Họ và tên: Lâm Mỹ Ngọc

MSSV: K214110843

BÀI LÀM

Câu 1:

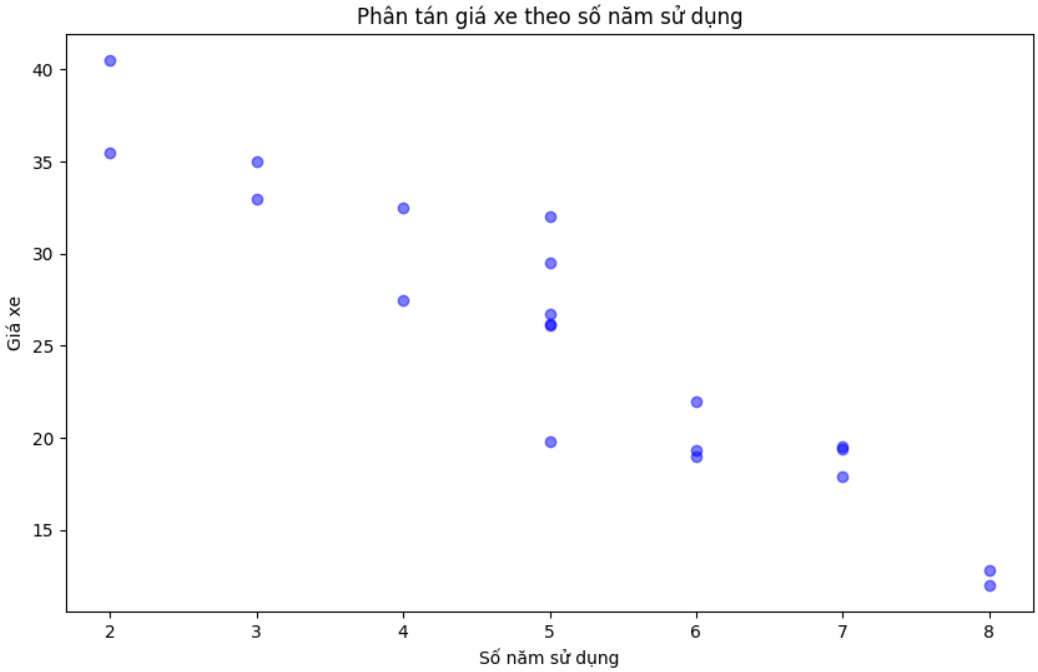
Theo em, liên quan đến ngành khoa học dữ liệu có các ngành nghề phổ biến sau:

- Khoa học dữ liệu: Phân tích dữ liệu để tìm kiếm thông tin, dự đoán xu hướng và đưa ra giải pháp cho các vấn đề kinh doanh.
- Chuyên gia phân tích dữ liệu: Thu thập, xử lý và phân tích dữ liệu để hỗ trợ việc ra quyết định trong doanh nghiệp.
- Khai phá dữ liệu: Áp dụng các kỹ thuật học máy để tìm kiếm các mẫu và xu hướng ẩn trong dữ liệu.
- Kỹ sư dữ liệu: Xây dựng và duy trì hệ thống cơ sở dữ liệu, công cụ và quy trình để thu thập, lưu trữ và xử lý dữ liệu.
- Kỹ sư máy học: tập trung vào việc xây dựng và phát triển các mô hình học máy và các công cụ phân tích dữ liệu.

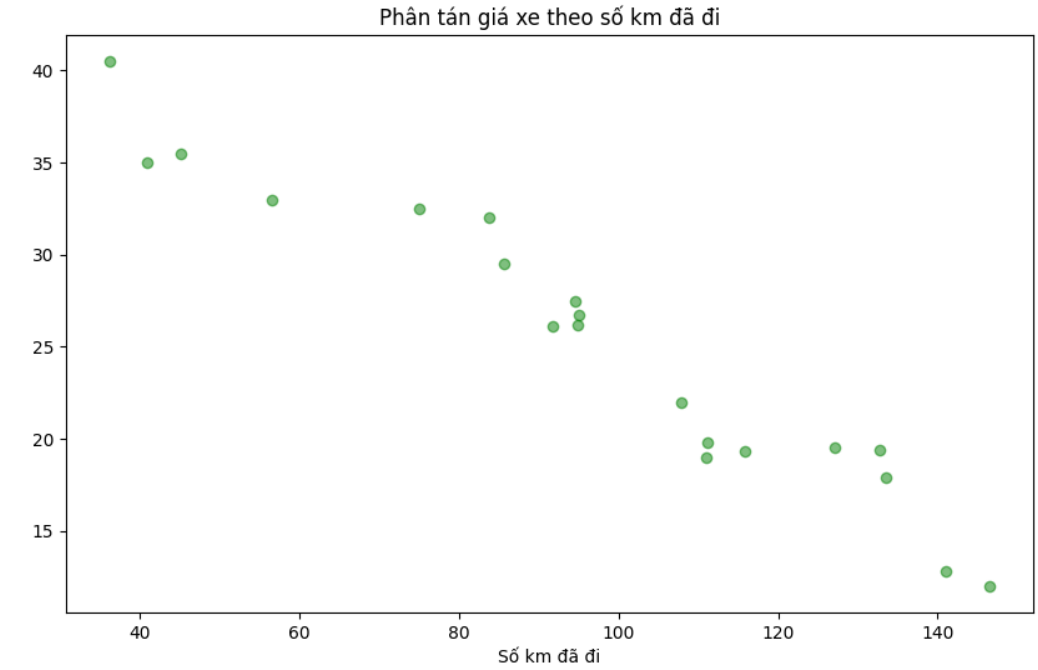
Để đáp ứng yêu cầu công việc cần chuẩn bị các kiến thức về thống kê , lập trình, kiến thức chuyên môn của ngành (chẳng hạn như phân tích dữ liệu tài chính thì cần có kiến thức về khách hàng trong lĩnh vực tài chính, quy trình hoạt động,…) và kiến thức về phân tích dữ liệu (quy trình phân tích, các mô hình phân tích và công cụ phân tích như SPSS, mô hình học máy, học sâu,…)

