Foreign\_tourist

**1. Giới thiệu dữ liệu:**

Tệp CSV có 1895 dòng. Các cột chính bao gồm:

* **Năm**: Năm thu thập dữ liệu.
* **Tháng**: Tháng của năm.
* **Chỉ tiêu**: Chỉ tiêu khách du lịch (đường không, đường biển, đường bộ, theo quốc gia/vùng lãnh thổ).
* **Ước tính mỗi tháng**: Số lượng khách du lịch ước tính trong tháng.
* **Tổng lượt khách cả năm**: Tổng số lượt khách của năm.
* **So với tháng trước (%)**: Tỷ lệ phần trăm so với tháng trước.
* **So sánh tháng này ở năm trước (%)**: Tỷ lệ phần trăm so với cùng tháng năm trước.
* **So sánh tổng lượt khách với cùng kỳ năm trước (%)**: Tỷ lệ so sánh tổng lượt khách với cùng kỳ năm trước.

**2. Vấn đề tiền xử lý:**

* **Kiểm tra dữ liệu thiếu (missing values)**: Một số cột có giá trị thiếu, đặc biệt là "So sánh tháng này ở năm trước (%)" và "So sánh tổng lượt khách với cùng kỳ năm trước (%)".
* **Dữ liệu kiểu object**: Cần chuyển đổi các cột như "Ước tính mỗi tháng" và "Tổng lượt khách cả năm" sang kiểu số.
* **Kiểm tra outlier**: Đặc biệt với các cột tỷ lệ phần trăm (%), có thể có giá trị bất thường.
* **Biến dummy (nếu cần)**: "Chỉ tiêu" có thể được mã hóa để sử dụng trong mô hình dự báo.
* **Đa cộng tuyến**: Kiểm tra sự tương quan giữa các biến.

## ARIMA là gì?

## Mô hình ARIMA là gì?

* **ARIMA** (Autoregressive Integrated Moving Average) là một mô hình thống kê chuyên dùng trong phân tích chuỗi thời gian.
* Nó gồm 3 thành phần chính:
  1. **AR (Autoregressive)** – Hồi quy tự hồi: Chuỗi hiện tại phụ thuộc vào các giá trị trước đó.
  2. **I (Integrated)** – Sai phân: Ta sẽ lấy sai phân dữ liệu (tức tính mức thay đổi giữa các thời điểm) để biến chuỗi trở nên ổn định (stationary).
  3. **MA (Moving Average)** – Trung bình trượt: Chuỗi hiện tại phụ thuộc vào nhiễu (phần sai số) của những thời điểm trước.

Khi ghi ARIMA(p, d, q), ta hiểu:

* **p**: Bậc của AR (số giá trị trễ của chuỗi đầu vào).
* **d**: Số lần lấy sai phân (để biến dữ liệu thành dạng tĩnh, không còn xu hướng tăng/giảm quá mạnh).
* **q**: Bậc của MA (số giá trị trễ của nhiễu).

## 2. Tại sao dùng ARIMA để dự đoán số lượng khách du lịch?

* Trong chuỗi thời gian về du lịch (ví dụ: số du khách hàng tháng/năm), thường có **xu hướng** (trend) tăng/giảm theo mùa, theo năm.
* ARIMA sẽ “tự động” xử lý:
  + **Phần xu hướng**: nhờ bước tích hợp (d) là lấy sai phân, giúp chuỗi được ổn định hơn.
  + **Độ trễ theo thời gian**: thành phần AR cho phép mô hình “nhìn” về các kỳ trước.
  + **Tính ngẫu nhiên**: thành phần MA giúp hiệu chỉnh biến động ngẫu nhiên.

Ngoài ra, nếu dữ liệu có tính **mùa vụ mạnh** (ví dụ khách du lịch tăng đột biến vào mùa hè), thì mô hình ARIMA có thể được mở rộng thành **SARIMA** (Seasonal ARIMA) để xử lý đặc tính mùa vụ. Nhưng nhìn chung, nền tảng cơ bản là ARIMA.

