Chapter 2: Visual Description of Data

EL Mechry, EL Koudouss

Fordham University

February 5, 2021

Introduction

In this chapter we will cover:

- Frequency distributions.
- Histograms.
- Bar Charts
- Line Graphs
- Pie Charts
- Scatter Diagrams

Frequency distribution

Data usually comes raw: not arranged or organized in any meaningful manner.

The frequency distribution is a table that:

- Divides the data values into classes, and
- Shows the number of observed values that fall into each class.

Frequency distribution: Example

The following data shows the age at death of 58 English monarchs.

30 34 55 40 68 43 15 13 67 77 24 46 47 35 68 50 57 59 70 90
48 50 32 24 65 41 16 69 82 68 24 67 49 52 66 42 44 51 43 82
45 48 27 33 62 53 78 19 71 56 56 65 56 59 33 33 72 50

To make sense of this data, we start by sorting it from smallest to largest.

13	15	16	19	24	24	24	27	30	32	33	33	33	34	35
40	41	42	43	43	44	45	46	47	48	48	49	50	50	50
51	52	53	55	56	56	56	57	59	59	62	65	65	66	67
67	68	68	68	69	70	71	72	77	78	82	82	90		

Now you can see that few kings died within the [10-20), [20-30), and [80-90) age ranges, while most kings died in the [30-40) , [40,50), and [50,60) age ranges.

Frequency distribution: Example, continued

The frequency distribution, a step by step construction:

- Sort the data
- Divide the data to classes or categories. The number of classes chosen is a judgment call.
- Classes must be mutually exclusive (they don't overlap) and exhaustive (they cover the whole range of the data).
- Count the *frequency* of observations in each class. That is, count the number of data values falling within each class.
- Assemble the classes and their corresponding frequencies in a table, see the table below.

Frequency distribution: Example, continued

Age Range	Frequency
$[10, 20)$	4
$[20, 30)$	4
$[30, 40)$	7
$[40, 50)$	12
$[50, 60)$	13
$[60, 70)$	10
$[70, 80)$	5
$[80, 90)$	2
$[90, 100)$	1

Relative and Cumulative Frequency Distributions

- Relative Frequency Distribution:
  - Describes the proportion or percentage of data values that fall within each category.
  - Useful in comparing two groups of unequal size.
- Cumulative Frequency Distribution
  - Lists the number of observations that are within or below each of the classes.
- Cumulative Relative Frequency Distribution
  - Shows the percentage of observations falling at or below a certain class limit.

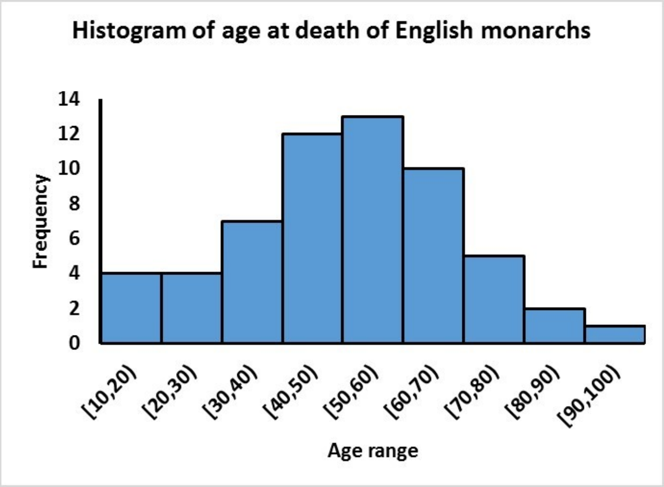
Frequency distribution: Example, continued

Age Range	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
$[10, 20)$	4	4	6.9	6.9
$[20, 30)$	4	8	6.9	13.8
$[30, 40)$	7	15	12.1	25.9
$[40, 50)$	12	27	20.7	46.6
$[50, 60)$	13	40	22.4	69
$[60, 70)$	10	50	17.2	86.2
$[70, 80)$	5	55	8.6	94.8
$[80, 90)$	2	57	3.4	98.3
$[90, 100)$	1	58	1.7	100

The Histogram

- The histogram is a visualization of the frequency distribution.
- It represents each class in the frequency distribution with a rectangle proportional in length to the corresponding frequency.
- That is, a class with frequency 5 gets represented by a rectangle of length 5 and a class with frequency 10 gets represented by a rectangle of length 10.

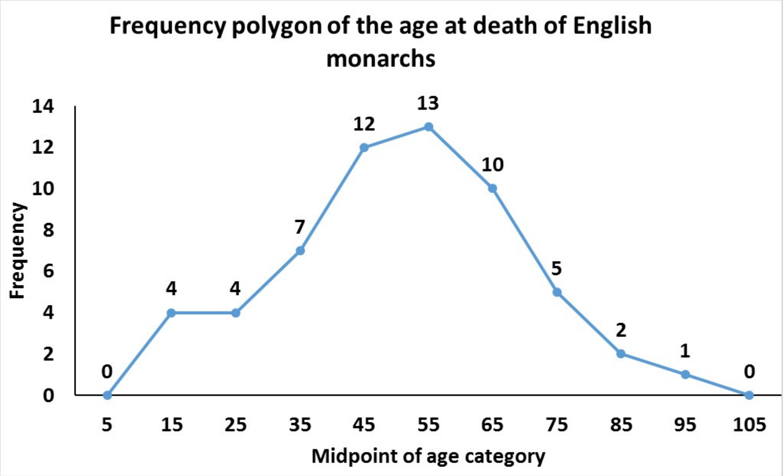
A histogram of the age at death of English monarchs



The Frequency Polygon

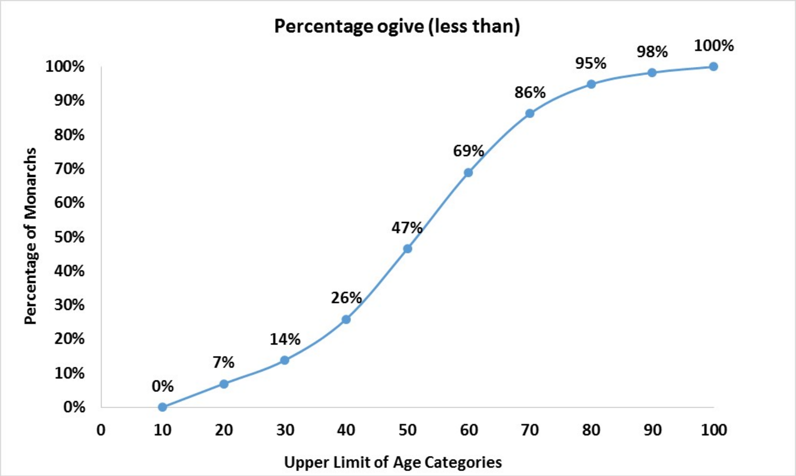
- The frequency polygon consists of line segments connecting the midpoints of each class with its corresponding frequency.
- Relative frequencies or percentages may also be used in constructing the frequency polygon.
- Empty classes are included at each end so the curve intersects the horizontal axis.

The Frequency Polygon: Age at Death

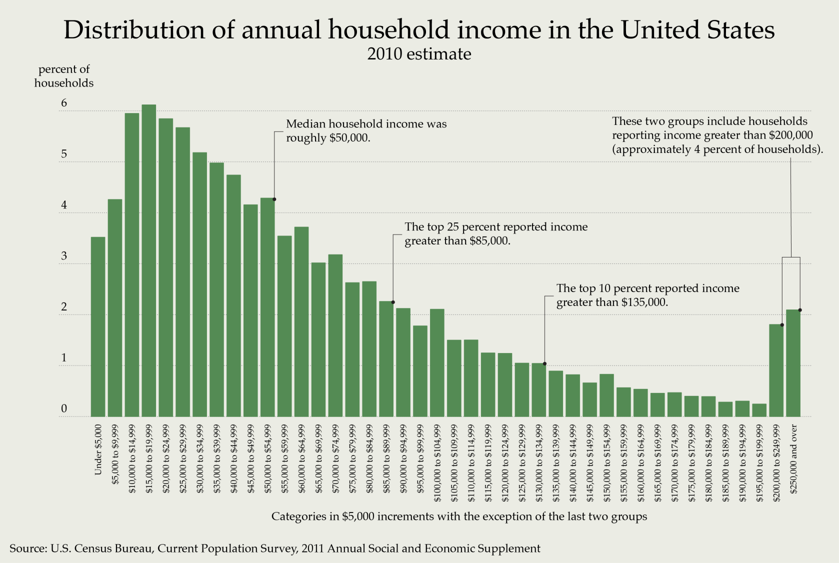


The Ogive: Age at Death

The ogive, a graphical display providing cumulative values for frequencies, relative frequencies, or percentages



Example 1: Histogram of US Income Distribution



Example 2: Student Loans

Using data in the table below:

- Construct frequency distribution table. Use intervals starting at 18900, incrementing by 2200.
- Add relative and cumulative frequencies to the table you created above
- Plot the frequency distribution in a Histogram

Table 1: Average student loans per student by state

UT: 18921	NM: 18969	NV: 20211	CA: 21382	AZ: 22609	LA: 23025	OK: 23430
WY: 23708	HI: 24554	WA: 24804	FL: 24947	CO: 25064	NC: 25218	AR: 25344
TN: 25510	KS: 25521	MO: 25844	KY: 25939	SD: 26023	ID: 26091	OR: 26106
MS: 26177	TX: 26250	NE: 26278	VA: 26432	GA: 26518	AK: 26742	WV: 26854
MT: 26946	ND: 27425	MD: 27457	NY: 27822	NJ: 28318	WI: 28810	IL: 28984
VT: 29060	SC: 29163	IN: 29222	OH: 29353	MA: 29391	AL: 29425	MI: 29450
IA: 29732	CT: 29750	ME: 30908	MN: 31579	RI: 31841	PA: 33264	NH: 33410
DE: 33808	DC: 40885					

Example 2: Student Loans, continued

Loans
$[18900, 21100)$
$[21100, 23300)$
$[23300, 25500)$
$[25500, 27700)$
$[27700, 29900)$
$[29900, 32100)$
$[32100, 34300)$
$[34300, 36500)$
$[36500, 38700)$
$[38700, 40900)$

Example 2: Student Loans, continued

Loans	Frequency
$[18900, 21100)$	3
$[21100, 23300)$	3
$[23300, 25500)$	8
$[25500, 27700)$	17
$[27700, 29900)$	13
$[29900, 32100)$	3
$[32100, 34300)$	3
$[34300, 36500)$	0
$[36500, 38700)$	0
$[38700, 40900)$	1

Example 2: Student Loans, continued

Loans	Frequency	Cumulative Frequency
$[18900, 21100)$	3	3
$[21100, 23300)$	3	6
$[23300, 25500)$	8	14
$[25500, 27700)$	17	31
$[27700, 29900)$	13	44
$[29900, 32100)$	3	47
$[32100, 34300)$	3	50
$[34300, 36500)$	0	50
$[36500, 38700)$	0	50
$[38700, 40900)$	1	51

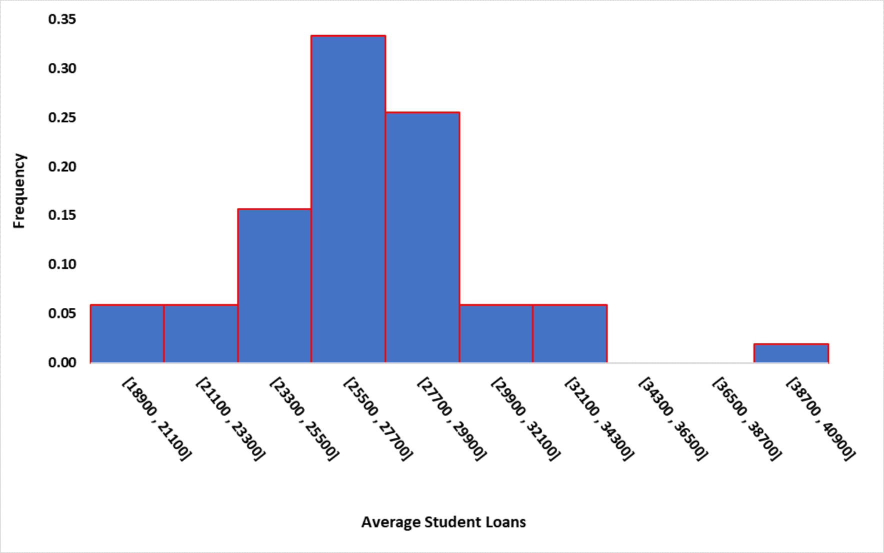
Example 2: Student Loans, continued

Loans	Frequency	Cumulative Frequency	Relative Frequency
$[18900, 21100)$	3	3	0.06
$[21100, 23300)$	3	6	0.06
$[23300, 25500)$	8	14	0.16
$[25500, 27700)$	17	31	0.33
$[27700, 29900)$	13	44	0.25
$[29900, 32100)$	3	47	0.06
$[32100, 34300)$	3	50	0.06
$[34300, 36500)$	0	50	0.00
$[36500, 38700)$	0	50	0.00
$[38700, 40900)$	1	51	0.02

Example 2: Student Loans, continued

Loans	Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Freq.
$[18900, 21100)$	3	3	0.06	0.06
$[21100, 23300)$	3	6	0.06	0.12
$[23300, 25500)$	8	14	0.16	0.28
$[25500, 27700)$	17	31	0.33	0.61
$[27700, 29900)$	13	44	0.25	0.86
$[29900, 32100)$	3	47	0.06	0.92
$[32100, 34300)$	3	50	0.06	0.98
$[34300, 36500)$	0	50	0.00	0.98
$[36500, 38700)$	0	50	0.00	0.98
$[38700, 40900)$	1	51	0.02	1

Example 2, continued



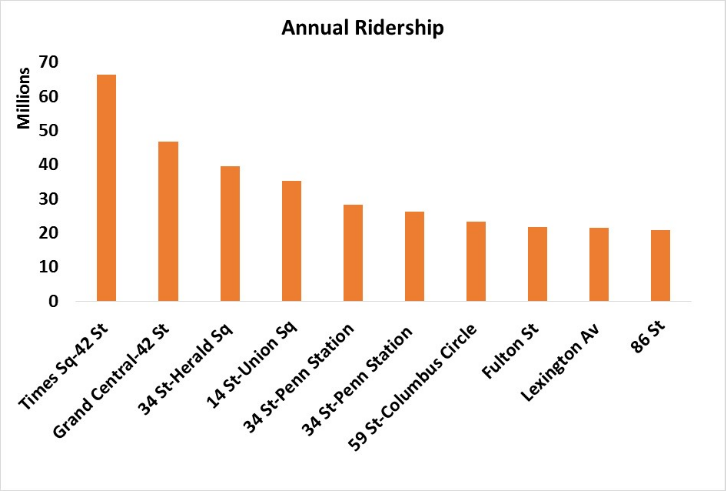
The Bar Chart

The bar chart is usually used to represent qualitative or categorical data.

Example: Consider the following table from MTA on the ten busiest subway stations in 2015.

Station	Annual Ridership
Times Sq-42 St	66,359,208
Grand Central-42 St	46,737,564
34 St-Herald Sq	39,541,865
14 St-Union Sq	35,320,623
34 St-Penn Station	28,309,160
34 St-Penn Station	26,147,434
59 St-Columbus Circle	23,299,666
Fulton St	21,671,684
Lexington Av	21,407,792
86 St	20,890,828

The 10 busiest NYC subway stations, 2015

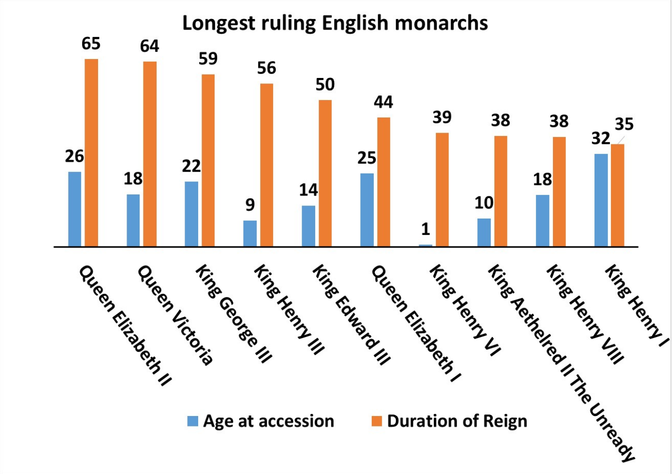


Bar Chart: Another example

The table below shows age at accession and duration of reign for the 10 longest ruling monarchs.

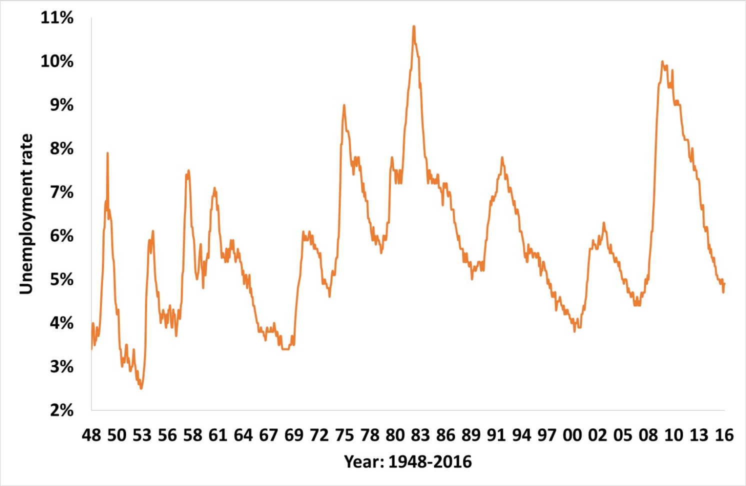
Monarch	Age at accession	Duration of Reign
Queen Elizabeth II	26	65
Queen Victoria	18	64
King George III	22	59
King Henry III	9	56
King Edward III	14	50
Queen Elizabeth I	25	44
King Henry VI	1	39
King Aethelred II The Unready	10	38
King Henry VIII	18	38
King Henry I	32	35

Bar chart: duration of reign & age at accession

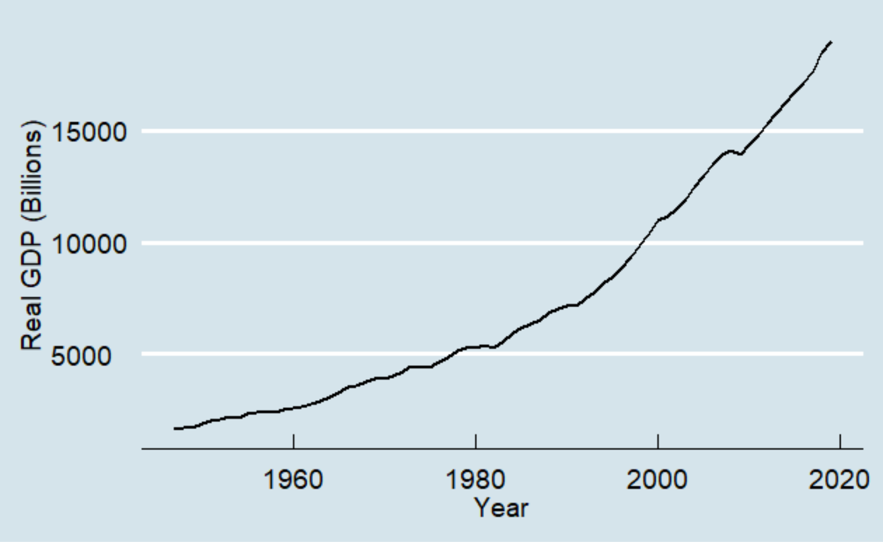


Line Graph: US unemployment

The line graph is good for simultaneously showing values of two quantitative variables.



Line Graph: US Real GDP



The Pie Chart

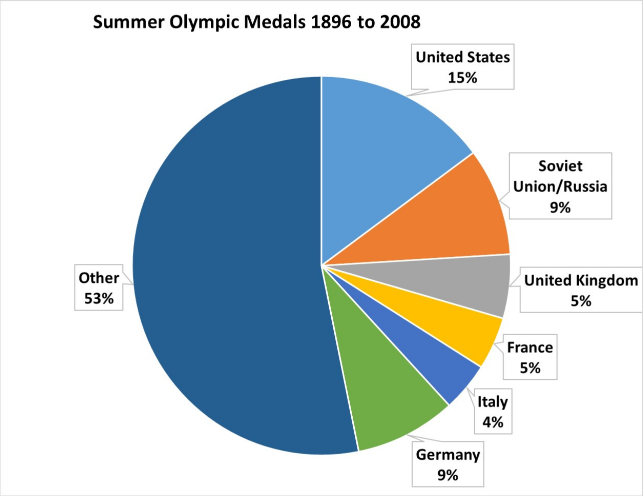
- The pie chart is a circular display divided into sections based on the number of observations within each category.
- It is constructed dividing the 360 degrees of a circle relatively among the categories being compared.
- The angle used for each piece of the pie can be calculated as: Number of degrees for the category = Relative value of the category $\times$ 360

The Pie Chart: Example

The following table shows data for summer Olympic medals from 1896 to 2008.

