**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

**ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH**

**A picture containing clipart

Description generated with very high confidence**

**ĐỀ CƯƠNG**

**SỬ DỤNG NGÔN NGỮ PYTHON VÀO GIẢI QUYẾT CÁC BÀI TOÁN KINH TẾ**

* **GVHD: PGS.TS Nguyễn Đình Thuân**
* **SVTH :**

**15520987 – Trần Văn Tùng**

**15520708 – Lê Thị Đỗ Quyên**

**Tp.Hồ Chí Minh, Ngày 21 tháng 5 năm 2018**

**Mục Lục**

[I. Giới thiệu ngôn ngữ Python 3](#_Toc519021158)

[1. Giới thiệu: 3](#_Toc519021159)

[2. Cài đặt 4](#_Toc519021160)

[3. Sử dụng 5](#_Toc519021161)

[II. Biện Pháp Thống Kê Mô Tả Và Hiển Thị Dữ Liệu 8](#_Toc519021162)

[1. Dữ liệu dân số Việt Nam 1990 - 2018 8](#_Toc519021163)

[2. Tỷ Lệ Dân Thành Thị 2010 - 2016 12](#_Toc519021164)

[III. Suy Diễn Thống Kê 16](#_Toc519021165)

[1. Kiểm Định Trung Bình Một Mẫu 16](#_Toc519021166)

[a) Dữ liệu GDP Việt Nam 16](#_Toc519021167)

[b) Dữ liệu Huy Chương Vàng Việt Nam 17](#_Toc519021168)

[2. Kiểm Định Trung Bình Hai Mẫu 19](#_Toc519021169)

[a) Dữ Liệu Du Lịch Việt Nam 19](#_Toc519021170)

[3. Phân tích phương sai (Analysis of Variance) 21](#_Toc519021171)

[a) Tỷ Lệ Hộ Nghèo theo khu vực 21](#_Toc519021172)

[b) Thu Nhập Bình Quân Theo Ngành Kinh Tế 24](#_Toc519021173)

[4. Kiểm Định Chi -Square 26](#_Toc519021174)

[a) Bình Quân Thu Nhập Theo Địa Phương và Nguồn Thu 26](#_Toc519021175)

[b) Tỷ Lệ Thất Nghiệp Ở Thành Thị và Nông Thôn Theo Vùng 28](#_Toc519021176)

[IV. Phân Tích Hồi Quy Tuyến Tính 29](#_Toc519021177)

[1. Hồi Quy Tuyến Tính Đơn 29](#_Toc519021178)

[a) Dữ Liệu Giáo Dục Việt Nam 29](#_Toc519021179)

[b) Dữ liệu Diện Tích Sản Xuất Nông Nghiệp 31](#_Toc519021180)

[2. Hồi Quy Tuyến Tính Bội 33](#_Toc519021181)

[b) Dữ Liệu Thu Nhập 33](#_Toc519021182)

[V. Kỹ Thuật Dự Báo 35](#_Toc519021183)

[1. Dự báo với mô hình ARIMA 35](#_Toc519021184)

[a) Chỉ số đô la 35](#_Toc519021185)

[b) Dữ Liệu Việt Nam 37](#_Toc519021186)

[VI. Tài Liệu Tham Khảo 39](#_Toc519021187)

