ỦY BAN NHÂN DÂN TP HỒ CHÍ MINH

**TRƯỜNG ĐẠI HỌC SÀI GÒN**

**KHOA CÔNG NGHỆ THÔNG TIN**



**BÁO CÁO ĐỒ ÁN**

ĐỀ TÀI: DỰ ĐOÁN LƯỢNG

TRUY CẬP TRANG WEB BẰNG MÔ HÌNH ARIMA

Giảng viên hướng dẫn: **Ts. Trịnh Tấn Đạt**

Sinh viên thực hiện: **Trần Ngọc Vũ**

MSSV: **3121410581**

**NIÊN KHÓA: 2021 - 2025**

**MỤC LỤC**

Lời cảm ơn

Giới thiệu đề tài

1. Tổng quan về đề tài
2. Mô tả bài toán

Chuỗi thời gian

Mô hình ARIMA

1. Lý thuyết
   1. Auto regression
   2. Moving average
   3. Intergated
2. Xây dựng mô hình
   1. Khảo sát tính dừng
   2. ACF và PACF
   3. Lựa chọn tham số cho mô hình ARIMA
3. Khảo sát
   1. Đọc hiểu kết quả mô hình ARIMA

Dự báo

Đánh giá

Tổng kết

**LỜI CẢM ƠN**

Chúng em xin chân thành cảm ơn Khoa Công nghệ thông tin, trường Đại học Sài Gòn đã tạo điều kiện cho chúng em được thực hiện đồ án môn học "Ngôn ngữ lập trình Python". Chúng em cũng xin chân thành cảm ơn thầy Trịnh Tấn Đạt, thầy đã giảng dạy cho chúng em những kiến thức cần thiết cho môn học này. Những kiến thức đó đã giúp cho chúng em rất nhiều trong quá trình làm đồ án báo cáo môn học. Chúng em xin chân thành cám ơn quý Thầy cô trong Khoa đã tận tình giảng dạy và trang bị cho chúng em những kiến thức cần thiết trong thời gian qua. Mặc dù nhóm đã cố gắng hoàn thành đồ án môn học với tất cả nổ lực của từng thành viên trong nhóm, nhưng đồ án chắc chắn không tránh khỏi những thiếu sót nhất định, rất mong nhận được sự cảm thông, chia sẻ và tận tình đóng góp chỉ bảo của quý Thầy Cô.

1. **MỞ ĐẦU**
2. **Lý do chọn đề tài:**

Khoa học dữ liệu là lĩnh vực nghiên cứu dữ liệu nhằm khai thác những thông tin chuyên sâu có ý nghĩa đối với hoạt động kinh doanh. Đây là một phương thức tiếp cận đa ngành, kết hợp những nguyên tắc và phương pháp thực hành của các lĩnh vực toán học, thống kê, trí tuệ nhân tạo và kỹ thuật máy tính để phân tích khối lượng lớn dữ liệu. Khoa học dữ liệu giúp các quản lý đưa ra những quyết định tốt hơn và nhanh hơn, dựa trên các bằng chứng vững chắc từ dữ liệu. Khoa học dữ liệu cũng trao quyền cho các nhà ra quyết định với khả năng dự đoán các xu hướng trong tương lai và vạch ra một kế hoạch nhằm đạt được các mục tiêu kinh doanh. Ngành khoa học dữ liệu có công dụng rộng rãi trong nhiều lĩnh vực khác nhau, từ thương mại điện tử, y tế, tài chính cho đến giáo dục và giải trí. Ngành khoa học dữ liệu đang phát triển mạnh mẽ và được săn đón trên thị trường lao động hiện nay.

