**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](#_Toc514630774)

[1. Giới thiệu: 3](#_Toc514630775)

[2. Cài đặt 4](#_Toc514630776)

[3. Sử dụng 6](#_Toc514630777)

[II. Data Visualization And Descriptive Statistical Measures 8](#_Toc514630778)

[1.Tính giá trị. 8](#_Toc514630779)

[- a. Purchase Order.xlsx 8](#_Toc514630780)

[b.Dữ liệu Việt Nam: 10](#_Toc514630781)

[2.Vẽ Histogram, Box Plot 12](#_Toc514630782)

[a. Purchase Order.xlsx 12](#_Toc514630783)

[b. Dữ liệu Việt Nam: 13](#_Toc514630784)

[III. Statistical Inference 14](#_Toc514630785)

[1. One-Sample Hypothesis Tests 15](#_Toc514630786)

[a. Vacation Survey1.xlsx 15](#_Toc514630787)

[b. DuLieuVietNamVeDanSoNam2010-2016.csv 15](#_Toc514630788)

[2. Two-Sample Hypothesis Tests 16](#_Toc514630789)

[3. Analysis of Variance (ANOVA) 18](#_Toc514630790)

[a. Insurance Survey1.xlsx 18](#_Toc514630791)

[b. Dân số Việt Nam từ 2000-2016 19](#_Toc514630792)

[4. Chi-Square Test for Independence 19](#_Toc514630793)

[a. Insurance Survey1.xlsx 19](#_Toc514630794)

[b.DuLieuVietNamVeDanSoNam2010-2016 20](#_Toc514630795)

[IV. Regression Analysis 21](#_Toc514630796)

[1. Simple Linear Regression 21](#_Toc514630797)

[a. Home Market Value.xlsx 21](#_Toc514630798)

[b. DLSX.csv 23](#_Toc514630799)

[2. Multiple Linear Regression 25](#_Toc514630800)

[V. Forecasting Techniques 28](#_Toc514630801)

[1. Statistical Forecasting Models 28](#_Toc514630802)

[- a. Tablet Computer Sales.xlsx 28](#_Toc514630803)

[b.Dữ liệu Việt Nam 30](#_Toc514630804)

[2. Regression Forecasting with Causal Variables 31](#_Toc514630805)

[a. Coal Production.xlsx 31](#_Toc514630806)

[b. Dữ Liệu Việt Nam.csv 33](#_Toc514630807)

# Giới thiệu ngôn ngữ Python

## Giới thiệu:

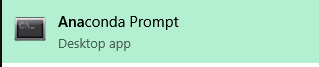
* Python là một ngôn ngữ lập trình thông dịch do Guido van Rossum tạo ra năm 1990, nó được xem là ngôn ngữ có hình thức rất sáng sủa, cấu trúc rõ ràng, thuận tiện cho người mới học lập trình. Cấu trúc của Python còn cho phép người sử dụng viết mã lệnh với số lần gõ phím tối thiểu, như nhận định của chính 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 dàng đọc và tìm hiểu.
  + Tốc độ**: 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 liệu khá lớn nh ằm cung cấp giao diện cho tất cả các CSDL thương mại lớn.
  + Thuận tiện**: Python** được biên dịch và chạy trên tất cả các nền tảng lớn hiện nay.
  + **Mở rộng**: Với tính năng này, 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 hoạ di động một cách tự nhiên và sống động.
* **Nhược điểm:**
* Python không có các thuộc tính như :protected,private hay public, không có vòng lặp do…while và switch….case.
* Python mặc dù nhanh hơn so với PHP, nhưng lại không nhanh hơn so với C++, Java.

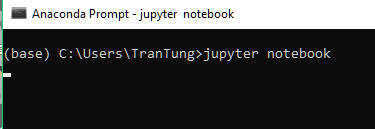
## Cài đặt

* Tiến hành cài đặt và sử dụng Python
* Cài đặt Anacoda [download](https://www.anaconda.com/download/)

