EDA

- In 1962, John W. Tukey (Figure 1-1) called for a reformation of statistics in his seminal paper “The Future of Data Analysis” [Tukey-1962].
- He proposed a new scientific discipline called *data analysis* that included statistical inference as just one component.
- Tukey forged links to the engineering and computer science communities and his original tenets are surprisingly durable and form part of the foundation for data science.
- The field of exploratory data analysis was established with Tukey’s 1977 now-classic book Exploratory Data Analysis.

*This presentation is based on the book ‘Practical Statistics for Data Scientist’ by Peter Bruce & Andrew Bruce

Types of Data

- In 1946 paper*, Psychologist Stanley Smith Stevens developed the best-known classification with four levels, or scales, of measurement: ***nominal, ordinal, interval, and ratio***.
- In 1977 Mosteller and Tukey came with seven levels:
  - Names
  - Grades (ordered labels like beginner, intermediate, advanced)
  - Ranks (orders with 1 being the smallest or largest, 2 the next smallest or largest, and so on)
  - Counted fractions (bound by 0 and 1)
  - Counts (non-negative integers)
  - Amounts (non-negative real numbers)
  - Balances (any real number)
- Another seven types are explained in the top-red article of Jeff Hale*:
  - Useless
  - Nominal
  - Binary
  - Ordinal
  - Count
  - Time
  - Interval

* Articles with asterix on can be found on drive

KEY TERMS FOR DATA TYPES

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, polychotomous

Binary

A special case of categorical data with just two categories of values (0/1, true/false).

Synonyms

dichotomous, logical, indicator, boolean

Ordinal

Categorical data that has an explicit ordering.

Synonyms

ordered factor

Rectengular 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 essentially a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables).

- Data Frames and Indexes: In Python, with the pandas library, the basic rectangular data structure is a DataFrame object. By default, an automatic integer index is created for a DataFrame based on the order of the rows.*
  - Be careful when sorting: index may move with the rows

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 in the 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 study.

Synonyms

dependent variable, response, target, output

Records

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

Synonyms

case, example, instance, observation, pattern, sample

Nonrectangular Data

- ***Time series*** data records successive measurements of the same variable. It is the raw material for statistical forecasting methods, and it is also a key component of the data produced by devices — the Internet of Things.

- ***Spatial data*** structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures. In the object representation, the focus of the data is an object (e.g., ahouse) and its spatial coordinates. The field view, by contrast, focuses on small units of space and the value of a relevant metric (pixel brightness, for example).

- ***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, may represent connections between people on the network. Distribution hubs connected by roads are an example of a physical network. Graph structures are useful for certain types of problems, such as network optimization and recommender systems.

- The basic data structure in data science is a rectangular matrix in which rows are records and columns are variables (features).

Estimates of Location

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

- At first glance, summarizing data might seem fairly trivial: just take the mean of the data (see “Mean”). In fact, while the mean is easy to compute and expedient to use, it may not always be the best measure for a central value. For this reason, statisticians have developed and promoted several alternative estimates to the mean.

- ***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

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

- 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.
  - For example, in international diving the top and bottom scores from five judges are dropped, and the final score is the average of the three remaining judges . This makes it difficult for a single judge to manipulate the score, perhaps to favor his country's contestant. Trimmed means are widely used, and in many cases, are preferable to use instead of the ordinary mean:

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

- Another type of mean is a ***weighted mean***, which you calculate by multiplying each data value by a weight and dividing their sum by the sum of the weights.
  - 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.

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

Median and Robust Estimates

- The ***median*** is the middle number on a sorted list of the data.

- Compared to the mean, which uses all observations, the median depends only on the values in the center of the sorted data. While this might seem to be a disadvantage, since the mean is much more sensitive to the data
  - Let's say we want to look at typical household incomes in neighborhoods around Lake Washington in Seattle. In comparing the Medina neighborhood to the Windermere neighborhood, using the mean would produce very different results because Bill Gates lives in Medina. If we use the median, it won't matter how rich Bill Gates is — the position of the middle observation will remain the same.

- The median is robust to outliers (extreme cases)

Outliers

- The median is referred to as a robust estimate of location since it is 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.

- The exact definition of an outlier is somewhat subjective, but a rule of thumb is 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3).

- Being an outlier in itself does not make a data value invalid or erroneous (as in the previous example with Bill Gates).

- *ANOMALY DETECTION: In contrast to typical data analysis, where outliers are sometimes informative and sometimes a nuisance, in anomaly detection the points of interest are the outliers, and the greater mass of data serves primarily to define the "normal" against which anomalies are measured.*

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.

- ***At the heart of statistics lies variability:*** measuring it, reducing it, distinguishing random from real variability, identifying the various sources of real variability, and making decisions in the presence of it.

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.

Synonyms

mean-squared-error

Standard deviation

The square root of the variance.

Synonyms

l2-norm, Euclidean norm

Mean absolute deviation

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

Synonyms

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.

Synonyms

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.

Synonyms

quantile

Interquartile range

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

Synonyms

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.
- 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 variance and standard deviation are especially sensitive to outliers since they are based on the squared deviations.

Estimates Based on Percentiles

- 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

- In a data set, the Pth ***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.
  - Median is the 50^{th} percentile, ***Q1*** is 25^{th} and ***Q3*** is 75^{th}

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

Exploring the Data Distribution

KEY TERMS FOR EXPLORING THE DISTRIBUTION

Boxplot

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

Synonyms

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.

