**Họ và tên:** Lâm Mỹ Ngọc

**MSSV:** K214110843

**BÀI LÀM**

**Câu 1:**

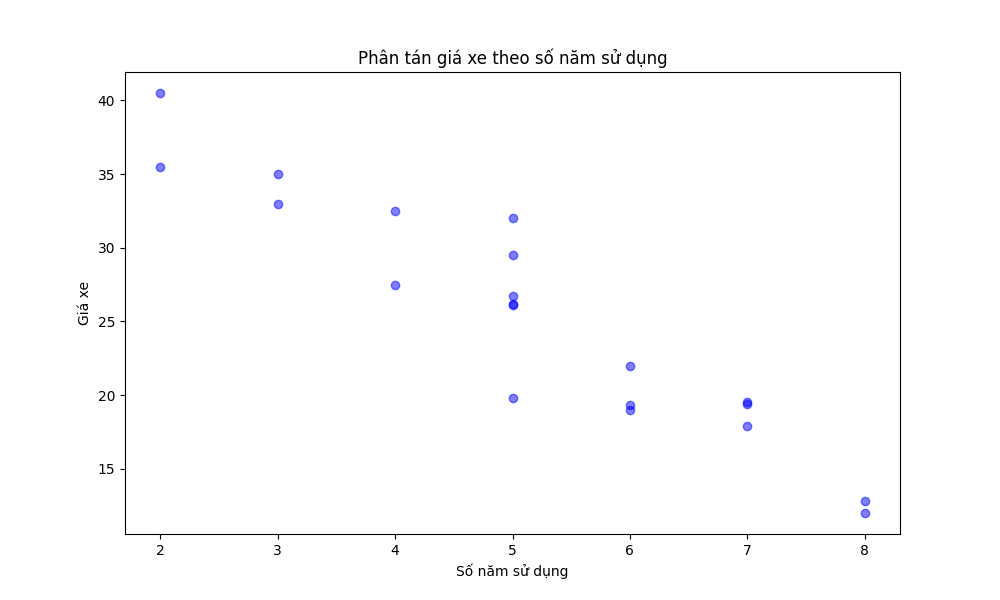
Theo em, liên quan đến ngành khoa học dữ liệu có các ngành nghề phổ biến sau:

* 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.

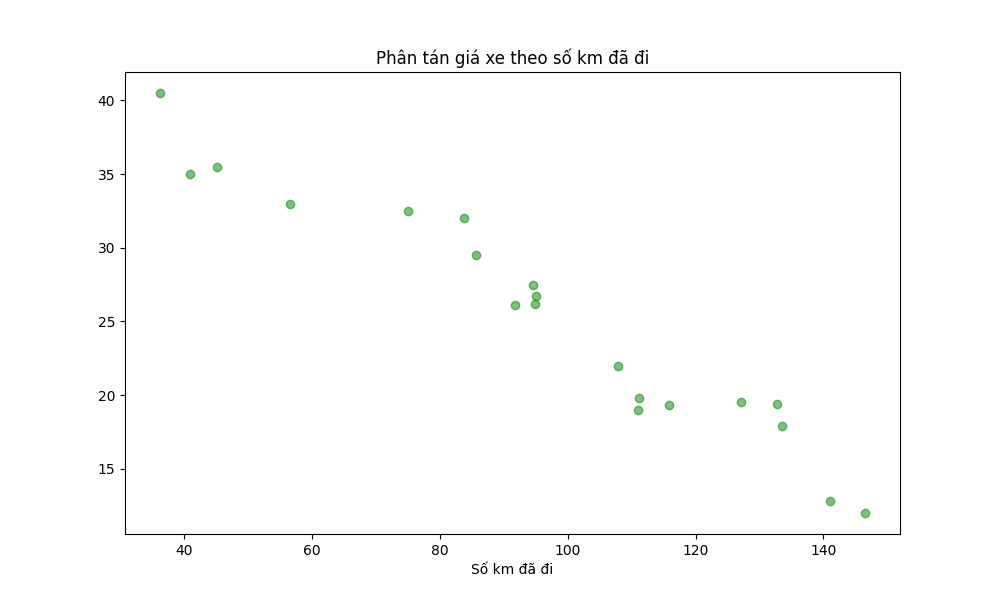
Để đáp ứng yêu cầu công việc cần chuẩn bị các kiến thức về thống kê , lập trình, kiến thức chuyên môn của ngành (chẳng hạn như phân tích dữ liệu tài chính thì cần có kiến thức về khách hàng trong lĩnh vực tài chính, quy trình hoạt động,…) và kiến thức về phân tích dữ liệu (quy trình phân tích, các mô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

**Câu 2:**

1. Biểu đồ phân tán

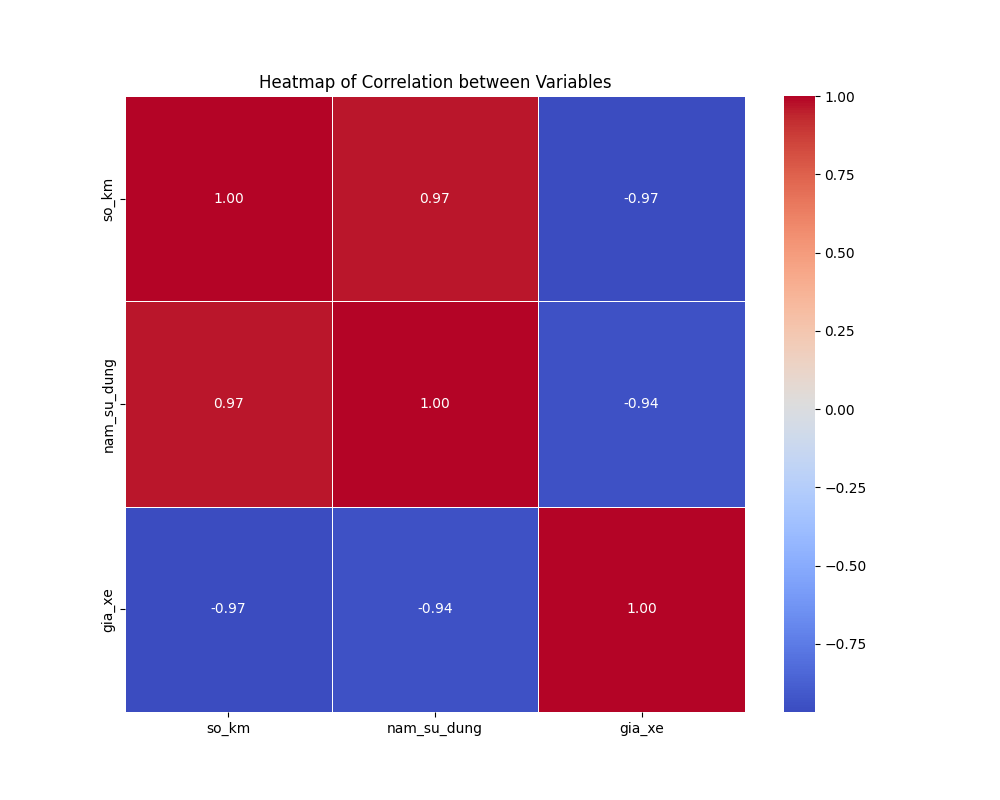


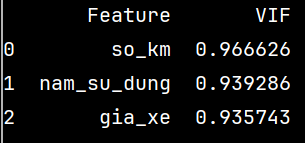
Biểu đồ thể hiện mối tương quan âm giữa giá xe và số năm sử dụng, tức là số năm sử dụng càng nhiều thì giá xe càng giảm và ngược lại.



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.

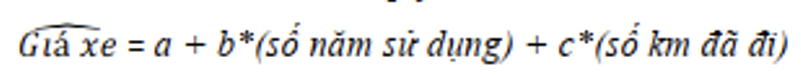
1. Biểu đồ Heatmap





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.

