Descriptive Statistics

Ex 1. A marketing consultant observed 50 consecutive shoppers at a grocery store. Here are the amounts that each shopper spent (in dollars).

18.71	32.82	37.52	33.26	6.90
31.99	39.28	69.49	19.55	12.66
27.07	63.85	34.76	20.89	16.55
23.85	30.54	40.80	52.36	15.01
14.35	14.52	20.58	33.80	13.72
36.22	29.15	43.97	45.58	15.33
21.13	14.55	13.67	61.57	18.30
20.91	64.30	11.34	18.22	17.15
2.32	26.04	28.76	8.04	9.45
19.54	11.63	6.61	12.95	10.26

It is really hard to get any information staring at a data set like this. We have to summary the data somehow.

(1) Frequency table. We could separate the range of the data into several intervals and count how many cases in each interval.

	Freq.	Relative Freq.
0<=x<10	5	0.10
10<=x<20	19	0.38
20<=x<30	9	0.18
30<=x<40	9	0.18
40<=x<50	3	0.06
50<=x<60	1	0.02
60<=x<70	4	0.08

Sometimes half a dozen figures will reveal, as with a lightning-flash, the importance of a subject which ten thousand labored words, with the same purpose in view had left at last but dim and uncertain.

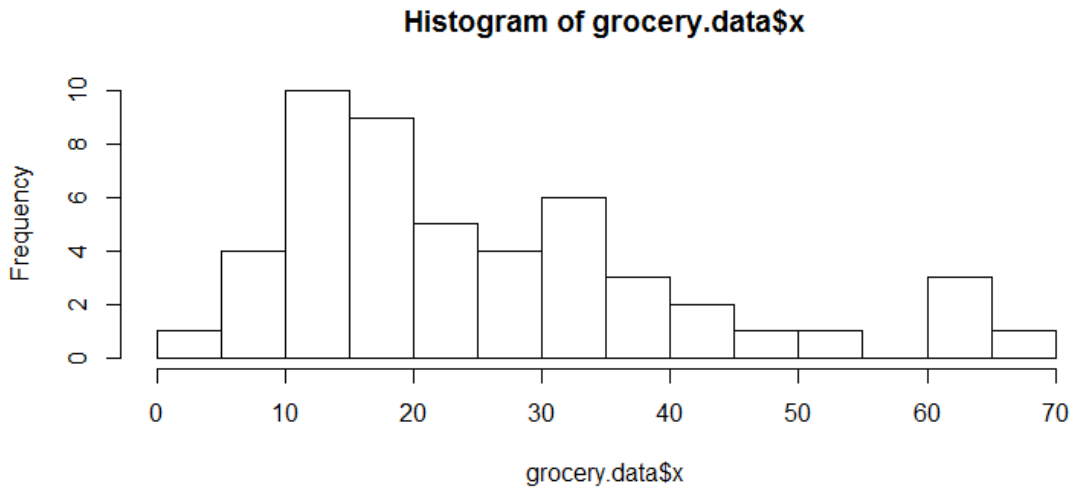
Mark Twain -- *Life on the Mississippi*, 1883.

(2) Stem-and-leaf plot

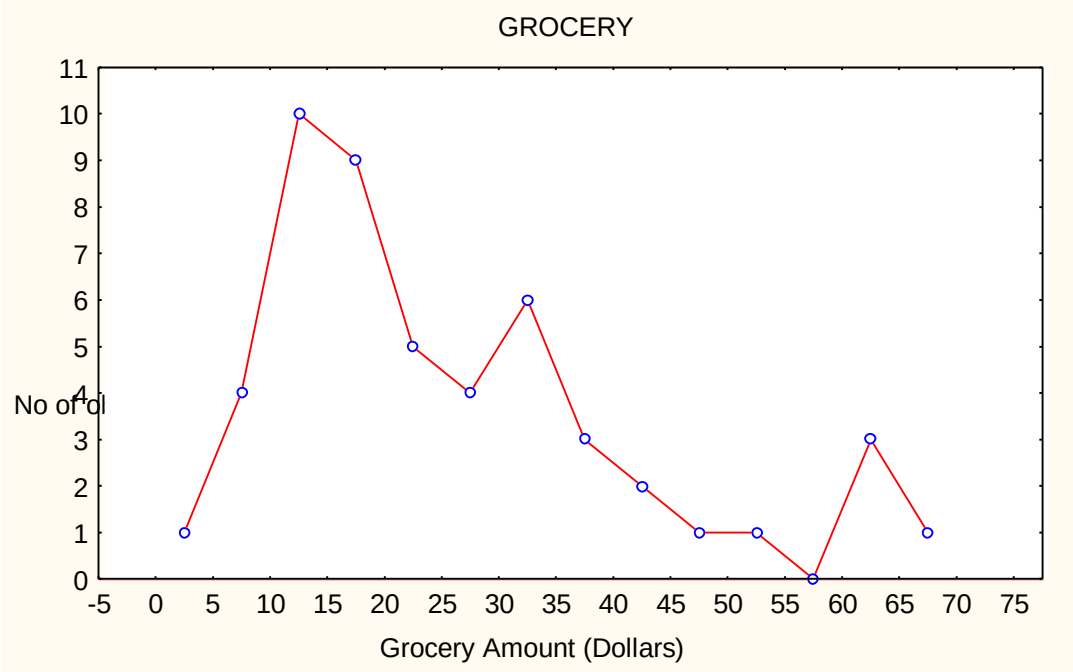
The same data set, separate into intervals of 0-4, 5-9, 10-14, 15-19, ….

