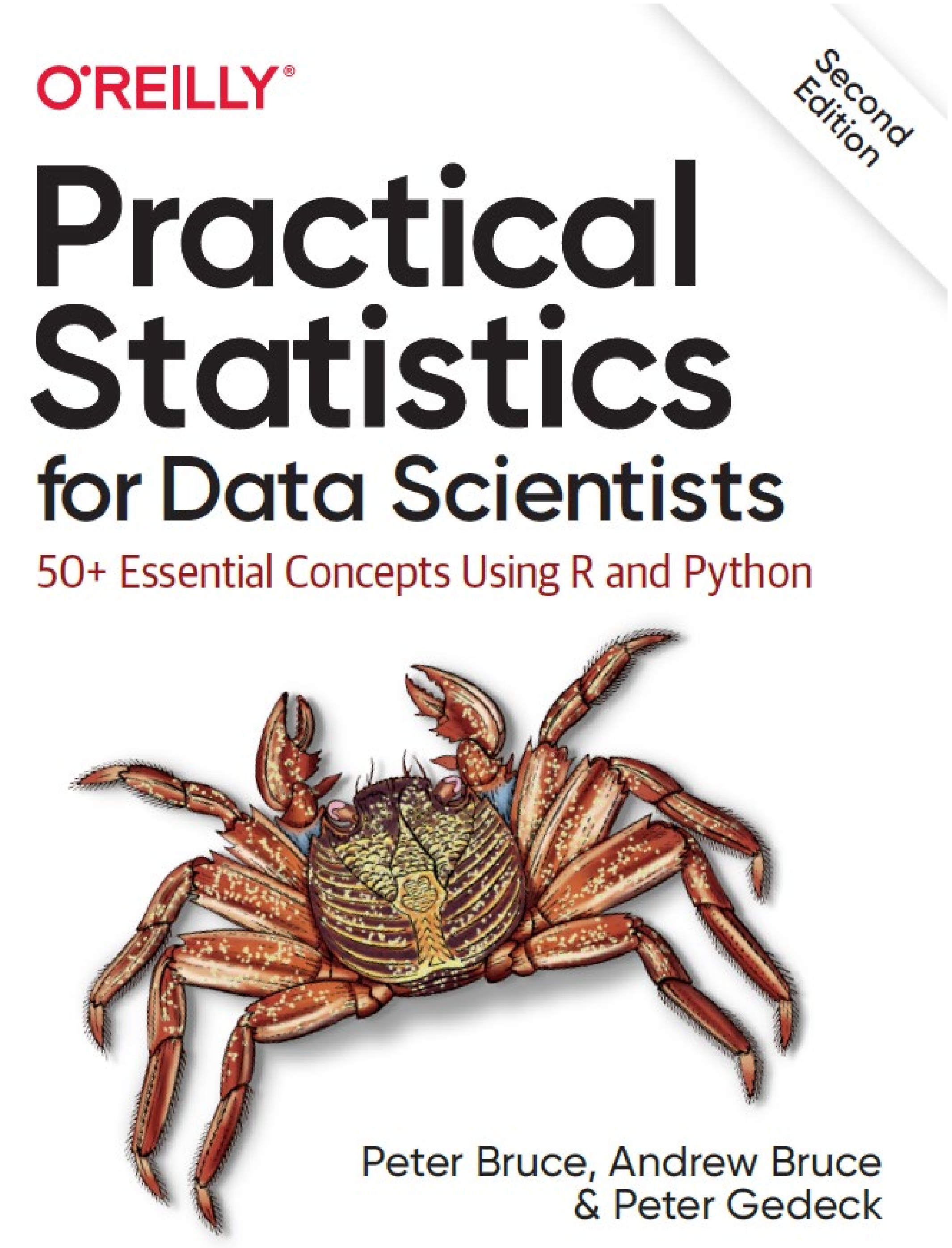
PRACTICAL STATISTICS FOR DATA SCIENTISTS

Reza Barahmand



O'REILLY®

Second Edition

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python

Peter Bruce, Andrew Bruce & Peter Gedeck

Exploratory Data Analysis

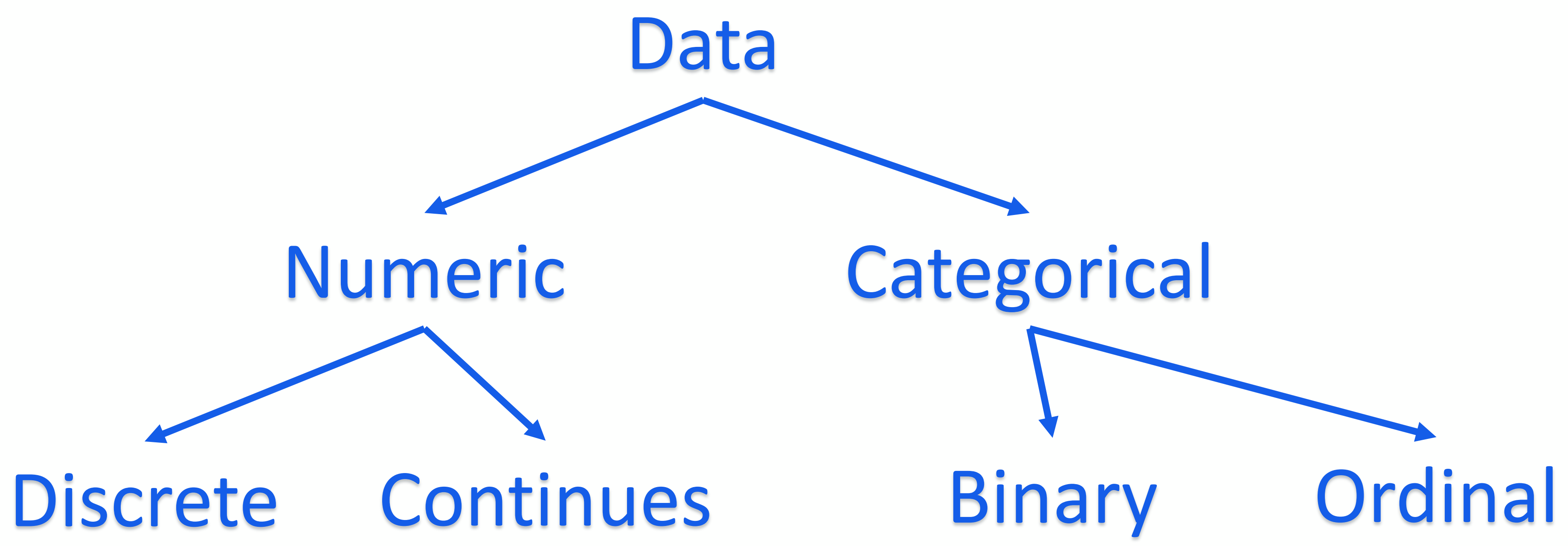
The first and most important step in any project based on data is to look at the data. By summarizing and visualizing the data, you can gain valuable intuition and understanding of the project.

Data Data Data !!!

Large Amount of Data

- Internet of Things
- Images
- Texts
- Clickstreams
- …

Data Types



Categorical Data not Text !

- Act as a signal telling software how statistical procedures should behave.
- Storage and indexing can be optimized.
- The possible values a given categorical variable can take are enforced in the software.

Major Challenge

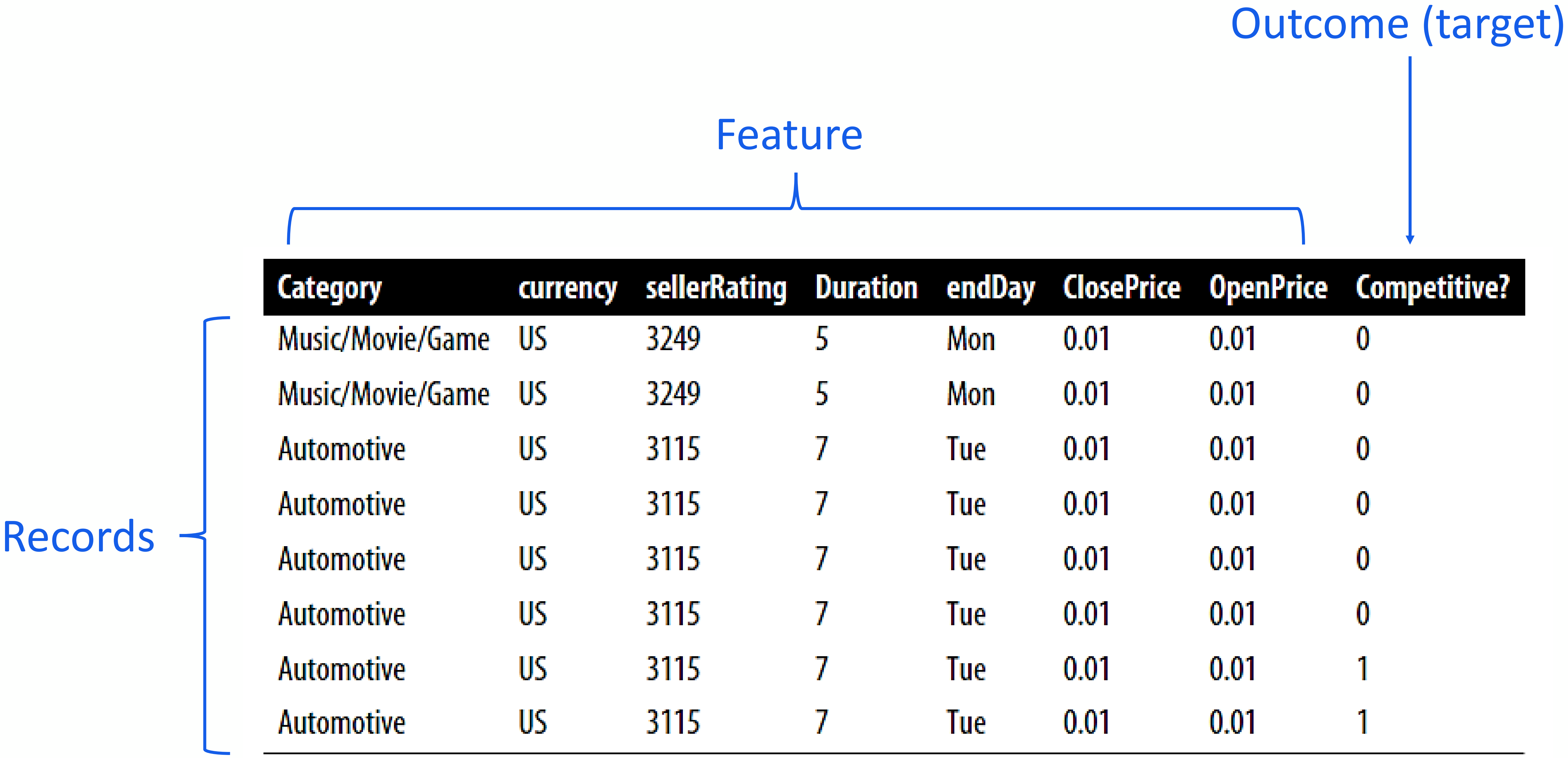
A major challenge of data science is to harness this torrent of raw data into actionable information.