1. Ước lượng hệ số mô hình hồi quy

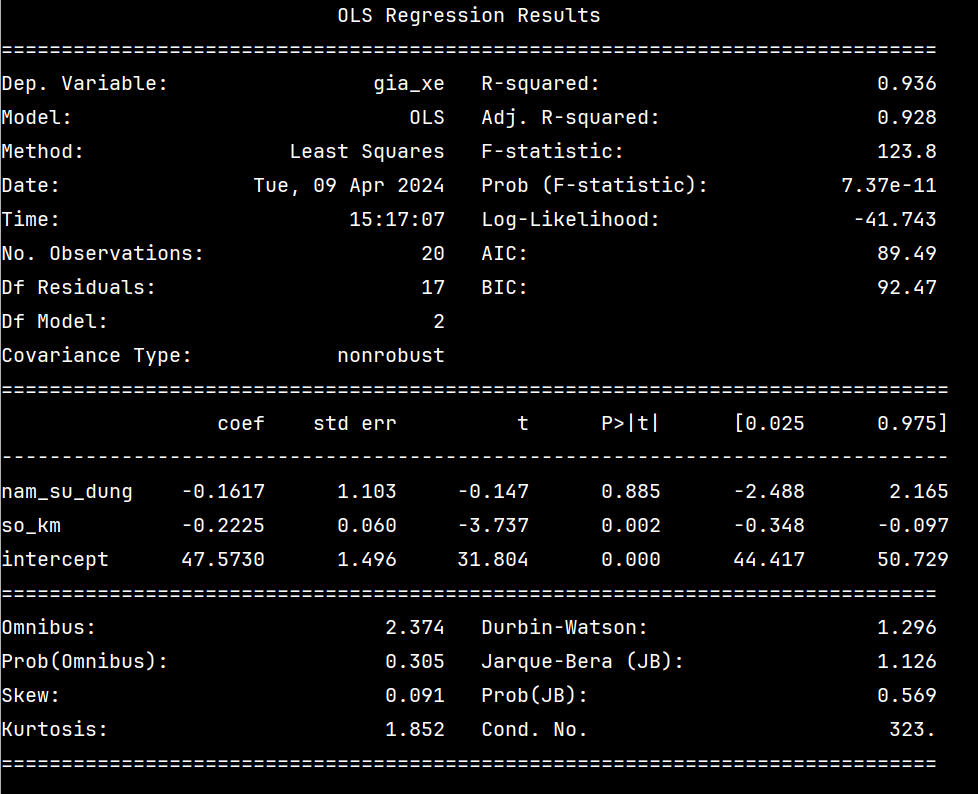


a = 47.5730

b = - 0.1617

c = - 0.2225

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



1. 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).

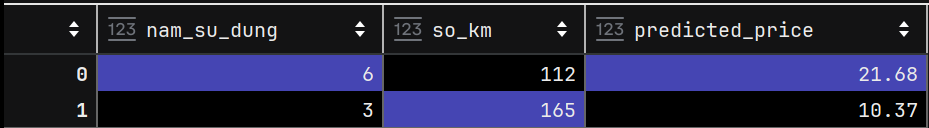
1. Kiểm định độ phù hợp của mô hình

**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.

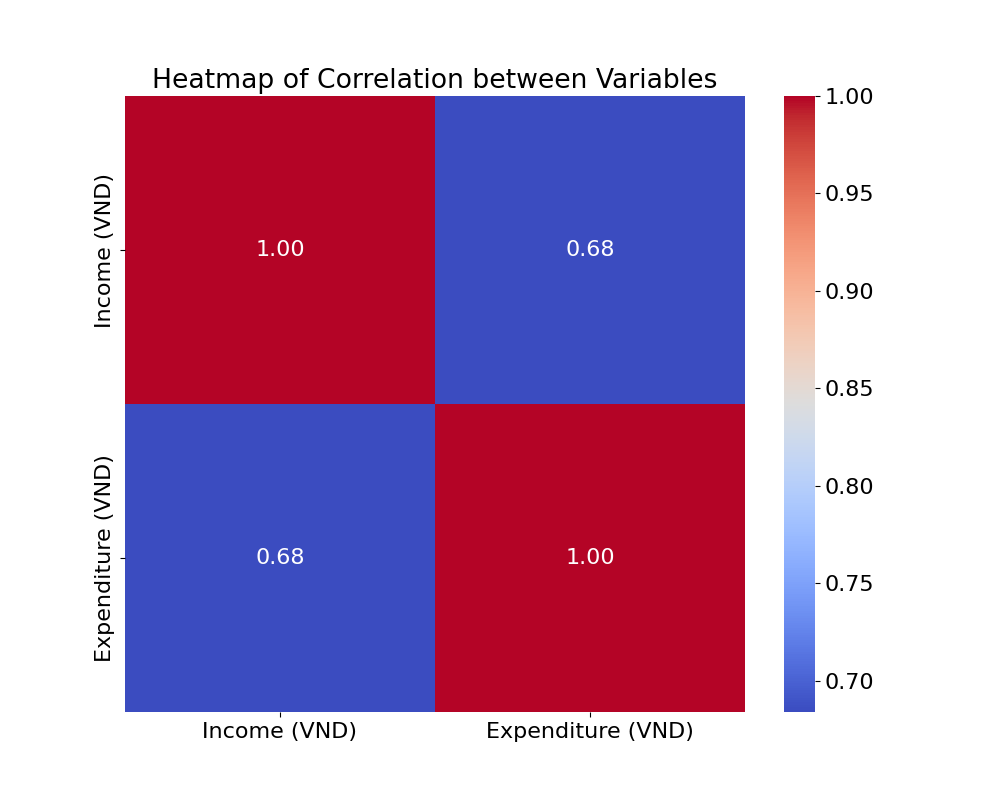
1. Lưu mô hình và thực hiện dự đoán



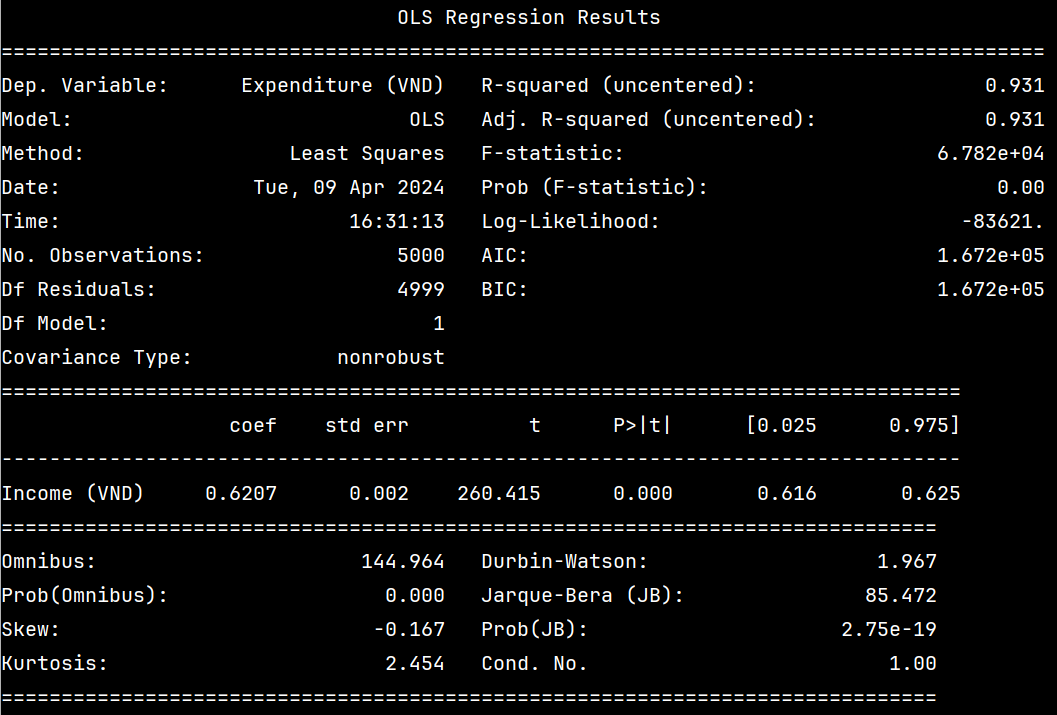
* Xe sử dụng 6 năm, đi được 112000Km, có giá dự đoán là 21.68 (đơn vị tiền tệ)
* Xe sử dụng 3 năm, đi được 165000Km, có giá dự đoán là 10.37 (đơn vị tiền tệ)

**Câu 3:**

1. Heatmap



1. Fit mô hình



* **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.

