# MSV = 2251262593

# Họ và tên: Mai Ánh Dương - 2251262593

# Link github:

**Bước 1: Xác định "Đề bài"**

1. **Tính toán:** Đề bài = MSV % 20 + 1 Đề bài = Đề 14

# Bước 2: Trả lời câu hỏi

1. **Giải thích khái niệm “hồi quy giả” (spurious regession) khi làm việc với các chuỗi thời gian không dừng. Hậu quả của nó là gì ?**

Hồi quy giả là việc 2 thuộc tính của 1 dữ liệu có thể có độ tương quan nhưng thực ra không liên quan đến nhau. Trong chuỗi thời gian không dừng, có thể do 2 dữ liệu có cùng xu hướng nhưng thực ra do chúng tự phát triển chứ không liên qua đến nhau. Hậu quả cho nó là nhầm lẫn quan hệ của các thuộc tính đó, gây ra hệ số thống kê giả, mối quan hệ không tồi tại

1. **Trình bày về các tiêu chí thông tin như AIC (Akaike Information Criterion) và BIC (Bayesian Information Criterion). Chúng được sử dụng như thế nào trong việc lựa chọn mô hình ARMA(p,q) tốt nhất?**

Cả hai chỉ số AIC và BIC là 2 chỉ số đánh giá mô hình dựa trên độ tương quan, chỉ số càng thấp, thể hiện mô hình càng tốt đối với dữ liệu.

Có thể sử dụng AIC và BIC cho từng tham số có thể của mô hình ARMA(p,q) cho cùng 1 bộ dữ liệu và đánh giá dự vào 2 chỉ số

1. **Giả sử bạn ước lượng một mô hình AR(2) cho một chuỗi thời gian sản lượng bia hàng tháng (đã được làm dừng nếu cần). Làm thế nào để bạn kiểm tra xem các hệ số của mô hình (φ1,φ2) có ý nghĩa thống kê hay không? Khoảng tin cậy của các hệ số này cho biết điều gì?**

Để kiểm tra xem các hệ số của mô hình có ý nghĩa thống kê hay không, ta có thể p-value. P-value là xác xuất quan sát được một giá trị thống kê trong dữ liệu không liên qua đến giá trị dự đoán. Còn khoảng tin cậy của hệ số cho biết liệu giá trị thức của hệ số có thể nằm trong khoảng nào, nếu khoảng đó chứa 0 nghĩa là hệ số đó có khả năng bằng 0 (Hệ số không liên qua giá trị dự đoán.

**Thực nghiệm**

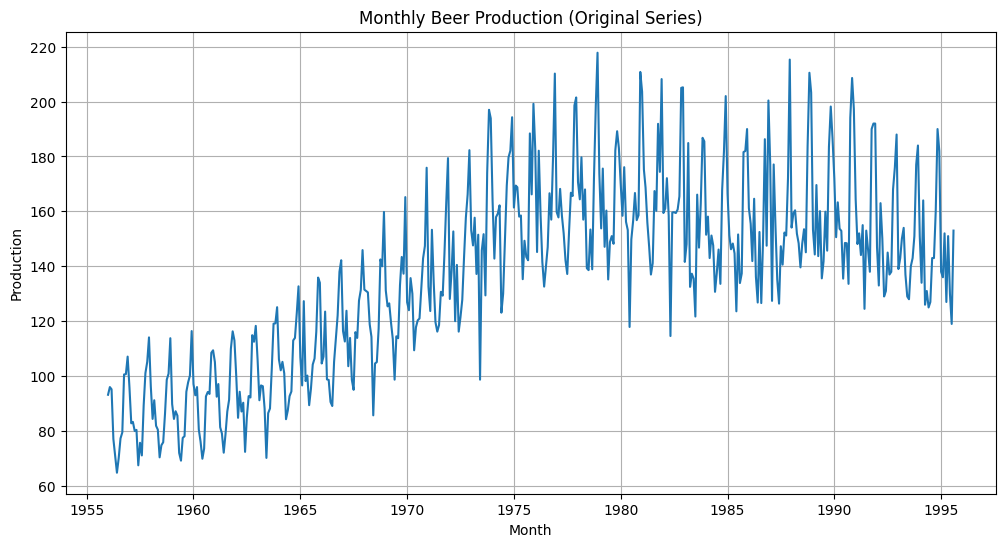
**o Dữ liệu: Monthly Beer Production in Australia**

**Link Kaggle:** [**https://www.kaggle.com/datasets/kalilurrahman/monthly-beer**](https://www.kaggle.com/datasets/kalilurrahman/monthly-beer) **-production-in-australia**

