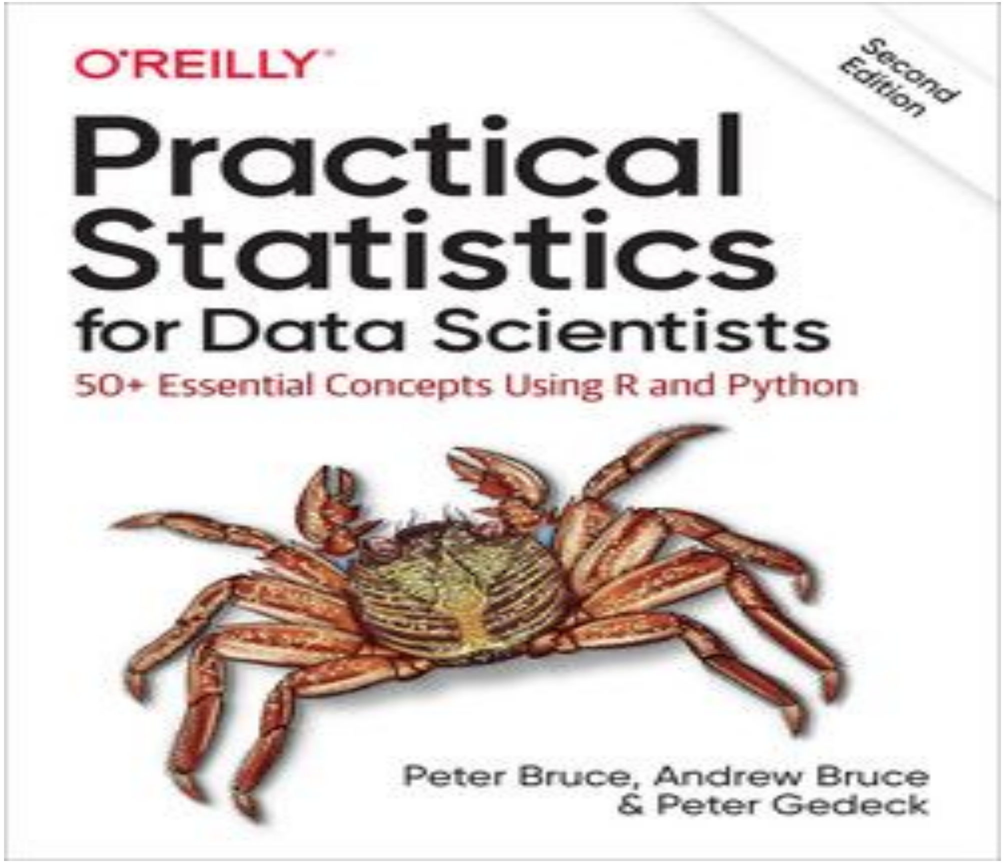


O'REILLY

Second Edition

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python

Peter Bruce, Andrew Bruce
& Peter Gedeck

Github link

https://github.com/gedeck/practical-statistics-for-data-scientists

Exploratory Data Analysis

John Tukey, the eminent statistician whose ideas developed over 50 years ago form the foundation of data science.

The field of exploratory data analysis was established with Tukey’s 1977 now-classic book Exploratory Data Analysis.

Tukey presented simple plots (e.g., boxplots, scatterplots) that, along with summary statistics (mean, median, quantiles), paint a picture of a data set.