Python được tạo ra bởi Guido van Rossum và phát hành lần đầu tiên vào năm 1991. Python là một ngôn ngữ lập trình bậc cao. Python là một ngôn ngữ lập trình phổ biến và linh hoạt, được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, như phát triển web, phân tích dữ liệu, trí tuệ nhân tạo và học máy. Python có cú pháp đơn giản và dễ đọc, hỗ trợ nhiều mô hình lập trình, và có một thư viện tiêu chuẩn phong phú. Python cũng có nhiều thư viện bên ngoài cho các mục đích cụ thể, như numpy, pandas, matplotlib và scikit-learn cho khoa học dữ liệu. Khoa học dữ liệu là một lĩnh vực nghiên cứu và ứng dụng các kỹ thuật thống kê, toán học và lập trình để khám phá, phân tích và trích xuất thông tin từ dữ liệu. Python là một trong những ngôn ngữ lập trình được ưa chuộng nhất trong khoa học dữ liệu, bởi vì nó có khả năng xử lý dữ liệu lớn và phức tạp, hỗ trợ các công cụ học máy và trực quan hóa dữ liệu, và có một cộng đồng lớn và tích cực.

Trong quá trình học tập và nghiên cứu, bản thân em rất có hứng thú với phân tích dữ liệu và dự đoán (forecasting). Bản thân em cũng có rất nhiều câu hỏi về khoa học dữ liệu như : liệu tương lai có thể dữ đoán được, hành vi người dùng liệu có thể mô tả được thông quá các bảng dữ liệu nhàm chán,... .Trong bối cảnh công nghệ thông tin phát triển như hiện nay, các trang web mọc lên như nấm sau mưa. Việc nắm bắt lượng truy cập người dùng là cân thiết để thực hiện các chiến lược marketing, sales,... vừa hiệu quả và ít tốn kém. Thế nên em quyết định xây dựng mô hình ARIMA để dự đoán lượng truy cập trong tương lai.

1. **Mục đích - Mục tiêu của đề tài:**

- Mục đích:

+ Có được các kĩ năng và kiến thức về Python

+ Tìm hiểu về 3 thư viện : numpy, pandas và matplotlib

+ Áp dụng kiến thức đã học

- Mục tiêu:

+ Làm quen với phân tích dữ liệu

+ Tiếp xúc với mô hình ARIMA

1. **Bố cục của đề tài:**

Nội dung gồm 2 phần:

Phần A: Mở đầu

Phần này bao gồm các thông tin về đề tài:

+ Tên đề tài

+ Mục đích

+ Lý do xây dựng đề tài

Phần B: Nội dung

+ Mô hình ARIMA

+ Xây dựng mô hình ARIMA

1. Chúng ta có một bộ dữ liệu chuỗi thời gian. Sau đó trải qua làm sạch và phân tích rồi áp dụng bộ dữ liệu vào mô hình ARIMA. Lựa chọn mô hình tốt nhất, đánh giá và đưa ra kết luận tổng quát cuối cùng.

**CHUỖI THỜI GIAN**

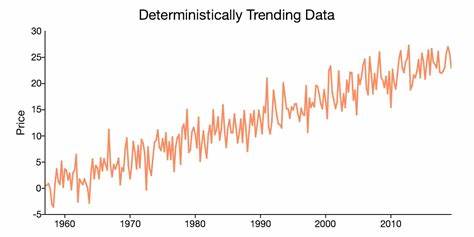
Chuỗi thời gian (time series) là một chuỗi các điểm dữ liệu được thu thập theo thứ tự thời gian. Mỗi điểm dữ liệu đại diện cho một giá trị của một đại lượng được quan sát tại một thời điểm cụ thể. Chuỗi thời gian được sử dụng rộng rãi trong nhiều lĩnh vực khác nhau, chẳng hạn như kinh tế, tài chính, khoa học khí tượng, y tế và nhiều lĩnh vực khác.