Density plot

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

Percentiles and Boxplots

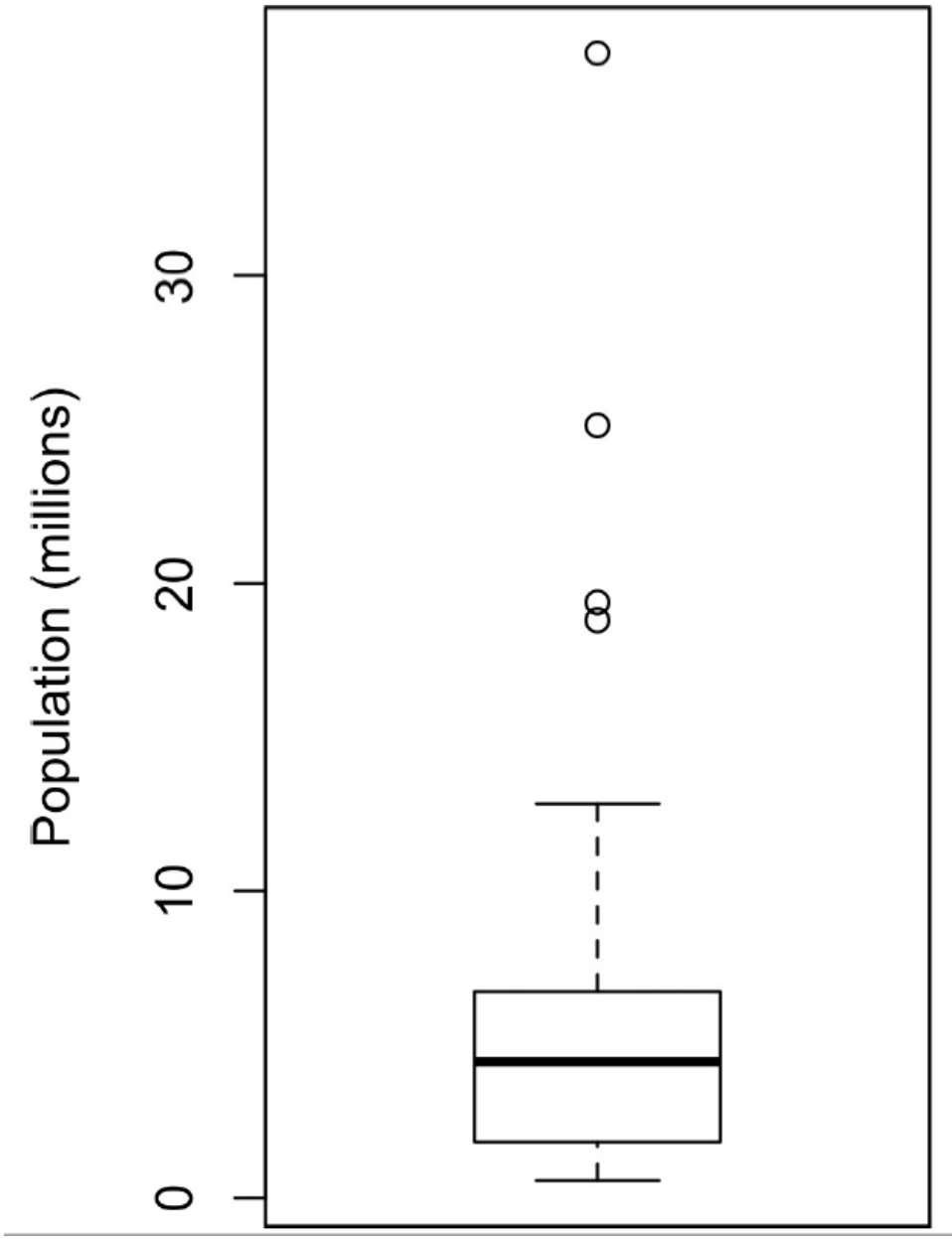
- Percentiles are also valuable to summarize the entire distribution. It is common to report the quartiles (25th, 50th, and 75th percentiles) and the deciles (the 10th, 20th, …, 90th percentiles).

Table 1-4. Percentiles of murder rate by state

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

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

- 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.



Frequency Table

- 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.
- 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. It they are too small, the result is too granular and the ability to see bigger pictures is lost.