John Tukey

Elements of Structured Data

Data comes from many sources: Sensors,Video,events etc

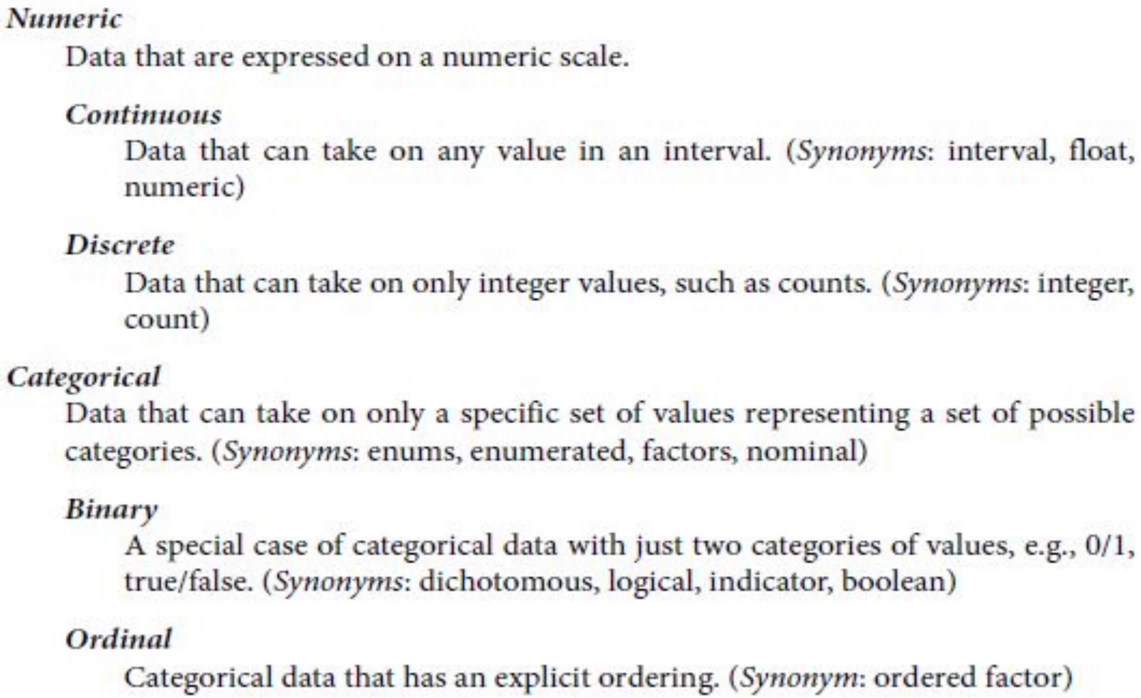
Data is unstructured : Image,Text,Clickstream

A major challenge of data science is to harness this torrent of raw data into actionable information. To apply the statistical concepts , unstructured raw data must be processed and manipulated into a structured form.

One of the commonest forms of structured data is a table with rows and columns—as data might emerge from a relational database

Data Types

Data type is important to help determine the type of visual display, data analysis, or statistical model



Numeric

Data that are expressed on a numeric scale.

Continuous

Data that can take on any value in an interval. (*Synonyms*: interval, float, numeric)

Discrete

Data that can take on only integer values, such as counts. (*Synonyms*: integer, count)

Categorical

Data that can take on only a specific set of values representing a set of possible categories. (*Synonyms*: enums, enumerated, factors, nominal)

Binary

A special case of categorical data with just two categories of values, e.g., 0/1, true/false. (*Synonyms*: dichotomous, logical, indicator, boolean)

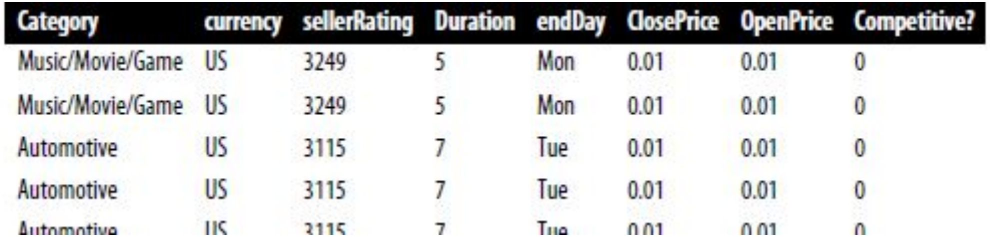
Ordinal

Categorical data that has an explicit ordering. (*Synonym*: ordered factor)

Rectangular Data

The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table.