# Giới thiệu ngôn ngữ Python

## Giới thiệu:

* Python là một ngôn ngữ lập trình thông dịch do Guido van Rossum tạo ra năm 1990, nó được xem là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu, như nhận định của chính Guido van Rossum trong một bài phỏng vấn ông. Python hoàn toàn tạo kiểu động và dùng cơ chế cấp phát bộ nhớ tự động, do vậy nó tương tự như Perl, Ruby, Scheme, Smalltalk, và Tcl. Python được phát triển trong một dự án mã mở, do tổ chức phi lợi nhuận Python Software Foundation quản lý. Python là một ngôn ngữ lập trình thông dịch (interpreted), hướng đối tượng (object-oriented), và là một ngôn ngữ bậc cao (high-level)  ngữ nghĩa động (dynamic semantics).
* Ưu điểm :
  + Vừa hướng thủ tục (procedural-oriented), vừa hướng đối tượng (object-oriented).
  + Hỗ trợ module và hỗ trợ gói (package).
  + Xử lý lỗi bằng ngoại lệ (Exception).
  + Kiểu dữ liệu động ở mức cao.
  + Có các bộ thư viện chuẩn và các module ngoài, đáp ứng tất cả các nhu cầu lập trình.
  + Đơn giản: cú pháp đơn giản giúp cho người lập trình dễ học và tìm hiểu
  + Python có tốc độ xử lý nhanh.
  + Tương tác: chế độ tương tác cho phép người lập trình thử nghiệm tương tác sửa lỗi của các đoạn mã.
  + Chất lượng: thư viện có tiêu chuẩn cao, Python có khối cơ sở dữ kiệu khá lớn, nhằm cung cấp giao diện cho các cơ sở dữ liệu thương mại lớn. Đồng thời Python được biên dịch và chạy trên tất cả nền tảng lớn hiện nay.
  + Mở rộng: Python cho phép người lập trình có thể thêm hoặc tùy chỉnh các công cụ nhằm tối đa hiệu quả có thể đạt được trong công việc.
  + GUI programming: giúp cho việc thực hiện ảnh minh họa di động một cách tự nhiên sống động.
* Nhược điểm:
  + Python không có các thuộc tính như: protected, private hay public, không có vòng lặp do .. while, switch .. case .
  + Python mặc dù nhanh hơn PHP, nhưng không nhanh so với C++ và Java.
* Giới thiệu trình soạn thảo sử dụng cho báo cáo. Trong phạm vi phần báo cáo, chúng tôi sử dụng ngôn ngữ Python để tính hành các câu lênh, và trình soạn thảo Jupyter Notebook để hỗ trợ việc làm thực hiện các câu lệnh.

## Cài đặt

* Tiến hành cài đặt và sử dụng Python.
* Cài đặt Anacoda:

|  |
| --- |
| A screenshot of a cell phone  Description generated with very high confidence |

* Sau khi cài đặt thành công, tiến hành mở Anaconda Prompt:

|  |
| --- |
|  |

* Tiến hành khỏi động jupyter notebook:

|  |
| --- |
|  |

* Giao diện thao tác chính :

|  |
| --- |
|  |

## Sử dụng

* Tạo một file python mới:

|  |
| --- |
|  |

* Các thao tác:

|  |
| --- |
|  |

* + Chức năng File:

|  |
| --- |
|  |

* + Các chức năng Edit:

|  |
| --- |
|  |

* Thực hiện thao tác trên để chạy toàn bộ đoạn code:

|  |
| --- |
|  |

* Và còn nhiều chức năng khác.

# Biện Pháp Thống Kê Mô Tả Và Hiển Thị Dữ Liệu

## 1. Dữ liệu dân số Việt Nam 1990 - 2018

* Dữ liệu: Data\_DanSoVietNam.xlsx
* Code : Descriptive Statistical Measures\_DanSoVietNam.ipynb
* Phát biểu bài toán: dữ liệu đưa vào cung cấp số liệu về dân số của Việt Nam trong khoảng thời gian 1990 - 2017. Qua đó chúng tôi thực hiện các phép cơ bản để đưa ra nhận xét về dân số Việt Nam. Tiến hành tính các giá trị: Count, Min, Max, Mean, Median, Mode, Quantile, Range, Mode, Variance, Standard Deviation, Coefficient of Deviation, Skewness, Kurtosis (đơn vị Người).
* Đầu tiên tiến hành import các thư viện cần thiết:

|  |
| --- |
|  |

* Đọc file dữ liệu excel :

|  |
| --- |
|  |

* Các giá trị:

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

* Giải thích số liệu:
  + Max: cho biết giá trị lớn nhất của các quan sát.
  + Min: cho biết giá trị nhỏ nhất của các quan sát.
  + Median: cho biết giá trị trung vị của các quan sát.
  + Mean: cho biết giá trị trung bình của các quan sát.
  + Var (phương sai): dùng để đo lường mức độ phân tán của tập các giá trị quan sát
  + Std (độ lệch chuẩn): đo độ phân tán dữ liệu xung quanh giá trị trung bình của nó.
  + Skewness: cho biết dạng phân phối của các giá trị quan sát, với skew < 0: các giá trị quan sát sẽ tập trung chủ yếu vào các giá trị nhỏ nhất.
  + Kurtosis: đánh giá dỉnh của đừng cong quan sát với dạng phân phối chuẩn, với kur < 0 đường cong có dạng hẹp hơn hay tương đối bằng phẳng.
* Minh họa dữ liệu bằng biểu đồ:
* Biểu đồ đường:

|  |
| --- |
|  |

|  |
| --- |
|  |

*Biểu đồ đường thể hiện dân số Việt Nam 1990 – 2018*

* Biểu đồ Boxplot

|  |
| --- |
|  |

|  |
| --- |
|  |

*Biểu đồ boxplot thể hiện dân số Việt Nam 1990 – 2018*

* Nhận xét: qua các phép tính cơ bản ta thấy dân số Việt Nam tăng nhanh trong giai đoạn 1990 - 2018, và là nước có dân số cao so với diện tích trong khu vực và thế giới.

## 2. Tỷ Lệ Dân Thành Thị 2010 - 2016

- Dữ liệu: Data\_TiLeDanThanhThi.xlsx

**-** Code: Descriptive Statistical Measures\_TyLeDanThanhThi.ipynb

- Phát biểu bài toán: dữ liệu đưa vào cung cấp số liệu về tỷ lệ dân thành thị của Việt Nam trong khoảng thời gian 2010 - 2016 . Qua đó chúng tôi thực hiện các phép cơ bản để đưa ra nhận xét về dân số Việt Nam. Tiến hành tính các giá trị: Count, Min, Max, Mean, Median, Mode, Quantile, Range, Mode, Variance, Standard Deviation, Coefficient of Deviation, Skewness, Kurtosis (đơn vị Người).

* Đầu tiên tiến hành import các thư viện cần thiết:

|  |
| --- |
|  |

* Đọc file dữ liệu excel :

|  |
| --- |
|  |

* Các giá trị:

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

* Giải thích số liệu:
  + Max: cho biết giá trị lớn nhất của các quan sát.
  + Min: cho biết giá trị nhỏ nhất của các quan sát.
  + Median: cho biết giá trị trung vị của các quan sát.
  + Mean: cho biết giá trị trung bình của các quan sát.
  + Var (phương sai): dùng để đo lường mức độ phân tán của tập các giá trị quan sát
  + Std (độ lệch chuẩn): đo độ phân tán dữ liệu xung quanh giá trị trung bình của nó.
  + Skewness: cho biết dạng phân phối của các giá trị quan sát, với skew < 0: các giá trị quan sát sẽ tập trung chủ yếu vào các giá trị nhỏ nhất.
  + Kurtosis: đánh giá dỉnh của đừng cong quan sát với dạng phân phối chuẩn, với kur < 0 đường cong có dạng hẹp hơn hay tương đối bằng phẳng.
* Minh họa dữ liệu bằng biểu đồ:
* Biểu đồ đường:

|  |
| --- |
|  |

|  |
| --- |
|  |

*Biểu đồ đường thể hiện Tỷ lệ dân thành thị Việt Nam 2010 - 2016*

* Biểu đồ Boxplot

|  |
| --- |
|  |

|  |
| --- |
|  |

*Biểu đồ boxplot thể hiện Tỷ lệ dân thành thị Việt Nam 2010 – 2016*

* Nhận xét: Thông qua các số liệu đã tính toán ở trên, và dữ liệu đã được hiển thị ở trên, ta thấy tỷ lệ dân thành thị ở Việt Nam còn thấp. Tuy nhiên đã tăng mạnh trong những năm qua, đặt biệt tăng vọt từ năm 2013.