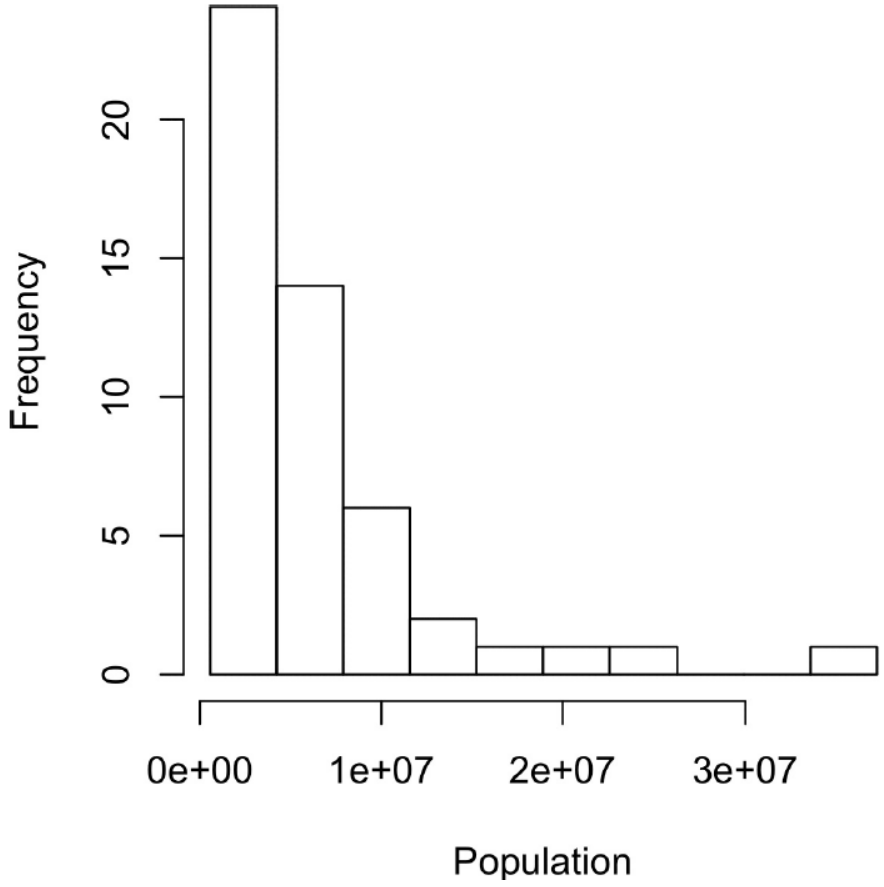
Table 1-5. A frequency table of population by state

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,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

Histograms

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

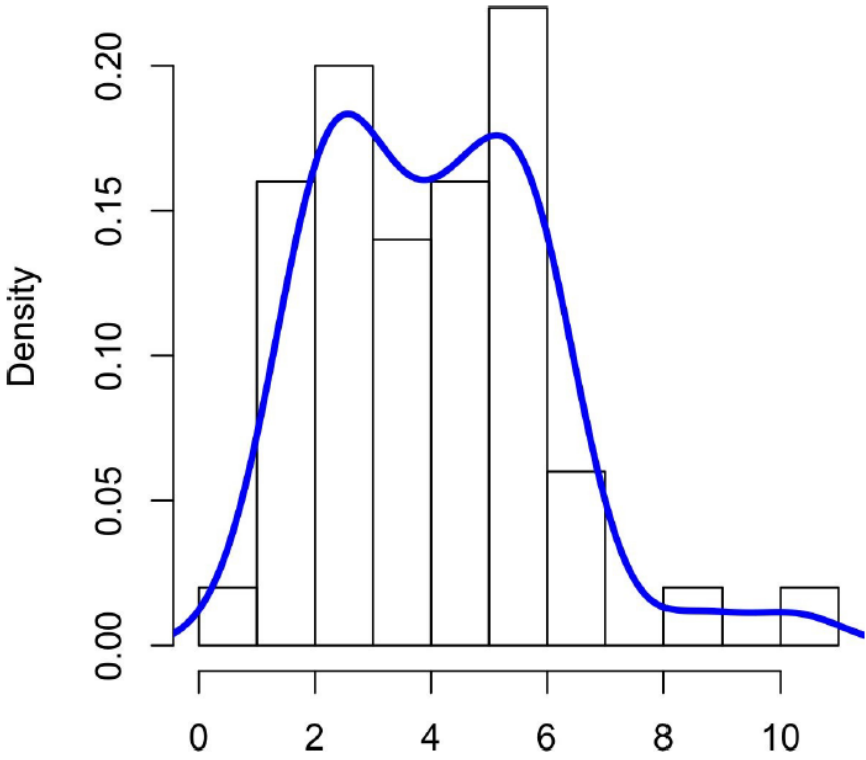
- In general, histograms are plotted such that:
  - Empty bins are included in the graph.
  - Bins are equal width.
  - Number of bins (or, equivalently, bin size) is up to the user.
  - Bars are contiguous — no empty space shows between bars, unless there is an empty bin.



- ***STATISTICAL MOMENTS***: 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 displays