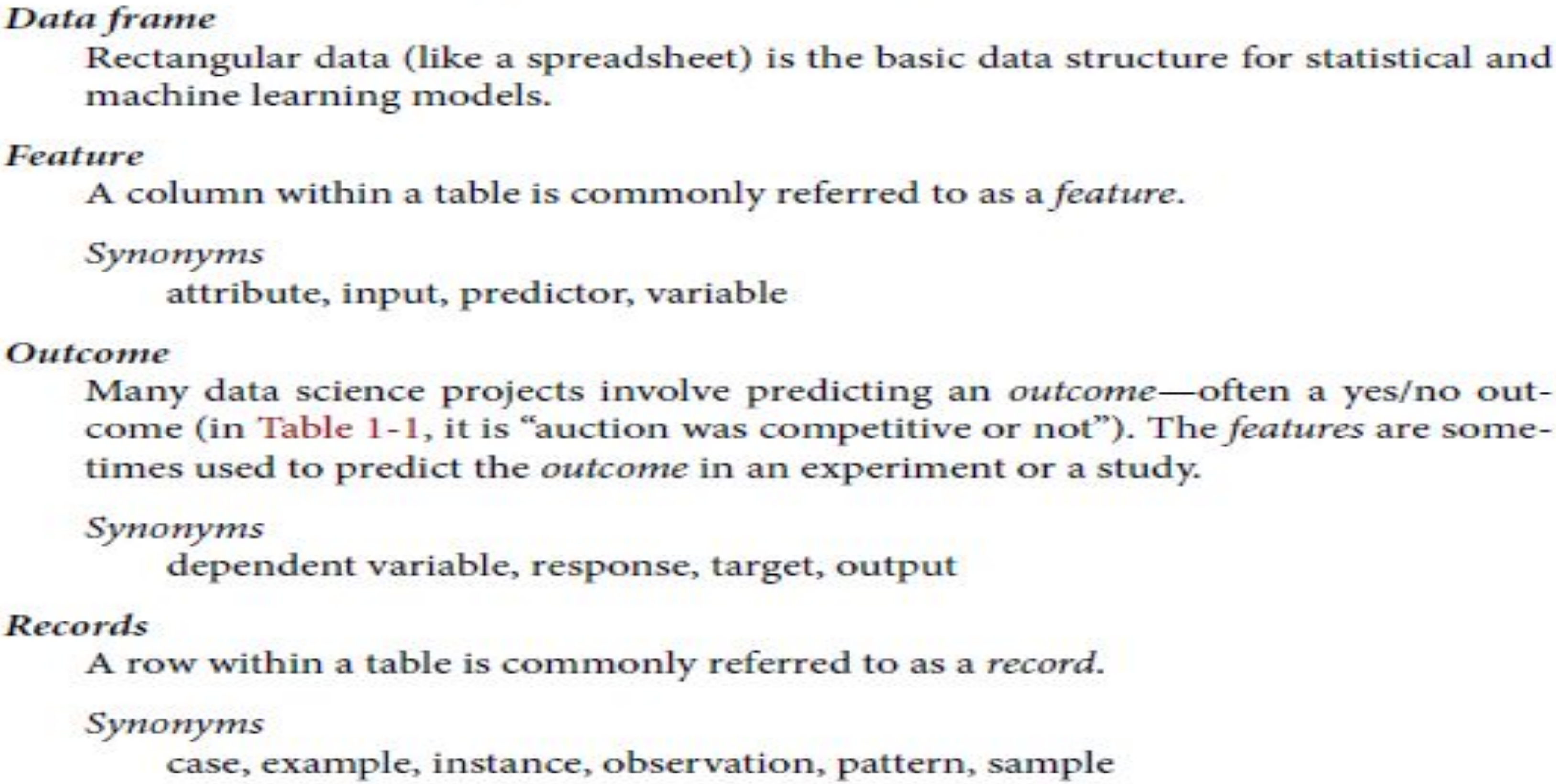
Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables); data frame is the specific format in R and Python.



Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0

A typical data frame format

Key Terms for Rectangular Data



Data frame

Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.

Feature

A column within a table is commonly referred to as a *feature*.

Synonyms

attribute, input, predictor, variable

Outcome

Many data science projects involve predicting an *outcome*—often a yes/no outcome (in Table 1-1, it is "auction was competitive or not"). The *features* are sometimes used to predict the *outcome* in an experiment or a study.

Synonyms

dependent variable, response, target, output

Records

A row within a table is commonly referred to as a *record*.

Synonyms

case, example, instance, observation, pattern, sample

Nonrectangular Data Structures

Spatial data structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures

Graph (or network) data structures are used to represent physical, social, and abstract relationships. For example, a graph of a social network, such as Facebook or LinkedIn

Each of these data types has its specialized methodology in data science.

Estimates of Location

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

Mean

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

Synonym

average

Weighted mean

The sum of all values times a weight divided by the sum of the weights.

Synonym

weighted average

Median

The value such that one-half of the data lies above and below.

Synonym

50th percentile

Percentile

The value such that *P* percent of the data lies below.

Synonym

quantile

Weighted median

The value such that one-half of the sum of the weights lies above and below the sorted data.

Trimmed mean

The average of all values after dropping a fixed number of extreme values.

Synonym

truncated mean

Robust

Not sensitive to extreme values.

Mean

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

The formula to compute the mean for a set of n values x1, x2, ..., xn is:

Mean = x =Σi=1n xi / 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 you calculate by dropping a fixed number of sorted values at each end and then taking an average of the remaining values.

A trimmed mean eliminates the influence of extreme values.

Representing the sorted values by x 1 , x 2 , ..., 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:

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

Weighted Mean

calculate by multiplying each data value xi by a user-specified weight wi and dividing their sum by the sum of the weights. The formula for a weighted mean is:

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

Median and Robust Estimates

The median is the middle number on a sorted list of the data. If there is an even number of data values, the middle value is one that is not actually in the data set, but rather the average of the two values that divide the sorted data into upper and lower halves.

Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data.

Reflection : median is a better metric for location?

Outlier

median is referred to as a robust estimate:-not influenced by outliers (extreme cases) that could skew the results.

An outlier is any value that is very distant from the other values in a data set

Reflection : Is median the only robust estimate of location?