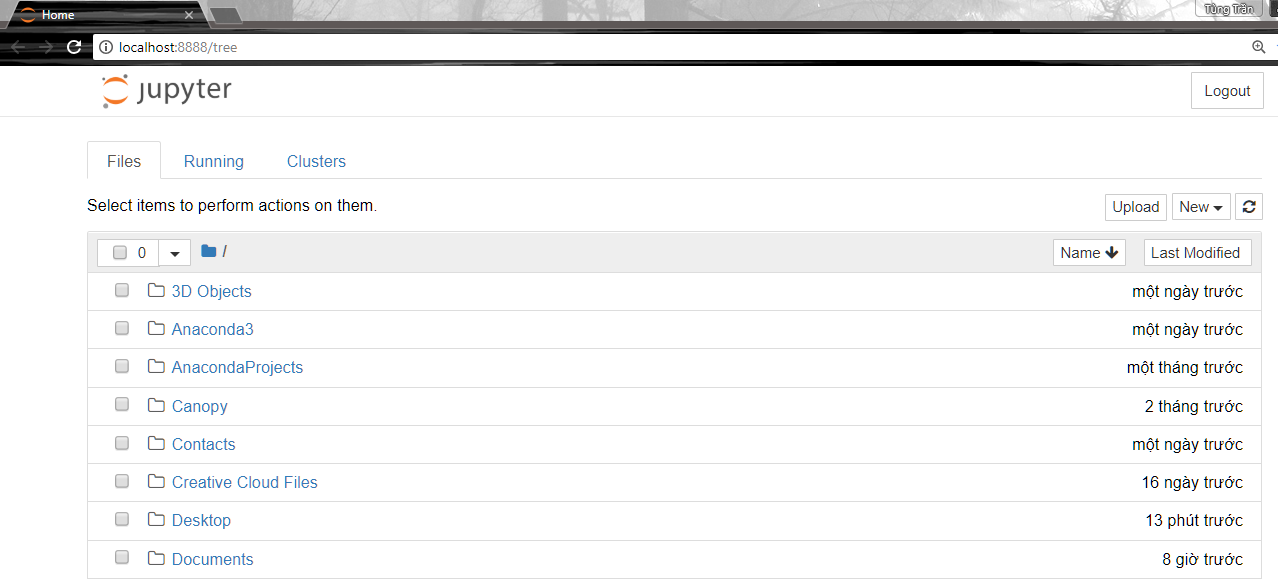


* Sau khi cài đặt thành công, tiến hành mở Anaconda Prompt



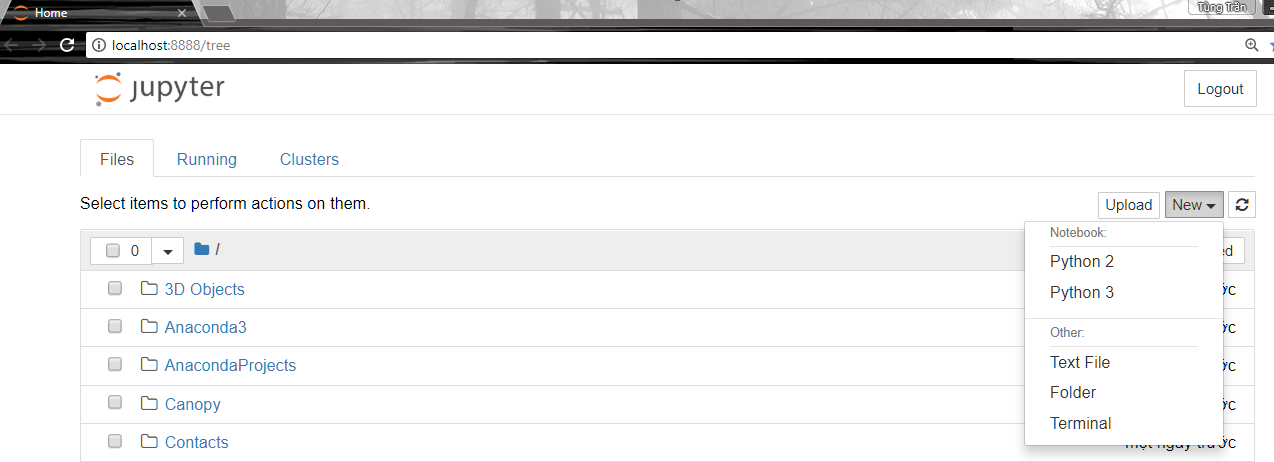


* Giao diện thao tác chính :

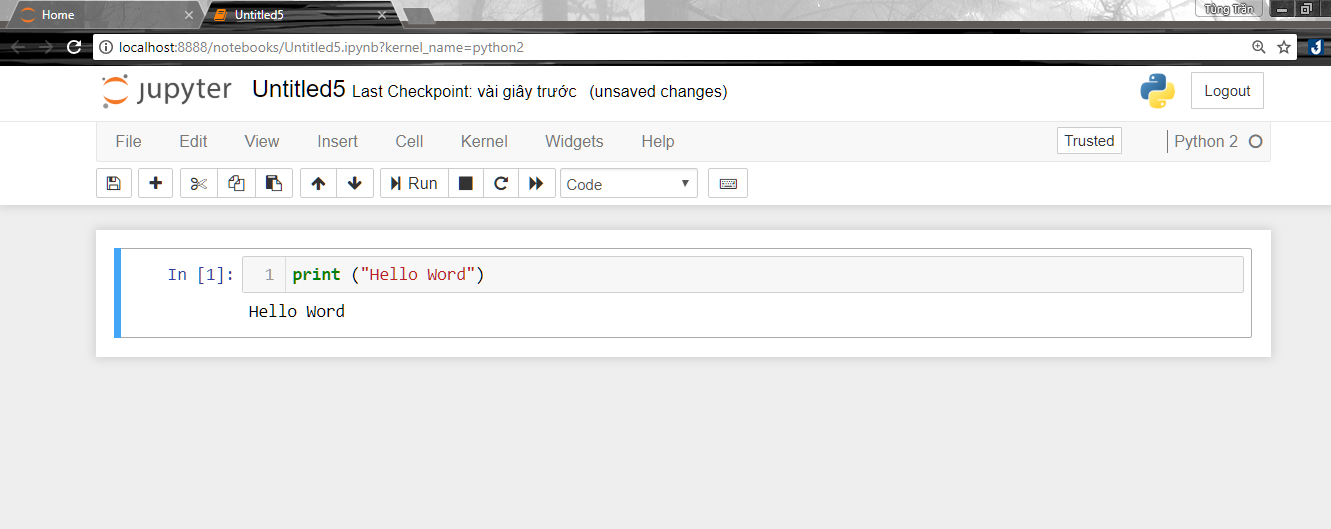


## Sử dụng

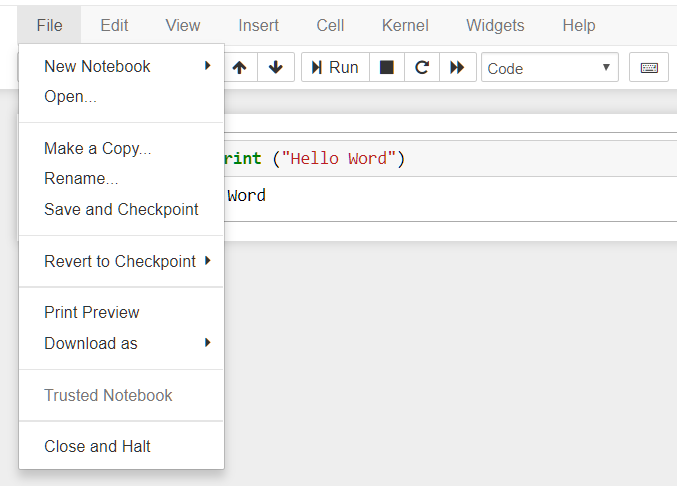
* Tạo một file python mới:



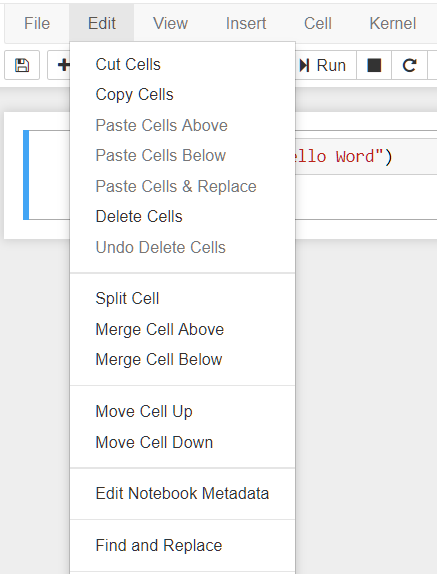
* Các thao tác:



* + Chức năng File:



* Các chức năng Edit:

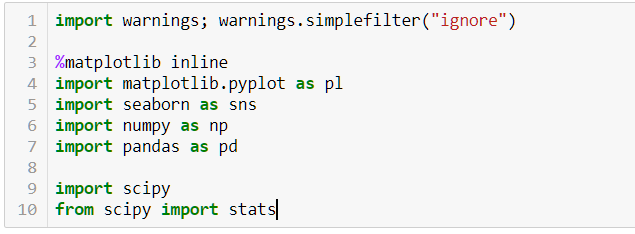


* Và còn nhiều chức năng khác.