* 1. **Đọc hiểu dữ liệu**

Dữ liệu có 2 trường: Month: Thời gian từng lần đo giá trị, cách nhau 1 tháng

Monthly beer production: Giá trị bia bán được trong tháng đó

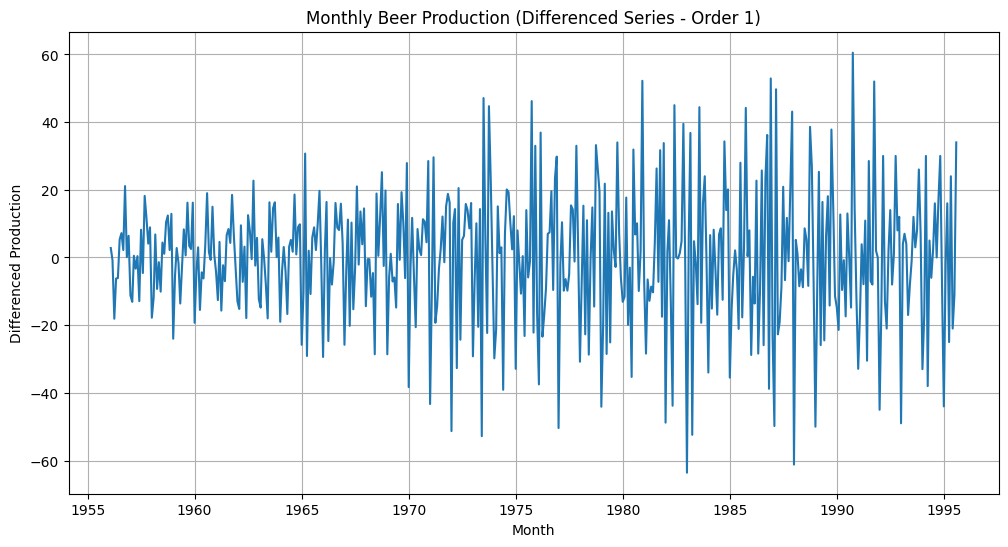
****

* 1. **Xác định tính dừng**

Sử dụng ADF để xác định tính dừng



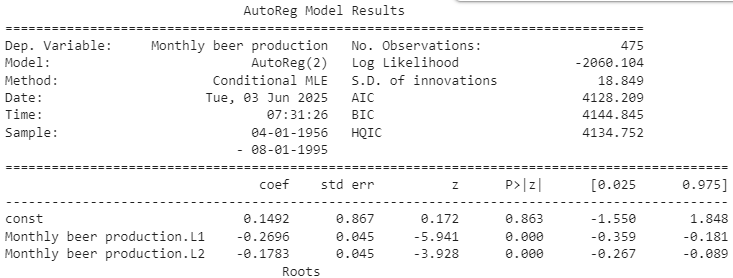
Giá trị p-value > 0.05 --> Dữ liệu không có tính dừng --> Lấy sai phân bậc 1



****Giá trị p-value < 0.05 --> Dữ liệu có tính dừng

* 1. **Dùng mô hình AR(2)**

Sử dụng thư viện statsmodels chạy AR(2) với dữ liệu sai phân bậc 1 hiển thị kết quả



Có thể thấy giá trị p-value của 2 hệ số đều bằng 0 và khoảng tin cậy đều không chứa 0 --> Cả hai hệ số có liên quan đến giá trị dự đoán

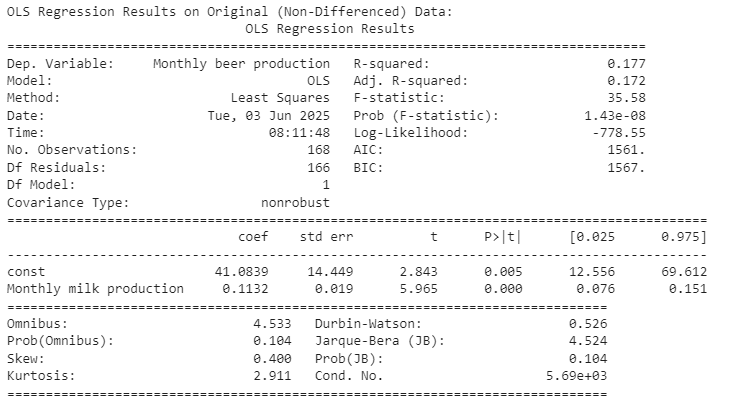
**o (Thêm cho phần hồi quy giả): Có thể xem xét thêm Monthly Milk Production (https://www.kaggle.com/datasets/gauravduttakiit/monthly-milk-production) và thử hồi quy hai chuỗi này với nhau (sau khi kiểm tra tính dừng).**

* 1. **Thêm dữ liệu**

Trong dữ liệu được thêm có khoảng thời gian ngắn hơn, ta sẽ thử nghiệm với khoảng thời gian ngắn hơn đấy

* + 1. **Kiểm tra tính dừng**

Để thử tính hồi quy giả, ta dùng dữ liệu không dừng ban đầu của cả 2 để hồi quy 2 chuổi với nhau, ở đây ta dùng milk dự đoán cho beer



Dựa trên bảng phân tích, ta thấy dù p-value và khoảng tin cậy rất đẹp nhưng R-squared khá cao và đây la dữ liệu không dừng nên rất có thể đây là trường hợp “hồi quy giả”