The basic metric for location is the mean, but it can be sensitive to extreme values (outlier).

- Other metrics (median, trimmed mean) are less sensitive to outliers and unusual distributions and hence are more robust.

Estimates of Variability

Location is just one dimension in summarizing a feature. A second dimension, variability,also referred to as dispersion, measures whether the data values are tightly clustered or spread out.

Key Terms for Variability Metrics

Deviations The difference between the observed values and the estimate of location.

Synonyms : errors, residuals

Variance The sum of squared deviations from the mean divided by n – 1 where n is the number of data values.

Synonym : mean-squared-error

Standard deviation : The square root of the variance.

Mean absolute deviation

The mean of the absolute values of the deviations from the mean.

Synonyms

l1-norm, Manhattan norm

Median absolute deviation from the median

The median of the absolute values 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.

Synonym

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.

Synonym

quantile

Interquartile range

The difference between the 75th percentile and the 25th percentile.

Synonym

IQR

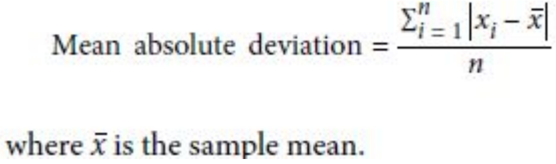
Deviations, between the estimate of location and the observed data.

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.

The sum of the deviations from the mean is precisely zero.

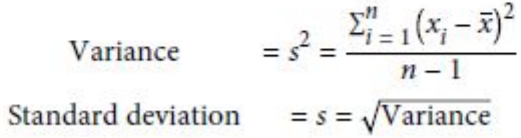
A simple approach is to take the average of the absolute values of the deviations from the mean.

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:



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

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



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

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

Reflection : standard deviation is preferred in statistics over the mean absolute deviation?

Note :

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.

Degrees of Freedom, and n or n – 1?

Median Absolute Deviation

The median absolute deviation from the median or MAD:

Median absolute deviation = Median x1 − m , x2 − m , ..., xN − m

where m is the median. Like the median, the MAD is not influenced by extreme values.

Estimates Based on Percentiles

Statistics based on sorted (ranked) data are referred to as order statistics.

A different approach to estimating dispersion is based on looking at the spread of the sorted data

range: the difference between the largest and smallest numbers

But the range is extremely sensitive to outliers and not very useful as a general measure of dispersion in the data.

Percentile

To avoid the sensitivity to outliers, we can look at the range of the data after dropping values from each end.

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).

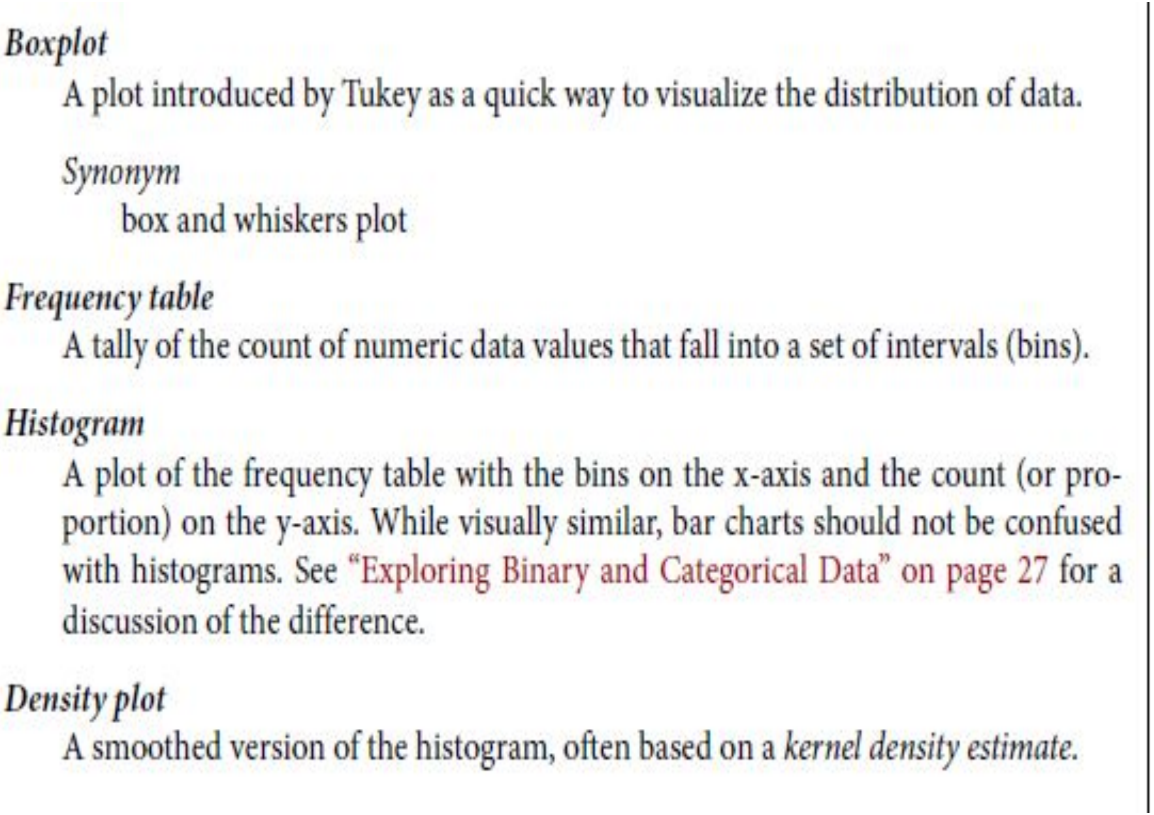
Note that the median is the same thing as the 50th percentile.