ARIMA thường được dùng nhiều trong kinh tế, tài chính, và dự báo nhu cầu.

## 3. Quy trình để dự đoán số lượng khách du lịch bằng ARIMA

1. **Thu thập và làm sạch dữ liệu**
   * Thu thập số liệu khách du lịch trong quá khứ (theo tháng/năm).
   * Kiểm tra dữ liệu bị thiếu, ngoại lệ, đột biến...
2. **Khám phá và kiểm tra tính dừng của chuỗi**
   * Plot dữ liệu (hoặc thống kê) để xem xu hướng, tính mùa vụ.
   * Dùng các phép kiểm định (ADF test, KPSS test) để xem dữ liệu đã dừng chưa.
   * Nếu không dừng, ta **lấy sai phân** (1 lần, 2 lần...) đến khi nào chuỗi tương đối ổn định.
3. **Xác định p, d, q**
   * **d**: số lần sai phân cần thiết (xác định từ bước trên).
   * **p**: xem xét biểu đồ PACF (Partial Autocorrelation) để chọn bậc AR.
   * **q**: xem xét biểu đồ ACF (Autocorrelation) để chọn bậc MA.
   * Hoặc sử dụng tiêu chí AIC/BIC để so sánh nhiều bộ (p, d, q) và chọn mô hình tối ưu.
4. **Xây dựng mô hình và ước lượng tham số**
   * Sử dụng các thư viện như statsmodels (Python) hoặc các công cụ thống kê khác.
   * Mô hình sẽ tự ước lượng tham số AR, MA từ dữ liệu.
5. **Kiểm tra mô hình**
   * Kiểm tra phần dư (residuals) xem có còn tự tương quan (autocorrelation) hay không.
   * Nếu phần dư vẫn chưa là “white noise” (nhiễu trắng), ta tiếp tục chỉnh (p, d, q).
6. **Dự báo**
   * Khi mô hình ổn định, ta dự báo số lượng khách du lịch cho giai đoạn tiếp theo (tháng/năm tiếp theo).
   * Kết quả dự báo thường dưới dạng khoảng tin cậy (confidence interval), ví dụ: “Tháng 1 năm sau, dự đoán X khách ± sai số”.

## Lưu ý cá nhân khi dùng ARIMA

* **Ưu điểm:**
  + Mô hình tương đối đơn giản, dễ giải thích.
  + Thường cho kết quả dự báo tốt khi chuỗi có tính mùa vụ hoặc xu hướng không quá phức tạp (có thể mở rộng thành SARIMA, SARIMAX...).
* **Nhược điểm:**
  + Yêu cầu chuỗi thời gian cần có tính dừng, phải xử lý bằng sai phân. Nếu chuỗi có yếu tố phi tuyến (phi tuyến mạnh) hay mang tính mùa vụ phức tạp thì ARIMA nguyên bản đôi khi không đủ tốt.
  + Việc chọn p, d, q tối ưu có thể mất thời gian nếu dữ liệu lớn hoặc phức tạp.
* **Ý kiến cá nhân (chính kiến riêng):**
  + Theo mình, ARIMA phù hợp cho những dự án vừa và nhỏ, dữ liệu ít biến động đột biến (shock) hoặc không có yếu tố phi tuyến nặng. Với các chuỗi mang tính mùa vụ rõ (như doanh thu tháng, nhiệt độ tuần...), nên mở rộng sang SARIMA.

