

Kỹ năng số cơ bản

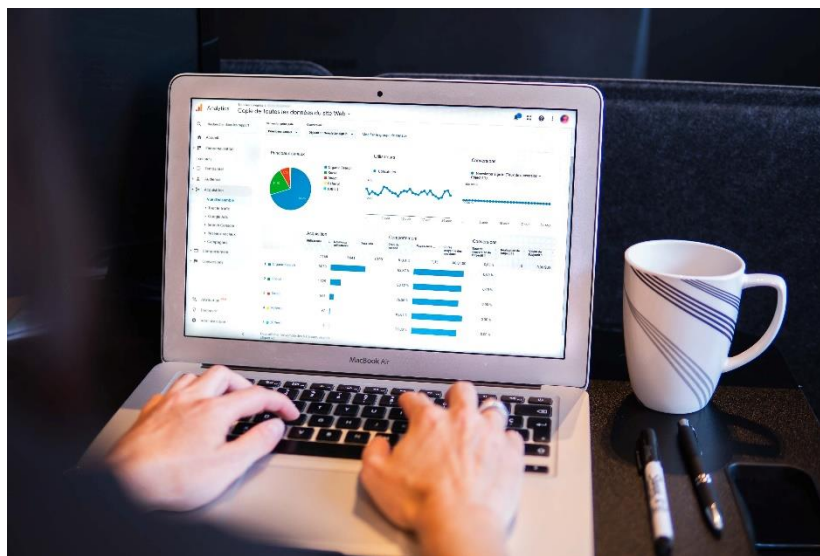
Đề cương bài giảng



TÀI LIỆU ĐÀO TẠO GIẢNG VIÊN

NGUỒN VỀ KHÓA HỌC

Khoa học dữ liệu: Ứng dụng Excel trong xử lý và quản lý dữ liệu



Khoa học Dữ liệu là cách gọi tên mới cho một khái niệm mà loài người chúng ta đã quen thuộc từ lâu. Cùng với sự phát triển của nhân loại, việc xử lý dữ liệu đã thay đổi rất nhiều. Trước kia, khi chưa xuất hiện các hệ thống đếm, loài người đã sử dụng những kí hiệu tượng hình, các cách thức biểu đạt thô sơ bằng các hình vẽ. Sau đó, với sự giúp đỡ của bàn tính, hệ thống sổ sách ghi chép, hệ thống lưu trữ, hệ thống kế toán, các phương pháp in ấn, ..v.v..con người đã đạt được những tiến bộ vượt bậc trong việc chế ngự và sử dụng dữ liệu. Có thể nói dữ liệu ở mọi nơi, mọi ngành nghề, công việc đều ít hay nhiều cần phải làm việc với dữ liệu để có thể đưa ra báo cáo, dựa trên những báo cáo này con người có thể đưa ra các quyết định phù hợp với tình hình thực tế. Ngày nay với sự bùng nổ của dữ liệu từ máy tính điện tử, điện thoại di động, các phương tiện giao thông, internet thì việc một cá nhân trong xã hội ít hay nhiều đều phải nắm được một số thao tác xử lý dữ liệu cơ bản, sử dụng được phân tích từ các dữ liệu đó để phục vụ cuộc sống và kinh doanh.

Bên cạnh đó, chuyển đổi số cũng góp phần thúc đẩy quá trình số hóa dữ liệu, từ lượng dữ liệu khổng lồ do con người và máy móc sinh ra và tích lũy hàng ngày nên ngay từ năm 2017 thời báo The Economist đã đưa ra dự đoán “Nguồn tài nguyên có giá trị nhất của loài người không phải là dầu mỏ mà là dữ liệu. Dữ liệu sẽ trở thành dầu mỏ mới”. Như vậy, việc nhận thức và sử dụng dữ liệu một cách hợp lý sẽ là lợi thế cạnh tranh không chỉ cho các tổ chức mà còn của từng cá nhân.

Chào mừng các bạn đã tham gia khóa học “Đại cương về Khoa học Dữ liệu” để trở thành những người tiên phong trong việc khai thác và sử dụng nguồn “dầu mỏ mới” này. Thông qua khóa học, bạn sẽ được giới thiệu và làm quen với lĩnh vực Khoa học dữ liệu, một lĩnh vực hứa hẹn sẽ là nền tảng cho hầu hết các ngành nghề, đời sống xã hội trong tương lai không xa. Khóa học sẽ cung cấp cho bạn các khái niệm và định nghĩa cơ bản, từ đó trợ giúp bạn có được nền tảng sẵn sàng để tiến xa hơn với những khóa học chuyên sâu hơn về sau. Hi vọng những kiến thức từ khóa học sẽ giúp ích cho bạn trong việc phát triển sự nghiệp bản thân.

Yêu cầu đầu vào: Nắm được thao tác sử dụng căn bản trên Microsoft Excel

ĐẠI CƯƠNG VỀ KHOA HỌC DỮ LIỆU

Nội dung

Tiết 1. Làm quen với Khoa học Dữ liệu	5
1.1 Giới thiệu về KHDL và sự cần thiết của KHDL trong kỷ nguyên số	5
1.2 Các nguồn dữ liệu	7
1.3 Minh họa 1 file dữ liệu bằng Excel	7
1.4 Một số thao tác cơ bản với dữ liệu trên Excel	8
1.4.1 Định dạng file, định dạng theo ngôn ngữ	8
1.4.2 Sắp xếp, lọc, dữ liệu không hoàn chỉnh	9
1.4.3 Dữ liệu sinh mới, nội suy dữ liệu, làm sạch dữ liệu	10
Tiết 2. Khám phá dữ liệu	11
2.1 Các loại câu hỏi về dữ liệu	11
2.2 Một số hàm thông dụng	11
2.3 Hiển thị trực quan trên bảng tính Excel	12
Tiết 3. Trực quan hóa dữ liệu (phần 1)	14
3.1 Trực quan hóa bằng biểu đồ	14
3.2 Trục X, Y, biểu diễn điểm, biểu diễn đường, đa đường	14
3.3 Minh họa biểu đồ bằng đường xu hướng	15
3.4 Minh họa với biểu đồ cột liên cụm, biểu đồ cột xếp chồng và biểu đồ tròn	16
3.5 Minh họa với biểu đồ phân tán để xác định tương quan	17
Tiết 4. Trực quan hóa dữ liệu (phần 2)	18
4.1 Một số biểu đồ phổ biến sử dụng trong KHDL	18
4.1.1 Minh họa với biểu đồ Tần suất	18
4.1.2 Minh họa với biểu đồ Hộp và dải dữ liệu trung bình	18
4.2 Biểu đồ 3 chiều	19
4.2.1 Biểu diễn trên không gian ba chiều	19
4.2.2 Pivot table / pivot chart	19
Tiết 5. Thực hành với dữ liệu bảng biểu trong Excel	21
5.1 Các thao tác cơ bản với dữ liệu	21
5.2 Trực quan hóa dữ liệu trên bảng tính	24
5.3 Trực quan hóa dữ liệu bằng biểu đồ	24
Tiết 6. Thống kê mô tả	25

6.1 Các loại biến số trong thống kê.....	25
6.2 Tổng thể và mẫu.....	25
6.3 Đo lường xu hướng tập trung.....	25
6.3.1 Trung bình, trung vị và yếu vị.....	26
6.3.2 Phân phối chuẩn, phân phối lệch trái/phải	26
6.4 Tính biến thiên của dữ liệu	27
6.4.1 Phạm vi giá trị	27
6.4.2 Phương sai tổng thể / phương sai 1 mẫu	27
6.4.3 Độ lệch chuẩn	28
6.4.4 Sai số chuẩn	28
Tiết 7. Thống kê kết hợp.....	29
7.1 Hệ số tương quan	29
7.2 Kiểm định giả thuyết.....	29
7.2.1 T-Test / Z-Test	30
7.2.2 T-Test trung bình hai mẫu.....	30
7.2.3 T-Test cặp đôi	31
7.2.5 Hồi quy.....	31
Tiết 8. Thực hành với thống kê.....	33
8.1 Thống kê mô tả	33
8.2 Thống kê kết hợp	33

Giới thiệu về khóa học:

Tài liệu được coi như một phần hỗ trợ cho khóa học online được cung cấp trên nền tảng congdanso.edu.vn. Tài liệu này được thiết kế bám sát với nội dung chương trình do Microsoft phát triển và tập trung vào ba nội dung chính như sau:

Phần 1. Giới thiệu về Khoa học dữ liệu: Phần này được trình bày trong tiết học đầu tiên với mục đích giúp người học có khái niệm về KHDL cũng như giới thiệu Excel như bộ công cụ để làm việc với KHDL xuyên suốt toàn khóa học

Phần 2: Nền tảng về Khoa học dữ liệu: Toàn bộ nội dung sẽ được trình bày trong tiết 2 đến tiết 5, trong đó, từ tiết 2 đến tiết 4 sẽ tập trung trình bày về các thao tác xử lý cơ bản với dữ liệu và một số công cụ hiển thị dữ liệu một cách trực quan. Tiết 5 sẽ đề cập chi tiết về phần thực hành

Phần 3: Nhập môn về Thống kê: Hai tiết 6, 7 sẽ đề cập về các lý thuyết thống kê có sử dụng các hàm thống kê trong Excel làm nền tảng công cụ mô phỏng. Tiết 8 sẽ là bài thực hành chi tiết. Để nắm vững được kiến thức, học viên cần làm theo đúng hướng dẫn trong phần thực hành.

Tiết 1. Làm quen với Khoa học Dữ liệu

Khoa học Dữ liệu là một lĩnh vực không mới, có thể xem nó như một nhánh con của Khoa học Máy tính hay Công nghệ Thông tin nhưng thay vì tập trung vào việc nghiên cứu và giải quyết các quy trình nghiệp vụ trên thực tế dựa vào máy tính thì khoa học dữ liệu tập trung vào phân tích, xử lý dữ liệu và dựa trên các kết quả này để con người đưa ra quyết định. Để trở thành một Nhà Khoa học Dữ liệu bạn không nhất thiết phải là một Kỹ sư Thông tin hay tốt nghiệp một ngành thuộc khối Công nghệ Kỹ thuật. Bạn sẽ trở thành Nhà Khoa học Dữ liệu trong lĩnh vực chuyên môn hẹp của bạn. Như vậy, nếu bạn là một nhân viên y tế, bạn sẽ có thêm một chức danh nghề nghiệp là Nhà Khoa học Dữ liệu - Y tế. Nếu bạn là chuyên gia Xã hội Học thì bạn có thể thêm chức danh nghề nghiệp là Nhà Khoa học Dữ liệu - Xã hội học vào trong danh thiếp của mình.

