Stats + Python

Week 1 - Day 04

Descriptive Statistics

Vs.

Inferential Statistics

Basic statistics

Mean/Average

Median

Mode

Min/Max

Quartiles/Percentiles

Values are sorted from min to max

Median - split in 2

Quartile - split in 4

1st quartile

2nd quartile = median

3rth quartile

Median - split in 2

Quartile - split in 4

Percentile - split in 100

1st quartile = 25th percentile

2nd quartile = 50th percentile = median

3rth quartile = 75 percentile

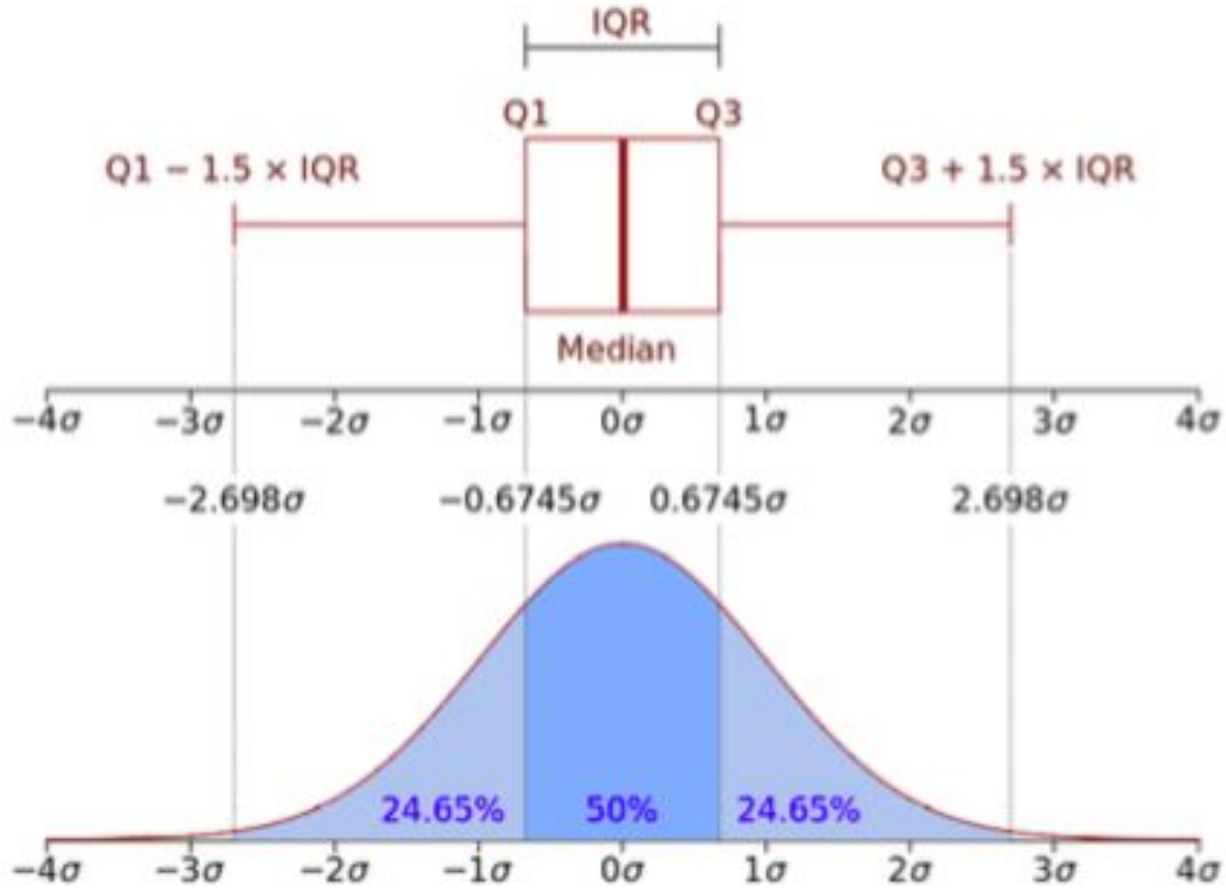
What’s the value of the 5th percentile?

What’s the value of the 95th percentile?

InterQuartile Range (IQR)

=

3th quartile - 1st quartile


IQR
Q1
Q3
Q1 − 1.5 × IQR
Q3 + 1.5 × IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
−2.698σ
−0.6745σ
0.6745σ
2.698σ
24.65%
50%
24.65%

Measures of Dispersion

How spread is our data?

[98, 99, 101, 102]

Vs.

