Họ và tên: Ngô Quang Minh  
Lớp: 64TTNT2  
MSV: 2251262618

**Câu 1. Thế nào là bộ nhớ dài và bộ nhớ ngắn của chuỗi thời gian, liên hệ với ACF**

**Bộ nhớ ngắn (Short memory)** nghĩa là: Giá trị của chuỗi hiện tại chỉ bị ảnh hưởng bởi một vài giá trị gần đây, còn những giá trị xa hơn trong quá khứ thì gần như không ảnh hưởng gì tới giá trị hiện tại

**Bộ nhớ dài (Long memory)** nghĩa là: Giá trị hiện tại vẫn chịu ảnh hưởng rất nhiều bởi những giá trị xa xưa trong quá khứ, tức là lượng thông tin được lưu trữ lâu dài qua các chuỗi

Để liên hệ với hàm ACF thì ta cần biết hàm ACF là gì trước. Bản chất của hàm ACF là đo mức độ tương quan giữa giá trị hiện tại và giá trị cách nó k bước trước đó. Vậy thì với **bộ nhớ ngắn (short memory)**, mức đó tương quan sẽ giảm nhanh, nghĩa là chỉ vài bước trước bước hiện tại thì mức độ tương quan đã giảm đi gần như không còn. Còn đối với **bộ nhớ dài (long memory)** thì độ tương quan giảm rất chậm kể cả khi xét tại thời điểm rất xa so với giá trị hiện tại

**Câu 2. Phương pháp san bằng mũ Holt-Winters để xử lý chuỗi thời gian có xu hướng và mùa vụ**

**Mục đích:** Khi dự báo chuỗi thời gian, dữ liệu thường có nhiễu (noise) chính là các biến động ngẫu nhiên khó dự đoán. Vậy thì phương pháp san bằng mũ giúp làm mượt chuỗi dữ liệu, giảm ảnh hưởng của những biến động bất thường.

**Ý tưởng:** Gán trọng số giảm dần theo cấp số nhân cho các giá trị trong quá khứ, nghĩa là giá trị gần giá trị hiện tại sẽ được coi trọng hơn còn giá trị trong quá khứ xa thì sẽ càng ít ảnh hưởng.

**Ứng dụng trong Holt-Winters:** holt-wintersmở rộng san bằng mũ cho chuỗi có xu hướng và mùa vụ bằng cách san bằng cho từng thành phần: Level (Mức trung bình hiện tại), Trend, Seasonality. Mỗi thành phần sẽ được cập nhật theo công thức có hệ số làm mượt (alpha, beta, gamma) để cân bằng dữ liệu mới và quá khứ.

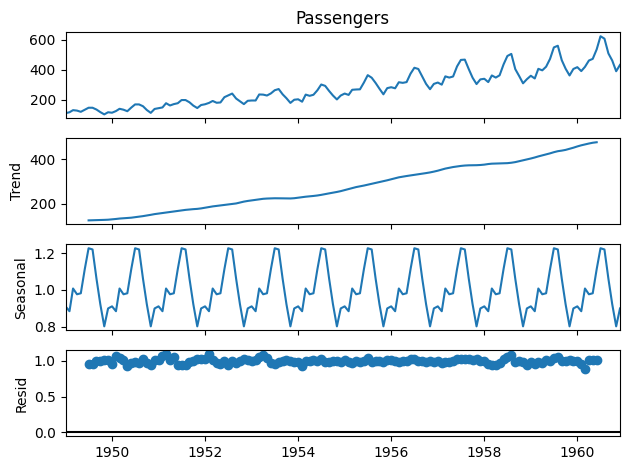
**Câu 3. Dự báo số lượng cuộc gọi đến một tổng đài hỗ trọ khách hàng hàng ngày**

**A. Thu thập dữ liêu**

Dataset được thập từ Kaggle về International Airline Passengers từ năm (1949-1960) với 144 mẫu dữ liệu được thu thâp và gồm 2 trường thông tin chính là Month thể hiện mốc thời gian mẫu đó được ghi và Passengers là số lượng hành khách có trong tháng đó. Mặc dù sử dụng bộ dữ liệu về airline passengers nhưng bộ dữ liệu này hoàn toàn có thể áp dụng cho dự báo cuộc gọi tổng đài vì cả hai đều là ở dạng timeseries data với seasonal patterns tương tự