Câu 2:

a) Biểu đồ phân tán

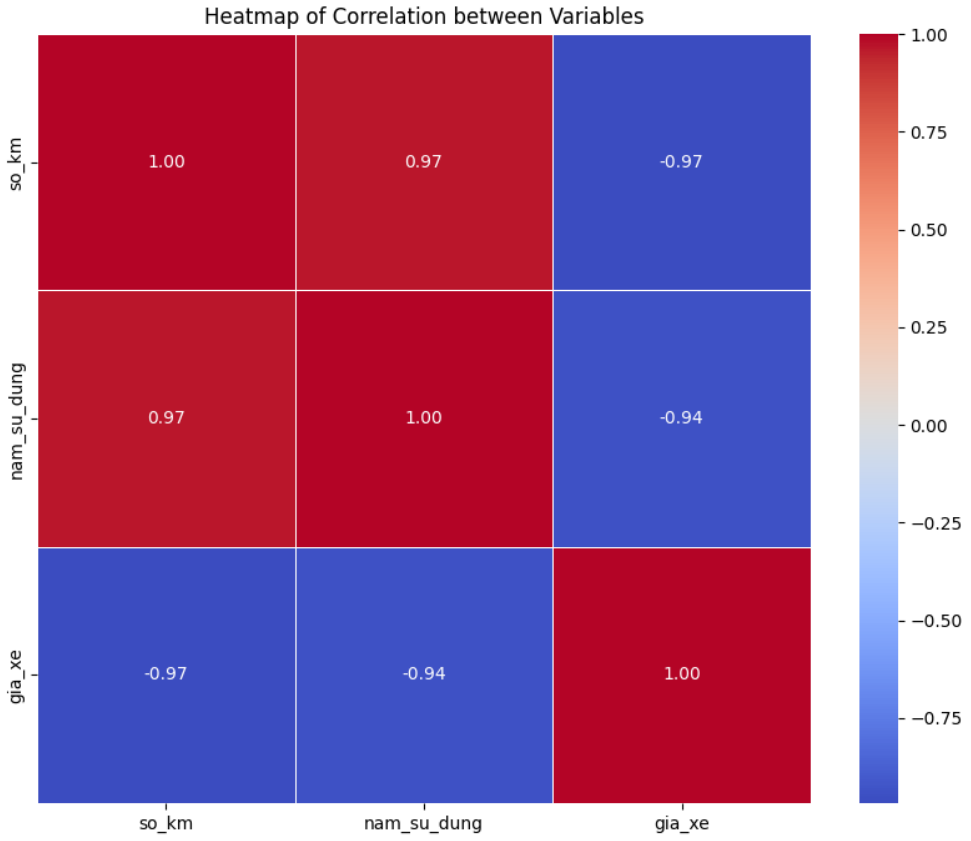


Biểu đồ thể hiện mối tương quan âm giữa giá xe và số năm sử dụng, tức là số năm sử dụng càng nhiều thì giá xe càng giảm và ngược lại.



Biểu đồ thể hiện mối tương quan âm giữa giá xe và số km đã đi, tức là số km đã đi càng tăng thì giá xe càng giảm và ngược lại.

b) Biểu đồ Heatmap



	Feature	VIF
0	so_km	0.966626
1	nam_su_dung	0.939286
2	gia_xe	0.935743

Kiểm tra đa cộng tuyến cho thấy chỉ số VIF của các biến < 1 nên có thể kết luận không có sự đa cộng tuyến giữa các biến.