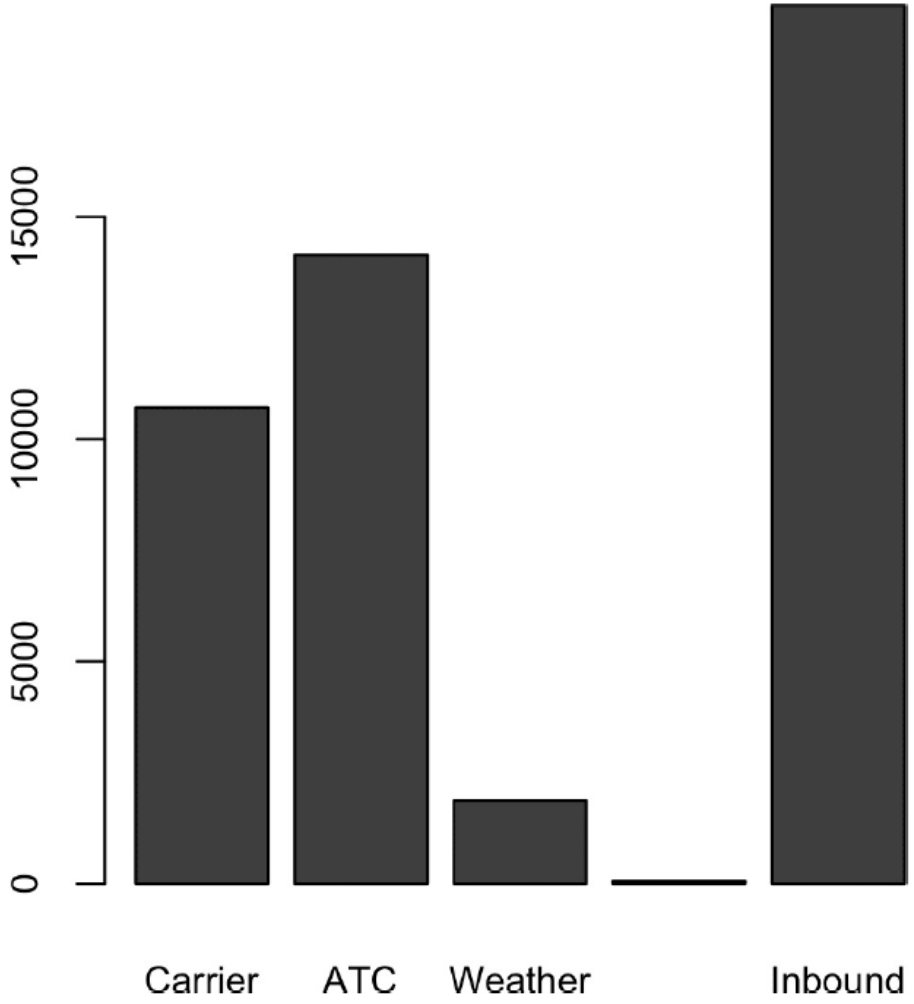
Density Estimates

- 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 (see [Duong-2001] for a short tutorial*).
- A key distinction from the histogram is the scale of the y-axis: a density plot corresponds to plotting the histogram as a proportion rather than counts.



Exploring Binary and Categorical Data

- ***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.
- Note that a bar chart resembles a histogram;
  - 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 and Expected Value

- The ***mode*** is the value — or values in case of a tie — that appears most often in the data.

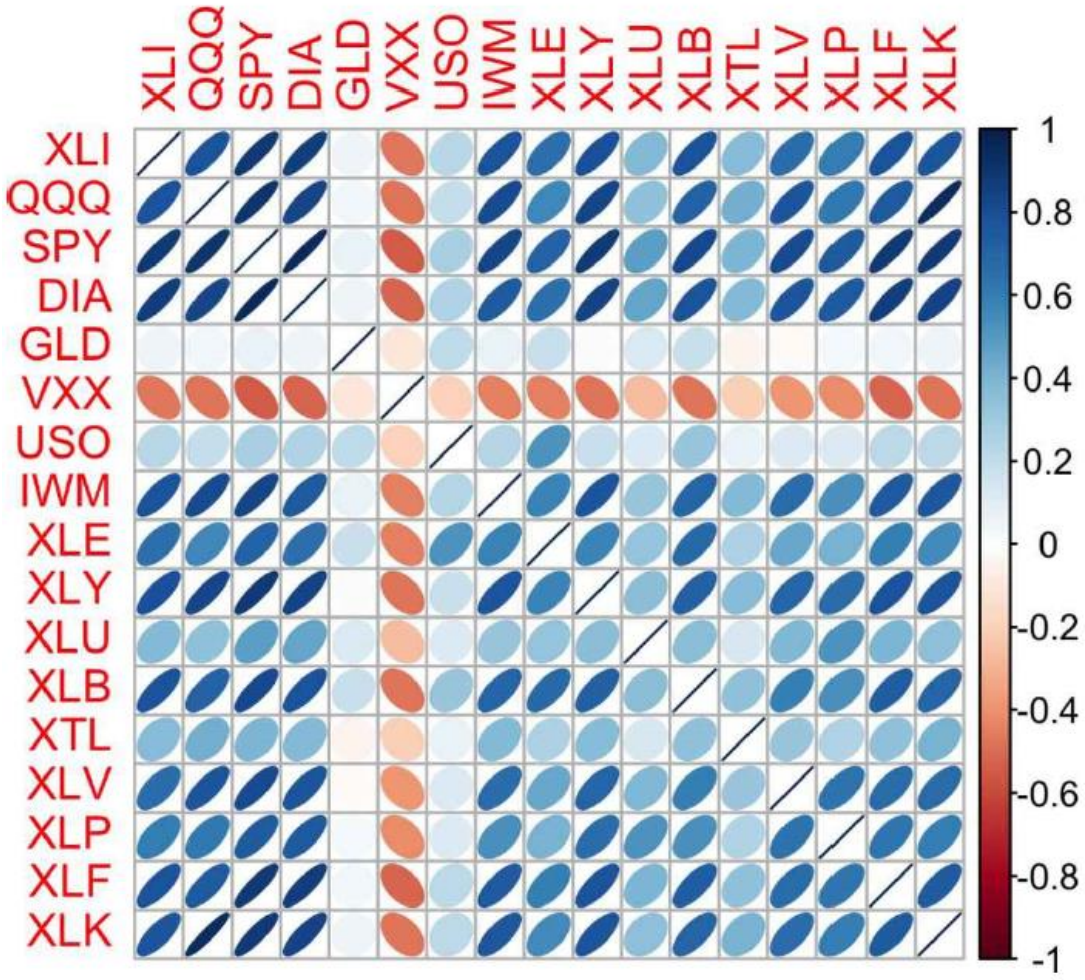
- *A special type of categorical data* is data in which the categories represent or can be mapped to discrete values on the same scale.
- This data can be summed up, in a single ***expected value*** which is a form of weighted mean in which the weights are probabilities.
  - Multiply each outcome by its probability of occurring.
  - Sum these values.