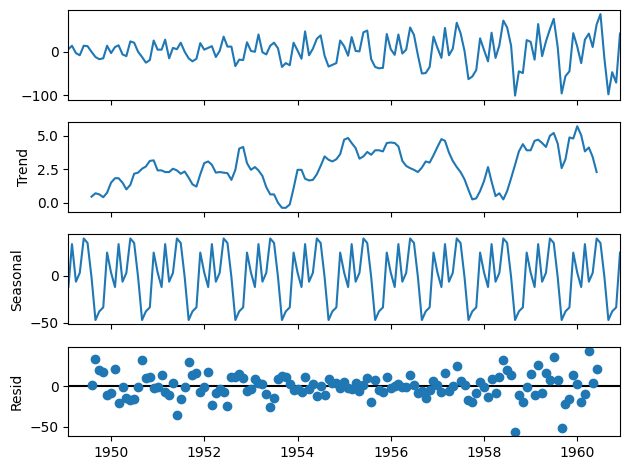
**B. Khám phá dữ liệu (EDA)**

Trước tiên ta cần trực quan hoá dữ liệu chuỗi thời gian bằng biểu đồ đường. Để nhìn thấy tổng quan xu hướng, quan sát tính mùa vụ của bộ dữ liệu. Trong Hình 1 giá trị Passengers này chính là các cuộc gọi đến tổng đài, ta có thể thấy xu hướng của bộ dữ liệu đang có xu hướng gia tăng theo thời gian, hơn nữa ta có thể thấy tính mùa vụ của bộ dữ liệu lặp lại một cách rất đều đặn qua các năm.



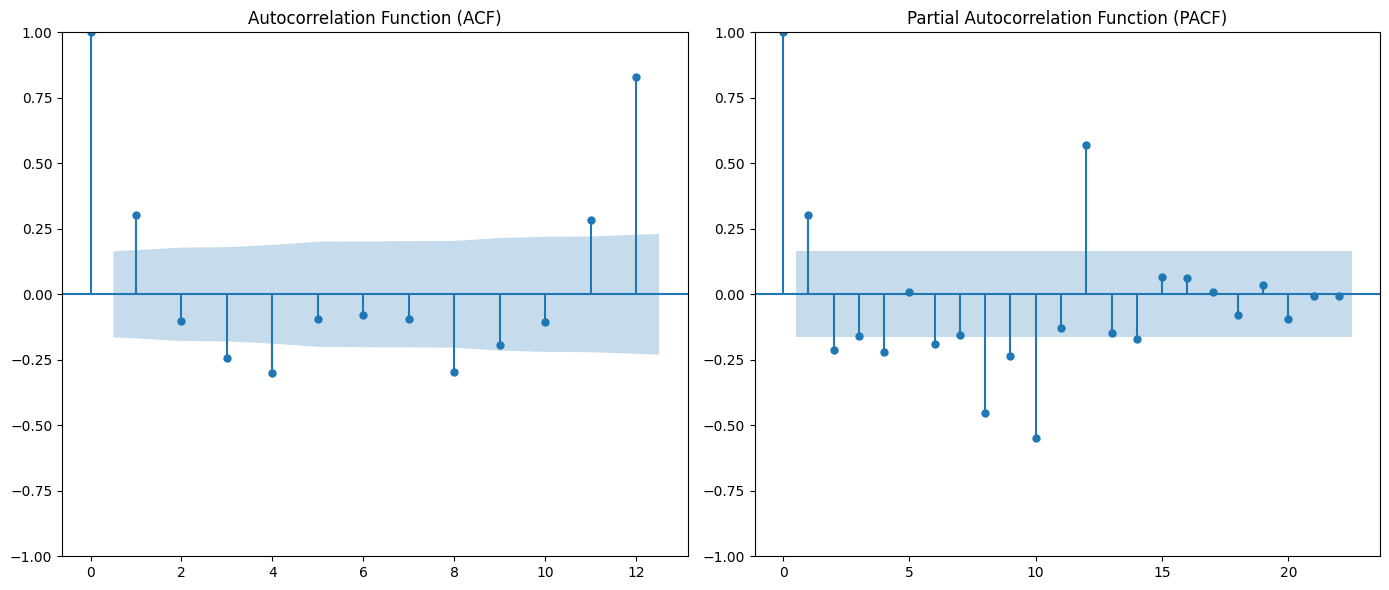
Hình 1: Trực quan dữ liệu trước khi xử lý