c) Ước lượng hệ số mô hình hồi quy

$$\widehat{Giá\ xe} = a + b*(số\ năm\ sử\ dụng) + c*(số\ km\ đã\ đi)$$

a = 47.5730

b = - 0.1617

c = - 0.2225

⇨ **Giá xe = 47.5730 – 0.1617*(số năm sử dụng) – 0.225*(số km đã đi)**

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 gia_xe   R-squared:                       0.936
Model:                            OLS   Adj. R-squared:                  0.928
Method:                 Least Squares   F-statistic:                     123.8
Date:                Tue, 09 Apr 2024   Prob (F-statistic):           7.37e-11
Time:                        15:17:07   Log-Likelihood:                -41.743
No. Observations:                  20   AIC:                             89.49
Df Residuals:                      17   BIC:                             92.47
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
nam_su_dung    -0.1617      1.103     -0.147      0.885      -2.488       2.165
so_km          -0.2225      0.060     -3.737      0.002      -0.348      -0.097
intercept      47.5730      1.496     31.804      0.000      44.417      50.729
==============================================================================
Omnibus:                        2.374   Durbin-Watson:                   1.296
Prob(Omnibus):                  0.305   Jarque-Bera (JB):                1.126
Skew:                           0.091   Prob(JB):                        0.569
Kurtosis:                       1.852   Cond. No.                         323.
==============================================================================
```

d) Ý nghĩa hệ số

Hệ số hồi quy của 'số năm sử dụng' (-0.1617): Đây là độ dốc của đường hồi quy đối với biến 'số năm sử dụng'. Giá trị âm của hệ số này (-0.1617) cho biết rằng khi số năm sử dụng tăng lên 1 đơn vị, giá xe sẽ giảm đi 0.1617 đơn vị, trong điều kiện các yếu tố khác không thay đổi. Điều này ngụ ý rằng càng lâu xe được sử dụng, giá trị của nó sẽ giảm do ảnh hưởng của việc sử dụng và mòn trên thời gian.

Hệ số hồi quy của 'số km đã đi' (-0.225): Đây cũng là độ dốc của đường hồi quy, nhưng đối với biến 'số km đã đi'. Giá trị âm của hệ số này (-0.225) cho biết rằng khi số km đã đi tăng lên 1 đơn vị, giá xe sẽ

giảm đi 0.225 đơn vị, trong điều kiện các yếu tố khác không thay đổi. Điều này ngụ ý rằng càng nhiều km đã đi, xe có xu hướng mất giá nhanh hơn do ảnh hưởng của việc sử dụng và mòn trên quãng đường.

Hệ số chặn (intercept) (47.5730): Đây là giá trị của biến phụ thuộc (giá xe) khi tất cả các biến độc lập (số năm sử dụng và số km đã đi) đều bằng 0. Trong trường hợp này, ý nghĩa của hệ số chặn là giá xe của một xe mới (số năm sử dụng và số km đã đi đều là 0).

e) Kiểm định độ phù hợp của mô hình

R-squared = 0.936, cho thấy rằng 93,6% sự thay đổi của biến phụ thuộc được giải thích bởi các biến độc lập trong mô hình. Đây là một giá trị R-squared khá cao, cho thấy mô hình có khả năng giải thích tốt sự biến đổi của biến phụ thuộc.

Adj. R-squared = 0.928, gần bằng R-squared. Điều này cho thấy việc thêm các biến độc lập vào mô hình không ảnh hưởng đáng kể đến khả năng giải thích của mô hình.

F-statistic = 123,8 và **p-value = 7,37e-11**. p-value rất nhỏ, nhỏ hơn mức ý nghĩa 0,05, cho thấy mô hình có ý nghĩa thống kê.

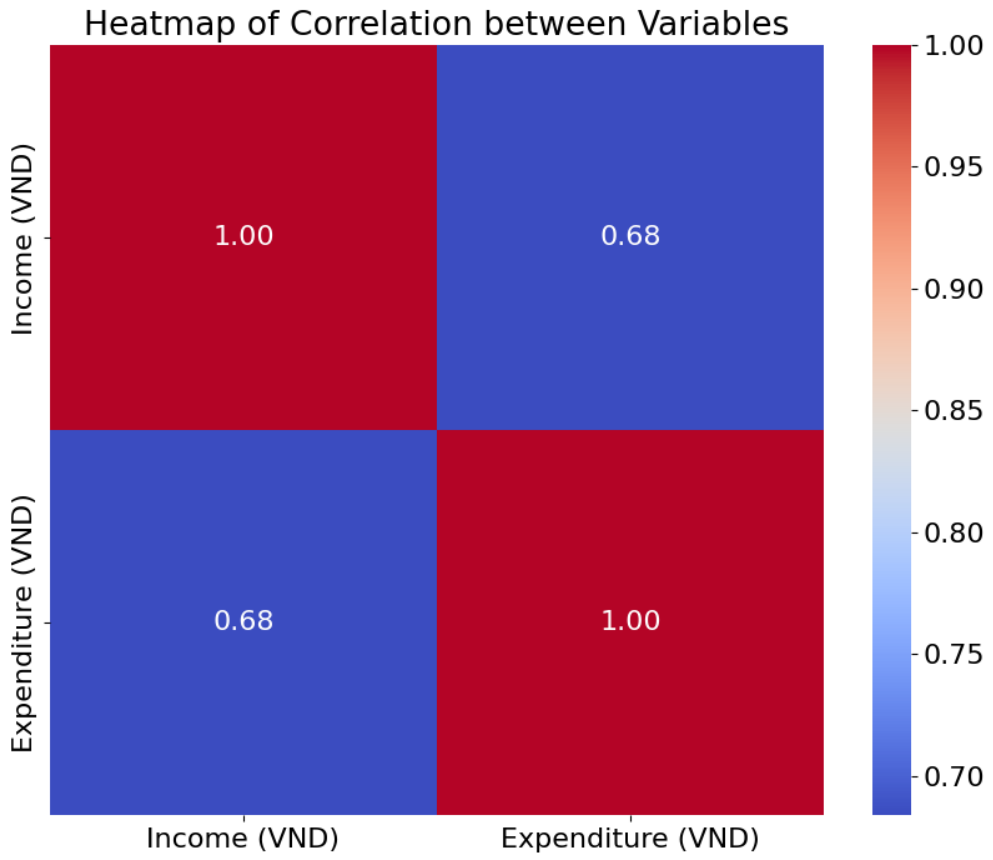
f) Lưu mô hình và thực hiện dự đoán

	nam_su_dung	so_km	predicted_price
0	6	112	21.68
1	3	165	10.37

- Xe sử dụng 6 năm, đi được 112000Km, có giá dự đoán là 21.68 (đơn vị tiền tệ)
- Xe sử dụng 3 năm, đi được 165000Km, có giá dự đoán là 10.37 (đơn vị tiền tệ)

Câu 3:

a) Heatmap


Heatmap of Correlation between Variables
Income (VND)
Expenditure (VND)
1.00
0.68
0.68
1.00
Income (VND)
Expenditure (VND)
1.00
0.95
0.90
0.85
0.80
0.75
0.70

b) Fit mô hình

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      Expenditure (VND)   R-squared (uncentered):                   0.931
Model:                            OLS   Adj. R-squared (uncentered):              0.931
Method:                 Least Squares   F-statistic:                          6.782e+04
Date:                Tue, 09 Apr 2024   Prob (F-statistic):                        0.00
Time:                        16:31:13   Log-Likelihood:                         -83621.
No. Observations:                5000   AIC:                                  1.672e+05
Df Residuals:                    4999   BIC:                                  1.672e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Income (VND)     0.6207      0.002    260.415      0.000       0.616       0.625
==============================================================================
Omnibus:                      144.964   Durbin-Watson:                   1.967
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               85.472
Skew:                          -0.167   Prob(JB):                     2.75e-19
Kurtosis:                       2.454   Cond. No.                         1.00
==============================================================================
```

⇨ **Inxpenditure = 0.627*Income**

Hệ số hồi quy (0.627): Đây là độ dốc của đường hồi quy đối với biến 'Income'. Giá trị dương của hệ số này (**0.627**) cho biết rằng khi số năm sử dụng tăng lên 1 đơn vị, giá xe sẽ tăng lên 0.627 đơn vị. Điều này ngụ ý rằng thu nhập càng cao thì chi tiêu càng nhiều.

Kiểm định độ phù hợp của mô hình

R-squared = 0.931, cho thấy rằng 93,1% sự thay đổi của biến phụ thuộc được giải thích bởi các biến độc lập trong mô hình. Đây là một giá trị R-squared khá cao, cho thấy mô hình có khả năng giải thích tốt sự biến đổi của biến phụ thuộc.

Adj. R-squared = 0.928, gần bằng R-squared. Điều này cho thấy việc thêm các biến độc lập vào mô hình không ảnh hưởng đáng kể đến khả năng giải thích của mô hình.