The data doesn’t always start in this form: unstructured data (e.g., text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data.

Data Structures

Rectangular Data

- Two-dimensional matrix.
- Data frame in python & R
- Features
- Outcome
- Records



Outcome (target)

Feature

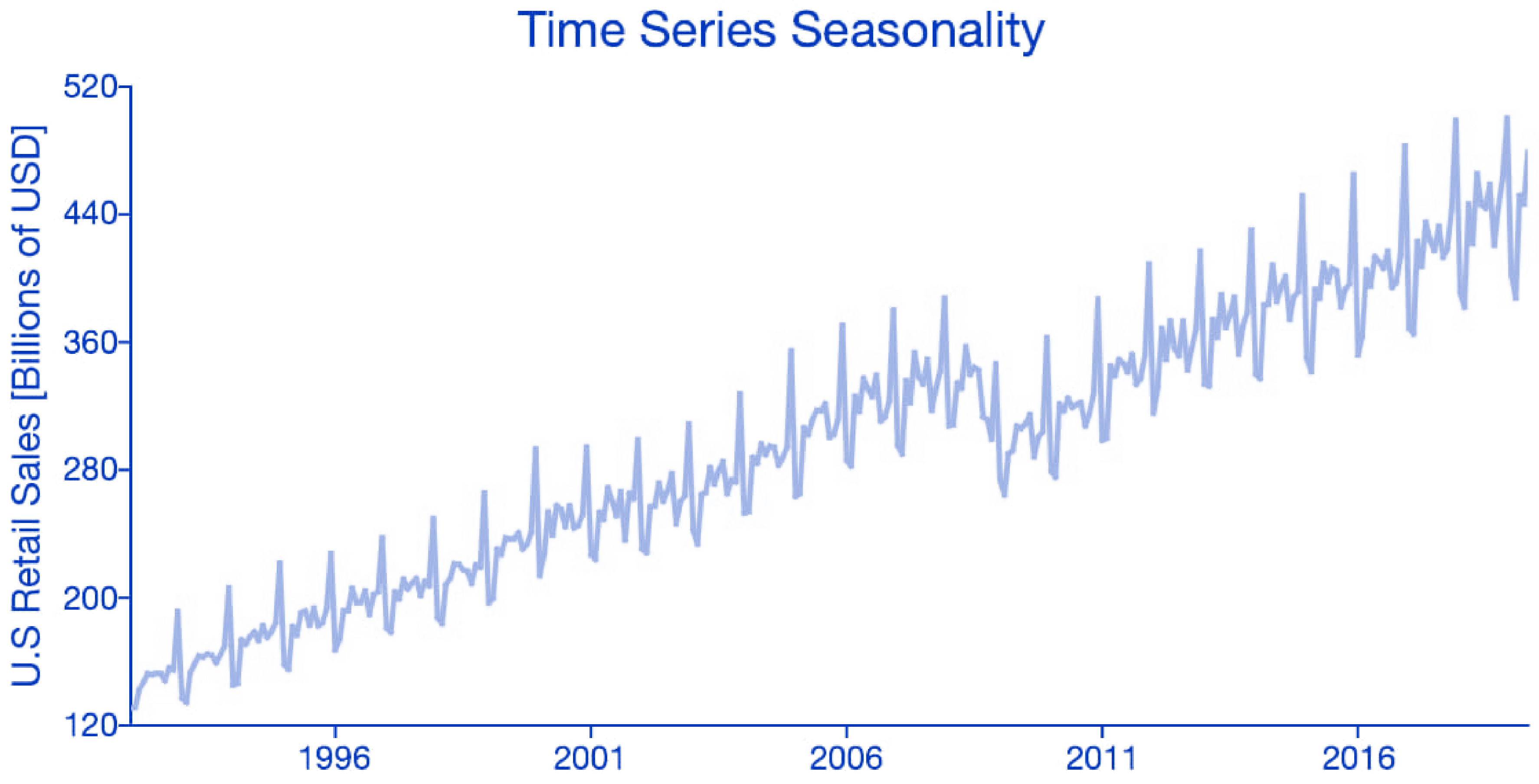
Records

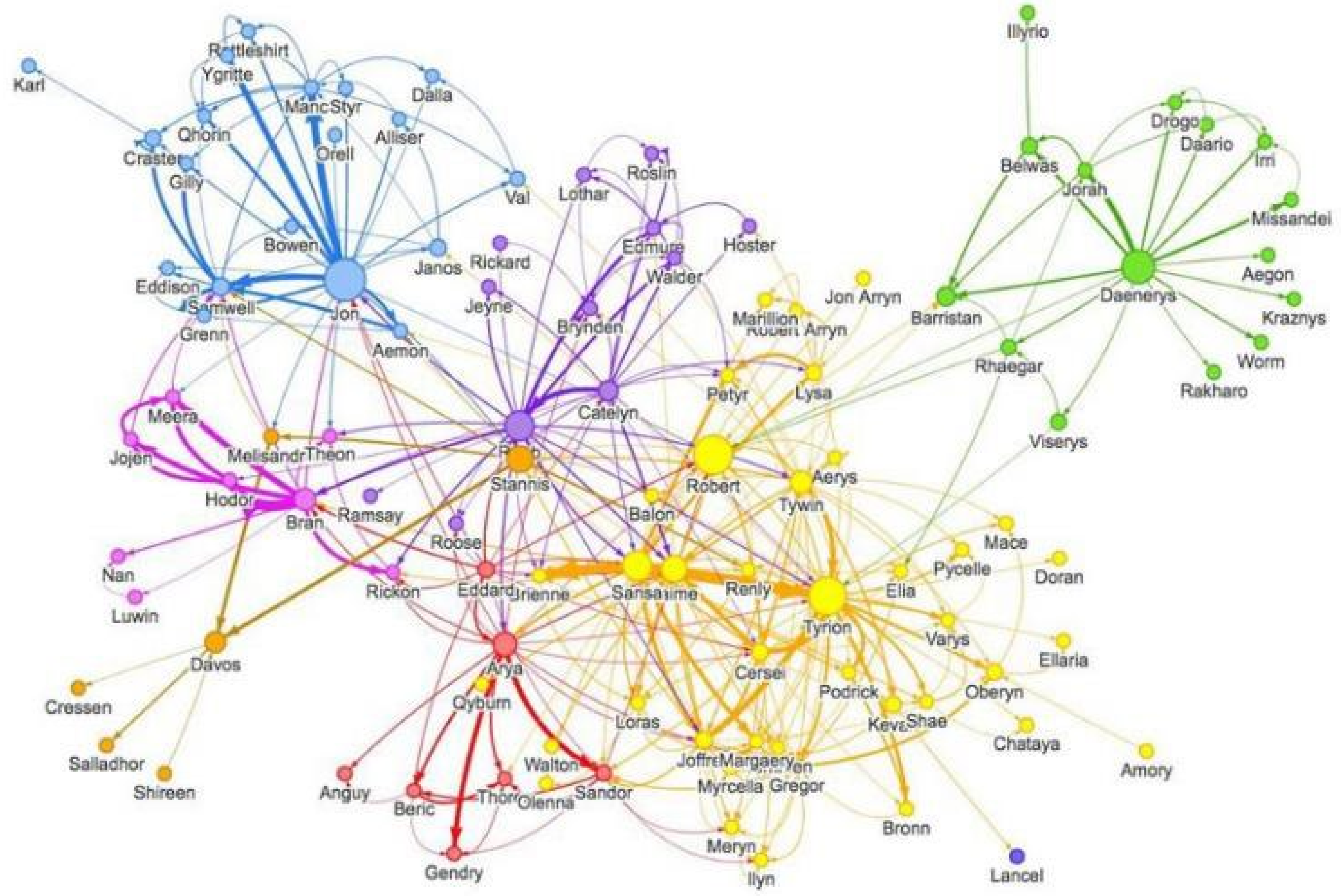
Category	currency	sellerRating	Duration	endDay	ClosePrice	OpenPrice	Competitive?
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Music/Movie/Game	US	3249	5	Mon	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	0
Automotive	US	3115	7	Tue	0.01	0.01	1
Automotive	US	3115	7	Tue	0.01	0.01	1

Data Structures

Nonrectangular Data

- Time series data
- Mapping and location analytics
- Graph (or network) data structures
    - Network optimization
    - Recommender systems





Estimations

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).