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

e.g.{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.

Exploring the Data Distribution

Each of the estimates we’ve covered 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.



Boxplot

A plot introduced by Tukey as a quick way to visualize the distribution of data.

Synonym

box and whiskers plot

Frequency table

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

Histogram

A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis. While visually similar, bar charts should not be confused with histograms. See "Exploring Binary and Categorical Data" on page 27 for a discussion of the difference.

Density plot

A smoothed version of the histogram, often based on a *kernel density estimate*.

Percentiles and Boxplots

Box Plots: based on percentiles and give a quick way to visualize the distribution of data.

Frequency Table and Histogram

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.

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

Note

In statistical theory, location and variability are referred to as the first and second moments of a distribution. The third and fourth moments are called skewness and kurtosis. Skewness refers to whether the data is skewed to larger or smaller values, and kurtosis indicates the propensity of the data to have extreme values. Generally, metrics are not used to measure skewness and kurtosis; instead, these are discovered through visual display.

Density Plot

It 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.

Density Plot

Density plots are used to observe the distribution of a variable in a dataset. It plots the graph on continuous interval or time-period. This is also known as Kernel density plot.

Density plots are a variation of Histograms. It charts the values from a selected column as equally binned distributions. It uses kernel smoothing to smoothen out noise. Thus, the plots are smooth across bins and are not affected by the number of bins created, which helps create a more defined distribution shape. The peaks of a Density Plot help display where values are concentrated over the interval.
Normal distribution curves are an example of density plots.

1.2

Exploring Binary and 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.

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 Chart : common visual tool for displaying a single categorical variable. Categories are listed on the x-axis, and frequencies or proportions on the y-axis.

Note : 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.

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 value—or values in case of a tie—that appears most often in the data.

The mode is a simple summary statistic for categorical data, and it is generally not used for numeric data.

Expected value

The expected value is calculated as follows:

1. Multiply each outcome by its probability of occurrence.

2. Sum these values.

EV = 0 . 05 300 + 0 . 15 50 + 0 . 80 0 = 22 . 5

correlation

Correlation coefficient

A metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1).

Correlation matrix

A table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables.

Pearson’s correlation coefficient

Pearson’s correlation coefficient, we multiply deviations from the mean for variable 1 times those for variable 2, and divide by the product of the standard deviations:

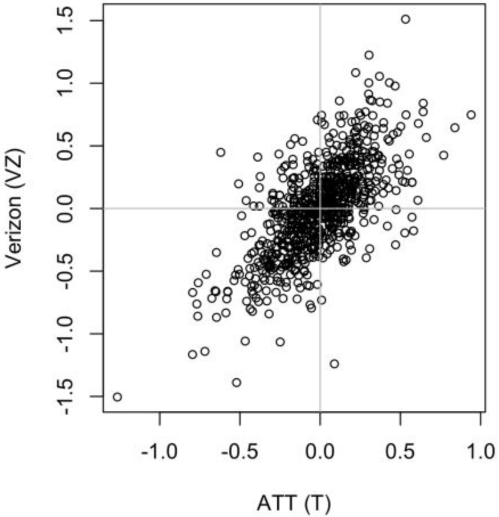
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

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.

Scatter Plot

Scatterplots are a straightforward way to visualize the data distribution in a XY plane, especially when we are looking for trends or clusters. But when you have a dataset with a large number of points, many of these data points can overlap. This overalpping effect can make difficult to see any trends or clusters.



Exploring Two or More Variables

One Variable : mean ,variance etc

Two Variable : Correlation

Multivariate : depends on the nature of the data: numeric versus categorical.

Key Terms

Contingency table : 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 plot : A plot showing the density of two numeric variables like a topographical map.