Tiếp theo, ta cần kiểm ta các đặc trưng của bộ dữ liệu xem có giá trị thiếu hay không và quan sát các chỉ số như: Trung bình, trung vị, độ lệch chuẩn, MinMax của của bộ dữ liệu. Qua kiểm tra ta có thể thấy được bộ dữ liệu này không có giá trị thiếu nên ta sẽ tiếp tục tới bước tiếp theo là kiểm tra xem bộ dữ liệu này có tỉnh dừng không. Vì chuỗi thời gian cần phải có tính dừng để các mô hình dự báo cổ điển như SARIMA, ARIMA hoạt động hiệu quả. Trong phần này ta sẽ dùng Augmented Dickey-Fuller (ADF) để kiểm tra tính dừng. Chuỗi dừng nghĩa là các đặc tính thông kê như (trung bình, phương sai, hiệp phương sai) không thay đổi theo thời gian, để đánh giá được tính dừng của bộ dữ liệu ta sẽ sử dụng giá trị p-value. Nếu giá trị p-value < 0.05 ta có thể kết luận là có chuỗi dừng và phù hợp với các mô hình dự báo cổ điển. Ngược lại, ta sẽ phải thực hiện sai phân để giúp chuỗi trở nên có tính dừng. Do tính được chỉ số p-value > 0.05 nên ta cần thực hiện sai phân. Trong Hình 2 là kết quả sau khi thực hiện.



Hình 2: Kết quả sau khi thực hiện sai phân

Như đã trình bày trên câu 1 ta sẽ thử tính chỉ số ACF để đánh giá bộ dữ liệu



Hình 3: Chỉ số ACF và PACF

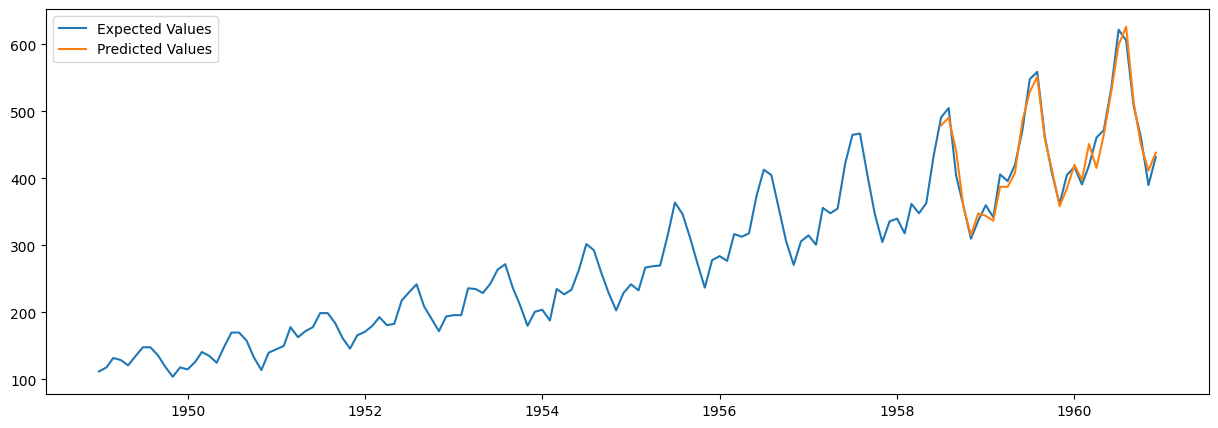
**C. Xây dựng và huấn luyện**

Trong bước này ta sẽ thực hiện huấn luyện với 2 mô hình là SARIMA và HOLT-WINTERS. Để huấn luyện ta sẽ cần chia bộ dữ liệu thành tập train và test, trong bài này ta sẽ chia theo tỉ lệ 7:3 với 70% dữ liệu cho tập train và 30% dữ liệu cho tập test.

train\_size = int(len(df)\*0.7)

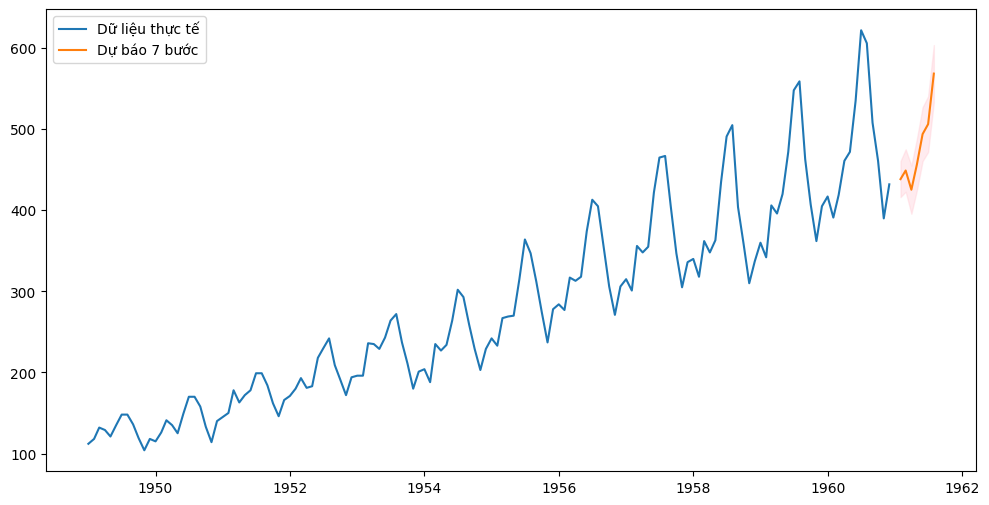
train, test = df[:train\_size], df[train\_size:]