# Data Visualization And Descriptive Statistical Measures

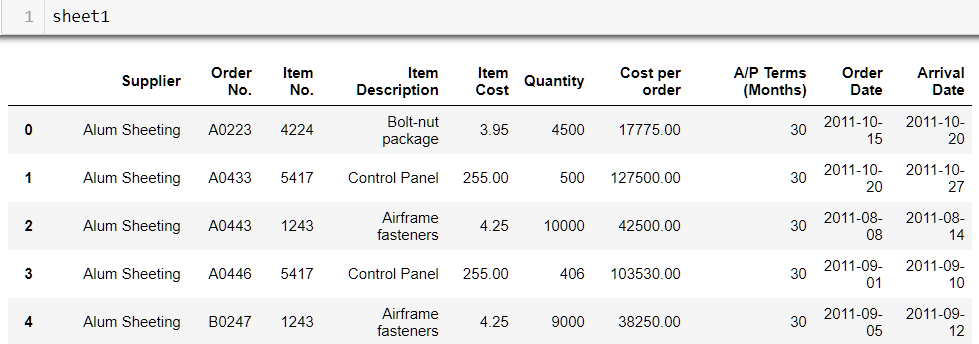
## 1.Tính giá trị.

* a. [Purchase Order](https://github.com/tranvantung241/data).xlsx
* Bài tập được thực hiện trên tập dữ liệu: [Purchase Order](https://github.com/tranvantung241/data).xlsx
* Tính các giá trị: Count, Min, Max, Mean, Median, Mode, Quantile, Range, Mode, Variance, Standard Deviation, Coefficient of Deviation, Skewness, Kurtosis.
* Đầu tiên tiến hành import các thư viện cần thiết

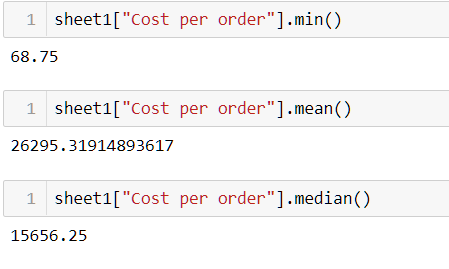


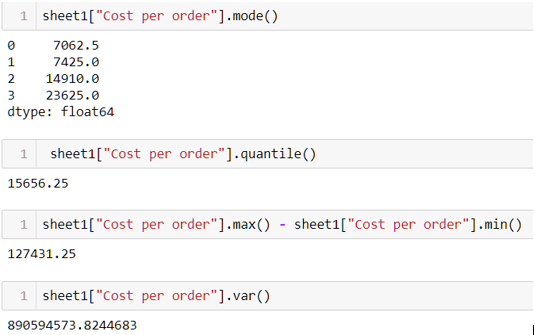
* Đọc file excle [Purchase Order](https://github.com/tranvantung241/data) và xem dữ liệu :

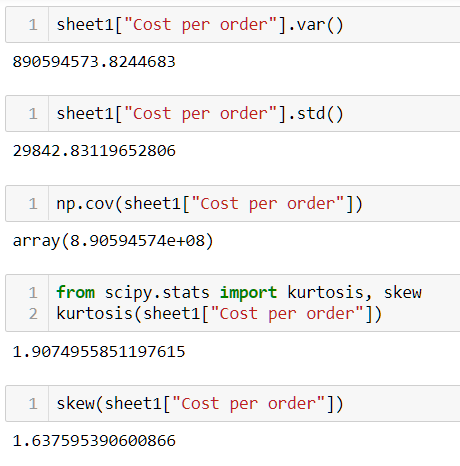




* Các giá trị:

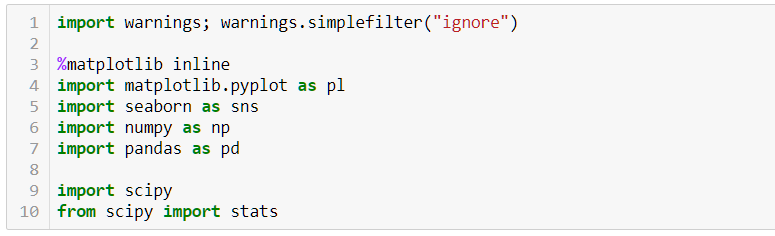






## b.[Dữ liệu Việt Nam:](https://github.com/tranvantung241/data)

* Đầu tiên tiến hành import các thư viện cần thiết



* Đọc fil dữ liệu:

|  |
| --- |
| sheet1 = pd.read\_excel(r"C:/Users/TranTung/Desktop/data/DuLieuVietNamVeDanSoNam2010-2016.xlsx", sheetname=0) |

* Kết quả:

|  |
| --- |
|  |

|  |
| --- |
|  |

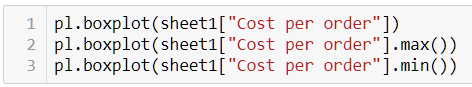
## 2.Vẽ Histogram, Box Plot

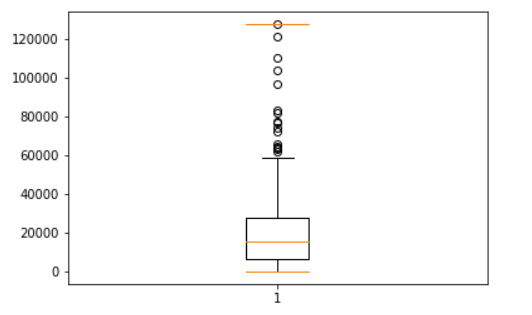
## a. [Purchase Order](https://github.com/tranvantung241/data).xlsx

**BoxPlot:**

* Thêm thư viện matplotlib để tiến hành vẽ đồ thị:

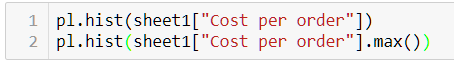


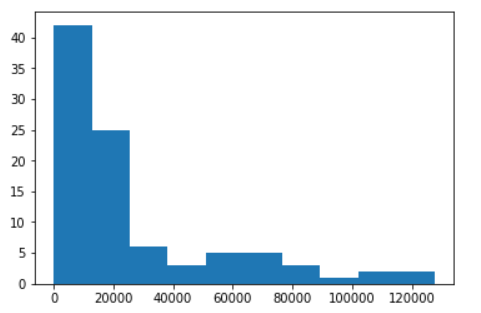




**Histogram**

* Tiến hành vẽ cột dữ liệu Cost per order lên histogram:

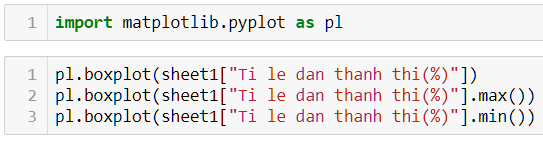


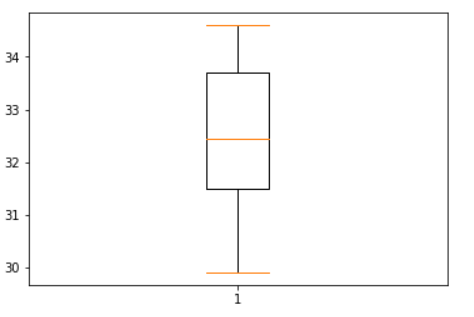


## b. [Dữ liệu Việt Nam:](https://github.com/tranvantung241/data)

**BoxPlot:**

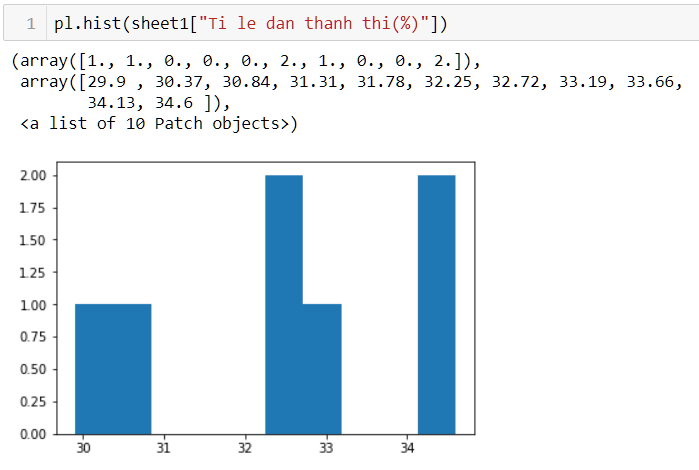
* Thêm thư viện matplotlib để tiến hành vẽ đồ thị:





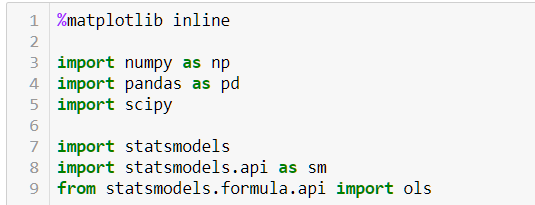
**Histogram**

* Tiến hành vẽ cột dữ liệu Ti le dan thanh thi(%) lên histogram:



# Statistical Inference

* Bài tập được thực hiện trên tập dữ liệu: [Insurance Survey](https://github.com/tranvantung241/data): One-Sample Hypothesis Tests, Analysis of Variance (ANOVA), Chi-Square Test for Independence
* Đầu tiên import các thư viện cần thiết



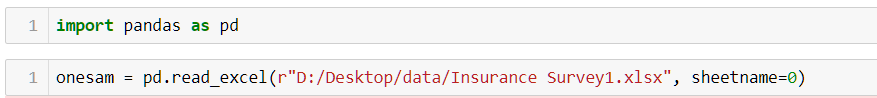
* Đọc file excel:

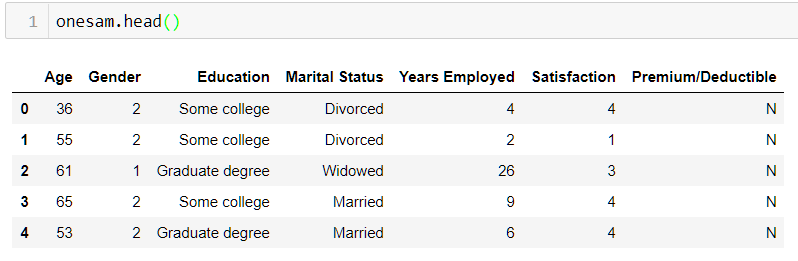


## One-Sample Hypothesis Tests

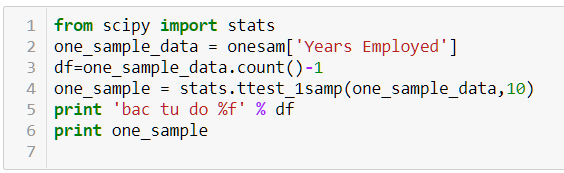
## a. Vacation Survey1.xlsx

* Bài tập được thực hiện trên tập dữ liệu: Vacation Survey1.xlsx
* Giả thuyết:
  + H0 : “Giá trị trung bình Years Employed >= 10”
  + H1: “Giá trị trung bình Years Employed < 10”
* Chúng ta tiến hành so sánh giá trị trung bình của **Years Employed** với 10
* Đầu tiên import thư viện vào đọc giữ liệu:





* Tiếp theo ta tính one-sample test:



* Kết quả:

|  |
| --- |
|  |

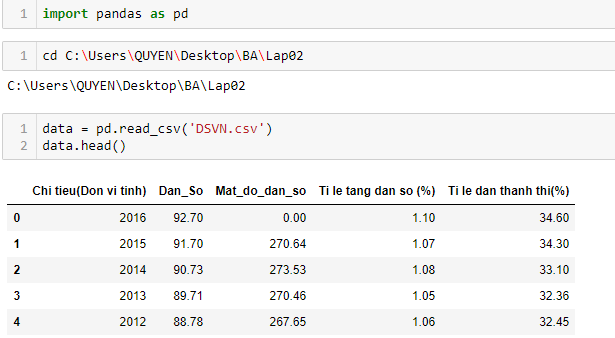
* Kết luận: pvalue: 0.153 > 0.05 , nên chấp nhận giả thuyết H0 . Giá trị trung bình của Years Employed có lớn hơn hoặc bằng 10.

b. DuLieuVietNamVeDanSoNam2010-2016.csv

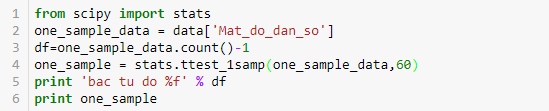
* Bài tập được thực hiện trên tập dữ liệu:

DuLieuVietNamVeDanSoNam2010-2016.csv

* Giả thuyết:
  + H0 : “Giá trị trung bình Dân số Việt Nam từ năm 2010- 2016 >= 60 triệu người”
  + H1: “Giá trị trung bình Dân số Việt Nam từ năm 2010-2016 < 60 triệu người”
* Chúng ta tiến hành so sánh giá trị trung bình của Dân số Việt Nam năm 2016 với 60 triệu người
* Đầu tiên import thư viện vào đọc giữ liệu:



* Tiếp theo ta tính one-sample test:



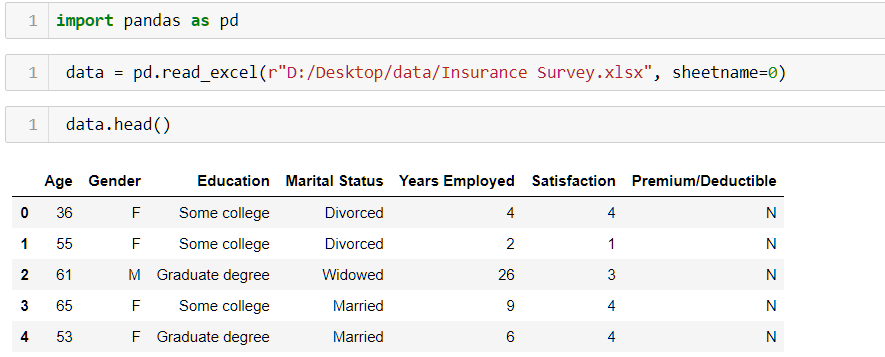
* Kết Quả:



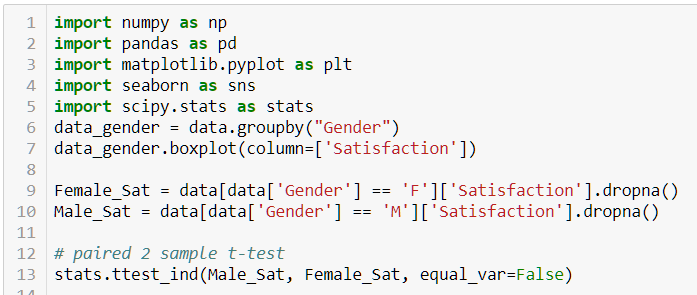
* Kết luận: pvalue: 0.004419 < 0.05 , nên từ chối giả thuyết H0 . Giá trị trung bình của Dân số Việt Nam từ năm 2010- 2016 lớn hơn hoặc bằng 60 triệu người.