[1, 2, 198, 199]

Range = max-min

Variance = var

Standard deviation = std = sqrt(var)

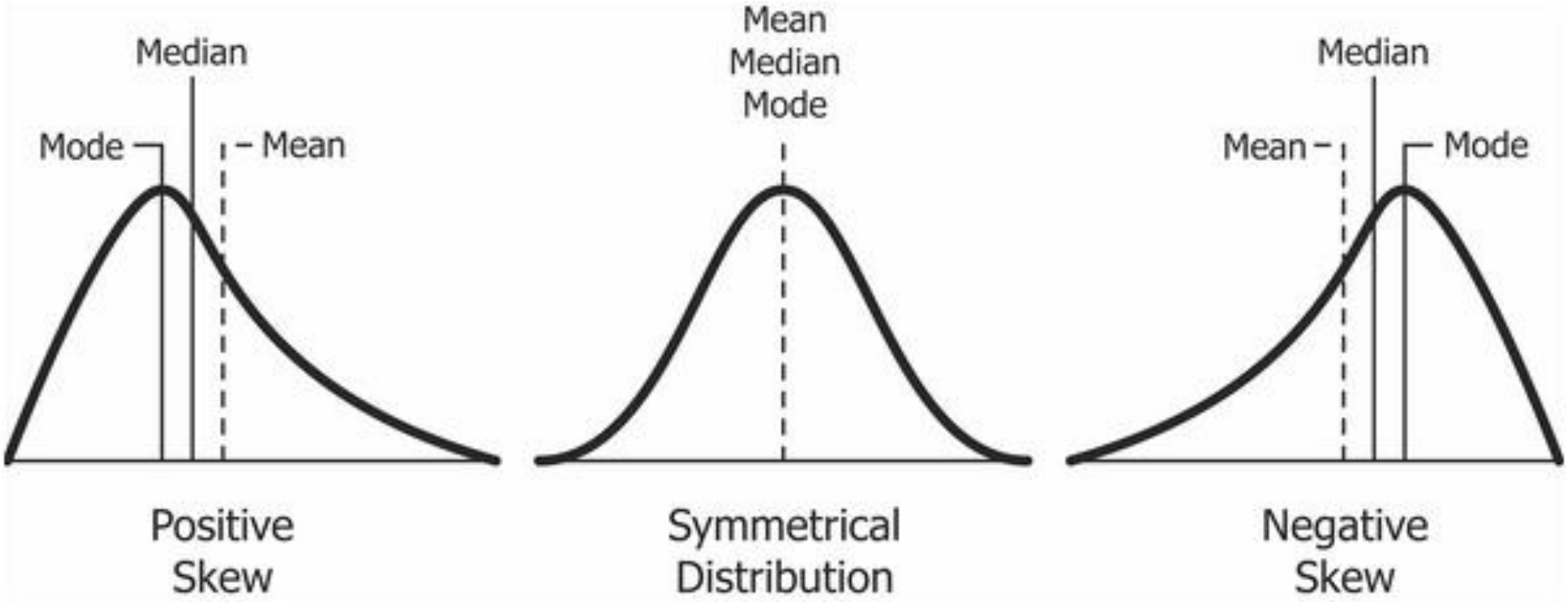
[98, 99, 101, 102], mean = 100

[-2, 1, 1, 2]