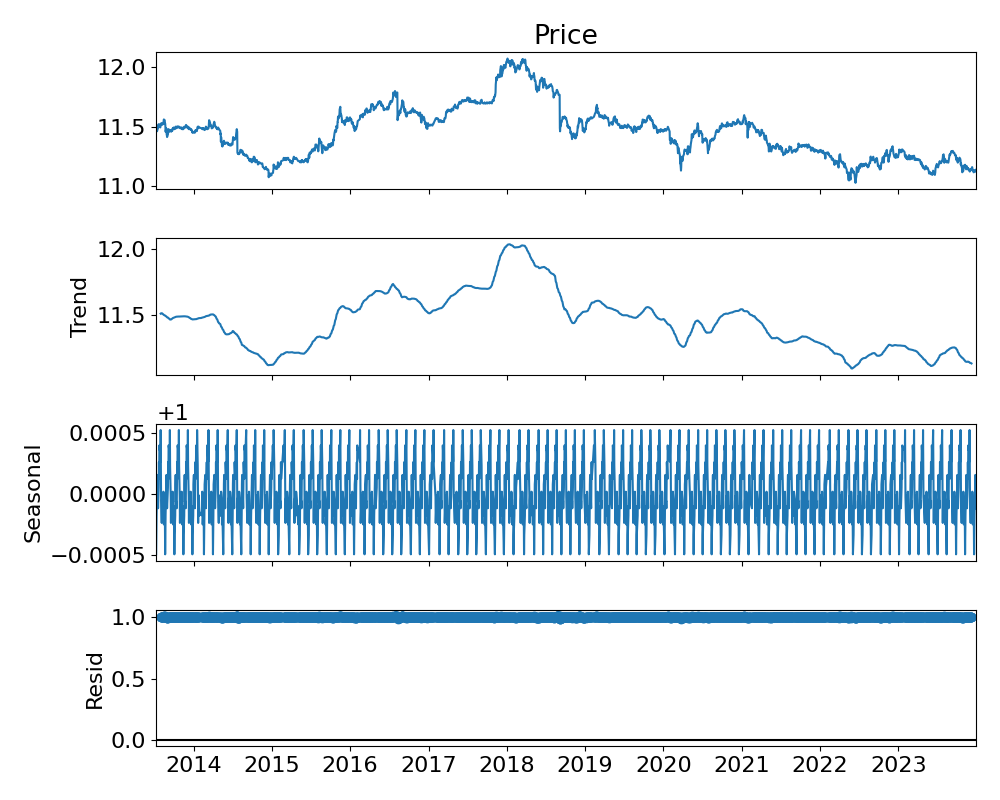
**F-statistic = 6.792e+04** và **p-value = 0,** p-value rất nhỏ, nhỏ hơn mức ý nghĩa 0,05, cho thấy mô hình có ý nghĩa thống kê.

1. Lưu mô hình và dự đoán

**Câu 4:**

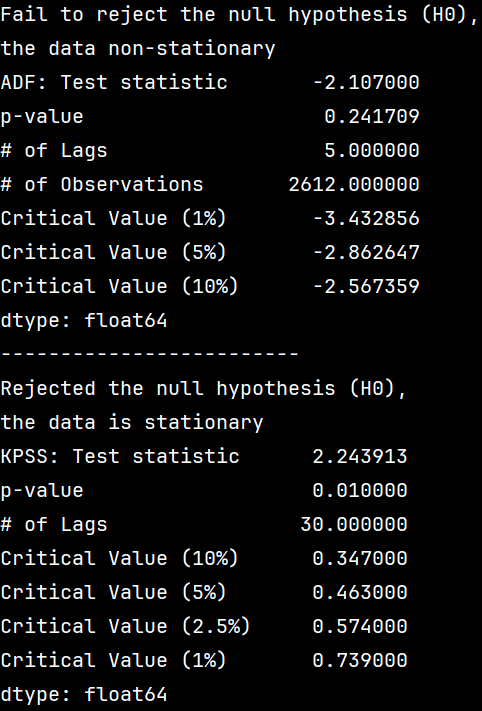
Dữ liệu được lấy trực tiếp từ website **Investing.com**

1. Phân rã dữ liệu

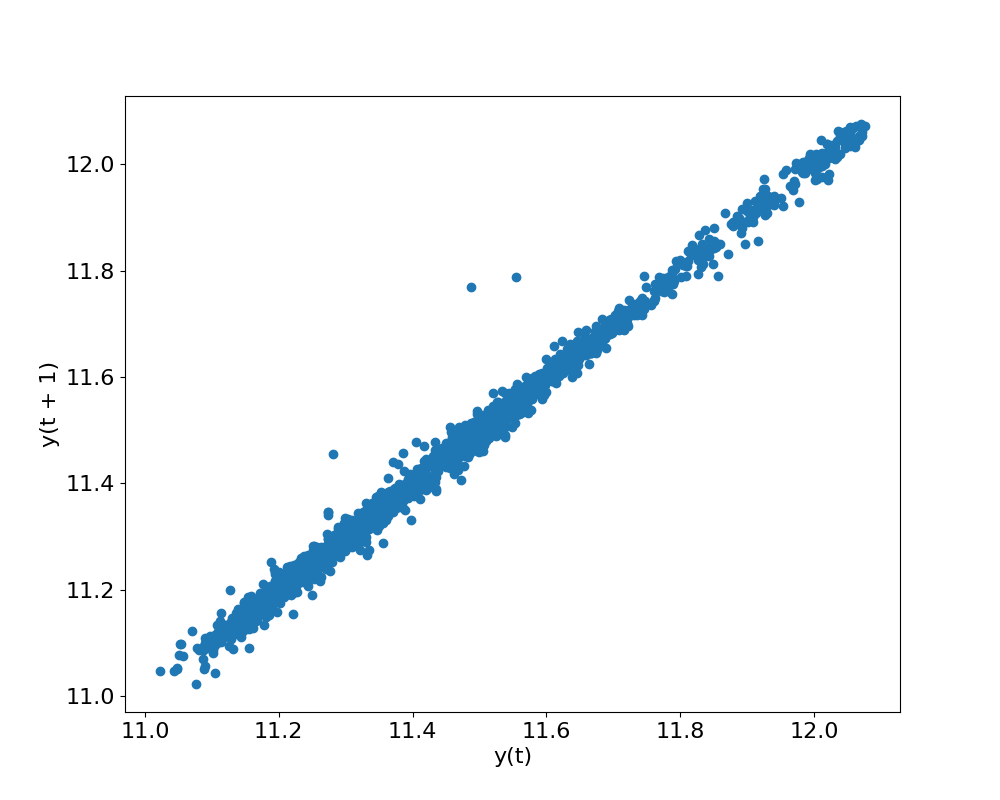


Dữ liệu có có sự biến động đáng kể theo thời gian, nhìn chung có xu hướng giảm nhẹ. Dữ liệu có tính mùa vụ với một chu kỳ lặp lại nhất định