Frequency	Stem & Leaf
1.00	0 * 2
4.00	0 . 6689
10.00	1 * 0112233444
9.00	1 . 556788899
5.00	2 * 00013
4.00	2 . 6789
6.00	3 * 012334
3.00	3 . 679
2.00	4 * 03
1.00	4 . 5
1.00	5 * 2
.00	5 .
3.00	6 * 144
1.00	6 . 9

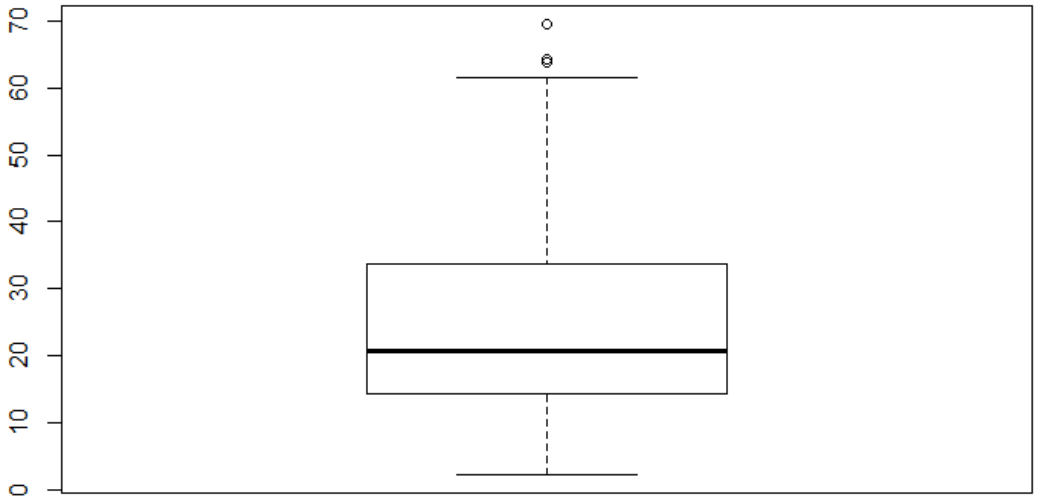
(3) Histogram



(4) Frequency Polygons



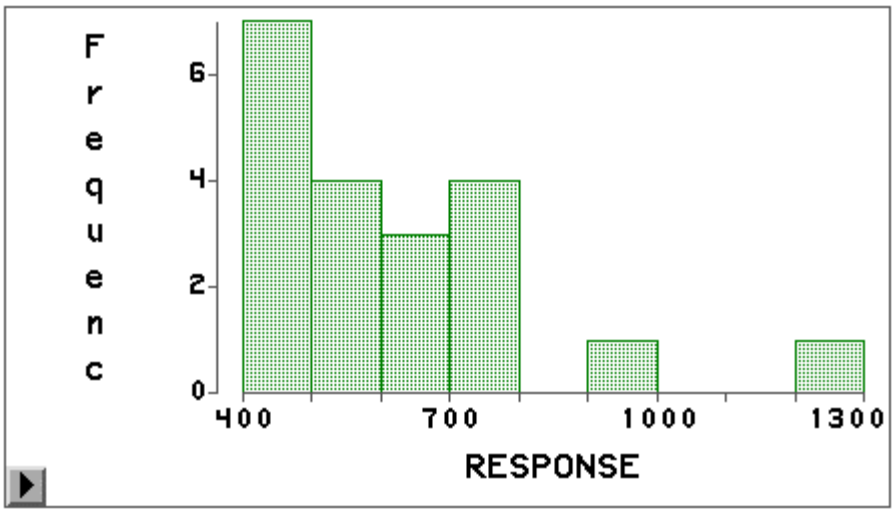
(5) Box Plot



Example: Response time

(https://cos.northeastern.edu/mathematics/~ding/statlab/Comparison/Comp_dist.html)

observations	response time	observations	response time
1	1270	11	660
2	600	12	500
3	710	13	440
4	600	14	490
5	720	15	490
6	930	16	490
7	770	17	550
8	720	18	490
9	490	19	550
10	440	20	500



