ĐẠI HỌC TÔN ĐỨC THẮNG
TON DUC THANG UNIVERSITY

Trường Đại học Tôn Đức Thắng

Pandas, Sklearn

Giảng viên: Tiến sĩ Bùi Thanh Hùng
Trưởng Lab Khoa học Phân tích dữ liệu và Trí tuệ nhân tạo
Giám đốc chương trình Hệ thống thông tin
Đại học Thủ Dầu Một
Email: tuhungphe@gmail.com
Website: https://sites.google.com/site/hungthanhbui1980/

Tutorial Content

- Overview of Python Libraries for Data Scientists
- Reading Data; Selecting and Filtering the Data; Data manipulation, sorting, grouping, rearranging
- Plotting the data
- Descriptive statistics
- Inferential statistics

Python Libraries for Data Science

Many popular Python toolboxes/libraries:

- NumPy
- SciPy
- Pandas
- SciKit-Learn

All these libraries are installed on the SCC

Visualization libraries

- matplotlib
- Seaborn

and many more …

Python Libraries for Data Science

NumPy:

- introduces objects for multidimensional arrays and matrices, as well as functions that allow to easily perform advanced mathematical and statistical operations on those objects

- provides vectorization of mathematical operations on arrays and matrices which significantly improves the performance

- many other python libraries are built on NumPy

Link: http://www.numpy.org/

Python Libraries for Data Science

SciPy:

- collection of algorithms for linear algebra, differential equations, numerical integration, optimization, statistics and more

- part of SciPy Stack

- built on NumPy

Link: https://www.scipy.org/scipylib/

Python Libraries for Data Science

Pandas:

- adds data structures and tools designed to work with table-like data (similar to Series and Data Frames in R)

- provides tools for data manipulation: reshaping, merging, sorting, slicing, aggregation etc.

- allows handling missing data

Link: http://pandas.pydata.org/

Python Libraries for Data Science

SciKit-Learn:

- provides machine learning algorithms: classification, regression, clustering, model validation etc.

- built on NumPy, SciPy and matplotlib

Link: http://scikit-learn.org/

Python Libraries for Data Science

matplotlib:

- python 2D plotting library which produces publication quality figures in a variety of hardcopy formats
- a set of functionalities similar to those of MATLAB
- line plots, scatter plots, barcharts, histograms, pie charts etc.
- relatively low-level; some effort needed to create advanced visualization

Link: https://matplotlib.org/

Python Libraries for Data Science

Seaborn:

- based on matplotlib

- provides high level interface for drawing attractive statistical graphics

- Similar (in style) to the popular ggplot2 library in R

Link: https://seaborn.pydata.org/

Loading Python Libraries

In [ ]:
```python
#Import Python Libraries
import numpy as np
import scipy as sp
import pandas as pd
import matplotlib as mpl
import seaborn as sns
```

Press `Shift+Enter` to execute the *jupyter* cell

Reading data using pandas

In [ ]:

```
#Read csv file
df = pd.read_csv("http://rcs.bu.edu/examples/python/data_analysis/Salaries.csv")
```

Note: The above command has many optional arguments to fine-tune the data import process.

There is a number of pandas commands to read other data formats:

```
pd.read_excel('myfile.xlsx',sheet_name='Sheet1',
index_col=None, na_values=['NA'])
pd.read_stata('myfile.dta')
pd.read_sas('myfile.sas7bdat')
pd.read_hdf('myfile.h5','df')
```

Exploring data frames

In [3]:

```
#List first 5 records
df.head()
```

Out[3]

	rank	discipline	phd	service	sex	salary
0	Prof	B	56	49	Male	186960
1	Prof	A	12	6	Male	93000
2	Prof	A	23	20	Male	110515
3	Prof	A	40	31	Male	131205
4	Prof	B	20	18	Male	104800

Hands-on exercises

- ✓ Try to read the first 10, 20, 50 records;
- ✓ Can you guess how to view the last few records;

Hint:

Data Frame data types

Pandas Type	**Native Python Type**	**Description**
object	string	The most general dtype. Will be assigned to your column if column has mixed types (numbers and strings).
int64	int	Numeric characters. 64 refers to the memory allocated to hold this character.
float64	float	Numeric characters with decimals. If a column contains numbers and NaNs(see below), pandas will default to float64, in case your missing value has a decimal.
datetime64, timedelta[ns]	N/A (but see the datetime module in Python's standard library)	Values meant to hold time data. Look into these for time series experiments.