Một số đặc điểm của chuỗi thời gian bao gồm:

* Xu hướng (Trend): Chuỗi thời gian có 1 xu hướng tăng hoặc giảm dài hạn
* Mùa vụ (Seasonality): Chuỗi thời gian có thể có một mô hình lặp lại theo chu kỳ cố định, chẳng hạn như hàng năm, hàng quý hoặc hàng tháng.
* Độ biến động (Volatility): Chuỗi thời gian có thể có độ biến động cao hoặc thấp, tức là giá trị của chuỗi thay đổi nhiều hoặc ít theo thời gian.
* T**ự tương quan (Autocorrelation):** Các giá trị trong chuỗi thời gian có thể tương quan với các giá trị trước đó của chính nó.

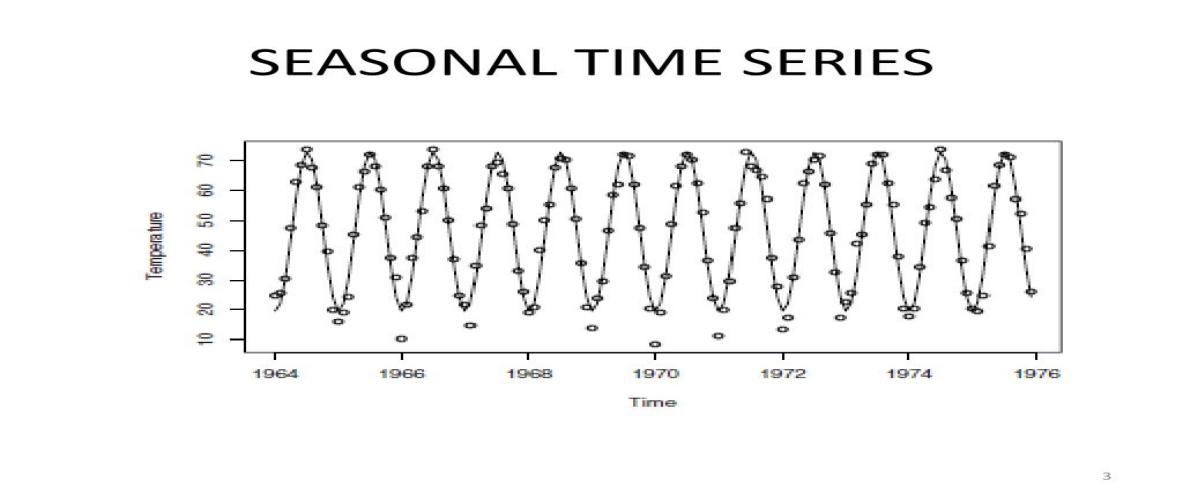
Các đặc điểm này có thể được phân tích và mô hình hóa để dự đoán các giá trị trong tương lai của chuỗi thời gian.

Chuỗi thời gian có xu hướng:



Ta có thể thấy chuỗi có đà tăng liên tục (có thể giảm) trong dài hạn

Chuỗi thời gian có tính mùa vụ:



Ta có thể thấy chuỗi thời gian lặp lại chu kỳ 1 năm. Cứ đạt đỉnh sau đó giảm rồi đạt đỉnh lần nữa sau 1 năm. Tương tự với đáy.

**MÔ HÌNH ARIMA**

1. Lý thuyết:

ARIMA là viết tắt của Autoregressive Intergrated Moving Average nghĩa là *tự hồi quy kết hợp trung bình động.* Đây là một mô hình phân tích thống kê sử dụng dữ liệu chuỗi thời gian để hiểu tốt hơn tập dữ liệu hoặc dự đoán xu hướng tương lai. Mô hình này có thể được hiểu bằng cách phân tích từng thành phần của nó như sau:

* 1. Auto regression:

Auto regression kí hiệu là AR. Đây là thành phần tự hồi quy dưa trên các giá trị bị trễ hoặc trước đó của nó. Độ trễ bậc ***p*** chính là giá trị lùi về quá khứ ***p*** bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ ***p***. Cụ thể, quá trình AR(***p***) của chuỗi được biểu diễn như bên dưới:

AR(p) =

* 1. Moving average
  2. Interagated

1. Xây dựng mô hình:
2. Khảo sát: