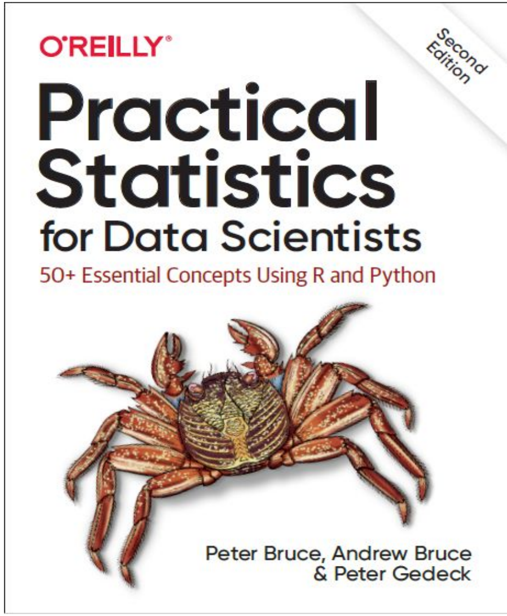
Practical Statistics for Data Scientists
50+ Essential Concepts Using R and Python

Exploratory Data Analysis



O'REILLY®

Second Edition

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python

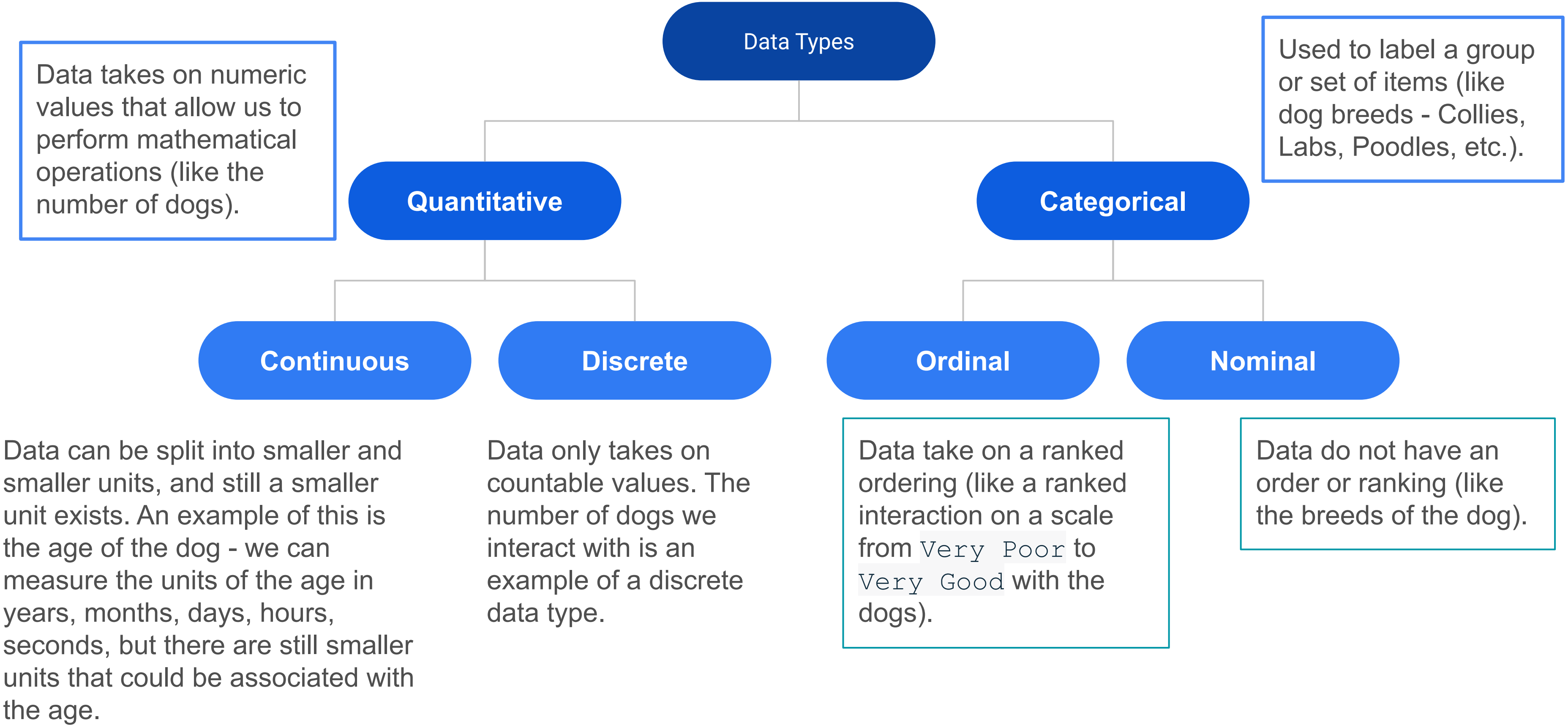
Peter Bruce, Andrew Bruce & Peter Gedeck