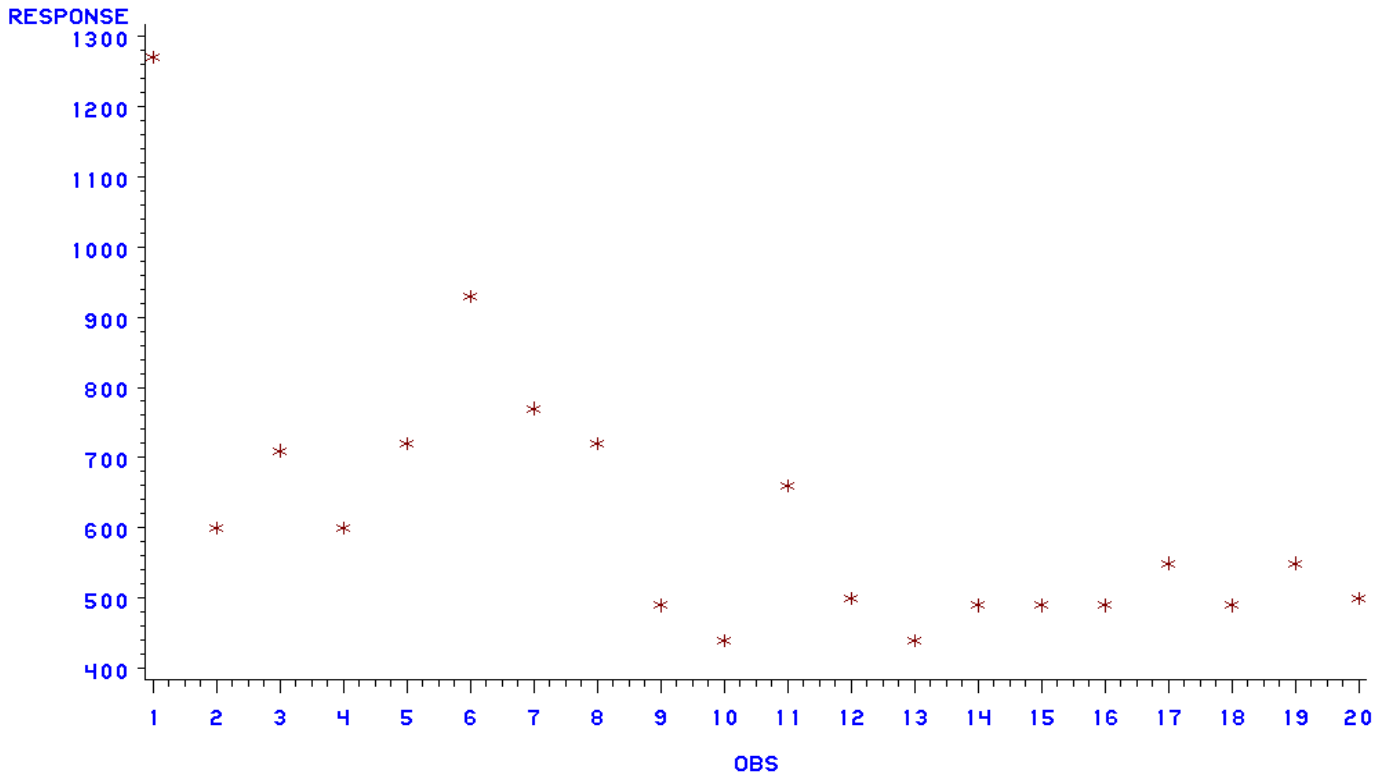
Variation in Response Times

- Judging from the data set, would the following response time be considered ordinary, too small or too large?
- 580 Ok
- 490 Ok
- 20 Too small
- 2000 Too big
- 750 Ok
- 1060 A little big, Ok.