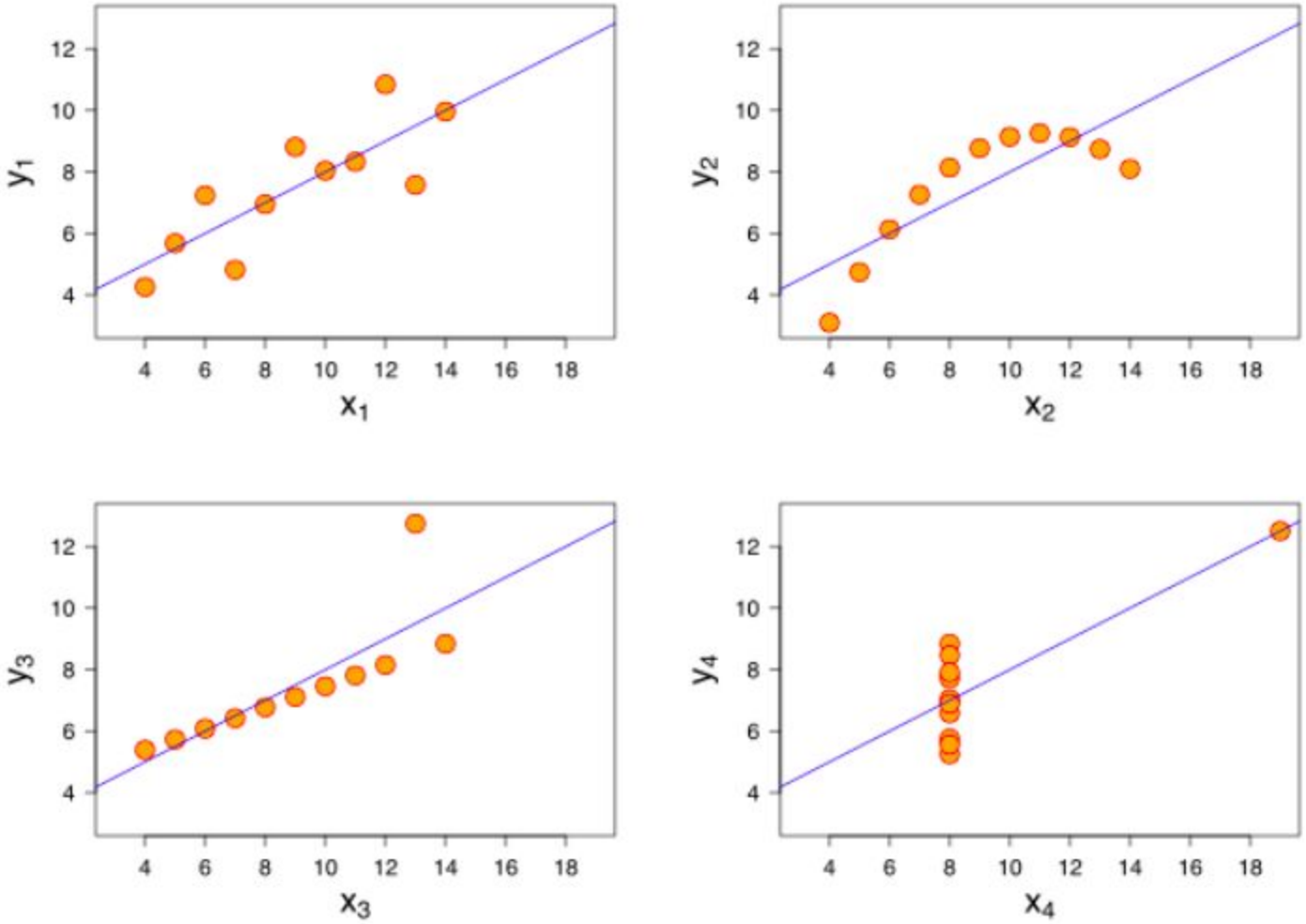
[4, 1, 1, 4]