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.

Estimation of Location

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

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

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

- A trimmed mean eliminates the influence of extreme values.
- rimming the bottom and top 10% (a common choice)
- It is robust to extreme values in the data, but uses more data to calculate the estimate for location.

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

- Some values are intrinsically more variable than others, and highly variable observations are given a lower weight.
- The data collected does not equally represent the different groups that we are interested in measuring. (unbalanced class)

Estimation of Location

1, 3, 3, **6**, 7, 8, 9

$$\text{Median} = \underline{\underline{6}}$$

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

$$\text{Median} = (4 + 5) \div 2$$

$$= \underline{\underline{4.5}}$$

Median (50th percentile)

- 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.
- Robust to outliers

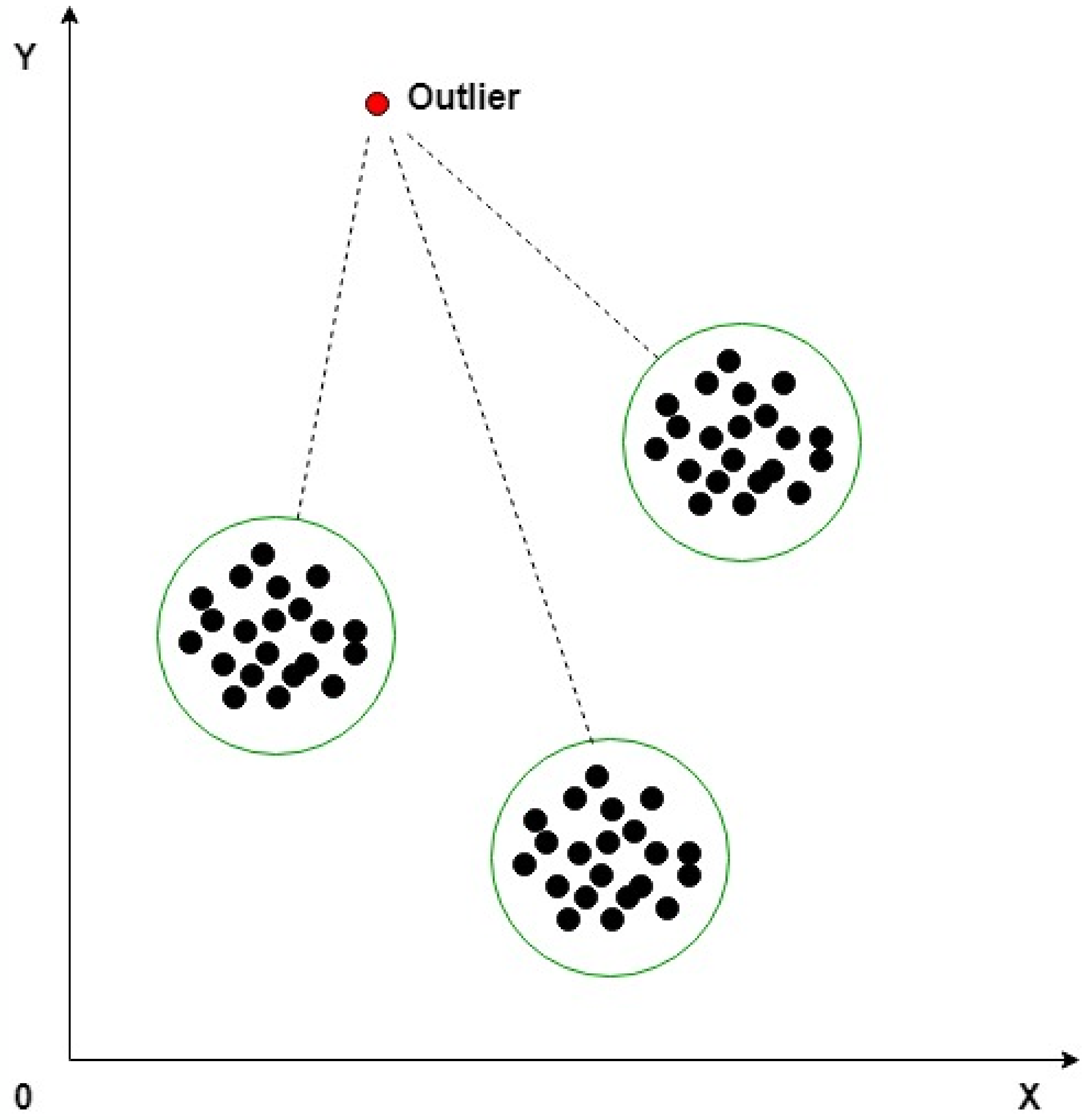
Weighted Median

- First sort the data.
- Each data value has an associated weight.
- The weighted median is a value such that the sum of the weights is equal for the lower and upper halves of the sorted list.
- Robust to outliers

Estimation of Location

Outliers

- An outlier is any value that is very distant from the other values in a data set.
- Being an outlier in itself does not make a data value invalid or erroneous (as in the previous example with Bill Gates).
- Outliers are often the result of data errors such as mixing data of different units (kilometers versus meters) or bad readings from a sensor.



Anomaly Detection

- Kind of data science project that the points of interest are the outliers, and the greater mass of data serves primarily to define the “normal” against which anomalies are measured.

Estimation of Variability

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

- Absolute deviation help us not to offset positive deviations by negative ones.
- Sensitive to outliers.

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

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

- The standard deviation is much easier to interpret than the variance since it is on the same scale as the original data.
- Standard deviation and Variance are preferred in statistics over the mean absolute deviation.
- Both are Sensitive to outliers.

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

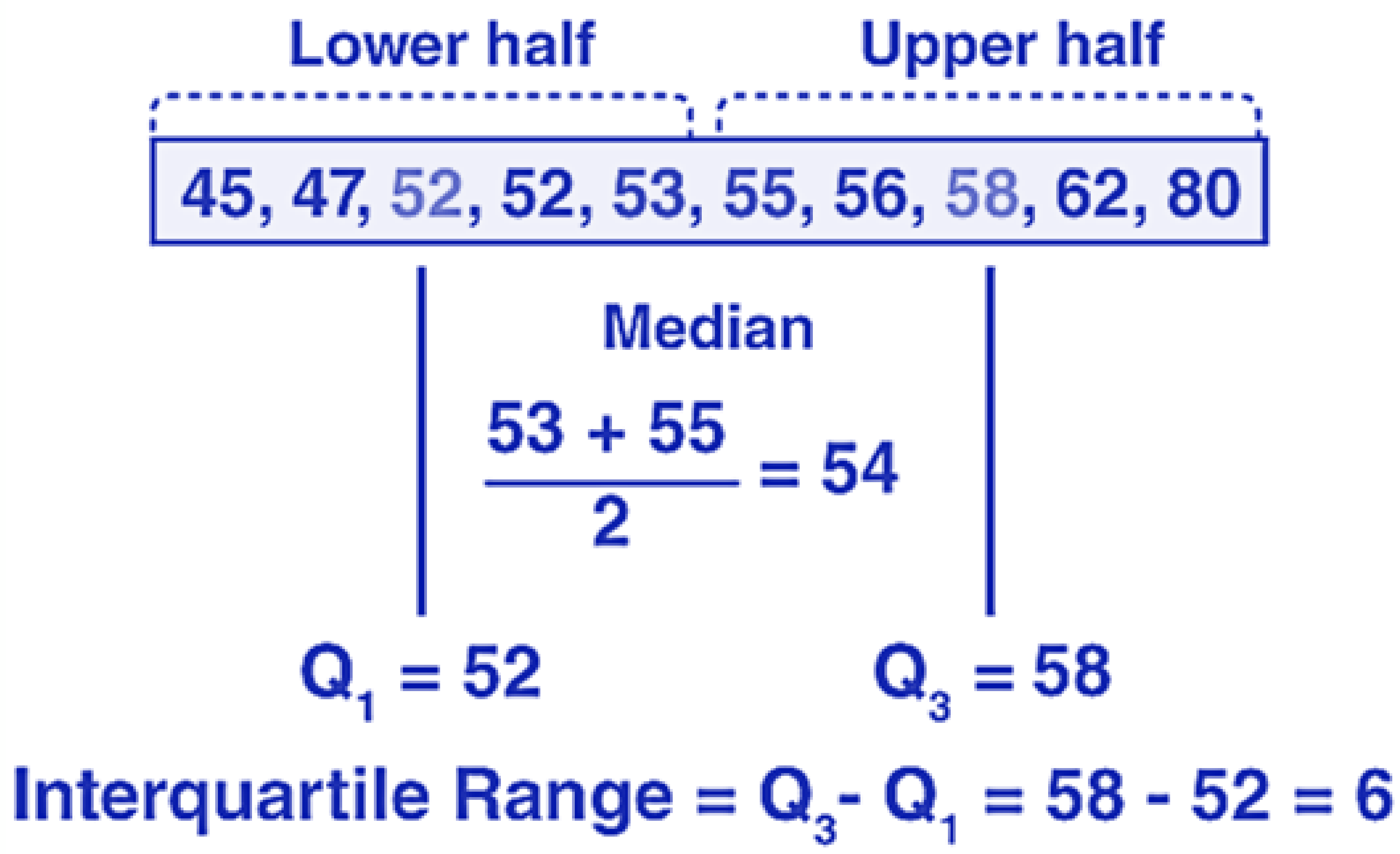
- Robust estimate of variability.
- multiplied by a constant scaling factor to put the MAD on the same scale as the standard deviation in the case of a normal distribution (commonly used factor of 1.4826)

