

Nhập môn phân tích dữ liệu

Instructor: Đặng Hải Đăng

Mục tiêu khóa học

- ▶ Giúp người học hiểu được bản chất và tầm quan trọng của Khoa học dữ liệu trong công việc và cuộc sống.
- ▶ Trang bị cho người học các công cụ cơ bản trên Excel sử dụng trong Khoa học dữ liệu.
- ▶ Định hướng người học tìm tòi, phân tích, xử lý dữ liệu trong chính lĩnh vực công tác của mình.

Nội dung giảng dạy

- ▶ Phần 1. Giới thiệu về Khoa học Dữ liệu (1 tiết)
- ▶ Phần 2. Khám phá dữ liệu với Excel
 - Tiết 2: Khám phá dữ liệu
 - Tiết 3: Trực quan hóa bằng biểu đồ
 - Tiết 4: Các biểu đồ sử dụng trong KHDL
 - Tiết 5: Thực hành khám phá dữ liệu
- ▶ Phần 3. Nhập môn về Thống kê
 - Tiết 6: Thống kê mô tả
 - Tiết 7: Thống kê kết hợp
 - Tiết 8: Thực hành với thống kê

Tiết 1. Giới thiệu về Khoa học Dữ liệu

- ▶ Khoa học Dữ liệu là một nhánh con của Khoa học Máy tính
- ▶ Khoa học Dữ liệu sẽ là một chức danh nghề nghiệp phụ trong tương lai
- ▶ Khoa học dữ liệu tập trung vào phân tích, xử lý các dữ liệu dưới dạng con người đọc được và chuyển đổi thành các thông tin có ý nghĩa, giúp cho quá trình ra quyết định của con người.
- ▶ Ai cũng có thể trở thành Chuyên gia, Nhà Khoa học Dữ liệu trong lĩnh vực chuyên môn hẹp của mình

Phần 1. Giới thiệu về Khoa học Dữ liệu

- ▶ Khoa học Dữ liệu là một nhánh con của Khoa học Máy tính
- ▶ Khoa học Dữ liệu sẽ là một chức danh nghề nghiệp phụ trong tương lai
- ▶ Khoa học dữ liệu tập trung vào phân tích, xử lý các dữ liệu dưới dạng con người đọc được và chuyển đổi thành các thông tin có ý nghĩa, giúp cho quá trình ra quyết định của con người.
- ▶ Ai cũng có thể trở thành Chuyên gia, Nhà Khoa học Dữ liệu trong lĩnh vực chuyên môn hẹp của mình

Các nguồn sinh dữ liệu

- ▶ Nguồn dữ liệu do cá nhân sinh ra: Là các số liệu do cơ thể, hoạt động cá nhân, hoạt động khi tham gia làm việc, giao tiếp xã hội
- ▶ Nguồn dữ liệu do tổ chức sinh ra: Là nguồn dữ liệu của tổ chức sinh ra phục vụ mục đích của tổ chức ấy
- ▶ Nguồn dữ liệu do hệ thống sinh ra: Là nguồn dữ liệu do hệ thống tự động sinh ra trong quá trình hoạt động của máy móc

Dữ liệu do cá nhân phát sinh

Thời gian	Các Hoạt động	Các loại dữ liệu được sinh ra
5 giờ 30	Tỉnh dậy, tắt chuông điện thoại, xem các thông tin về sức khỏe trên smart watch. Đọc và trả lời email, tương tác với mạng xã hội	Thông tin về thời gian ngủ, lượng bước chân, số calo tiêu thụ trong ngày. Thông tin gửi và nhận trên các nền tảng mạng xã hội
6 giờ	Tập thể dục	Thông tin lượng calo đốt cháy
6 giờ 45	Lái xe đến chỗ làm việc	Thông tin GPS trên xe về quãng đường di chuyển, camera hành trình. Thời điểm vào ra điểm kiểm soát vé của căn hộ của Tom / công ty của Tom
9 – 12 giờ	Làm các công việc văn phòng trên bộ công cụ Microsoft Office, gửi và nhận email, sử dụng các phần mềm nghiệp vụ cho nhân viên	Dữ liệu check-in time, dữ liệu từ bộ công cụ Microsoft Office, dữ liệu phát sinh trên phần mềm nghiệp vụ
12 giờ	Đi ăn trưa, quẹt thẻ credit card, thẻ thành viên thân thiết của quán ăn	Dữ liệu thông tin chuyển khoản, dữ liệu thông tin trên thẻ khách hàng thân thiết
13 giờ	Chơi games giải trí tại công ty	Các thông tin dữ liệu người chơi trên hệ thống
14 giờ - 17 giờ	Tiếp tục làm việc	Các dữ liệu mới
1715 – 19 giờ	Lái xe trở về nhà, đi siêu thị gần nhà	Ngoài dữ liệu sinh ra của xe hơi, dữ liệu tài khoản ngân hàng, dữ liệu mua sắm được cập nhật thêm trên hệ thống của siêu thị
20 giờ	Xem tivi, Netflix, lướt internet, mạng xã hội	Các chương trình truyền hình phát sinh dữ liệu, hệ thống lưu lại dữ liệu về lịch sử xem phim, dữ liệu duyệt web, mạng xã hội

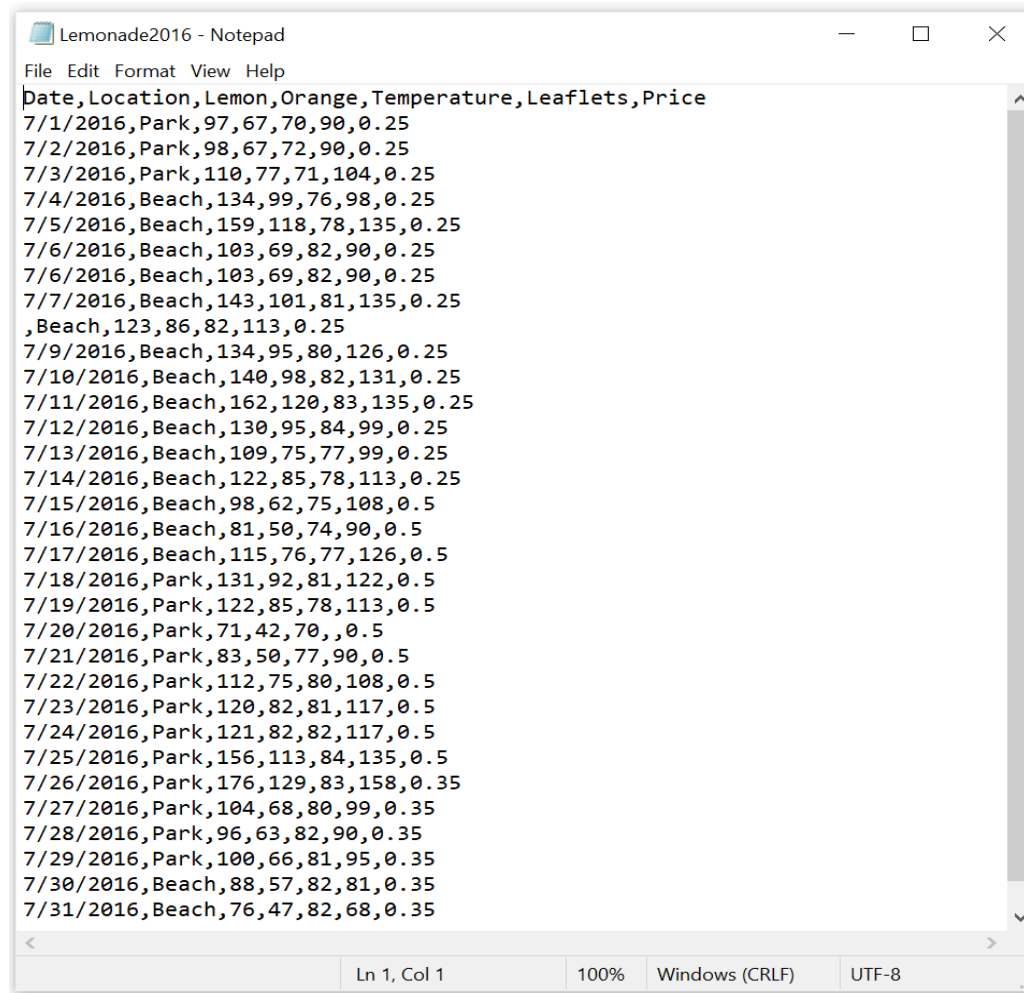
Dữ liệu dẫn xuất

No	Name	Weight (kg)	Height (cm)
1	Alice	58	161
2	Bob	63	189
3	Crist	99	168
4	Dave	81	183
5	Emmy	84	167

ABI
22.7
17.8
35.2
24.3
29.9

Category
Normal
Underweight
Obese
Normal
Overweight

Mở 1 file dữ liệu dạng text



```
Lemonade2016 - Notepad
File Edit Format View Help
Date,Location,Lemon,Orange,Temperature,Leaflets,Price
7/1/2016,Park,97,67,70,90,0.25
7/2/2016,Park,98,67,72,90,0.25
7/3/2016,Park,110,77,71,104,0.25
7/4/2016,Beach,134,99,76,98,0.25
7/5/2016,Beach,159,118,78,135,0.25
7/6/2016,Beach,103,69,82,90,0.25
7/6/2016,Beach,103,69,82,90,0.25
7/7/2016,Beach,143,101,81,135,0.25
,Beach,123,86,82,113,0.25
7/9/2016,Beach,134,95,80,126,0.25
7/10/2016,Beach,140,98,82,131,0.25
7/11/2016,Beach,162,120,83,135,0.25
7/12/2016,Beach,130,95,84,99,0.25
7/13/2016,Beach,109,75,77,99,0.25
7/14/2016,Beach,122,85,78,113,0.25
7/15/2016,Beach,98,62,75,108,0.5
7/16/2016,Beach,81,50,74,90,0.5
7/17/2016,Beach,115,76,77,126,0.5
7/18/2016,Park,131,92,81,122,0.5
7/19/2016,Park,122,85,78,113,0.5
7/20/2016,Park,71,42,70,,0.5
7/21/2016,Park,83,50,77,90,0.5
7/22/2016,Park,112,75,80,108,0.5
7/23/2016,Park,120,82,81,117,0.5
7/24/2016,Park,121,82,82,117,0.5
7/25/2016,Park,156,113,84,135,0.5
7/26/2016,Park,176,129,83,158,0.35
7/27/2016,Park,104,68,80,99,0.35
7/28/2016,Park,96,63,82,90,0.35
7/29/2016,Park,100,66,81,95,0.35
7/30/2016,Beach,88,57,82,81,0.35
7/31/2016,Beach,76,47,82,68,0.35
Ln 1, Col 1 100% Windows (CRLF) UTF-8
```