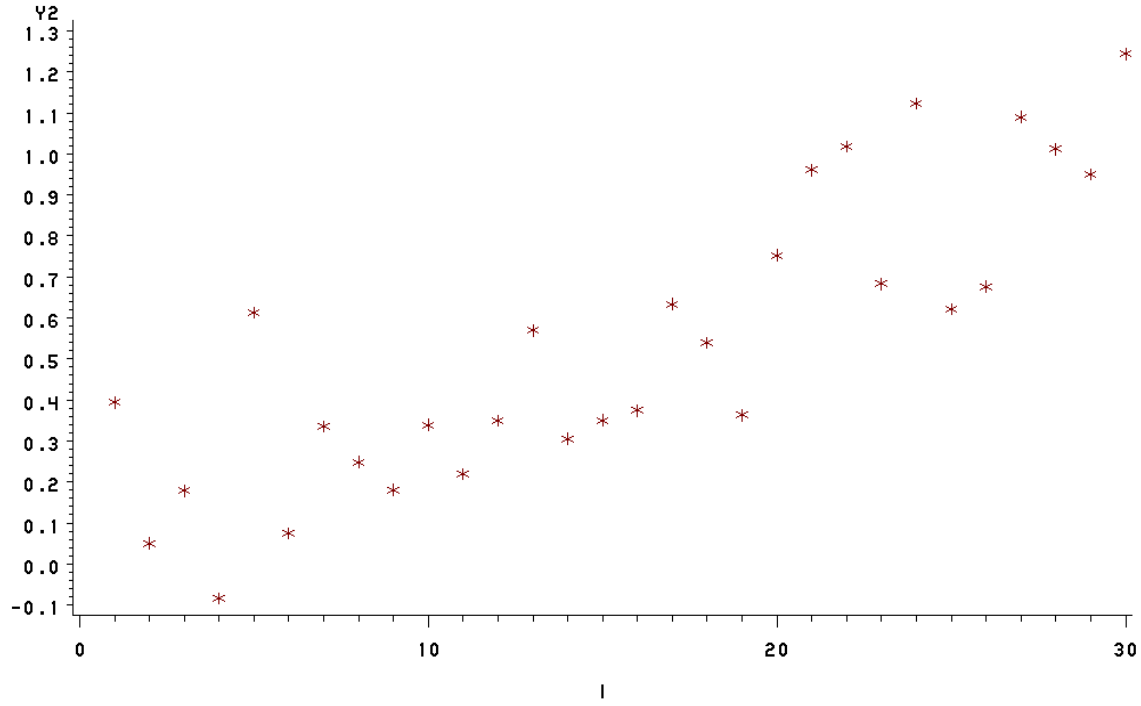
- Summary counts of response times used to construct the histogram

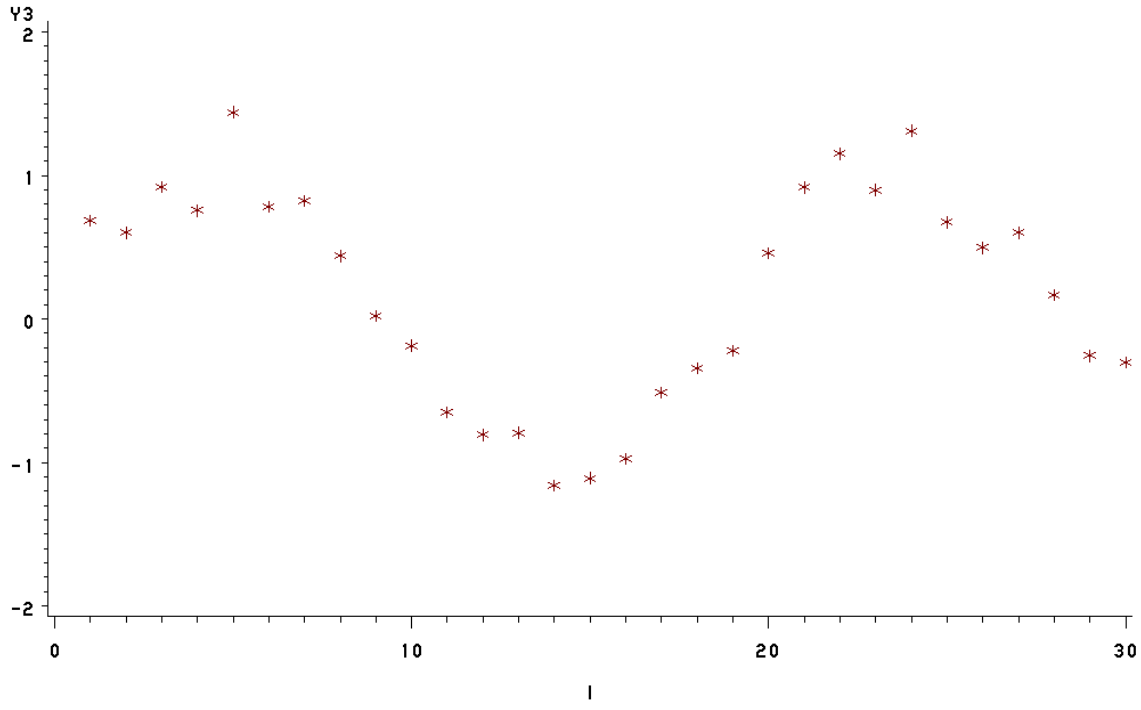
response time	frequency
400-499	7
500-599	4
600-699	3
700-799	4
800-899	0
900-999	1
1000-1099	0
1100-1199	0
1200-1299	1