sum([4, 1, 1, 4]) / len([4, 1, 1, 4]) = 2.4

sqrt(2.4) = 1.58

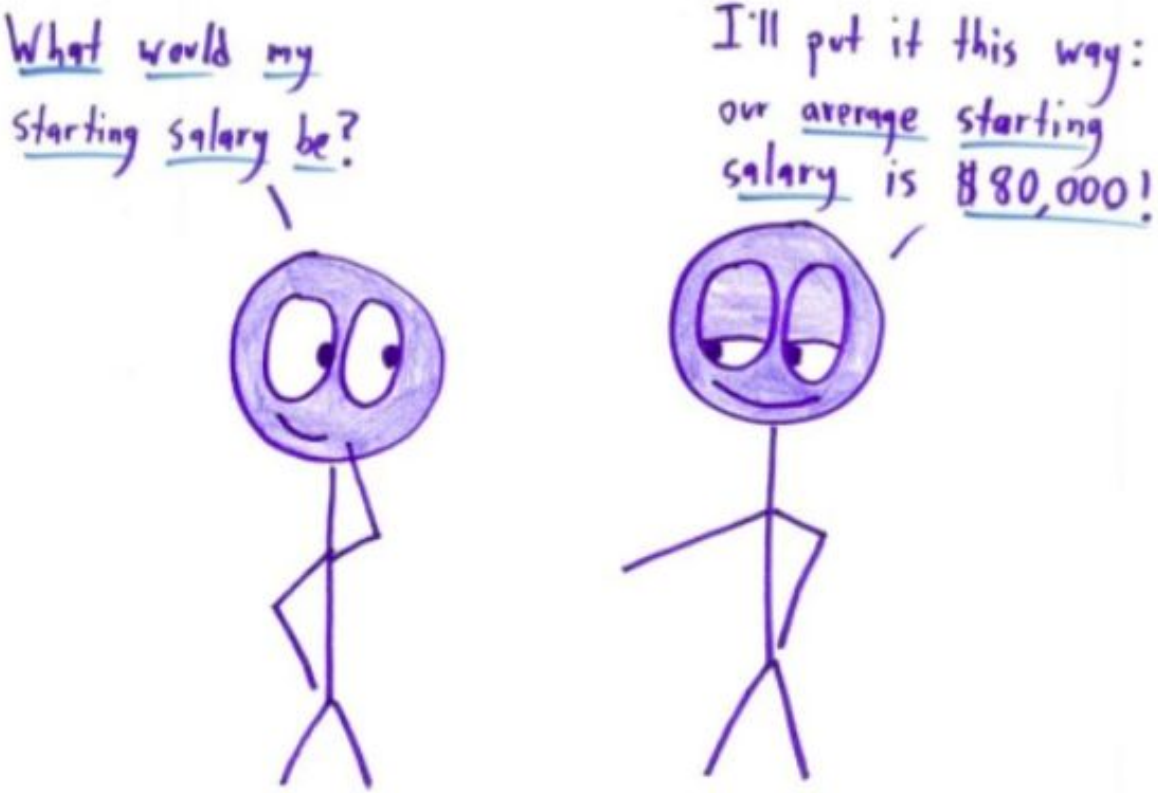

Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

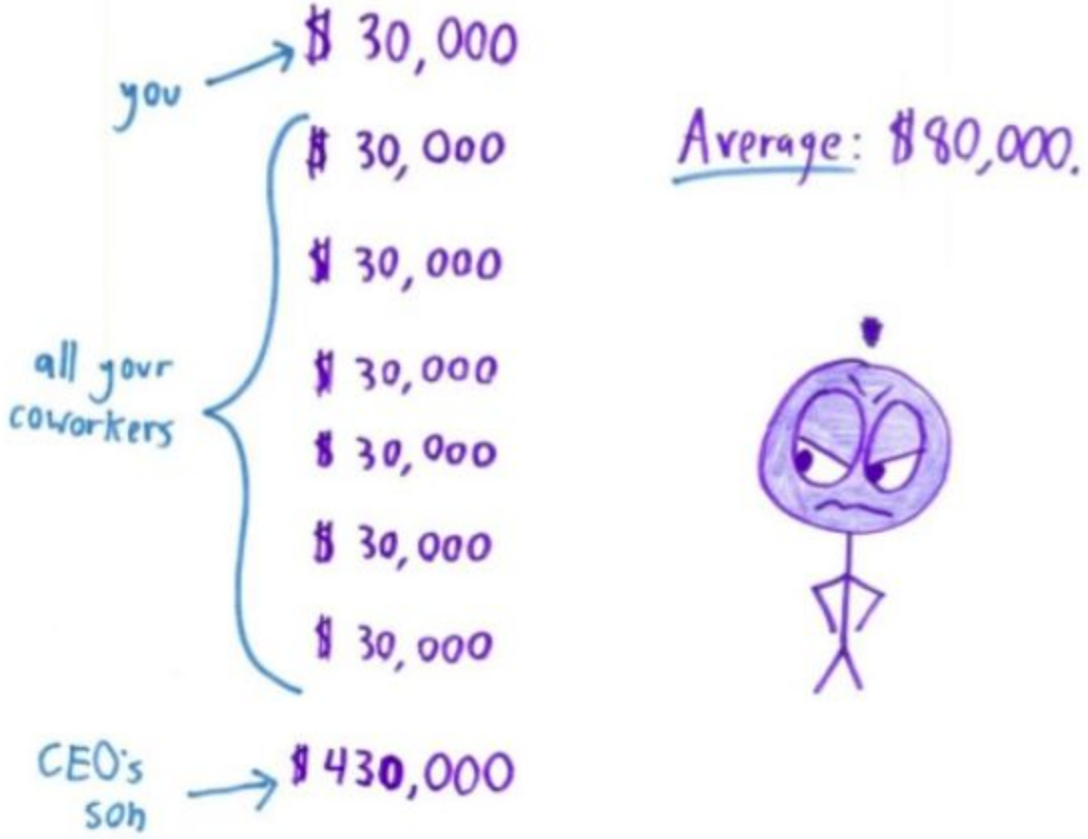
What's wrong with summary statistics?