# Suy Diễn Thống Kê

* Suy diễn thống kê là phương pháp dùng để thống kê dữ liệu, kiểm định dữ liệu có đúng với giả thuyết hay không. Suy diễn thống kê gồm có 4 loại: So sánh trung bình của một tổng thể với một giá trị cụ thể (One – Sample Hypoothesis Test), (), Phân tích phương sai (Anova), kiểm định Chi -Square xét sự độc lập của hai nhóm tổng thể.

## Kiểm Định Trung Bình Một Mẫu

### Dữ liệu GDP Việt Nam

* Dữ liệu: Data\_GDPVietNam.csv
* Code : One\_Sample T-Test\_GDPVietNam.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp thu nhập bình quân đầu người trên năm (GDP) của Việt Nam từ 1994 - 2017 (đơn vị USD).
* Phát biểu bài toán, chúng tôi đưa giả giả thuyết như sau:
  + H0 : “Giá trị trung bình GDP của Việt Nam trong khoảng 1994 – 2017 bằng 2000(USD)”
  + H1: “Giá trị trung bình GDP của Việt Nam trong khoảng 1994 – 2017 khác 2000 (USD)”
* Đầu tiên import thư viện vào đọc giữ liệu:

|  |
| --- |
|  |

* Đọc dữ liệu:

|  |
| --- |
|  |

* Tiếp theo ta tính one-sample test:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận: pvalue: 4.42e-06 < 0.05 , nên ta từ chối giả thuyết H0 . Từ đó ta kết luận thu nhập bình quân người trên năm của Việt Nam 1990 -2017 khác 2000(USD)
* Minh họa dữ liệu bằng đồ thị:

|  |
| --- |
|  |

*Biểu đồ GDP Việt Nam 1990 – 2017*

* Nhận xét : dựa vào đồ thị ta thấy thu nhập bình quân của người Việt Nam còn khá thấp so với mặt bằng chung, vì Việt Nam là nước đang phát triển, nhưng GDP chưa cao. Thu nhập bình quân tăng mạnh thừ năm 2005 đến nay, cho thấy Việt Nam đang trên đà phát triển.

### Dữ liệu Huy Chương Vàng Việt Nam

* Dữ liệu: Data\_HuyChuongVang.csv
* Code : One\_Sample T-Test\_HuyChuongVangVietNam.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp số huy chương vàng cấp thế giới mà Việt Nam đã đạt được trong 2002 – 2015.
* Phát biểu bài toán, chúng tôi đưa giả giả thuyết như sau:
  + H0 : “Giá trị trung bình số huy chương vàng Việt Nam đạt được trong khoảng 2002– 2015 bằng 40”
  + H1: “Giá trị trung bình số huy chương vàng Việt Nam đạt được trong khoảng 2002– 2015 khác 40”
* Đầu tiên import thư viện vào đọc giữ liệu:

|  |
| --- |
|  |

* Đọc dữ liệu:

|  |
| --- |
|  |

* Tiếp theo ta tính one-sample test:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận: pvalue: 0.03 < 0.05 , nên ta từ chối giả thuyết H0 . Từ đó ta kết luận giá trị trung bình số huy chương vàng cấp thế giới mà Việt Nam đạt được từ 2002 - 2015 khác 40 huy chương vàng.
* Minh họa dữ liệu bằng đồ thị:

|  |
| --- |
|  |

*Biểu đồ số huy chương vàng Thế giới Việt Năm 2002- 2015*

* Nhận xét : dựa vào đồ thị số huy chương vàng thế giới đạt được của Việt Nam biến động. Và giá trị trung bình số huy chương thế giới của Việt Nam còn thấp.