Country	Medals	Percentage	Angle
United States	4335	15	53°
Soviet Union/Russia	2687	9	33°
United Kingdom	1594	5	20°
France	1314	4	16°
Italy	1228	4	15°
Germany	2526	9	31°
Other	15532	53	191°
Total	29216	100	360°

The Pie Chart: Olympic Medals 1896 to 2008



Pie Chart, Example

The table below shows U.S. federal spending by category. Use this information to create a pie chart.

Sector	Total Spending
Pensions	984
Health Care	1106
Education	120
Defense	115
Welfare	368
Protection	33
Transportation	93
General Government	43
Other Spending	52
Interest	241
Total	3155

Pie Chart example, continued

Sector	Spending
Pensions	984
Health Care	1106
Education	120
Defense	115
Welfare	368
Protection	33
Transportation	93
General Government	43
Other Spending	52
Interest	241
Total	3155

Pie Chart example, continued

Sector	Spending	Share
Pensions	984	0.312
Health Care	1106	0.351
Education	120	0.038
Defense	115	0.036
Welfare	368	0.117
Protection	33	0.010
Transportation	93	0.029
General Government	43	0.014
Other Spending	52	0.016
Interest	241	0.076
Total	3155	1.000

Pie Chart example, continued

Sector	Spending	Share	Angle
Pensions	984	0.312	112
Health Care	1106	0.351	126
Education	120	0.038	14
Defense	115	0.036	13
Welfare	368	0.117	42
Protection	33	0.010	4
Transportation	93	0.029	11
General Government	43	0.014	5
Other Spending	52	0.016	6
Interest	241	0.076	27
Total	3155	1.000	360

The scatter diagram, or Scatterplot

The scatter plot is a chart of two variables plotted against each other. Usually the scatter plot tells a story, most often revealing a correlation (positive or negative) in a large amount of data.

The two variables are referred to as the *dependent* variable, y, and the *independent* variable, x.

We are usually interested in how x predicts y.

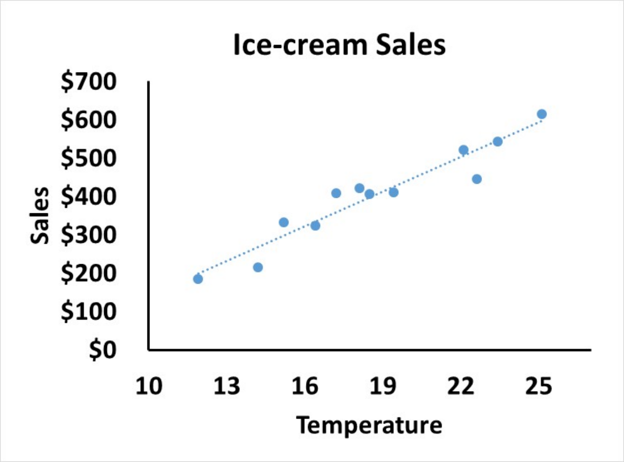
We can fit a “best fit” line in the scatter plot.

The direction of the best fit line determines whether the relationship between the two variables is *direct (positive)*, *inverse (negative)*, or nonexistent. The shape of the line determines whether the relationship is *linear* or *non-linear (curvilinear)*

Temperature C°	Ice Cream Sales	Hot Chocolate Sales	Onion Sales	Electricity Bill
11.9°	185	603	463	500
14.2°	215	549	475	450
15.2°	332	533	457	300
16.4°	325	473	501	250
17.2°	408	455	465	200
18.1°	421	448	499	150
18.5°	406	445	457	150
19.4°	412	443	501	180
22.1°	522	386	444	250
22.6°	445	381	472	300
23.4°	544	296	496	450
25.1°	614	273	450	500