Sau khi học xong học phần này bạn sẽ nắm được định nghĩa về Khoa học dữ liệu, những công cụ cần thiết để chuẩn bị cho chức danh nghề nghiệp mới. Thực ra những công cụ này bạn đã ít nhiều biết về nó, có điều bạn chưa để ý đến việc sử dụng một số tính năng của nó hoặc bạn đã từng dùng nó, bạn đã biết về nó nhưng bạn chưa đặt nó vào bối cảnh của Khoa học dữ liệu.

1.1 Giới thiệu về KHDL và sự cần thiết của KHDL trong kỷ nguyên số

Được Thời báo The Economist mệnh danh là “dầu mỏ mới” trong kỷ nguyên số, việc nhận thức được tầm quan trọng và sử dụng được dữ liệu trong công việc và cuộc sống sẽ là một trong những kỹ năng cần thiết đối với chúng ta. Ngày nay, các thiết bị điện tử cá nhân như smart phone, tablet, smart watch, laptop hay thậm chí các thiết bị gia dụng, thiết bị điện tử gia đình, xe hơi đều là những thiết bị tiêu thụ thông tin. Hãy cùng tưởng tượng quá trình sinh/ tiêu thụ thông tin trong một ngày của một người trưởng thành như Tom có 1 số thiết bị như smart watch, smart phone, laptop và một chiếc xe hơi.

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	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
23 giờ	Tắt đèn, đi ngủ	Dữ liệu về thông tin sức khỏe được ghi lại

Bảng 1.1 Minh họa quá trình sinh/tiêu thụ dữ liệu của một cá nhân

Như các bạn đã thấy, đây chỉ là những hoạt động cơ bản nhất trong một ngày của người trưởng thành. Ngoài con người là đối tượng tiêu thụ dữ liệu thì các loại máy móc công nghiệp, hệ thống Internet, viễn thông, ngân hàng, thị trường chứng khoán, bệnh viện, trường học, khu công nghiệp, hệ thống quân sự, v.v. cũng trao đổi với nhau rất nhiều loại dữ liệu. Khoa học dữ liệu nói một cách đơn giản là một ngành khoa học nghiên cứu về việc xử lý một khối lượng lớn dữ liệu, phát hiện ra các mối quan hệ và sử dụng hiệu quả các kết quả phân tích đầu ra dựa trên một số bộ công cụ. Trong khuôn khổ của khóa học này, chúng ta sẽ sử dụng Microsoft Excel như là một bộ công cụ chính để làm việc với dữ liệu.

1.2 Các nguồn dữ liệu

Nguồn dữ liệu do cá nhân sinh ra: Là các số liệu do cơ thể, hoạt động hàng ngày của từng người

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

Với khối lượng dữ liệu khổng lồ hàng ngày được sinh ra và tích lũy, có thể nói dầu mỏ trên trái đất là nguồn tài nguyên có hạn chế nhưng dữ liệu thì không có giới hạn.

1.3 Minh họa 1 file dữ liệu bằng Excel

Cũng giống như toàn bộ thế giới kỹ thuật số được thể hiện bằng các ký tự 0 và 1 thì thành phần cơ bản của dữ liệu là 1 bảng bao gồm hàng và cột. Hàng thể hiện giá trị cụ thể ứng với từng đối tượng, cột thể hiện các giá trị của các thuộc tính của đối tượng ấy. Ta có bảng chiều cao, cân nặng, chỉ số BMI cũng như phân loại về cân nặng của mỗi người như bảng sau đây:

No	Name	Weight (kg)	Height (cm)	BMI	Category
1	Alice	58	161	22.7	Normal
2	Bob	63	189	17.8	Underweight
3	Crist	99	168	35.2	Obese
4	Dave	81	183	24.3	Normal
5	Emmy	84	167	29.9	Overweight

Bảng 1.2 Một dữ liệu mẫu về cân nặng của một nhóm người

Dữ liệu thực tế cũng được tổ chức dưới dạng bảng như trên có điều số hàng nhiều hơn. Khi bạn làm với dữ liệu thực tế bạn sẽ gặp những bảng dạng này, có khác chẳng là tên số hàng, số cột nhiều hơn, các hạng mục hay định dạng dạng dữ liệu khác đi.

Các hàng còn được gọi là record (bản ghi), các cột gọi là field (trường). Như trong bảng trên ta có thể thấy chỉ với ba trường dữ liệu Name, Weight, Height ta có thể suy ra tiếp các thông tin như chỉ số BMI và đánh giá sơ bộ được cân nặng và tình trạng sức khỏe của người đó.

Như vậy bạn có thể thấy vẻ đẹp của KHDL nằm ở chỗ đó, chỉ từ một ít thông tin, phát hiện được ra các mối quan hệ dựa trên các dữ liệu sẵn có và bạn lại sinh ra được dữ liệu mới, thông tin mới sẵn sàng để cho việc ra quyết định.

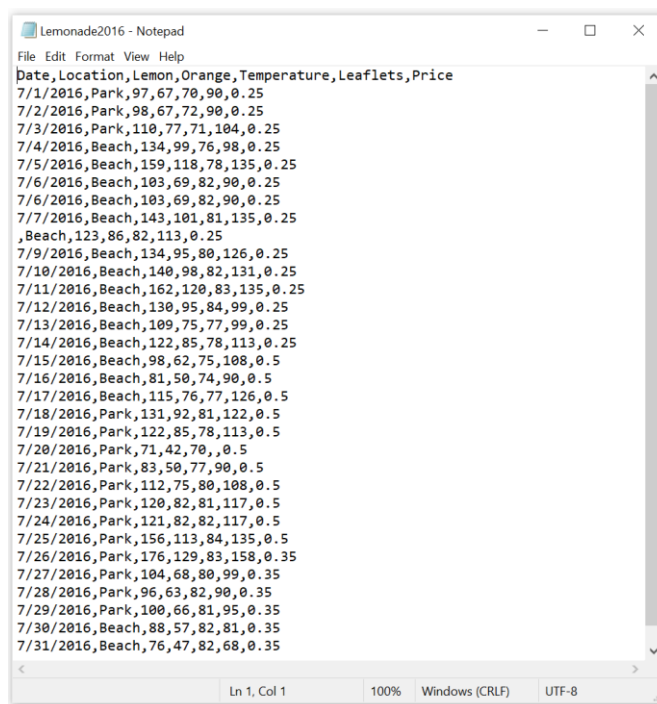
1.4 Một số thao tác cơ bản với dữ liệu trên Excel

1.4.1 Định dạng file, định dạng theo ngôn ngữ

Ngay khi bạn mở 1 file mới trên Excel, hệ thống sẽ tạo một bảng tính mới trống, không có dữ liệu, việc của bạn sẽ là nhập liệu vào đó. File Excel sẽ có đuôi file là .xlsx

Còn một cách khác là bạn nhập (import) dữ liệu từ nguồn khác vào Excel, Excel hiểu được rất nhiều loại định dạng. Về sau khi bạn có điều kiện nghiên cứu sâu hơn về dữ liệu thì sẽ có rất nhiều các loại định dạng file, quan trọng là phải hiểu được cấu trúc của các file dữ liệu đó. Các hệ thống máy móc sản sinh ra dữ liệu ở định dạng khác con người, việc của chúng ta là phải có các phần mềm hoặc thư viện mở rộng để import vào dưới dạng người đọc được. Excel thực sự là một công cụ mạnh để làm điều đó.