## Kiểm Định Trung Bình Hai Mẫu

### Dữ Liệu Du Lịch Việt Nam

* Được thực hiện trên tập dữ liệu : Data\_DuLich\_TwoSample.xlsx
* Code: Two\_Sample\_Final-DuLich.ipynb
* Giả thuyết:
  + H0: Giá trị trung bình của số lượt khách quốc tế từ năm 2015 đến tháng 4/2018 đến tham quan Việt Nam bằng đường hàng không và đường biển **không có** sự chênh lệch nhau nhiều.
  + H1: Giá trị trung bình của số lượt khách quốc tế đến tham quan Việt Nam bằng đường hàng không và đường biển **có** sự chênh lệch nhau nhiều..
* Đầu tiên ta thêm thư viện và đọc file dữ liệu:

|  |
| --- |
|  |

* Tính two-sample:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận: p-value: 1.243530517018681e-13 < 0.05 , nên từ chối chấp nhận giả thuyết H0. Giá trị trung bình của số lượt khách quốc tế đến tham quan Việt Nam bằng đường hàng không và đường biển có sự chênh lệch nhau nhiều.

## Phân tích phương sai (Analysis of Variance)

### **Tỷ Lệ Hộ Nghèo theo khu vực**

* Dữ liệu: Data\_TyLeHoNgheo\_Anova.csv
* Code: Anova\_TyLeHoNgheo.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp tỷ lệ hộ nghèo của 4 khu vực : Bắc trung bộ và duyên hải miền trung, tây nguyên, đông nam bộ, đồng bằng sông cửu long trong giai đoạn 1998 – 2016. Tỷ lệ hộ nghèo được tính theo thu nhập bình quân 1 người 1 tháng của hộ gia đình theo chuẩn nghèo của Chính phủ giai đoạn 2011-2016 được cập nhật theo chỉ số giá tiêu dùng như sau:2010: 400 nghìn đồng đối với khu vực nông thôn và 500 nghìn đồng đối với khu vực thành thị.; 2012: 530 nghìn đồng đối với khu vực nông thôn và 660 nghìn đồng đối với khu vực thành thị.;2013: 570 nghìn đồng đối với khu vực nông thôn và 710 nghìn đồng đối với khu vực thành thị.; 2014: 605 nghìn đồng đối với khu vực nông thôn và 750 nghìn đồng đối với khu vực thành thị.; 2015: 615 nghìn đồng đối với khu vực nông thôn và 760 nghìn đồng đối với khu vực thành thị.; 2016: 630 nghìn đồng đối với khu vực nông thôn và 780 nghìn đồng đối với khu vực thành thị.
* Phát biểu bài toán: tiến hành tính toán và xem xét có sự khác biệt về tỷ lệ hộ nghèo giữa khác khu vực, và đưa ra sự khác biệt như thế nào của mỗi nhóm. Đưa ra giả thuyết:
  + H0 : Không có sự khác biệt về tỷ lệ hộ nghèo trên từng nhóm khu vực.
  + H1: Có sự khác biệt về về tỷ lệ hộ nghèo trên từng nhóm khu vực.
* Tiến hành import các thư viện và thực hiện các câu lệnh theo hình dưới đây:

|  |
| --- |
|  |

* Đọc dữ liệu:

|  |
| --- |
|  |

* Minh họa dữ liệu:

|  |
| --- |
|  |

|  |
| --- |
|  |

* Phân tích phương sai Anova:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận : ta thấy pValue = 0.000022 < 0.05 (mức ý nghĩa), nên ta bác bỏ giả thuyết H0. Suy ra, có sự khác biệt về tỷ lệ hộ nghèo giữa các nhóm vùng với nhau.
* Tiến hành kiểm tra sự khác biệt của mỗi nhóm với nhau:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận :
  + Ta thấy quan hệ giữa nhóm “Bac Trung Bo Va Duyen Hai Mien Trung ” với nhóm “Dong Nam Bo” có khoảng tin cậy lower và upper đề nhỏ hơn 0, nên ta có thể kết luận sự khác biệt của hai nhóm này có ý nghĩa thống kê.
  + Tương tự như “Dong Nam Bo ” và “Tay Nguyen” có khoảng tin cậy lower và upper đều lớn hơn 0 nên sự khác biệt của hai nhóm này có ý nghĩa thống kê.
  + Còn các nhóm còn lại không có khoảng tin cậy lower và upper đều nhỏ hơn 0, hoặc đều lớn hơn 0 nên không thể kết luận chúng khác biệt nhau.