It is also possible to compute a trimmed standard deviation analogous to the trimmed mean

Estimation of Variability

Estimates Based on Percentiles

- Statistics based on sorted (ranked) data are referred to as order statistics.
- Range
  - Sensitive to outliers
  - look at the range of the data after dropping values from each end.
- Min / Max
  - To identify outliers
- Percentile or Quantile
- Interquartile range (IQR)
  - Different between 25th percentile and 75th percentile
  - For large data sets it is computationally expensive → Zhang-Wang



Lower half

Upper half

45, 47, 52, 52, 53, 55, 56, 58, 62, 80

Median

$$\frac{53 + 55}{2} = 54$$

$Q_1 = 52$

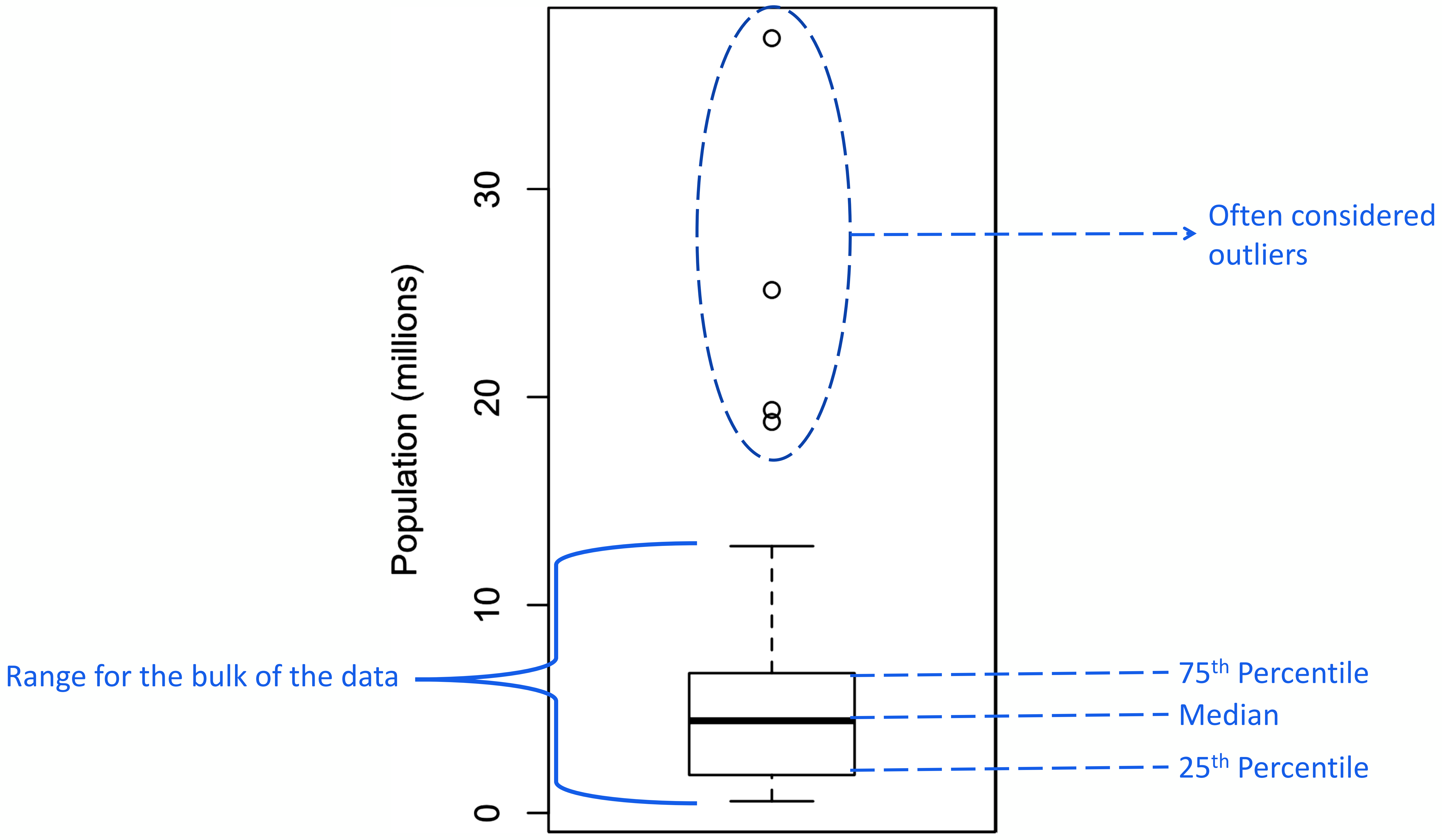
$Q_3 = 58$

Interquartile Range = $Q_3 - Q_1 = 58 - 52 = 6$

Exploring the Data Distribution

Boxplots

- Based on Percentiles
- Visualize The Distribution of Data
- Whiskers
- Median
- IQR
- Visual detection of outliers