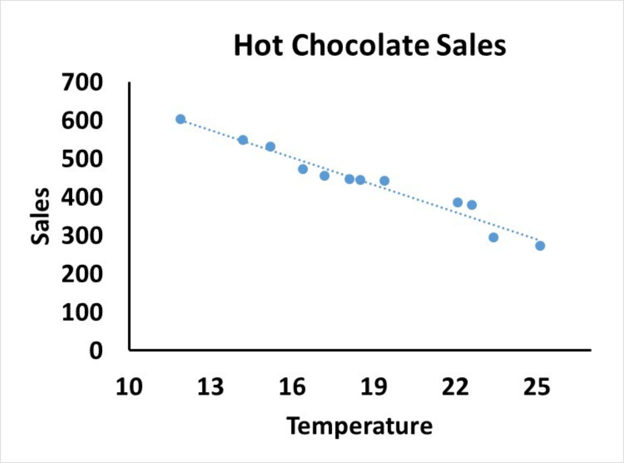
Positive Correlation: Example

The figure below shows a direct (positive) relationship between ice-cream sales and temperature. A direct relationship exists when variables increase and decrease together.



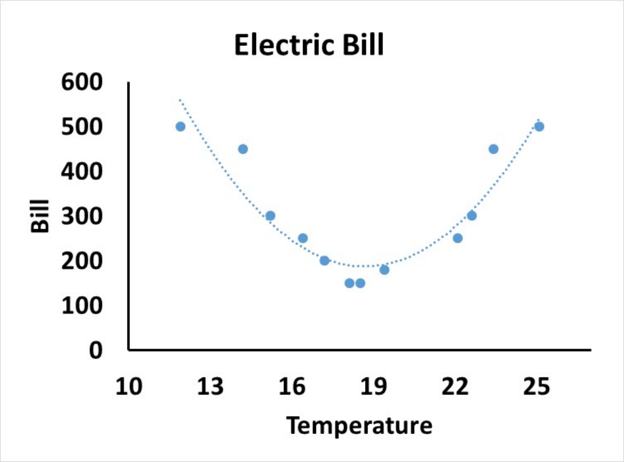
Negative Correlation: Example

The figure below shows an inverse (negative) relationship between hot chocolate sales and temperature. A inverse relationship exists when variables increase and decrease in opposite directions.



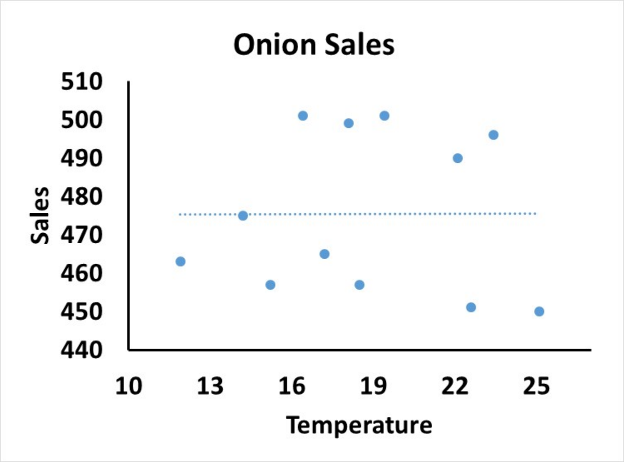
Nonlinear relationships: Example

The figure below shows the relationship between temperature and the cost of electricity. Basically, it shows electric bills are high in extreme temperatures, both cold and hot.



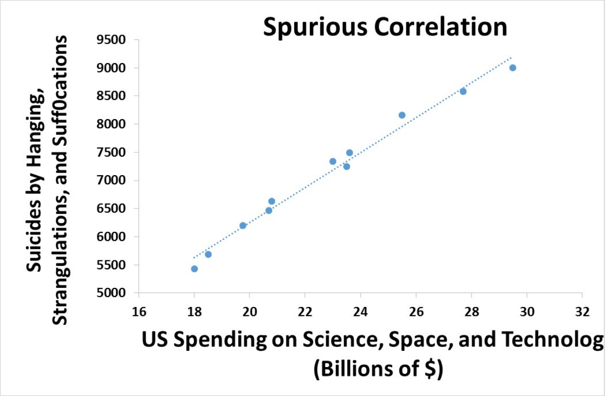
No Correlation: Example

The figure below shows the relationship between onions' sales and temperature. Basically, it shows onion sales are not related to temperature.



Correlation versus causation

If one thing causes the other, then they are most certainly correlated. However, just because two things are correlated doesn't mean one causes the other. When the correlation of two variables is theoretically implausible, we say there is a spurious correlation between these variables.



Examples of some spurious correlations

Discover more fascinating spurious correlations