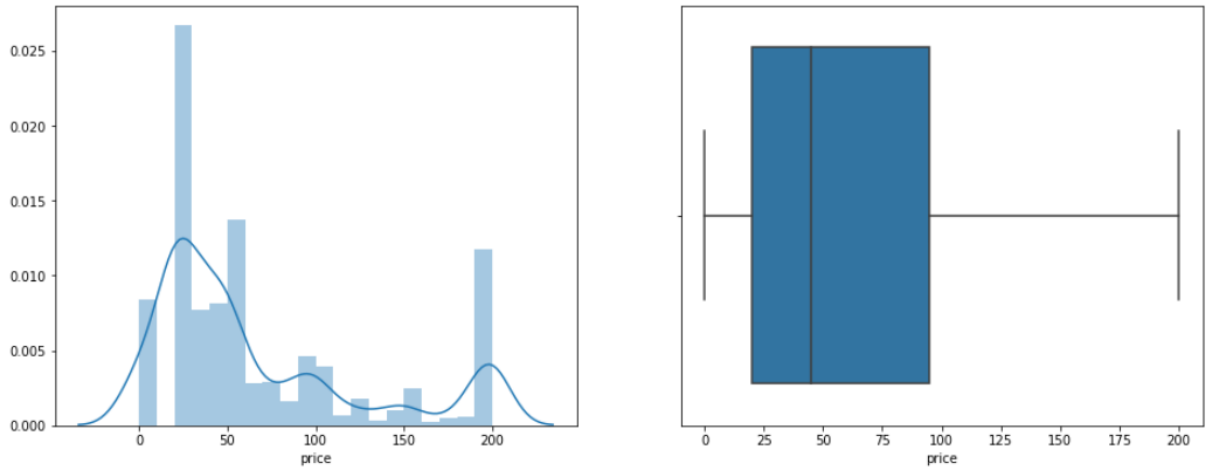
e) Vẽ biểu đồ pie chart cho biết tỷ lệ các khóa học Free và Paid trên Udemy, kết quả gợi ý như hình sau (0.5 điểm):



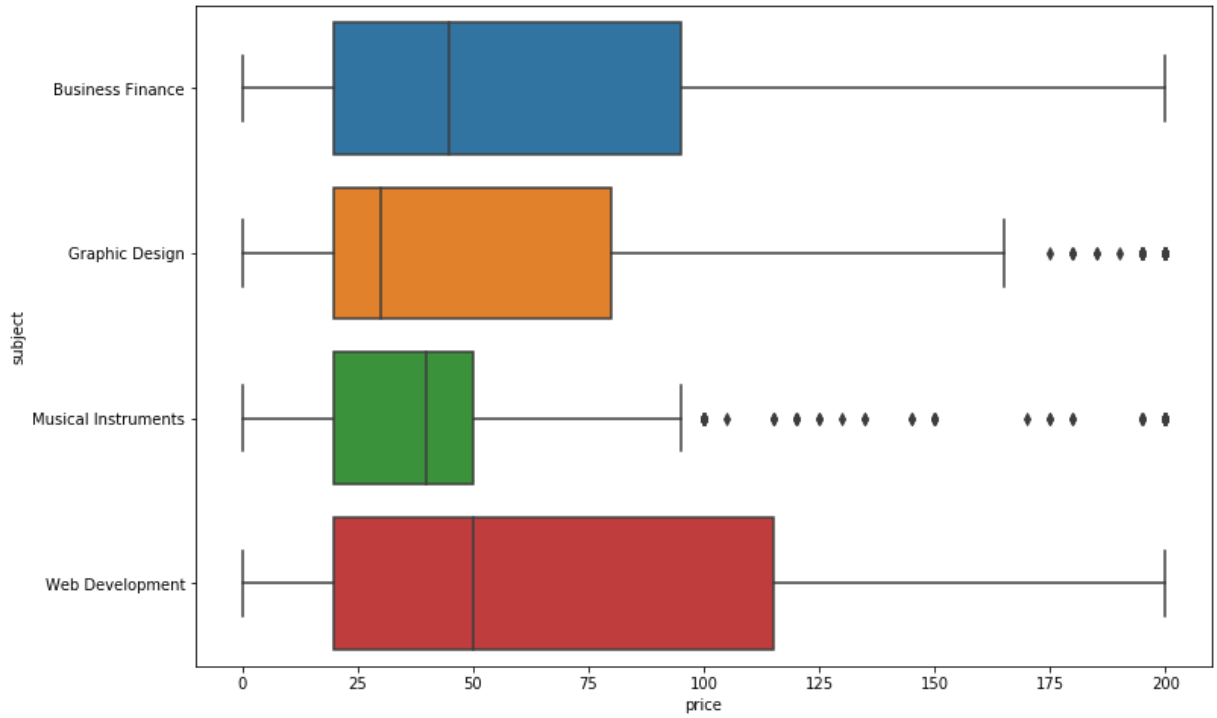
f) Vẽ biểu đồ thống kê số lượng khóa học theo từng level, có phân loại (miễn phí và trả phí) như hình sau, rồi cho nhận xét (1 điểm)