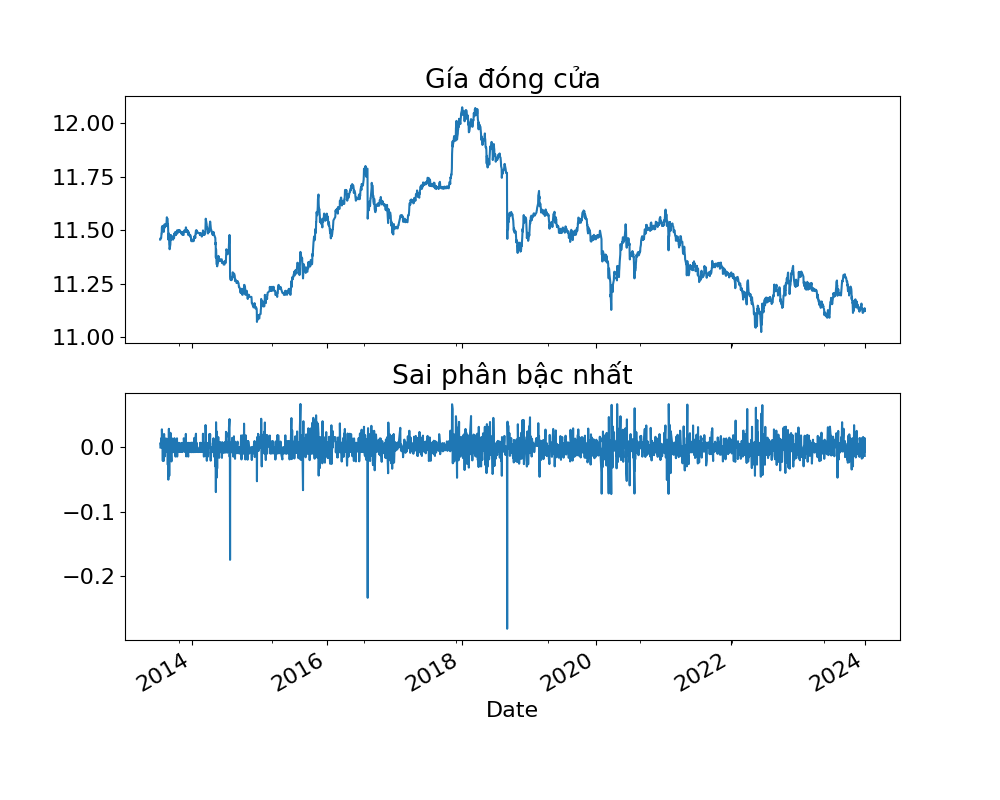
1. Kiểm định tính dừng, tương quan

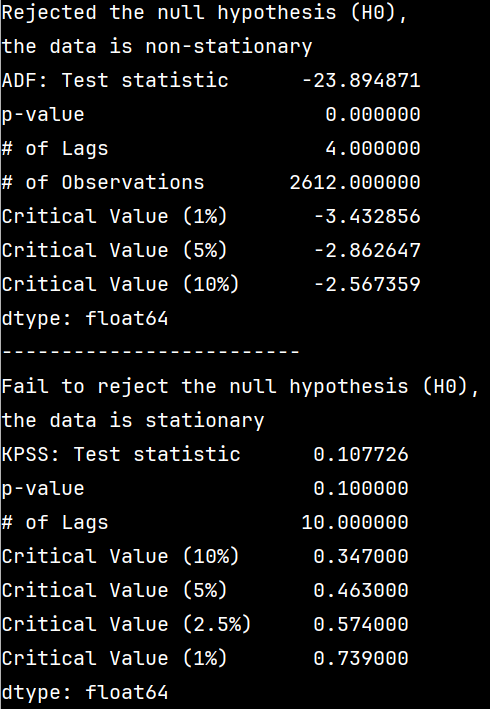


Đối với kiểm định ADF, p-value = 0.24 > 0.05 nên không thể bác bỏ Ho (chuỗi không dừng). Đối với kiểm đinh KPSS, p-value = 0.01 < 0.05 nên có thể bác bỏ Ho (chuỗi dừng). Từ 2 kiểm định trên có thể kết luận rằng chuỗi dữ liệu chưa dừng.



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.

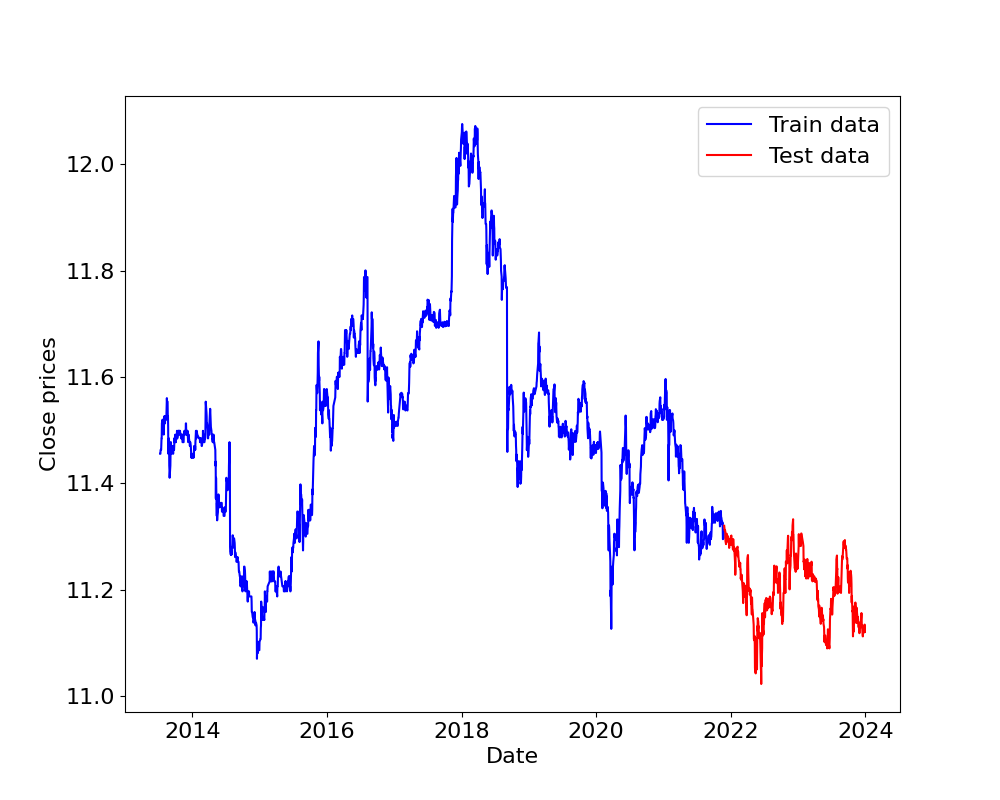




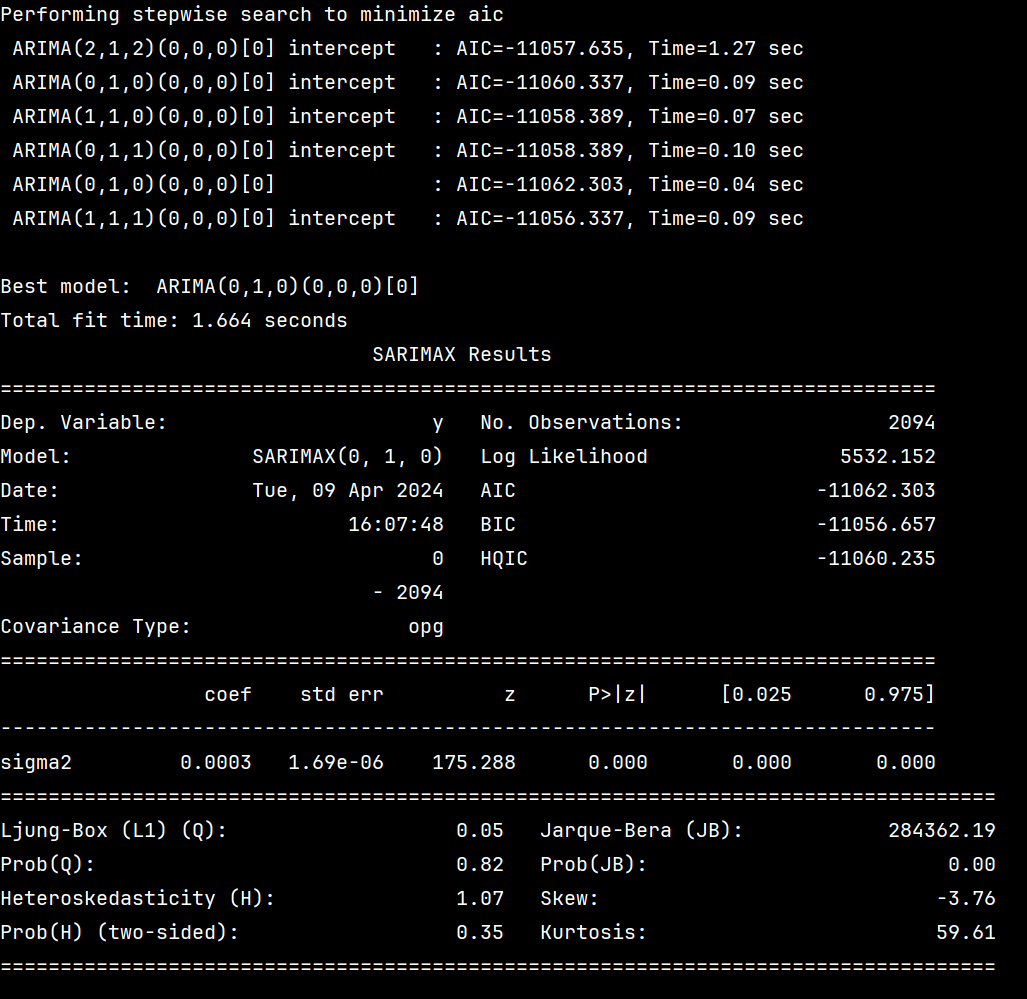
Sai phân bậc 1 để chuyển dữ liệu về chuỗi dừng, kiểm định bằng ADF và KPSS cho thấy chuỗi đã dừng.