Introduction

We'll discuss :

- Measures of center and spread.
- Common shapes that data takes on and how to handle outliers
- Rectangular Data
- Nonrectangular Data Structures
- Mean
- Median and Robust Estimates
- Standard Deviation and Related Estimates
- Estimates Based on Percentiles

Data Types



Data Types

Quantitative

Data takes on numeric values that allow us to perform mathematical operations (like the number of dogs).

Continuous

Data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

Discrete

Data only takes on countable values. The number of dogs we interact with is an example of a discrete data type.

Categorical

Used to label a group or set of items (like dog breeds - Collies, Labs, Poodles, etc.).

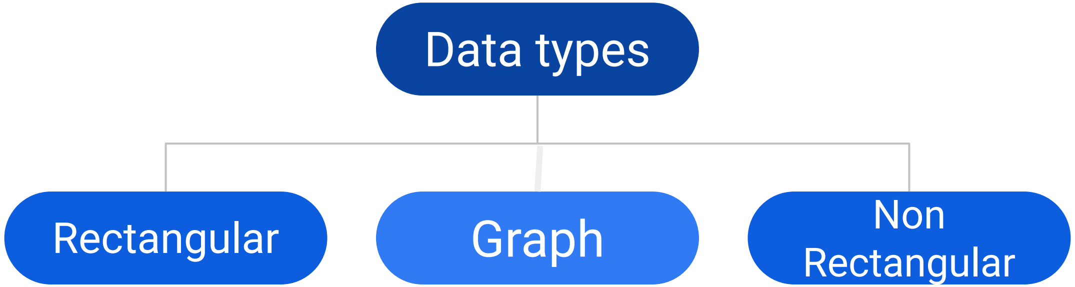
Ordinal

Data take on a ranked ordering (like a ranked interaction on a scale from `Very Poor` to `Very Good` with the dogs).

Nominal

Data do not have an order or ranking (like the breeds of the dog).

Data Types



Key Terms for Estimates of Location

Analyzing Quantitative Data

Four Aspects for Quantitative Data

There are four main aspects to analyzing **Quantitative** data.

1. Measures of `Center`
    a. `Mean`
    b. `Median`
    c. `Mode`
2. Measures of `Spread`
3. The `Shape` of the data.
4. `Outliers`

The Mean

The mean is often called the average or the **expected value** in mathematics. We calculate the mean by adding all of our values together and dividing by the number of values in our dataset.

Example: Number of dogs I see in a coffee Shop in a week

Mon	Tue	Wed	Thu	Fri	Sat	Sun
5	3	8	3	15	45	9

$$\frac{5 + 3 + 8 + 3 + 15 + 45 + 9}{7} = 12.57 \text{ dogs}$$

Trimmed Mean

A variation of the mean is a trimmed mean, which you calculate by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.

Example:

Below is a set of score on a standardized test, compute the mean and the 10% trimmed mean for the scores:

425	475	450	600	800	575	550	500	150	425

Weighted mean

You calculate by multiplying each data value xi by a user-specified weight wi and dividing their sum by the sum of the weights.

Example

You take three 100-point exams in your statistics class and score 80, 80 and 95. The last exam is much easier than the first two, so your professor has given it less weight. The weights for the three exams are:

- Exam 1: 40 % of your grade. (Note: 40% as a decimal is .4.)
- Exam 2: 40 % of your grade.
- Exam 3: 20 % of your grade.

What is your final weighted average for the class?

Weighted mean

1. Multiply the numbers in your data set by the weights:
   .4(80) = 32
   .4(80) = 32
   .2(95) = 19
2. Add the numbers up. 32 + 32 + 19 = 83.
3. All of the weights add up to 1 (.4 + .4 + .2) so you would divide your answer (83) by 1:
4. 83 / 1 = 83.

Median

The **median** splits our data so that 50% of our values are lower and 50% are higher

Median for Odd Values

If we have an **odd** number of observations, the **median** is simply the number in the **direct middle**.

Mon	Tue	Wed	Thu	Fri	Sat	Sun
5	3	8	3	15	45	9

3	3	5	8	9	15	45

Median

Median

Median for Even Values

If we have an **even** number of observations, the **median** is the **average of the two values in the middle**.