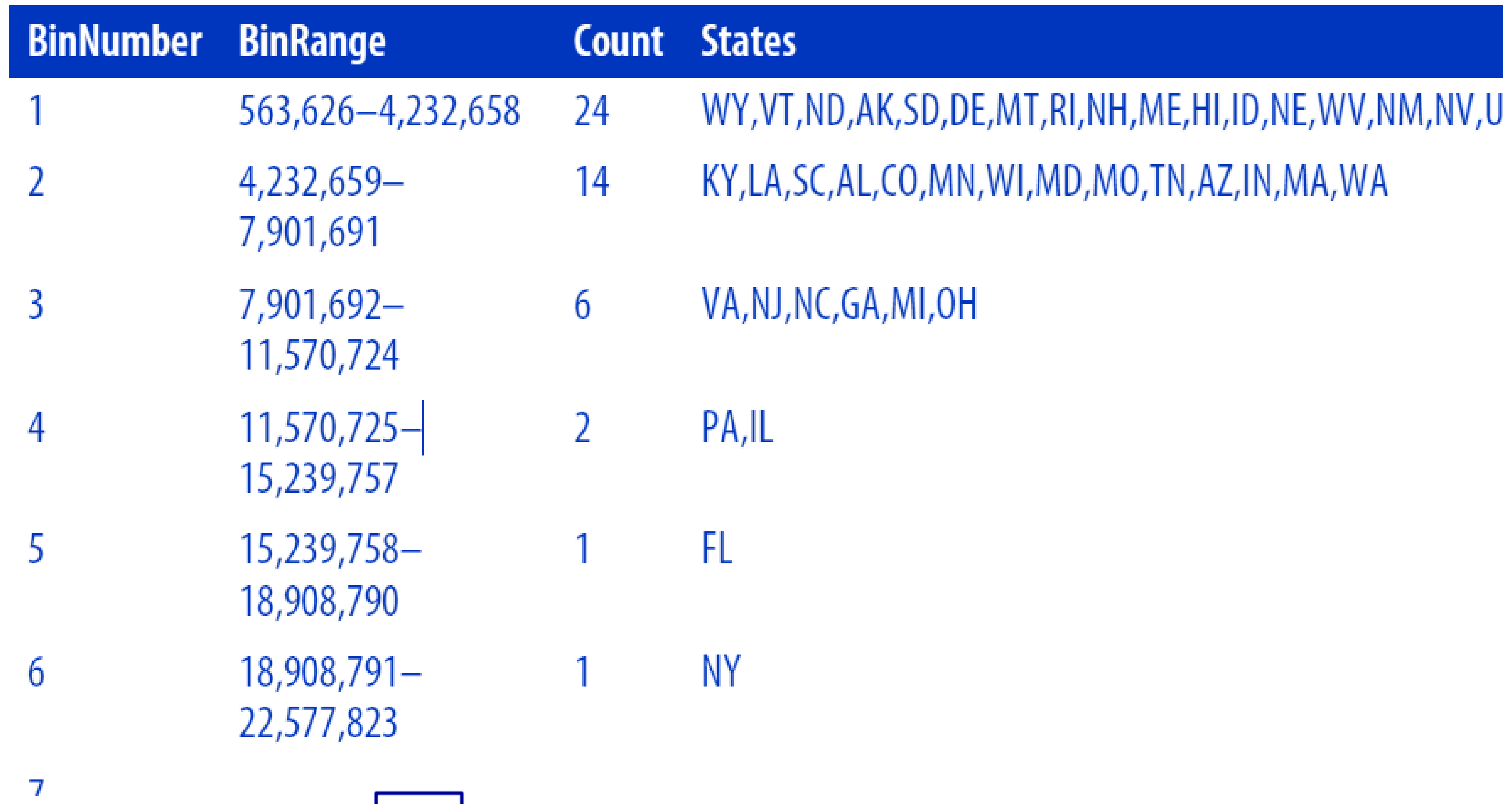


Exploring the Data Distribution

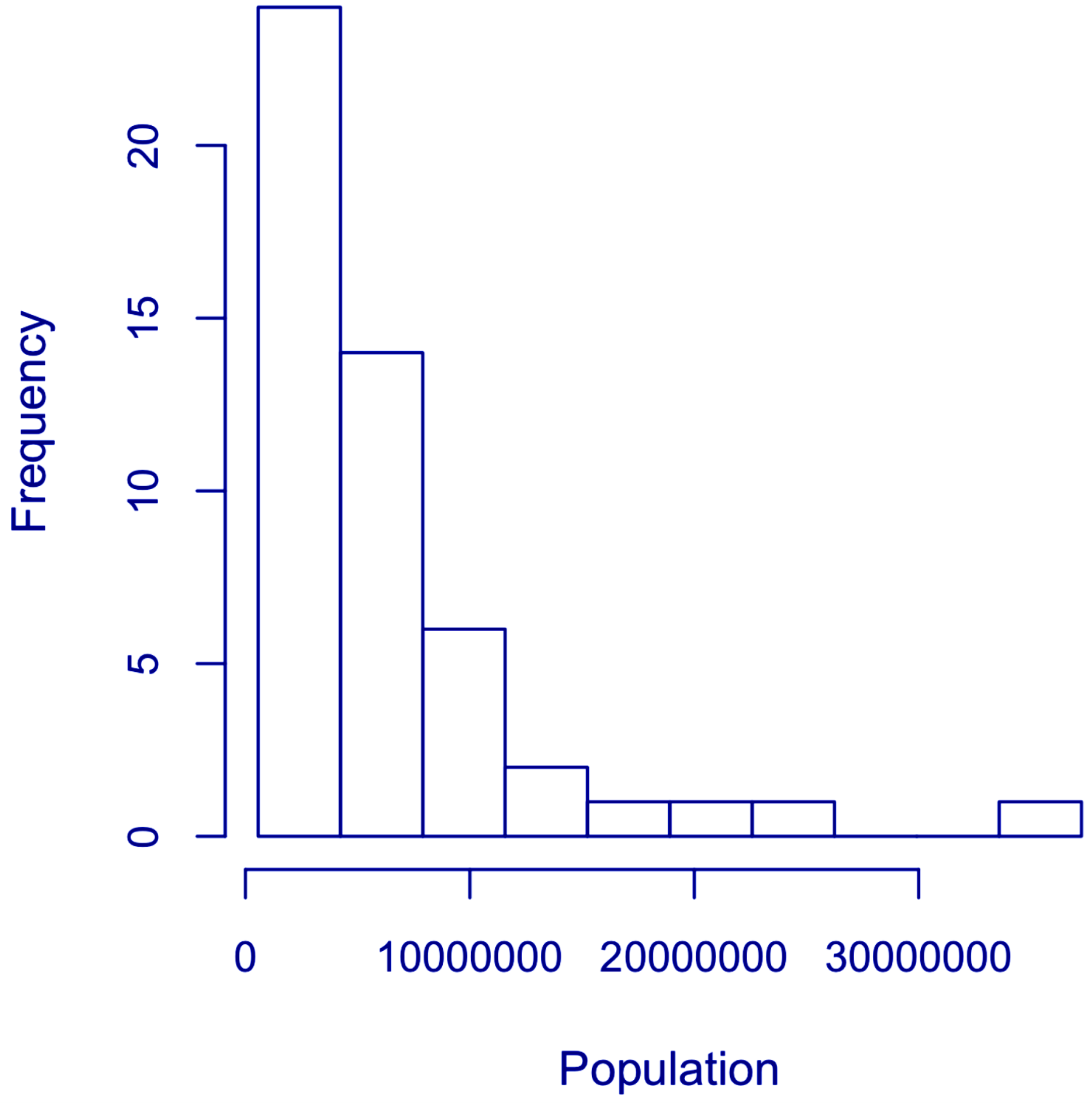
Frequency Tables and Histograms

- Divides up the variable range into equally spaced segments and tells us how many values fall within each segment.
- Bin Sizes
  - Too Small → Ability to see the big picture is lost
  - Too Big → Important features of the distribution can be obscured
- Histogram (Visualized Frequency Table)

- Converting numeric data to categorical data is an important and widely used step in data analysis since it reduces the complexity (and size) of the data. This aids in the discovery of relationships between features, particularly at the initial stages of an analysis.



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,U
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			



Exploring the Data Distribution

Statistical Moment

- 1st Moment of Distribution
  Location

- 2nd Moment of Distribution
  Variability

- 3rd Moment of Distribution
  Skewness

$$\tilde{\mu}_3 = \frac{\sum_i^N \left(X_i - \bar{X}\right)^3}{(N-1) * \sigma^3}$$