## Two-Sample Hypothesis Tests

* Được thực hiện trên tập dữ liệu : [Insurance Survey.xlsx](https://github.com/tranvantung241/data)
* Giả thuyết:
  + H0: Giá trị trung bình trên nhóm F và nhóm M với độ hài lòng là **không có** sự khác biệt.
  + H1: Giá trị trung bình trên nhóm F và nhóm M với độ hài lòng là **có** sự khác biệt.
* Đầu tiên ta thêm thư viện và đọc file dữ liệu:



* Tính two-sample:



* Kết quả:

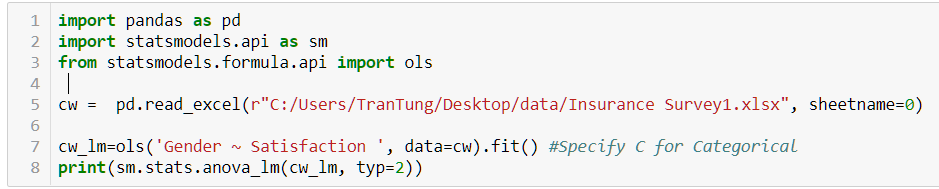
|  |
| --- |
|  |

* Kết luận: p-value: 0.97576 > 0.05 , nên chấp nhận giả thuyết H0, Giá trị trung bình trên nhóm F và nhóm M với độ hài lòng là **không có** sự khác biệt.

## Analysis of Variance (ANOVA)

## a. Insurance Survey1.xlsx

* Bài tập được thực hiện trên tập dữ liệu: Insurance Survey1.xlsx
* Giả thuyết :
  + H0 : Không có sự khác biệt về mức độ hài lòng trên từng nhóm giới tính
  + H1: Có sự khác biệt về mức độ hài lòng trên từng nhóm giới tính
* Chúng ta tiến hành import các thư viện và thực hiện các câu lệnh theo hình dưới đây:



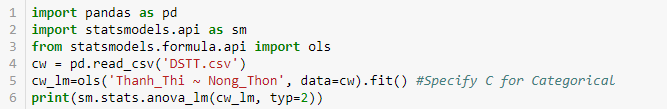
* Kết quả:

|  |
| --- |
|  |

* + Kết luận: p\_value: 0.976 > 0.05 , nên ta từ chối giả thuyết H0 : “Không có sự khác biệt về mức độ hài lòng trên từng nhóm giới tính” . Suy ra có sự khác biệt giữa nhóm giới tính với sự hài lòng.

b. Dân số Việt Nam từ 2000-2016

* Bài tập được thực hiện trên tập dữ liệu: DSTT.csv
* Giả thuyết :
  + H0 : Không có sự khác biệt về dân số Thành Thị(Trieu nguoi) và Nông Thôn từ năm 2000-2016
  + H1: Có sự khác biệt về dân số Thành Thị(Trieu nguoi) và Nông Thôn(Trieu nguoi) từ năm 2000-2016
* Chúng ta tiến hành import các thư viện và thực hiện các câu lệnh theo hình dưới đây:



* Kết quả:



* Kết luận: p\_value: 2.870410e-09 < 0.05 , nên ta chấp nhận giả thuyết:
  + H0 : “Không có sự khác biệt về dân số Thành Thị(Trieu nguoi) và Nông Thôn từ năm 2000-2016
  + ” . Suy ra không có sự khác biệt về dân số Thành Thị(Trieu nguoi) và Nông Thôn từ năm 2000-2016

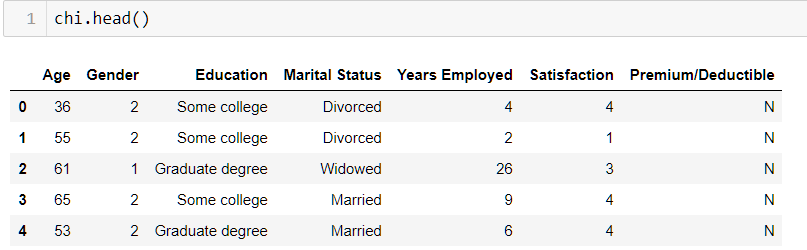
## Chi-Square Test for Independence

## a. Insurance Survey1.xlsx

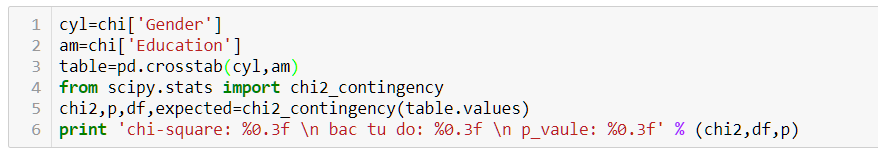
Bài tập được thực hiện trên tập dữ liệu: Insurance Survey1.xlsx

* Giả thuyết :
  + H0 : Hai biến Gender và Education độc lập với nhau
  + H1: Hai biến Gender và Education phụ thuộc với nhau
* Chúng ta tiến hành thêm thư viện và đọc dữ liệu từ file excel:

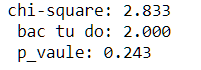




* Chúng ta tiến hành thực hiện phép tính Chi-Square trên 2 cột dữ liệu : Gender và Education



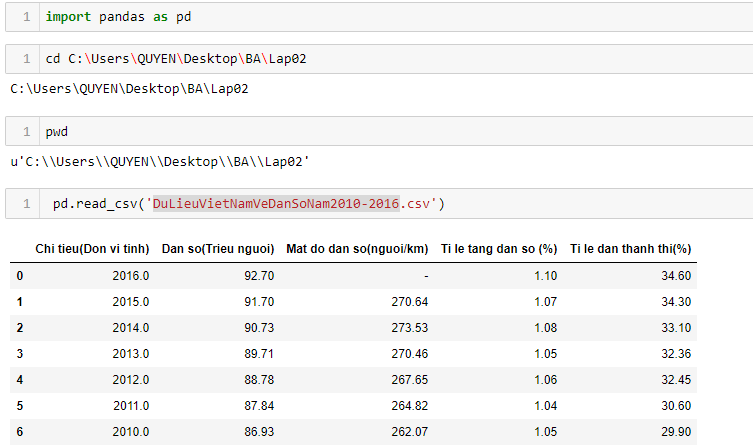
* Kết quả:

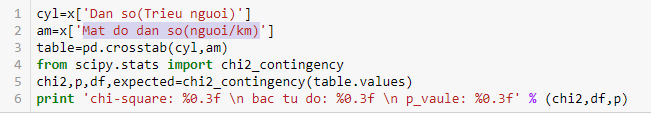


* Kết luận : p\_value = 0.243 > 0.05 , nên chúng ta từ chối giả thuyết H0 = “Hai biến Gender và Education độc lập với nhau” . Suy ra hai biến có liên hệ với nhau (phụ thuộc nhau)

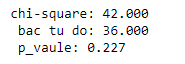
b.DuLieuVietNamVeDanSoNam2010-2016

Bài tập được thực hiện trên tập dữ liệu:DuLieuVietNamVeDanSoNam2010-2016.csv

* Giả thuyết :
  + H0 : Hai biến Dan so(Trieu nguoi) và Mat do dan so(nguoi/km) độc lập với nhau
  + H1: Hai biến Dan so(Trieu nguoi)và Mat do dan so(nguoi/km)phụ thuộc với nhau
* Chúng ta tiến hành thêm thư viện và đọc dữ liệu:
* 
* Chúng ta tiến hành thực hiện phép tính Chi-Square trên 2 cột dữ liệu : Dan so(Trieu nguoi) và Mat do dan so(nguoi/km)



-Kết Quả:



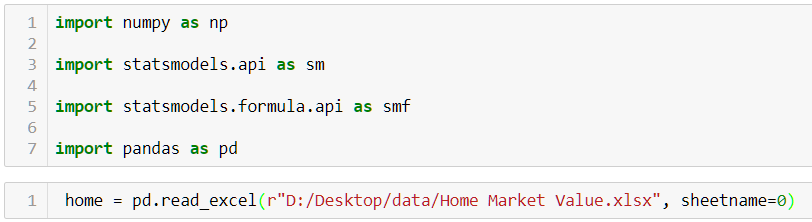
* Kết luận : p\_value = 0.227 > 0.05 , nên chúng ta từ chối giả thuyết H0 = “Hai biến Dan so(Trieu nguoi) và Mat do dan so(nguoi/km) độc lập với nhau” . Suy ra hai biến có liên hệ với nhau (phụ thuộc nhau)

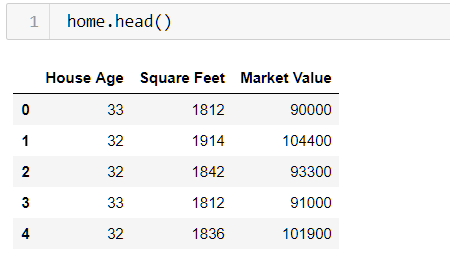
# Regression Analysis

## Simple Linear Regression

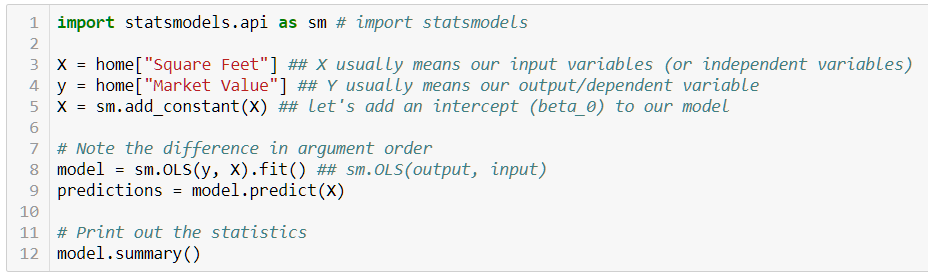
## a. Home Market Value.xlsx

* Bài tập được thực hiện trên tập dữ liệu: Home Market Value.xlsx
* Giả thuyết:
  + H0 = “Phương trình tìm được không có ý nghĩa, và giá ngôi nhà không phụ thuộc vào diện tích nhà.”
  + H1 = “Phương trình tìm được có ý nghĩa, giá ngôi nhà phụ thuộc vào diện tích nhà.”
* 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:

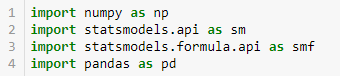
|  |
| --- |
|  |

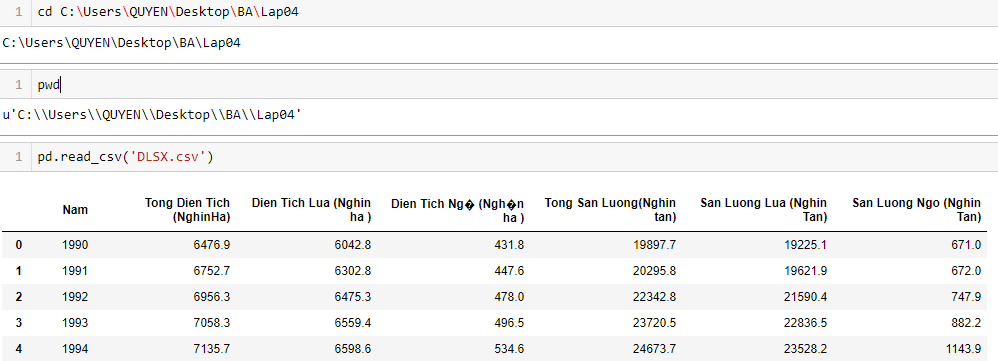
|  |
| --- |
|  |

* Kết luận:
  + Phương trình tìm được là: Y = 3.267e+04 + 35.036 b1
  + Ta thấy p-value: 3.80e-08< 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được và giá bán ngôi nhà có phụ thuộc vào diện tích ngôi nhà.