Violin plot : Similar to a boxplot but showing the density estimate.

Binning

Binning is a technique of data aggregation used for grouping a dataset of N values into less than N discrete groups. In this article we are considering only the case of datasets build up of (x,y) points distributed on a XY plane, but this technique is applicable in other cases. This technique is based on extremely simple concepts.

- the XY plane is uniformly tiled with polygons (squares, rectangles or hexagons).
- the number of points falling in each bin (tile) are counted and stored in a data structure.
- the bins with count > 0 are plotted using a color range (heatmap) or varying their size in proportion to the count.

Hexagonal Binning

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

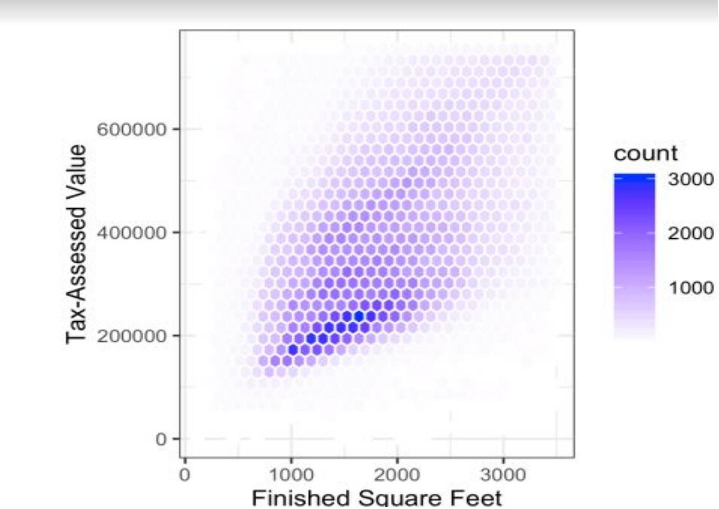
hexagonal binning:Rather than plotting points, which would appear as a monolithic dark cloud, we grouped the records into hexagonal bins and plotted the hexagons with a color indicating the number of records in that bin.

Hexagonal binning

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.

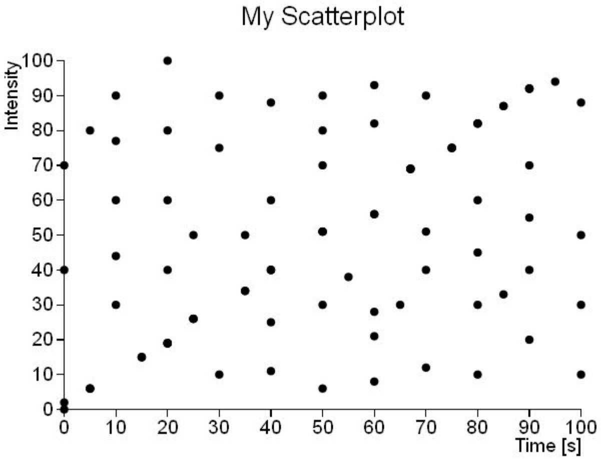
Hexagonal binning is a plot of two numeric variables with the records binned into hexagons.

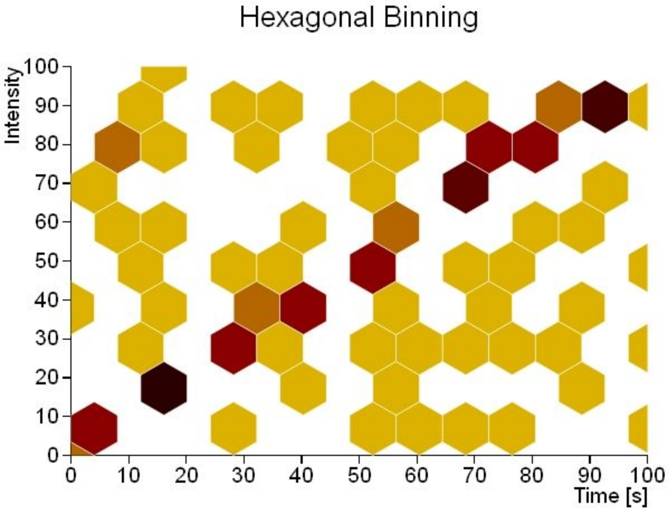
Rather than plotting points, records are grouped into hexagonal bins and color indicating the number of records in that bin.