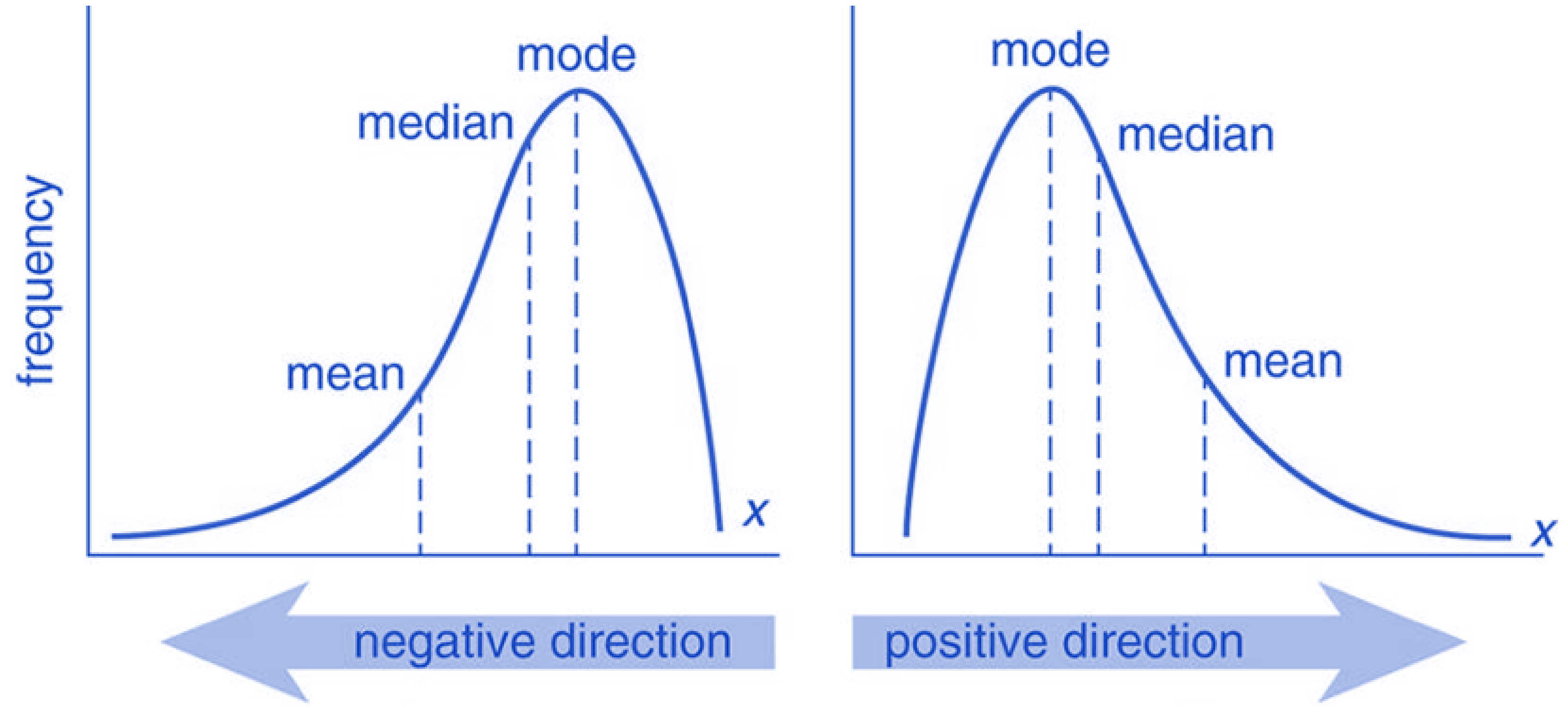
$\tilde{\mu}_3$ = skewness

N = number of variables in the distribution

X_i = random variable

$\bar{X}$ = mean of the distribution

σ = standard deviation



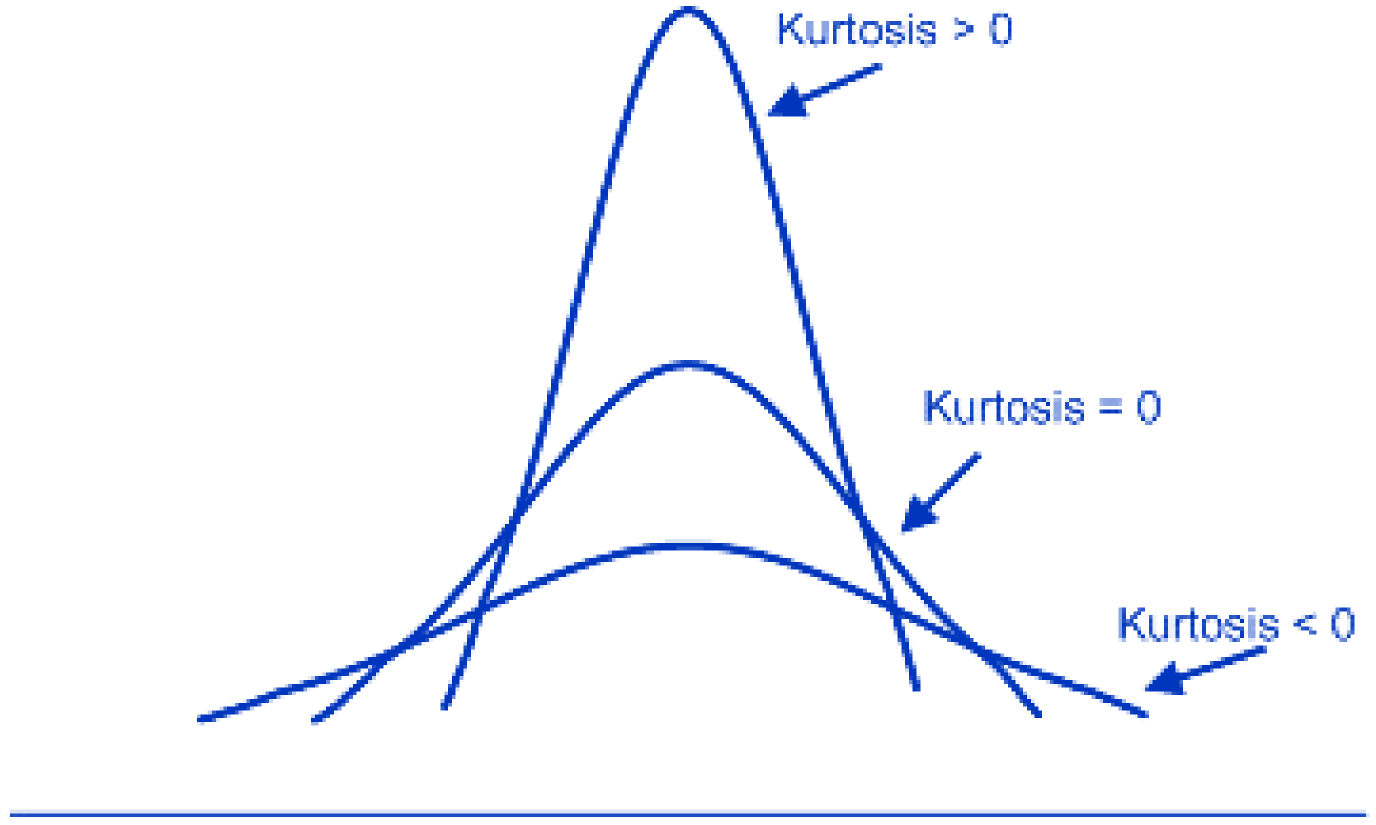
- 4th Moment of Distribution
  kurtosis

$$\text{Kurt} = \frac{\mu_4}{\sigma^4}$$

Kurt = kurtosis

μ_4 = fourth central moment

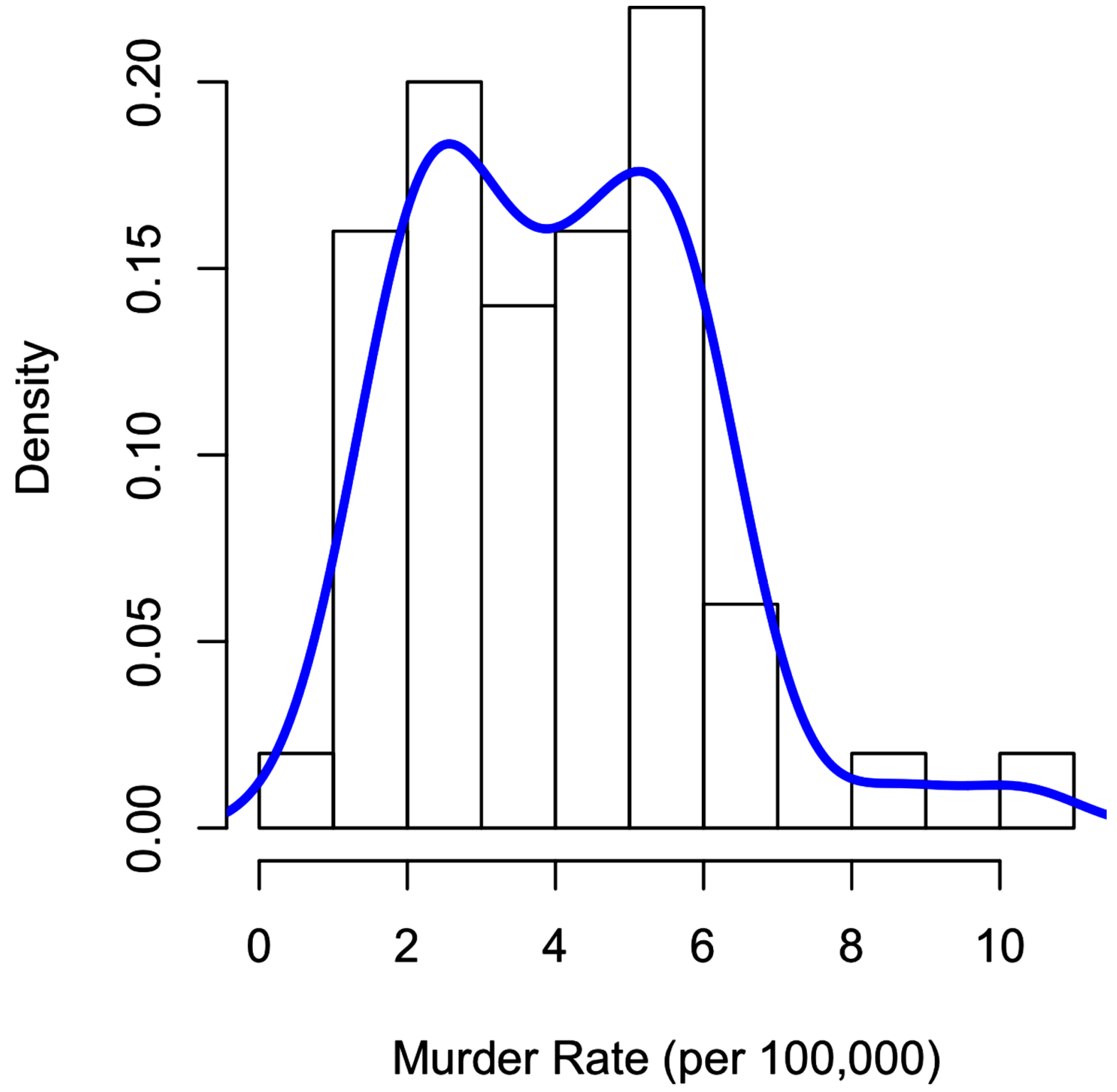
σ^4 = standard deviation



Exploring the Data Distribution

Density Plots and Estimates

- Distribution of data values as a continuous line.
- Total area under the density curve = 1
- Density plot corresponds to plotting the histogram as a proportion rather than counts.
- It requires a function to estimate a plot based on the data (Estimate Function)
- For many data science problems, there is no need to worry about the various types of density estimates; it suffices to use the base functions.



