

# Statistical Machine Learning for Data Science- BAD702

**Prepared By,  
Dr. Anitha DB  
Associate Professor & Head  
Department of CSE-Data Science  
ATME College of Engineering, Mysuru**

[GitHub - gedeck/practical-statistics-for-data-scientists: Code repository for O'Reilly book](https://github.com/gedeck/practical-statistics-for-data-scientists)

## Module-1

**Exploratory Data Analysis:** Estimates of locations and variability, Exploring data distributions, Exploring binary and categorical data, Exploring two or more variables.

**Textbook: Chapter 1**

## Module-2

**Data and Sampling Distributions:** Random sampling and bias, selection bias, sampling distribution of statistic, bootstrap, confidence intervals, data distributions: normal, long tailed, student's-t, binomial, Chi-square, F distribution, Poisson and related distributions.

**Textbook: Chapter 2**

## Module-3

**Statistical Experiments and Significance Testing:** A/B testing, hypothesis testing, resampling, statistical significance & p-values, t-tests, multiple testing, degrees of freedom.

**Textbook: Chapter 3**

## Module-4

Multi-Arm Bandit algorithm, power and sample size, factor variables in regression, interpreting the regression equation, Regression diagnostics, Polynomial and Spline Regression.

**Textbook: Chapter 3 & 4**

## Module-5

**Discriminant Analysis:** Covariance Matrix, Fisher's Linear discriminant, Generalized Linear Models, Interpreting the coefficients and odd ratios, Strategies for Imbalanced Data.

**Textbook: Chapter 5**

## Exploratory Data Analysis is the first step in any of the Data Science Projects

### Topics

- 1.Estimates of locations** – Mean, Weighted mean, Median, Weighted median, Trimmed mean, Robust, Outlier
- 2.Estimates of variability-** Deviations, Variance, Standard deviation, Mean absolute deviation, Median absolute deviation from the median, Range, Order statistics, Percentile, Interquartile range
- 3.Exploring data distributions-** Boxplot, Frequency table, Histogram, Density plot
- 4.Exploring binary and categorical data-** Mode, Expected value, Bar charts, Pie charts, Correlation(Correlation coefficient, Correlation matrix, Scatterplot)
- 5.Exploring two or more variables-** Contingency tables, Hexagonal binning, Contour plots, Violin plots

**Textbook: Chapter 1**

## Topic 1: Estimates of locations

Variables with measured or count data might have thousands of distinct values.

A basic step in exploring the data is getting a “typical value” for each feature (variable): Gives an estimate of where most of the data is located (i.e., its Central Tendency).

### Key Terms for Estimates of Location

- i. **Mean:** The sum of all values divided by the number of values (Average).
- ii. **Weighted mean:** The sum of all values times a weight divided by the sum of the weights (weighted Average).
- iii. **Median:** The value such that one-half of the data lies above and below (50th percentile).
- iv. **Weighted median:** The value such that one-half of the sum of the weights lies above and below the sorted data.
- v. **Trimmed mean:** The average of all values after dropping a fixed number of extreme values(truncated mean)
- vi. **Robust:** Not sensitive to extreme values(resistant).
- vii. **Outlier:** A data value that is very different from most of the data(extreme value)

## Mean

The most basic estimate of location is the mean, or average value.

The mean is the sum of all the values divided by the number of values.

Consider the following set of numbers: {3 5 1 2}.

The mean is  $(3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75$ .

You will encounter the symbol  $\bar{x}$  (pronounced “x-bar”) to represent the mean of a sample from a population.

The formula to compute the mean for a set of  $n$  values  $x_1, x_2, \dots, x_N$  is:

$$\text{Mean} = \bar{x} = \frac{\sum_i^n x_i}{n}$$

$N$  (or  $n$ ) refers to the total number of records or observations

## Trimmed Mean

A variation of the mean is a trimmed mean, which can be calculated by dropping a fixed number of sorted values at each end and then taking an average of the remaining values. Representing the sorted values by  $x_{(1)}, x_{(2)}, \dots, x_{(n)}$  where  $x_{(1)}$  is the smallest value and  $x_{(n)}$  the largest, the formula to compute the trimmed mean with  $p$  smallest and largest values omitted is:

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n - 2p}$$

A trimmed mean eliminates the influence of extreme values.

## Weighted Mean

Another type of mean is a weighted mean, which can be calculated by multiplying each data value  $x_i$  by a weight  $w_i$  and dividing their sum by the sum of the weights. The formula for a weighted mean is:

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^n w_i x_i}{\sum_i w_i}$$

There are **two main motivations** for using a weighted mean:

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight. For example, if we are taking the average from multiple sensors and one of the sensors is less accurate, then we might downweight the data from that sensor.
- The data collected does not equally represent the different groups that we are interested in measuring. For example, because of the way an online experiment was conducted, we may not have a set of data that accurately reflects all groups in the user base. To correct that, we can give a higher weight to the values from the groups that were underrepresented.



## Median and Robust Estimates

- The **median** is the middle value in a sorted list of numbers.
- If there's an even number of values, it's the average of the two middle numbers.

## Median vs. Mean

- The **mean** uses all data values, making it **sensitive to outliers**.
- The **median** only uses the center values, so it's **more robust** when data is skewed or contains extreme values.

## Real-Life Example

- In neighborhoods like **Medina**, where **Bill Gates** lives, the **mean income** is inflated due to extreme wealth.
- The **median income** gives a **better idea of a typical household** because it isn't affected by one very large value.

## Weighted Median

- Like a weighted mean, you can calculate a **weighted median**.
- Each value has a weight, and the weighted median splits the total weight into two equal halves.(sum of the weights is equal for the lower and upper halves of the sorted list)
- It's also **robust to outliers**, just like the regular median.



## Outliers

- The **median** is often referred to as a *robust estimate of location* because it is not influenced by **outliers**—extreme values that differ significantly from the rest of the data.
- An **outlier** is any data point that is far away(very distant) from the other values in a dataset.
- Being an outlier does **not necessarily** mean that a data value is incorrect or invalid. For example, in the case of income data, someone like **Bill Gates** would be a legitimate outlier due to his extreme wealth.
- However, outliers are **often caused by errors**, such as mixing units (e.g., kilometers versus meters), data entry mistakes, or sensor malfunctions. In these situations, the **mean** can be misleading, since it is sensitive to extreme values. In contrast, the **median** remains a reliable indicator of central tendency, even in the presence of such anomalies.
- Regardless of their cause, **outliers should always be identified and investigated**, as they may reveal important issues or insights within the data.
- The **median** is not the only robust estimate of location. In fact, a **trimmed mean** is widely used to avoid the influence of outliers. For example, trimming the bottom and top 10% (a common choice) of the data will provide protection against outliers in all but the smallest data sets. The **trimmed mean** can be thought of as a **compromise between the median and the mean**: it is robust to extreme values in the data, but uses more data to calculate the estimate for location.

## Example: Location Estimates of Population and Murder Rates

Table 1-2 shows the first few rows in the data set containing population and murder rates (in units of murders per 100,000 people per year) for each state.

Table 1-2. A few rows of the data.frame state of population and murder rate by state

	State	Population	Murder rate
1	Alabama	4,779,736	5.7
2	Alaska	710,231	5.6
3	Arizona	6,392,017	4.7
4	Arkansas	2,915,918	5.6
5	California	37,253,956	4.4
6	Colorado	5,029,196	2.8
7	Connecticut	3,574,097	2.4
8	Delaware	897,934	5.8

Compute the mean, trimmed mean, and median for the population using R: state

```
> state <- read.csv(file="/Users/andrewbruce1/book/state.csv")
> mean(state[["Population"]])
[1] 6162876
> mean(state[["Population"]], trim=0.1)
[1] 4783697
> median(state[["Population"]])
[1] 4436370
```

The mean is bigger than the trimmed mean, which is bigger than the median. This is because the trimmed mean excludes the largest and smallest five states (trim=0.1 drops 10% from each end).

If we want to compute the average murder rate for the country, we need to use a weighted mean or median to account for different populations in the states. Since base R doesn't have a function for weighted median, we need to install a package such as matrixStats:

```
> weighted.mean(state[["Murder.Rate"]], w=state[["Population"]])  
[1] 4.445834  
> library("matrixStats")  
> weightedMedian(state[["Murder.Rate"]], w=state[["Population"]])  
[1] 4.4
```

In this case, the weighted mean and median are about the same.

## KEY IDEAS

- The basic metric for location is the mean, but it can be sensitive to extreme values (outlier).
- Other metrics (median, trimmed mean) are more robust to outliers

## Topic 2- Estimates of variability

### Variability

**Location** is just one aspect of summarizing a dataset. A second critical dimension is **variability**, also known as **dispersion**, which describes how spread out or tightly clustered the data values are.

At the core of statistics lies the concept of **variability**—understanding and managing it is central to all statistical reasoning. This includes:

- **Measuring** variability,
- **Reducing** it when possible,
- **Distinguishing** between random variation and meaningful differences,
- **Identifying** sources of real variability,
- And **making decisions** in the presence of uncertainty.

Understanding variability is essential not only for accurate data analysis but also for making sound inferences and predictions.

### Key Terms for Variability Metrics

**Deviations** :The difference between the observed values and the estimate of location(errors, residuals).

**Variance** :The sum of squared deviations from the mean divided by  $n - 1$  where  $n$  is the number of data values(mean-squared-error).

**Standard deviation**: The square root of the variance(l2-norm, Euclidean norm)

**Mean absolute deviation**: The mean of the absolute value of the deviations from the mean(l1-norm, Manhattan norm).

**Median absolute deviation from the median**: The median of the absolute value of the deviations from the median.

**Range**: The difference between the largest and the smallest value in a data set.

**Order statistics**: Metrics based on the data values sorted from smallest to biggest(ranks).

**Percentile**: The value such that  $P$  percent of the values take on this value or less and  $(100-P)$  percent take on this value or more(quantile).

**Interquartile range**: The difference between the 75th percentile and the 25th percentile(IQR).

### Standard Deviation and Related Estimates

The most widely used estimates of variation are based on the differences, or deviations, between the estimate of location and the observed data. For a set of data {1, 4, 4}, the mean is 3 and the median is 4. The deviations from the mean are the differences:  $1 - 3 = -2$ ,  $4 - 3 = 1$ ,  $4 - 3 = 1$ . These deviations tell us how dispersed the data is around the central value.

One way to measure variability is to estimate a typical value for these deviations. Averaging the deviations themselves would not tell us much — the negative deviations offset the positive ones. In fact, the sum of the deviations from the mean is precisely zero. Instead, a simple approach is to take the average of the absolute values of the deviations from the mean.

In the preceding example, the absolute value of the deviations is {2 1 1} and their average is  $(2 + 1 + 1) / 3 = 1.33$ . This is known as the mean absolute deviation and is computed with the formula

$$\text{Mean absolute deviation} = \frac{\sum_{i=1}^n |x_i - \bar{x}|}{n}$$

where  $\bar{x}$  is the sample mean



### Standard Deviation and Related Estimates

- The best-known estimates for variability are the **variance and the standard deviation**, which are based on squared deviations. The variance is an average of the squared deviations, and the standard deviation is the square root of the variance.

$$\text{Variance} = s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

- The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data.
- The standard deviation is preferred in statistics over the mean absolute deviation (mathematically, working with squared values is much more convenient than absolute values, especially for statistical models)
- Neither the variance, the standard deviation, nor the mean absolute deviation is robust to outliers and extreme values. The variance and standard deviation are especially **sensitive to outliers** since they are based on the squared deviations.
- A **robust estimate** of variability is the median absolute deviation from the median or MAD



## Standard Deviation and Related Estimates

$$\text{Median absolute deviation} = \text{Median}(|x_1 - m|, |x_2 - m|, \dots, |x_N - m|)$$

Where  $m$  is the median. Like the median, the MAD is not influenced by extreme values. It is also possible to compute a trimmed standard deviation analogous to the trimmed mean .

## Estimates Based on Percentiles

A different approach to estimating dispersion is based on looking at the spread of the sorted data. Statistics based on sorted (ranked) data are referred to as order statistics.

The most basic measure is the **range**: the difference between the largest and smallest number.

The minimum and maximum values themselves are useful to know, and helpful in identifying outliers, but the range is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data.

To **avoid the sensitivity to outliers**, we can look at the range of the data after dropping values from each end. Formally, these types of estimates are based on differences between **percentiles**.

In a data set, the  $P$ th percentile is a value such that at least  $P$  percent of the values take on this value or less and at least  $(100 - P)$  percent of the values take on this value or more.

For example, to find the 80th percentile, sort the data. Then, starting with the smallest value, proceed 80 percent of the way to the largest value. Note that the median is the same thing as the 50th percentile. The percentile is essentially the same as a quantile, with quantiles indexed by fractions (so the .8 quantile is the same as the 80th percentile).

## Estimates Based on Percentiles

A common measurement of variability is the difference between the 25th percentile and the 75th percentile, called the interquartile range (or IQR).

Here is a simple example: 3,1,5,3,6,7,2,9.

We sort these to get 1,2,3,3,5,6,7,9.

The 25th percentile is at 2.5, and the 75th percentile is at 6.5, so the interquartile range is  $6.5 - 2.5 = 4$ .

If we have an even number of data ( $n$  is even), then the percentile is ambiguous under the preceding definition. In fact, we could take on any value between the order statistics  $x_{(j)}$  and  $x_{(j+1)}$  where  $j$  satisfies:

$$100 * \frac{j}{n} \leq P < 100 * \frac{j+1}{n}$$

Estimates Based on Percentiles

Example: Variability Estimates of State Population

Table 1-3 shows the first few rows in the data set containing population and murder rates for each state.

Table 1-3. A few rows of the data.frame state of population and murder rate by state

	State	Population	Murder rate
1	Alabama	4,779,736	5.7
2	Alaska	710,231	5.6
3	Arizona	6,392,017	4.7
4	Arkansas	2,915,918	5.6
5	California	37,253,956	4.4
6	Colorado	5,029,196	2.8
7	Connecticut	3,574,097	2.4
8	Delaware	897,934	5.8

Using R’s built-in functions for the standard deviation, interquartile range (IQR), and the median absolution deviation from the median (MAD), we can compute estimates of variability for the state population data:

```

> sd(state[["Population"]])
[1] 6848235
> IQR(state[["Population"]])
[1] 4847308
> mad(state[["Population"]])
[1] 3849870

```

Each of the estimates we've discussed sums up the data in a single number to describe the location or variability of the data. It is also useful to explore how the data is distributed overall

### Key Terms for Exploring the Distribution

**Boxplot** :A plot introduced by Tukey as a quick way to visualize the distribution of data (Box and whiskers plot).

**Frequency table** : A tally of the count of numeric data values that fall into a set of intervals (bins).

**Histogram axis**: A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y axis

**Density plot** : A smoothed version of the histogram, often based on a kernel density estimate

### Percentiles and Boxplots

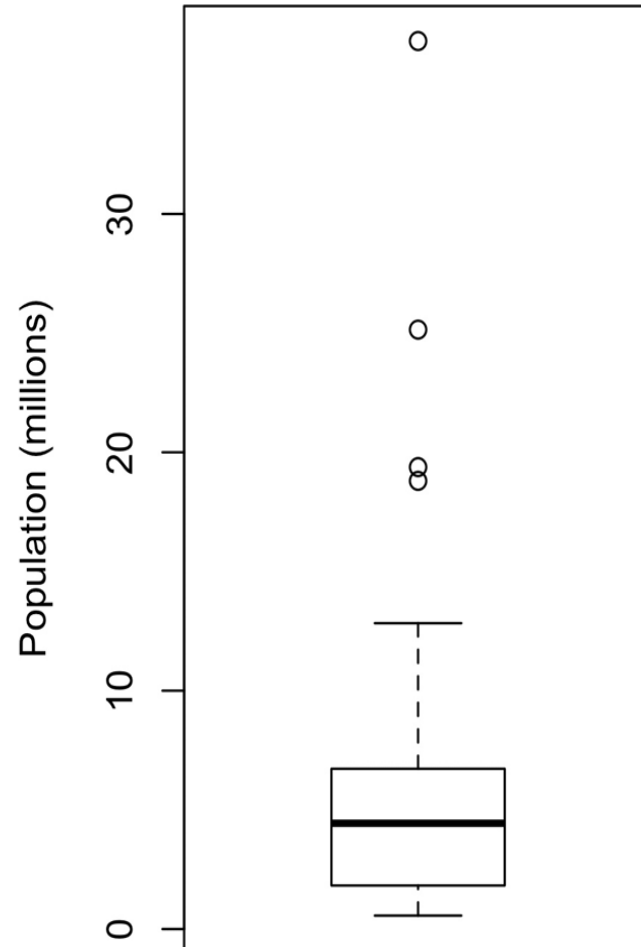
Percentiles are valuable to summarize the entire distribution. It is common to report the quartiles (25th, 50th, and 75th percentiles) and the deciles (the 10th, 20th, ..., 90th percentiles). Percentiles are especially valuable to summarize the tails (the outer range) of the distribution.

```
quantile(state[["Murder.Rate"]], p=c(.05, .25, .5, .75, .95))
  5%   25%   50%   75%   95%
1.600 2.425 4.000 5.550 6.510
```

Table 1-4. Percentiles  
of murder rate by  
state

5%	25%	50%	75%	95%
1.60	2.42	4.00	5.55	6.51

The median is 4 murders per 100,000 people, although there is quite a bit of variability: the 5th percentile is only 1.6 and the 95th percentile is 6.51



Boxplots, introduced by Tukey [Tukey-1977], are based on percentiles and give a quick way to visualize the distribution of data.

Figure 1-2 shows a boxplot of the population by state produced by R

```
boxplot(state[["Population"]]/1000000, ylab="Population (millions)")
```

- The top and bottom of the box are the 75th and 25th percentiles, respectively.
- The median is shown by the horizontal line in the box.
- The dashed lines, referred to as whiskers, extend from the top and bottom to indicate the range for the bulk of the data.
- Any data outside of the whiskers is plotted as single points



## Topic 3- Exploring the Data Distribution

Frequency Table and Histograms : A frequency table of a variable divides up the variable range into equally spaced segments, and tells us how many values fall in each segment.

Table 1-5 shows a frequency table of the population by state computed in R

```
breaks <- seq(from=min(state[["Population"]]),
              to=max(state[["Population"]]), length=11)
pop_freq <- cut(state[["Population"]], breaks=breaks,
               right=TRUE, include.lowest = TRUE)
table(pop_freq)
```

- The least populous state is Wyoming, with 563,626 people (2010 Census) and the most populous is California, with 37,253,956 people.
- 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 bigger pictures is lost

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
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

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

To create a histogram corresponding to Table 1-5 in R, use the hist function with the breaks argument:

```
hist(state[["Population"]], breaks=breaks)
```

The histogram is shown in Figure 1.3.

In general, histograms are plotted such that:

- Empty bins are included in the graph.
- Bins are equal width.
- Number of bins (or, equivalently, bin size) is up to the user.
- Bars are contiguous — no empty space shows between bars, unless there is an empty bin

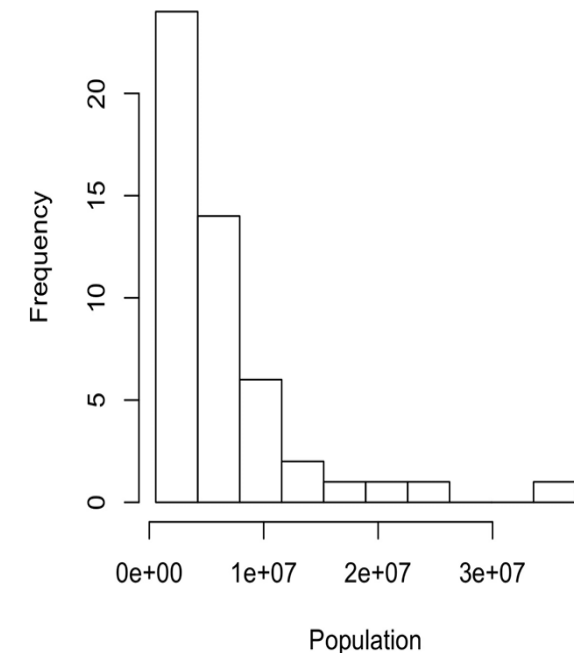


Figure 1-3. Histogram of state populations

### Density Estimates

Density plot shows the distribution of data values as a continuous line.

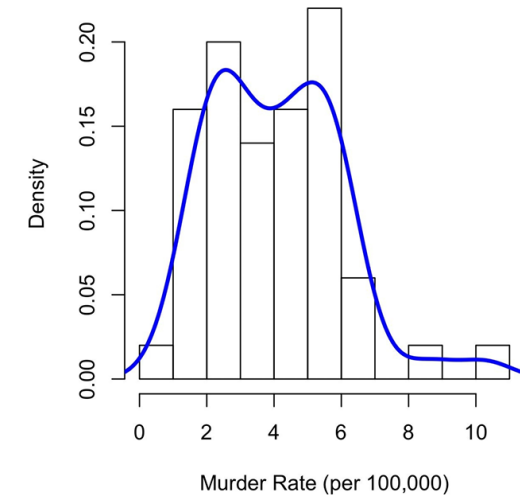
A density plot can be thought of as a smoothed histogram, although it is typically computed directly from the data through a kernel density estimate .

Figure 1-4 displays a density estimate superposed on a histogram.

In R, we can compute a density estimate using the density function:

```
hist(state[["Murder.Rate"]], freq=FALSE)
```

```
lines(density(state[["Murder.Rate"]]), lwd=3, col="blue")
```



A key distinction from the histogram plotted in Figure 1-3 is the scale of the y axis: a density plot corresponds to plotting the histogram as a proportion rather than counts (In R use the argument `freq=FALSE`)

### 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)

## Key Terms for Exploring Categorical Data

**Mode:** The most commonly occurring category or value in a data set.

**Expected value :**When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.

**Bar charts:** The frequency or proportion for each category plotted as bars.

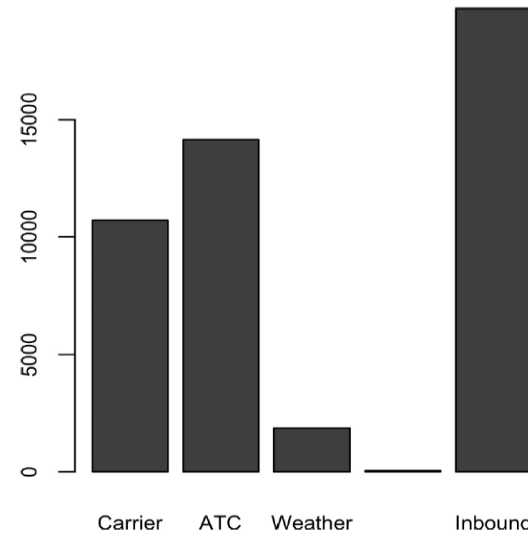
**Pie charts:** The frequency or proportion for each category plotted as wedges in a pie

Getting a summary of a binary variable or a categorical variable with a few categories is a easy matter.

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 are a common visual tool for displaying a single categorical variable, often seen in the popular press. Categories are listed on the x-axis, and frequencies or proportions on the y-axis.

Figure 1-5 shows the airport delays per year by cause for Dallas/Fort Worth, and it is produced with the R function `barplot`:

```
barplot(as.matrix(dfw)/6, cex.axis=.5)
```

**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.



### Mode

The mode is the value — or values in case of a tie — that appears most often in the data. For example, the mode of the cause of delay at Dallas/Fort Worth airport is “Inbound.” As another example, in most parts of the United States, the mode for religious preference would be Christian. The mode is a simple summary statistic for categorical data, and it is generally not used for numeric data.

### Expected Value

A special type of categorical data is data in which the categories represent or can be mapped to discrete values on the same scale. A marketer for a new cloud technology, for example, offers two levels of service, one priced at \$300/month and another at \$50/month. The marketer offers free webinars to generate leads, and the firm figures that 5% of the attendees will sign up for the \$300 service, 15% for the \$50 service, and 80% will not sign up for anything. This data can be summed up, for financial purposes, in a single “expected value,” which is a form of **weighted mean** in which the weights are probabilities.

The expected value is calculated as follows:

1. Multiply each outcome by its probability of occurring.
2. Sum these values.



In the cloud service example, the expected value of a webinar attendee is thus \$22.50 per month, calculated as follows:

$$EV = (0.05)(300) + (0.15)(50) + (0.80)(0) = 22.5$$

The expected value is really a form of weighted mean: it adds the ideas of future expectations and probability weights, often based on subjective judgment.

Expected value is a fundamental concept in business valuation and capital budgeting — for example, the expected value of five years of profits from a new acquisition, or the expected cost savings from new patient management software at a clinic.

### 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.

### Key Terms for Exploring Two or More Variables

- **Contingency tables:** A tally of counts between two or more categorical variables.
  - **Hexagonal binning:** A plot of two numeric variables with the records binned into hexagons.
  - **Contour plots:** A plot showing the density of two numeric variables like a topographical map.
  - **Violin plots:** Similar to a boxplot but showing the density estimate.
- 
- Familiar estimators like mean and variance look at variables one at a time (*univariate analysis*).
  - Correlation analysis is an important method that compares two variables (*bivariate analysis*).
  - In this section we look at additional estimates and plots, and at more than two variables (*multivariate analysis*).
- 
- Like univariate analysis, bivariate analysis involves both computing summary statistics and producing visual displays.
  - The appropriate type of bivariate or multivariate analysis depends on the nature of the data: numeric versus categorical.

### Hexagonal Binning and Contours (Plotting Numeric versus Numeric Data)

Scatterplots are fine when there is a relatively small number of data values.

The plot of stock returns in Figure 1-7 involves only about 750 points. For data sets with hundreds of thousands or millions of records, a scatterplot will be too dense, so we need a different way to visualize the relationship.

To illustrate, consider the data set `kc_tax`, which contains the tax-assessed values for residential properties in King County, Washington. In order to focus on the main part of the data, we strip out very expensive and very small or large residences using the `subset` function:

```
kc_tax0 <- subset(kc_tax, TaxAssessedValue < 750000 & SqFtTotLiving>100 &
                  SqFtTotLiving<3500)
nrow(kc_tax0)
[1] 432733
```

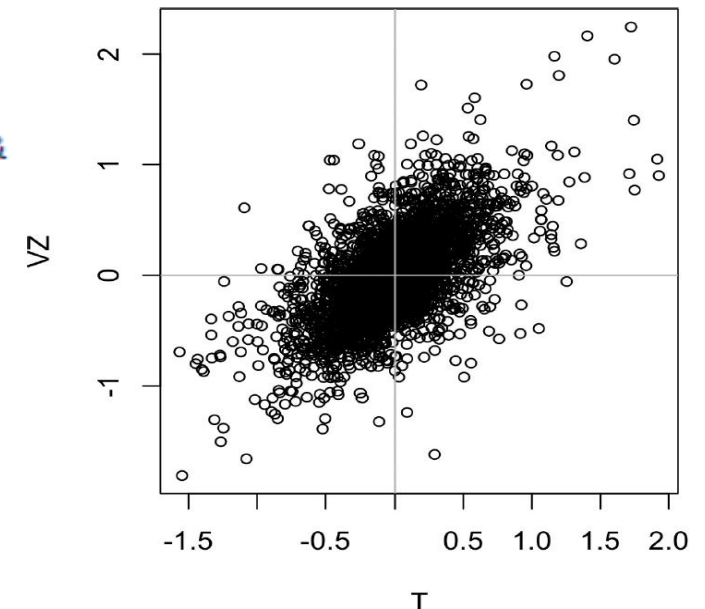


Figure 1-7. Scatterplot between returns for ATT and Verizon

Figure 1-8 is a *hexagon binning* plot of the relationship between the finished square feet versus the tax-assessed value for homes in King County.

Rather than plotting points, which would appear as a monolithic dark cloud, we grouped the records into hexagonal bins and plotted the hexagons with a color indicating the number of records in that bin.

In this chart, the positive relationship between square feet and tax-assessed value is clear.

An interesting feature is the hint of a second cloud above the main cloud, indicating homes that have the same square footage as those in the main cloud, but a higher tax-assessed value.

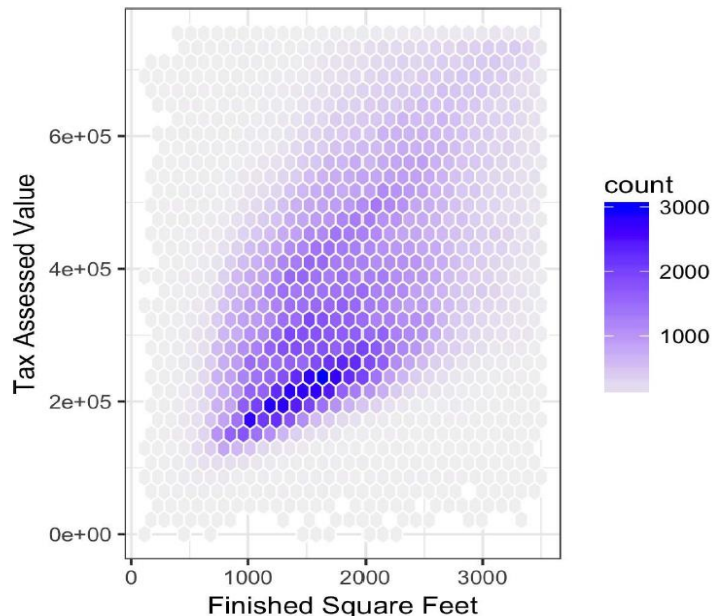


Figure 1-8. Hexagonal binning for tax-assessed value versus finished square feet

Figure 1-8 was generated by the powerful R package `ggplot2`, developed by Hadley Wickham [[ggplot2](#)]. `ggplot2` is one of several new software libraries for advanced exploratory visual analysis of data; see “[Visualizing Multiple Variables](#)”.

```
ggplot(kc_tax0, (aes(x=SqFtTotLiving, y=TaxAssessedValue))) +
  stat_binhex(colour="white") +
  theme_bw() +
  scale_fill_gradient(low="white", high="black") +
  labs(x="Finished Square Feet", y="Tax Assessed Value")
```



**Figure 1-9** uses contours overlaid on a scatterplot to visualize the relationship between two numeric variables. The contours are essentially a topographical map to two variables; each contour band represents a specific density of points, increasing as one nears a “peak.” This plot shows a similar story as **Figure 1-8**: there is a secondary peak “north” of the main peak. This chart was also created using `ggplot2` with the built-in `geom_density2d` function.

```
ggplot(kc_tax0, aes(SqFtTotLiving, TaxAssessedValue)) +
  theme_bw() +
  geom_point(alpha=0.1) +
  geom_density2d(colour="white") +
  labs(x="Finished Square Feet", y="Tax Assessed Value")
```

- Other types of charts are used to show the relationship between two numeric variables, including *heat maps*.
- Heat maps, hexagonal binning, and contour plots all give a visual representation of a two-dimensional density.
- In this way, they are natural analogs to histograms and density plots.

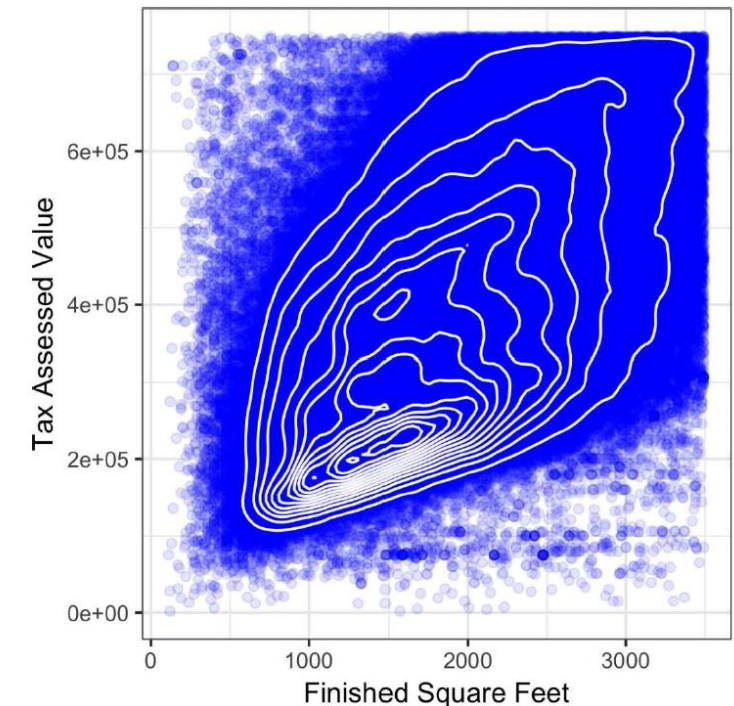


Figure 1-9. Contour plot for tax-assessed value versus finished square feet

### Two Categorical Variables

A useful way to summarize two categorical variables is a contingency table — a table of counts by category. **Table 1-8** shows the contingency table between the grade of a personal loan and the outcome of that loan. This is taken from data provided by Lending Club, a leader in the peer-to-peer lending business. The grade goes from A (high) to G (low). The outcome is either paid off, current, late, or charged off (the balance of the loan is not expected to be collected). This table shows the count and row percentages.

High-grade loans have a very low late/charge-off percentage as compared with lower-grade loans. Contingency tables can look at just counts, or also include column and total percentages.

Pivot tables in Excel are perhaps the most common tool used to create contingency tables.

In R, the `CrossTable` function in the `descr` package produces contingency tables, and the following code was used to create **Table 1-8**:

```
library(descr)
x_tab <- CrossTable(lc_loans$grade, lc_loans$status,
                    prop.c=FALSE, prop.chisq=FALSE, prop.t=FALSE)
```

*Table 1-8. Contingency table of loan grade and status*

Grade	Fully paid	Current	Late	Charged off	Total
A	20715	52058	494	1588	74855
	0.277	0.695	0.007	0.021	0.161
B	31782	97601	2149	5384	136916
	0.232	0.713	0.016	0.039	0.294
C	23773	92444	2895	6163	125275
	0.190	0.738	0.023	0.049	0.269
D	14036	55287	2421	5131	76875
	0.183	0.719	0.031	0.067	0.165
E	6089	25344	1421	2898	35752
	0.170	0.709	0.040	0.081	0.077
F	2376	8675	621	1556	13228
	0.180	0.656	0.047	0.118	0.028
G	655	2042	206	419	3322
	0.197	0.615	0.062	0.126	0.007
Total	99426	333451	10207	23139	466223

## Topic 5- Exploring Two or More Variables

### Categorical and Numeric Data

Boxplots are a simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.

For example, we might want to compare how the percentage of flight delays varies across airlines.

Figure 1-10 shows the percentage of flights in a month that were delayed where the delay was within the carrier's control.

```
boxplot(pct_delay ~ airline, data=airline_stats, ylim=c(0, 50))
```

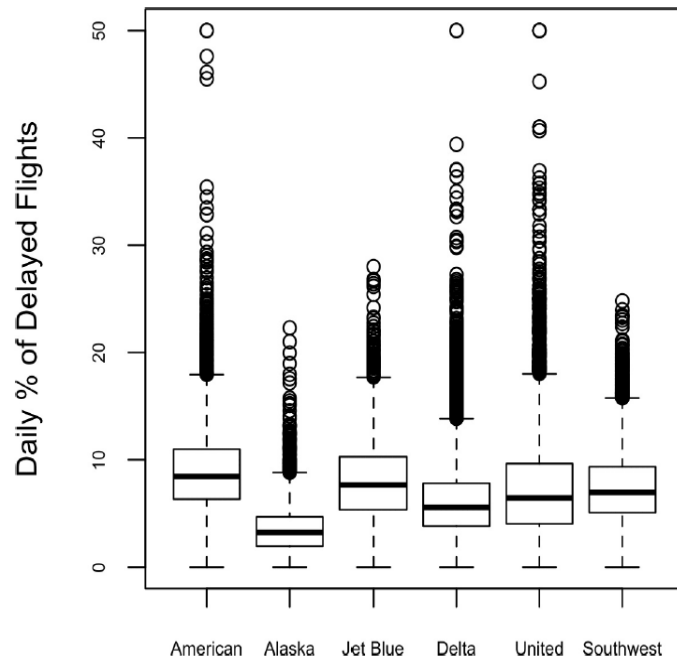


Figure 1-10. Boxplot of percent of airline delays by carrier

**Alaska** stands out as having the fewest delays, while American has the most delays: the lower quartile for American is higher than the upper quartile for Alaska.



## Topic 5- Exploring Two or More Variables

A *violin plot* is an enhancement to the boxplot and plots the density estimate with the density on the y-axis. The density is mirrored and flipped over and the resulting shape is filled in, creating an image resembling a violin. The advantage of a violin plot is that it can show nuances in the distribution that aren't perceptible in a boxplot. On the other hand, the boxplot more clearly shows the outliers in the data. In ggplot2, the function `geom_violin` can be used to create a violin plot as follows:

```
ggplot(data=airline_stats, aes(airline, pct_carrier_delay)) +
  ylim(0, 50) +
  geom_violin() +
  labs(x="", y="Daily % of Delayed Flights")
```

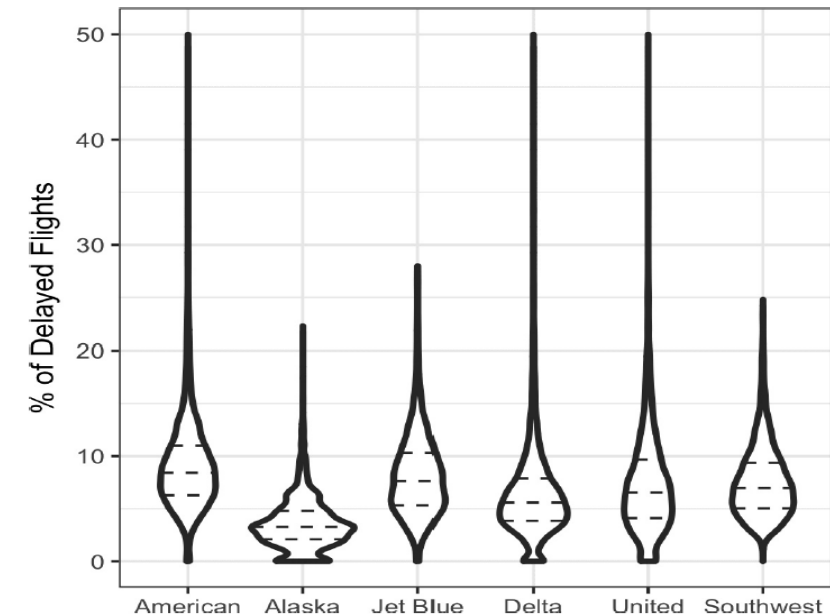


Figure 1-11. Combination of boxplot and violin plot of percent of airline delays by carrier

The corresponding plot is shown in Figure 1-11. The violin plot shows a concentration in the distribution near zero for Alaska, and to a lesser extent, Delta. This phenomenon is not as obvious in the boxplot. You can combine a violin plot with a boxplot by adding `geom_boxplot` to the plot (although this is best when colors are used).