Correlation

- Variables X and Y are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y. If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.
- Variables can have an association that is not linear, in which case the correlation coefficient may not be a useful metric.

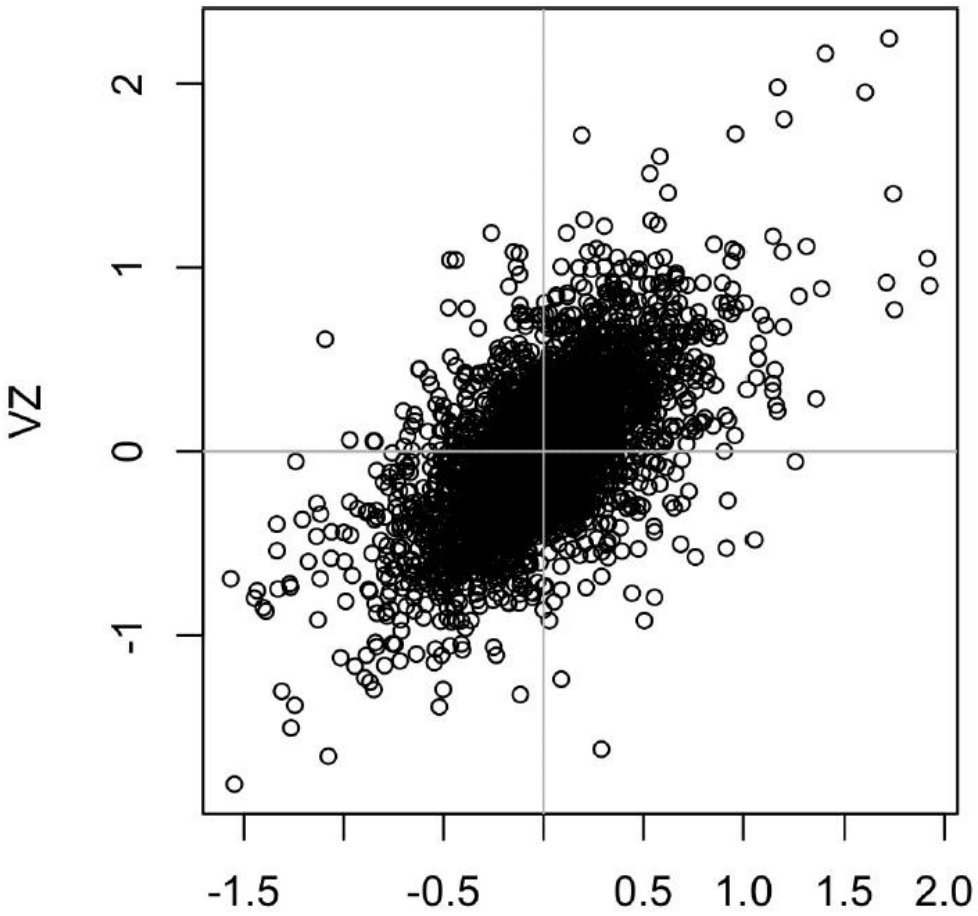
	T	CTL	FTR	VZ	LVLT
T	1.000	0.475	0.328	0.678	0.279
CTL	0.475	1.000	0.420	0.417	0.287
FTR	0.328	0.420	1.000	0.287	0.260
VZ	0.678	0.417	0.287	1.000	0.242
LVLT	0.279	0.287	0.260	0.242	1.000

- The orientation of the ellipse indicates whether two variables are positively correlated or negatively correlated.
- The shading and width of the ellipse indicate the strength of the association: thinner and darker ellipses correspond to stronger relationships.
- Like the mean and standard deviation, the correlation coefficient is sensitive to outliers in the data.



Scatterplots

- The standard way to visualize the relationship between two measured data variables is with a scatterplot.