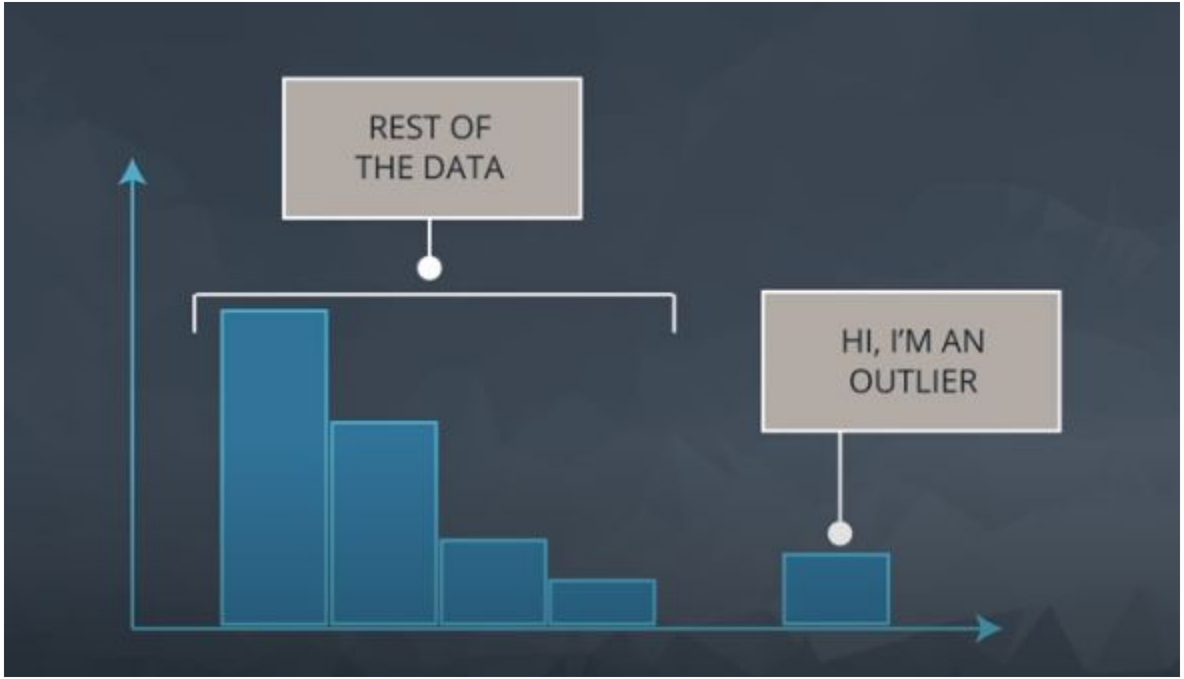
5	8	3	2	1	3	10	105

1	2	3	3	5	8	10	105

$$\frac{3+5}{2} = 4$$

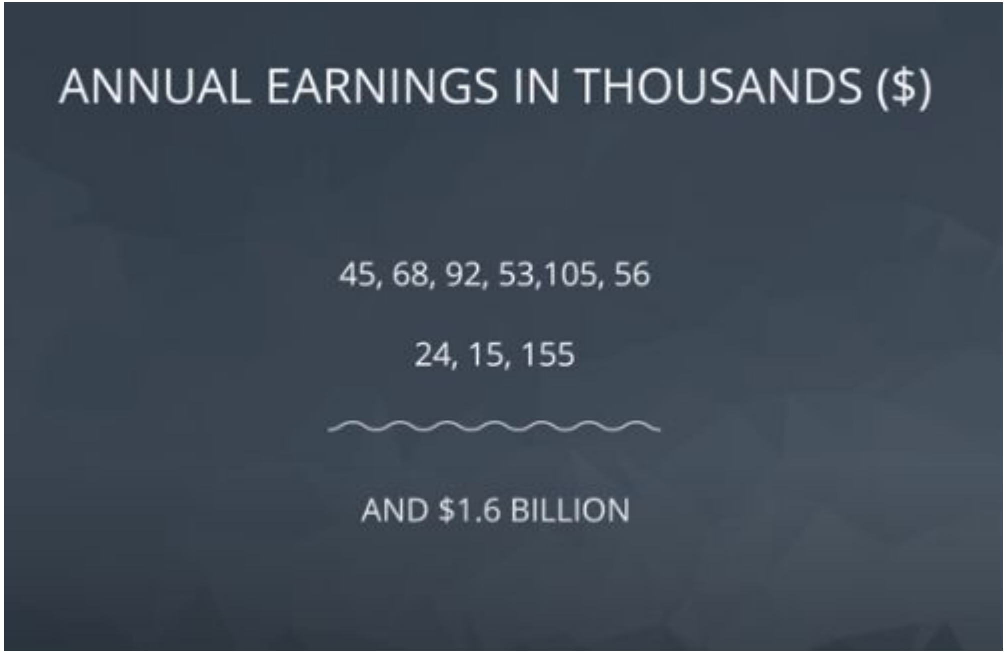
Outliers

Outliers are points that fall very far from the rest of our data points. This influences measures like the mean and standard deviation much more than measures associated with the five-number summary.