Mở 1 file dữ liệu trên Excel

The screenshot displays an Excel spreadsheet titled "Lemonade2016 - Excel". The spreadsheet contains a table with the following data:

Date	Location	Lemon	Orange	Temperature	Leaflets	Price
7/1/2016	Park	97	67	70	90	0.25
7/2/2016	Park	98	67	72	90	0.25
7/3/2016	Park	110	77	71	104	0.25
7/4/2016	Beach	134	99	76	98	0.25
7/5/2016	Beach	159	118	78	135	0.25
7/6/2016	Beach	103	69	82	90	0.25
7/6/2016	Beach	103	69	82	90	0.25
7/7/2016	Beach	143	101	81	135	0.25
7/7/2016	Beach	123	86	82	113	0.25
7/9/2016	Beach	134	95	80	126	0.25
7/10/2016	Beach	140	98	82	131	0.25
7/11/2016	Beach	162	120	83	135	0.25
7/12/2016	Beach	130	95	84	99	0.25
7/13/2016	Beach	109	75	77	99	0.25
7/14/2016	Beach	122	85	78	113	0.25
7/15/2016	Beach	98	62	75	108	0.5
7/16/2016	Beach	81	50	74	90	0.5
7/17/2016	Beach	115	76	77	126	0.5
7/18/2016	Park	131	92	81	122	0.5
7/19/2016	Park	122	85	78	113	0.5
7/20/2016	Park	71	42	70	0.5	
7/21/2016	Park	83	50	77	90	0.5
7/22/2016	Park	112	75	80	108	0.5
7/23/2016	Park	120	82	81	117	0.5
7/24/2016	Park	121	82	82	117	0.5
7/25/2016	Park	156	113	84	135	0.5
7/26/2016	Park	176	129	83	158	0.35
7/27/2016	Park	104	68	80	99	0.35
7/28/2016	Park	96	63	82	90	0.35
7/29/2016	Park	100	66	81	95	0.35
7/30/2016	Beach	88	57	82	81	0.35
7/31/2016	Beach	76	47	82	68	0.35

Kho dữ liệu mở dành cho KHDL

- ▶ Có rất nhiều kho dữ liệu mở trên thế giới dành cho cộng đồng những nhà KHDL được liệt kê dưới đây:
- ▶ Data.gov - Trang dữ liệu mở của chính phủ Hoa Kỳ
- ▶ Data.gov.in - Trang dữ liệu mở của chính phủ Ấn Độ
- ▶ Data.worldbank.org - Trang dữ liệu mở của Ngân hàng thế giới
- ▶ Data.world – Dữ liệu cho các nhà báo, nhà kinh doanh và nhiều đối tượng khác
- ▶ Kaggle.com – Trang dữ liệu chia sẻ được nhiều nhà KHDL sử dụng nhất

Một số thao tác cơ bản với dữ liệu

- ▶ Sắp xếp
- ▶ Lọc dữ liệu
- ▶ Thao tác với dữ liệu không hoàn chỉnh
- ▶ Dữ liệu sinh mới
- ▶ Nội suy dữ liệu
- ▶ Làm sạch dữ liệu

Tiết 2. Khám phá dữ liệu

► Kịch bản dữ liệu:

- Rosie là một học sinh cấp hai. Cô ấy đã dành kỳ nghỉ hè của mình, cố gắng tìm cách kiếm tiền, và Rosie đã chọn làm một công việc tại quầy bán đồ uống. Rosie làm đồ uống để bán cho mọi người.
- Bởi vì là một người khá thông minh, nhạy bén, cô ấy biết rằng nếu cô ấy tận dụng dữ liệu về doanh số bán đồ uống của mình thì cô ấy có thể sẽ thành công hơn trong tương lai.
- Vậy nên, Rosie cẩn thận ghi lại tất cả dữ liệu liên quan đến việc bán đồ uống và lưu trữ dữ liệu đó trong bảng tính Excel dưới dạng tệp csv. Chúng ta sẽ cùng tìm hiểu và phân tích cách Rosie làm việc với dữ liệu để hiểu một số khái niệm cơ bản về khám phá dữ liệu.

Các loại câu hỏi về dữ liệu

- ▶ Tôi đã bán được bao nhiêu ly nước chanh hoặc doanh thu của tôi là bao nhiêu? Đây là loại thông tin mô tả về dữ liệu
- ▶ Tiếp đó cô ấy muốn xem các loại dữ liệu có tính liên kết hay không, có quan hệ nguyên nhân - kết quả nào giữa các dữ liệu hay không như:
 - ▶ Yếu tố nhiệt độ / số lượng tờ rơi phát ra có ảnh hưởng đến doanh số bán hàng hay không?
- ▶ Ngoài ra cô ấy còn có các câu hỏi loại so sánh như:
 - ▶ Doanh số nước chanh và nước cam có chênh lệch hay không và mức chênh lệch là bao nhiêu?
 - ▶ Doanh thu giữa công viên và bãi biển có chênh lệch hay không và nơi nào bán hàng nhiều hơn?
- ▶ Cuối cùng là loại câu hỏi mang tính chất dự đoán với dữ liệu về doanh số bán hàng trước đó liệu Rosie có thể dự đoán sẽ bán được bao nhiêu trong những ngày kế tiếp hay không.

Một số hàm thông dụng trong Excel

- ▶ Hàm tổng hợp: count, sum, average, min, max..v....v..v
 - ▶ Công thức = [cell \$X\$i] operand [cell \$Y\$j]
 - ▶ Công thức = function_name(danh sách tham số)
- ▶ Minh họa (Làm sạch dữ liệu & tính toán):
 - ▶ Count:
 - ▶ Lọc trùng
 - ▶ Điền khuyết thiếu
 - ▶ =COUNT([Date])

Hiển thị trực quan trên bảng tính Excel

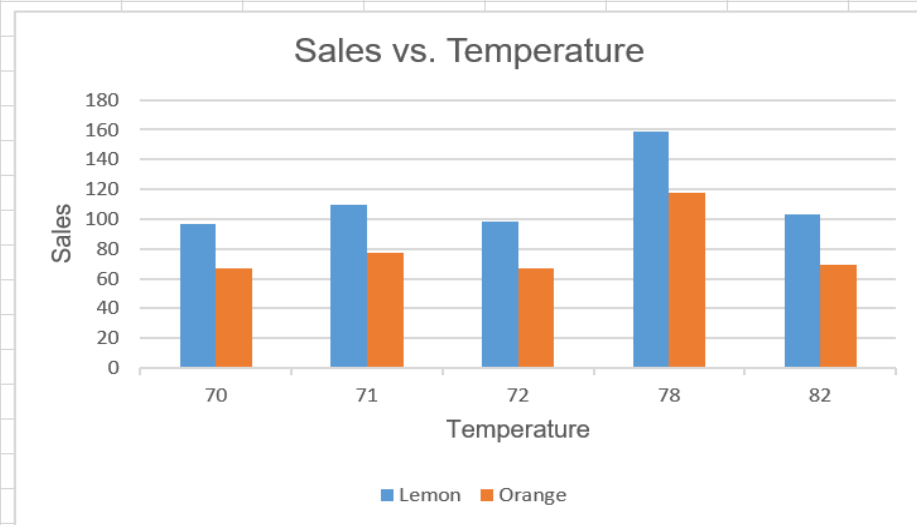
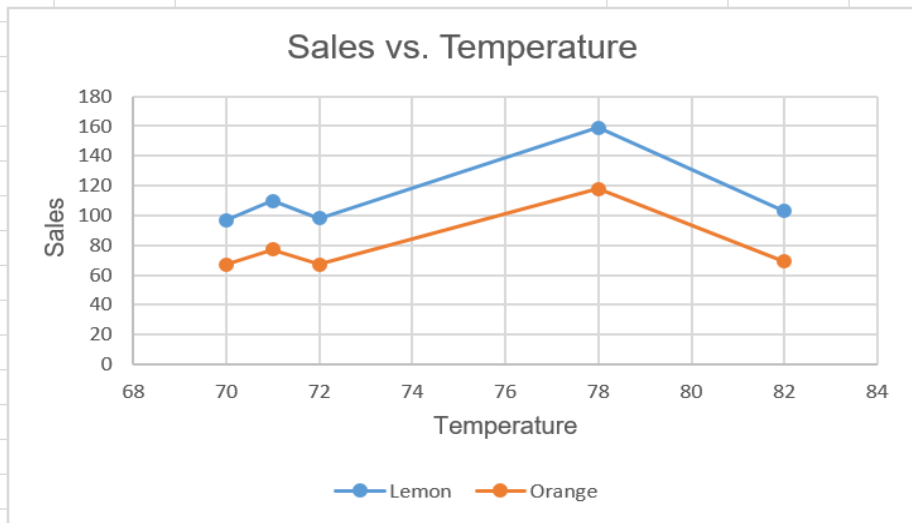
Thứ tự	Tên định dạng	Cột dữ liệu sử dụng	Thao tác
1	Data Bars	Revenue	Home → Conditional Formating → Gradient Field
2	Color Scale	Temperature	Home → Conditional Formating → Color Scales
3	Icon Sets	Leaflets	Home → Conditional Formating → Icon Sets → Rating (Star icon)
4	Top/Bottom Rules	Sales	Home → Conditional Formating → Top/Bottom Rules → Top 10% / Bottom 10%