KEY IDEAS FOR CORRELATION

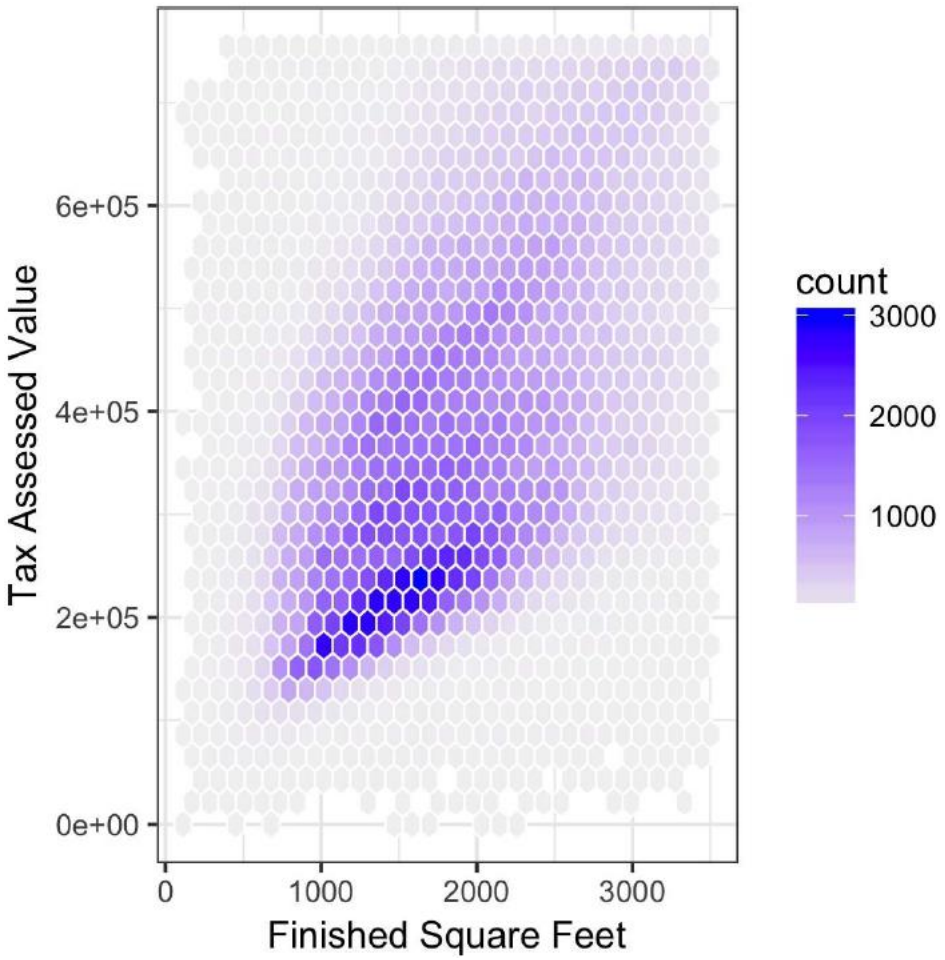
- The correlation coefficient measures the extent to which two variables are associated with one another.
- When high values of v1 go with high values of v2, v1 and v2 are positively associated.
- When high values of v1 are associated with low values of v2, v1 and v2 are negatively associated.
- The correlation coefficient is a standardized metric so that it always ranges from –1 (perfect negative correlation) to +1 (perfect positive correlation).
- A correlation coefficient of 0 indicates no correlation, but be aware that random arrangements of data will produce both positive and negative values for the correlation coefficient just by chance.
- For correlation between binary variables, see Phi coefficient by Karl Pearson.
- For correlation between categorical variables, see Cramer's V by Herald Cramer. (based on a variation of Pearson's Chi-Square Test)

Exploring Two or More Variables

- Familiar estimators like mean and variance look at variables one at a time (univariate analysis).
- Correlation analysis (see “Correlation”) 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).
- The appropriate type of bivariate or multivariate analysis depends on the nature of the data: numeric versus categorical.

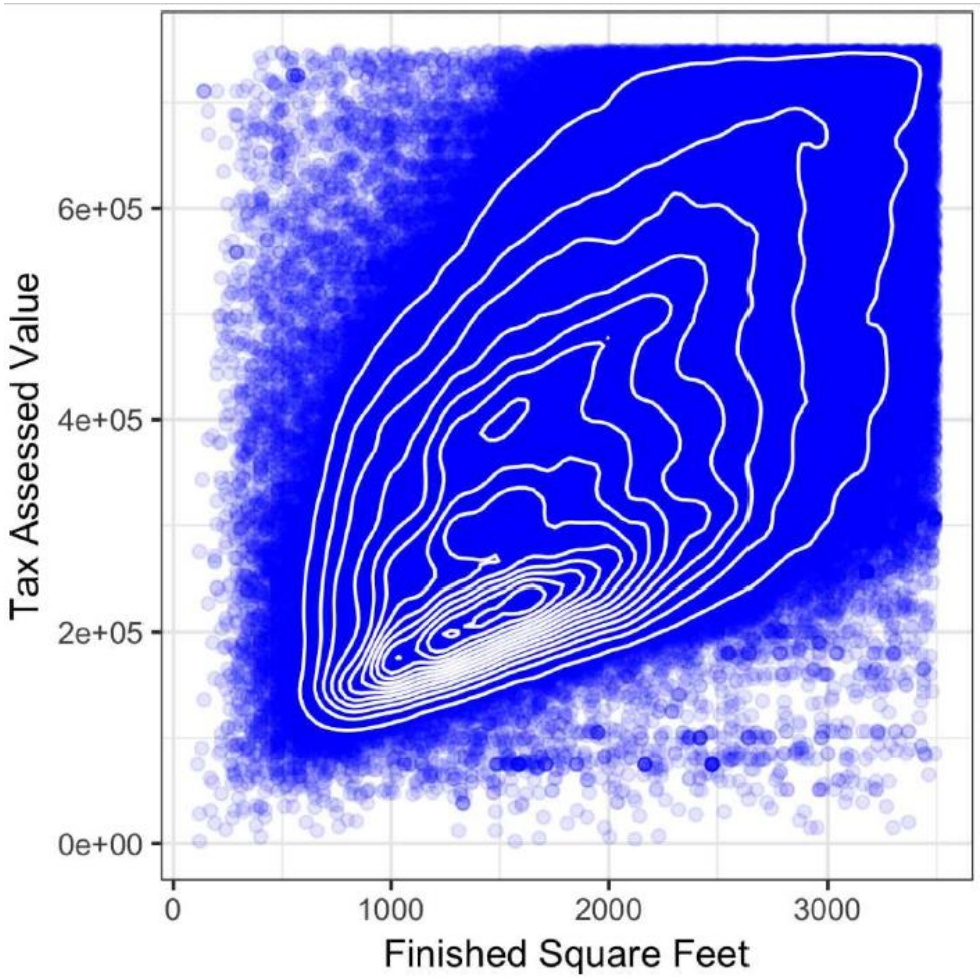
Hexagonal Binning and Contours (Plotting Numeric versus Numeric Data)

- Scatterplots are fine when there is a relatively small number of data values.
- For data sets with hundreds of thousands or millions of records, a scatterplot will be too dense.
- 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.
  - 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.