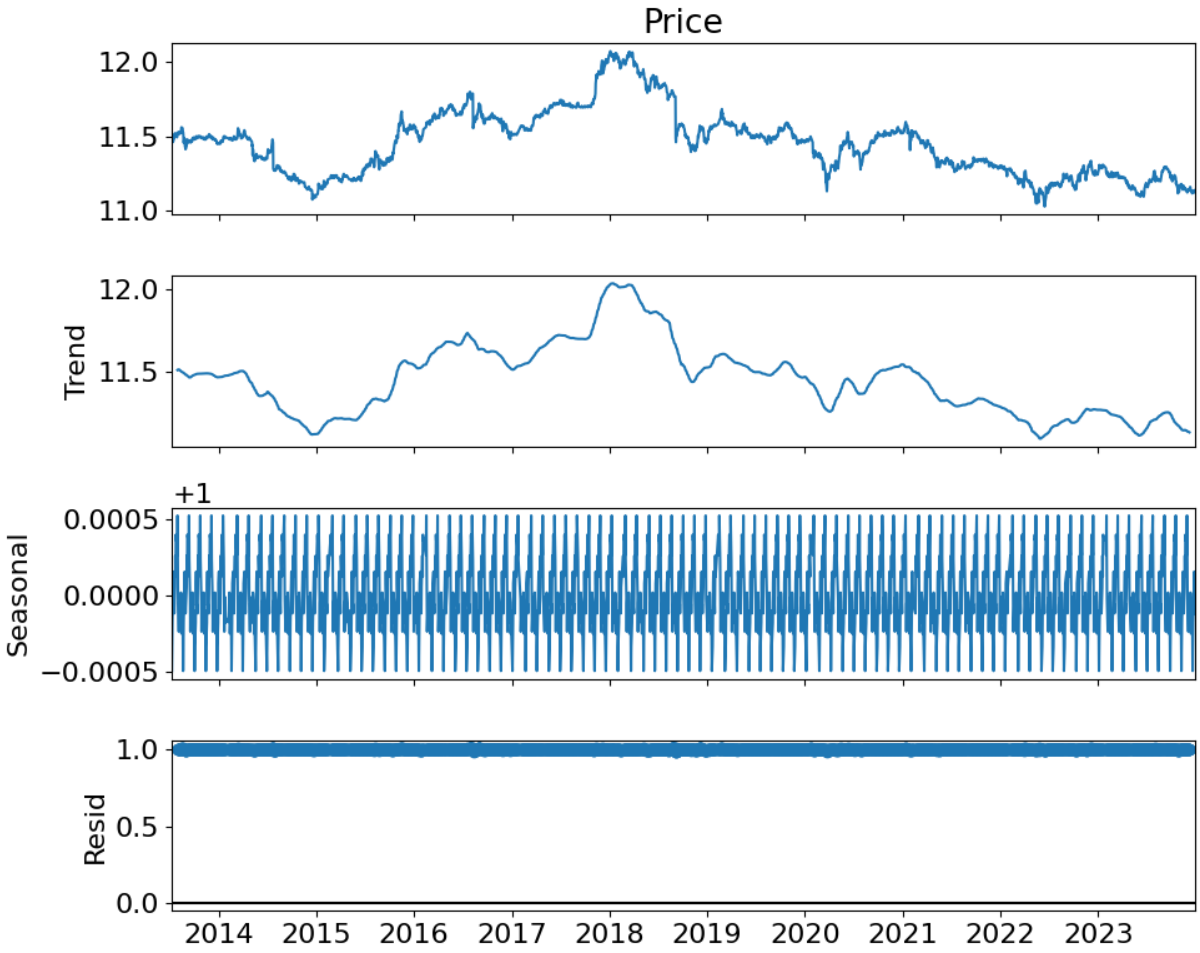
F-statistic = 6.792e+04 và **p-value = 0,** p-value rất nhỏ, nhỏ hơn mức ý nghĩa 0,05, cho thấy mô hình có ý nghĩa thống kê.

c) Lưu mô hình và dự đoán

Câu 4:

Dữ liệu được lấy trực tiếp từ website **Investing.com**

a) Phân rã dữ liệu

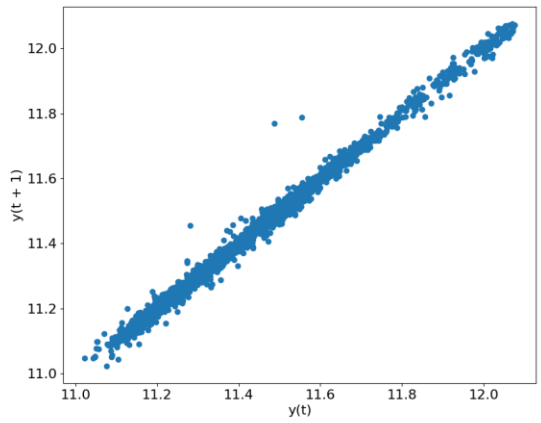


Dữ liệu có có sự biến động đáng kể theo thời gian, nhìn chung có xu hướng giảm nhẹ. Dữ liệu có tính mùa vụ với một chu kỳ lặp lại nhất định