### **Thu Nhập Bình Quân Theo Ngành Kinh Tế**

* Dữ liệu: Data\_ThuNhapTheoNganh\_Anovaa.csv
* Code: Anova\_ThuNhapTheoNganh.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp thu nhập bình quân của lao động làm công ăn lương trong khu vực nhà nước theo 4 khu vực : thông tin và truyền thông, giáo dục và đào tạo, y tế, vui chơi giải trí 2005 - 2015. Đưa ra giả thuyết:
  + H0 : Không có sự khác biệt về thu nhập bình quân của lao động làm công ăn lươn của 4 ngành.
  + H1: Có sự khác biệt về thu nhập bình quân của lao động làm công ăn lươn của 4 ngành.
* Tiến hành import các thư viện và thực hiện các câu lệnh theo hình dưới đây:

|  |
| --- |
|  |

* Đọc dữ liệu:

|  |
| --- |
|  |

* Minh họa dữ liệu:

|  |
| --- |
|  |

|  |
| --- |
|  |

* Phân tích phương sai Anova:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận : ta thấy pValue = 0.01 < 0.05 (mức ý nghĩa), nên ta bác bỏ giả thuyết H0. Suy ra, có sự khác biệt về bình quân thu nhập của các lao động làm công ăn lương nhà nước theo 4 ngành ở trên.
* Tiến hành kiểm tra sự khác biệt của mỗi nhóm với nhau:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận :
  + Ta thấy quan hệ giữa nhóm “GiaoDucVaDaoTao” với nhóm “ThongTinTruyenThong” có khoảng tin cậy lower và upper đều lớn hơn 0, nên ta có thể kết luận sự khác biệt của hai nhóm này có ý nghĩa thống kê.
  + Tương tự như “ThongTinTruyenThong” và “VuiChoiGiaiTri” có khoảng tin cậy lower và upper đều nhỏ hơn 0 nên sự khác biệt của hai nhóm này có ý nghĩa thống kê.
  + Tương tự như “ThongTinTruyenThong” và “Yte” có khoảng tin cậy lower và upper đều nhỏ hơn 0 nên sự khác biệt của hai nhóm này có ý nghĩa thống kê.
  + Còn các nhóm còn lại không có khoảng tin cậy lower và upper đều nhỏ hơn 0, hoặc đều lớn hơn 0 nên không thể kết luận chúng khác biệt nhau.

## Kiểm Định Chi -Square

### Bình Quân Thu Nhập Theo Địa Phương và Nguồn Thu

* Dữ Liệu: Data\_ThuNhapThangNam2016\_ChiSquare.csv
* Code: Chi-Quare\_ThuNhapThang.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp hai nhóm tổng thể mà ta cần xét hai nhóm có độc lập với nhau hay không. Trong dữ liệu này có hai nhóm đó là địa phương và nguồn thu.
* Phát biểu bài toán: xét sự độc lập của hai nhóm địa phương và nguồn thu. Xét xem chúng độc lập hay phụ thuộc nhau. Ta có giả thuyết:
  + H0 : Hai nhóm địa phương và nguồn thu độc lập với nhau.
  + H1: Hai nhóm địa phương và nguồn thu phụ thuộc với nhau.
* Chúng ta tiến hành thêm thư viện :

|  |
| --- |
|  |

* Chuyển dữ liệu từ file excel dưới dạng ma trận trong Python, việc này được thực hiện thủ công:

|  |
| --- |
|  |

* Thực hiện Chi Square:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận : giá trị Chi-Square : 323.89, pvalue = 7.55e-69 < 0.05 (mức ý nghĩa). Cho nên ta bác bỏ giả thuyết H0. Và điều đó cũng cho biết rằng giá trị nguồn thu (thu nhập ) của người dân mỗi vùng khác nhau sẽ phụ thuộc theo từng vùng mà họ sinh sống.