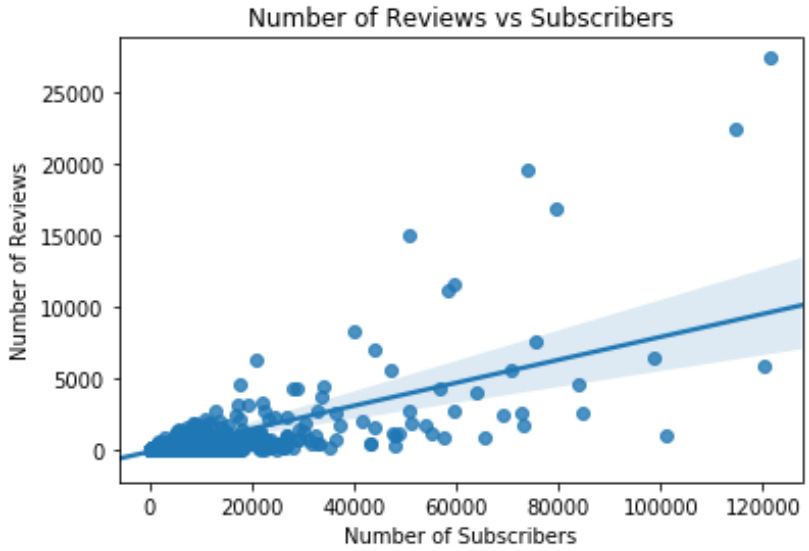
g) Vẽ biểu đồ phân phối tần suất các khóa học theo học phí (price) và biểu đồ box plot của price, kết quả gợi ý như hình sau và cho nhận xét : (1 điểm)



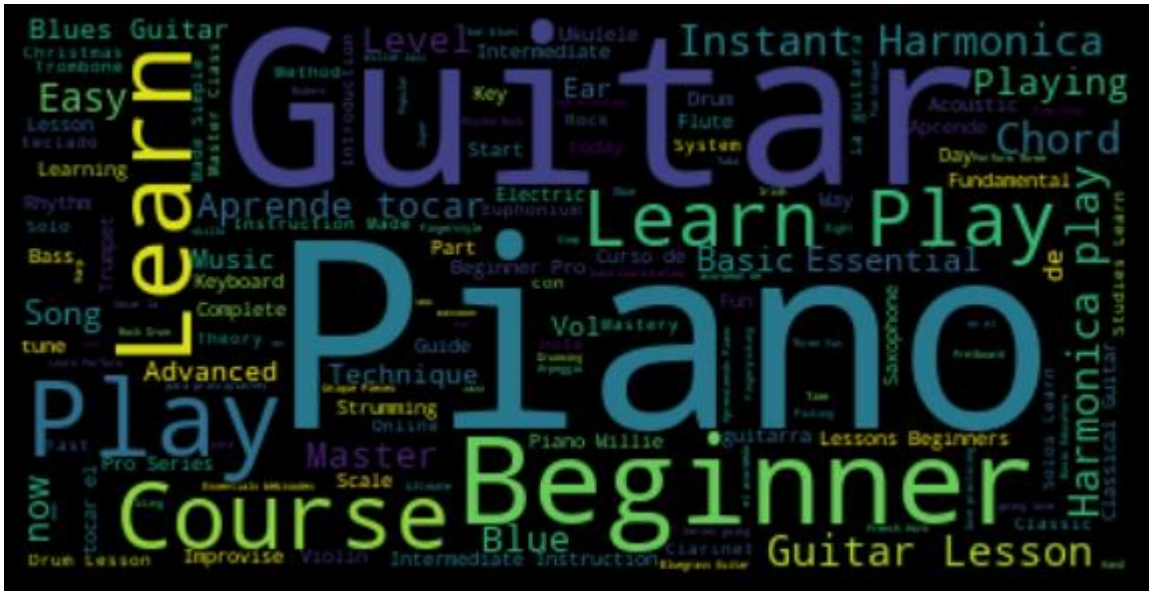
h) Vẽ biểu đồ thể hiện boxplot giá khóa học giữa các thể loại và cho nhận xét (0.75 điểm)



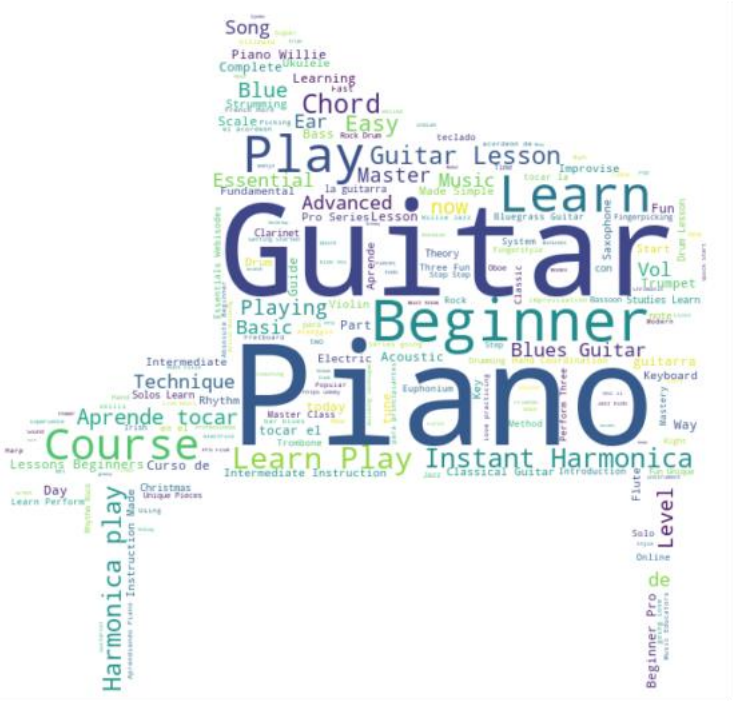
i) Vẽ biểu đồ thể hiện tương quan giữa **num_reviews** và **num_subscribers** (*chỉ chọn những khóa có num_reviews < 15000*) như hình sau rồi cho nhận xét (0.5 điểm)



j) Tạo biểu đồ Wordcloud dựa trên **course_title** với subject là **Musical Instruments**, có kết quả gợi ý như sau : (0.5 điểm)



k) Cho tập tin hình ảnh **piano.jpg**, hãy tạo biểu đồ có kết quả gợi ý như hình sau : (0.25 điểm)

