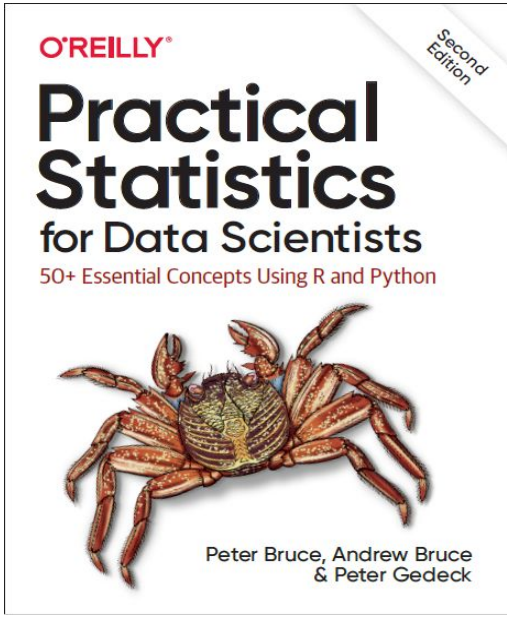
Practical Statistics for Data Scientists
50+ Essential Concepts Using R and Python

Exploratory Data Analysis

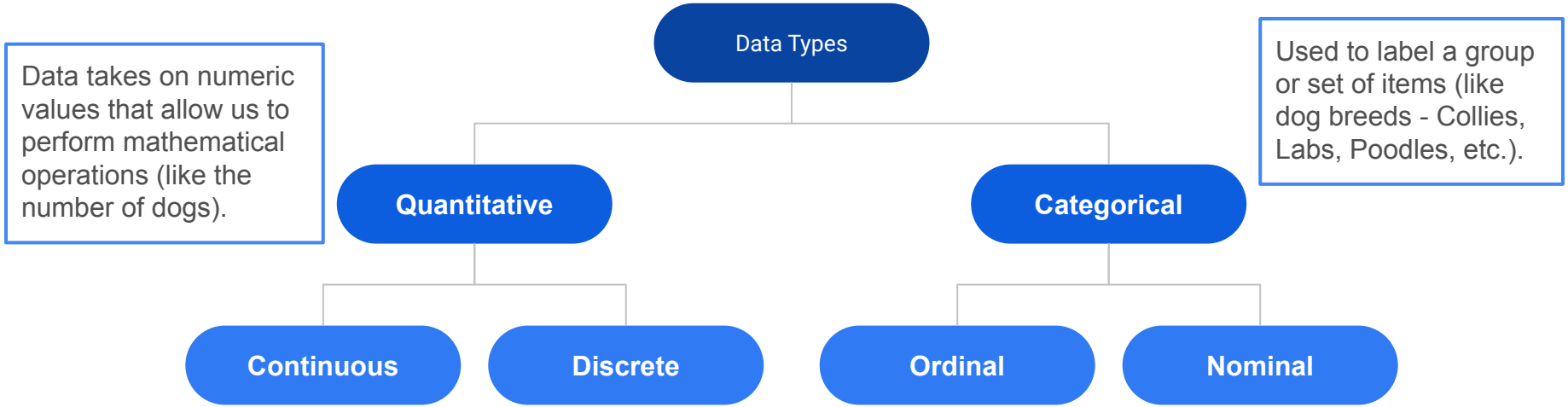


Introduction

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

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

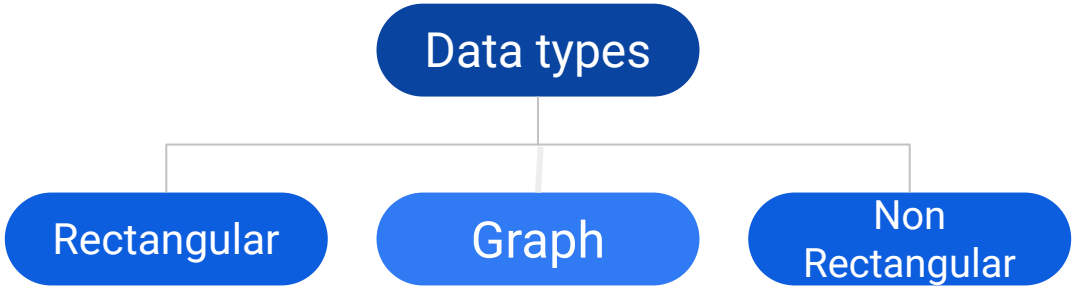
Data can be split into smaller and smaller units, and still a smaller unit exists. An example of this is the age of the dog - we can measure the units of the age in years, months, days, hours, seconds, but there are still smaller units that could be associated with the age.

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

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

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**.

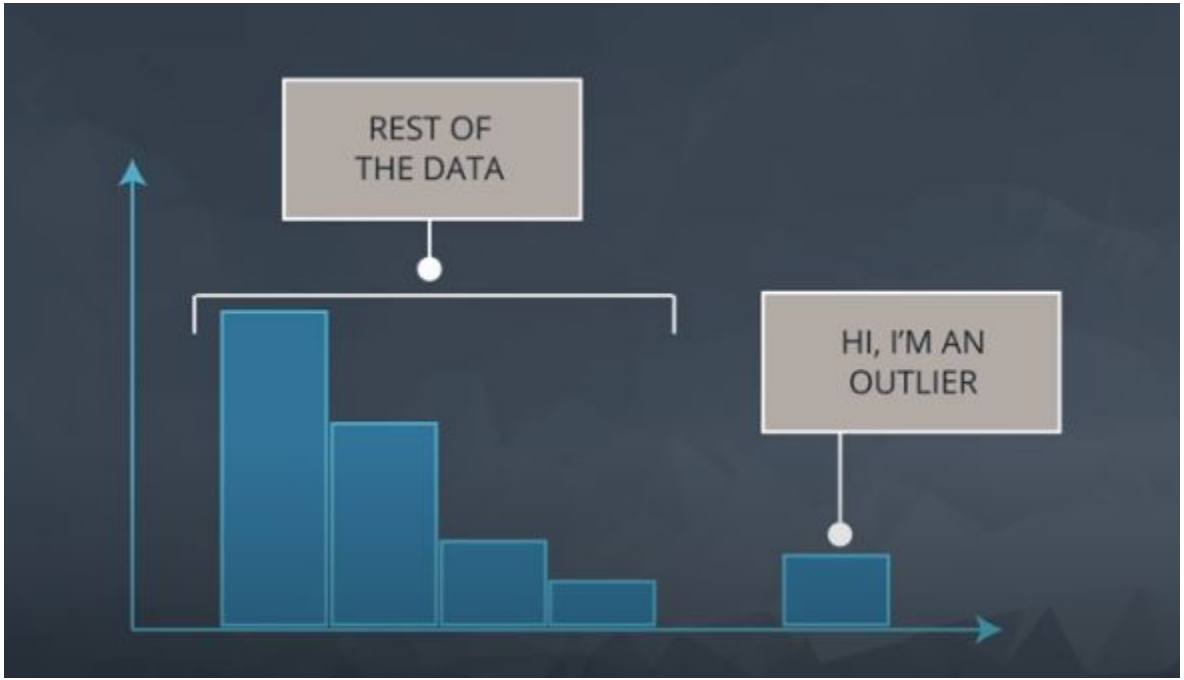
5	8	3	2	1	3	10	105

1	2	3	3	5	8	10	105

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

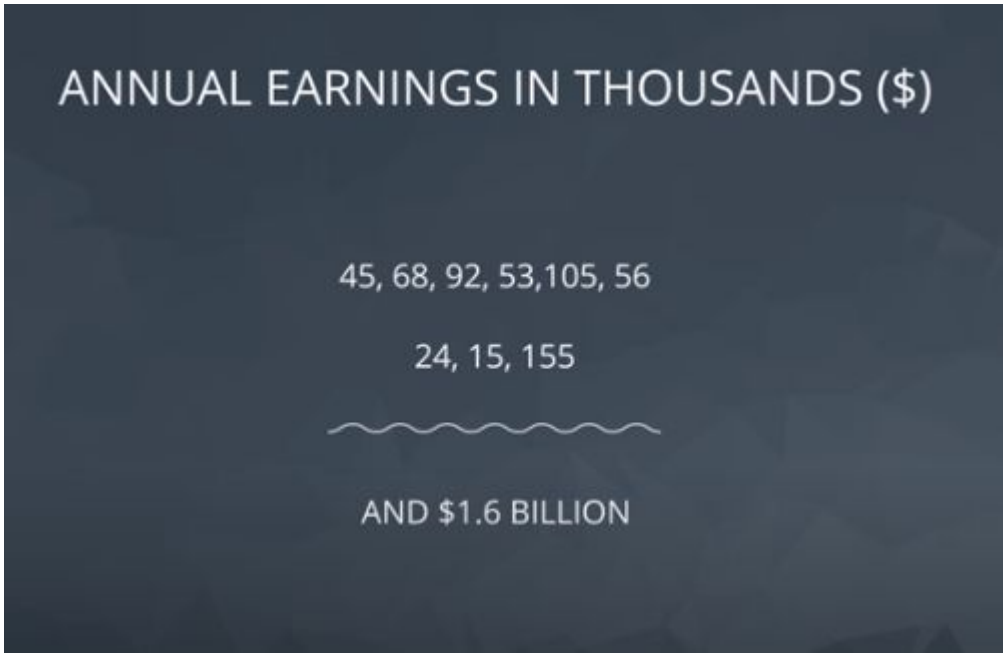
Outliers

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



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:

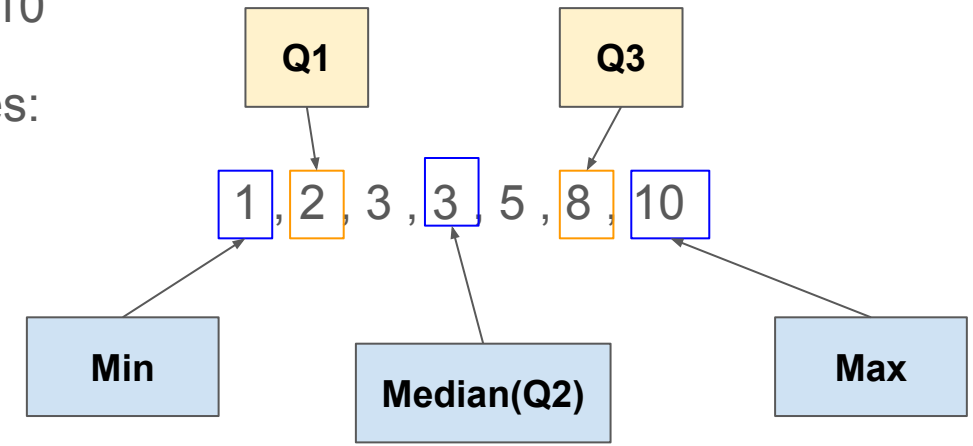
1. **Minimum:** The smallest number in the dataset.
2. **Q1:** The value such that 25% of the data fall below.
3. **Q2:** The value such that 50% of the data fall below.
4. **Q3**: The value such that 75% of the data fall below.
5. **Maximum:** The largest value in the dataset.

Calculating the 5 Number Summary

Example:

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

1. order your values:



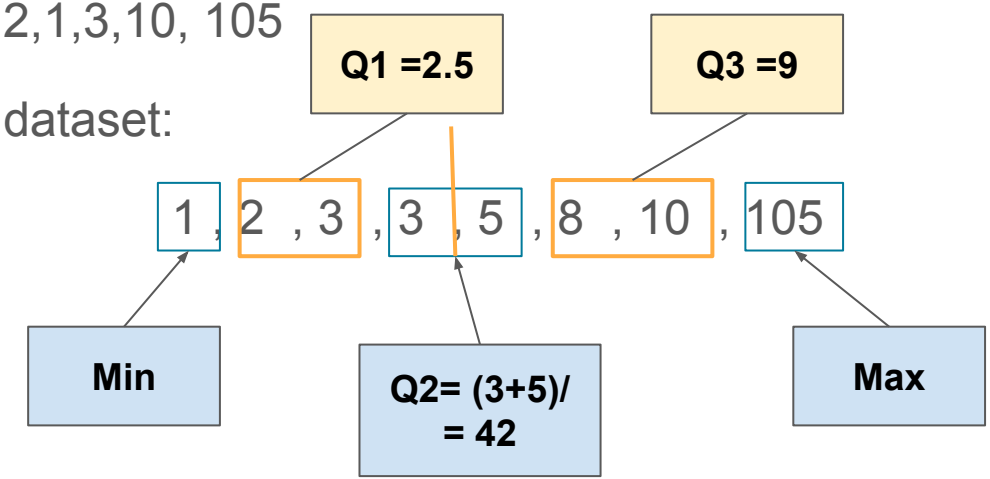
Q1 and Q3 : The Medians of the data on either side of Q2

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

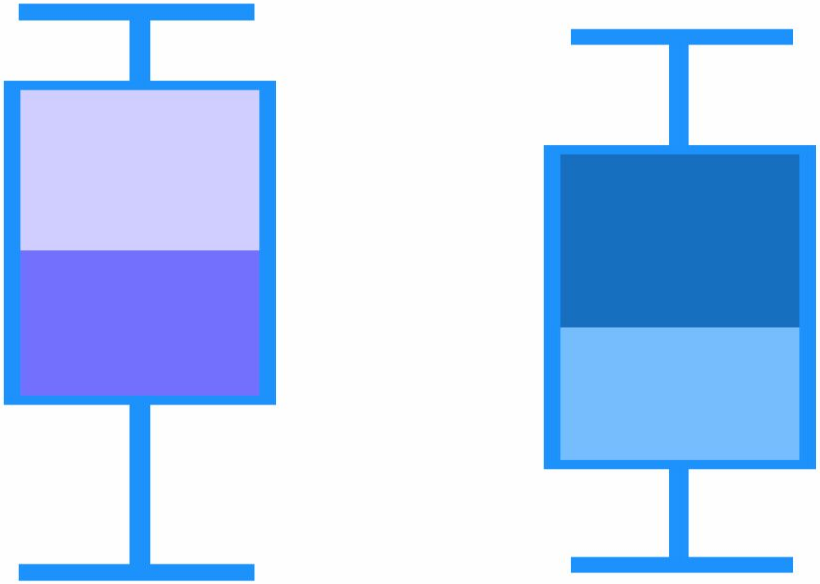
The **interquartile range** is calculated as the difference between $Q3$ and $Q1$

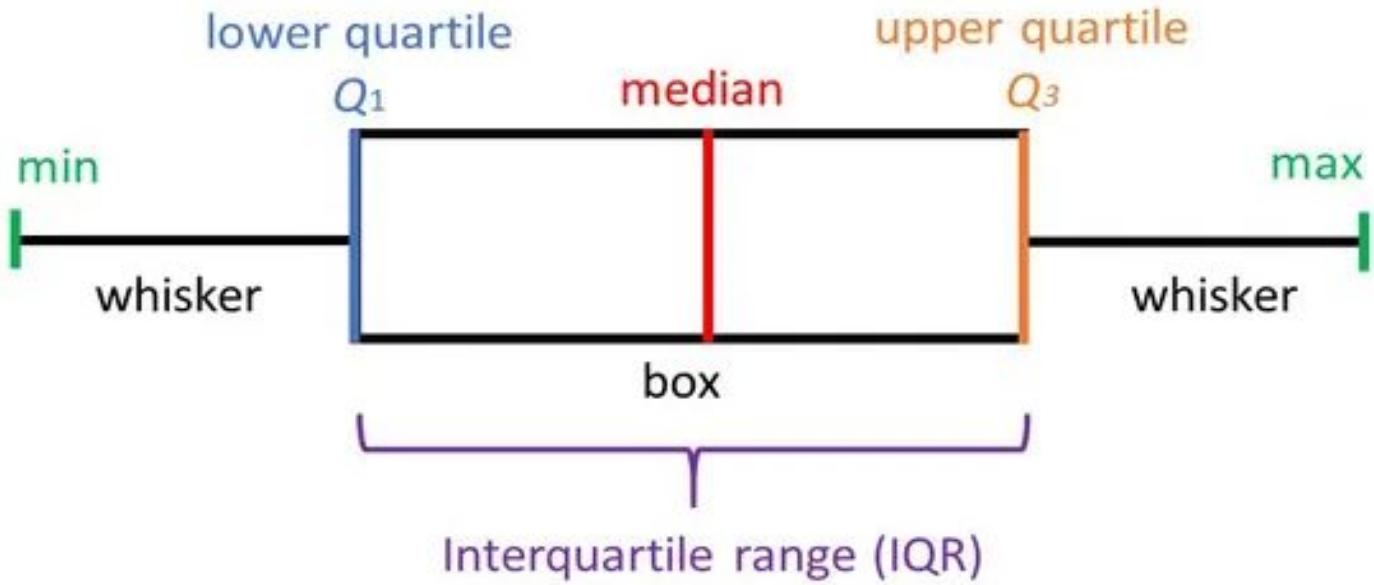
Exploring the Data Distribution
15/01/2021

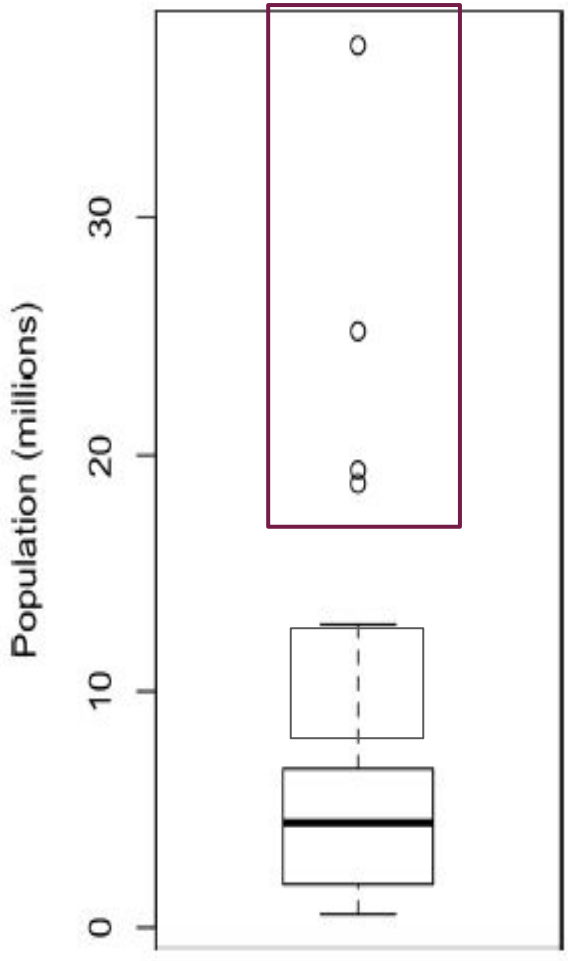
Exploring the Data Distribution

Boxplot

Exploring the Data Distribution




lower quartile
Q_1
median
upper quartile
Q_3
min
max
whisker
whisker
box
Interquartile range (IQR)



Any data outside of the whiskers is plotted as single points or circles (often considered outliers)

whiskers

Frequency Tables and Histograms

A frequency table of a variable divides up the variable range into equally spaced segments and tells us how many values fall within each segment

Table 1-5. A frequency table of population by state

BinNumber	BinRange	Count	States
1	563,626–4,232,658	24	WY,VT,ND,AK,SD,DE,MT,RI,NH,ME,HI,ID,NE,WV,NM,NV,UT,KS,AR,MS,IA,CT,OK,OR
2	4,232,659–7,901,691	14	KY,LA,SC,AL,CO,MN,WI,MD,MO,TN,AZ,IN,MA,WA
3	7,901,692–11,570,724	6	VA,NJ,NC,GA,MI,OH
4	11,570,725–15,239,757	2	PA,IL
5	15,239,758–18,908,790	1	FL
6	18,908,791–22,577,823	1	NY
7	22,577,824–26,246,856	1	TX
8	26,246,857–29,915,889	0	
9	29,915,890–33,584,922	0	
10	33,584,923–37,253,956	1	CA

Frequency Tables and Histograms

- The least populous state is Wyoming,, and the most populous is California. T
- This gives us a range of 37,253,956 – 563,626 = 36,690,330, which we must divide up into equal size bins—let's say 10 bins.
- With 10 equal size bins, each bin will have a width of 3,669,033, so the first bin will span from 563,626 to 4,232,658. By contrast, the top bin, 33,584,923 to 37,253,956, has only one state: California. The two bins immediately below California are empty, until we reach Texas.
- It is important to include the empty bins; the fact that there are no values in those bins is useful information.
- It can also be useful to experiment with different bin sizes. If they are too large, important features of the distribution can be obscured.
- If they are too small, the result is too granular, and the ability to see the bigger picture is lost.

Histograms

- A histogram is a way to visualize a frequency table, with bins on the x-axis and the data count on the y-axis.

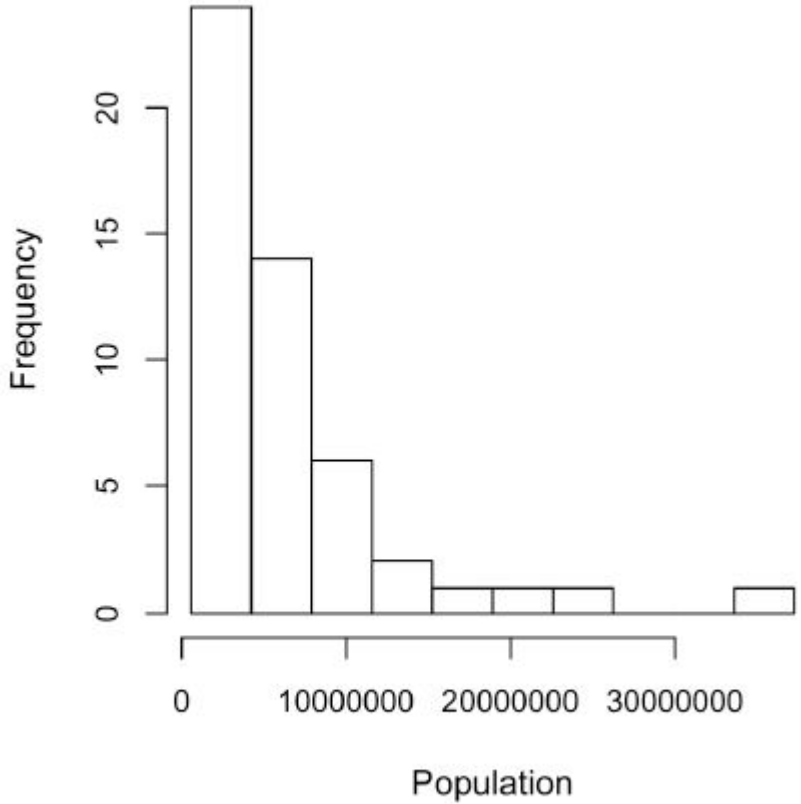


Figure 1-3. Histogram of state populations

Density Plots and Estimates

- Related to the histogram is a density plot, which shows the distribution of data values as a continuous line.

- A density plot can be thought of as a smoothed histogram, although it is typically computed directly from the data through a kernel density estimate

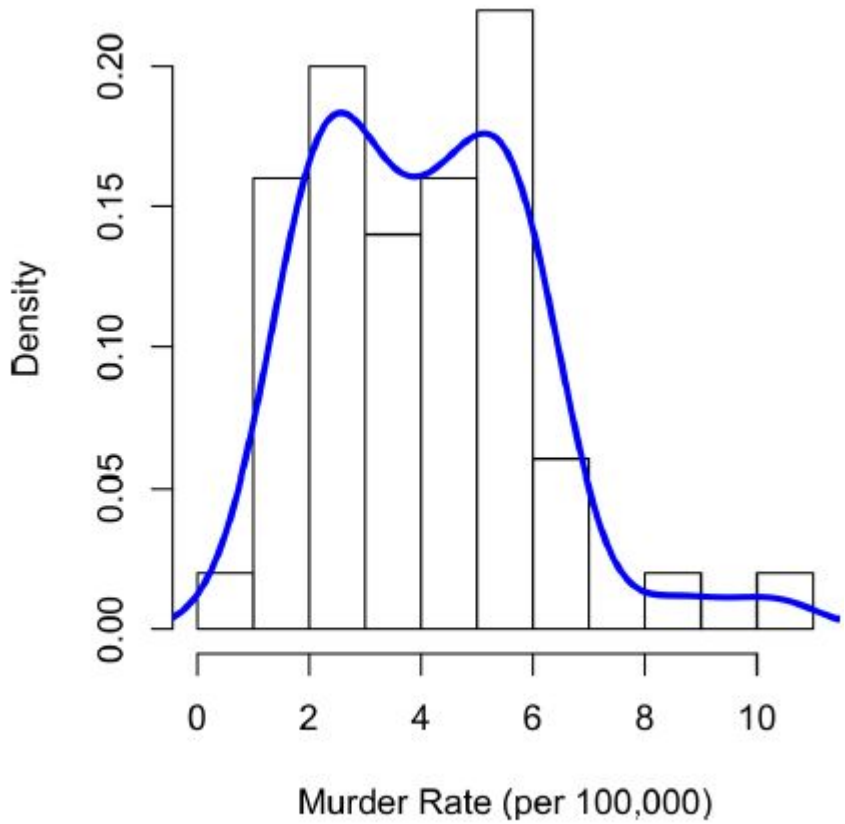


Figure 1-4. Density of state murder rates

Key Ideas

• A frequency histogram plots frequency counts on the y-axis and variable values on the x-axis; it gives a sense of the distribution of the data at a glance.

• A frequency table is a tabular version of the frequency counts found in a histogram.

• A boxplot—with the top and bottom of the box at the 75th and 25th percentiles, respectively—also gives a quick sense of the distribution of the data; it is often used in side-by-side displays to compare distributions.

• A density plot is a smoothed version of a histogram; it requires a function to estimate a plot based on the data (multiple estimates are possible, of course).

Exploring Binary and Categorical Data

Getting a summary of a binary variable or a categorical variable with a few categories is a fairly easy matter: we just figure out the proportion of 1s, or the proportions of the important categories.

For example, Table 1-6 shows the percentage of delayed flights by the cause of delay at Dallas/Fort Worth Airport since 2010.

Delays are categorized as being due to factors under carrier control, air traffic control (ATC) system delays, weather, security, or a late inbound aircraft.

Table 1-6. Percentage of delays by cause at Dallas/Fort Worth Airport

Carrier	ATC	Weather	Security	Inbound
23.02	30.40	4.03	0.12	42.43

Bar charts

- Bar charts, seen often in the popular press, are a common visual tool for displaying a single categorical variable.

- Categories are listed on the x-axis, and frequencies or proportions on the y-axis

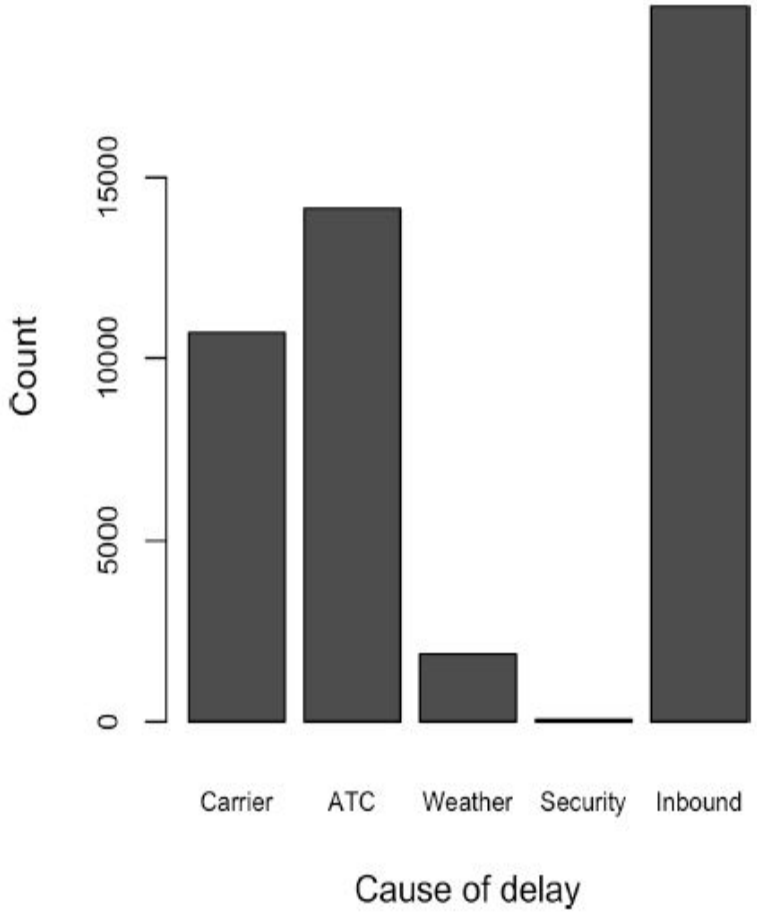


Figure 1-5. Bar chart of airline delays at DFW by cause

Note that a bar chart resembles a histogram; in a bar chart the x-axis represents different categories of a factor variable, while in a histogram the x-axis represents values of a single variable on a numeric scale.

In a histogram, the bars are typically shown touching each other, with gaps indicating values that did not occur in the data. In a bar chart, the bars are shown separate from one another.

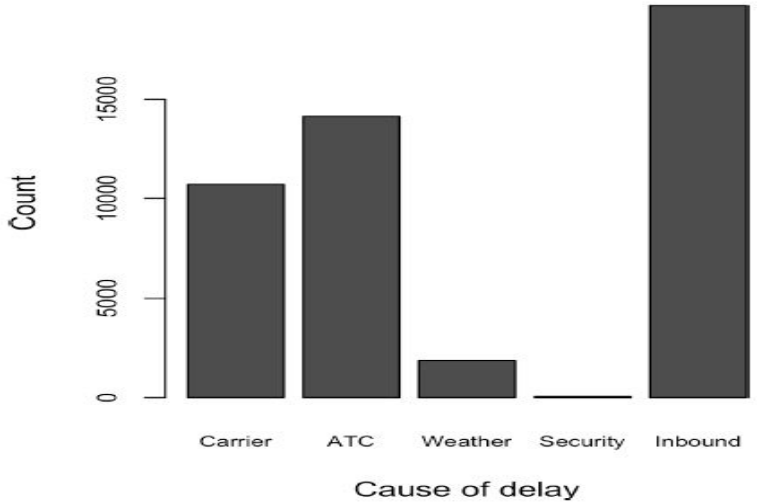
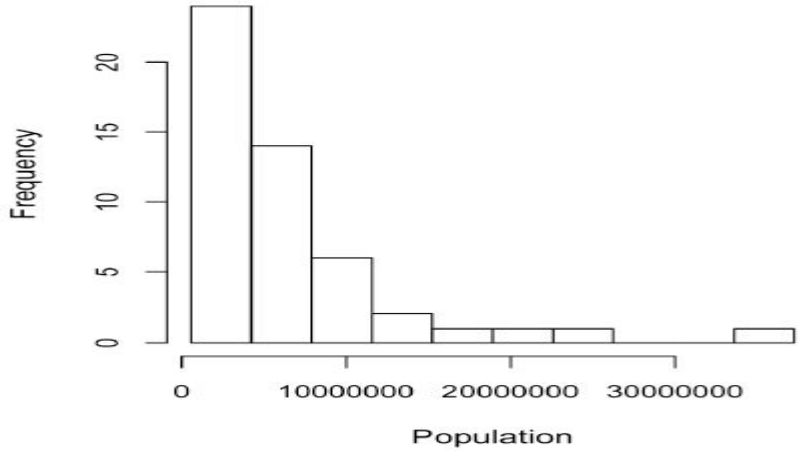


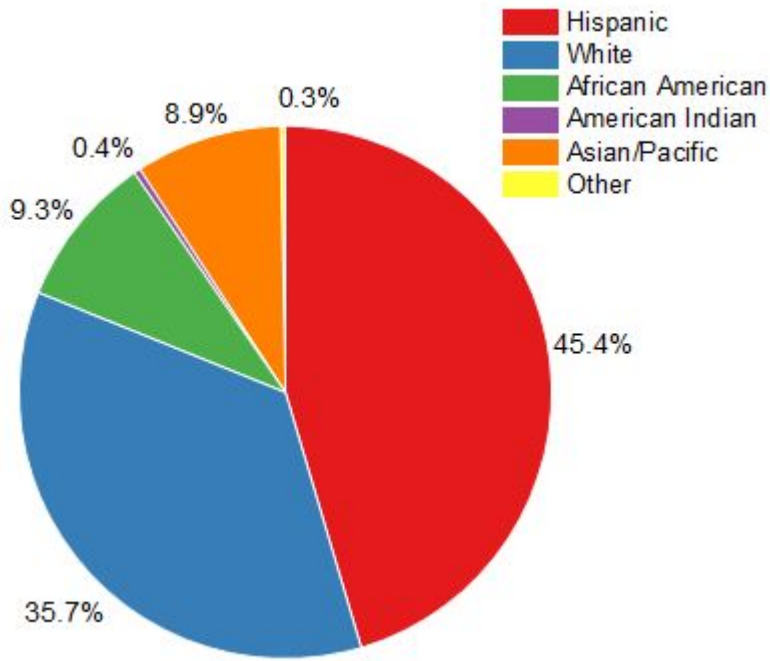
Figure 1-5. Bar chart of airline delays at DFW by cause



-3. Histogram of state populations

Pie charts

Pie charts are an alternative to bar charts, although statisticians and data visualization experts generally eschew pie charts as less visually informative



Mode

The **mode** is the most frequently observed value in our dataset.

There might be multiple modes for a particular dataset or no mode at all.

No Mode

If all observations in our dataset are observed with the same frequency, there is no mode. If we have the dataset:

1, 1, 2, 2, 3, 3, 4, 4

There is no mode because all observations occur the same number of times.

Mode

Many Modes

If two (or more) numbers share the maximum value, then there is more than one mode. If we have the dataset:

1, 2, 3, 3, 3, 4, 5, 6, 6, 6, 7, 8, 9

There are two modes 3 and 6, because these values share the maximum frequencies at 3 times, while all other values only appear once.

Expected Value

The expected value (EV) is an anticipated value for an investment at some point in the future. In statistics and probability analysis, the expected value is calculated by multiplying each of the possible outcomes by the likelihood each outcome will occur and then summing all of those values. By calculating expected values, investors can choose the scenario most likely to give the desired outcome.

Example of Expected Value (EV)

To calculate the EV for a single discrete random variable, you must multiply the value of the variable by the probability of that value occurring. Take, for example, a normal six-sided die. Once you roll the die, it has an equal one-sixth chance of landing on one, two, three, four, five, or six. Given this information, the calculation is straightforward:

$$\left(\frac{1}{6} \times 1\right) + \left(\frac{1}{6} \times 2\right) + \left(\frac{1}{6} \times 3\right) + \left(\frac{1}{6} \times 4\right) + \left(\frac{1}{6} \times 5\right) + \left(\frac{1}{6} \times 6\right) = 3.5$$

If you were to roll a six-sided die an infinite amount of times, you see the average value equals 3.5.

Probability

The probability that an event will happen is the proportion of times it will occur if the situation could be repeated over and over, countless times. Most often this is an imaginary construction, but it is an adequate operational understanding of probability.

Key Ideas

- Categorical data is typically summed up in proportions and can be visualized in a bar chart.
- Categories might represent distinct things (apples and oranges, male and female), levels of a factor variable (low, medium, and high), or numeric data that has been binned.
- Expected value is the sum of values times their probability of occurrence, often used to sum up factor variable levels.

Correlation

Correlation is a statistical term describing **the degree to which two variables move in coordination with one another**. If the two variables move in the same direction, then those variables are said to have a positive correlation. If they move in opposite directions, then they have a negative correlation.

Example:

As a hypothetical example, assume that an analyst needs to calculate the correlation for the following two data sets:

X: (41, 19, 23, 40, 55, 57, 33)

Y: (94, 60, 74, 71, 82, 76, 61)

Correlation

There are three steps involved in finding the correlation. The first is to add up all the X values to find SUM(X), add up all the Y values to fund SUM(Y) and multiply each X value with its corresponding Y value and sum them to find SUM(X,Y):

SUM(X) = (41 + 19 + 23 + 40 + 55 + 57 + 33) = 268

SUM(Y) = (94 + 60 + 74 + 71 + 82 + 76 + 61) = 518

SUM(X,Y) = (41 x 94) + (19 x 60) + (23 x 74) + ... (33 x 61) = 20,391

Correlation

The next step is to take each X value, square it, and sum up all these values to find SUM(x^2). The same must be done for the Y values:

SUM(X^2) = (41^2) + (19^2) + (23^2) + ... (33^2) = 11,534

SUM(Y^2) = (94^2) + (60^2) + (74^2) + ... (61^2) = 39,174

Noting that there are seven observations, n, the following formula can be used to find the correlation coefficient, r:

$$r = \frac{n \times (\sum(X,Y) - (\sum(X) \times \sum(Y)))}{\sqrt{(n \times \sum(X^2) - \sum(X)^2) \times (n \times \sum(Y^2) - \sum(Y)^2)}}$$

where:

$r =$ Correlation coefficient

$n =$ Number of observations

Correlation

In this example, the correlation would be:

r = (7 x 20,391 - (268 x 518) / SquareRoot((7 x 11,534 - 268^2) x (7 x 39,174 - 518^2)) = 3,913 / 7,248.4 = 0.54

Note:

A perfect positive correlation means that the correlation coefficient is exactly 1. This implies that as one security moves, either up or down, the other security moves in lockstep, in the same direction. A perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no linear relationship at all.

Scatterplots

The standard way to visualize the relationship between two measured data variables is with a scatterplot. The x-axis represents one variable and the y-axis another, and each point on the graph is a record

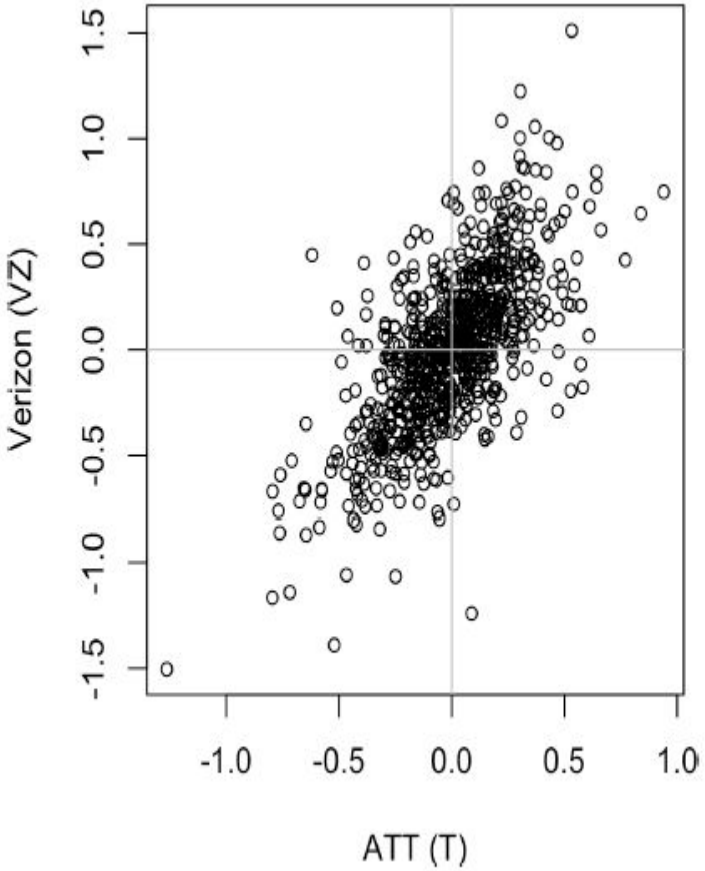


Figure 1-7. Scatterplot of correlation between returns for ATT and Verizon

Exploring Two or More Variables

Hexagonal Binning and Contours (Plotting Numeric Versus Numeric Data)