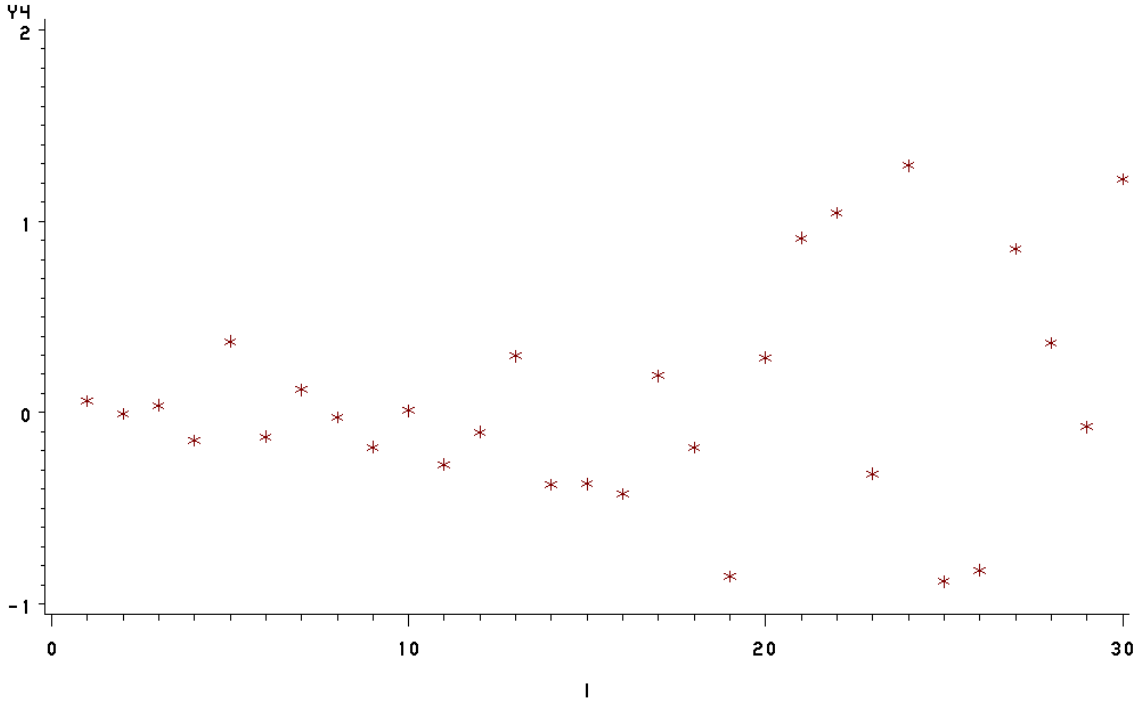
Scatterplot of the response time

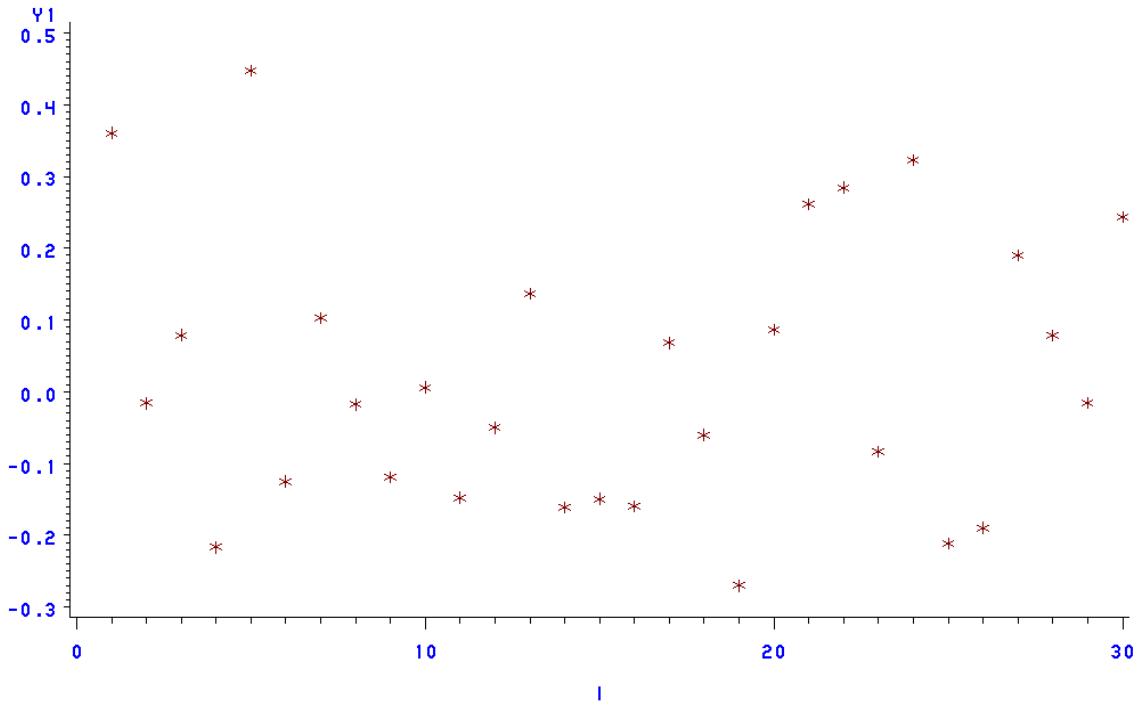


Common patterns in some scatter plots





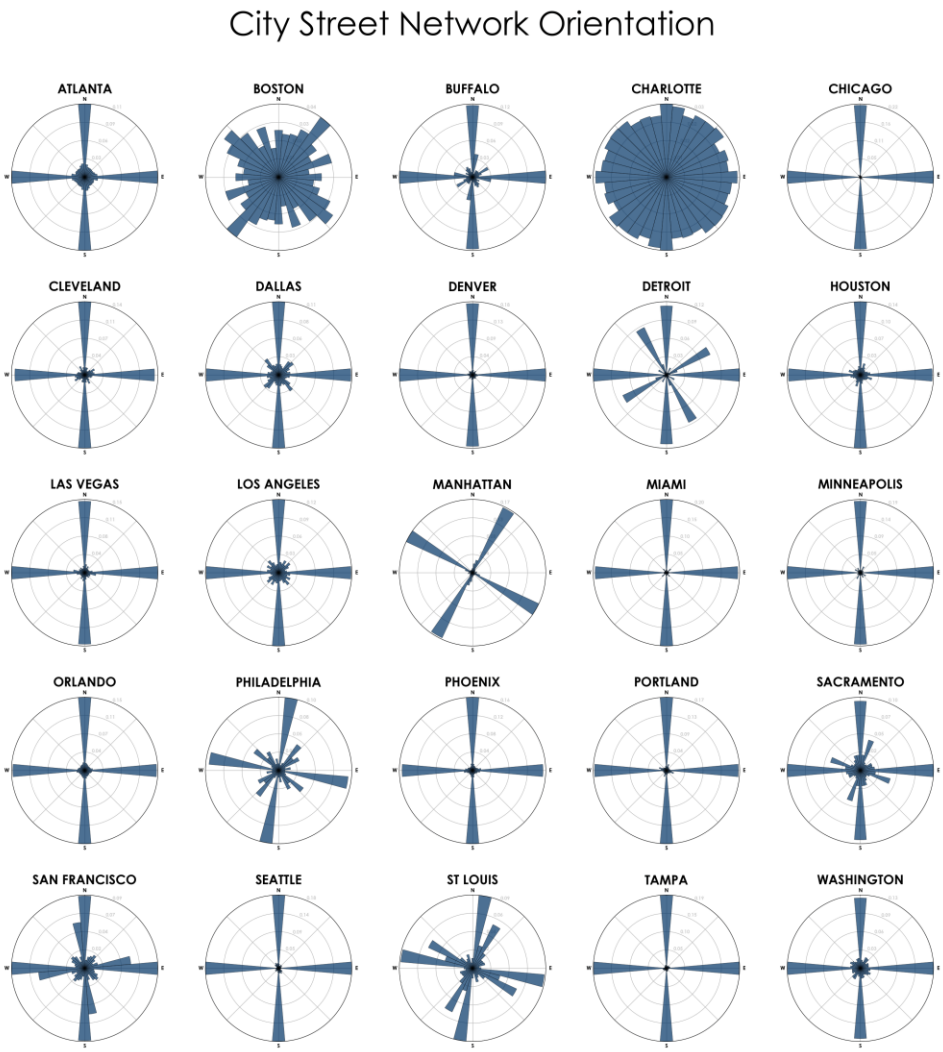

Y4
2
1
0
-1
0
10
20
30
I


Y1
0.5
0.4
0.3
0.2
0.1
0.0
-0.1
-0.2
-0.3
0
10
20
30
I

Data Visualization

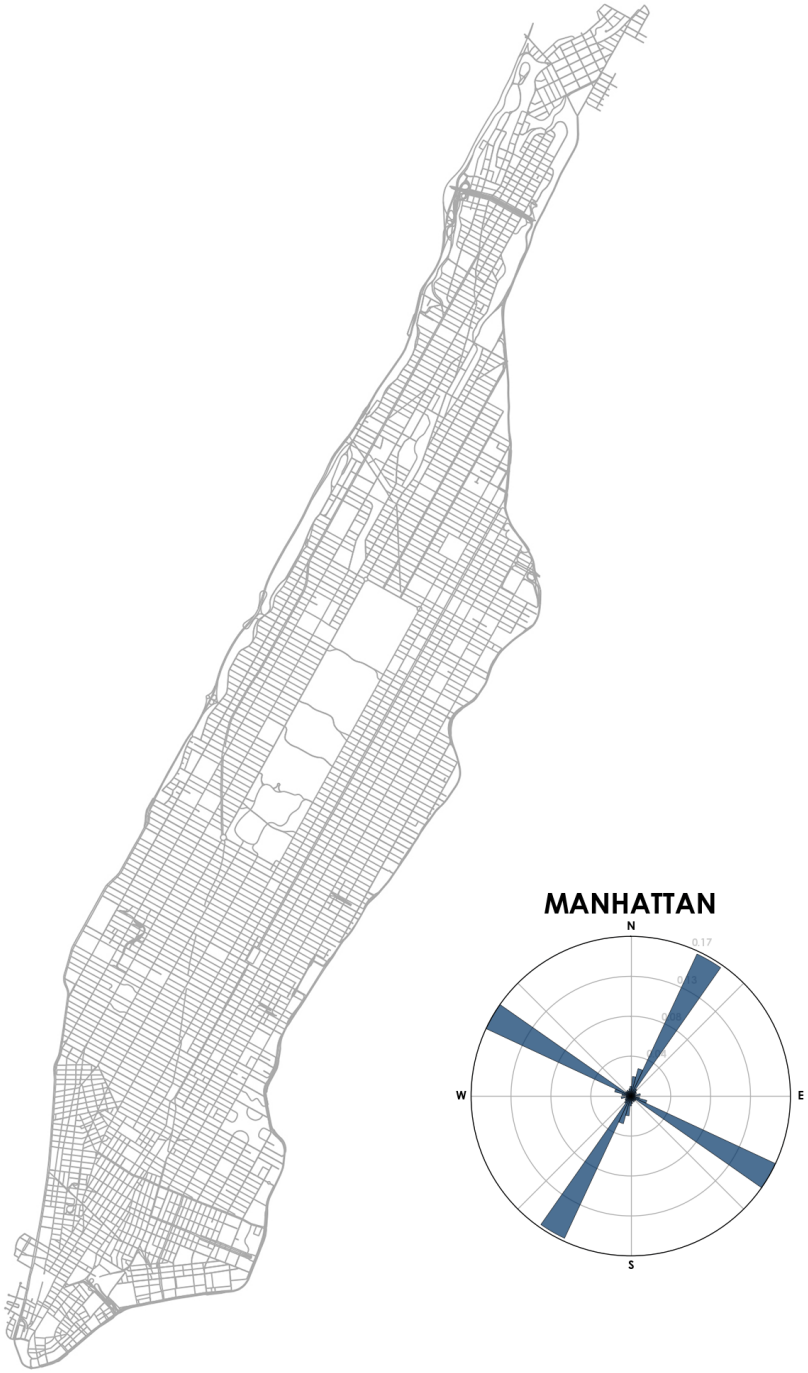
Above are some very basic common plots to display data. Nowadays, data visualization itself is an important evolving research topic. People are inventing ways to display high-dimensional data that can we can visualize patterns. The techniques (e.g. heatmap) are continuously invented and then programmed into common statistical software such as R. You are encouraged to read and learn about those techniques on your own.

Following is an example by Professor Boeing to visualize the distribution of city street orientations. (https://geoffboeing.com/2018/07/comparing-city-street-orientations/)



Each of the cities above is represented by a polar histogram (aka rose diagram) depicting how its streets orient. Each bar's *direction* represents the compass bearings of the streets (in that histogram bin) and its