Outliers

Example

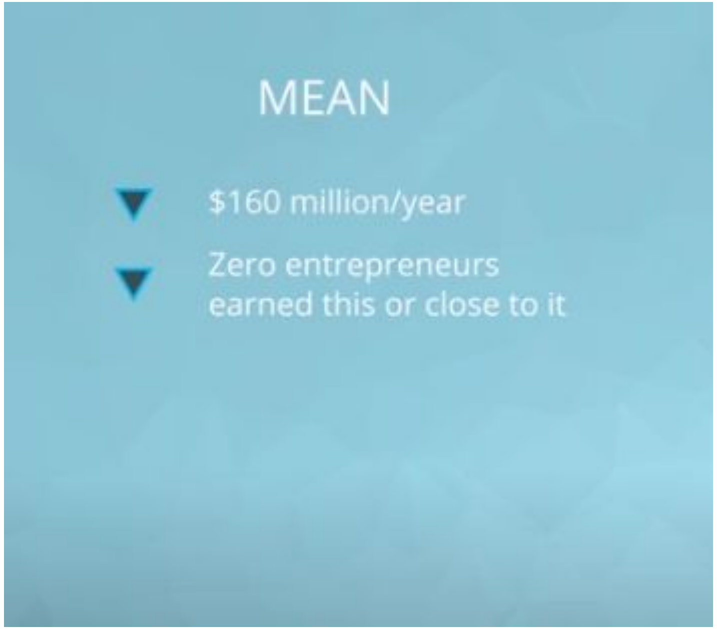


ANNUAL EARNINGS IN THOUSANDS ($)

45, 68, 92, 53,105, 56

24, 15, 155

AND $1.6 BILLION



MEAN

- $160 million/year
- Zero entrepreneurs earned this or close to it

Outliers

Common Techniques

When outliers are present we should consider the following points.

1. Noting they exist and the impact on summary statistics.

2. If typo - remove or fix

3. Understanding why they exist, and the impact on questions we are trying to answer about our data.(anomaly detection deal with this idea)

4. Reporting the 5 number summary values is often a better indication than measures like the mean and standard deviation when we have outliers.

5. Be careful in reporting. Know how to ask the right questions.

Outliers

Outliers Advice

Below are guidelines for working with any column (random variable) in your dataset.

1. Plot your data to identify if you have outliers.

2. Handle outliers accordingly via the previous methods.

3. If no outliers and your data follow a normal distribution - use the mean and standard deviation to describe your dataset, and report that the data are normally distributed.

4. If you have skewed data or outliers, use the five-number summary to summarize your data and report the outliers.

Outliers

Side note

If you aren't sure if your data are normally distributed, there are plots called normal quantile plots and statistical methods like the Kolmogorov-Smirnov test that are aimed to help you understand whether or not your data are normally distributed.

Standard Deviation and Variance

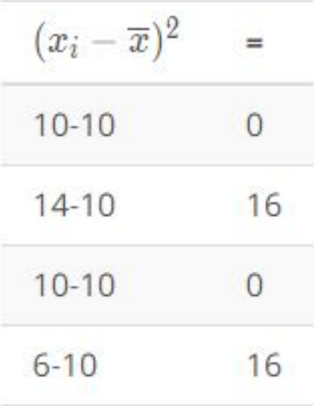
The **standard deviation** is defined as **the average distance of each observation from the mean**

How to Calculate Standard Deviation

Dataset = 10, 14, 10, 6

1. Calculate the mean $(\sum_{i=1}^{4} x_i)/n$ = 40/4 = 10

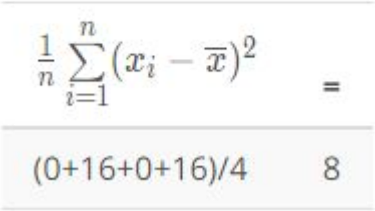
2. Calculate the distance of each observation from the mean and square the value



$(x_i - \overline{x})^2$	=
10-10	0
14-10	16
10-10	0
6-10	16

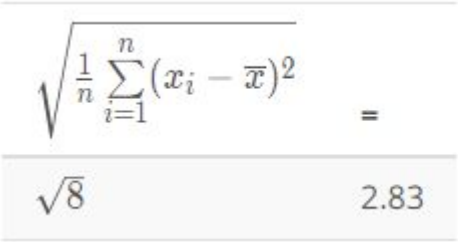
Standard Deviation and Variance

3. Calculate the **variance**, the average squared difference of each observation from the mean



$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$	=
(0+16+0+16)/4	8

4. Calculate the **standard deviation**, the square root of the variance



$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}$	=
$\sqrt{8}$	2.83

is on average, how far each point in our dataset is from the mean.