Contour Plot

- 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 previous, there is a secondary peak “north” of the main peak.



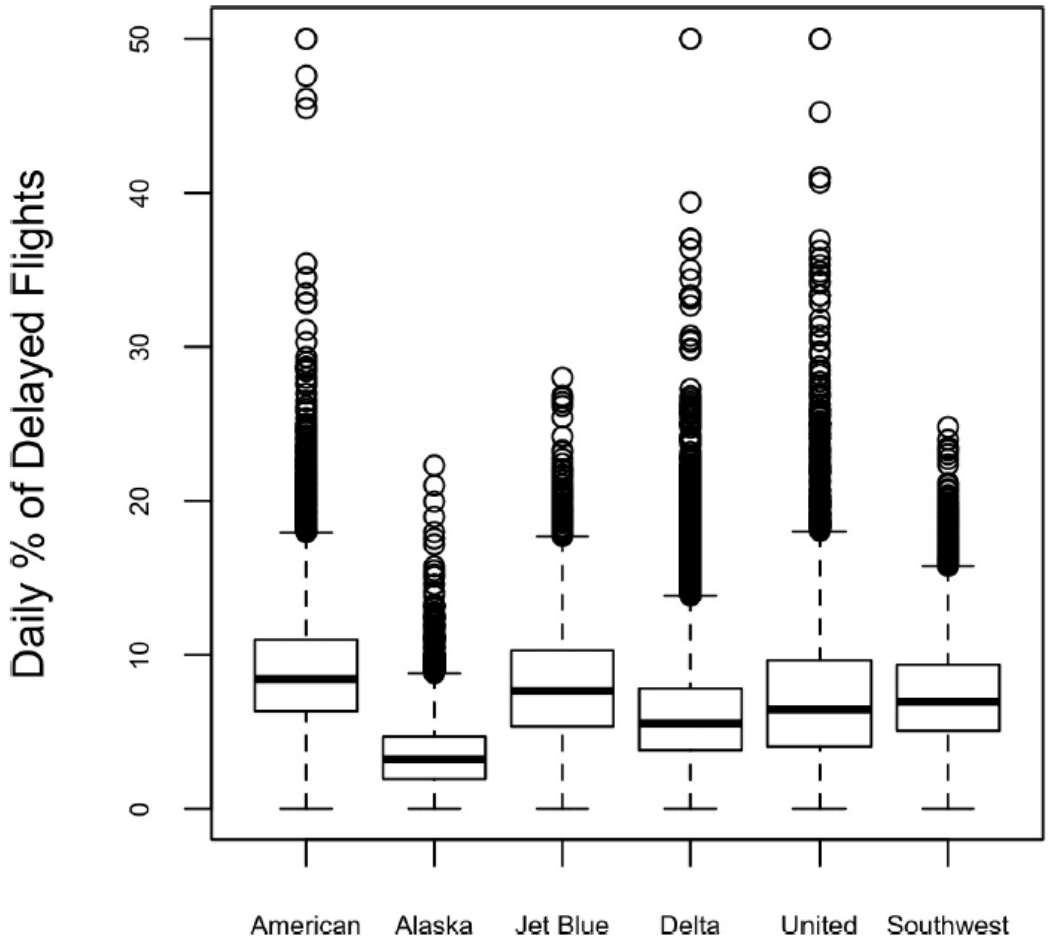
Two Categorical Variables

- A useful way to summarize two categorical variables is a contingency table — a table of counts by category.
- Pivot tables in Excel are perhaps the most common tool used to create contingency tables.