### Tỷ Lệ Thất Nghiệp Ở Thành Thị và Nông Thôn Theo Vùng

* Dữ Liệu: Data\_TyLeThatNghiepTTNT\_ChiSquare.csv
* Code: Chi-Quare\_TyLeThatNghiepTheoVung.ipynb
* Mô tả dữ liệu: dữ liệu cung cấp hai nhóm tổng thể mà ta cần xét hai nhóm có độc lập với nhau hay không. Trong dữ liệu này có hai nhóm đó là Vùng và Loại Đô Thị.
* Phát biểu bài toán: xét sự độc lập của hai nhóm vùng và loại đô thị. Xét xem chúng độc lập hay phụ thuộc nhau. Ta có giả thuyết:
  + H0 : Hai nhóm vùng và loại đô thị độc lập với nhau.
  + H1: Hai nhóm vùng và loại đô thị phụ thuộc với nhau.
* Chúng ta tiến hành thêm thư viện :

|  |
| --- |
|  |

* Chuyển dữ liệu từ file excel dưới dạng ma trận trong Python, việc này được thực hiện thủ công:

|  |
| --- |
|  |

* Thực hiện Chi Square:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận : giá trị Chi-Square : 1.12, pvalue = 0.95 > 0.05 (mức ý nghĩa). Cho nên ta chấp nhận giả thuyết H0. Và điều đó cũng cho biết rằng tỷ lệ thất nghiệp theo loại đô thị sẽ độc lập với từng vùng miền.

# Phân Tích Hồi Quy Tuyến Tính

## Hồi Quy Tuyến Tính Đơn

### Dữ Liệu Giáo Dục Việt Nam

* Dữ liệu: Data \_GiaoDuc\_SimpleLinear.xlsx
* Code: [Simple Linear Regression \_ GiaoDuc.ipynb](https://github.com/tranvantung241/BA-with-Python/blob/master/CodePython/Simple%20Linear%20Regression%20_%20Home%20Market%20Value.ipynb)
* Giả thuyết:
  + H0 = “Phương trình tìm được không có ý nghĩa, trình độ giảng dạy của giáo viên thường không quyết định đầu ra/số lượng tốt nghiệp của sinh viên.”
  + H1 = “Phương trình tìm được có ý nghĩa, trình độ giảng dạy của giáo viên thường quyết định đầu ra/số lượng tốt nghiệp của sinh viên.”
* Tiến hành thêm thư viện vào đọc file dữ liệu :

|  |
| --- |
|  |

|  |
| --- |
|  |

* Tiến hành các phép tính tính hồi quy tuyến tính, và tìm ra phương trình có nghĩa:

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

* Kết luận:
  + Phương trình tìm được là: Y = -10.4141+ 0.1769 b1
  + Ta thấy p-value: 2.46e-06< 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được và trình độ giảng dạy của giáo viên thường quyết định đầu ra/số lượng tốt nghiệp của sinh viên.

### b) Dữ liệu Diện Tích Sản Xuất Nông Nghiệp

* Tập dữ liệu: Data\_DLSX\_SimpleLinear.csv
* Code: Linear Regression\_DLSX.ipynb
* Giả thuyết:
  + H0 = “Phương trình tìm được không có ý nghĩa, và Tổng sản lượng lúa thu được không phụ thuộc vào Tổng diện tích đất trồng.”
  + H1 = “Phương trình tìm được có ý nghĩa, Tổng sản lượng lúa thu được không phụ thuộc vào Tổng diện tích đất trồng.”
* Tiến hành thêm thư viện vào đọc file dữ liệu :

|  |
| --- |
|  |

|  |
| --- |
|  |

* Tiến hành các phép tính tính hồi quy tuyến tính, và tìm ra phương trình có nghĩa:

|  |
| --- |
|  |

-Kết Quả:

|  |
| --- |
|  |

* Kết luận:
  + Phương trình tìm được là: Y = -6.676e+04 + 12.6977 b1
  + Ta thấy p-value: 1.02e-15< 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được và Tổng sản lượng lúa thu được phụ thuộc vào Tổng diện tích đất trồng