Scatter plot vs hexagon Binning

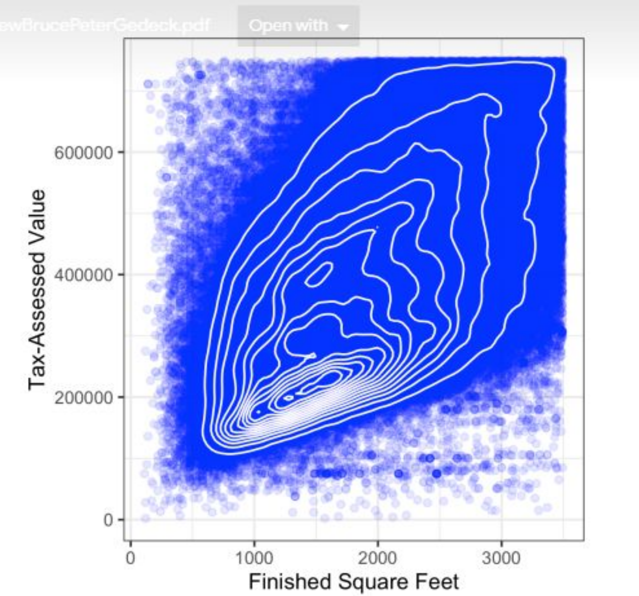
White bins are the bins where there is no data (count = 0).a linear trend is visible.





Contours

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."



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

Two Categorical Variables

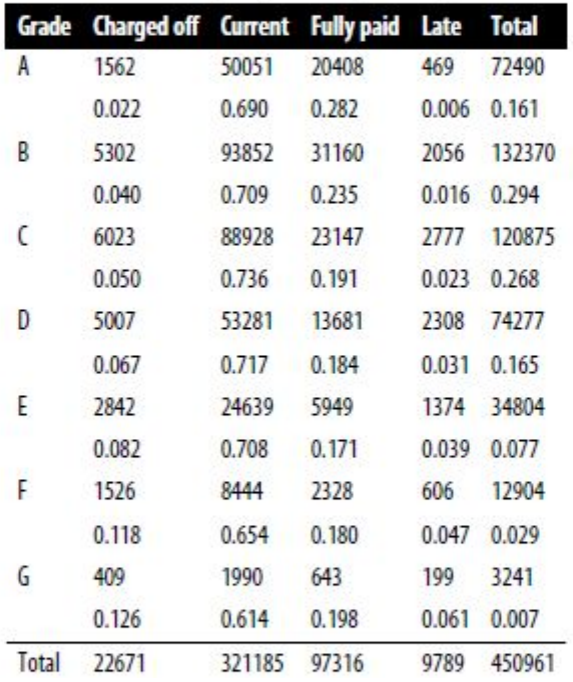
A useful way to summarize two categorical variables is a contingency table—a table of counts by category. Table shows the contingency table between the grade of a personal loan and the outcome of that loan.

Contingency tables can look only at counts, or

they can also include column and total percentages.

Pivot tables in Excel are perhaps the most common

tool used to create contingency tables.



Grade	Charged off	Current	Fully paid	Late	Total
A	1562	50051	20408	469	72490
	0.022	0.690	0.282	0.006	0.161
B	5302	93852	31160	2056	132370
	0.040	0.709	0.235	0.016	0.294
C	6023	88928	23147	2777	120875
	0.050	0.736	0.191	0.023	0.268
D	5007	53281	13681	2308	74277
	0.067	0.717	0.184	0.031	0.165
E	2842	24639	5949	1374	34804
	0.082	0.708	0.171	0.039	0.077
F	1526	8444	2328	606	12904
	0.118	0.654	0.180	0.047	0.029
G	409	1990	643	199	3241
	0.126	0.614	0.198	0.061	0.007
Total	22671	321185	97316	9789	450961