## Tóm tắt cho bài thuyết trình

* **Giới thiệu mô hình ARIMA**: Là công cụ truyền thống nhưng vô cùng “lâu đời” và hiệu quả để phân tích chuỗi thời gian.
* **Cấu trúc mô hình**: AR (hồi quy tự hồi), I (sai phân), MA (trung bình trượt).
* **Quy trình**: (1) Kiểm tra dữ liệu -> (2) Sai phân -> (3) Tìm p, d, q -> (4) Huấn luyện -> (5) Kiểm tra -> (6) Dự báo.
* **Ứng dụng cho khách du lịch**: Số liệu khách du lịch tháng/năm, xử lý xu hướng và mùa vụ, mô hình cho ra dự báo + khoảng tin cậy.
* **Kết luận**: ARIMA vẫn rất hiệu quả trong tình huống dữ liệu không quá phức tạp hoặc đã được làm sạch xu hướng và mùa vụ cẩn thận.

### ****Phân tích Kết Quả Mô Hình ARIMA (4,1,5):****

#### **1. Tổng quan mô hình:**

* **Mô hình được chọn:** ARIMA(4,1,5)
* **Số quan sát:** 1874
* **Thời gian huấn luyện:** 45.431 giây
* **Log Likelihood:** -11672.1
* **AIC (Akaike Information Criterion):** 23364.2 (Thấp là tốt)
* **BIC (Bayesian Information Criterion):** 23419.55
* **HQIC:** 23384.59

#### **2. Phân tích các tham số:**

| **Tham số** | **Giá trị (coef)** | **P-value** | **Giải thích** |
| --- | --- | --- | --- |
| **ar.L1** | -1.3507 | 0.000 | Thành phần tự hồi quy bậc 1 có ý nghĩa |
| **ar.L2** | -1.5930 | 0.000 | Thành phần tự hồi quy bậc 2 có ý nghĩa |
| **ar.L3** | -1.2621 | 0.000 | Thành phần tự hồi quy bậc 3 có ý nghĩa |
| **ar.L4** | -0.9065 | 0.000 | Thành phần tự hồi quy bậc 4 có ý nghĩa |
| **ma.L1** | 0.5036 | 0.000 | Thành phần trung bình trượt bậc 1 có ý nghĩa |
| **ma.L2** | 0.5913 | 0.000 | Thành phần trung bình trượt bậc 2 có ý nghĩa |
| **ma.L3** | 0.0749 | 0.004 | Thành phần trung bình trượt bậc 3 có ý nghĩa |
| **ma.L4** | -0.0080 | 0.709 | Không có ý nghĩa (P-value > 0.05) |
| **ma.L5** | -0.6979 | 0.000 | Thành phần trung bình trượt bậc 5 có ý nghĩa |

* **Sigma2:** 1.524e+04 – Phương sai của sai số (càng nhỏ càng tốt).

#### **3. Các kiểm định thống kê:**

| **Kiểm định** | **Giá trị** | **Ý nghĩa** |
| --- | --- | --- |
| **Ljung-Box (L1):** | 7.43 | Kiểm định phần dư, nếu nhỏ cho thấy phần dư là white noise (tốt). |
| **Prob(Q):** | 0.01 | Phần dư có thể chưa hoàn toàn là white noise. |
| **Heteroskedasticity (H):** | 0.91 | Dữ liệu không có phương sai thay đổi. |
| **Prob(H):** | 0.26 | Không có heteroskedasticity (giá trị này > 0.05 là tốt). |
| **Skew (Độ lệch):** | 1.62 | Dữ liệu hơi lệch phải. |
| **Kurtosis:** | 6.81 | Phân phối dữ liệu có đuôi dài (kurtosis > 3). |

### ****4. Đánh giá mô hình:****

* **P-value cho tất cả các thành phần ar (1 đến 4) và ma (1 đến 3, 5)** đều rất nhỏ (p < 0.05), cho thấy các tham số này có ý nghĩa.
* **Thành phần ma.L4 có p-value = 0.709 (không có ý nghĩa thống kê).** Có thể thử loại bỏ hoặc giảm bậc của MA để tinh chỉnh mô hình.
* **AIC/BIC thấp:** Mô hình đang khá tốt nhưng có thể thử nghiệm thêm các giá trị (p, d, q) để giảm AIC/BIC hơn nữa.