File mẫu để thực hiện các bài thực hành cũng như được đề cập trên các video có định dạng .csv (comma-separate value). Đây là định dạng sử dụng rất phổ biến trên các hệ thống máy móc, dung lượng file nhỏ, tuy nhiên bạn sẽ gặp khó khăn khi đọc nó. Nó không phải dùng cho con người mà dùng để kết xuất từ các hệ thống máy móc. Đúng như tên gọi của nó, các trường dữ liệu được phân tách bởi dấu phẩy, các bản ghi được bố trí trên các dòng khác nhau. File gốc nguyên bản csv chúng ta sử dụng sẽ có dạng như sau (mở bằng ứng dụng 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
```

Hình 1.1 Xem file .csv trên Notepad

Bạn sẽ thấy một chút bối rối khi nhìn thấy dữ liệu dạng này đúng không. Đừng lo, bây giờ hãy thử mở nó bằng Excel xem sao.

Date	Location	Lemon	Orange	Temperature	Price
7/1/2016	Park	97	67	70	90
7/2/2016	Park	98	67	72	90
7/3/2016	Park	100	77	73	104
7/4/2016	Beach	104	99	76	98
7/5/2016	Beach	109	118	78	135
7/6/2016	Beach	103	69	82	90
7/7/2016	Beach	103	69	82	90
7/8/2016	Beach	143	201	82	135
7/9/2016	Beach	123	86	82	113
7/10/2016	Beach	134	95	80	126
7/11/2016	Beach	140	96	82	131
7/12/2016	Beach	182	120	83	135
7/13/2016	Beach	130	95	84	99
7/14/2016	Beach	109	75	77	99
7/15/2016	Beach	122	85	78	113
7/16/2016	Beach	96	62	75	108
7/17/2016	Beach	81	50	74	90
7/18/2016	Beach	115	76	77	126
7/19/2016	Park	111	82	81	122
7/20/2016	Park	122	85	78	113
7/21/2016	Park	71	42	70	90
7/22/2016	Park	88	50	77	90
7/23/2016	Park	112	75	80	108
7/24/2016	Park	120	82	81	117
7/25/2016	Park	131	82	82	117
7/26/2016	Park	156	113	84	135
7/27/2016	Park	136	129	83	158
7/28/2016	Park	104	98	80	99
7/29/2016	Park	96	63	82	90
7/30/2016	Park	100	46	81	95
7/31/2016	Beach	88	57	82	81
7/31/2016	Beach	76	47	82	68

Hình 1.2 File .csv trong Excel

Một điều hơi khó chịu đối với các nhà KHDH như chúng ta về định dạng dữ liệu, hệ thống đo lường của chúng ta không được chuẩn hóa trên toàn bộ thế giới. Riêng đối với định dạng ngày tháng thì chúng ta đã có vô số các định dạng theo cấu trúc ngày/ tháng/ năm hoặc tháng/ngày/năm hoặc năm/tháng/ngày. Đó còn chưa kể đến việc hiển thị tháng là chữ hay là số, hiển thị năm là 4 hay 2 chữ số.

Riêng đối với các con số, ở định dạng dấu phẩy theo hệ đo lường này nó là dấu phân cách hàng nghìn, nhưng sang hệ đo lường khác nó lại là dấu phẩy thập phân. Bạn phải rất cẩn thận khi làm việc với các định dạng khác do các chuẩn đo lường khác nhau của các quốc gia.

Có rất nhiều kho dữ liệu mở trên thế giới dành cho cộng đồng những nhà KHDH 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 KHDH sử dụng nhất

1.4.2 Sắp xếp, lọc, dữ liệu không hoàn chỉnh

Dữ liệu đôi khi cần phải sắp xếp, sắp xếp là một phần bản chất của con người, thường là sắp xếp để dễ quan sát và để tìm kiếm thông tin thuận tiện hơn. Việc sắp xếp theo thứ tự trong tăng hay giảm của bảng chữ cái, theo độ lớn của số hoặc theo ngày là những công việc thường xuyên phải làm.

Bên cạnh đó, đôi khi với một lượng dữ liệu lớn mà chúng ta muốn tìm kiếm hay thao tác với tập con của dữ liệu đó, ta cần phải biết các hàm lọc (filter) để có thể làm việc với các đối tượng ở phạm vi tìm kiếm hẹp và liên quan đến vấn đề phải giải quyết hơn.

Dữ liệu, xét trong điều kiện lý tưởng luôn luôn đầy đủ nhưng trên thực tế, thường là không đầy đủ. Việc không đầy đủ có thể do nhiều nguyên nhân như thuật toán ghi dữ liệu bị lỗi, rớt gói tin trên mạng, thao tác sơ sót của người sử dụng. Dữ liệu bị thiếu có thể là các bản ghi hoặc có khi chỉ là một trường nào đó.

Bên cạnh dữ liệu bị thiếu còn khả năng là có dữ liệu nhưng không đúng định dạng hoặc nằm ở các khoảng ngoại lai dẫn đến việc xử lý dữ liệu sẽ bị sai theo nội dung sai của dữ liệu.

Trong mọi trường hợp, ta cần phải xem xét cẩn trọng dữ liệu và áp dụng một số biện pháp để hoàn thiện hoặc chỉnh sửa dữ liệu theo những cách thức nhằm giảm thiểu sự sai lệch của dữ liệu. Excel cũng cung cấp một số công cụ chiến lược như thế sẽ được đề cập trong phần thực hành.

1.4.3 Dữ liệu sinh mới, nội suy dữ liệu, làm sạch dữ liệu

Như trên ví dụ trong bảng 1.2, các bạn có thể nhìn thấy dữ liệu BMI là dữ liệu được sinh mới từ dữ liệu cân nặng chia cho bình phương chiều cao. Dữ liệu phân loại từ suy dinh dưỡng đến béo phì được căn cứ theo các khoảng dữ liệu của BMI.

Như đã nói ở trên, dữ liệu nhiều khi không hoàn chỉnh, ta cần phải có cách thức để nội suy và sinh ra được dữ liệu một cách phù hợp. Đối với dữ liệu loại liên tục thì rất dễ, chỉ cần điền số còn thiếu. Đối với dữ liệu dạng số thì cách thường dùng thì ta có thể sử dụng giá trị trung bình.

Làm sạch dữ liệu là cách thức tổng hợp giúp chúng ta có thể loại bỏ việc dư thừa dữ liệu, điền được dữ liệu khuyết thiếu, dữ liệu được định dạng không đúng, dữ liệu nằm ngoài khoảng, phục hồi ở mức tối đa các dữ liệu mất mát theo các chiến lược phù hợp. Đầu ra của dữ liệu được làm sạch sẽ được sử dụng cho những phân tích dữ liệu ở các giai đoạn sau.

Tiết 2. Khám phá dữ liệu

Trong tiết này chúng ta sẽ làm quen với 1 kịch bản về dữ liệu do Microsoft xây dựng và sẽ được sử dụng xuyên suốt trong khóa học này.