Tiết 3. Trực quan hóa bằng biểu đồ

- ❑ Khi vẽ bất cứ biểu đồ nào, chúng ta phải thực hiện các bước sau:
 - ▶ Lựa chọn các cột dữ liệu trên bảng để thực hiện biểu diễn (có thể thêm bớt các cột dữ liệu đưa vào biểu đồ sau)
 - ▶ Chọn loại biểu đồ cần biểu diễn
 - ▶ Tạo loại biểu đồ được lựa chọn và tùy biến hiển thị
 - ▶ Thêm chú thích cho rõ ràng

Biểu đồ doanh số bán hàng theo nhiệt độ

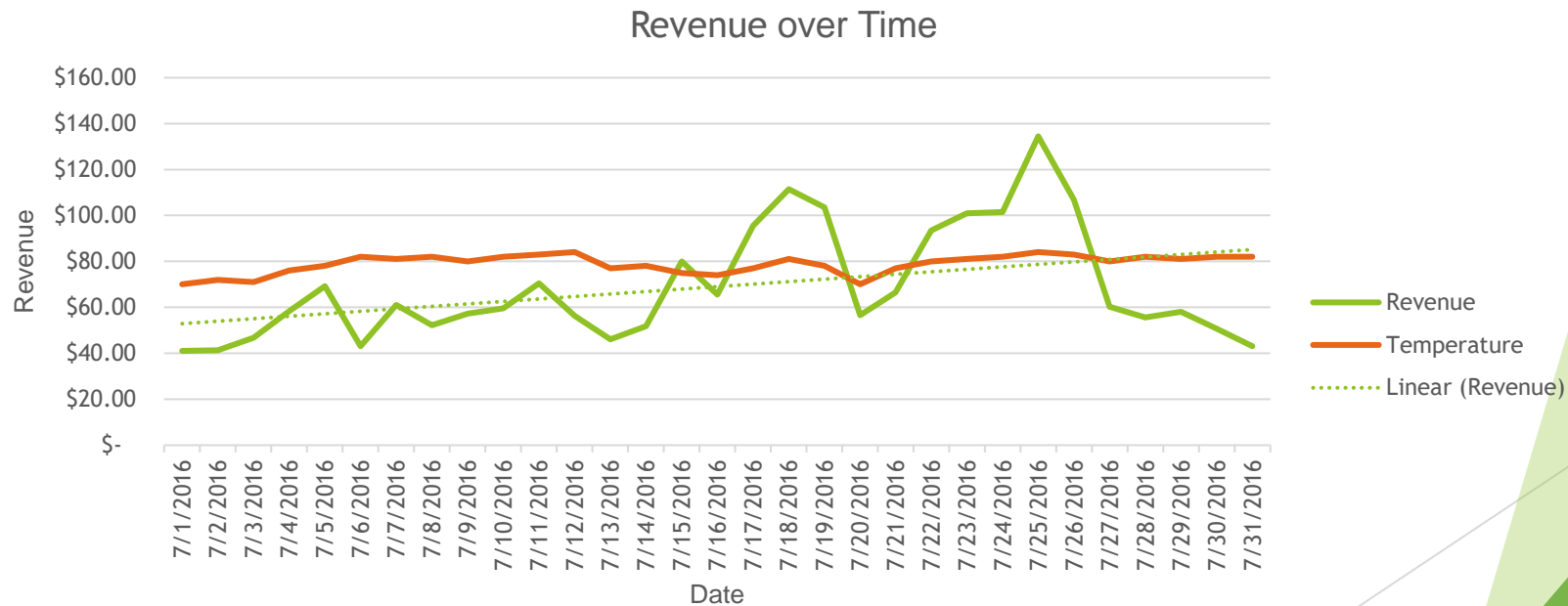
Temperature	Lemon	Orange
70	97	67
71	110	77
72	98	67
78	159	118
82	103	69