### ****5. Hướng cải thiện:****

1. **Kiểm tra lại phần dư (Residuals):**  
   Nếu phần dư không phải white noise, có thể thực hiện tuning lại tham số hoặc chuyển sang **SARIMA** nếu dữ liệu có tính mùa vụ.
2. **Loại bỏ tham số không ý nghĩa:**
   * Thử huấn luyện lại mô hình với ARIMA(4,1,4) (loại bỏ ma.L4).
3. model = ARIMA(data, order=(4,1,4))
4. result = model.fit()
5. print(result.summary())
6. **Tuning thêm:**
   * Dùng auto\_arima với phạm vi lớn hơn cho (p, d, q).
7. auto\_model = auto\_arima(data, seasonal=False, start\_p=0, start\_q=0,
8. max\_p=6, max\_q=6, trace=True)

### ****Kết luận:****

* Mô hình ARIMA(4,1,5) hoạt động khá tốt với dữ liệu của bạn.
* Có thể tối ưu thêm bằng cách loại bỏ các thành phần không ý nghĩa (ma.L4).
* Nếu AIC/BIC vẫn cao, có thể thử SARIMA để xử lý dữ liệu có tính mùa vụ.

**Phân tích biểu đồ dự báo ARIMA(4,1,5) và ARIMA(4,1,4):**

**1. Tổng quan về biểu đồ:**

* **Đường xanh dương:** Dữ liệu thực tế (quá khứ).
* **Đường cam:** Dự báo từ mô hình ARIMA(4,1,5).
* **Đường xanh lá:** Dự báo từ mô hình ARIMA(4,1,4).

**2. Những điều biểu đồ thể hiện:**

1. **Dữ liệu thực tế (Đường xanh dương):**
   * Có sự sụt giảm mạnh vào giai đoạn giữa (khoảng năm 2020-2021), có thể là do tác động của COVID-19 hoặc một yếu tố đặc biệt.
   * Sau năm 2022, dữ liệu tăng mạnh trở lại, cho thấy sự phục hồi.
   * Dữ liệu có tính chu kỳ rõ ràng với các đỉnh và đáy (tính mùa vụ hoặc dao động theo tháng).
2. **Dự báo ARIMA (Đường cam và xanh lá):**
   * **Cả hai mô hình ARIMA(4,1,5) và ARIMA(4,1,4)** cho thấy dự báo khá tương đồng, dao động nhẹ và ổn định trong khoảng thời gian dự báo (2024-2025).
   * **ARIMA(4,1,5)** có dự báo dao động nhiều hơn một chút so với ARIMA(4,1,4).

**3. Ý nghĩa của dự báo:**

* **Sự ổn định:**
  + Cả hai mô hình đều cho thấy sự ổn định trong tương lai gần (12 tháng tiếp theo).
  + Không có sự tăng trưởng hoặc sụt giảm mạnh nào được dự báo.
  + Dự báo dao động nhẹ, cho thấy sự phục hồi chậm nhưng ổn định.
* **Khả năng mô hình:**
  + Mô hình ARIMA đã bắt kịp xu hướng dữ liệu sau giai đoạn sụt giảm.
  + **Không có dấu hiệu dự báo sai lệch lớn**, chứng tỏ mô hình hoạt động tốt với dữ liệu.

**4. Kết luận:**

* **Dự báo ổn định, không có biến động lớn.**
* **ARIMA(4,1,5) và ARIMA(4,1,4)** cho kết quả tương đương, có thể chọn mô hình đơn giản hơn (ARIMA(4,1,4)) để tránh overfitting.
* Nếu mong muốn mô hình dự báo có xu hướng tăng/giảm rõ ràng hơn, có thể:
  + Kiểm tra và xử lý thêm tính mùa vụ (SARIMA).
  + Thử nghiệm các mô hình khác như **Prophet hoặc LSTM (deep learning)**.

Nếu bạn muốn đi sâu vào phân tích hoặc tinh chỉnh mô hình, mình sẵn sàng hỗ trợ thêm!