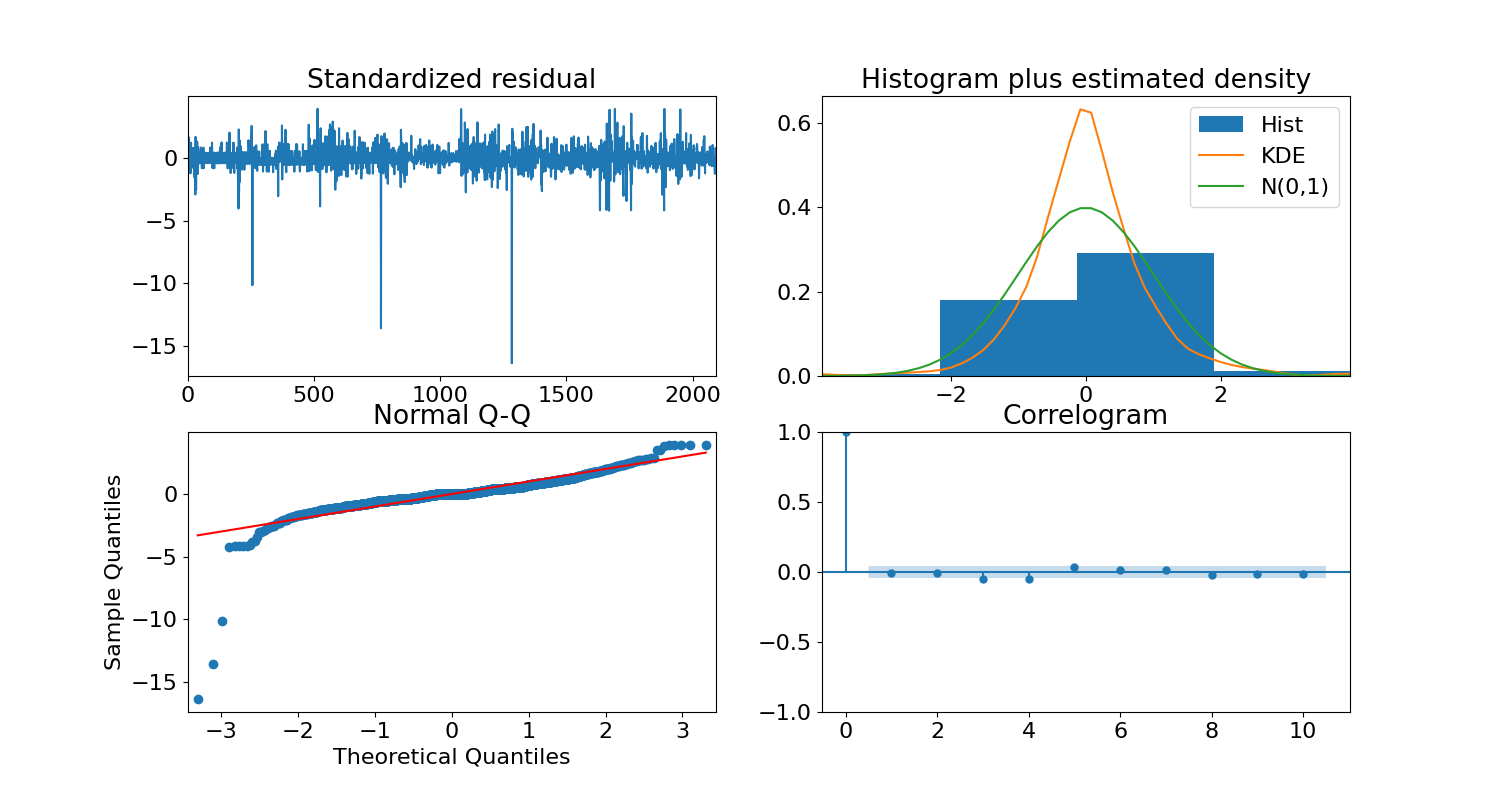
1. Xây dựng mô hình ARIMA

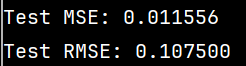
* Chia dữ liệu train và test theo tỷ lệ 80:20



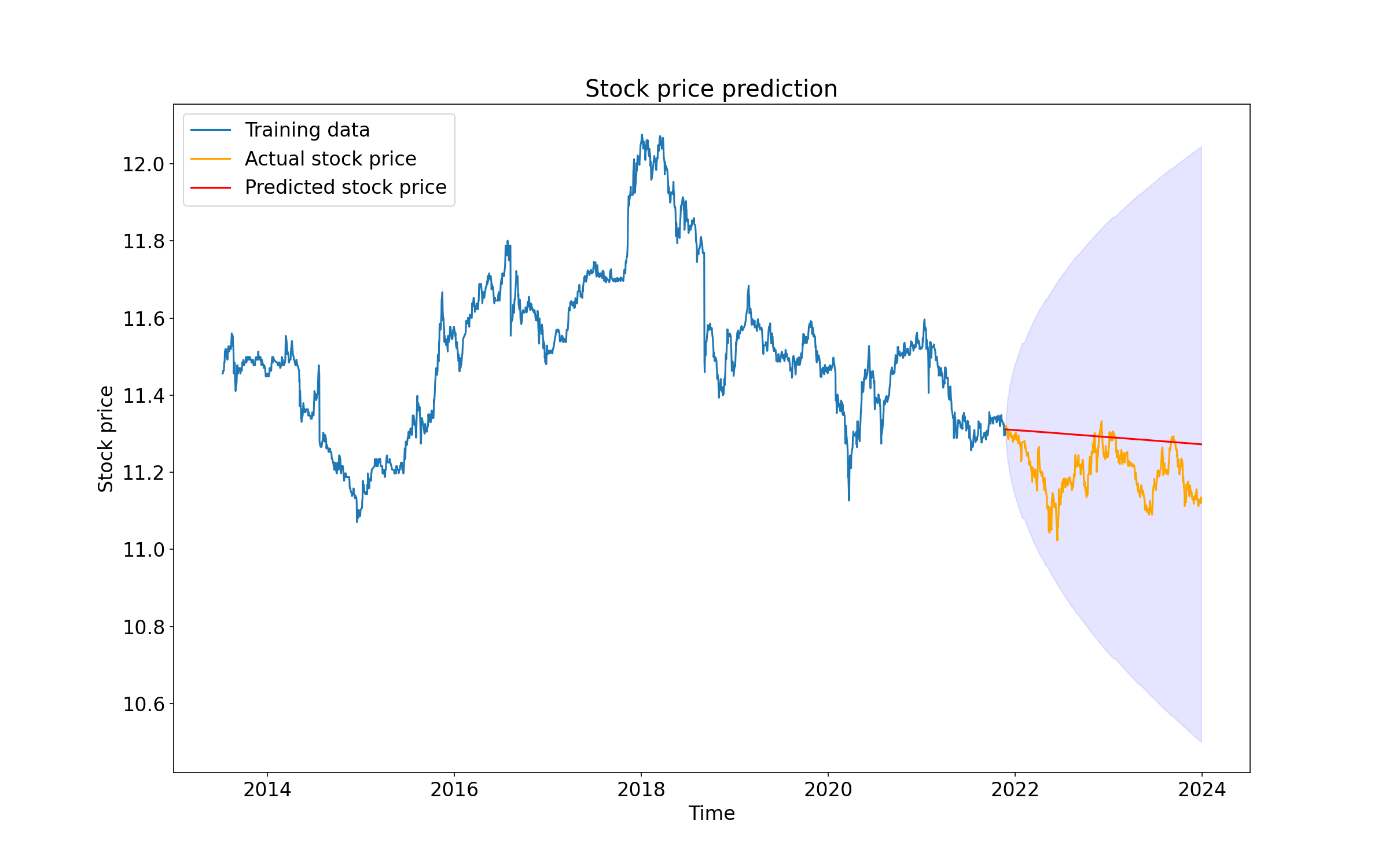
* Xác định tham số (p,d,q) = (0,1,0)