y1
x1
y2
x2
y3
x3
y4
x4
4
6
8
10
12
14
16
18

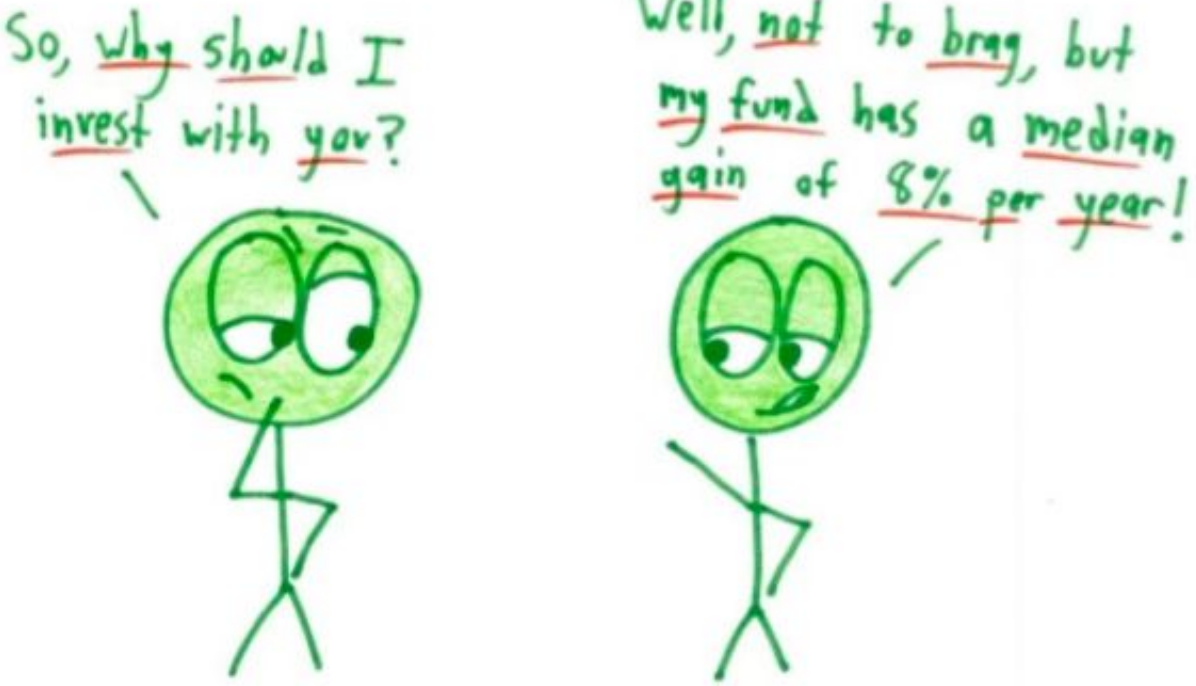
Property	Value
Mean of x	9
Sample variance of x	11
Mean of y	7.50
Sample variance of y	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression	0.67