Data Frame data types

In [4]:

```
#Check a particular column type
df['salary'].dtype
```

Out[4]: dtype('int64')

In [5]:

```
#Check types for all the columns
df.dtypes
```

Out[4]:

```
rank          object
discipline    object
phd           int64
service       int64
sex           object
salary        int64
dtype: object
```

Data Frames attributes

Python objects have *attributes* and *methods*.

df.attribute	description
dtypes	list the types of the columns
columns	list the column names
axes	list the row labels and column names
ndim	number of dimensions
size	number of elements
shape	return a tuple representing the dimensionality
values	numpy representation of the data

Hands-on exercises

- ✓ Find how many records this data frame has;
- ✓ How many elements are there?
- ✓ What are the column names?
- ✓ What types of columns we have in this data frame?

Data Frames methods

Unlike attributes, python methods have *parenthesis.*
All attributes and methods can be listed with a *dir()* function: `dir(df)`

df.method()	description
head([n]), tail([n])	first/last n rows
describe()	generate descriptive statistics (for numeric columns only)
max(), min()	return max/min values for all numeric columns
mean(), median()	return mean/median values for all numeric columns
std()	standard deviation
sample([n])	returns a random sample of the data frame
dropna()	drop all the records with missing values

Hands-on exercises

- ✓ Give the summary for the numeric columns in the dataset
- ✓ Calculate standard deviation for all numeric columns;
- ✓ What are the mean values of the first 50 records in the dataset? ***Hint:*** use head() method to subset the first 50 records and then calculate the mean

Selecting a column in a Data Frame

Method 1: Subset the data frame using column name:

df['sex']

Method 2: Use the column name as an attribute:

df.sex

Note: there is an attribute *rank* for pandas data frames, so to select a column with a name "rank" we should use method 1.

Hands-on exercises

- ✓ Calculate the basic statistics for the *salary* column;
- ✓ Find how many values in the *salary* column (use *count* method);
- ✓ Calculate the average salary;

Data Frames *groupby* method

Using "group by" method we can:

- Split the data into groups based on some criteria
- Calculate statistics (or apply a function) to each group
- Similar to dplyr() function in R

```
In [ ]: #Group data using rank
        df_rank = df.groupby(['rank'])
```

```
In [ ]: #Calculate mean value for each numeric column per each group
        df_rank.mean()
```

	phd	service	salary
rank			
AssocProf	15.076923	11.307692	91786.230769
AsstProf	5.052632	2.210526	81362.789474
Prof	27.065217	21.413043	123624.804348

Data Frames *groupby* method

Once groupby object is create we can calculate various

statistics for each group:

In [ ]:

```
#Calculate mean salary for each professor rank:
df.groupby('rank')[['salary']].mean()
```

	salary
rank	
AssocProf	91786.230769
AsstProf	81362.789474
Prof	123624.804348

Note: If single brackets are used to specify the column (e.g. salary), then the output is Pandas Series object. When double brackets are used the output is a Data Frame

Data Frames *groupby* method

groupby performance notes:

- no grouping/splitting occurs until it's needed. Creating the *groupby* object only verifies that you have passed a valid mapping
- by default the group keys are sorted during the *groupby* operation. You may want to pass sort=False for potential speedup:

In [ ]:

```
#Calculate mean salary for each professor rank:
df.groupby(['rank'], sort=False)[['salary']].mean()
```

Data Frame: filtering

To subset the data we can apply Boolean indexing. This indexing is commonly known as a filter. For example if we want to subset the rows in which the salary value is greater than $120K:

```
In [ ]: #Calculate mean salary for each professor rank:
        df_sub = df[ df['salary'] > 120000 ]
```

Any Boolean operator can be used to subset the data:

\> greater; >= greater or equal;

< less; <= less or equal;

== equal; != not equal;

```
In [ ]: #Select only those rows that contain female professors:
        df_f = df[ df['sex'] == 'Female' ]
```