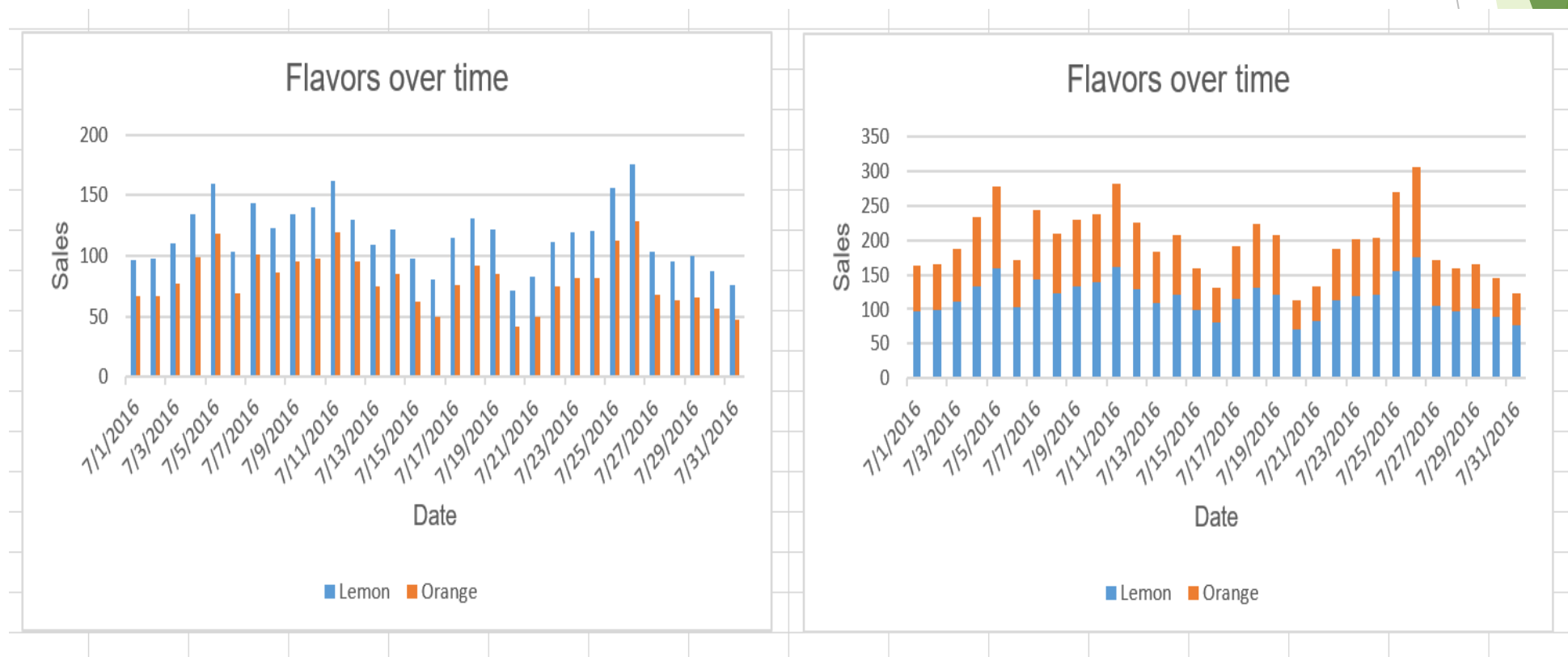
Minh họa biểu đồ bằng đường xu hướng



Thêm đường nhiệt độ vào biểu đồ doanh thu theo thời gian

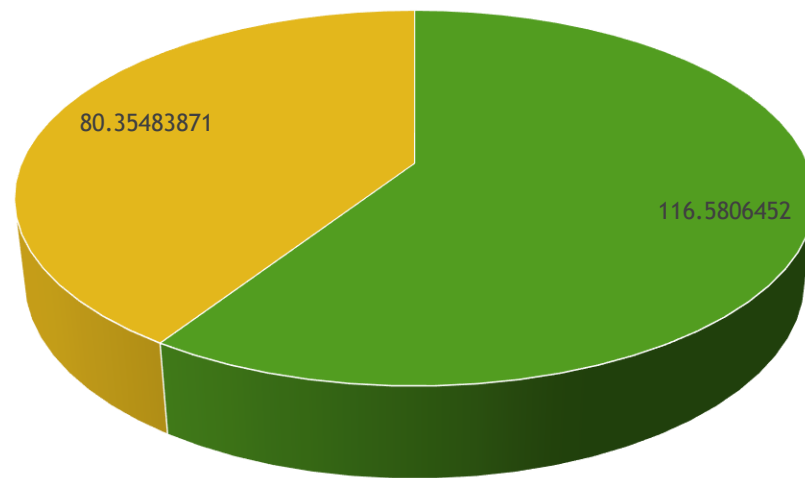


Minh họa với biểu đồ liên cụm, cột xếp chồng



Minh họa với biểu đồ tròn

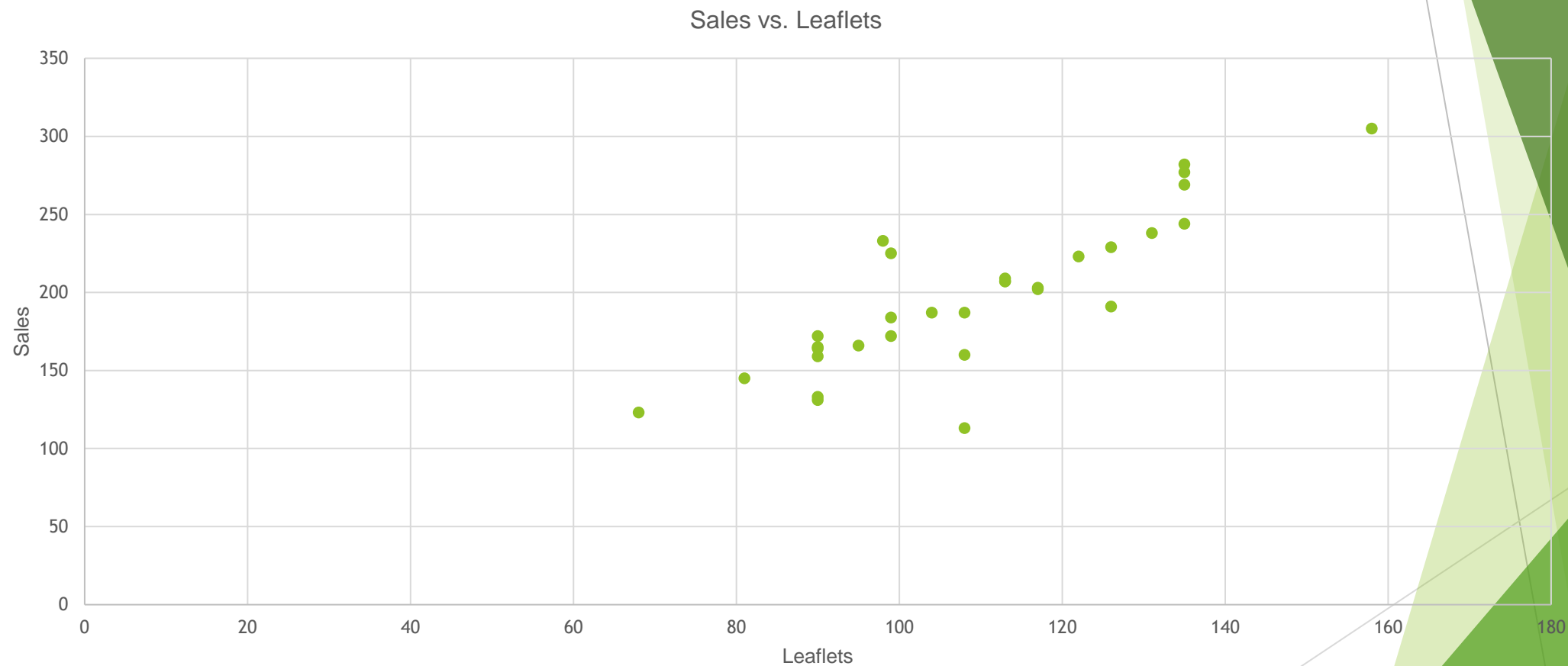
Flavor average sales



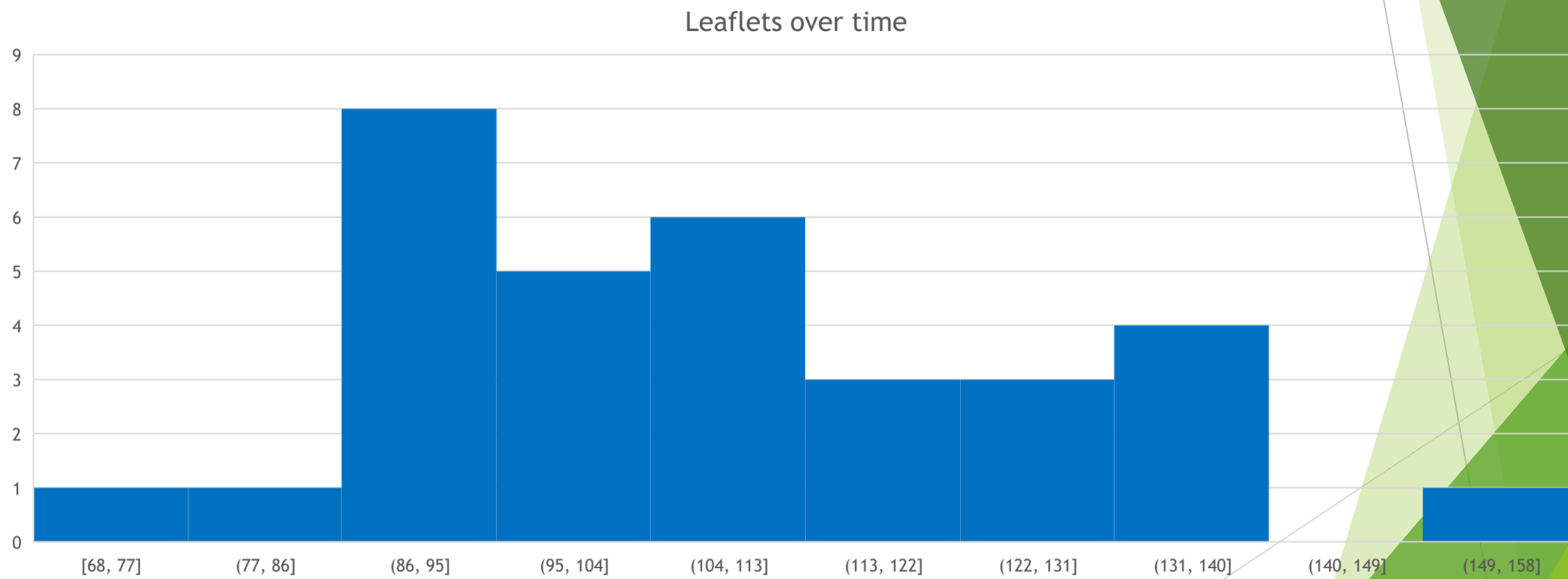
■ Lemon

■ Orange

Minh họa với biểu đồ phân tán

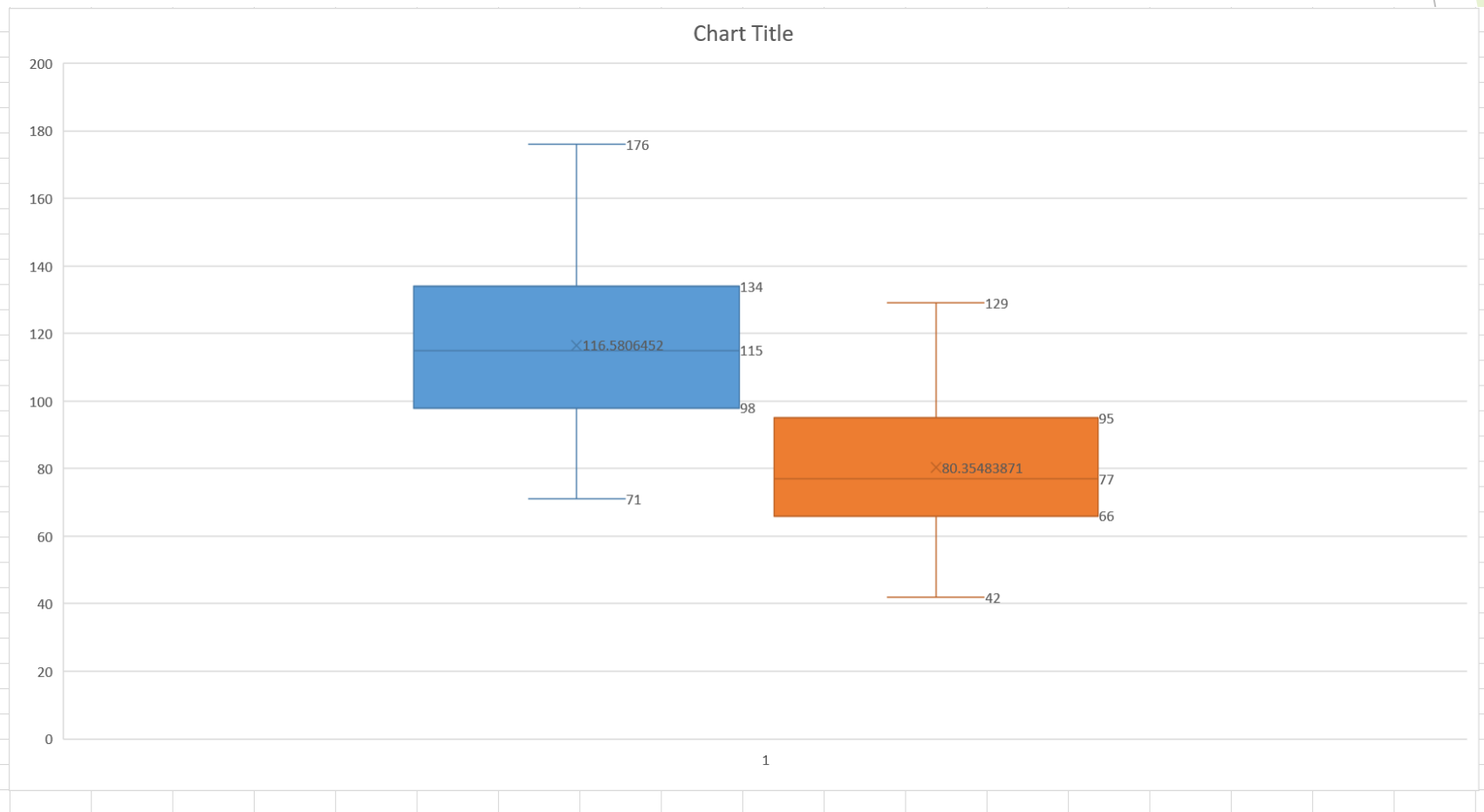


Tiết 4. Một số biểu đồ sử dụng phổ biến trong KHDL



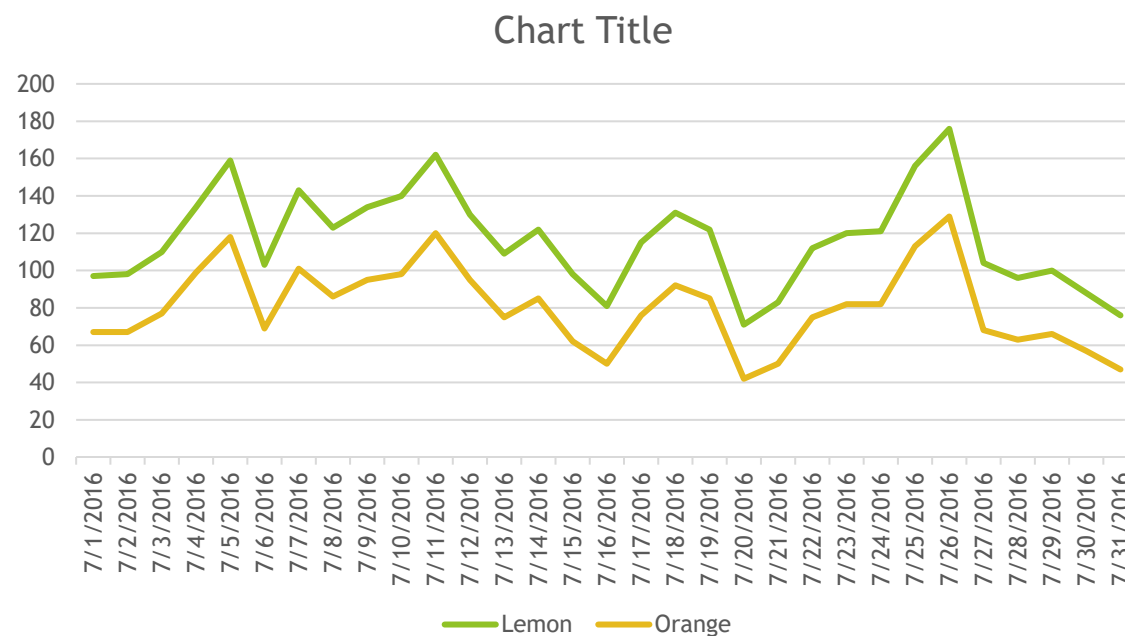
Biểu đồ hộp và dải dữ liệu

Lemon	Orange
97	67
98	67
110	77
134	99
159	118
103	69
143	101
123	86
134	95
140	98
162	120
130	95
109	75
122	85
98	62
81	50
115	76
131	92
122	85
71	42
83	50
112	75
120	82
121	82
156	113
176	129
104	68
96	63
100	66
88	57
76	47