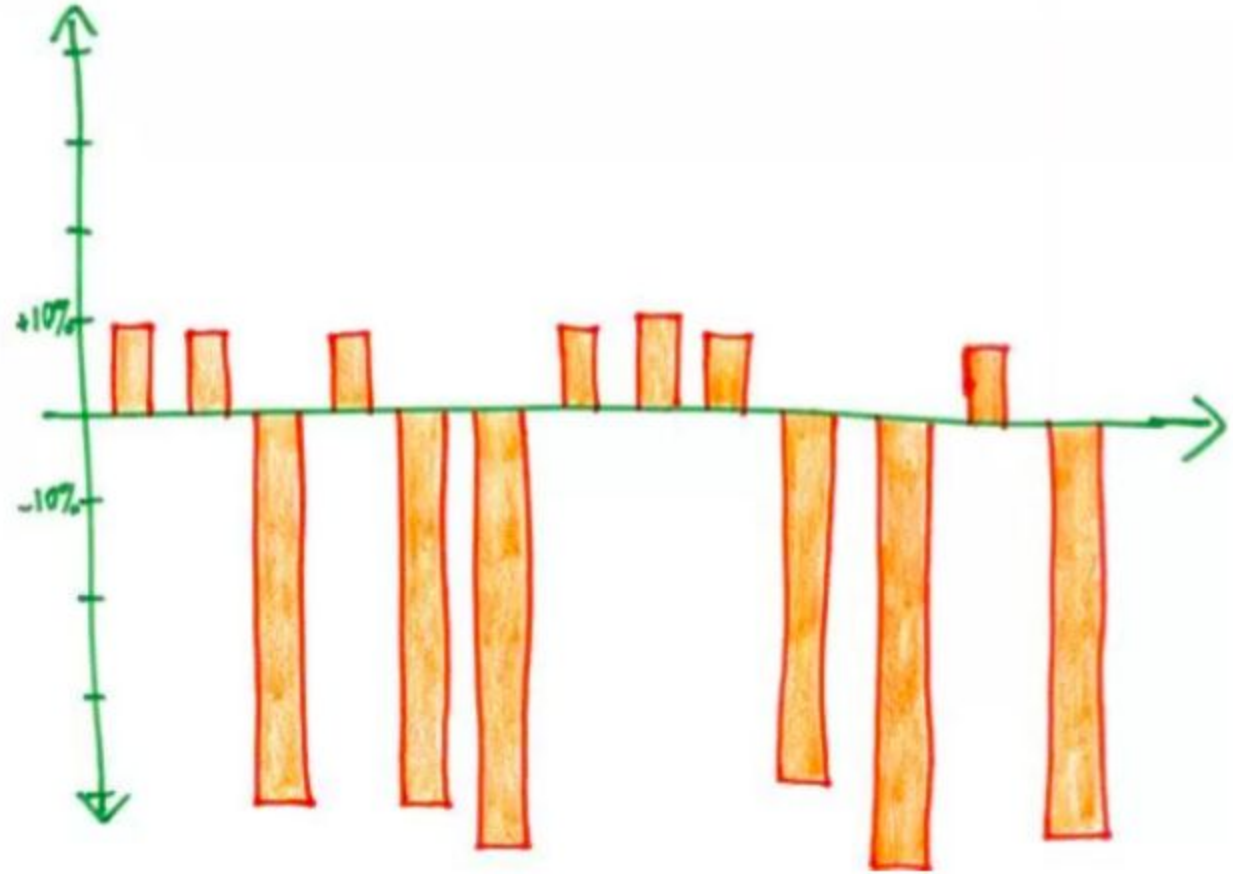
Mean


What would my starting salary be?
I'll put it this way: our average starting salary is $80,000!

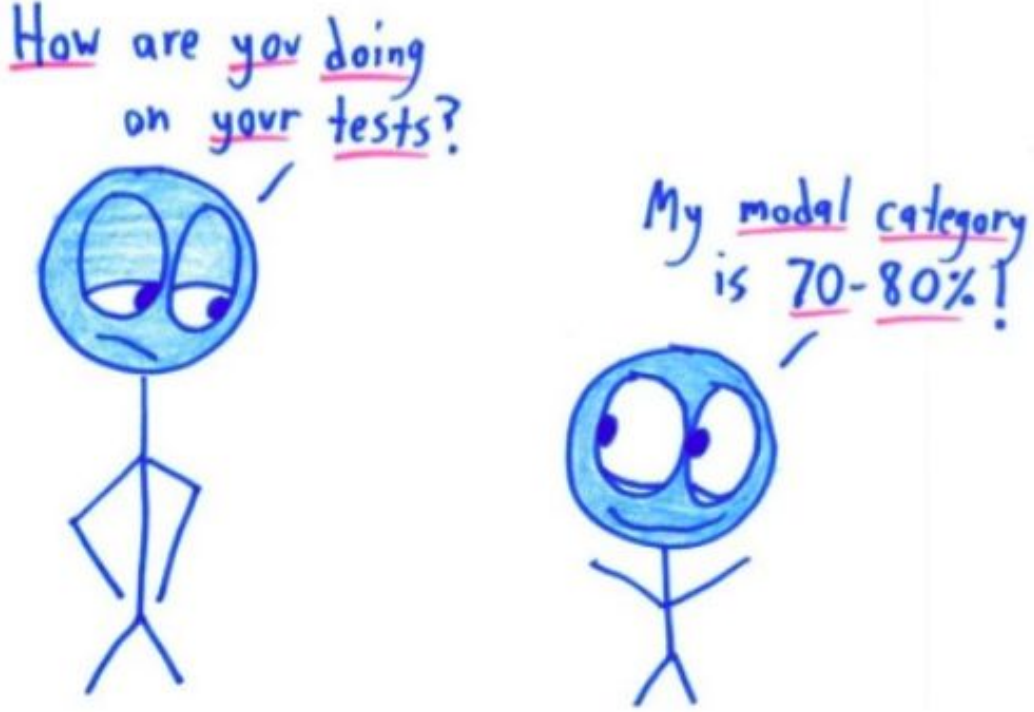

you
$ 30,000
all your coworkers
$ 30,000
$ 30,000
$ 30,000
$ 30,000
$ 30,000
$ 30,000
CEO's son
$ 430,000
Average: $80,000.

Median


So, why should I invest with you?
Well, not to brag, but my fund has a median gain of 8% per year!