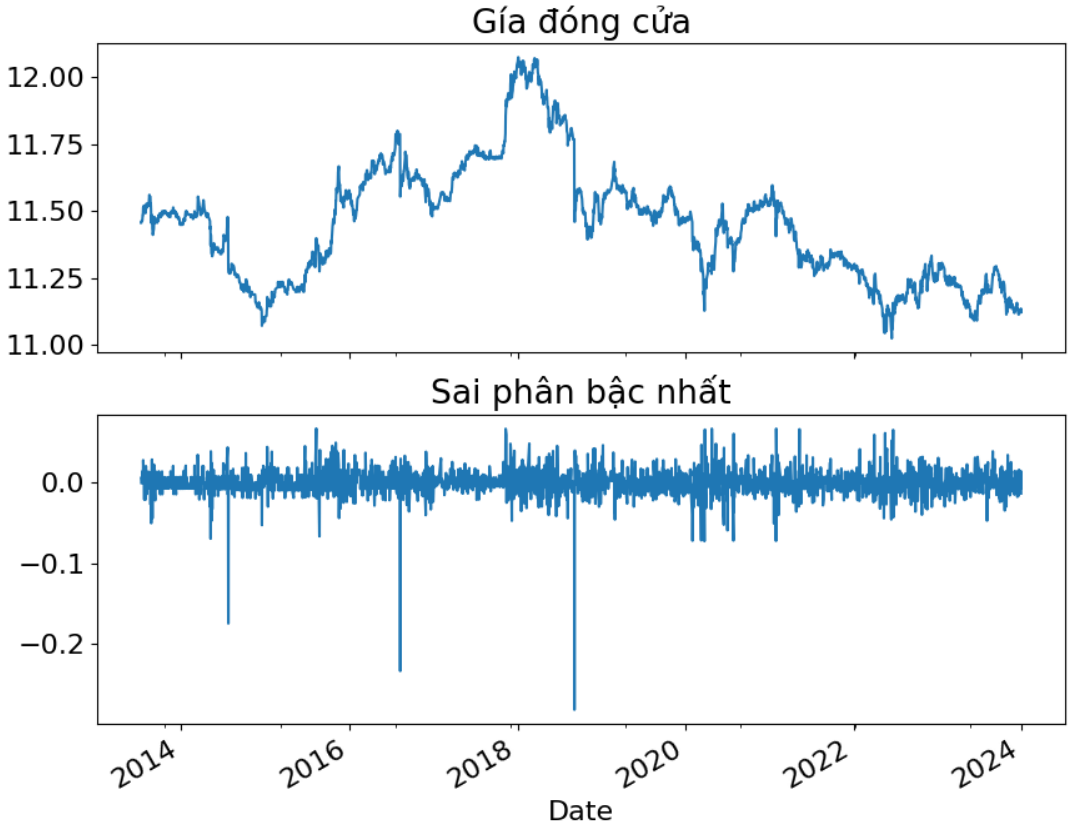
b) Kiểm định tính dừng, tương quan

```
Fail to reject the null hypothesis (H0),
the data non-stationary
ADF: Test statistic       -2.107000
p-value                    0.241709
# of Lags                  5.000000
# of Observations       2612.000000
Critical Value (1%)       -3.432856
Critical Value (5%)       -2.862647
Critical Value (10%)      -2.567359
dtype: float64
-------------------------
Rejected the null hypothesis (H0),
the data is stationary
KPSS: Test statistic      2.243913
p-value                   0.010000
# of Lags                30.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Đối với kiểm định ADF, p-value = 0.24 > 0.05 nên không thể bác bỏ Ho (chuỗi không dừng). Đối với kiểm đinh KPSS, p-value = 0.01 < 0.05 nên có thể bác bỏ Ho (chuỗi dừng). Từ 2 kiểm định trên có thể kết luận rằng chuỗi dữ liệu chưa dừng.



Biểu đồ tương quan cho thấy giá đóng cửa có sự tương quan dương với giá đóng cửa tại thời điểm trước đó. Tuy nhiên vẫn còn một số giá trị ngoại lai xuất hiện trong bộ dữ liệu có thể là do một số sự kiện bất thường xảy ra trong thời gian được phân tích.

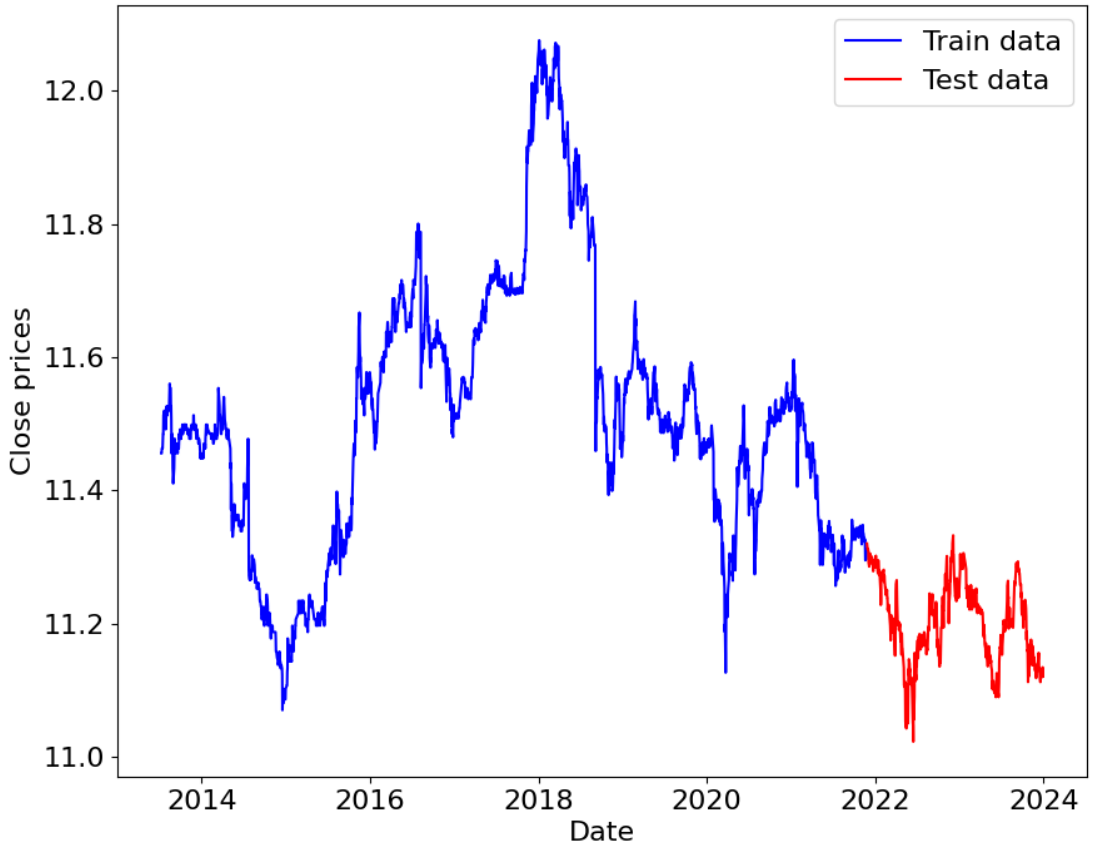

Gía đóng cửa
12.00
11.75
11.50
11.25
11.00
Sai phân bậc nhất
0.0
−0.1
−0.2
2014
2016
2018
2020
2022
2024
Date