Biểu đồ 3 chiều

- ▶ Biểu đồ 2 chiều: Xét doanh số bán hàng của Lemon và Orange theo Day
- ▶ Biểu đồ 3 chiều: Xét doanh số bán hàng của Lemon và Orange theo Day và Place

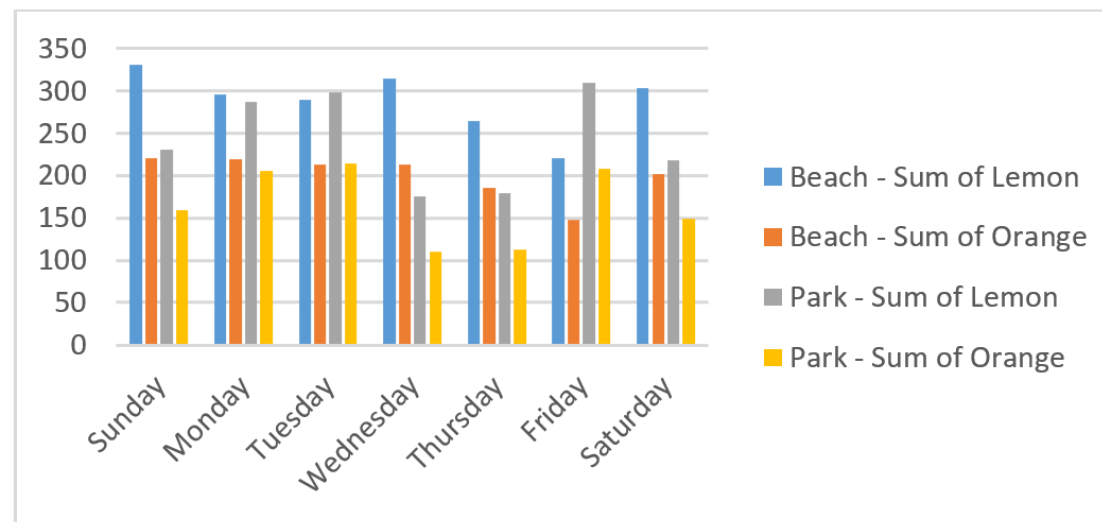


Pivot Table

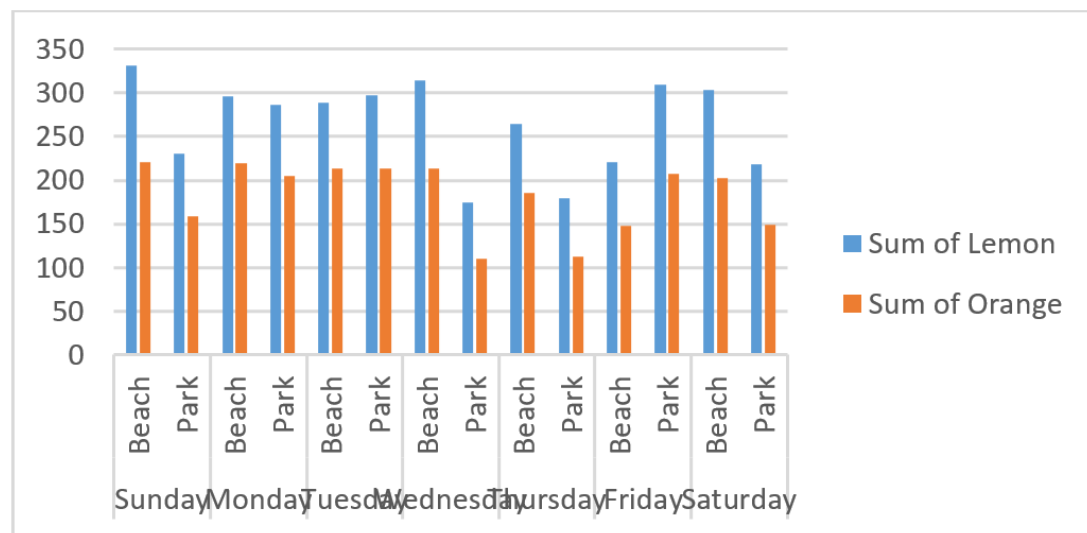
Row Labels	Sum of Lemon	Sum of Orange
Sunday	562	380
Beach	331	221
Park	231	159
Monday	583	424
Beach	296	219
Park	287	205
Tuesday	587	427
Beach	289	213
Park	298	214
Wednesday	490	323
Beach	315	213
Park	175	110
Thursday	444	299
Beach	265	186
Park	179	113
Friday	530	356
Beach	221	148
Park	309	208
Saturday	521	351
Beach	303	202
Park	218	149
Grand Total	3717	2560

Pivot Table sau khi chuyển Location sang vị trí cột

	Column Labels					
	Beach		Park		Total Sum of Lemon	Total Sum of Orange
Row Labels	Sum of Lemon	Sum of Orange	Sum of Lemon	Sum of Orange		
Sunday	331	221	231	159	562	380
Monday	296	219	287	205	583	424
Tuesday	289	213	298	214	587	427
Wednesday	315	213	175	110	490	323
Thursday	265	186	179	113	444	299
Friday	221	148	309	208	530	356
Saturday	303	202	218	149	521	351
Grand Total	2020	1402	1697	1158	3717	2560



Hình 4.3 Pivot chart với Location chuyển sang vị trí cột



Hình 4.4 Pivot chart với Location chuyển trở lại vị trí hàng

Tiết 5. Thực hành khám phá dữ liệu – Thao tác cơ bản

- ▶ Mở file Lab
- ▶ Thao tác định dạng
- ▶ Tính tổng số ngày
- ▶ Lọc dữ liệu trùng
- ▶ Tính toán dữ liệu sinh mới
- ▶ Một số hàm cơ bản trong Excel

Tiết 5. Thực hành khám phá dữ liệu – Trực quan hóa (1)

► Trực quan với bảng tính

Thứ tự	Tên định dạng	Cột dữ liệu sử dụng	Thao tác
1	Data Bars	Revenue	Home → Conditional Formatting → Gradient Field
2	Color Scale	Temperature	Home → Conditional Formatting → Color Scales
3	Icon Sets	Leaflets	Home → Conditional Formatting → Icon Sets → Rating (Star icon)
4	Top/Bottom Rules	Sales	Home → Conditional Formatting → Top/Bottom Rules → Top 10% / Bottom 10%

Tiết 5. Thực hành khám phá dữ liệu – Trực quan hóa (2)

► Trực quan với biểu đồ

Thứ tự	Biểu đồ	Cột dữ liệu sử dụng	Thao tác
1	Line	Date, Temperature, Revenue	Insert → 2D-Line → Thêm tiêu đề, nhãn, trend line
2	Clustered/Staked Column	Orange, Lemon	Insert → Clustered Column / Staked Column → Thêm tiêu đề, nhãn
3	3D Pie chart	Average(Lemon), Average(Orange)	Insert → 3D Pie chart Design → Switch Row/Column, thêm tiêu đề
4	Scatter	Sales, Leaflets	Insert → Statistic Chart → Scatter (biểu đồ đầu tiên), thêm tiêu đề, nhãn
5	Histogram	Leaflets	Insert → Histogram → Xác định số bin, thêm tiêu đề, nhãn trục
6	Box and Whisker	Orange, Lemon	Insert → Statistic Chart → Box and Whisker, Thêm tiêu đề, nhãn

Tiết 5. Thực hành khám phá dữ liệu – Pivot Chart/Table

- ▶ Trực quan với biểu đồ

Tiết 6. Thống kê mô tả

- ▶ Các loại biến số trong thống kê
- ▶ Tổng thể và mẫu
- ▶ Đo lường xu hướng tập trung
 - Trung bình, trung vị và yếu vị
 - Phân phối chuẩn, lệch trái, lệch phải
- ▶ Tính biến thiên của dữ liệu
 - Phạm vi
 - Phương sai
 - Độ lệch chuẩn
 - Sai số chuẩn