Categorical and Numeric Data

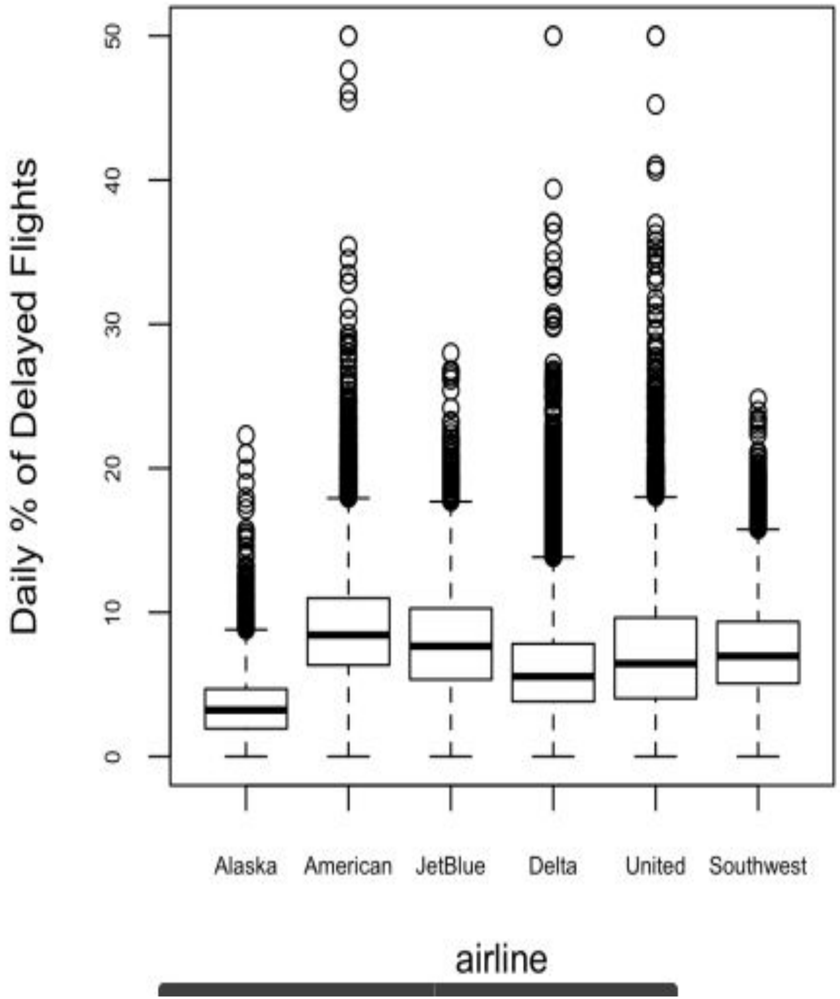
Box Plot are a simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.

A violin plot, introduced by [Hintze-Nelson-1998], 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.

Box plot

Visually compare the distributions of numeric variable grouped according to a categorical variable.



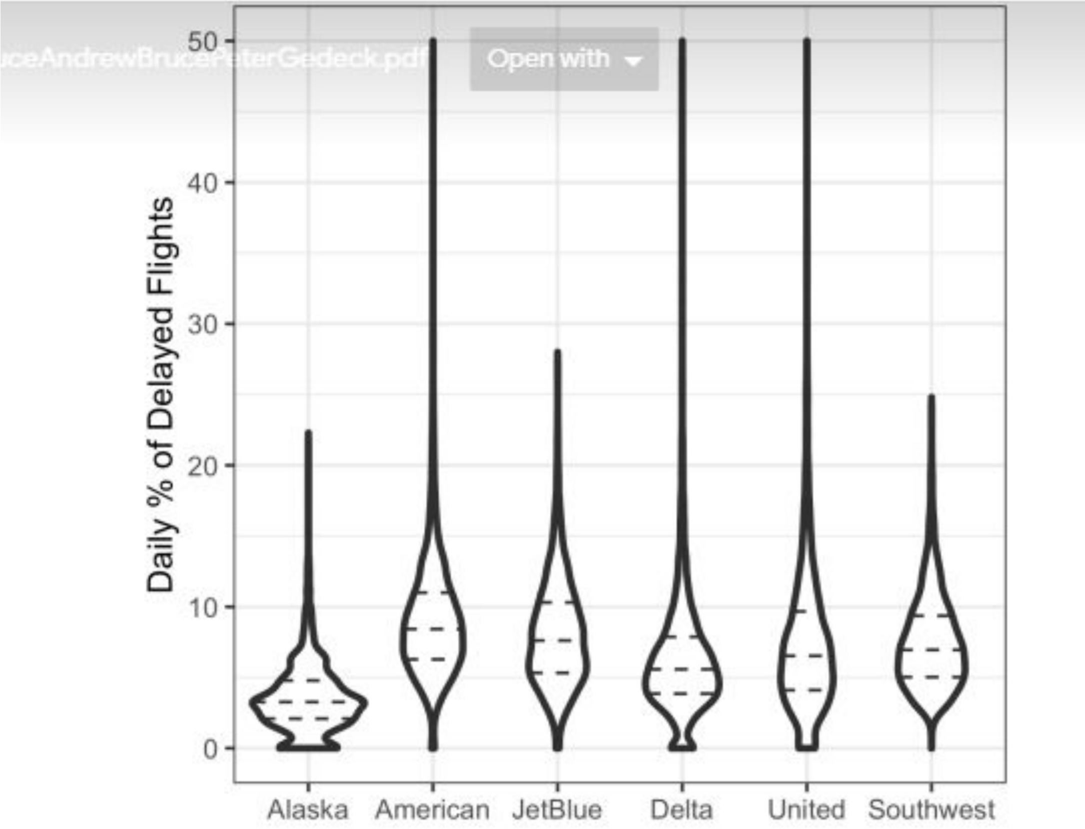


Figure 1-11. Violin plot of percent of airline delays by carrier

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.

Summary

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.