Estimates Based on Percentiles

Called also : Calculating the 5 Number Summary

The five-number summary consist of 5 values:

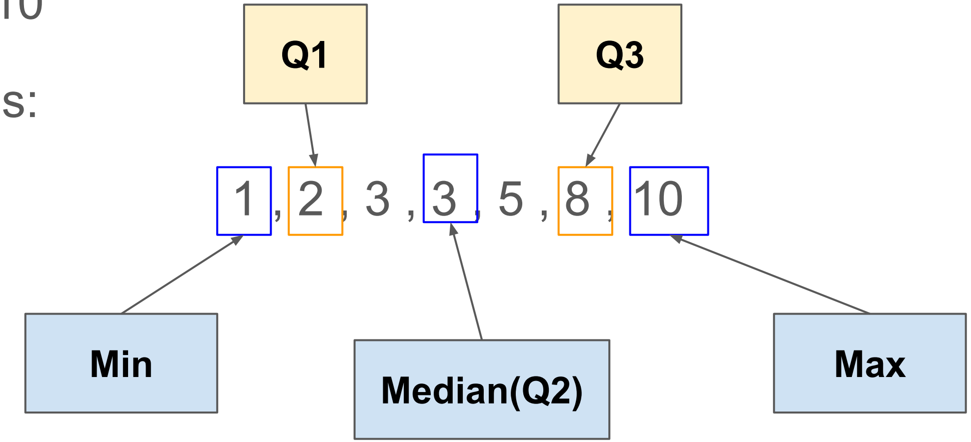
1. **Minimum:** The smallest number in the dataset.
2. **Q1:** The value such that 25% of the data fall below.
3. **Q2:** The value such that 50% of the data fall below.
4. **Q3**: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

Calculating the 5 Number Summary

Example:

Dataset: 5,8,3,2,1,3,10

1. order your values:



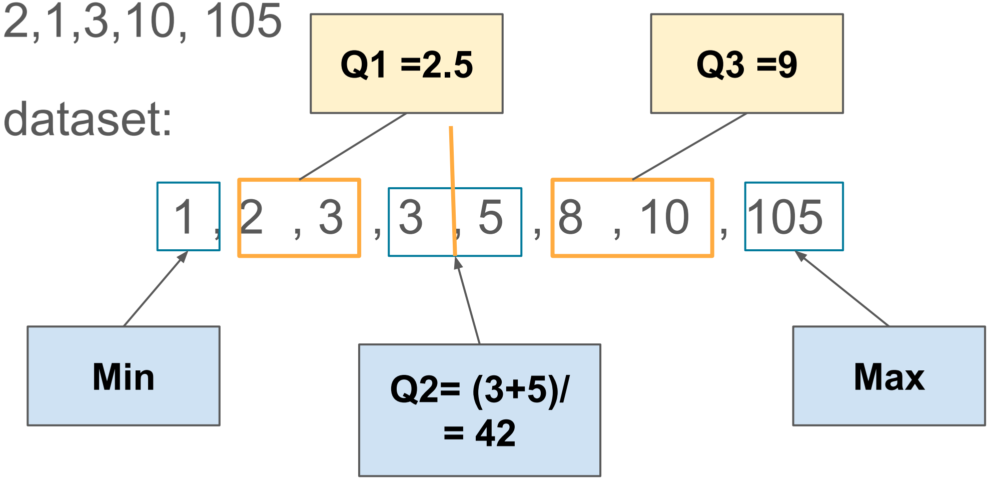
Q1 and Q3 : The Medians of the data on either side of Q2

Calculating the 5 Number Summary

Another example:

Dataset: 5,8,3,2,1,3,10, 105

1. Order the dataset:



Calculating the 5 Number Summary

Range

The **range** is then calculated as the difference between the **maximum** and the **minimum**.

IQR

The **interquartile range** is calculated as the difference between $\mathbf{Q3}$ and $\mathbf{Q1}$