Data Frames: Slicing

There are a number of ways to subset the Data Frame:

- one or more columns
- one or more rows
- a subset of rows and columns

Rows and columns can be selected by their position or label

Data Frames: Slicing

When selecting one column, it is possible to use single set of brackets, but the resulting object will be a Series (not a DataFrame):

In [ ]:

```
#Select column salary:
df['salary']
```

When we need to select more than one column and/or make the output to be a DataFrame, we should use double brackets:

In [ ]:

```
#Select column salary:
df[['rank','salary']]
```

Data Frames: Selecting rows

If we need to select a range of rows, we can specify the range using ":"

In [ ]:

```
#Select rows by their position:
df[10:20]
```

Notice that the first row has a position 0, and the last value in the range is omitted:
So for 0:10 range the first 10 rows are returned with the positions starting with 0 and ending with 9

Data Frames: method loc

If we need to select a range of rows, using their labels we can use method loc:

In [ ]:

```
#Select rows by their labels:
df_sub.loc[10:20,['rank','sex','salary']]
```

Out[ ]

	rank	sex	salary
10	Prof	Male	128250
11	Prof	Male	134778
13	Prof	Male	162200
14	Prof	Male	153750
15	Prof	Male	150480
19	Prof	Male	150500

Data Frames: method iloc

If we need to select a range of rows and/or columns, using their positions we can use method iloc:

In [ ]:
```
#Select rows by their labels:
df_sub.iloc[10:20,[0, 3, 4, 5]]
```

Out[ ]:

	rank	service	sex	salary
26	Prof	19	Male	148750
27	Prof	43	Male	155865
29	Prof	20	Male	123683
31	Prof	21	Male	155750
35	Prof	23	Male	126933
36	Prof	45	Male	146856
39	Prof	18	Female	129000
40	Prof	36	Female	137000
44	Prof	19	Female	151768
45	Prof	25	Female	140096

Data Frames: method iloc (summary)

```
df.iloc[0]   # First row of a data frame
df.iloc[i]   #(i+1)th row
df.iloc[-1]  # Last row
```

```
df.iloc[:, 0]   # First column
df.iloc[:, -1]  # Last column
```

```
df.iloc[0:7]           #First 7 rows
df.iloc[:, 0:2]        #First 2 columns
df.iloc[1:3, 0:2]      #Second through third rows and first 2 columns
df.iloc[[0,5], [1,3]]   #1st and 6th rows and 2nd and 4th columns
```

Data Frames: Sorting

We can sort the data by a value in the column. By default the sorting will occur in ascending order and a new data frame is return.

In [ ]:
```
# Create a new data frame from the original sorted by the column Salary
df_sorted = df.sort_values( by ='service')
df_sorted.head()
```

Out[ ]:

	rank	discipline	phd	service	sex	salary
55	AsstProf	A	2	0	Female	72500
23	AsstProf	A	2	0	Male	85000
43	AsstProf	B	5	0	Female	77000
17	AsstProf	B	4	0	Male	92000
12	AsstProf	B	1	0	Male	88000

Data Frames: Sorting

We can sort the data using 2 or more columns:

In [ ]:

```
df_sorted = df.sort_values( by =['service', 'salary'],
ascending = [True, False])
df_sorted.head(10)
```

Out[ ]:

	rank	discipline	phd	service	sex	salary
52	Prof	A	12	0	Female	105000
17	AsstProf	B	4	0	Male	92000
12	AsstProf	B	1	0	Male	88000
23	AsstProf	A	2	0	Male	85000
43	AsstProf	B	5	0	Female	77000
55	AsstProf	A	2	0	Female	72500
57	AsstProf	A	3	1	Female	72500
28	AsstProf	B	7	2	Male	91300
42	AsstProf	B	4	2	Female	80225
68	AsstProf	A	4	2	Female	77500

Missing Values

Missing values are marked as NaN

In [ ]:

```
# Read a dataset with missing values
flights =
pd.read_csv("http://rcs.bu.edu/examples/python/data_analysis/fl
ights.csv")
```

```
# Select the rows that have at least one missing value
flights[flights.isnull().any(axis=1)].head()
```