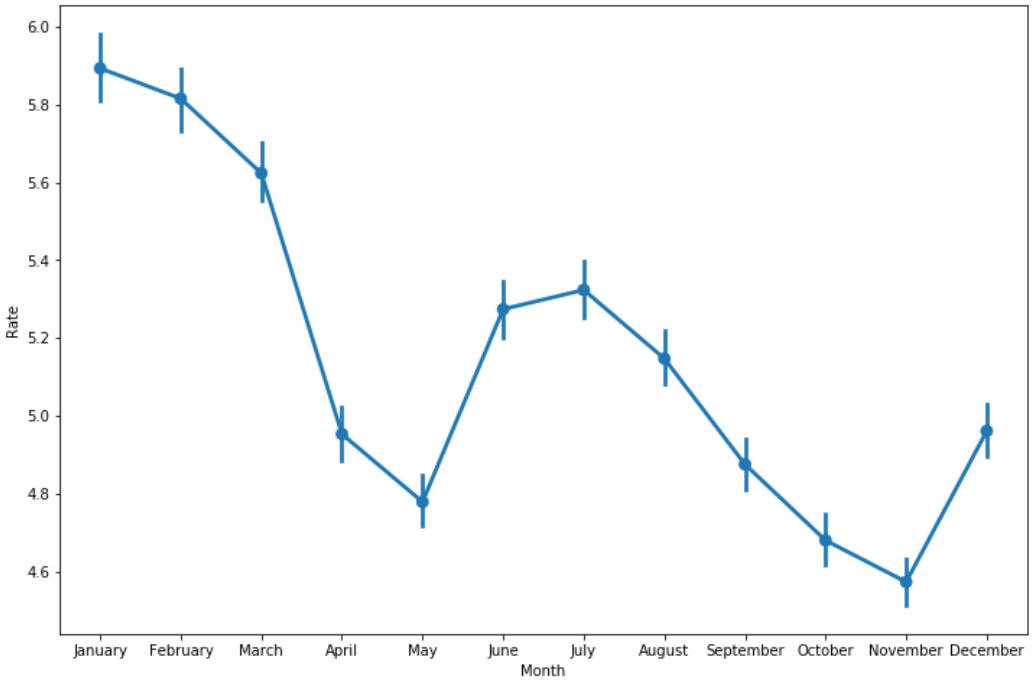


2. Trực quan hóa dữ liệu bản đồ (2.5 điểm)

Cho dữ liệu **us_unemployment_2016.csv** và **us-states.json**, thực hiện các yêu cầu sau :

a) Đọc dữ liệu, hiển thị thông tin chung của dữ liệu bao gồm : head, tail, info, describe (0.25 điểm)

b) Hiển thị tỷ lệ thất nghiệp trung bình các tháng trong năm 2016 rồi nhận xét, kết quả gợi ý như hình sau (0.75 điểm):

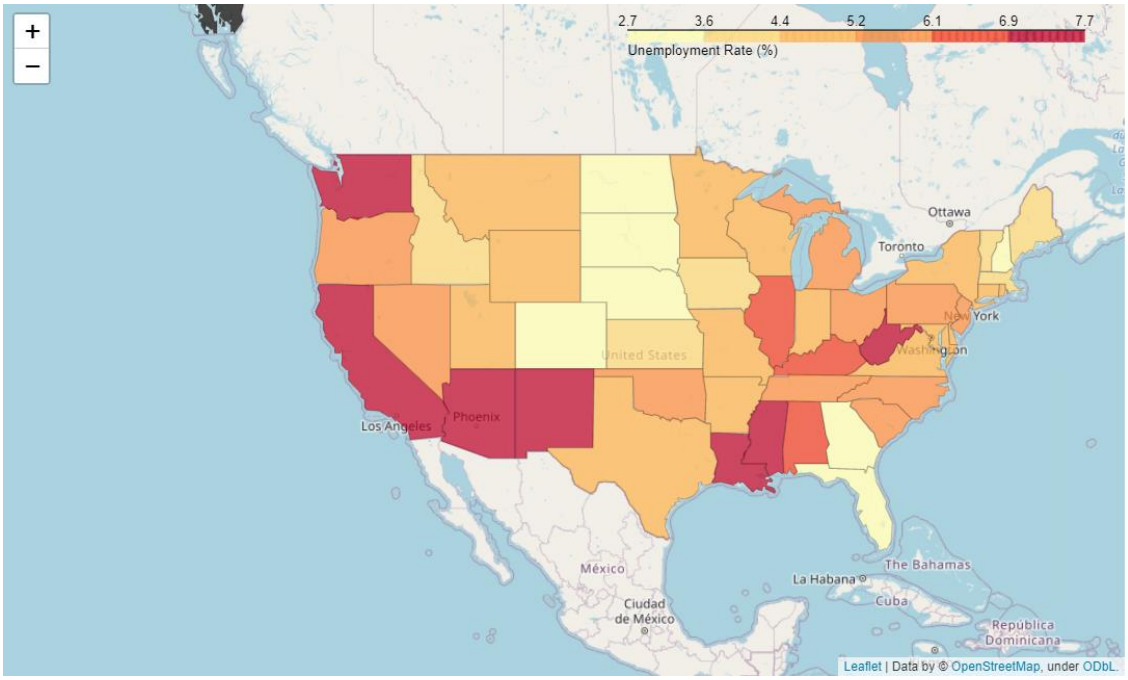


c) Tạo bản đồ với center là Mỹ và zoom level (zoom_start=5) gợi ý như hình sau : (0.25 điểm)

d) Nhóm dữ liệu State và tính trung bình (mean) theo Unemployment_Rate:

	State	Rate
0	AL	6.773756
1	AR	4.656667
2	AZ	7.732778
3	CA	6.920219
4	CO	3.416146

Sau đó tạo choropleth map theo Unemployment_Rate của từng tiểu bang theo gợi ý như hình sau : (1.25 điểm)



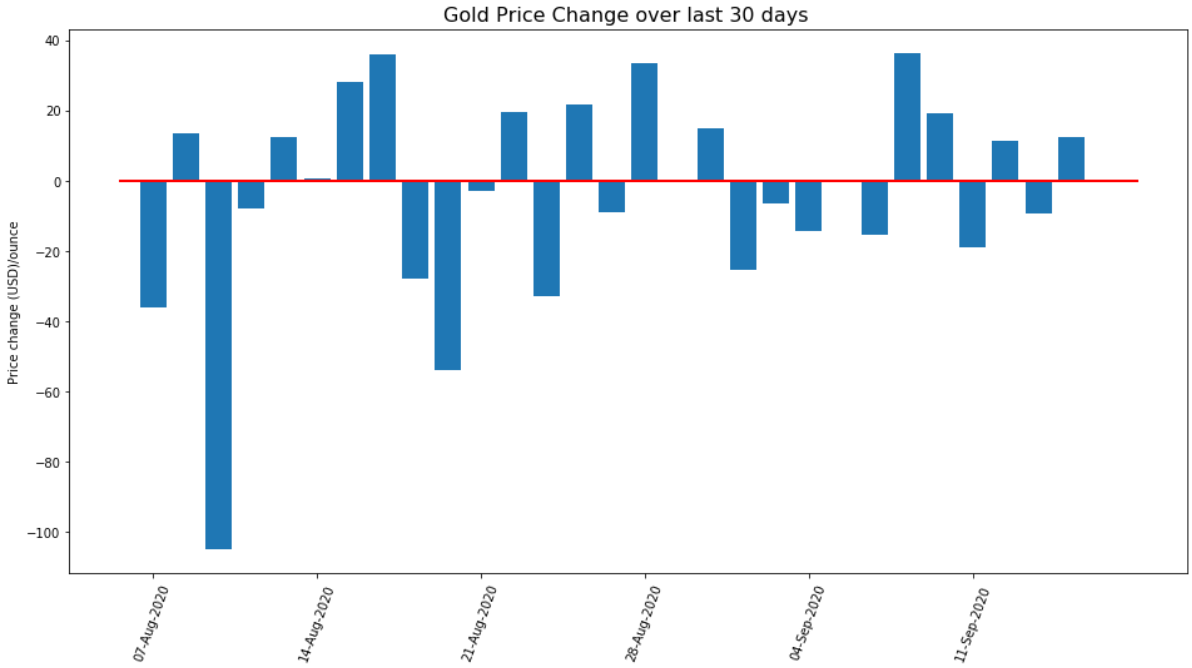
3. Xử lý dữ liệu và hiển thị biểu đồ giá vàng thế giới

SPDR là quỹ tín thác (ETF) vàng lớn nhất thế giới do State Street Global Advisors – một bộ phận của State Street Corporation, tập đoàn quản lý tài sản lớn thứ hai thế giới, điều hành. Dữ liệu giá vàng và lượng giao dịch (tấn) của quỹ SPDR đều được Quỹ cập nhật trên website (https://www.spdrgoldshares.com/usa/historical-data/) . Hãy thực hiện các yêu cầu sau :

- Download và đọc dữ liệu
- Thực hiện các xử lý dữ liệu cần thiết. Lưu ý các thời điểm nghỉ (" HOLIDAY"," NYSE Closed"," AWAITED") sẽ lấy giá đóng cửa của ngày trước thời điểm nghỉ.
- Vẽ biểu đồ giá đóng cửa của giá vàng qua thời gian (hình 1) ***hoặc*** Vẽ biểu đồ cho biết thay đổi giá của 30 ngày gần nhất – gợi ý: sử dụng hàm diff để tính chênh lệch giá mỗi ngày, trước khi vẽ (hình 2)

Hình 1:

Hình 2:



--- Chúc các bạn làm bài tốt ☺ ---