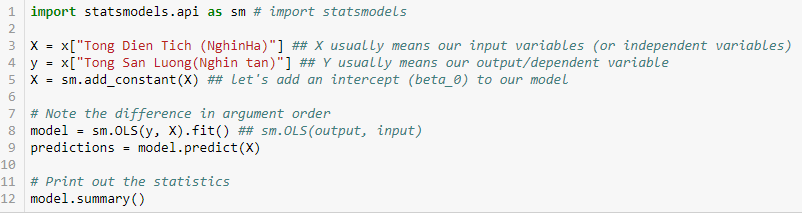
b. DLSX.csv

* Bài tập được thực hiện trên tập dữ liệu: DLSX.csv
* 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 :

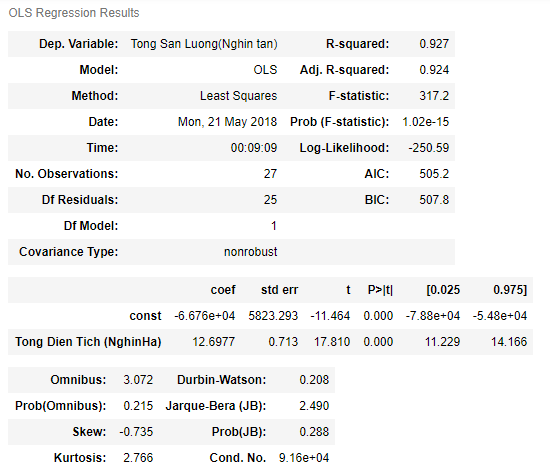




* Tiến hành các phép tính tính hồi quy tuyến tính, và tìm ra phương trình có nghĩa:



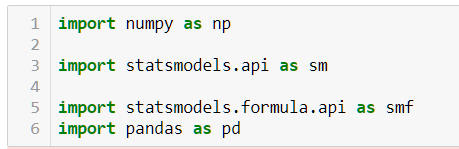
-Kết Quả:



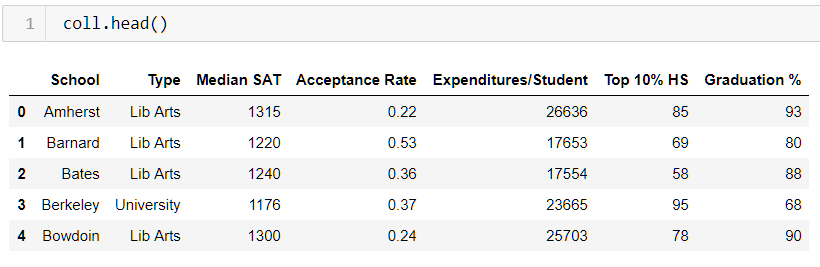
* Kết luận:
  + Phương trình tìm được là: Y = -6.676e+04 + 12.6977 b1
  + Ta thấy p-value: 1.02e-15< 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được và Tổng sản lượng lúa thu được phụ thuộc vào Tổng diện tích đất trồng

## Multiple Linear Regression

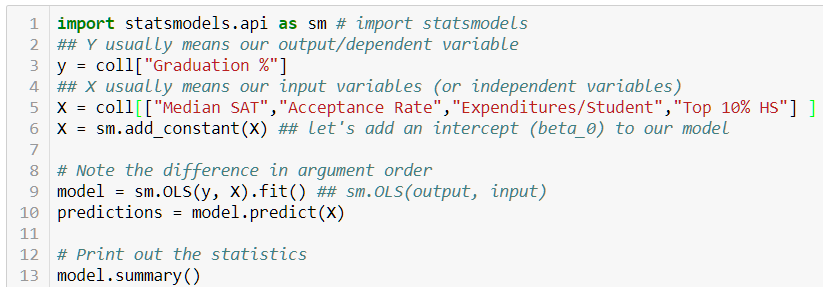
* Bài tập được thực hiện trên tập dữ liệu: [Colleges and Universities.xlsx](https://github.com/tranvantung241/data)
* Giả thuyết:
  + H0 = “Phương trình tìm được không có ý nghĩa, và **Graduation %** không phụ thuộc vào **Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS ”**
  + H1 = “Phương trình tìm được có ý nghĩa, và **Graduation %** phụ thuộc vào **Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS ”**
* Thêm thư viện và đọc file dữ liệu:







* Tiến hành tính Multiple Linear Regression



* Kết quả:

|  |
| --- |
|  |

|  |
| --- |
|  |

* **Kết luận:** 
  + Phương trình tìm được là: Y = 17.9210 + 0.0720\*b1 – 24.85\*b2 – 0.0001\* b3 – 0.01628\*b4
  + Ta thấy p-value: 6.33e-07 < 0.05 = > Từ chối H0
  + Kết luận: Phương trình tìm được có ý nghĩa, và **Graduation %** phụ thuộc vào **Median SAT, Acceptance Rate, Expenditures/Student, Top 10% HS**

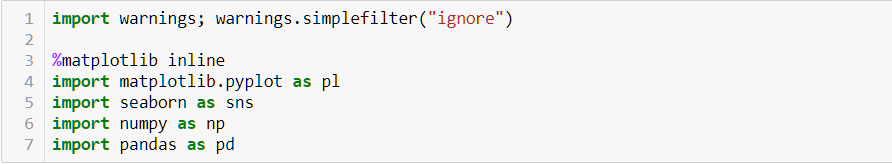
# Forecasting Techniques

## Statistical Forecasting Models

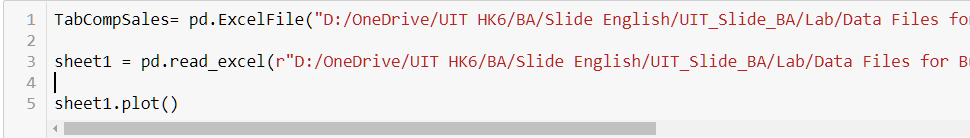
* a. [Tablet Computer Sales.xlsx](https://github.com/tranvantung241/data)

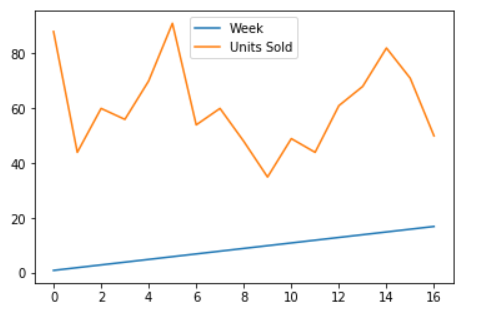
Moving Average

* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : [Tablet Computer Sales.xlsx](https://github.com/tranvantung241/data)
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:



* Đọc file dữ liệu: (TabCompSales= pd.ExcelFile(“/path”)):

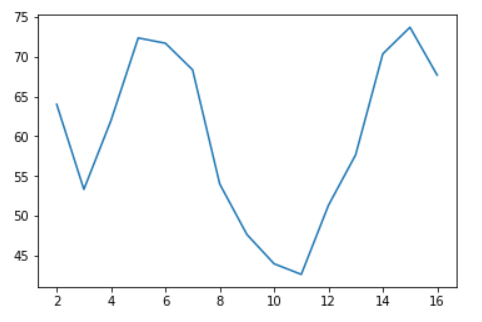




* Tiếp theo ta tính Moving Average của Units Sale với k =3:



|  |
| --- |
|  |



- So sánh với kết quả khi làm với Execl:

|  |
| --- |
|  |

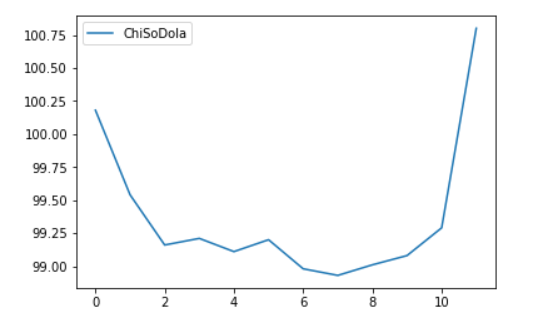
## b.[Dữ liệu Việt Nam](https://github.com/tranvantung241/data)