## Hồi Quy Tuyến Tính Bội

### Dữ Liệu Thu Nhập

* Dữ liệu: Data\_DienTich\_MultipleLinear.xlsx
* Code: Multiple Linear Regression-DienTichNha. .ipynb
* Giả thuyết:
  + H0 = “Phương trình tìm được không có ý nghĩa, và **Diện tích nhà ở** không phụ thuộc vào **Thu nhập bình quân đầu người,năng suất lao động, giới tính”**
  + H1 = “Phương trình tìm được có ý nghĩa, và **Diện tích nhà ở**  phụ thuộc vào **Thu nhập bình quân đầu người,năng suất lao động, giới tính”**
* Thêm thư viện và đọc file dữ liệu:

|  |
| --- |
|  |

* Đọc dữ liệu:

|  |
| --- |
|  |

|  |
| --- |
|  |

* Tiến hành tính Multiple Linear Regression

|  |
| --- |
|  |

* Kết quả:

|  |
| --- |
|  |

|  |
| --- |
|  |

* **Kết luận:** 
  + Phương trình tìm được là: Y = -94.9766 + 0.0070\*b1 – 0.0675\*b2 +3.3333\* b3 – 1.7891\*b4
  + Ta thấy p-value: 0.000106 < 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được có ý nghĩa, và **Diện tích nhà ở** phụ thuộc vào **Thu nhập bình quân đầu người,năng suất lao động, giới tính**

# Kỹ Thuật Dự Báo

## Dự báo với mô hình ARIMA

### Chỉ số đô la

* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : Data\_ChiSoDoLa.csv
* Code:  [Forecast\_Arima\_ChiSoDoLa.ipynb](https://github.com/tranvantung241/BA-with-Python/blob/master/CodePython/Simple%20linear%20regression_Coal%20Production.ipynb)
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:

|  |
| --- |
|  |

* Đọc file dữ liệu:

|  |
| --- |
|  |

* Dựa vào giá trị lịch sử trong tập dữ liệu, ta khởi tạo mô hình.Xét dự báo cho giai đoạn giá mua vàng trong tương lai là tháng 1 của năm 2018

|  |
| --- |
|  |

-Và chúng ta sẽ được kết quả dự báo tháng 1 của năm 2018

|  |
| --- |
|  |

|  |
| --- |
|  |

-Từ kết quả ta thấy chỉ số đô la trong tháng 1 của năm 2018 sẽ tăng trưởng với tốc độ chậm lại so với năm 2017 và chỉ số đô la sẽ giảm đi so với năm 2017

### b) Dữ Liệu Việt Nam

* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : Data\_GiaVangSJC.csv
* Code: [Forecast\_Arima\_GiaVangSJC.ipynb](https://github.com/tranvantung241/BA-with-Python/blob/master/CodePython/Simple%20linear%20regression_ChiSoDoLa-Copy2.ipynb)
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:

|  |
| --- |
|  |

* Đọc file dữ liệu: (*GiaVangSJC*= pd.read\_csv(“/path”)):

|  |
| --- |
|  |

* Dựa vào giá trị lịch sử trong tập dữ liệu, ta khởi tạo mô hình.Xét dự báo cho giai đoạn giá mua vàng trong tương lai là ngày 7/7/2018 của năm 2018

|  |
| --- |
|  |

-Bây giờ mô hình của chúng ta được khởi tạo, chúng tôi có thể đánh giá nó. Và chúng ta sẽ được kết quả dự báo của ngày 7/7/2018

|  |
| --- |
|  |

|  |
| --- |
|  |

* Từ kết quả ta thấy giá vàng trong ngày 7/7 của năm 2018 sẽ giảm với tốc độ chậm lại so với năm 2017 và giá vàng sẽ giảm đi so với ngày 6/7/2018.

# Tài Liệu Tham Khảo

* Giáo trình môn Phân Tích Dữ Liệu Kinh Doanh.
* Thư viện hỗ trợ trong Python : https://docs.python.org/3/index.html