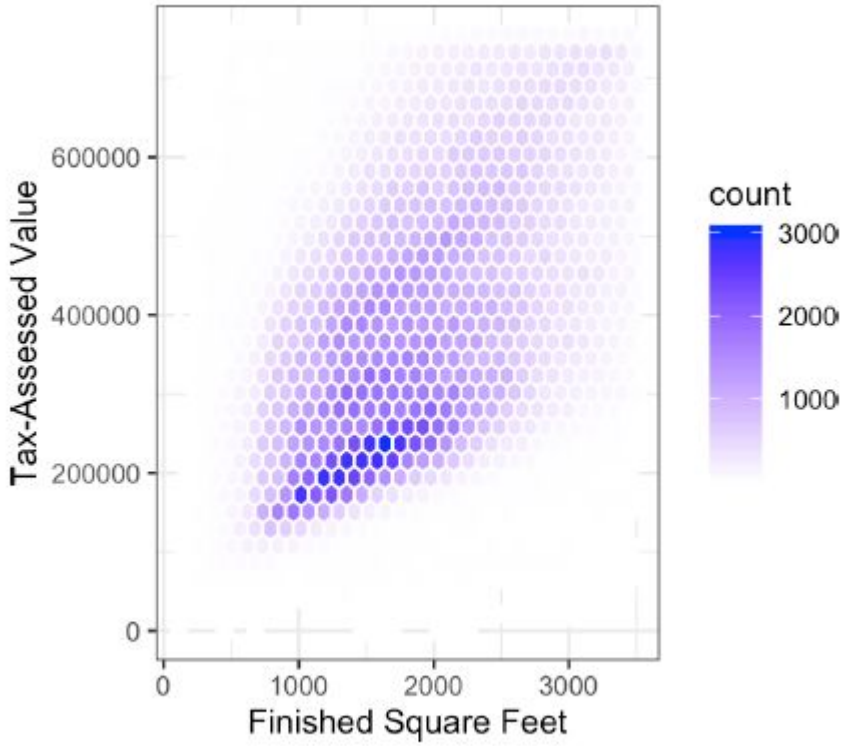


Figure 1-8. Hexagonal binning for tax-assessed value versus finished square feet

Visualizing Multiple Variables

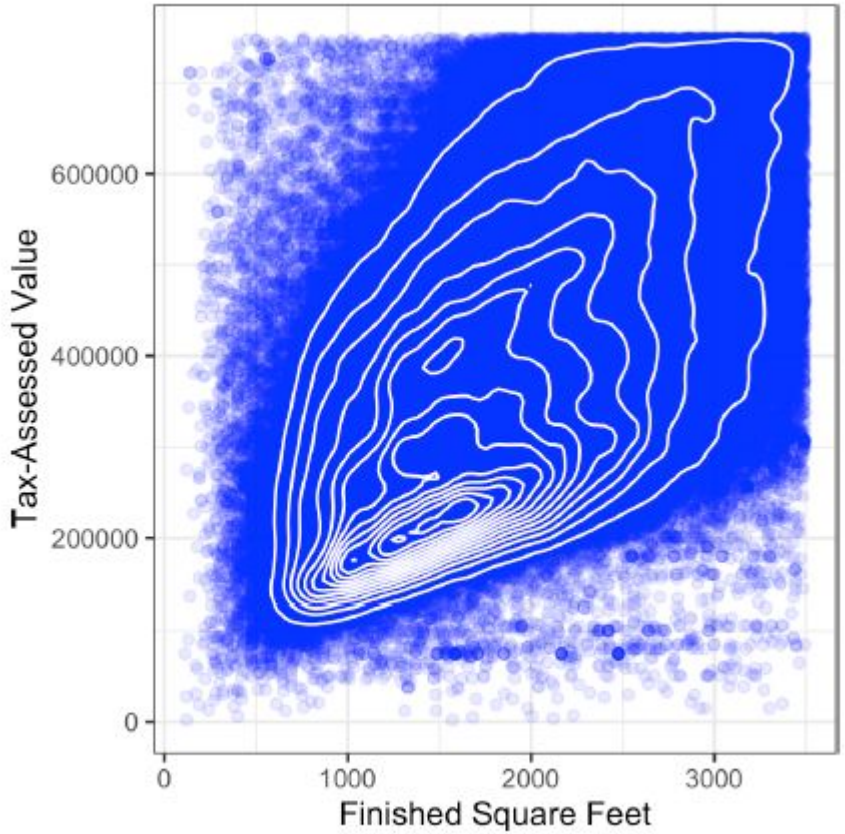
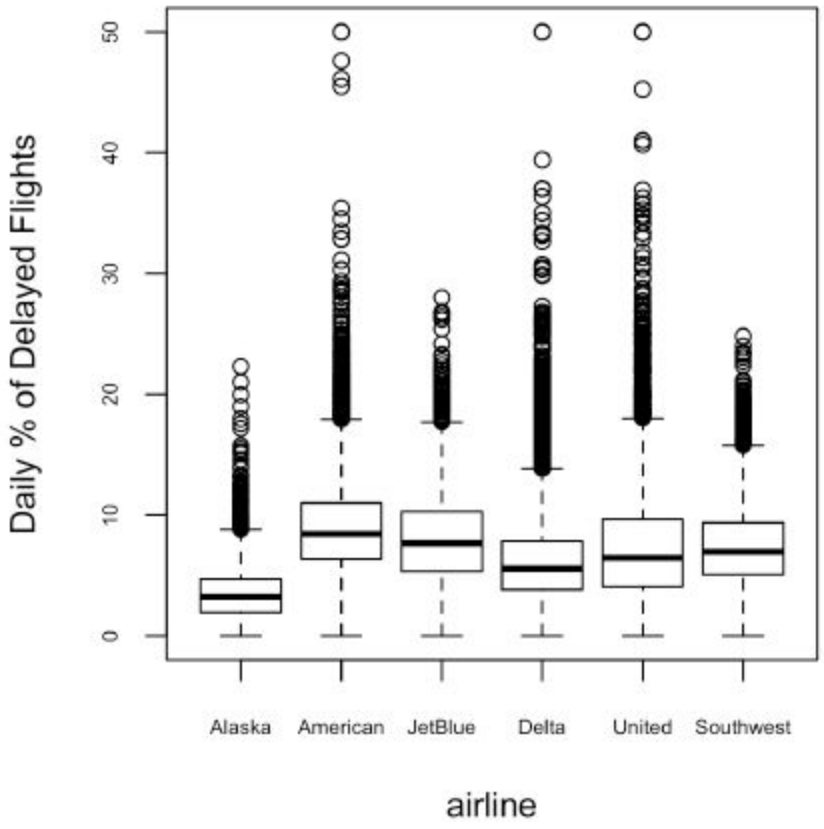