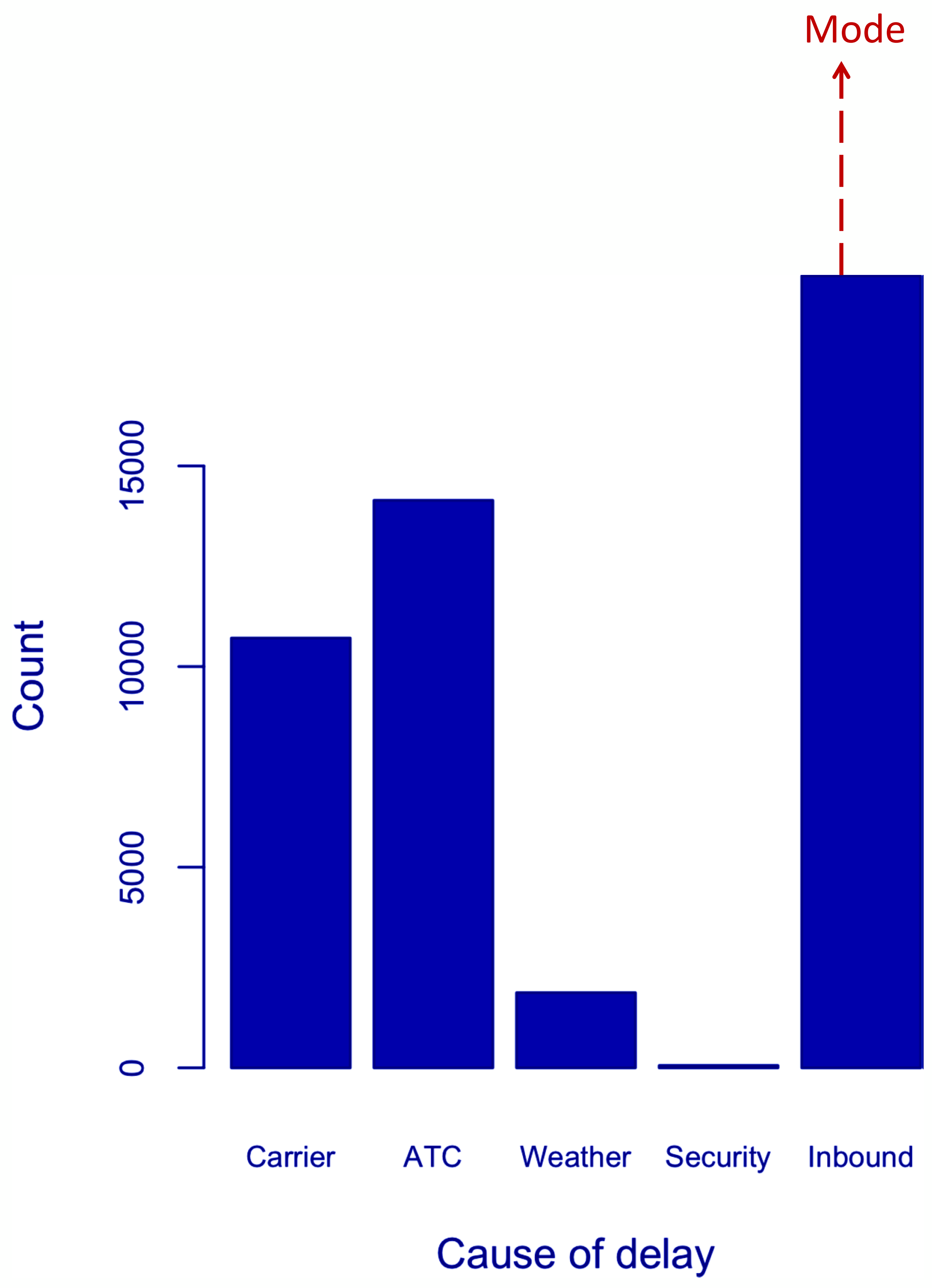
Exploring Binary and Categorical Data

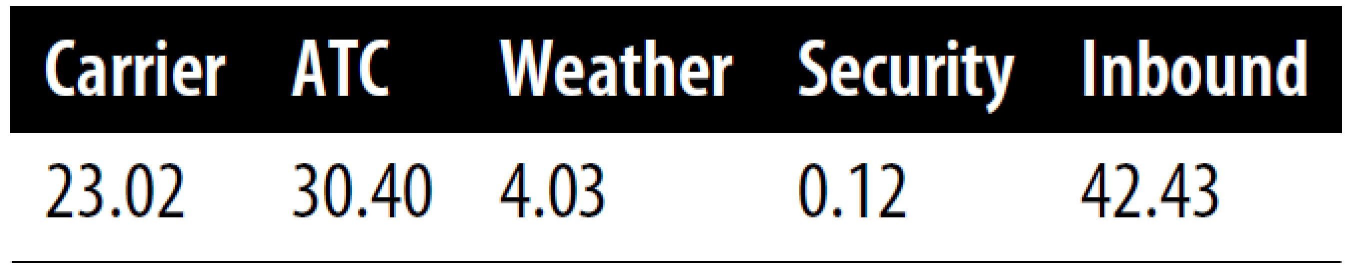
Bar Charts and Pie Charts

- Good for few categories.
- Just need to calculate the proportions of 1s
- Proportion of the important category
- Pie Charts are less visually informative in comparison to bar charts.

Mode

- The value that appears most often in the data
- The values in case of a tie that appear most often in the data
- It is generally not used for numeric data.





Carrier	ATC	Weather	Security	Inbound
23.02	30.40	4.03	0.12	42.43

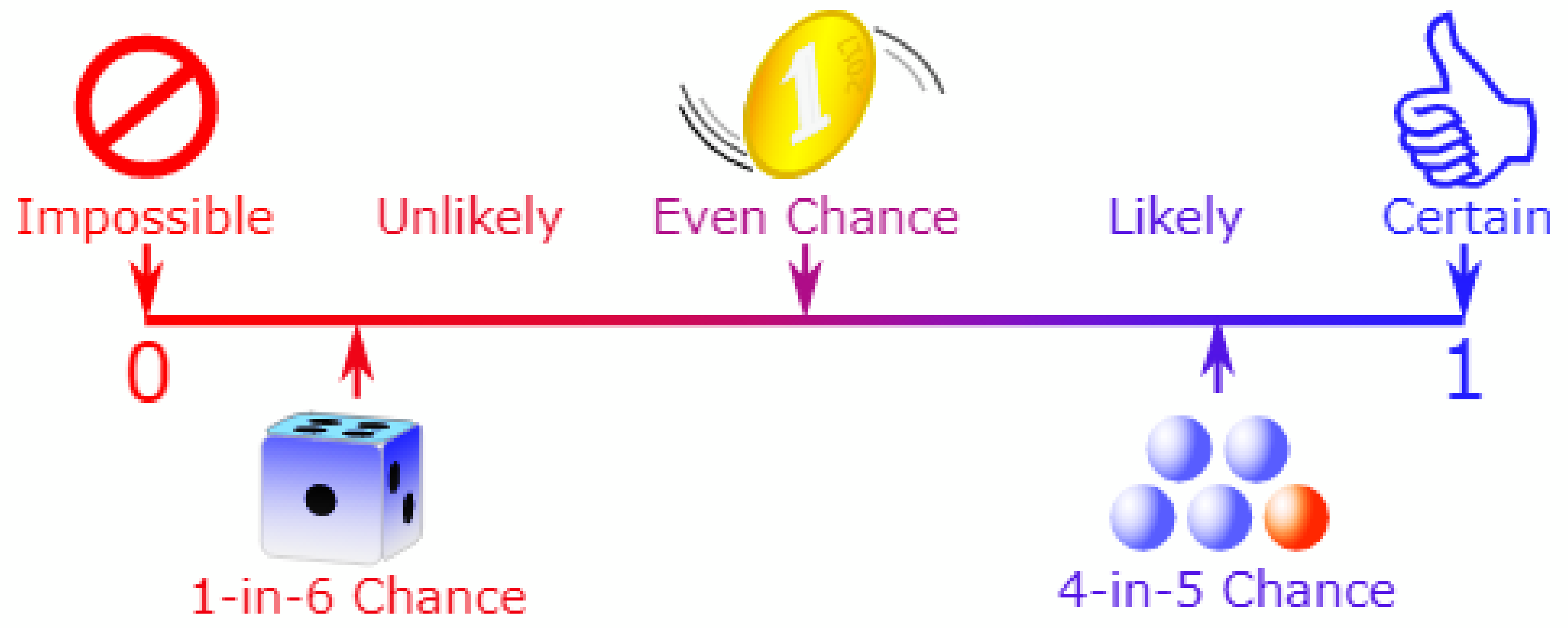
Exploring Binary and Categorical Data

Expected Value

- Is a form of weighted mean, in which the weights are probabilities.
- It adds the ideas of future expectations and probability weights, often based on subjective judgment.
- How to Calculate:
  - Multiply each outcome by its probability of occurrence.
  - Sum these values.
  - EV = (0.05)(300$) + (0.15)(50$) + (0.80)(0$) = 22 .5$

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.
- Example
  - If the odds that a team will win are 2 to 1, its probability of winning is 2/(2+1)= 2/3



Exploratory Data Analysis

Correlation

- Bivariate analysis
- Involves examining correlation among predictors, and between predictors and a target variable.
- Vector sum of products:
  - v1: {1, 2, 3} , v2: {4, 5, 6} , 1· 4 + 2 · 5 + 3 · 6 = 32
  - Reference to the resampling distribution in Permutation Test.
- Correlation coefficient is sensitive to outliers
- Spearman's rho or Kendall's tau Coefficient
  - Rank-based
  - Robust to outliers
  - Can handle certain types of nonlinearities
  - Rank-based estimates is mostly for smaller data sets and specific hypothesis tests.

Correlation Coefficient

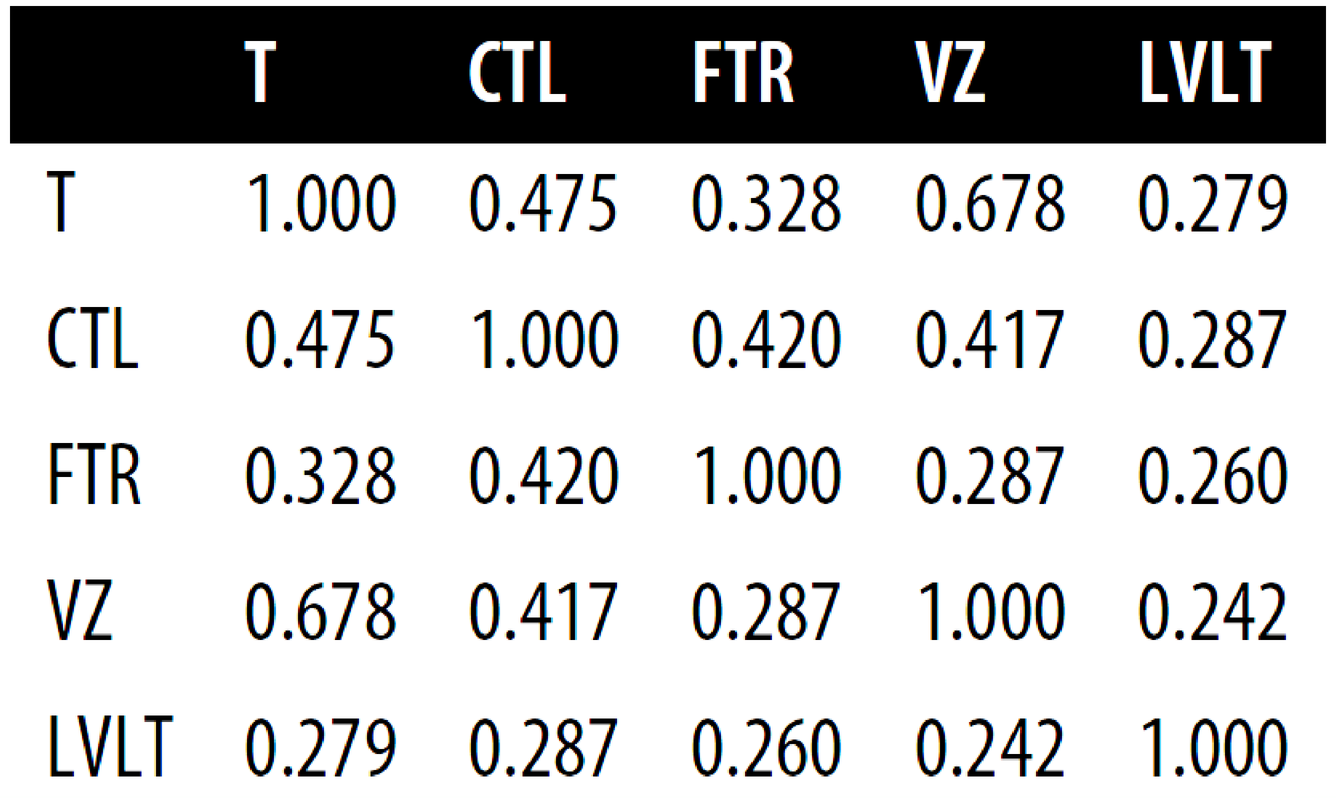
- Gives an estimate of the correlation between two variables.
- **Pearson's correlation coefficient:**
  - –1 (perfect negative correlation)
  - +1 (perfect positive correlation). — → Zero – No Correlation
  - Value Base
  - For non-linear correlation this metric is not useful

$$r = \frac{\overbrace{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}^{\text{Multiplication of deviations from the mean for variable 1 times those for variable 2}}}{\underbrace{(n-1)s_x s_y}_{\text{product of the standard deviations}}}$$