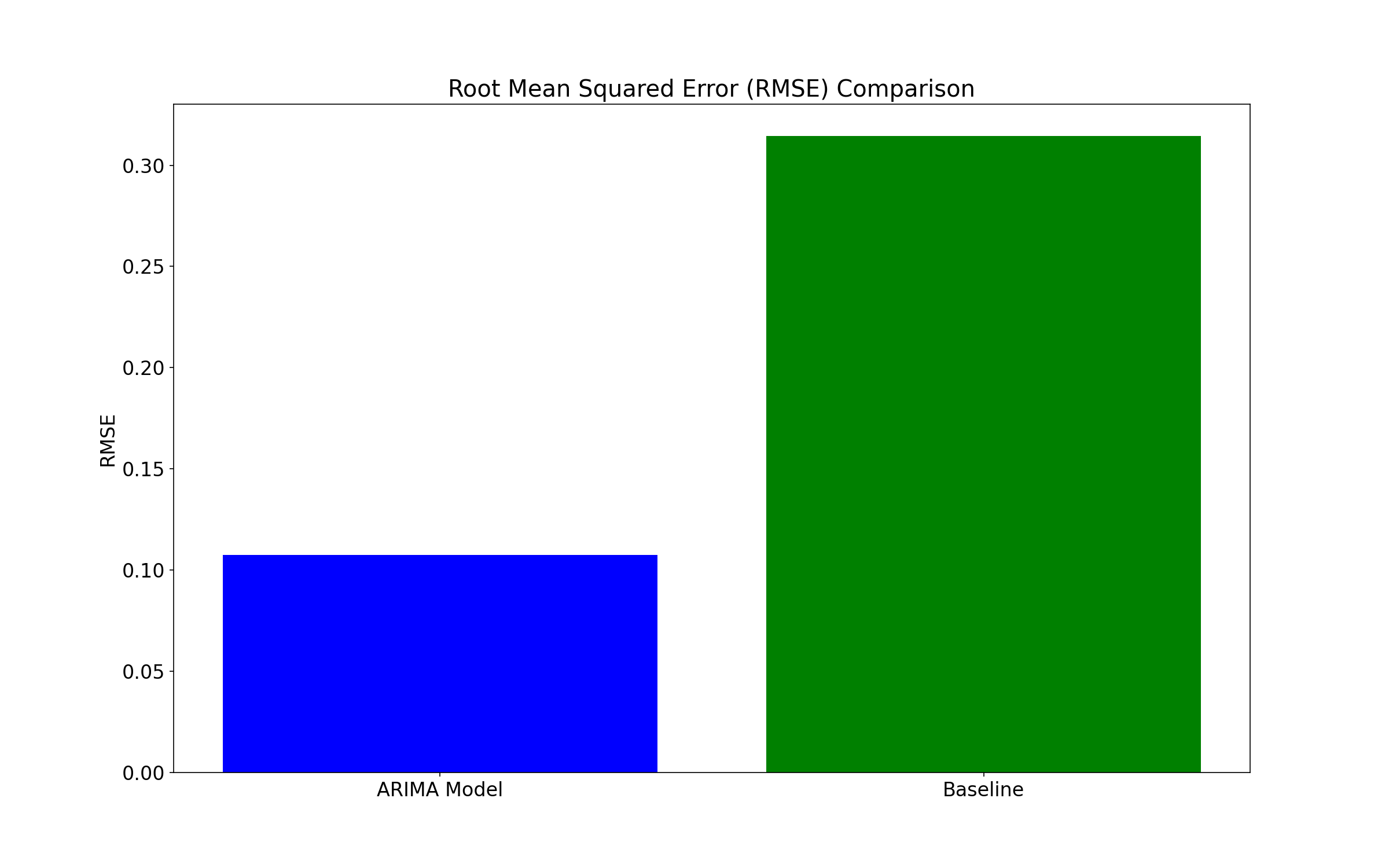






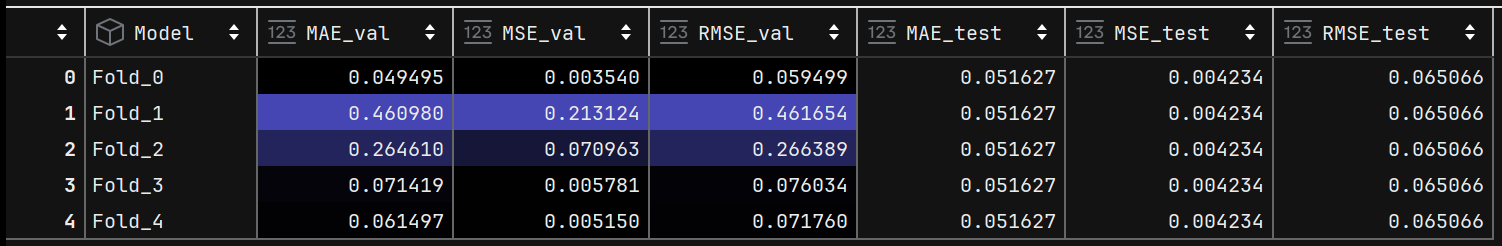
* Dự đoán và thực tế





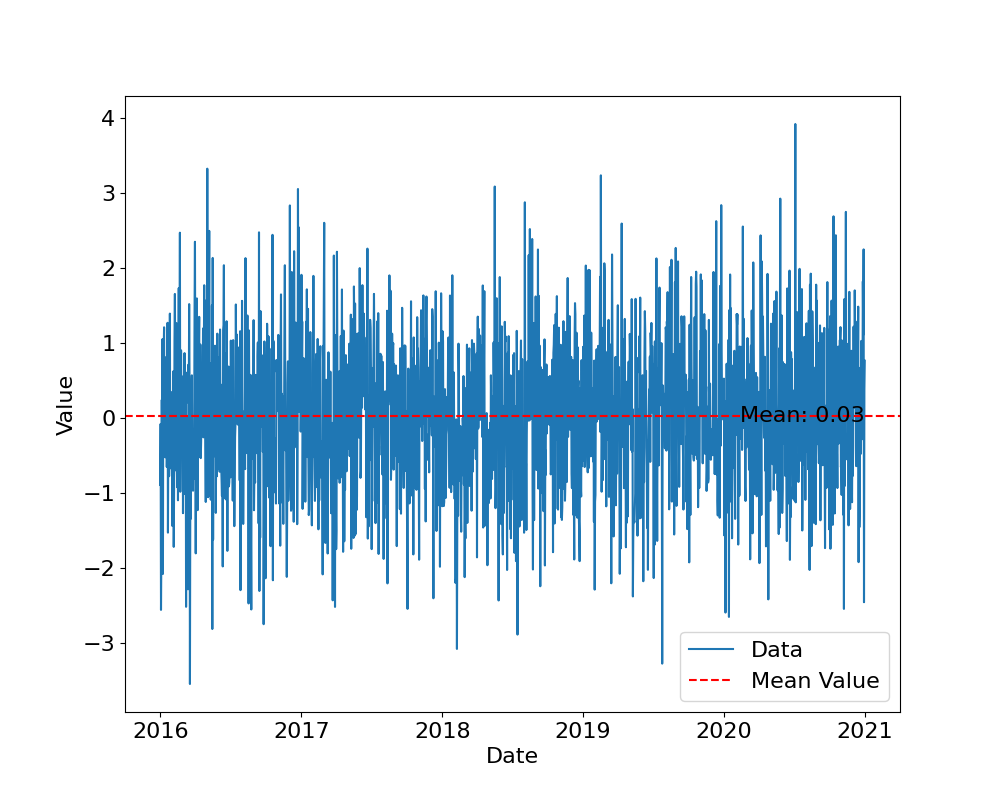
Mô hình hình có các chỉ số đánh giá nhỏ và có RMSE so với baseline thấp thể hiện hiệu suất dự đoán tốt hơn và đáng tin cậy hơn so với mô hình baseline.