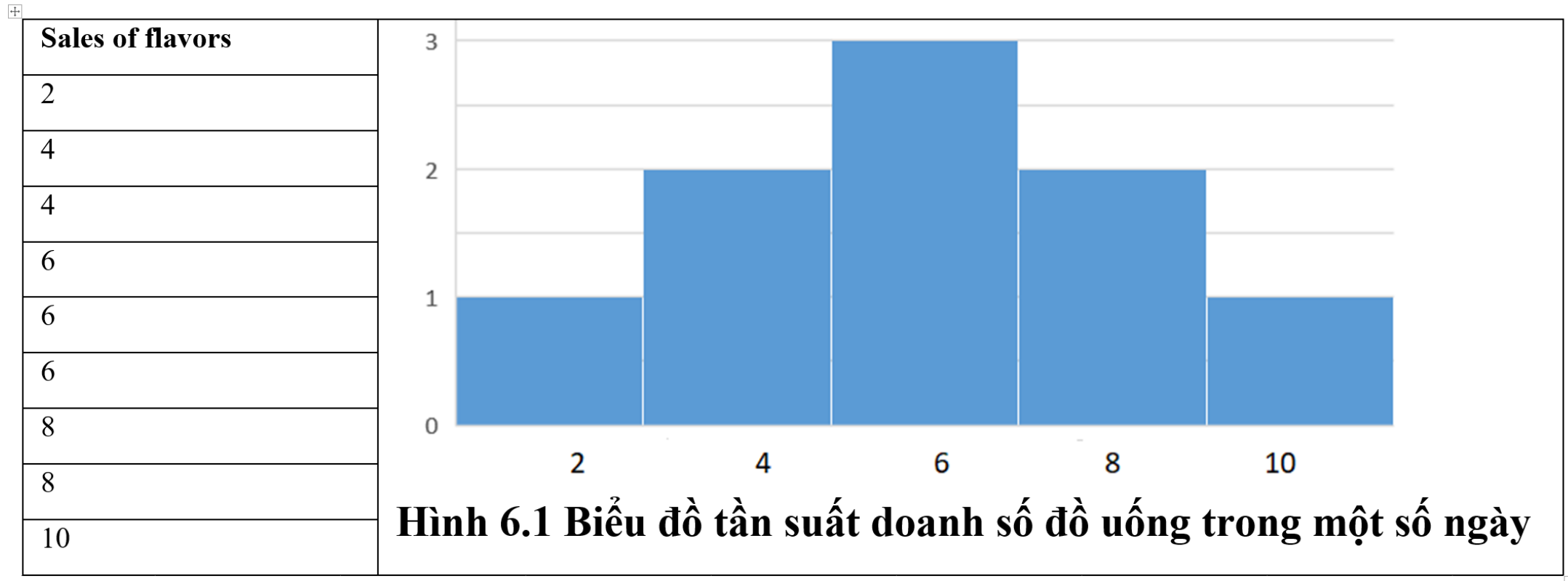
3 loại biến số trong thống kê

- ▶ **Biến liên tục:** Đó là các con số liên tục nằm trong một phạm vi và thường chúng ta sẽ tính được. Ví dụ nhiệt độ mỗi ngày mà Rosie bán nước chanh, nó thay đổi mỗi ngày nhưng nằm trong khoảng giá trị (khoảng giá trị của nhiệt độ mùa hè). Những biến nào có liên quan tới yếu tố thời gian thường là biến liên tục
- ▶ **Biến rời rạc:** Đây là những số nguyên rời rạc, riêng lẻ và thường chúng ta đếm thay vì tính. Ví dụ lượng tờ rơi mà Rosie phát ra mỗi ngày.
- ▶ **Biến số phân loại:** Thực ra là cách dán nhãn cho dữ liệu ví dụ Rosie bán ở công viên và bãi biển, chúng ta có thể hiểu rằng Park là biểu diễn số 1 và Beach là biểu diễn của số 2. Đây là cách con người chúng ta dán nhãn để dễ quản lý.

Tổng thể và mẫu

- ▶ Khi chúng ta thực hiện các phép thống kê chúng ta đi thu thập dữ liệu. Trong nhiều trường hợp chúng ta không đủ nguồn lực để lấy hết dữ liệu (do lượng dữ liệu khổng lồ, mất nhiều thời gian, công sức).
- ▶ Lúc đó chúng ta sẽ lấy những mẫu đại diện, đó là cách làm việc với thống kê và cũng là cách mà các nhà KHDL làm việc với dữ liệu. Mẫu dữ liệu lúc này chỉ là tập con của tổng thể. Dữ liệu tổng thể được ký hiệu X (Các dữ liệu trong tổng thể sẽ là X_1, X_2, \dots, X_N), dữ liệu con được ký hiệu là x (Các dữ liệu trong mẫu sẽ là x_1, x_2, \dots, x_n)

Đo lường xu hướng tập trung



Trung bình, trung vị và yếu vị

Công thức tính trung bình tổng thể: $\mu = \frac{\sum_{i=1}^N X_i}{N}$ và trung bình mẫu: $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$

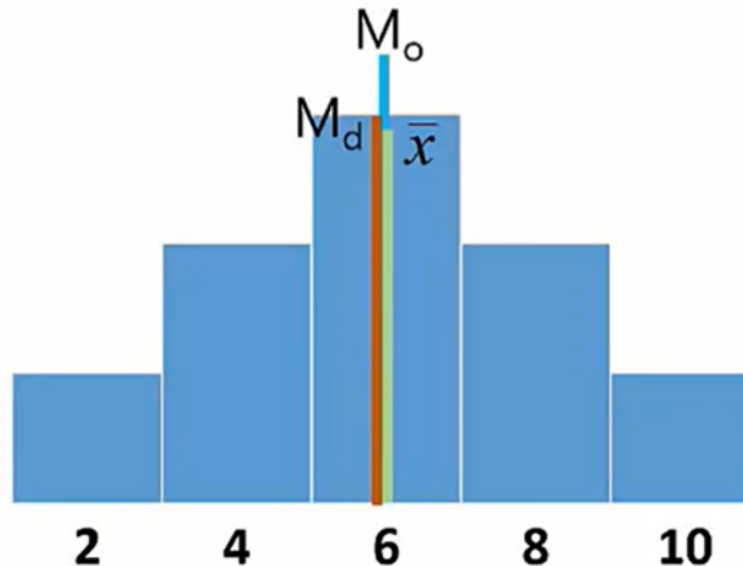
Trong trường hợp này ta tính được trung bình là 6.

Tiếp đến chúng ta tính đến giá trị trung vị (giá trị ở giữa nhất sau khi dữ liệu đã sắp xếp) theo công thức sau:

$Md = \frac{n+1}{2}$ với n là tổng số quan sát, số trung vị nằm ở vị trí thứ 5 và có giá trị là 6. Giá trị yếu vị được định nghĩa là giá trị xuất hiện nhiều nhất, trong trường hợp này yếu vị cũng là 6.

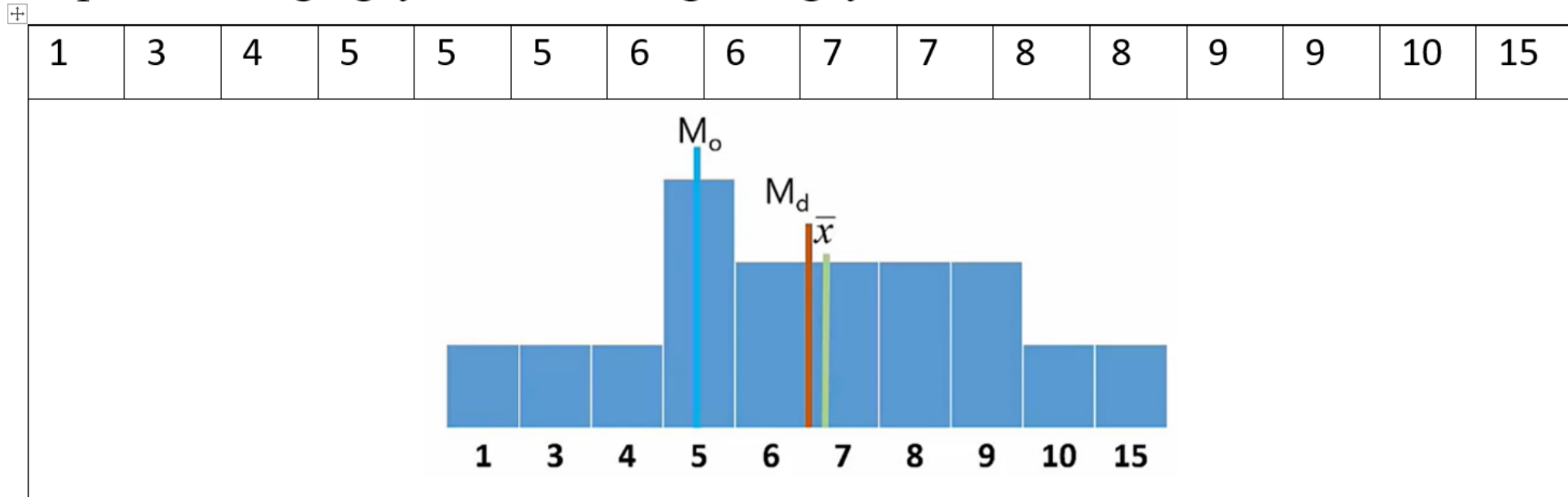
Phân phối chuẩn

Quan sát biểu đồ bên dưới ta thấy trông như ngẫu nhiên mà các giá trị trung bình, trung vị, yếu vị trùng nhau. Phân phối chúng ta quan sát được gọi là phân phối chuẩn, trong phân phối chuẩn thì hầu hết các giá trị bị hút về ở giá trị ở giữa và từ đó phân phối khá đồng đều sang hai bên:

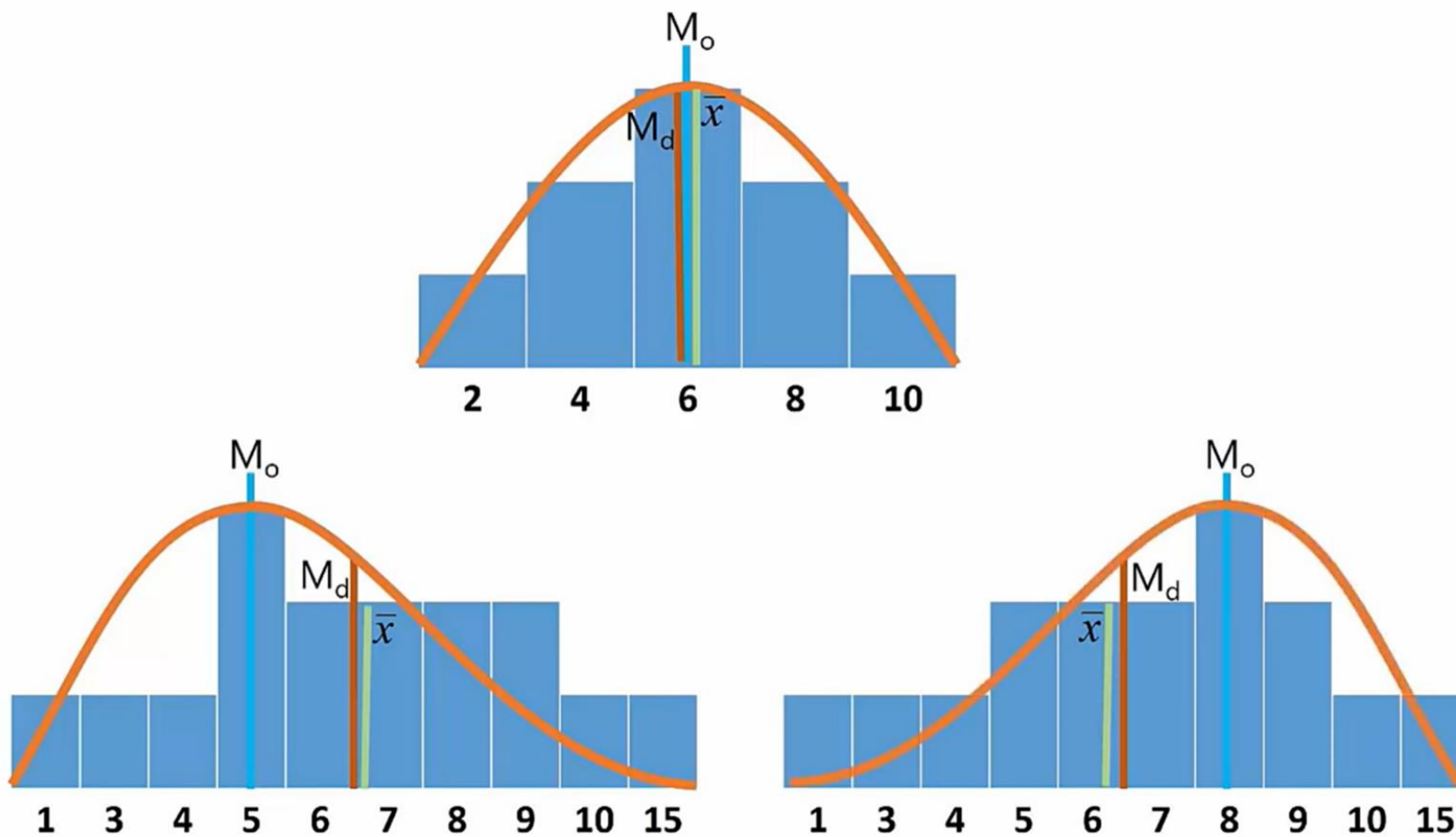


Phân phối lệch phải

Trên thực tế chúng ta luôn muốn tìm các giá trị trung bình, trung vị và yếu vị để tìm hiểu xem phân phối của chúng ta chuẩn đến đâu, trên thực tế thì không thể có phân phối chuẩn một cách hoàn hảo như hình trên. Hãy xem xét một ví dụ khác với các cột là số tờ rơi Sophie phát ra hàng ngày, liên tục trong 16 ngày.



Các dạng thức phân phối dữ liệu



Tính biến thiên của dữ liệu

- ▶ Phạm vi giá trị
- ▶ Phương sai
- ▶ Độ lệch chuẩn
- ▶ Sai số chuẩn

Phạm vi giá trị - Phương sai

- ▶ Cách tốt nhất để theo dõi độ phân tán của nó là theo dõi phạm vi giá trị. Phạm vi giá trị đơn giản là lấy giá trị lớn nhất trừ đi giá trị nhỏ nhất
- ▶ Phương sai:

Công thức tính phương sai - hay độ biến thiên của dữ liệu như sau:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Đây là công thức tính phương sai của tổng thể. Công thức tính phương sai với mẫu có dạng như dưới đây (thay vì chia cho n thì chia cho n-1, ước lượng chệch)

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

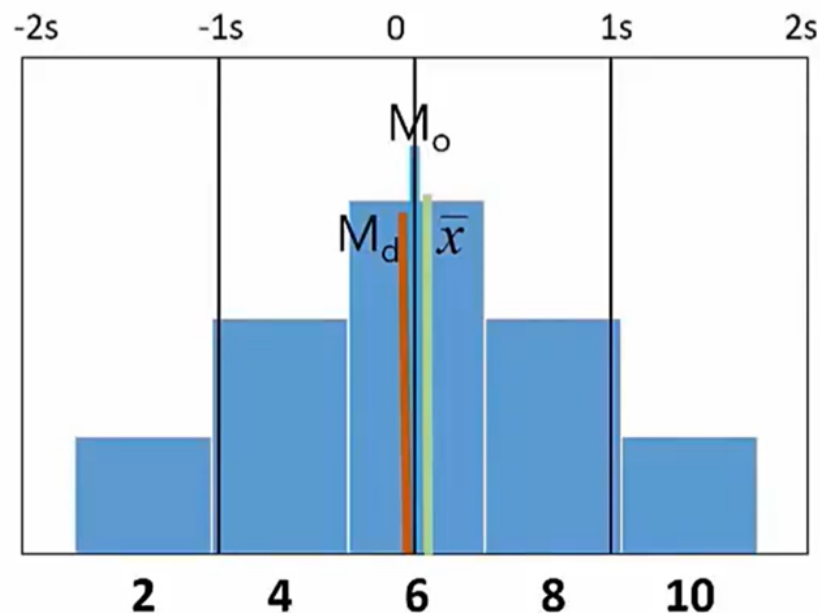
Áp dụng vào trường hợp cụ thể với ví dụ đầu tiết, ta có giá trị phương sai mẫu là:

$$s^2 = (6-2)^2 + (6-4)^2 + (6-4)^2 + (6-6)^2 + (6-6)^2 + (6-6)^2 + (8-6)^2 + (10-6)^2 / (9-1) = 6$$

Độ lệch chuẩn

$s = \sqrt{s^2} = 2.45$ được gọi là độ lệch chuẩn của mẫu.

Một điều thú vị với độ lệch chuẩn là nếu như bạn có một phân phối lệch chuẩn thì khi đặt nó vào một hệ thống chuẩn đo thì chúng ta sẽ hiểu được hệ thống chuẩn đo có ý nghĩa như thế nào. Chúng ta quay lại với ví dụ đầu tiết, với một phân phối chuẩn như thế này thì 68.2% dữ liệu sẽ rơi vào khoảng lệch chuẩn từ -1 đến +1, 94.5% rơi vào khoảng lệch chuẩn từ -2 đến +2 và 99.7% rơi vào khoảng lệch chuẩn từ -3 đến +3



Sai số chuẩn

- ▶ $SE = s / \sqrt{n} = 0.82$ được gọi là sai số chuẩn của mẫu và được sử dụng nhiều hơn là độ lệch chuẩn. Ví dụ bạn lấy rất nhiều mẫu trong một tổng thể và không phải lần nào cũng ra một số trung bình giống nhau. Sai số chuẩn cho chúng ta ước lượng được giá trị tính ra gần với số trung bình thực như thế nào.
- ▶

Tiết 7. Thống kê tổng hợp

- ▶ Hệ số tương quan
- ▶ Kiểm định giả thuyết
- ▶ T-Test/Z-Test
- ▶ T-Test Trung bình hai mẫu
- ▶ T-Test Cặp đôi
- ▶ Hồi quy

Hệ số tương quan

- ▶ Hệ số tương quan nằm trong khoảng -1 đến +1. Giá trị càng gần 1 thì các biến số càng liên kết với nhau. Nếu hệ số tương quan dương thì các biến số có mối quan hệ đồng biến, hệ số tương quan âm thì các biến số có mối quan hệ nghịch biến.
- ▶ Hãy tưởng tượng xem điều gì xảy ra khi Rosie tăng giá đồ uống. Chúng ta có thể thấy doanh số giảm xuống. Đó là ví dụ của tương quan nghịch biến. Giá tăng, doanh số giảm - lý thuyết kinh điển của kinh tế.
- ▶ Khi lượng tờ rơi được phát ra tăng lên, doanh số cũng tăng lên (nhiều người biết đến cửa hàng của Rosie hơn). Đây là ví dụ của tương quan đồng biến.

	<i>Lemon</i>	<i>Orange</i>	<i>Temperature</i>	<i>Leaflets</i>	<i>Price</i>	<i>Sale</i>	<i>Revenue</i>
Lemon	1						
Orange	0.996714	1					
Temperature	0.477345	0.453116	1				
Leaflets	0.90578	0.872198	0.31299028	1			
Price	-0.27053	-0.31808	-0.033574567	0.033303	1		
Sale	0.999309	0.999036	0.466616419	0.891188	-0.29257	1	
Revenue	0.469276	0.426955	0.339446501	0.603123	0.702259	0.450239	1

Kiểm định giả thuyết

- ▶ Rosie đã bán nước chanh được một thời gian và cô ấy muốn xem xét tại sao lại có sự khác nhau về lượng nước chanh được bán ra ở các tháng khác nhau. Thống kê liệu có giúp giải thích cho cô ấy được liệu doanh số hàng tháng chỉ là điều ngẫu nhiên hay có các yếu tố khác chi phối.
- ▶ Trên thực tế, chúng ta có 2 loại giả thuyết trong thống kê. Loại đầu tiên là “giả thuyết không” về cơ bản trong trường hợp này là không có khác biệt gì, sự chênh lệch doanh số bán hàng nước chanh chỉ là ngẫu nhiên.
- ▶ Loại giả thuyết thứ hai là “giả thuyết thay thế”: sự khác biệt về doanh số này là lớn hơn, hoặc nhỏ hơn hoặc đơn giản chỉ là có sự khác nhau ở đây.
- ▶ Giả thuyết thống kê chấp nhận trị số p vào khoảng 0.05 (tức là 5% trong phép thống kê được thực hiện có thể sai). Trong một số lĩnh vực khoa học con số p còn nhỏ hơn nữa ví dụ 0.01 hoặc thậm chí 0.001.