length represents the relative frequency of streets with those bearings. For example, in Manhattan we can clearly see the angled, primarily orthogonal street grid in its polar histogram:



How to use statistics?

(O) Ages at death for 8 women who divorced within 5 years of their first marriage: 32, 83, 71, 75, 45, 68, 56, 57

Ages at death for 5 women who celebrated Golden Anniversary with their first husband: 83, 72, 85, 94, 74

Does happy 1st marriage leads to longer lifespan?

a. selection bias
Second group at least >= 50 years + legal marriage age
One way to match two groups
1. divorced within 5 years
2. 1st marriage last atleast 5 years

Observational Study
No causation conclusion

(A) Should we use *Mean* or *Median* to measure: standard of living? Oil production? Accident cost for automobile insurance premium setting?

(B) HMO has higher client satisfaction score (mean or median?) than traditional health insurance company. Does this address the criticism that they pressure doctors to avoid expensive procedures <u>needed</u> by the patients? (See the article about one case of insurance coverage denial of cancer patient https://www.cnn.com/2018/08/15/health/cancer-survivor-insurance-denial-battle/index.html)

Since the percentage of population who'd require expensive procedures are very less, data is not normalized. To do this study, survey should be done for only ill people who requires this procedure. Also people who aren't happy with the insurance would just leave and get with another company, that means unsatisfied consumers would not be a part of this study, to avoid that, survey should be taken from all past and current consumers.

(C) Air (travel) safety, see paper in next two pages.
Which statistics should be used to measure safety?

How to average?
1. # fatalities / # flights
2. # fatalities / # mileages
3. # fatalities / # passengers * # mileages
4. # fatalities / # passengers * duration

Does technology make life more dangerous?
1. No fatal traffic accidents before automobiles
Is it safer before?
fatalities / #trips * person (safer before)
fatalities / duration * persons
fatalities / mileage * persons (safer now)

Standard of Living
Mathematical Property:
Robustness

Average:
Median: Robust (When salary is reported, we should always use median and not average, 'cause average can be highly affected due to outliers.