* Thực hiện trên tập dữ liệu ChiSoDoLa.csv:

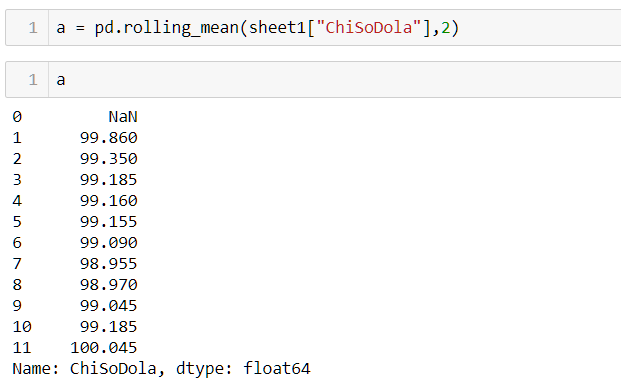
Moving Average

* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : ChiSoDoLa.xlsx
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:
* Đọc file dữ liệu: (TabCompSales= pd.ExcelFile(“/path”)):





* Tiếp theo ta tính Moving Average của Units Sale với k =2:

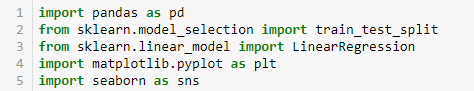


## Regression Forecasting with Causal Variables

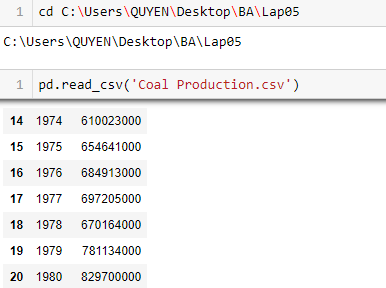
a. [*Coal Production*.xlsx](https://github.com/tranvantung241/data)

Simple linear regression

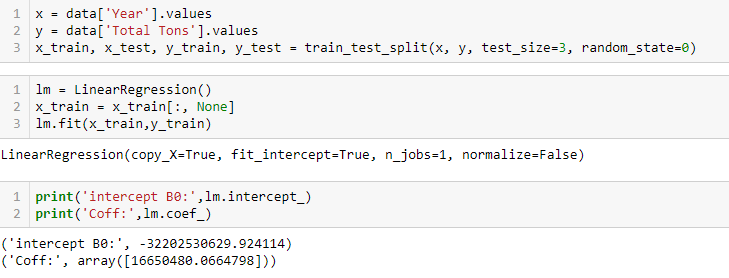
* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : [*Coal Production*.xlsx](https://github.com/tranvantung241/data)
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:



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

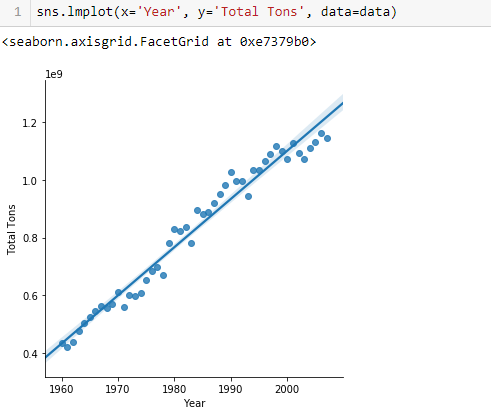


* Tiếp theo ta lấy giá trị trong tập dữ liệu và khởi tạo mô hình.Xét dự báo cho giai đoạn trong tương lai là 3(test\_size=3)

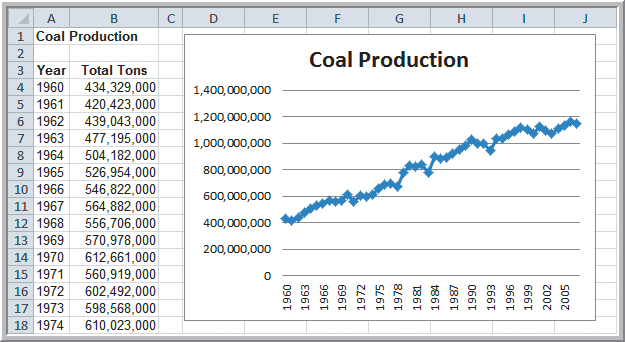


* Bây giờ mô hình của chúng ta được khởi tạo, chúng tôi có thể đánh giá nó. Đề cập trở lại phương trình hồi quy tuyến tính đơn giản ,ta có:

F=-32202530629K+16650480

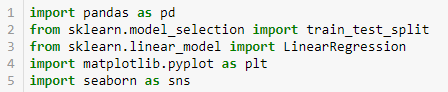


- So sánh với kết quả khi làm với Execl:

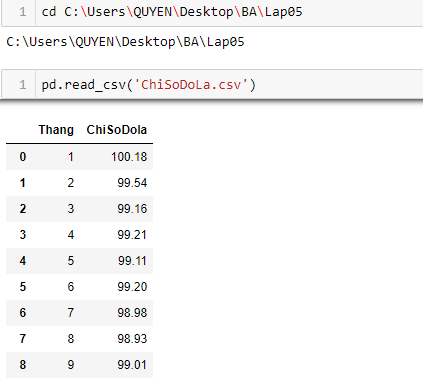


## b. Dữ Liệu Việt Nam.csv

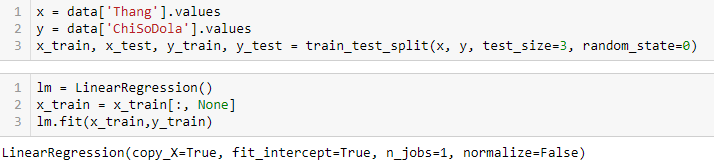
* Trong phần này sẽ tiến hành trên tập dữ liệu đơn giản : ChiSoDoLa.csv
* Đầu tiên sẽ tiến hành import các thư viện cần thiết:



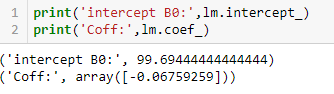
* Đọc file dữ liệu: (*ChiSoDoLa*= pd.read\_csv(“/path”)):



* Tiếp theo ta lấy giá trị trong tập dữ liệu và khởi tạo mô hình.Xét dự báo cho giai đoạn trong tương lai là 3(test\_size=3)



* Bây giờ mô hình của chúng ta được khởi tạo, chúng tôi có thể đánh giá nó. Đề cập trở lại phương trình hồi quy tuyến tính đơn giản ,ta có:



F=99.694K – 0.06759