Figure 1-9. Contour plot for tax-assessed value versus finished square feet


Daily % of Delayed Flights
0
10
20
30
40
50
Alaska
American
JetBlue
Delta
United
Southwest
airline

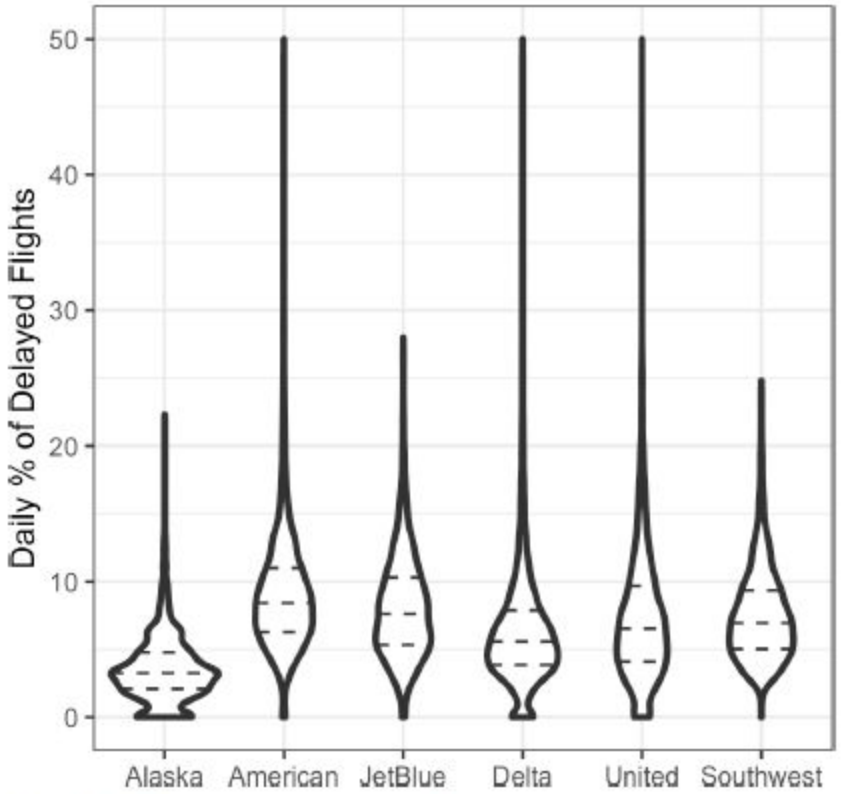


Figure 1-11. Violin plot of percent of airline delays by carrier

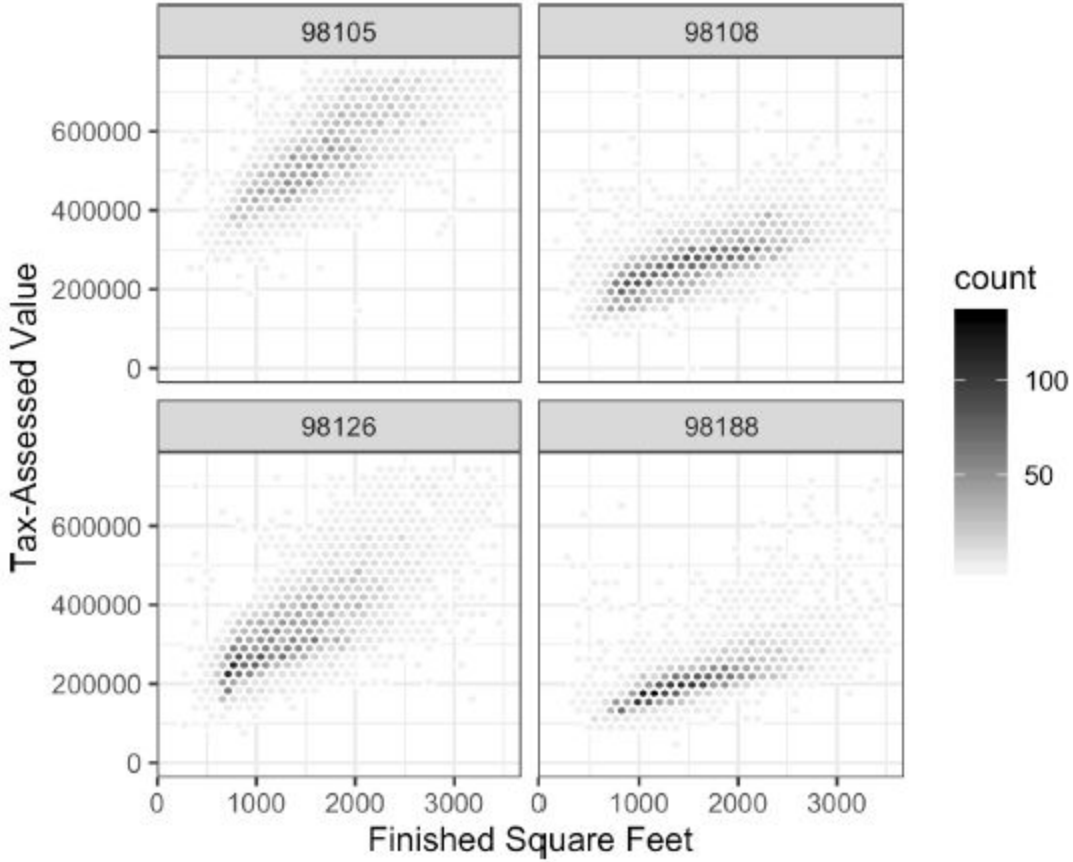


Figure 1-12. Tax-assessed value versus finished square feet by zip code

Summary

Exploratory data analysis (EDA), pioneered by John Tukey, set a foundation for the field of data science. The key idea of EDA is that the first and most important step in any project based on data is to look at the data. By summarizing and visualizing the data, you can gain valuable intuition and understanding of the project.

Project

https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/