Out[ ]:

	year	month	day	dep_time	dep_delay	arr_time	arr_delay	carrier	tailnum	flight	origin	dest	air_time	distance	hour	minute
330	2013	1	1	1807.0	29.0	2251.0	NaN	UA	N31412	1228	EWR	SAN	NaN	2425	18.0	7.0
403	2013	1	1	NaN	NaN	NaN	NaN	AA	N3EHAA	791	LGA	DFW	NaN	1389	NaN	NaN
404	2013	1	1	NaN	NaN	NaN	NaN	AA	N3EVAA	1925	LGA	MIA	NaN	1096	NaN	NaN
855	2013	1	2	2145.0	16.0	NaN	NaN	UA	N12221	1299	EWR	RSW	NaN	1068	21.0	45.0
858	2013	1	2	NaN	NaN	NaN	NaN	AA	NaN	133	JFK	LAX	NaN	2475	NaN	NaN

Missing Values

There are a number of methods to deal with missing values in the data frame:

df.method()	**description**
dropna()	Drop missing observations
dropna(how='all')	Drop observations where all cells is NA
dropna(axis=1, how='all')	Drop column if all the values are missing
dropna(thresh = 5)	Drop rows that contain less than 5 non-missing values
fillna(0)	Replace missing values with zeros
isnull()	returns True if the value is missing
notnull()	Returns True for non-missing values

Missing Values

- When summing the data, missing values will be treated as zero
- If all values are missing, the sum will be equal to NaN
- cumsum() and cumprod() methods ignore missing values but preserve them in the resulting arrays
- Missing values in GroupBy method are excluded (just like in R)
- Many descriptive statistics methods have *skipna* option to control if missing data should be excluded . This value is set to *True* by default (unlike R)

Aggregation Functions in Pandas

Aggregation - computing a summary statistic about each group, i.e.

- compute group sums or means
- compute group sizes/counts

Common aggregation functions:

min, max
count, sum, prod
mean, median, mode, mad
std, var

Aggregation Functions in Pandas

agg() method are useful when multiple statistics are computed per column:

In [ ]:

```
flights[['dep_delay','arr_delay']].agg(['min','mean','max'])
```

Out[ ]:

	dep_delay	arr_delay
min	-16.000000	-62.000000
mean	9.384302	2.298675
max	351.000000	389.000000

Basic Descriptive Statistics

df.method()	**description**
describe	Basic statistics (count, mean, std, min, quantiles, max)
min, max	Minimum and maximum values
mean, median, mode	Arithmetic average, median and mode
var, std	Variance and standard deviation
sem	Standard error of mean
skew	Sample skewness
kurt	kurtosis

Graphics to explore the data

Seaborn package is built on matplotlib but provides high level interface for drawing attractive statistical graphics, similar to ggplot2 library in R. It specifically targets statistical data visualization

To show graphs within Python notebook include inline directive:

In [ ]:

```
%matplotlib inline
```

Graphics

	description
distplot	histogram
barplot	estimate of central tendency for a numeric variable
violinplot	similar to boxplot, also shows the probability density of the data
jointplot	Scatterplot
regplot	Regression plot
pairplot	Pairplot
boxplot	boxplot
swarmplot	categorical scatterplot
factorplot	General categorical plot

Basic statistical Analysis

statsmodel and scikit-learn - both have a number of function for statistical analysis

The first one is mostly used for regular analysis using R style formulas, while scikit-learn is more tailored for Machine Learning.

statsmodels:

- linear regressions
- ANOVA tests
- hypothesis testings
- many more ...

scikit-learn:

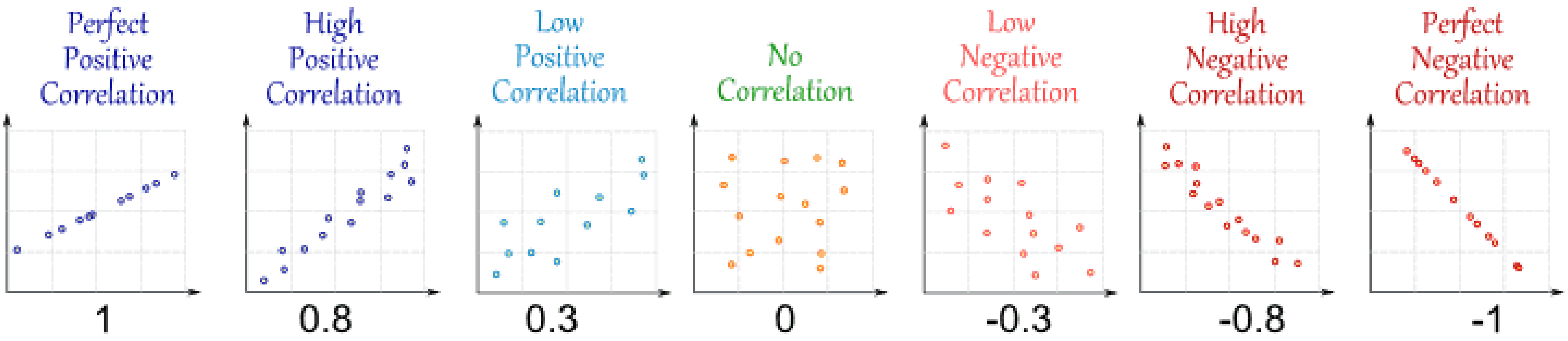
- kmeans
- support vector machines
- random forests
- many more ...

Tương quan (Correlation)

- Trong lý thuyết xác suất và thống kê, hệ số tương quan (Coefficient Correlation) cho biết độ mạnh của mối quan hệ tuyến tính giữa hai biến số ngẫu nhiên. Từ tương quan (Correlation) được thành lập từ **Co-** (có nghĩa "together") và Relation (quan hệ).
- Hệ số tương quan giữa 2 biến có thể dương (positive) hoặc âm (negative). Hệ số tương quan dương cho biết rằng giá trị 2 biến tăng cùng nhau còn hệ số tương quan âm thì nếu một biến tăng thì biến kia giảm.

Tương quan (Correlation)