```
Rejected the null hypothesis (H0),
the data is non-stationary
ADF: Test statistic      -23.894871
p-value                    0.000000
# of Lags                  4.000000
# of Observations       2612.000000
Critical Value (1%)       -3.432856
Critical Value (5%)       -2.862647
Critical Value (10%)      -2.567359
dtype: float64
-------------------------
Fail to reject the null hypothesis (H0),
the data is stationary
KPSS: Test statistic      0.107726
p-value                   0.100000
# of Lags                10.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
dtype: float64
```

Sai phân bậc 1 để chuyển dữ liệu về chuỗi dừng, kiểm định bằng ADF và KPSS cho thấy chuỗi đã dừng.

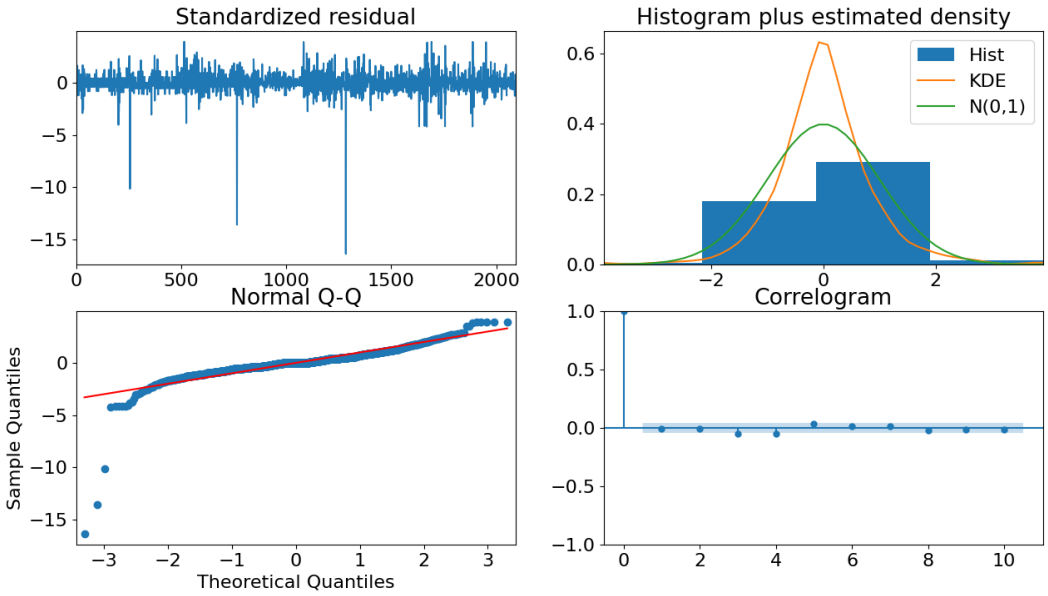
c) Xây dựng mô hình ARIMA

- Chia dữ liệu train và test theo tỷ lệ 80:20



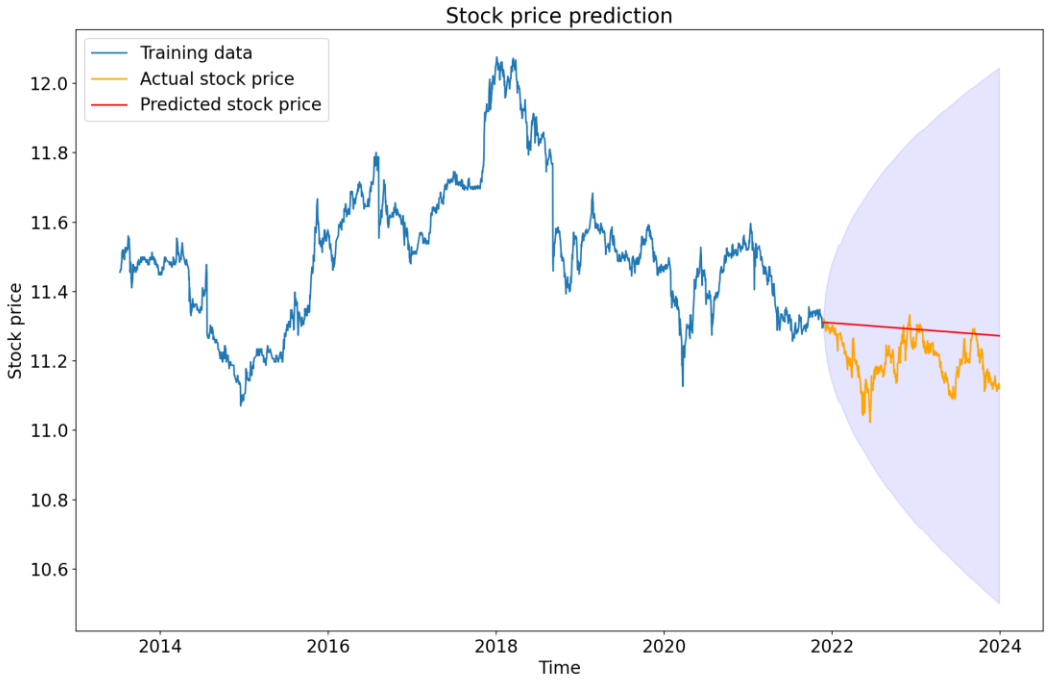
- Xác định tham số (p,d,q) = (0,1,0)

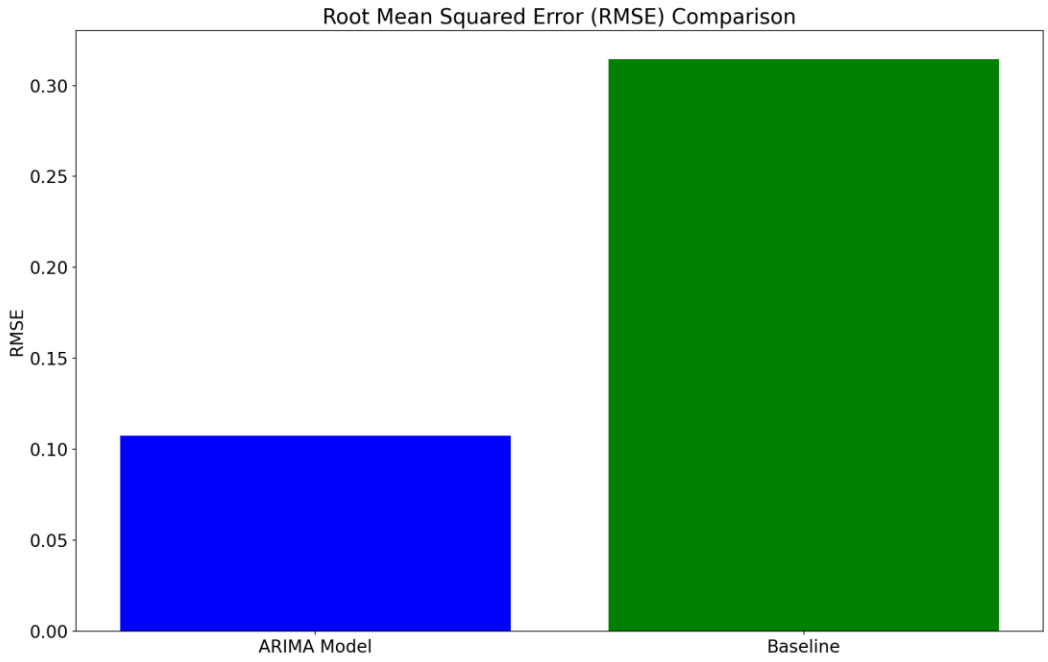
```
Performing stepwise search to minimize aic
 ARIMA(2,1,2)(0,0,0)[0] intercept   : AIC=-11057.635, Time=1.27 sec
 ARIMA(0,1,0)(0,0,0)[0] intercept   : AIC=-11060.337, Time=0.09 sec
 ARIMA(1,1,0)(0,0,0)[0] intercept   : AIC=-11058.389, Time=0.07 sec
 ARIMA(0,1,1)(0,0,0)[0] intercept   : AIC=-11058.389, Time=0.10 sec
 ARIMA(0,1,0)(0,0,0)[0]             : AIC=-11062.303, Time=0.04 sec
 ARIMA(1,1,1)(0,0,0)[0] intercept   : AIC=-11056.337, Time=0.09 sec

Best model:  ARIMA(0,1,0)(0,0,0)[0]
Total fit time: 1.664 seconds
                               SARIMAX Results
================================================================================
Dep. Variable:                      y   No. Observations:                 2094
Model:               SARIMAX(0, 1, 0)   Log Likelihood                5532.152
Date:                Tue, 09 Apr 2024   AIC                         -11062.303
Time:                        16:07:48   BIC                         -11056.657
Sample:                             0   HQIC                        -11060.235
                                - 2094
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
sigma2         0.0003   1.69e-06    175.288      0.000       0.000       0.000
===================================================================================
Ljung-Box (L1) (Q):                   0.05   Jarque-Bera (JB):            284362.19
Prob(Q):                              0.82   Prob(JB):                         0.00
Heteroskedasticity (H):               1.07   Skew:                            -3.76
Prob(H) (two-sided):                  0.35   Kurtosis:                        59.61
===================================================================================
```