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

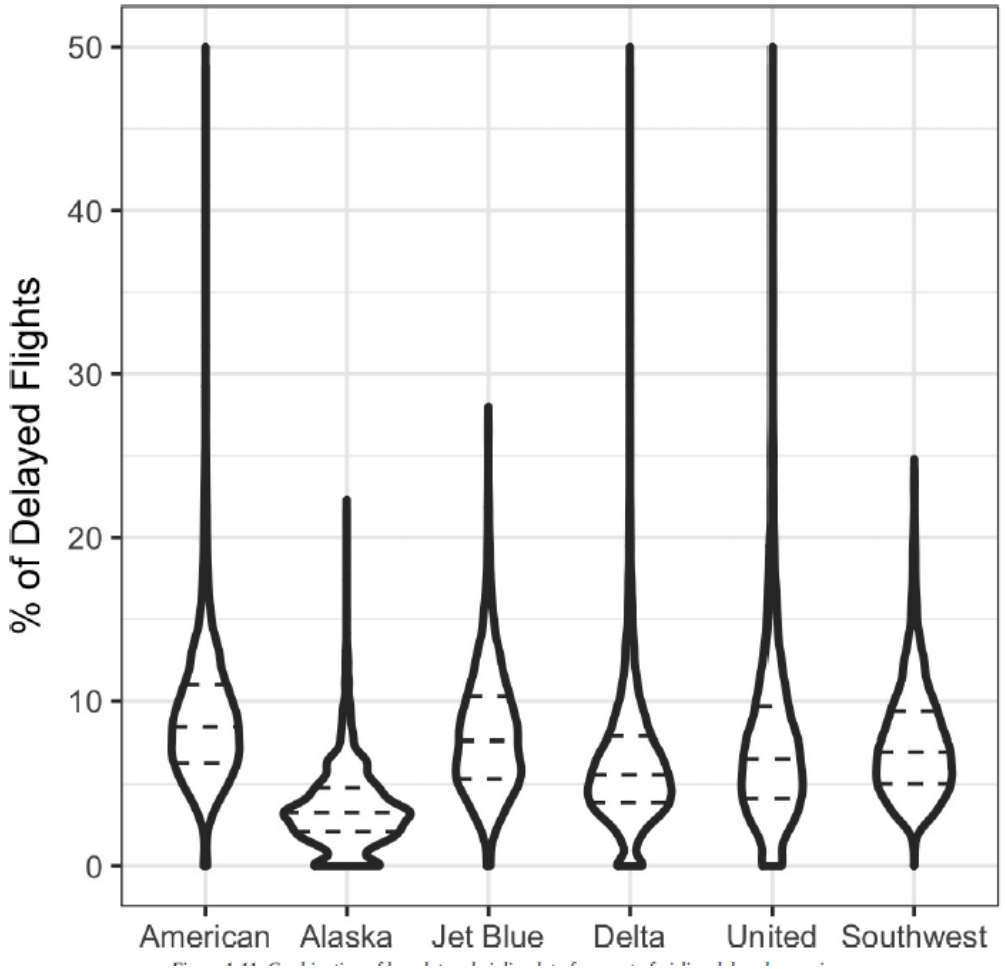
Categorical and Numeric Data

- Boxplots (see “Percentiles and Boxplots”) are a simple way to visually compare the distributions of a numeric variable grouped according to a categorical variable.



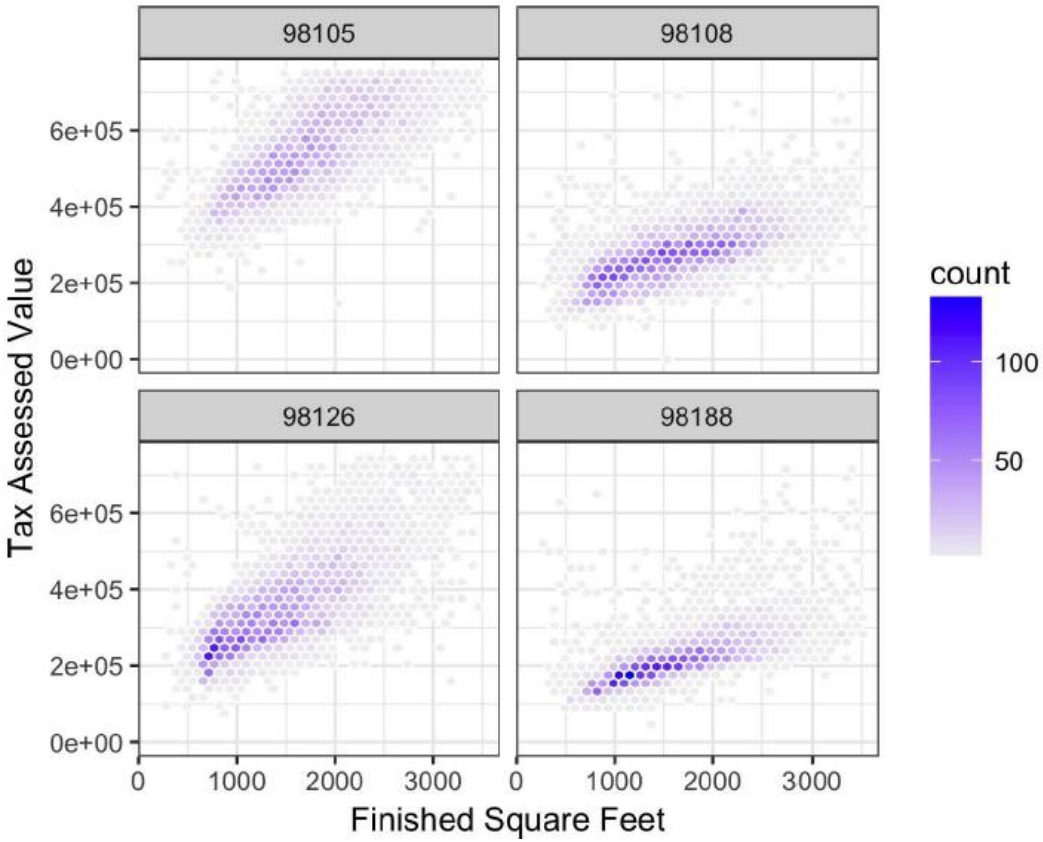
Categorical and Numeric Data

- A violin plot, is an enhancement to the boxplot and plots the density estimate with the density on the y-axis.
- 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.



Visualizing Multiple Variables

- The types of charts used to compare two variables are readily extended to more variables through the notion of conditioning.
- As an example, look back at page 28, 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, plot on the right accounts for the effect of location by plotting the data for a set of zip codes.



- This presentation is based on the book ‘Practical Statistics for Data Scientist’