- *Độ mạnh và hướng tương quan của 2 biến được mô tả như sau:*

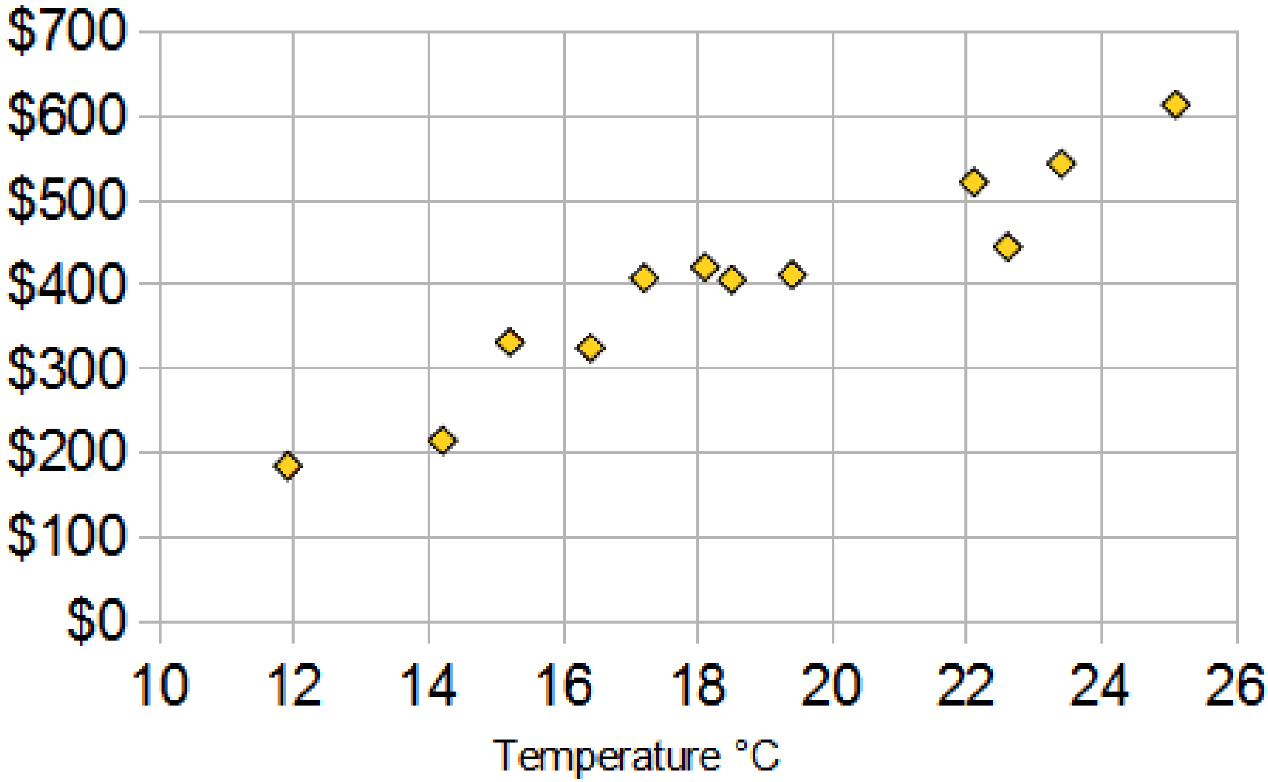


Hệ số tương quan có thể nhận giá trị từ -1 đến 1:

Có dữ liệu (bivariate) về nhiệt độ (Temperature) và doanh thu bán kem (Ice Cream Sales) như sau:

Ice Cream Sales vs Temperature

Temperature °C	**Ice Cream Sales**
14.2°	$215
16.4°	$325
11.9°	$185
15.2°	$332
18.5°	$406
22.1°	$522
19.4°	$412
25.1°	$614
23.4°	$544
18.1°	$421
22.6°	$445
17.2°	$408



Từ Scatter Plot, ta có thể thấy rằng nhiệt độ càng cao thì doanh thu bán kem càng cao. Trong dữ liệu trên, hệ số tương quan là **0.9575** và mối quan hệ giữa nhiệt độ và doanh số bán kem là rất mạnh. Hệ số tương quan dương nói rằng nhiệt độ tăng thì doanh số bán kem cũng tăng.

The z-score for the X value

The z-score for the Y value

Correlation coefficient

$$r = \frac{\sum (z_x z_y)}{n}$$

The number of pairs of scores

Hệ số tương quan Pearson

- (Pearson correlation coefficient, kí hiệu r) đo lường mức độ tương quan tuyến tính giữa hai biến. Nguyên tắc cơ bản, tương quan Pearson sẽ tìm ra một đường thẳng phù hợp nhất với mối quan hệ tuyến tính của 2 biến. Chính vì vậy, phân tích tương quan Pearson đôi khi còn được gọi là phân tích hồi quy giản đơn (nhưng khác nhau về mặt ý nghĩa).
- Hệ số tương quan Pearson (r) sẽ nhận giá trị từ +1 đến -1. r > 0 cho biết một sự tương quan dương giữa hai biến, nghĩa là nếu giá trị của biến này tăng thì sẽ làm tăng giá trị của biến kia và ngược lại. r < 0 cho biết một sự tương quan âm giữa hai biến, nghĩa là nếu giá trị của biến này tăng thì sẽ làm giảm giá trị của biến kia và ngược lại.
- Giá trị tuyệt đối của r càng cao thì mức độ tương quan giữa 2 biến càng lớn hoặc dữ liệu càng phù hợp với quan hệ tuyến tính giữa hai biến. Giá trị r bằng +1 hoặc bằng -1 cho thấy dữ liệu hoàn toàn phù hợp với mô hình tuyến tính.

Pearson's

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Bước 1: Tính trung bình của x và y
Bước 2: Tính độ lệch của mỗi giá trị của x với trung bình của x (lấy các giá trị của x trừ đi trung bình của x) và gọi là "**a**", làm tương tự như vậy với y và gọi là "**b**"
Bước 3: Tính: **a × b**, $\mathbf{a^2}$ và $\mathbf{b^2}$ cho mỗi giá trị
Bước 4: Tính tổng **a × b**, tổng $\mathbf{a^2}$ vả tổng $\mathbf{b^2}$
Bước 5: Chia tổng của **a × b** cho căn bậc 2 của [(sum a^2) × (sum b^2)]