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 nước chanh. Rosie làm nước chanh 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 nước chanh 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 nước chanh 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.

2.1 Các loại câu hỏi về dữ liệu

Sau khi ghi chép dữ liệu, Rosie rất có thể muốn hỏi những câu hỏi như sau:

- 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.

Đó là 4 loại câu hỏi mà chúng ta có thể hỏi về dữ liệu.

2.2 Một số hàm thông dụng

Khi làm việc với dữ liệu, chúng ta thường rất hay muốn dữ liệu tổng hợp để chúng ta có cái nhìn tổng quan về dữ liệu.

Ví dụ chúng ta muốn dùng hàm count (đếm) số ngày Rosie bán hàng trong kịch bản dữ liệu. Ta chỉ đơn giản chọn xuống hàng cuối cùng và sử dụng hàm **count** (Cẩn trọng với định dạng dữ liệu do cột Date có định dạng là Date, bạn phải chuyển về định dạng Decimal Number)

Đối với định dạng doanh thu, ta chọn Currency là \$ sau khi lựa chọn cột Revenue

Để tính trung bình lượng **Orange** được bán ra mỗi ngày, ta chọn lại hàng dưới cùng và cột **Orange** sử dụng hàm trung bình average.

Để tính lượng **Lemon** bán ra ta chỉ cần copy cell sang bên cột **Lemon**, Excel sẽ tự hiểu và tính toán cho chúng ta.

Để tính được các giá trị lớn nhất / nhỏ nhất của Temperature ta chọn lại hàng dưới cùng và dùng hàm **max/min**

Để tính tổng lượng **Lemon** bán ra ta có thể sử dụng hàm **sum** trong Excel đối với cột **Lemon**

Muốn tính doanh số của cả 2 loại đồ uống ta chỉ cần thêm 1 cột **Sales** và nhập công thức: **Sales = Lemon + Orange**

Muốn tính doanh thu bán hàng của **Lemon** và **Orange** chúng ta thêm 1 cột mới **Revenue** được tính bằng công thức sau: **Revenue = Sales * Price**

Như các bạn có thể thấy file dữ liệu của chúng được sinh thêm 2 cột dữ liệu mới

	A	B	C	D	E	F	G	H	I
1	Date	Location	Lemon	Orange	Temperature	Leaflets	Price	Sales	Revenue
2	7/1/2016	Park	97	67	70	90	0.25	164	41
3	7/2/2016	Park	98	67	72	90	0.25	165	41.25
4	7/3/2016	Park	110	77	71	104	0.25	187	46.75
5	7/4/2016	Beach	134	99	76	98	0.25	233	58.25
6	7/5/2016	Beach	159	118	78	135	0.25	277	69.25
7	7/6/2016	Beach	103	69	82	90	0.25	172	43
8	7/6/2016	Beach	103	69	82	90	0.25	172	43
9	7/7/2016	Beach	143	101	81	135	0.25	244	61
10	7/8/2016	Beach	123	86	82	113	0.25	209	52.25
11	7/9/2016	Beach	134	95	80	126	0.25	229	57.25
12	7/10/2016	Beach	140	98	82	131	0.25	238	59.5
13	7/11/2016	Beach	162	120	83	135	0.25	282	70.5
14	7/12/2016	Beach	130	95	84	99	0.25	225	56.25
15	7/13/2016	Beach	109	75	77	99	0.25	184	46
16	7/14/2016	Beach	122	85	78	113	0.25	207	51.75
17	7/15/2016	Beach	98	62	75	108	0.5	160	80
18	7/16/2016	Beach	81	50	74	90	0.5	131	65.5
19	7/17/2016	Beach	115	76	77	126	0.5	191	95.5
20	7/18/2016	Park	131	92	81	122	0.5	223	111.5
21	7/19/2016	Park	122	85	78	113	0.5	207	103.5
22	7/20/2016	Park	71	42	70	108	0.5	113	56.5
23	7/21/2016	Park	83	50	77	90	0.5	133	66.5
24	7/22/2016	Park	112	75	80	108	0.5	187	93.5
25	7/23/2016	Park	120	82	81	117	0.5	202	101
26	7/24/2016	Park	121	82	82	117	0.5	203	101.5
27	7/25/2016	Park	156	113	84	135	0.5	269	134.5
28	7/26/2016	Park	176	129	83	158	0.35	305	106.75
29	7/27/2016	Park	104	68	80	99	0.35	172	60.2
30	7/28/2016	Park	96	63	82	90	0.35	159	55.65
31	7/29/2016	Park	100	66	81	95	0.35	166	58.1
32	7/30/2016	Beach	88	57	82	81	0.35	145	50.75
33	7/31/2016	Beach	76	47	82	68	0.35	123	43.05

Bảng 2.1 Dữ liệu thêm hai cột Sales và Revenue

2.3 Hiện thị trực quan trên bảng tính Excel

Mặc dù hiển thị dưới dạng bảng đã giúp chúng ta dễ dàng đi rất nhiều khi quan sát dữ liệu nhưng vẫn có những cách trực quan hơn để quan sát dữ liệu trên bảng tính Excel.

Đó là những cách hiển thị giúp chúng ta xác định được những điểm khác thường, những dữ liệu ngoại lai hoặc những dữ liệu ở những cực khác nhau của dải dữ liệu. Điều này đặc biệt hữu ích khi chúng ta làm việc với những bảng dữ liệu khổng lồ. Những dữ liệu cần được lưu ý sẽ được thể hiện một cách nổi bật.

Định dạng có điều kiện (Conditional Formatting) là một bộ công cụ để giúp chúng ta thể hiện dữ liệu trực quan hơn.

Data Bar (Định dạng thanh dữ liệu) là một công cụ để giúp chúng ta hiển thị các dữ liệu theo màu sắc, ví dụ ta chọn một màu bất kì thì trong cột **Revenue** những ngày có doanh thu cao sẽ được thể hiện bằng phần tô màu chiếm tỉ trọng lớn (solid fill) trong background của ô dữ liệu đó.

Đối với cột **Temperature** ta có thể lựa chọn công cụ Color Scale với 2 tông màu Red/White, những ngày có nhiệt độ cao nóng hơn sẽ có màu đỏ sẫm hơn, những ngày lạnh hơn thì màu chuyển về gần màu trắng phớt hồng.

Phối hợp các cột **Temperature** và **Revenue** ta thấy cột **Temperature** có màu đỏ sẫm hơn và phần màu bên cột **Revenue** được tô màu phần lớn, ta có thể đưa ra giả thuyết là những ngày nóng thì doanh thu cao hơn. Đừng vội kết luận ngay, chúng ta còn phải kiểm định giả thuyết bằng các công thức chứ không thể dùng hình ảnh trực quan để kiểm chứng.