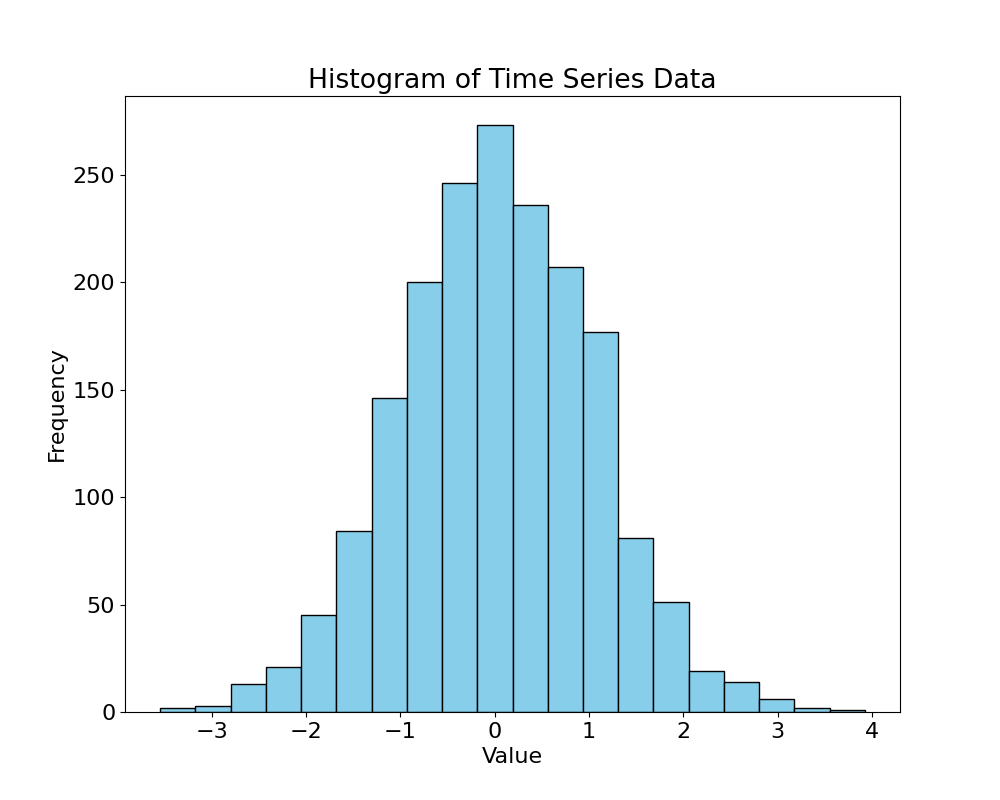
1. Expanding window



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.

1. Dữ liệu tập tin **Data.xls**

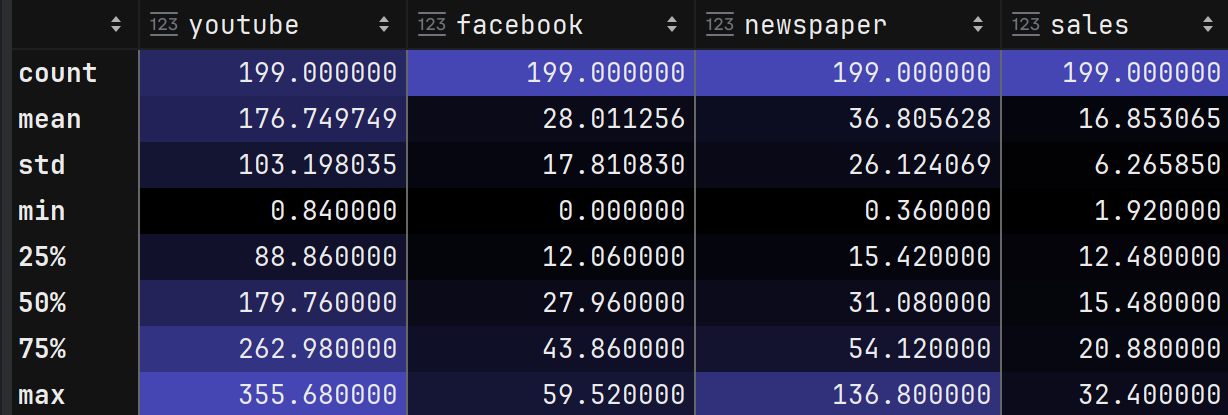




Từ hai biểu đồ trên cho thấy giá trị trung gần bằng 0 và dữ liệu có phân phối chuẩn. Do đó, kết luận đây là chuỗi nhiễu trắng.