T-Test/Z-Test

- ▶ Đầu tiên là T-Test trung bình một mẫu. Hãy tưởng tượng chúng ta đã biết về doanh số bán hàng của Rosie trong 5 năm vừa qua và xác định xem doanh số đạt được trong năm nay có thực sự khác với những gì được dự đoán theo lịch sử bán hàng và chúng ta sẽ dùng loại thống kê nào. Giả sử 180 ly nước chanh là số lượng trung bình cô ấy bán được trong tháng 7 trong năm qua (lấy mẫu tháng 7 năm nay để so sánh). Công thức để tính như sau:
- ▶
$$t = \frac{x - \mu}{s / \sqrt{n}}$$
- ▶ Phần này ta đề cập thêm khái niệm bậc tự do $df = n - 1$ là số lượng các mục trong dữ liệu có thể thay đổi mà vẫn cho ra giá trị trung bình không đổi. Thực ra chúng ta đang thực hiện kiểm định Z-Test thay vì T-Test vì chúng ta đang giả định chúng ta đã có sẵn thông tin tổng thể. Giả sử bình quân tháng 7 trong 5 năm qua Rosie bán được 180 ly nước chanh. Ta nhập công thức Z.Test("Dữ liệu cột Total Sales", 180) ta được con số 0.028 (< 0.05) nên nó có giá trị thống kê, trung bình bán hàng năm nay có khác biệt với năm ngoái.

T-Test trung bình 2 mẫu

t-Test: Two-Sample Assuming Equal Variances		
	Lemon	Orange
Mean	116.5806452	80.35483871
Variance	683.1182796	489.7698925
Observations	31	31
Pooled Variance	586.444086	
Hypothesized Mean Difference	0	
df	60	
t Stat	5.889393952	
P(T<=t) one-tail	9.39311E-08	
t Critical one-tail	1.670648865	
P(T<=t) two-tail	1.87862E-07	
t Critical two-tail	2.000297822	

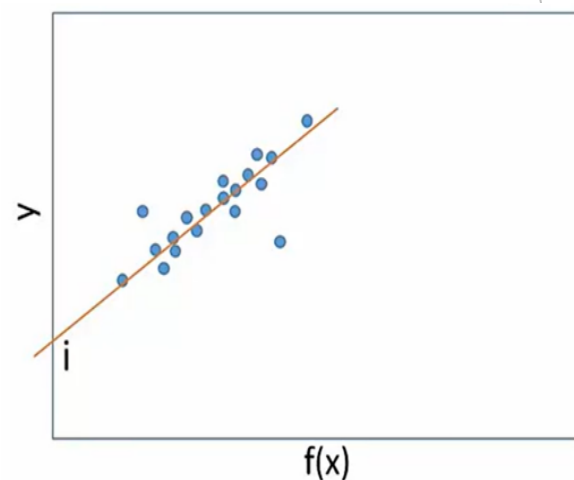
T-Test cặp đôi

- Vẫn sử dụng ví dụ trên, giả sử năm nay Rosie quyết định thay đổi chiến lược bán hàng bằng cách phát tờ rơi, điều mà cô ấy không làm trong những năm trước. Chúng ta sẽ tìm hiểu xem liệu sự thay đổi trong cách cô ấy đang làm có tạo ra sự khác biệt về mặt thống kê hay không. Việc này sẽ được thực hiện với kiểm định T-test cặp đôi.

t-Test: Paired Two Sample for Means		
	<i>Sale</i>	<i>Sale</i>
Mean	196.9354839	185.9677419
Variance	2325.929032	1638.032258
Observations	31	31
Pearson Correlation	0.942683115	
Hypothesized Mean Difference	0	
df	30	
t Stat	3.624233716	
P(T<=t) one-tail	0.000530027	
t Critical one-tail	1.697260887	
P(T<=t) two-tail	0.001060055	
t Critical two-tail	2.042272456	

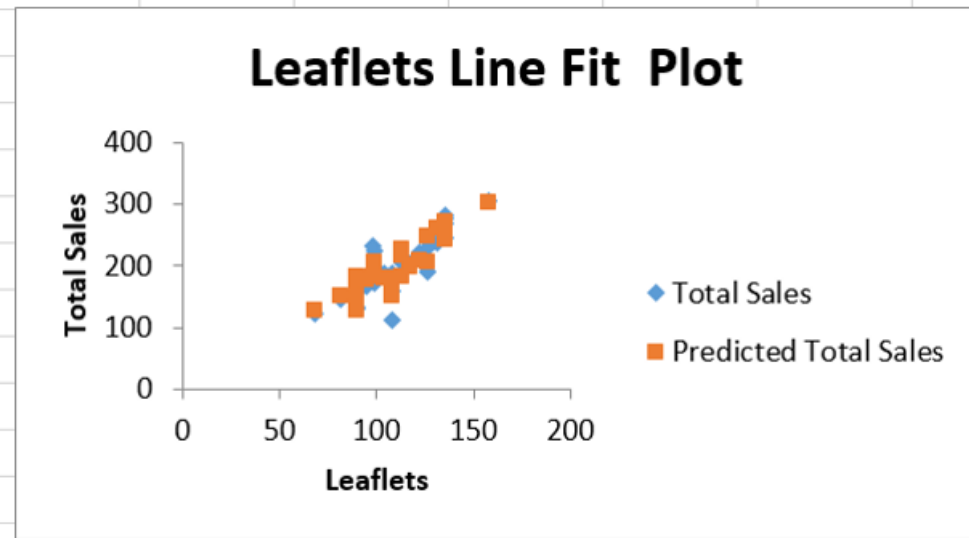
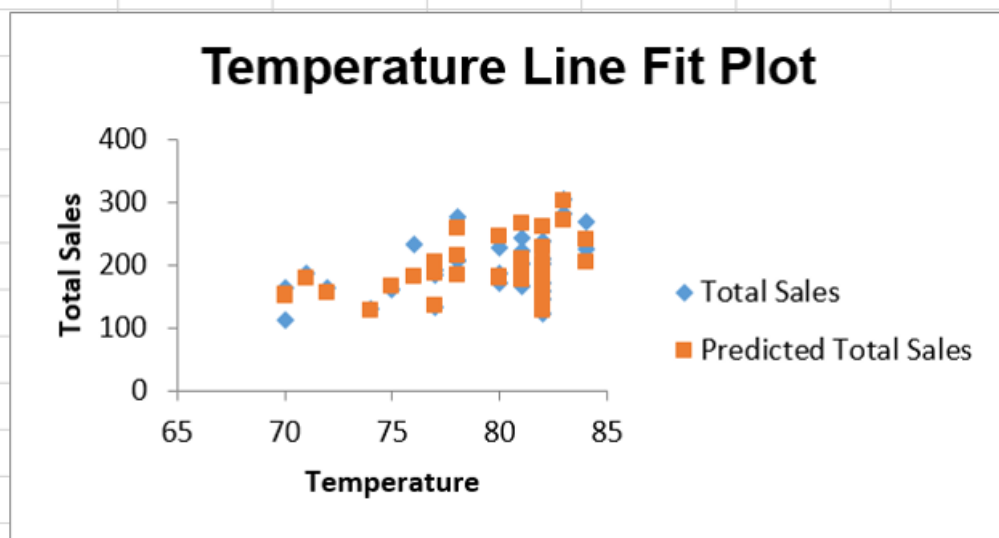
Hồi quy

- Ở phần cuối của khóa học chúng ta sẽ được học về hồi quy, ví dụ như sử dụng dữ liệu doanh số bán hàng trong quá khứ để dự đoán doanh số bán hàng trong tương lai. Nó rất hữu ích trong việc giúp xác định những biến nào trong dữ liệu của bạn có thể được sử dụng để dự đoán một số kết quả mà bạn quan tâm. Đối với ví dụ của chúng ta thì **Leaflets**, **Price** và **Temperature** liệu có thể được sử dụng để dự đoán **Sales** hay không. Về cơ bản chúng ta tìm dạng hàm số $y = f(x)$ và nếu đồ thị của nó có dạng như các điểm phân bố tập trung quanh một đường thẳng như thế này thì ta có thể kết luận là hàm tuyến tính thể hiện hàm số của đường thẳng đó chính là hàm hồi quy chúng ta cần tìm, các giá trị của y sẽ dao động xung quanh với biến thiên rất nhỏ. Giá trị i là giá trị chặn.



Hồi quy (tiếp theo)

Chúng ta có thể vẽ rất nhiều biểu đồ và nhận ra rằng trong số các biến số thì **Leaflets** và **Sales** có mối quan hệ chặt chẽ hơn trên biểu đồ so với **Temperature** và **Sales** và do đó **Leaflets** có ý nghĩa hơn so với **Temperature** trong việc dự đoán giá cả:



Chương 8. Thực hành với thống kê (1)

Thứ tự	Loại thống kê	Cột dữ liệu sử dụng	Thao tác
1	Descriptive Statistic	Orange, Lemon, Temperature, Leaflets, Price, Sales	Data → Data Analysis → Descriptive Statistic, chọn Label in First Rows, New Worksheet By, vùng dữ liệu cần mô tả (các cột dữ liệu)
2	Correlation	Orange, Lemon, Temperature, Leaflets, Price, Sales	Data → Data Analysis → Correlation

Chương 8. Thực hành với thống kê (2)

Thứ tự	Loại thống kê	Cột dữ liệu sử dụng	Thao tác
1	T-Test/Z-Test	Lemon	Nhập =Z.TEST(I2:I32, 180) trên một ô bất kỳ nằm ngoài vùng dữ liệu
2	T-Test Two-Sample	Orange, Lemon	Data → Data Analysis → T-Test Two-Sample Assuming Equal Variance, lựa chọn Labels, New Worksheet
3	T-Test Paired Two Sample for Means	Sale (năm 2015), Sale (năm 2014)	Data → Data Analysis → T-Test Paired Two Sample for Means, lựa chọn Labels, New Worksheet
4	Regression	Leaflets, Price, Temperature, Sales	Data → Data Analysis → Regression, lựa chọn Labels, New Worksheet, Residual, Residual Plots, Standardized Residuals, Line Fit Plots