## Topic 5- Exploring Two or More Variables

### Visualizing Multiple Variables

The types of charts used to compare two variables — scatterplots, hexagonal binning, and boxplots — are readily extended to more variables through the notion of *conditioning*.

As an example, look back at Figure 1-8, which showed the relationship between homes' finished square feet and tax-assessed values. We observed that there appears to be a cluster of homes that have higher tax-assessed value per square foot.

Diving deeper, Figure 1-12 accounts for the effect of location by plotting the data for a set of zip codes. Now the picture is much clearer: tax-assessed value is much higher in some zip codes (98112, 98105) than in others (98108, 98057).

This disparity gives rise to the clusters observed in Figure 1-8.

Figure 1-12 is created using ggplot2 and the idea of *facets*, or a conditioning variable (in this case zip code):

```
ggplot(subset(kc_tax0, ZipCode %in% c(98188, 98105, 98108, 98126)),  
       aes(x=SqFtTotLiving, y=TaxAssessedValue)) +  
  stat_binhex(colour="white") +  
  theme_bw() +  
  scale_fill_gradient( low="white", high="blue") +  
  labs(x="Finished Square Feet", y="Tax Assessed Value") +  
  facet_wrap("ZipCode")
```

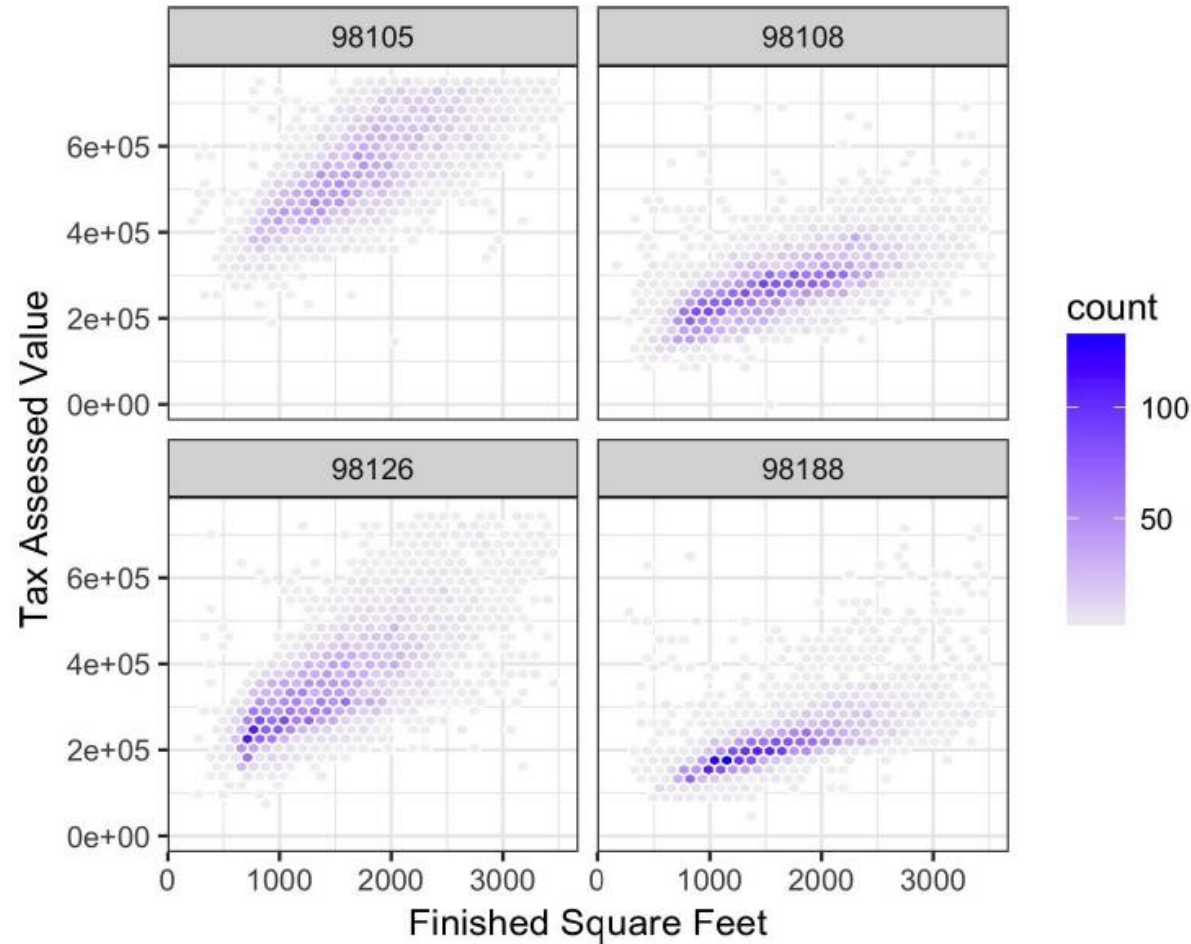


Figure 1-12. Tax-assessed value versus finished square feet by zip code

### KEY IDEAS

- Hexagonal binning and contour plots are useful tools that permit graphical examination of two numeric variables at a time, without being overwhelmed by huge amounts of data.
- Contingency tables are the standard tool for looking at the counts of two categorical variables.
- Boxplots and violin plots allow you to plot a numeric variable against a categorical variable.