2 Subtract Mean

3 Calculate ab, a² and b²

Temp °C	Sales	"a"	"b"	a×b	a²	b²
14.2	$215	-4.5	-$187	842	20.3	34,969
16.4	$325	-2.3	-$77	177	5.3	5,929
11.9	$185	-6.8	-$217	1,476	46.2	47,089
15.2	$332	-3.5	-$70	245	12.3	4,900
18.5	$406	-0.2	$4	-1	0.0	16
22.1	$522	3.4	$120	408	11.6	14,400
19.4	$412	0.7	$10	7	0.5	100
25.1	$614	6.4	$212	1,357	41.0	44,944
23.4	$544	4.7	$142	667	22.1	20,164
18.1	$421	-0.6	$19	-11	0.4	361
22.6	$445	3.9	$43	168	15.2	1,849
17.2	$408	-1.5	$6	-9	2.3	36
18.7	**$402**			**5,325**	**177.0**	**174,757**

1 Calculate Means

4 Sum Up

5 $$\frac{5{,}325}{\sqrt{177.0 \times 174{,}757}} = 0.9575$$

Pearson's

X	1	2	3	4	5
Y	10	20	30	40	50

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Let's compute the Pearson correlation coefficient between X and Y variables :

X	1	2	3	4	5
Y	10	20	30	40	50

X	Y	X^2	Y^2	XY
1	10	1	100	10
2	20	4	400	40
3	30	9	900	90
4	40	16	1600	160
5	50	25	2500	250
$\Sigma X = 15$	$\Sigma Y = 150$	$\Sigma X^2 = 55$	$\Sigma Y^2 = 5500$	$\Sigma XY = 550$

- Substituting the relevant values in formula we get :

$$r = \frac{5\times 550 - 15\times 150}{\sqrt{5\times 55-(15)^2}\times\sqrt{5\times 5500-(150)^2}}$$

$$\Rightarrow \quad r = \frac{2750-2250}{\sqrt{50}\times\sqrt{5000}} = 1$$

Hence there is perfect positive correlation between X and Y
i.e. If X increases, Y also increases and vice versa

Spearman

Sử dụng tương quan hạng Spearman để kiểm tra mối quan hệ giữa hai biến được xếp hạng hoặc một biến được xếp hạng và một biến đo lường. Có thể sử dụng tương quan hạng Spearman thay cho hồi quy/tương quan Pearson khi lo lắng về phân phối không chuẩn của dữ liệu. Tuy nhiên, điều này không phải thật luôn cần thiết.

Tương quan hạng Spearman được thực hiện dựa trên các giả định sau:

- Biến kiểm định có thể là dạng thứ tự, tỉ lệ, khoảng và có phân phối bất kì.
- Các biến phải thỏa mãn tính chất đơn điệu (monotonics). Đây là giả định quan trọng của tương quan hạng Spearman.

Spearman's

The difference between a pair of scores

$$r_s = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

The number of pairs of ranks

The number of pairs of ranks

	Marks									
English	56	75	45	71	62	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

Where d = difference between ranks and d^2 = difference squared.

We then calculate the following:

$$\sum d_i^2 = 25 + 1 + 9 + 1 + 16 + 1 + 1 = 54$$

We then substitute this into the main equation with the other information as follows:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2 - 1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

as n = 10. Hence, we have a ρ (or r_s) of 0.67. This indicates a strong positive relationship between the ranks individuals obtained in the maths and English exam. That is, the higher you ranked in maths, the higher you ranked in English also, and vice versa.

Point Biserial Correlation

Mean of Group 2

Mean of Group 1

Number of scores in Group 1

Number of scores in Group 2

$$r_{pb} = \frac{M_1 - M_2}{s_n}\sqrt{\frac{n_1 n_2}{n(n-1)}}$$

Standard deviation of the entire sample

Total number of scores in both groups

Phi Coefficient

Mean of Group 2

Mean of Group 1

Number of scores in Group 1

Number of scores in Group 2

$$r_{pb} = \frac{M_1 - M_2}{\sigma_n}\sqrt{\frac{n_1 n_2}{n^2}}$$

Standard deviation of the entire population

Total number of scores