Oil Production
Mathematically it's not very Robust, so we use average to find about the oil production

Auto Insurance Premium
Company POV: Average cost per accident
Consumer POV: Median Cost - 95 percentile
Comprise: Truncated Mean

Air Safety Seeks to Keep Up With Growth

New York Times Dec. 9, 1996.

Record Crash Death Toll in '96, But Statistically Travel Is Safer

By ADAM BRYANT

With three weeks to go in 1996, more passengers have died in airline crashes this year than in any other, even though statistics show that air travel is becoming safer over time.

This year's high death toll is in part a result of the continuing rapid increase in the number of flights worldwide and, with it, the chances for an accident.

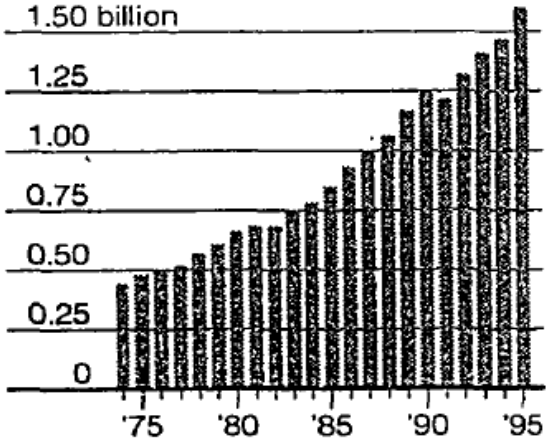
"Flying isn't becoming inherently more dangerous," said Stuart Matthews, president of the Flight Safety Foundation, an organization based in Alexandria, Va., that is supported by the airline industry. "But because we are getting significantly more flying, we're just going to see more and more accidents."

According to Airclaims, a London-based company that collects accident data, 1,187 passengers have been killed on commercial jet flights this year. That figure excludes deaths from terrorist acts and from crashes of long-troubled Soviet-built planes.

Industry experts are quick to note that annual numbers for passenger deaths are notoriously volatile. In 1984, the year before the previous record of 1,169 was set, just 2 passengers were killed worldwide in Western-built jets.

Continued on Page B10, Column 1

Passengers carried each year.



Source: Airclaims Limited

The New York Times

AIR SAFETY

Fatalities Rise, but Not the Risk of Flying

This year is already the deadliest on record for air travel, although flying remains the safest form of transportation. Even so, the risk is not the same throughout the world. Figures are for Western-built jets only.

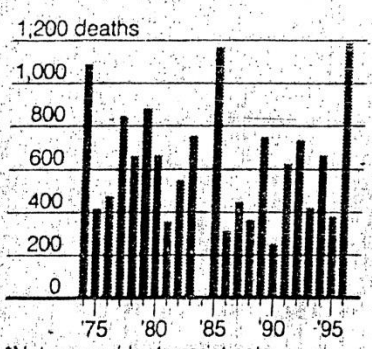
ACCIDENT RATES AROUND THE WORLD

For the five years ended June 30, 1996

REGION	FATAL AIRLINER ACCIDENTS	FLIGHTS (THOUSANDS)	FATAL ACCIDENTS PER MILLION FLIGHTS
Europe	8	17,261	**0.463**
Australia	0	1,958	**0.000**
North America and Caribbean	6	36,194	**0.166**
Africa	5	2,342	**2.135**
South and Central America	5	5,837	**0.857**
Asia	14	11,111	**1.260**
WORLD	**38**	**74,703**	**0.509**

YEARLY DEATH TOLLS VARY . . .

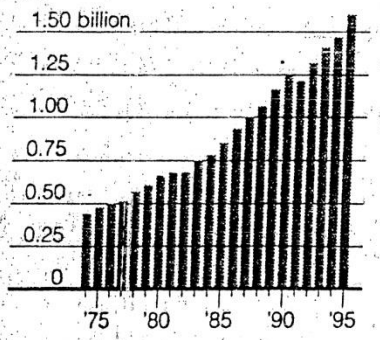
Passenger deaths.*



*Not caused by terrorist acts

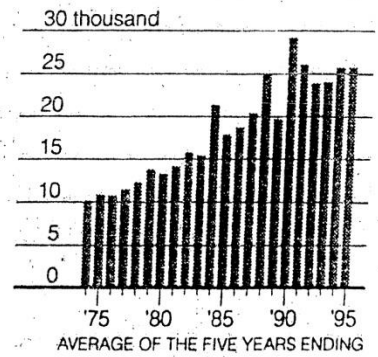
. . . BUT MORE PEOPLE ARE FLYING

Passengers carried each year.



. . . AND IT'S GETTING SAFER

Safe flights per passenger fatality.



Sources: Airclaims Limited (historical flight and accident data); Ronald Ashford, using Airclaims data (regional accident rates)

The New York Times