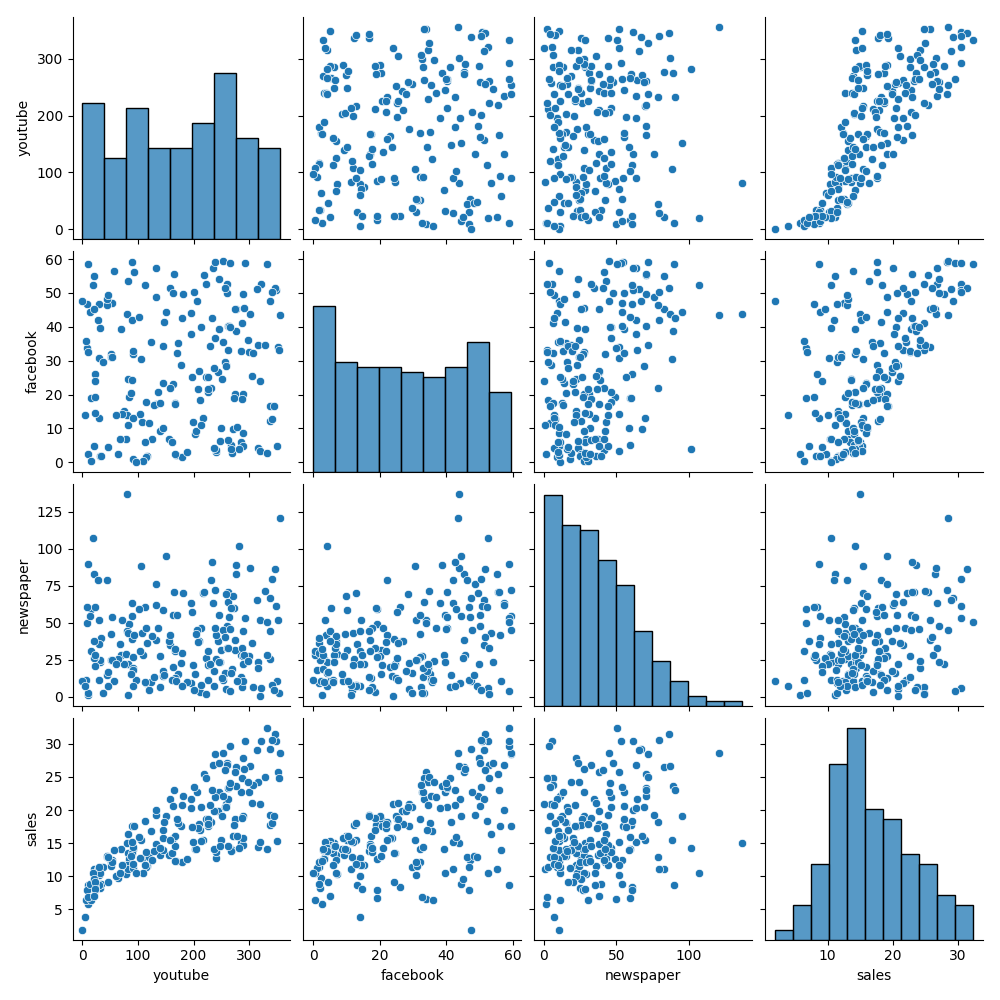
**Câu 5:**

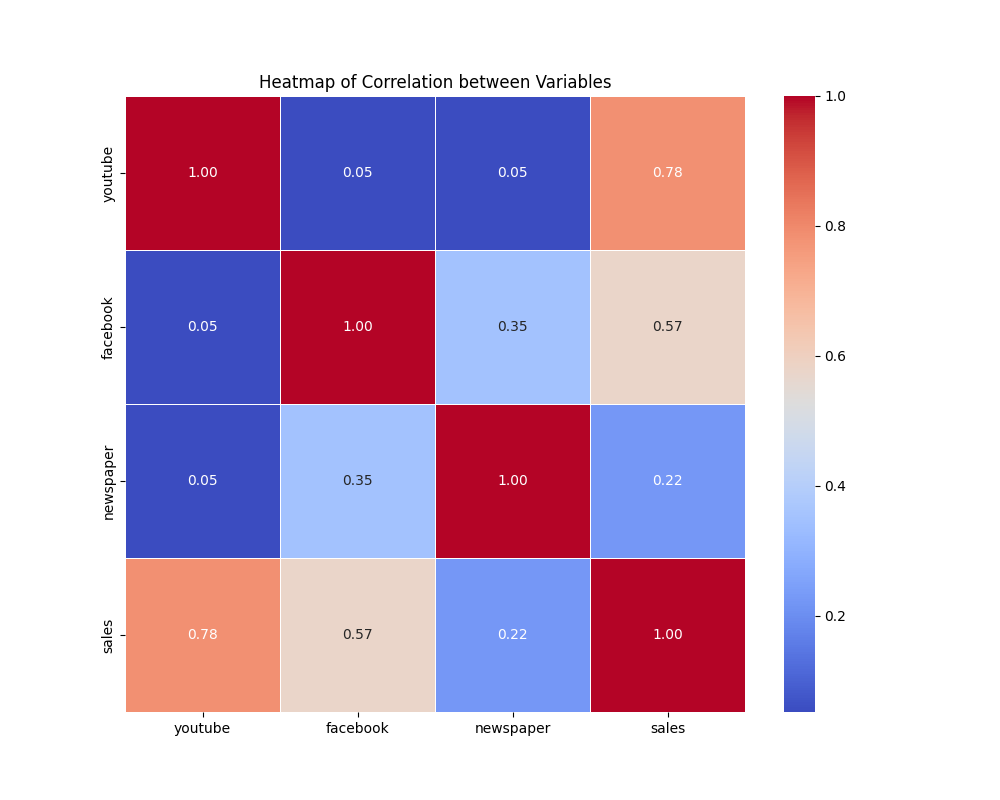
1. Mô tả



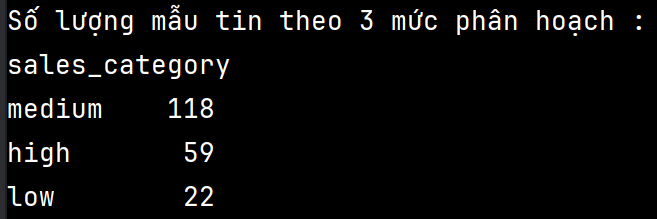
Bộ dữ liệu gồm 5 biến: Id, youtube, facebook, newspaper và sales với 199 dòng.

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

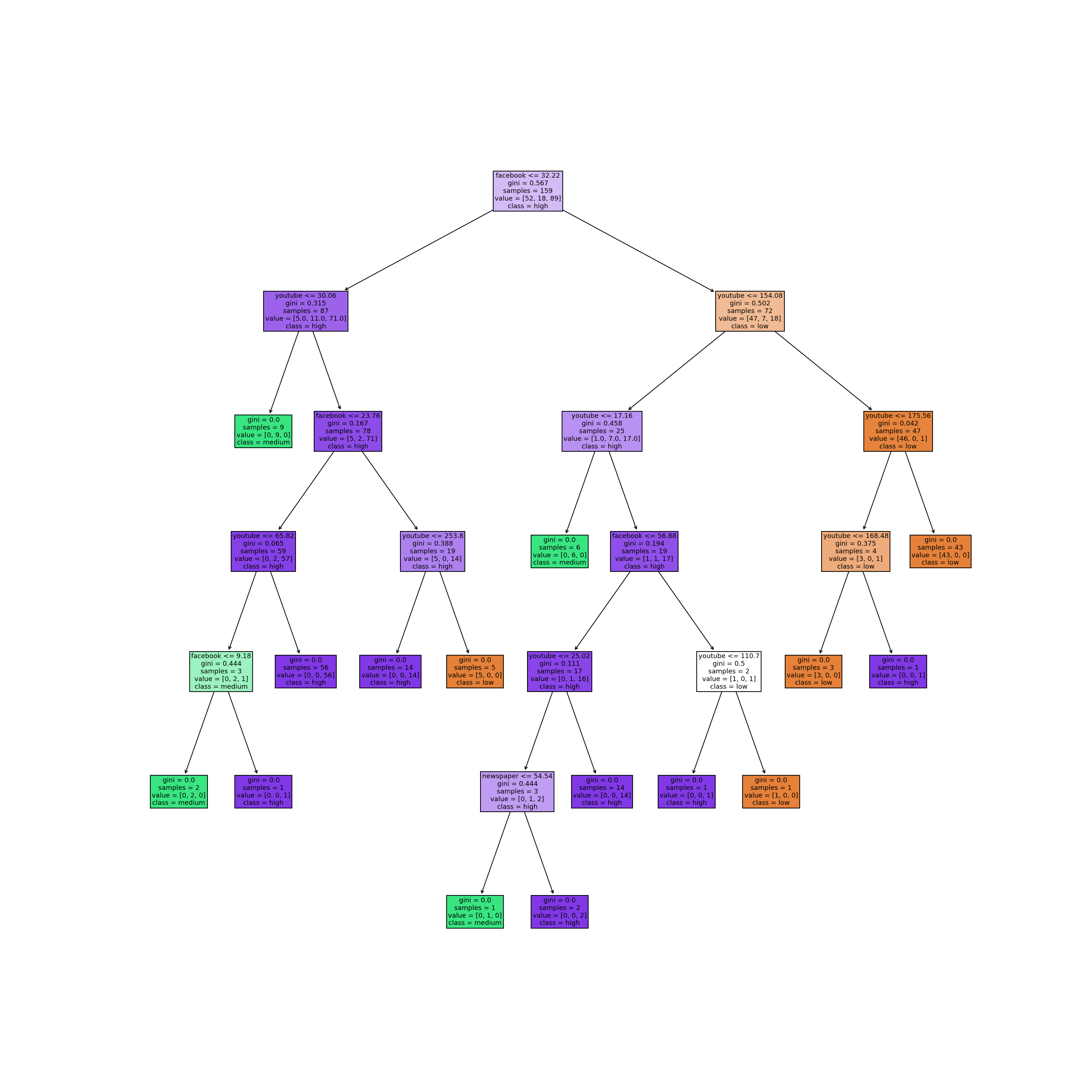




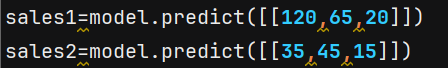
1. Số lượng mẫu tính theo 3 mức phân hoạch



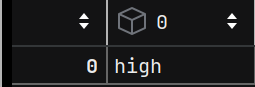
1. Trực quan hóa cây quyết định



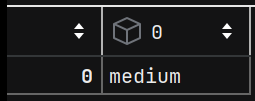
1. Dự báo mức doanh thu



* 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.