```
Test MSE: 0.011556
Test RMSE: 0.107500
```

- Dự đoán và thực tế





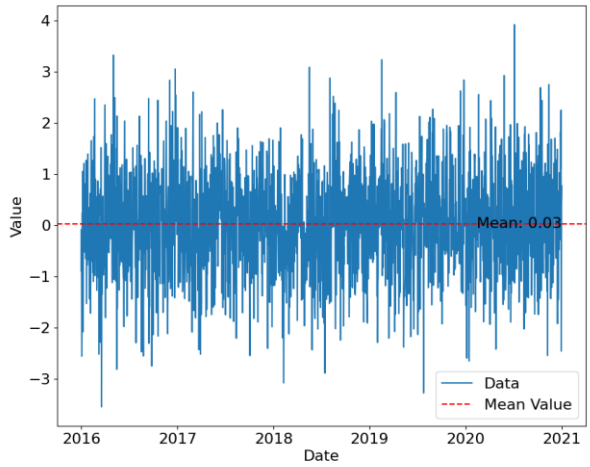
Mô hình hình có các chỉ số đánh giá nhỏ và có RMSE so với baseline thấp thể hiện hiệu suất dự đoán tốt hơn và đáng tin cậy hơn so với mô hình baseline.

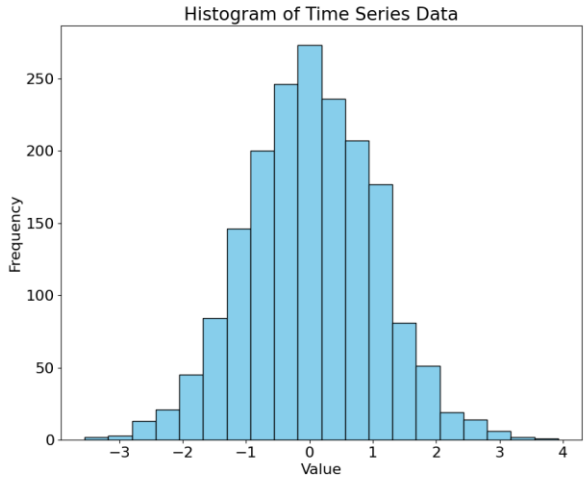
a) Expanding window

	Model	MAE_val	MSE_val	RMSE_val	MAE_test	MSE_test	RMSE_test
0	Fold_0	0.049495	0.003540	0.059499	0.051627	0.004234	0.065066
1	Fold_1	0.460980	0.213124	0.461654	0.051627	0.004234	0.065066
2	Fold_2	0.264610	0.070963	0.266389	0.051627	0.004234	0.065066
3	Fold_3	0.071419	0.005781	0.076034	0.051627	0.004234	0.065066
4	Fold_4	0.061497	0.005150	0.071760	0.051627	0.004234	0.065066

Từ bảng đánh giá có thể thấy mô hình của Fold_0 dự đoán hiểu quạ trên cả tập đánh giá và tập kiểm tra.

b) Dữ liệu tập tin **Data.xls**





Từ hai biểu đồ trên cho thấy giá trị trung gần bằng 0 và dữ liệu có phân phối chuẩn. Do đó, kết luận đây là chuỗi nhiễu trắng.

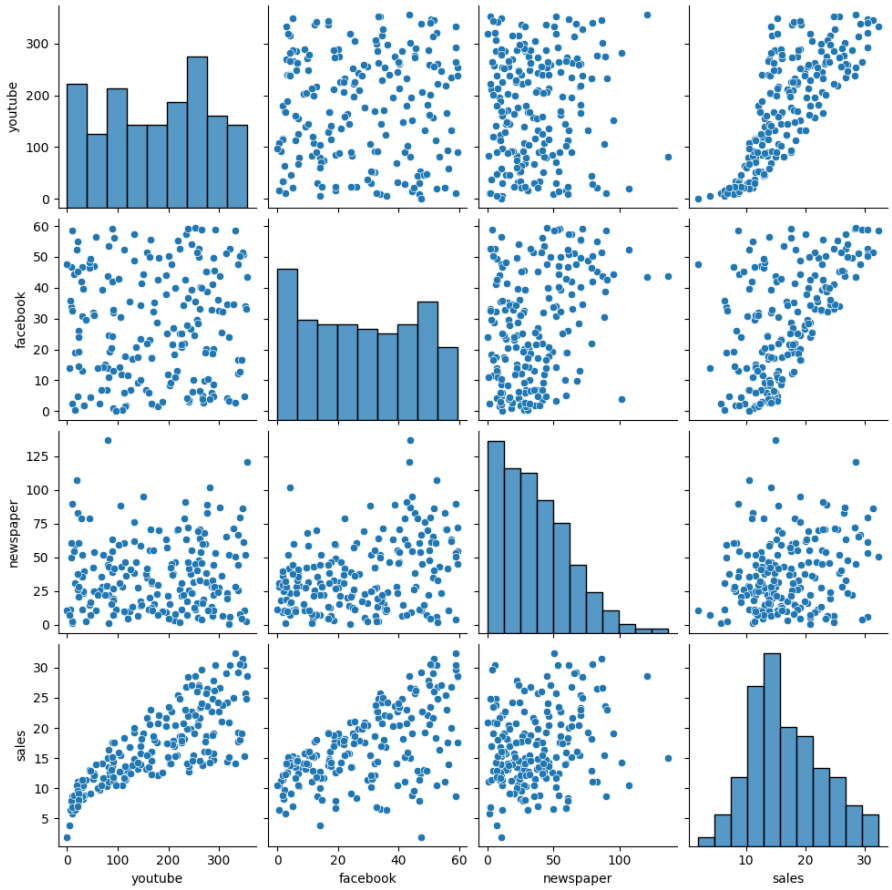
Câu 5:

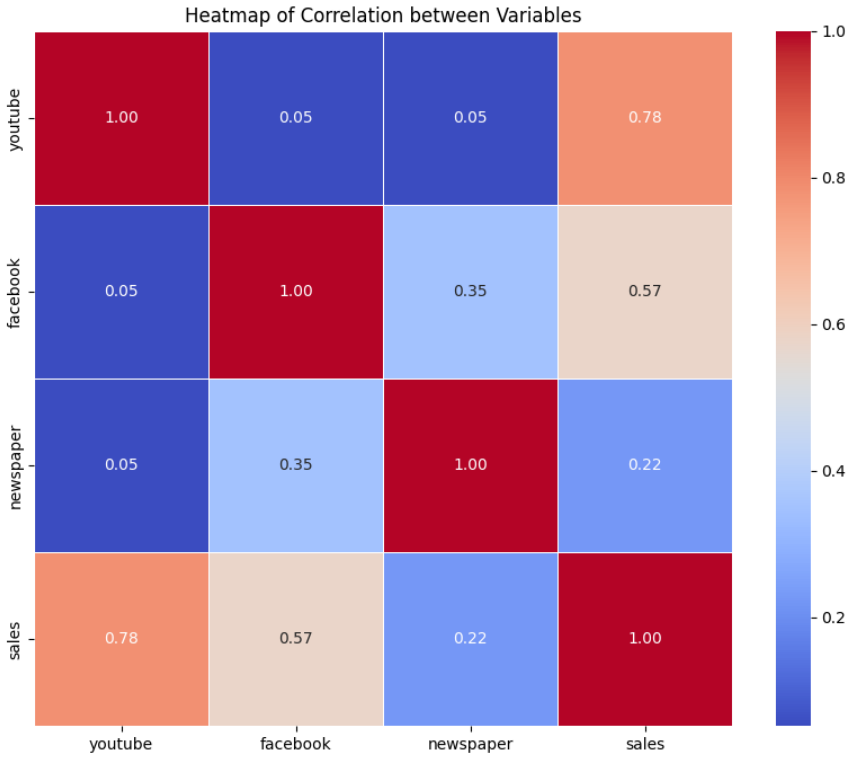
a) Mô tả

	youtube	facebook	newspaper	sales
count	199.000000	199.000000	199.000000	199.000000
mean	176.749749	28.011256	36.805628	16.853065
std	103.198035	17.810830	26.124069	6.265850
min	0.840000	0.000000	0.360000	1.920000
25%	88.860000	12.060000	15.420000	12.480000
50%	179.760000	27.960000	31.080000	15.480000
75%	262.980000	43.860000	54.120000	20.880000
max	355.680000	59.520000	136.800000	32.400000

Bộ dữ liệu gồm 5 biến: Id, youtube, facebook, newspaper và sales với 199 dòng.

b) Biểu đồ pairplot và heatmap. Giá trị tương quan cụ thể

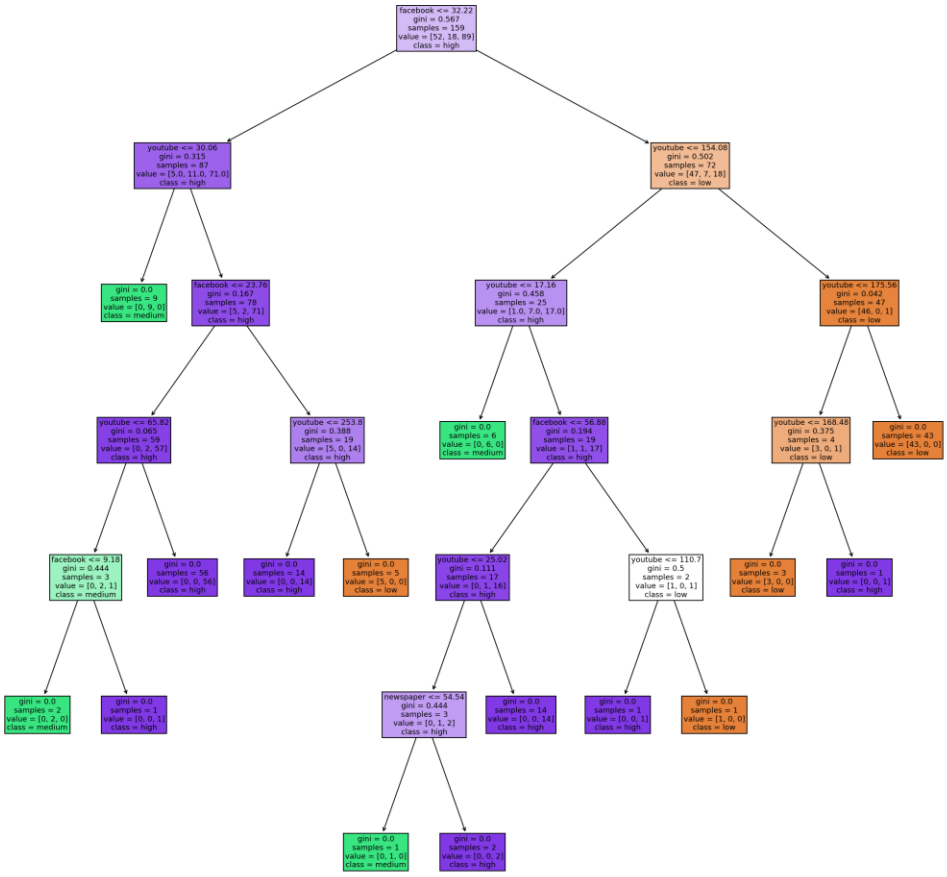

youtube
facebook
newspaper
sales
youtube
facebook
newspaper
sales



c) Số lượng mẫu tính theo 3 mức phân hoạch