Đối với bộ dữ liệu của chúng ta, các bạn có thể sử dụng thêm công cụ Icon Sets (bộ biểu tượng) như ngôi sao để hiển thị trực quan lượng tờ rơi phát ra với ngôi sao được tô màu hoàn toàn là những ngày phát được nhiều tờ rơi, những ngôi sao tô màu một nửa hoặc không được tô màu là những ngày phát được trung bình hoặc ít tờ rơi hơn.

Để xác định được các yếu tố ngoại lai ta sẽ sử dụng công cụ Top/Bottom Rules (nguyên tắc trên cùng / dưới cùng) với số % được đặt tùy biến.

Tiết 3. Trục quan hóa dữ liệu (phần 1)

3.1 Trục quan hóa bằng biểu đồ

Như bạn đã thấy, chúng ta có thể phân tích dữ liệu bằng cách sử dụng một số công thức, hàm tổng hợp và sau đó trục quan hóa dữ liệu vào các cột riêng lẻ. Tuy nhiên việc trục quan hóa bằng biểu đồ còn tốt hơn nhiều do con người chúng ta dễ dàng làm việc với hình vẽ hơn là bảng. Bước tiếp theo chúng ta nên tạo một số biểu đồ, đồ thị và các con số để trình bày trục quan chân thực những gì đang diễn ra trong dữ liệu.

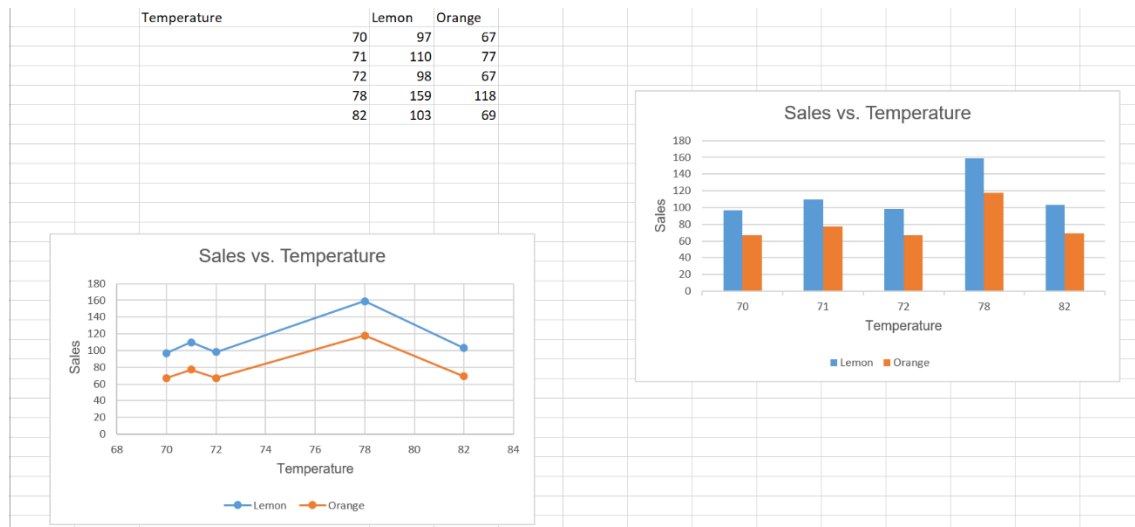
3.2 Trục X, Y, biểu diễn điểm, biểu diễn đường, đa đường

Tuy nhiên, trước khi chúng ta làm điều đó, điều chúng ta cần làm là hiểu một số khái niệm cơ bản trong biểu đồ. Thông thường khi chúng ta tạo biểu đồ, chúng ta sử dụng trục, và thông thường có một Trục X và trục Y trong hầu hết các biểu đồ. Không phải tất cả nhưng hầu hết các biểu đồ đều có X và Y và cách dễ nhớ là X là trục ngang và Y là trục dọc. Chúng ta có hai trục và tiếp theo chúng ta sẽ gắn nhãn trục để biết các trục biểu thị gì hoặc đại diện cho cái gì; các giá trị ở trục X, trong trường hợp này, nó sẽ là nhiệt độ và giá trị ở trục Y sẽ là doanh số bán hàng. Vì vậy, chúng ta đang so sánh hai dữ liệu khác nhau dọc theo các trục này và một trong những điều bạn sẽ làm là đánh dấu các điểm giá trị hơn là tìm hiểu một loại biểu đồ cụ thể nào. Các giá trị được đánh dấu trong trường hợp này là một giá trị của **Sales** ứng với một giá trị của **Temperature**. Chúng ta nên đặt tiêu đề trên cùng cho biểu đồ này để giúp người xem hiểu những gì biểu đồ này thực sự hiển thị mà không biết biểu đồ biểu diễn giá trị nào. Tỷ lệ cũng giữa các giá trị cũng rất quan trọng, tỷ lệ giữa trục X và Y không nhất thiết phải giống nhau. Nhưng chúng phải phù hợp với dữ liệu đang được biểu thị. Đôi khi có thể chuẩn hóa các giá trị để tất cả các giá trị xuất hiện dưới dạng giá trị từ 0 đến 1 chẳng hạn. Việc tiếp theo thông thường chúng ta sẽ làm là vẽ một đường đi qua các điểm giá trị đã được đánh dấu để được biểu đồ dạng đường. Và biểu đồ đường là một loại biểu đồ thực sự phổ biến.

Khi vẽ bất cứ biểu đồ nào, chúng ta phải thực hiện các bước sau:

1. Lựa các cột dữ liệu trên bảng để thực hiện biểu diễn (có thể thêm bớt các cột dữ liệu đưa vào biểu đồ sau)
2. Chọn loại biểu đồ cần biểu diễn
3. Tạo loại biểu đồ được lựa chọn và tùy biến hiển thị
4. Thêm chú thích cho rõ ràng

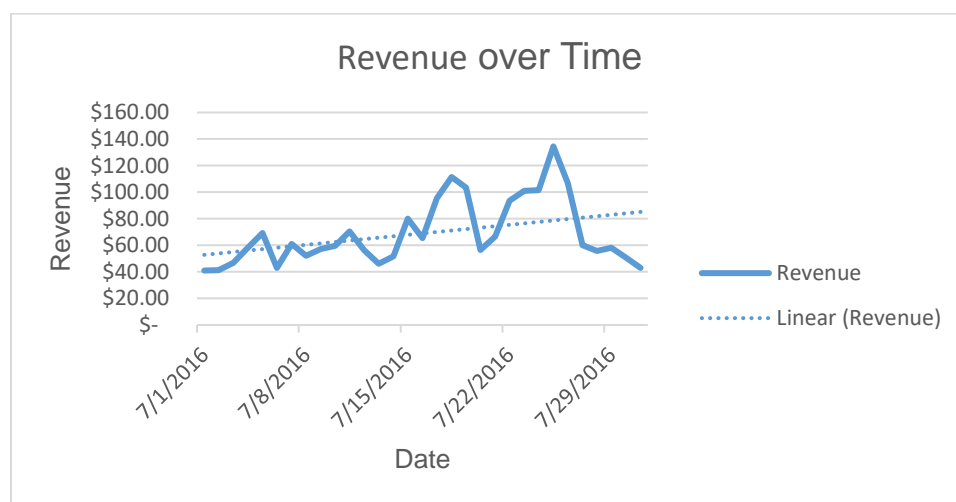
Quan sát biểu đồ doanh số của từng loại đồ uống **Orange** và **Lemon** so với **Temperature** được thể hiện bởi 2 đường riêng biệt. Ta cũng có thể sử dụng biểu đồ liên cụm (Stacked Column) để có cái nhìn so sánh trục quan về doanh số bán hàng của từng loại đồ uống.



Hình 3.1 Biểu đồ đường/liên cụm doanh số từng loại đồ uống theo nhiệt độ

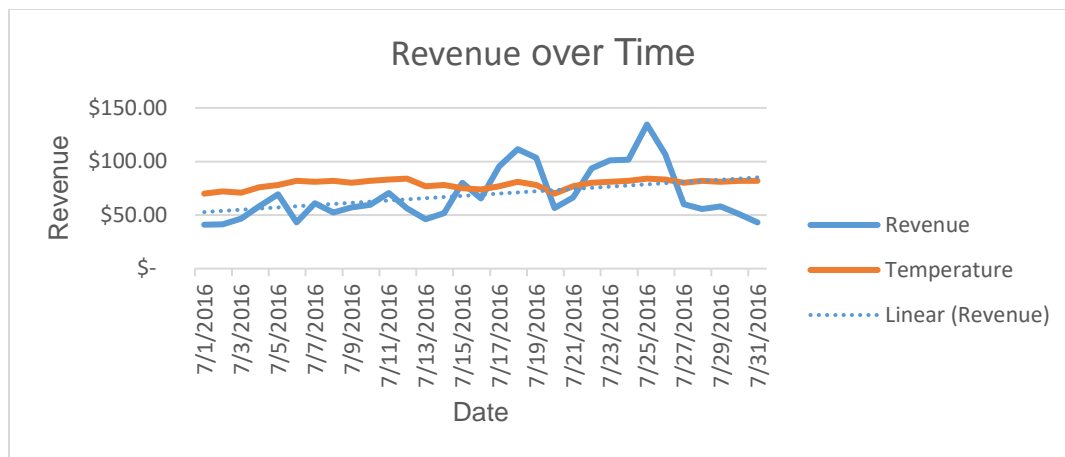
3.3 Minh họa biểu đồ bằng đường xu hướng

Sử dụng 2 cột Date và Revenue để vẽ biểu đồ sau (nhớ chú thích tên biểu đồ cũng như tên của các trục). Để quan sát được xu hướng của biểu đồ chúng ta thêm đường trendline



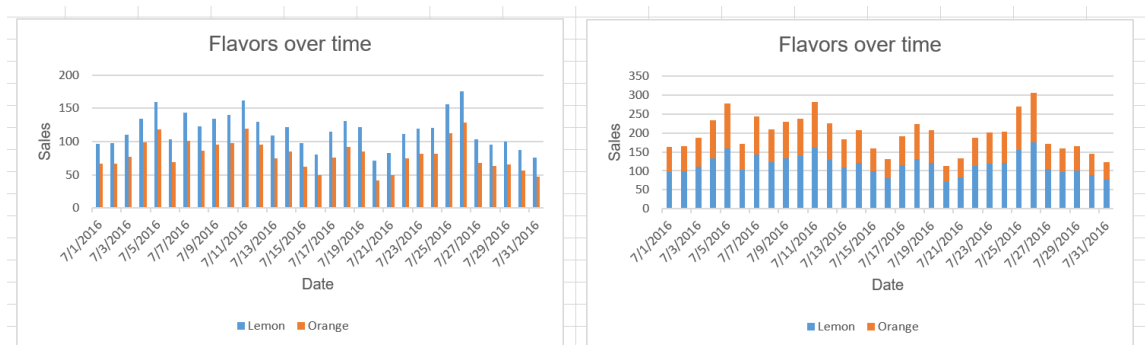
Hình 3.2 Biểu đồ doanh thu theo thời gian

Ta cũng có thêm các dữ liệu về nhiệt độ để xem có sự tương quan nào giữa doanh thu và nhiệt độ hay không.



Hình 3.3 Biểu đồ doanh thu, nhiệt độ theo thời gian

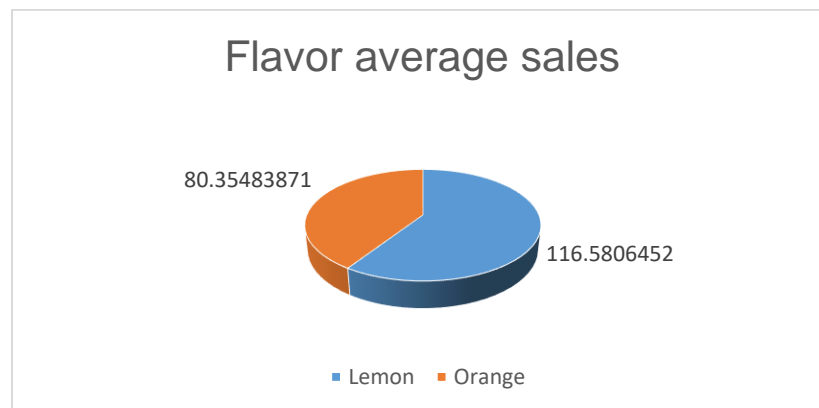
3.4 Minh họa với biểu đồ cột liên cụm, biểu đồ cột xếp chồng và biểu đồ tròn



Hình 3.4 Biểu đồ cột liên cụm / cột xếp chồng

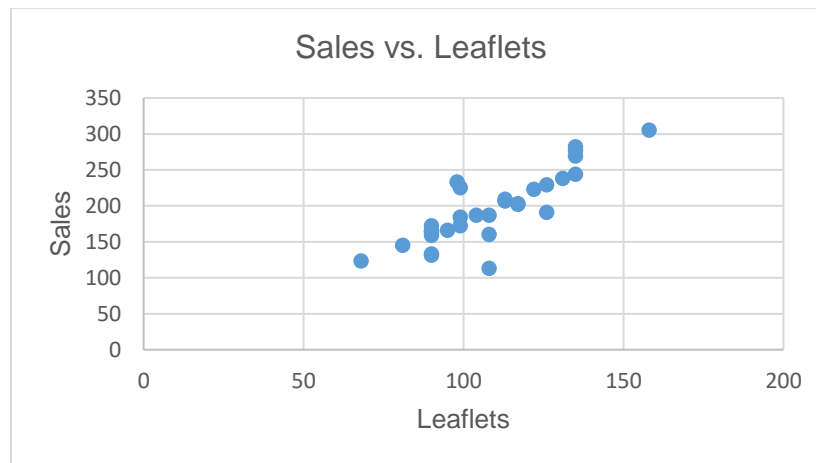
Biểu đồ bên trái hình 3.4 cho ta thấy sự so sánh doanh số bán hàng giữa 2 loại đồ uống. Biểu đồ bên phải lại cho thấy được cái nhìn tổng thể của từng loại đồ uống trong tổng doanh số

Biểu đồ trên hình 3.5 thể hiện so sánh trung bình doanh số của từng loại đồ uống



Hình 3.5 Biểu đồ trung bình doanh số của từng loại đồ uống

3.5 Minh họa với biểu đồ phân tán để xác định tương quan



Hình 3.5 Biểu đồ phân tán doanh số và tờ rơi phát ra

Biểu đồ phân tán này cho chúng ta thấy dường như các giá trị của doanh số dao động xung quanh một đường thẳng có xu thế hướng lên trên. Điều đó thể hiện có sự tương quan giữa các dữ liệu này. Ta sẽ đi sâu vào phân tích biểu đồ này trong phần sau.

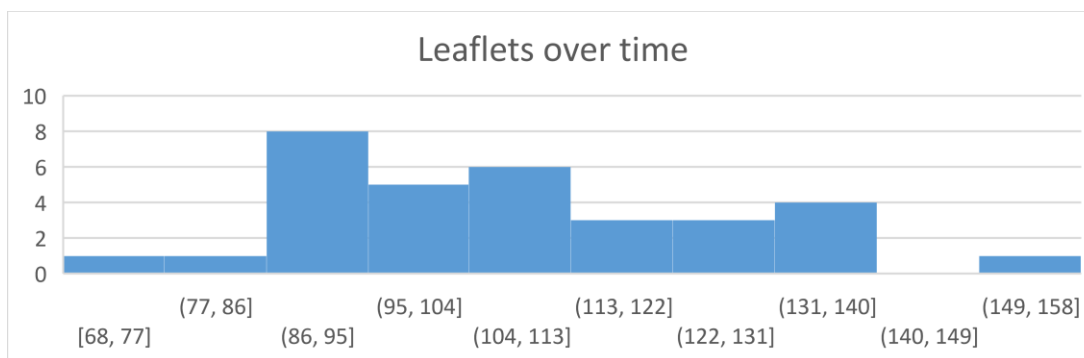
Tiết 4. Trực quan hóa dữ liệu (phần 2)

4.1 Một số biểu đồ phổ biến sử dụng trong KHDL

KHDL sử dụng rất nhiều các loại biểu đồ nhưng trong khuôn khổ của khóa học này chúng ta chỉ đề cập đến 2 loại biểu đồ đó là biểu đồ Tần suất và biểu đồ Hộp và dải dữ liệu trung bình

4.1.1 Minh họa với biểu đồ Tần suất

Đầu tiên ta đề cập đến biểu đồ tần suất với 10 ngăn, chúng ta lấy cột **Leaflets** làm bộ dữ liệu cho biểu đồ