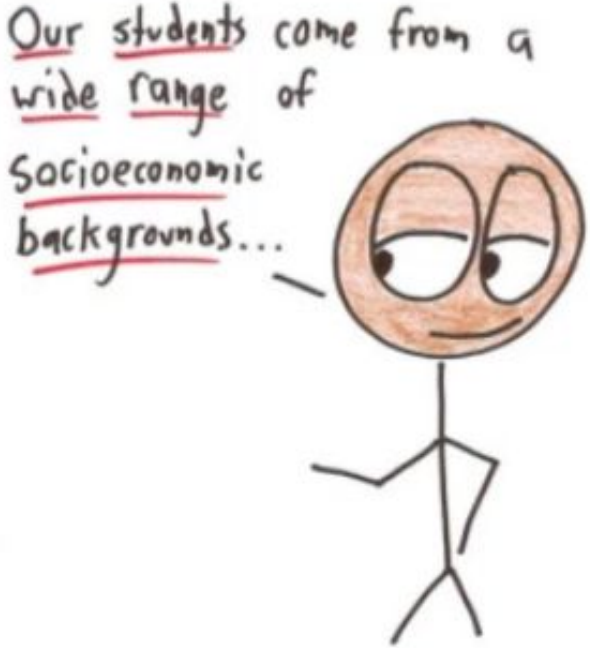

+10%
-10%

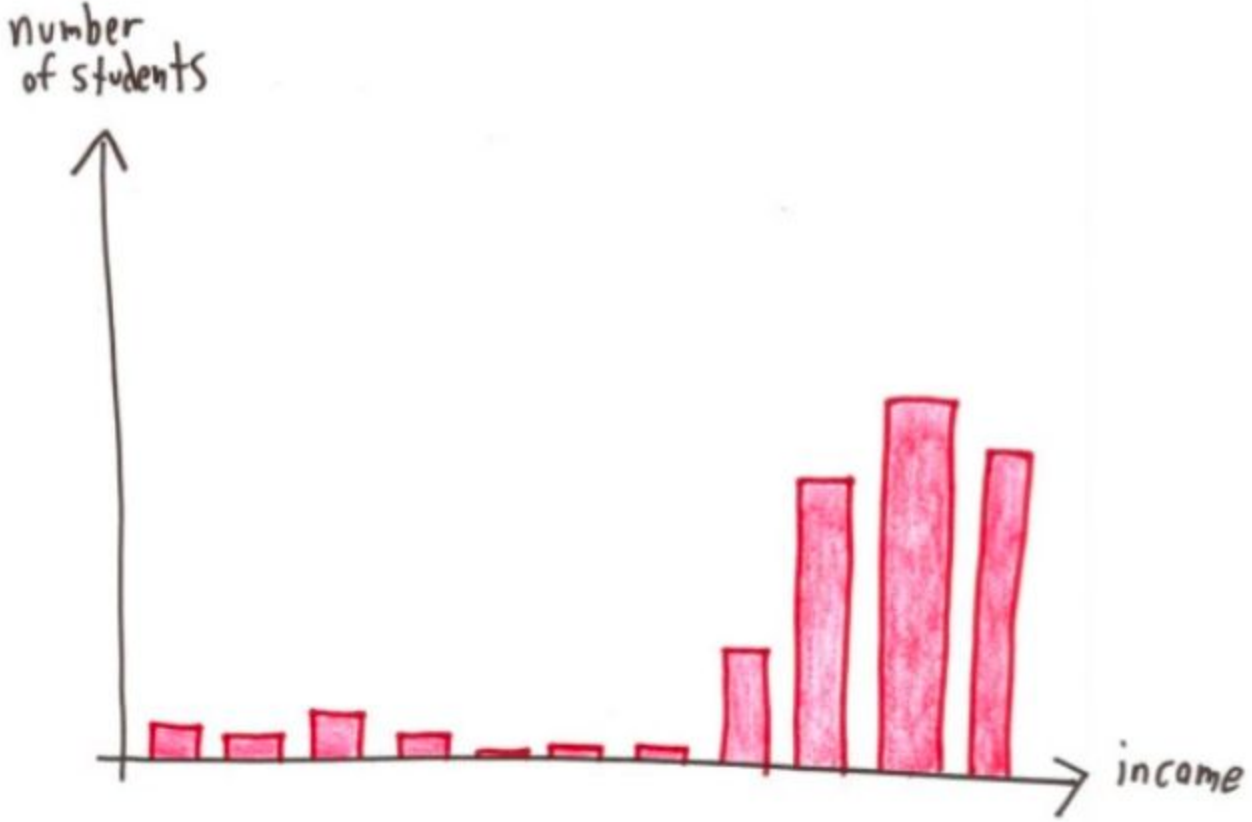
Mode



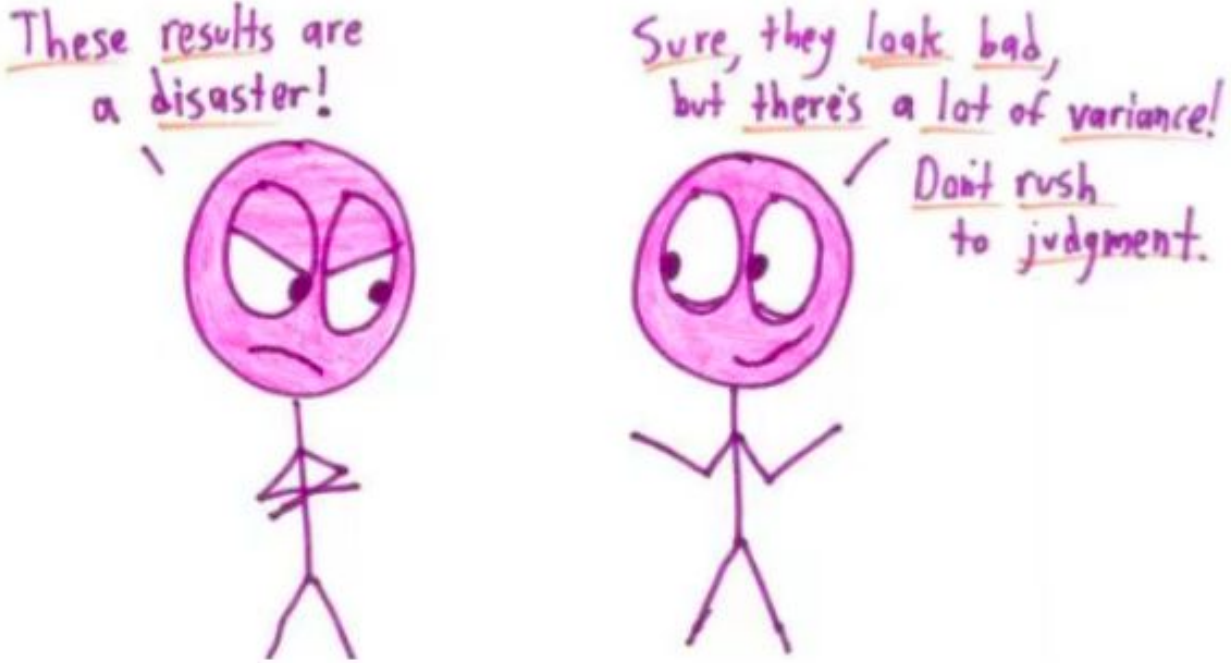
Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

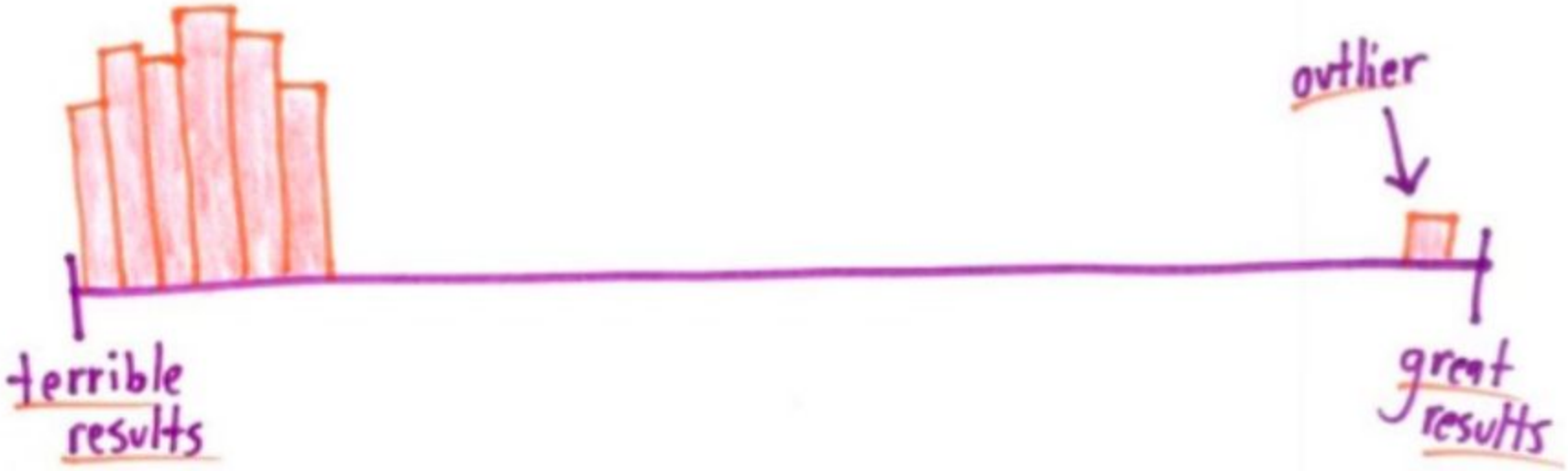
Range


Our students come from a wide range of socioeconomic backgrounds...


number
of students
income

Variance


These results are a disaster!
Sure, they look bad, but there's a lot of variance!
Don't rush to judgment.


outlier
terrible results
great results

Solution 1 = Data Visualization

Solution 2 = Use 3+ metrics

Stats + Python = Numpy

```
import numpy as np

np.mean([1,2,3])

np.median([1,2,3])

np.std([1,2,3])

np.percentile([1,2,3], 50)
```