```
Số lượng mẫu tin theo 3 mức phân hoạch :
sales_category
medium    118
high       59
low        22
```

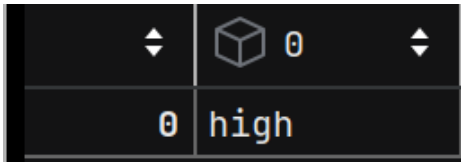
d) Trực quan hóa cây quyết định



e) Dự báo mức doanh thu

```
sales1=model.predict([[120,65,20]])
sales2=model.predict([[35,45,15]])
```

- youtube: 120, facebook: 65, newspaper: 20



- youtube: 35, facebook: 45, newspaper: 1

Câu 6:

Khi thực hiện kỹ thuật phân chia dữ liệu theo k-fold và cross-validation để đảm bảo mô hình có độ chính xác và khả năng tổng quát tốt, quy trình thực hiện như sau:

K-fold Cross-Validation:

- Chia dữ liệu thành k phần (k-folds).
- Lặp lại quá trình huấn luyện và kiểm tra k lần, mỗi lần chọn một fold làm tập kiểm tra và các fold còn lại làm tập huấn luyện.
- Tính trung bình độ chính xác của các lần kiểm tra để đánh giá hiệu suất của mô hình.

Xử lý Overfitting hoặc Underfitting:

- Overfitting: Khi mô hình quá phức tạp và hiệu suất trên tập huấn luyện cao nhưng trên tập kiểm tra thấp.
  → Giải pháp: Sử dụng kỹ thuật như regularization, giảm độ phức tạp của mô hình, thu thập thêm dữ liệu, hoặc sử dụng kỹ thuật early stopping.
- Underfitting: Khi mô hình quá đơn giản và không thể học được đặc điểm của dữ liệu.
  → Giải pháp: Tăng độ phức tạp của mô hình, thu thập thêm dữ liệu, hoặc chọn mô hình phù hợp hơn.