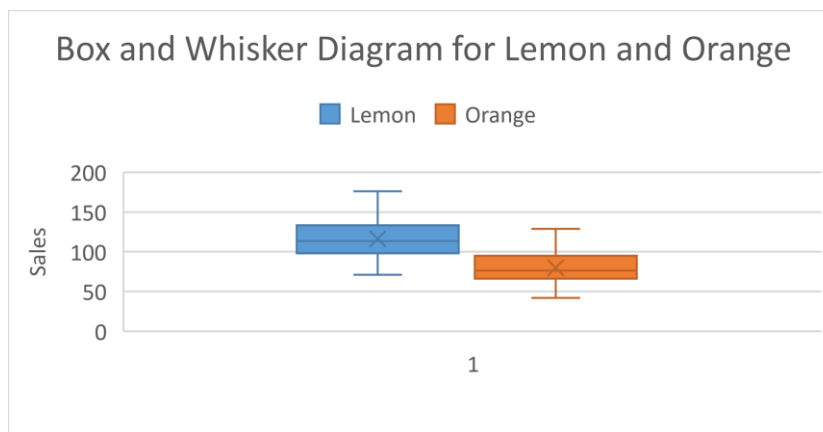


Hình 4.1 Biểu đồ phân tán về số tờ rơi phát ra

4.1.2 Minh họa với biểu đồ Hộp và dải dữ liệu trung bình

Qua quan sát ta có thể thấy không có ngày nào có số lượng tờ rơi phát ra nằm trong khoảng từ 140-149 tờ rơi.

Đối với biểu đồ Hộp và dải dữ liệu trung bình chúng ta lấy cột **Lemon** và **Orange** làm bộ dữ liệu cho biểu đồ. Ta có thể thấy kí hiệu x đại diện cho giá trị trung bình, điểm cao nhất, thấp nhất của hộp là tứ phân vị thứ 3 và tứ phân vị thứ nhất, đường kẻ chính giữa là giá trị trung vị, tận cùng mút của dải dữ liệu là các giá trị max/min.



Hình 4.2 Biểu đồ Hộp và dải dữ liệu cho từng loại đồ uống

4.2 Biểu đồ 3 chiều

4.2.1 Biểu diễn trên không gian ba chiều

Khi làm việc với dữ liệu lớn, đôi khi chúng ta cần phải có cách cắt lát, rút gọn và chiết xuất dữ liệu ở những dạng tóm tắt, mang tính chất tổng quan hơn, đó là khi chúng ta sử dụng biểu diễn trên không gian ba chiều. Hãy thử tưởng tượng Rosie quan tâm đến việc doanh số bán hàng của **Lemon** và **Orange** theo **Day**. Dữ liệu này có 2 khía cạnh, đó là doanh số đồ uống và theo ngày. Tuy nhiên khi Rosie quan tâm đến cả việc xác định doanh số bán hàng theo cả **Location** ở bãi biển hay công viên sẽ phát sinh thêm chiều thứ ba của dữ liệu.

4.2.2 Pivot table / pivot chart

Bộ dữ liệu chúng ta dùng ở đây sẽ gồm các cột **Lemon**, **Orange**, **Day** và **Location**. Lựa chọn các cột đó và lưu Pivot table trên 1 sheet mới

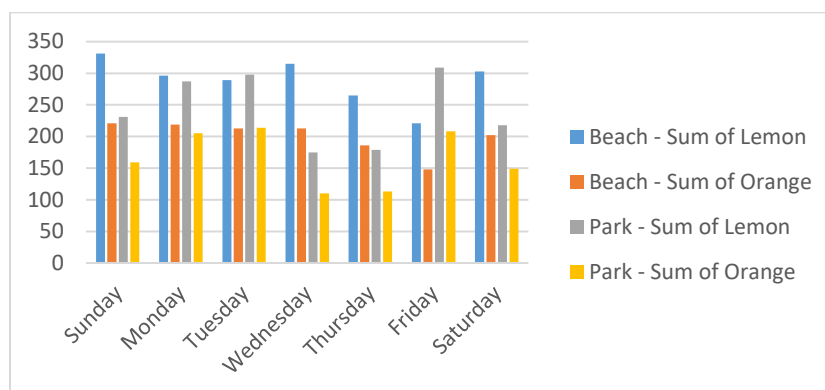
Row Labels	Sum of Lemon	Sum of Orange
Sunday	562	380
Beach	331	221
Park	231	159
Monday	583	424
Beach	296	219
Park	287	205
Tuesday	587	427
Beach	289	213
Park	298	214
Wednesday	490	323
Beach	315	213
Park	175	110
Thursday	444	299
Beach	265	186
Park	179	113
Friday	530	356
Beach	221	148
Park	309	208
Saturday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

Bảng 4.1 Pivot table trên dữ liệu được lựa chọn

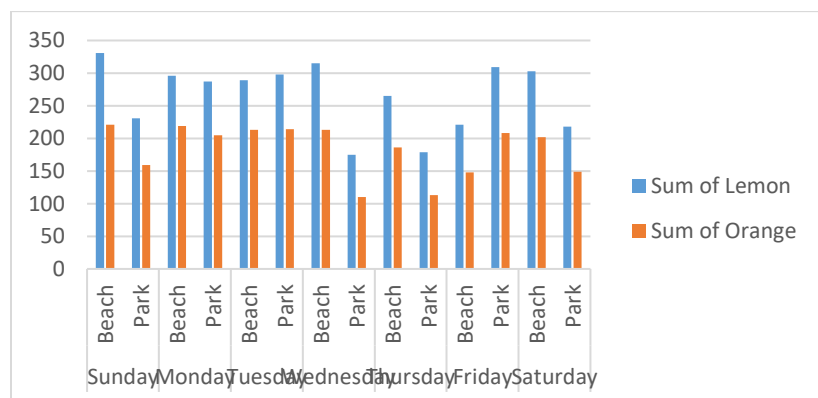
Column Labels						
Row Labels	Beach		Park		Total Sum of Lemon	Total Sum of Orange
	Sum of Lemon	Sum of Orange	Sum of Lemon	Sum of Orange		
Sunday	331	221	231	159	562	380
Monday	296	219	287	205	583	424
Tuesday	289	213	298	214	587	427
Wednesday	315	213	175	110	490	323
Thursday	265	186	179	113	444	299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

Bảng 4.2 Pivot table với Location chuyển sang vị trí cột

Bảng trên cho ta thấy sự khác biệt về góc nhìn dữ liệu khi chuyển Location sang vị trí cột. Từ bảng Pivot table trên ta có Pivot chart như dưới đây, tuy nhiên tại 1 ngày cụ thể có 4 cột dữ liệu khiến chúng ta cảm thấy khó khăn khi theo dõi dữ liệu.



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 với dữ liệu bảng biểu trong Excel

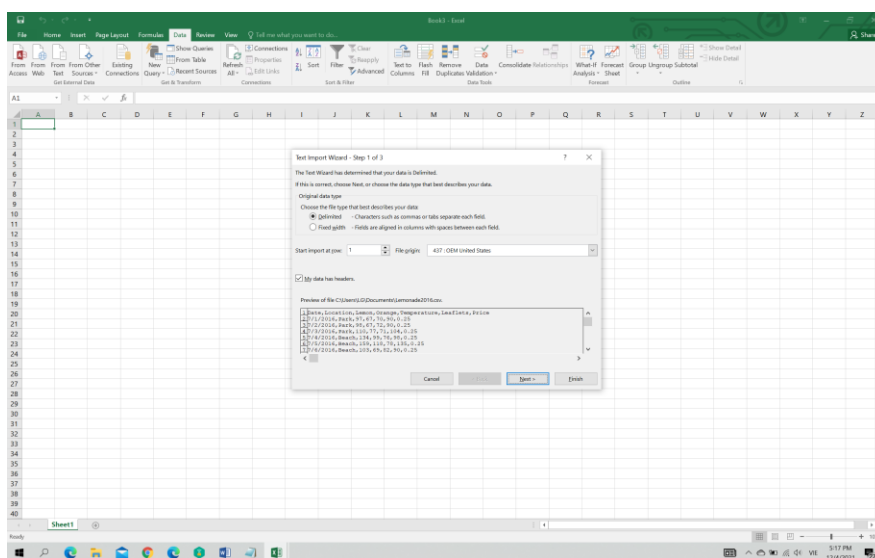
Trong tiết này chúng ta sẽ làm quen với 1 kịch bản về dữ liệu do Microsoft xây dựng và sẽ được sử dụng xuyên suốt trong khóa học này.

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 nước chanh. Rosie làm nước chanh 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 nước chanh 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 nước chanh 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. Trong phần tiếp theo, chúng ta sẽ xem xét dữ liệu của Rosie và những gì cô ấy có thể làm với tập dữ liệu đó.

5.1 Các thao tác cơ bản với dữ liệu

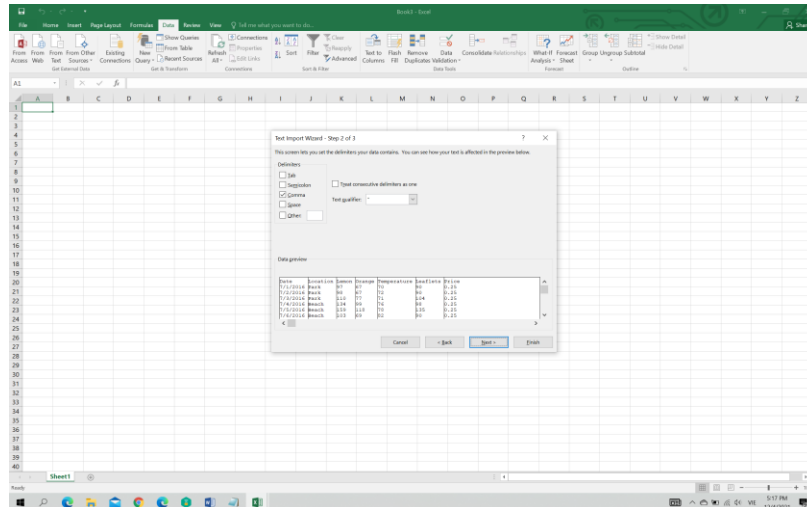
File csv được hỗ trợ mở tự động trong Excel tuy nhiên như không phải lúc nào các file dữ liệu cũng được hỗ trợ. Chúng ta hãy xem xét mở file .csv bằng cách sau

Trên thanh công cụ Data → From Text → Chọn đến vị trí của file Lemonade.csv



Hình 5.1 Mở file .csv theo cách định dạng text

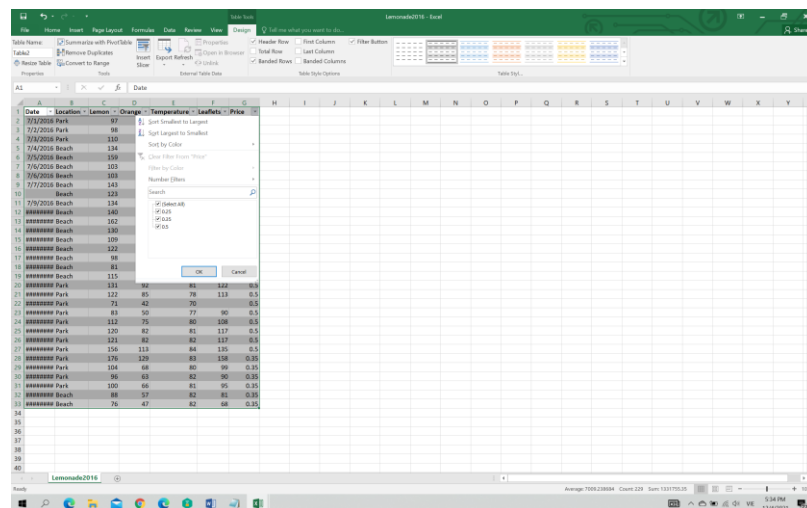
Ta sẽ gặp cửa sổ trên, nhớ rằng trong file csv ta có phần header (tiêu đề) và phần delimiter (dấu phân tách giữa các trường dữ liệu) nên bạn phải chọn vào những phần này. Ở cửa sổ tiếp theo vì chúng ta đã biết định dạng phân cách các trường là dấu comma nên các bạn phải chọn vào dấu comma



Hình 5.2 Chọn delimiter là comma

Cửa sổ tiếp theo sẽ cho bạn thấy định dạng ngày theo kiểu Mỹ MDY (Month/Day/Year) nhưng nếu bạn không quen thuộc bạn có thể chuyển đổi định dạng về ngôn ngữ địa phương quy định của bạn. Đối với các trường dữ liệu khác Excel có thể tự động nhận dạng kiểu dữ liệu, bạn có thể tùy biến sau. Quan trọng nhất bạn phải nhớ là hàng đầu tiên trong file dữ liệu này chính là nhãn của các trường dữ liệu. Sau khi bấm Finish thì dữ liệu của bạn sẽ sẵn sàng như trong Hình 1.2

Để có thể thao tác dễ dàng hơn với dữ liệu, chúng ta quay lại thanh công cụ, chọn Home → Format as Table, chúng ta sẽ có định dạng theo bảng rất dễ dàng theo để sử dụng với kéo thả và lựa chọn giá trị (Drop Down và Checkbox) để thao tác sắp xếp hoặc lọc dữ liệu. Như trong lựa chọn dưới đây ta có thể sắp xếp các giá trị tiền 1 cốc nước chanh từ thấp lên cao hoặc ngược lại cũng như chỉ lọc ra những bản ghi nào có giá tiền là 1,2 hoặc cả 3 yếu tố giá 0.25, 0.35 và 0.5



Hình 5.3 Dropdown và Checkbox trong định dạng bảng