Ta sẽ bắt đầu train với mô hình SARIMA và sau đây là kết quả thu được sau khi train và test với dữ liệu



Hình 4: Kết quả SARIMA

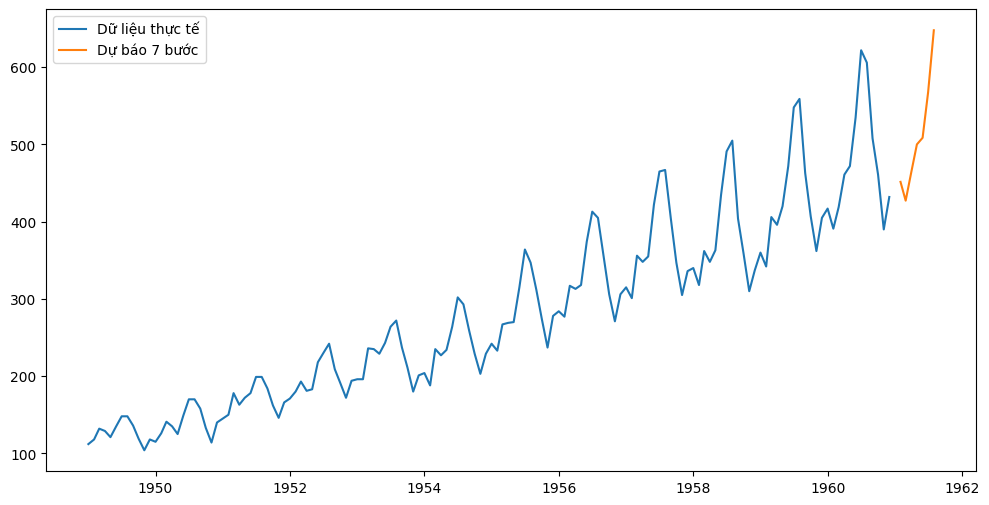
Ta có thể thấy mô hình đã dự báo bám sát được với xu hướng của bộ dữ liệu và với chỉ số R2 = 0.9521 thì khả năng dự đoán chính xác của mô hình là rất cao. Sau đây, ta sẽ thử cho mô hình dự báo 7 ngày trong tương lai để xem khả năng nắm bắt xu hướng của mô hình



Hình 5:SARIMA dự đoán 7 ngày

Qua biểu đồ ta thấy được mô hình đã dự đoán một cách chính xác và đã đưa ra được xu hướng của bộ dữ liệu.

Cũng tương tự như vậy ta sẽ thực hiện huấn luyện với holt-winters và cho dự đoán 7 ngày để đánh giá khả năng dự đoán của mô hình



Hình 6: Holt-winter dự đoán 7 ngày

Ta thấy cả 2 mô hình đã bám sát và dự đoán ra được xu hướng của bộ dữ liệu rất chính xác

**D. Đánh giá**

Bài toán dự báo số lượng cuộc gọi hàng ngày là một ví dụ điển hình của phân tích chuỗi thời gian với đặc điểm có xu hướng và tính mùa vụ rõ ràng. Qua quá trình khám phá dữ liệu, ta nhận thấy chuỗi số liệu thể hiện rõ ràng xu hướng tăng trưởng qua các năm.

Cả hai mô hình đều thể hiện khả năng dự báo tốt với dữ liệu hàng ngày, tuy nhiên mỗi mô hình có ưu điểm riêng phù hợp với từng hoàn cảnh cụ thể.

SARIMA có lợi thế khi dữ liệu có cấu trúc phức tạp, cần xử lý rõ các thành phần tự hồi quy và sai phân mùa vụ.

Holt-Winters thích hợp với các bài toán cần mô hình hóa nhanh, đơn giản, dễ hiểu và dễ điều chỉnh.