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
7/9/2016	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

Hình 5.4 Quan sát giá trị bị thiếu

Tiếp tục quan sát kỹ hơn chúng ta có thể có 1 số dữ liệu bị mất ở đây, đó là cột Date ở giữa giá trị ngày 7 và 9 tháng 7 năm 2016, ta có thể ngoại suy dữ liệu đó là ngày 8-7-2016.

Ngày 20-7-2016 cũng bị thiếu trường dữ liệu Leaflets. Trong tình huống này ta sẽ phải cân nhắc một chút và nhận ra rằng tờ rơi phải là số nguyên, nó có thể đặt là 0, giá trị min /max nhưng tốt nhất là cho nó giá trị trung bình. Bạn chỉ cần chọn cột này trong Excel là sẽ có giá trị trung bình ở thanh trạng thái dưới cùng màn hình. Vì số lượng tờ rơi là số nguyên nên ta làm tròn nó tại giá trị 109 và điền dữ liệu này vào ô còn trống.

The screenshot shows the Excel interface with the data from Figure 5.4. The date 7/8/2016 has been added, and the Leaflets value for that date is 109, which is the average of the Leaflets values for 7/7/2016 (135) and 7/9/2016 (113). The status bar at the bottom shows the average of the selected cells is 109.5, which is rounded to 109.

Hình 5.5 Giá trị trung bình của lượng tờ rơi phát ra

Tiếp tục lướt qua dữ liệu ta thấy dữ liệu ngày 7-6-2016 được nhập 2 lần, ta có thể sử dụng tính năng Remove Duplicate trong Data để xóa dữ liệu nhập trùng này.

Đưa trỏ chuột ra phía ngoài cùng ở ô I1 điền chữ **Sales**. Để thêm dữ liệu cho cột Sales ta chỉ cần chọn đến hàng đầu tiên dưới nhãn của cột **Sales** ở vị trí I2 (vị trí nhập liệu đầu tiên của cột Sales) nhập công thức = D2 + E2 (lần lượt là các giá trị đầu tiên của cột Lemon và

Orange). Sau đó đặt con chuột vào góc dưới cùng bên phải của ô I2 cho đến khi hiện ra dấu + thì giữ chuột và kéo xuống đến ô I31 rồi thả chuột.

Làm tương tự với ô J2 với công thức = I2 * H2 cho cột Revenue và thao tác tương tự đối với cột **Sales** để hoàn thiện các giá trị bên dưới

Chèn 1 cột Day nằm giữa Date và Location ở vị trí B2 sử dụng công thức như dưới đây:

=TEXT(WEEKDAY(A2),"dddd") sau đó thao tác tương tự như đối với cột **Sales** để hoàn thiện các giá trị bên dưới.

5.2 Trục quan hóa dữ liệu trên bảng tính

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%

Bảng 5.1 Thao tác trục quan hóa dữ liệu trên bảng tính

5.3 Trục quan hóa dữ liệu bằng 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 Column / 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

Bảng 5.2 Thao tác trục quan hóa dữ liệu trên biểu đồ

Tiết 6. Thống kê mô tả

Mặc dù đây không phải là một khóa học chuyên về thống kê nhưng là một nhà KHDL, bạn không thể không biết một số khái niệm cũng như một số hàm thống kê dùng trong KHDL. Trong trường hợp của Rosie, chúng ta có thể áp dụng thống kê để phát hiện ra một số quy luật thống kê để cải thiện doanh số bán đồ uống của cô ấy.

6.1 Các loại biến số trong thống kê

Trong thống kê mọi thứ đều xoay quanh các biến số. Có 3 loại biến số ta thường xuyên làm việc là: Biến số liên tục, biến số rời rạc và biến số phân loại

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.

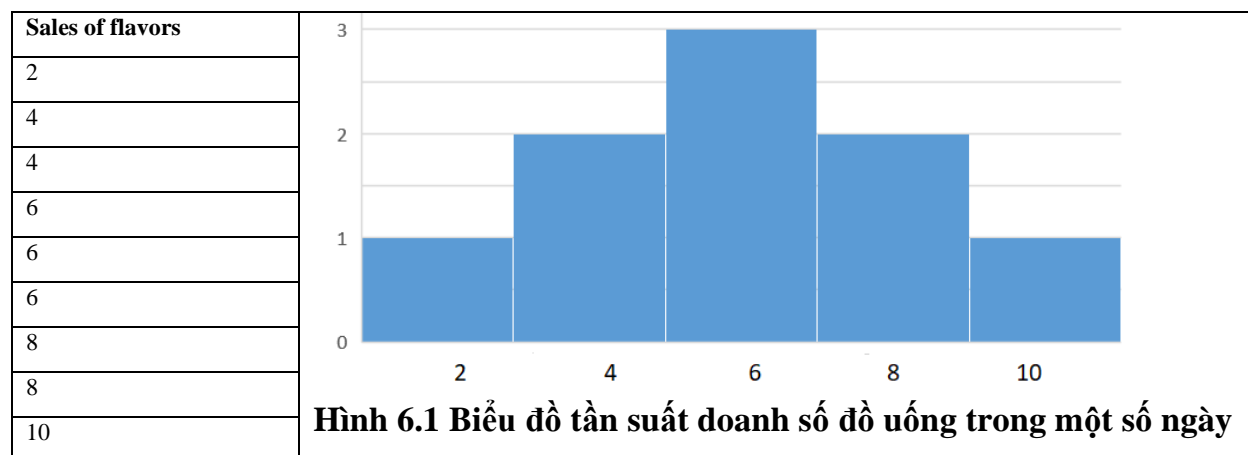
Biến số phân loại: Thực ra là cách dán nhãn cho dữ liệu ví dụ Rosie bán ở Park và Beach, chúng ta có thể hiểu rằng Park là biểu diễn số 1 và Beach là biểu diễn của số 2. Đây là cách con người chúng ta dán nhãn để dễ quản lý.

6.2 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)

6.3 Đo lường xu hướng tập trung

Loại thống kê đầu tiên chúng ta đề cập đến sẽ là thống kê mô tả dựa trên các thông tin cơ bản về dữ liệu. Hãy xem biểu đồ tần suất dưới đây:



6.3.1 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

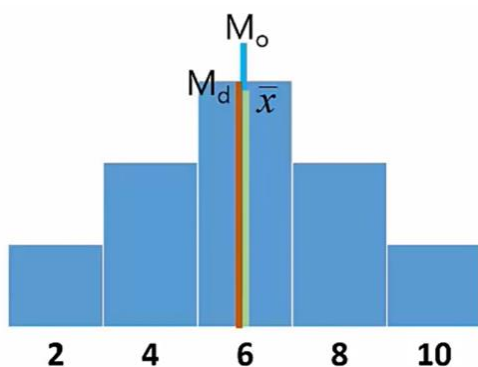
trung bình là 6.

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.

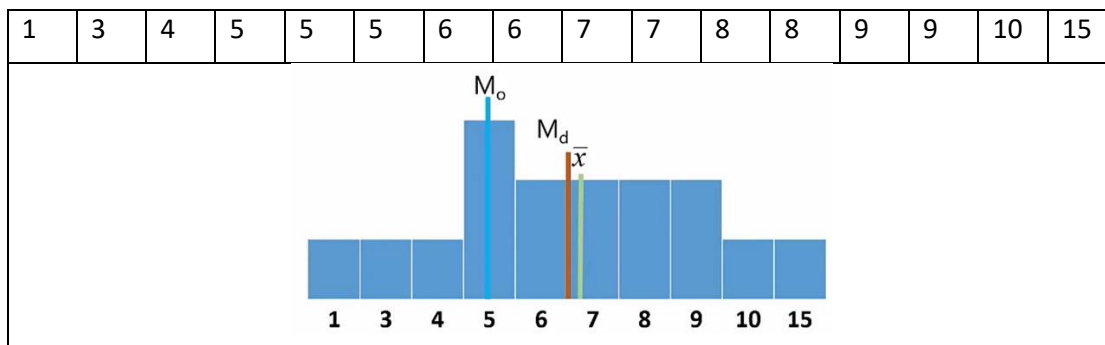
6.3.2 Phân phối chuẩn, phân phối lệch trái/phải

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:



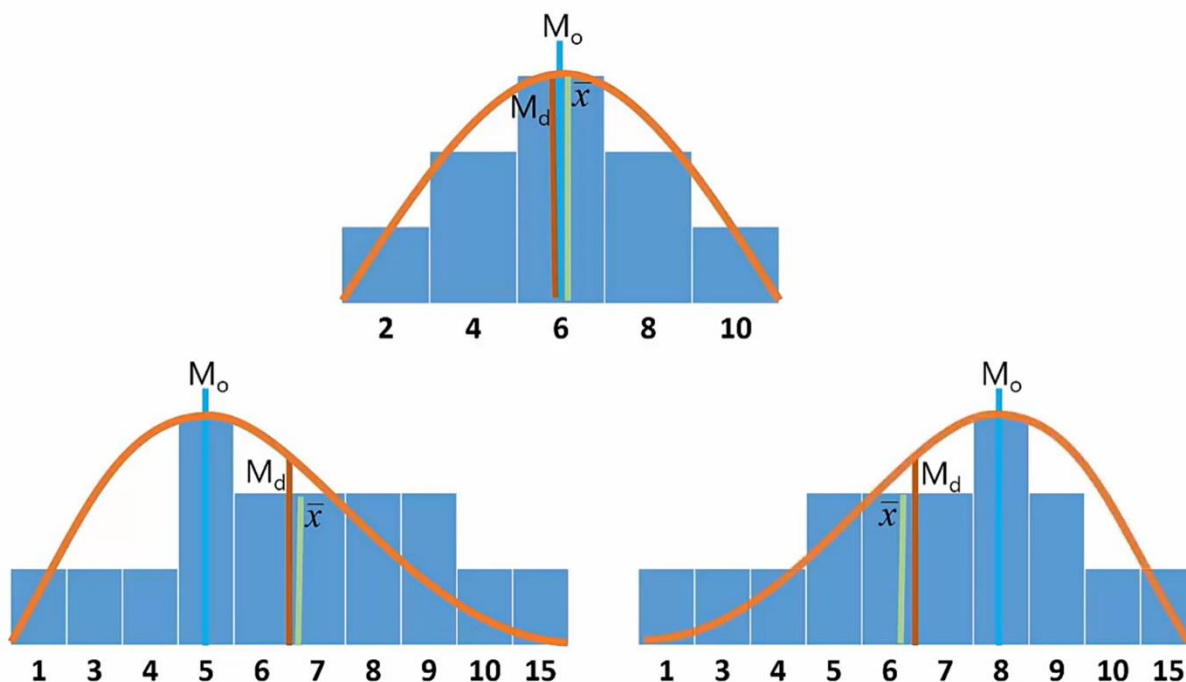
Hình 6.2 Một phân phối chuẩn của dữ liệu

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 Rosie phát ra hàng ngày, liên tục trong 15 ngày.



Hình 6.3 Một phân phối lệch phải của dữ liệu

Các giá trị trung bình, trung vị và yếu vị được tính lần lượt là 6.75, 6.5 và 5. Quan sát ta thấy giá trị trung vị và yếu vị nhỏ hơn giá trị trung bình. Để tổng quát hóa các dạng thức phân phối ta có hình vẽ dưới đây:



Hình 6.4 Các dạng thức phân phối chuẩn, phân phối lệch phải, phân phối lệch trái

6.4 Tính biến thiên của dữ liệu

Mặc dù chúng ta đã xác định được một số điểm quan trọng để biết được xu hướng tập trung của dữ liệu vẫn cần phải tính thêm tính biến thiên hay nói cách khác là độ chênh lệch giữa các dữ liệu.

6.4.1 Phạm vi giá trị

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.

6.4.2 Phương sai tổng thể / phương sai 1 mẫu

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)

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

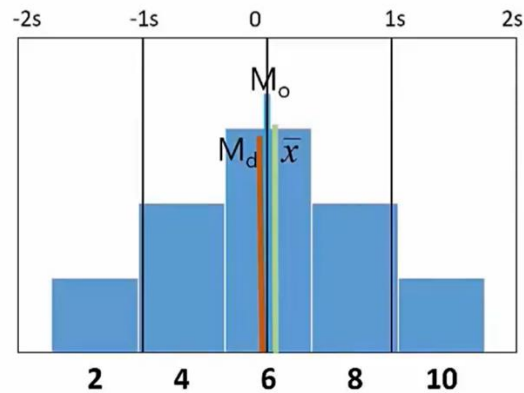
Áp dụng vào trường hợp cụ thể với ví dụ đầu tiết, ta có giá trị phương sai mẫu là:

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

6.4.3 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 tiên, 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



Hình 6.5 Khoảng lệch chuẩn của dữ liệu

6.4.4 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 kết hợp

Trong tiết trước chúng ta đã quan sát được một số thông tin về dữ liệu. Trong tiết này chúng ta sẽ tìm hiểu về thống kê kết hợp, làm thế nào chúng ta có thể nhìn thấy mối quan hệ giữa những yếu tố trong dữ liệu.

7.1 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 quay lại với ví dụ của chúng ta về cửa hàng đồ uống của Rosie và tìm hiểu xem điều gì có thể giúp cô ấy trong việc cải thiện doanh số. Sẽ có 3 biến số mà chúng ta quan tâm xem liệu chúng có liên quan đến **Sales** hay không. Đầu tiên là **Price**, sau đó là **Temperature** và cuối cùng là **Leaflets**. 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.

Đối với kịch bản nhiệt độ thay đổi ta không nhìn thấy sự thay đổi rõ trong doanh số bán hàng khi quan sát trong một khoảng thời gian. Căn bản là không có mối tương quan nào hay hệ số tương quan là không.

Cuối cùng 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.

7.2 Kiểm định giả thuyết

Chúng ta đã biết cách phát hiện các mối quan hệ tiềm năng giữa hai dữ liệu của bộ dữ liệu nhờ hệ số tương quan. Lưu ý hệ số tương quan không nhất thiết biểu thị quan hệ nguyên nhân – kết quả. Phần này chúng ta sẽ đề cập đến việc phân tích thống kê bắt đầu ở đâu và 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 .

Trong trường hợp ví dụ của chúng ta nếu chúng ta chấp nhận 5% tức là chúng ta chấp nhận rằng có 5% khả năng mà ở đó sự khác biệt chỉ là ngẫu nhiên hay cũng tương đương với

việc nếu chúng ta nhận được kết quả có ý nghĩa thống kê thì 5% khả năng là chúng ta đã sai.

Trong ví dụ của chúng ta nếu $p > 0.05$ tức là giả thuyết không của chúng ta đúng, chẳng có sự khác biệt ở đây, tất cả chỉ là ngẫu nhiên. Với $p < 0.05$ ta có thể loại bỏ “giả thuyết không” và sử dụng “giả thuyết thay thế”.

7.2.1 T-Test / Z-Test

Trong phần này chúng ta sẽ bàn luận về thống kê học so sánh, trong trường hợp cụ thể là so sánh hai bộ dữ liệu khác nhau từ hai mẫu khác nhau.

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 120 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: $t = \frac{\bar{x} - \mu}{s / \sqrt{n}}$

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.

7.2.2 T-Test trung bình hai mẫu

Ở phần trước chúng ta thực hiện kiểm định T-Test/ Z-Test cũng như so sánh tổng doanh số bán hàng của năm nay và năm trước. Trong trường hợp đó chúng ta chỉ nhìn vào doanh số bán hàng tức là chỉ xem xét một yếu tố. Trong trường hợp dữ liệu có nhiều hơn hai yếu tố thì chúng ta phải sử dụng công cụ khác. Đối với ví dụ này ta muốn so sánh hai dữ liệu doanh số của Orange và Lemon. Công thức tính T-Test trung bình hai mẫu như sau:

$$\frac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \quad s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

df: $(n_1 + n_2 - 2)$

Sau khi lựa chọn công cụ là T-Test trung bình hai mẫu với cột Lemon và Orange ta sẽ có bảng kết quả sau đây:

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	

Bảng 7.1 kết quả T-Test trung bình hai mẫu

Qua quan sát ta thấy giá trị t Stat mà ta nhận được lớn hơn t quan trọng kiểm định một bên và t quan trọng kiểm định hai bên, chỉ số thống kê p cho cả kiểm định 1 bên và 2 bên đều rất nhỏ. Như vậy giả thuyết thay thế được chấp nhận, có sự khác biệt về doanh số giữa nước chanh và nước cam.

7.2.3 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. Công thức toán học của kiểm định T-test cặp đôi:

$$t = \frac{\bar{d} - D_0}{s_d / \sqrt{n}}$$

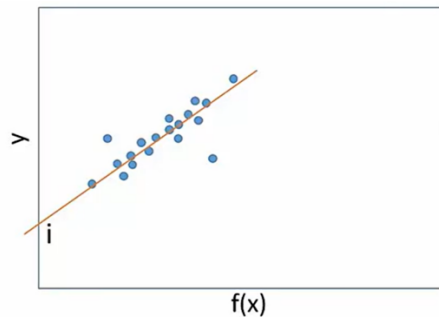
$$df = n - 1$$

Cách đọc giá trị p trên bảng cũng giống như cách đọc đối với kiểm định T-test trung bình hai mẫu.

7.2.5 Hồi quy

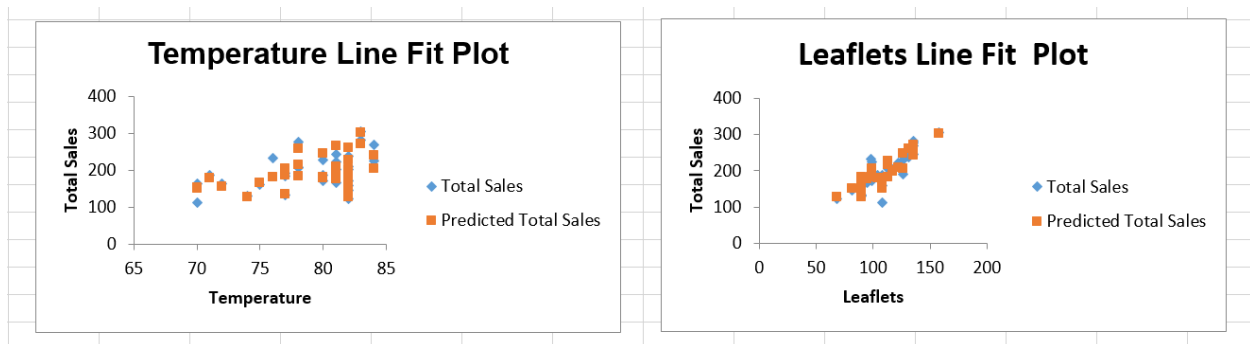
Ở phần cuối của khóa học chúng ta sẽ được học về hồi quy, ví dụ như sử dụng dữ liệu qua doanh số bán hàng trong quá khứ để dự đoán doanh số bán hàng trong tương lai. Nó rất hữu ích trong việc giúp xác định những biến nào trong dữ liệu của bạn có thể được sử dụng để dự đoán một số kết quả mà bạn quan 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.



Hình 7.1 Hình ảnh dữ liệu hồi quy

Chúng ta có thể vẽ rất nhiều đồ thị hàm số hồi quy 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:



Hình 7.2 So sánh dữ liệu hồi quy của Temperature và Leaflets so với Sales

Tiết 8. Thực hành với thống kê

8.1 Thống kê mô tả

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

Bảng 8.1 Các loại thống kê mô tả

8.2 Thống kê kết hợp

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	Sales (năm 2015), Sales (năm 2014)	Data → Data Analysis → T-Test Paired Two Sample for Means, lựa chọn Labels, New Worksheet
4	Correlation	Leaflets, Price, Temperature, Sales	Data → Data Analysis → Regression, lựa chọn Labels, New Worksheet, Residual, Residual Plots, Standardized Residuals, Line Fit Plots

Bảng 8.2 Các loại thống kê kết hợp