Exploratory Data Analysis

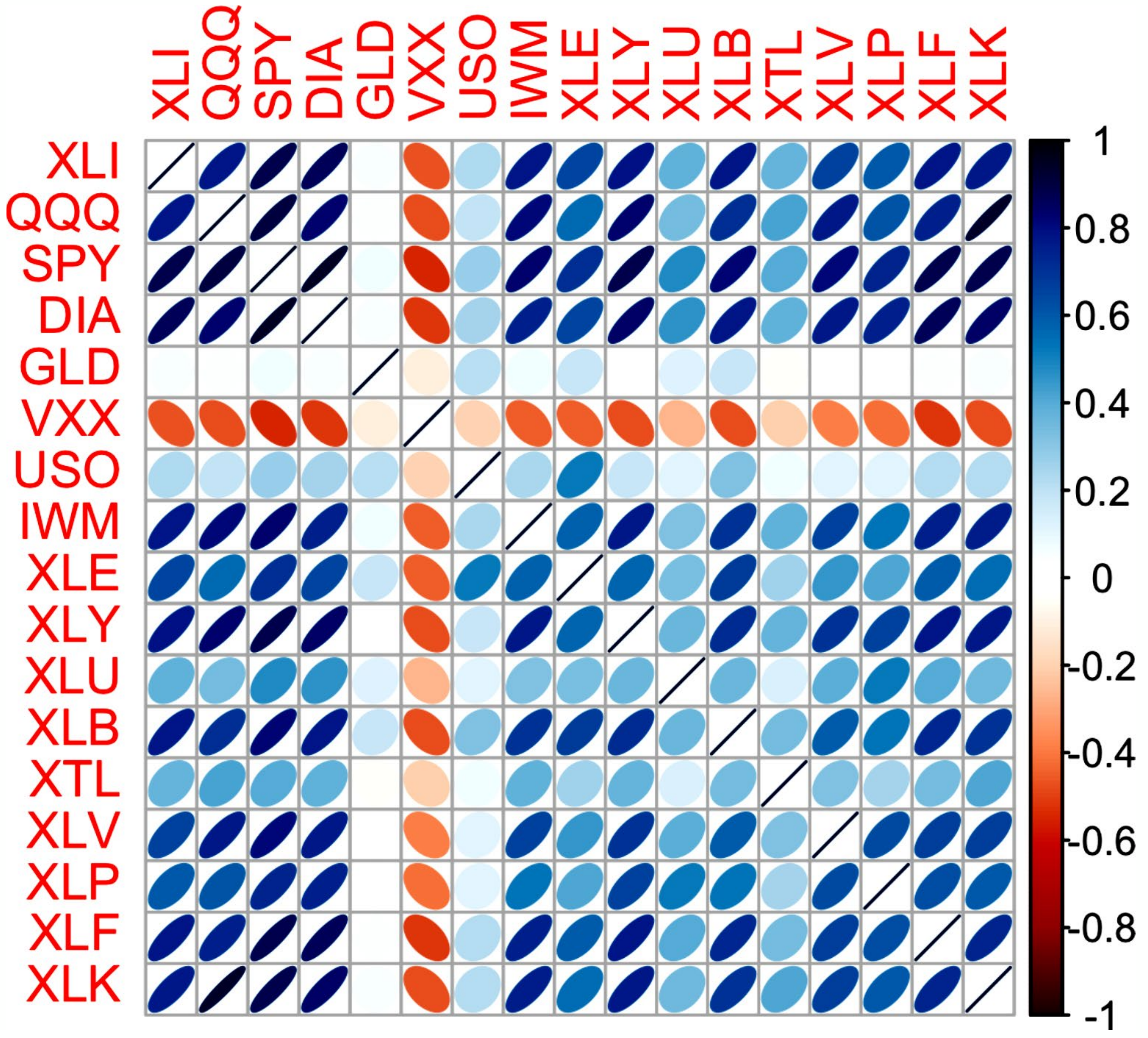
Correlation Matrix

- Shows the correlation between 2 variables



	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

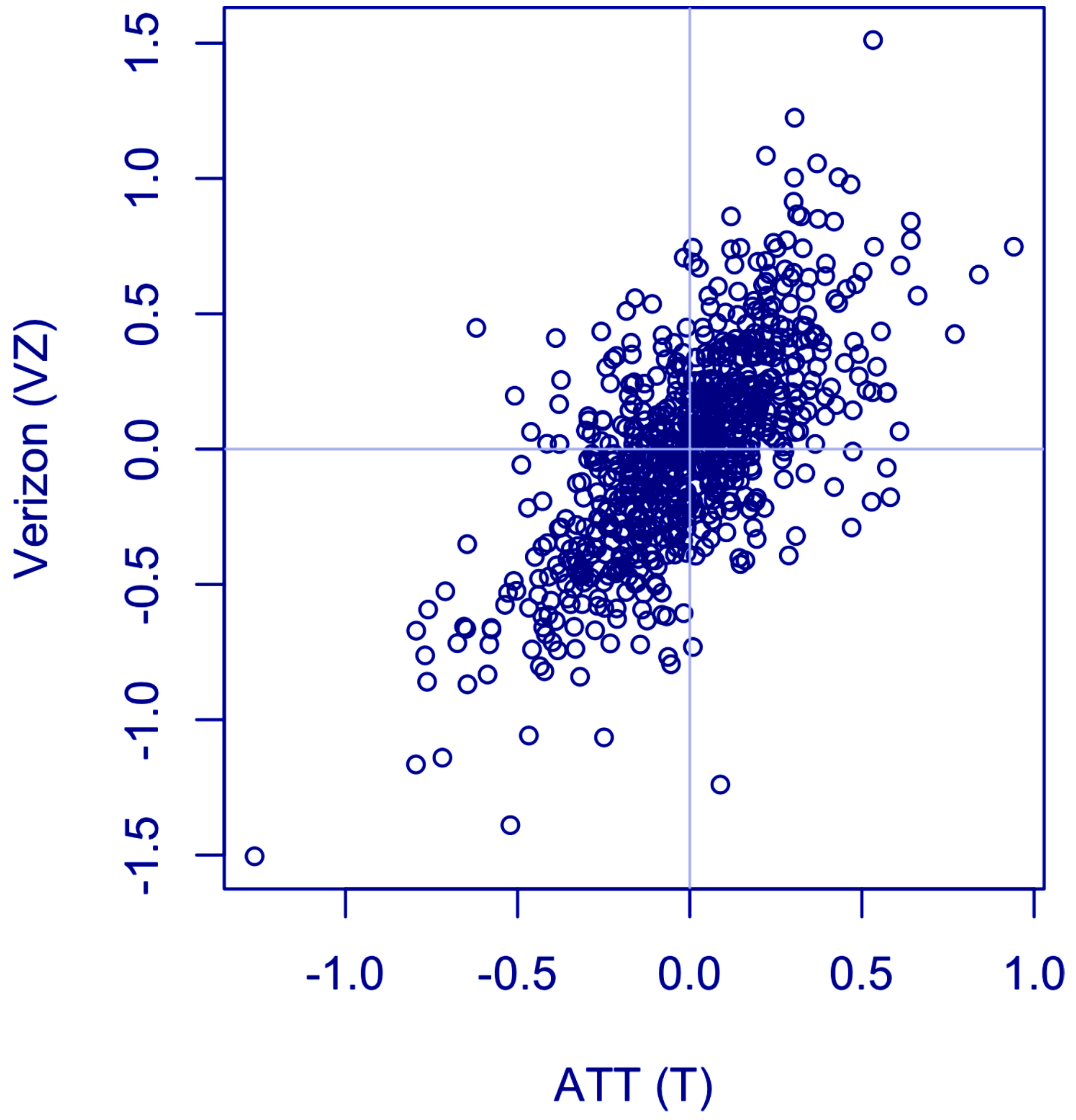
- Visualization of correlation matrices using **HEATMAP**



Exploratory Data Analysis

Scatterplot

- The standard way to visualize the relationship between two measured data variables.
- It actually is a single cell in correlation matrix with more information on both observation units.
- Difficult to identify details in the middle of the plot
- Best performance is on small data sets
  - Adding transparency to the points
  - Hexagonal binning and density plots

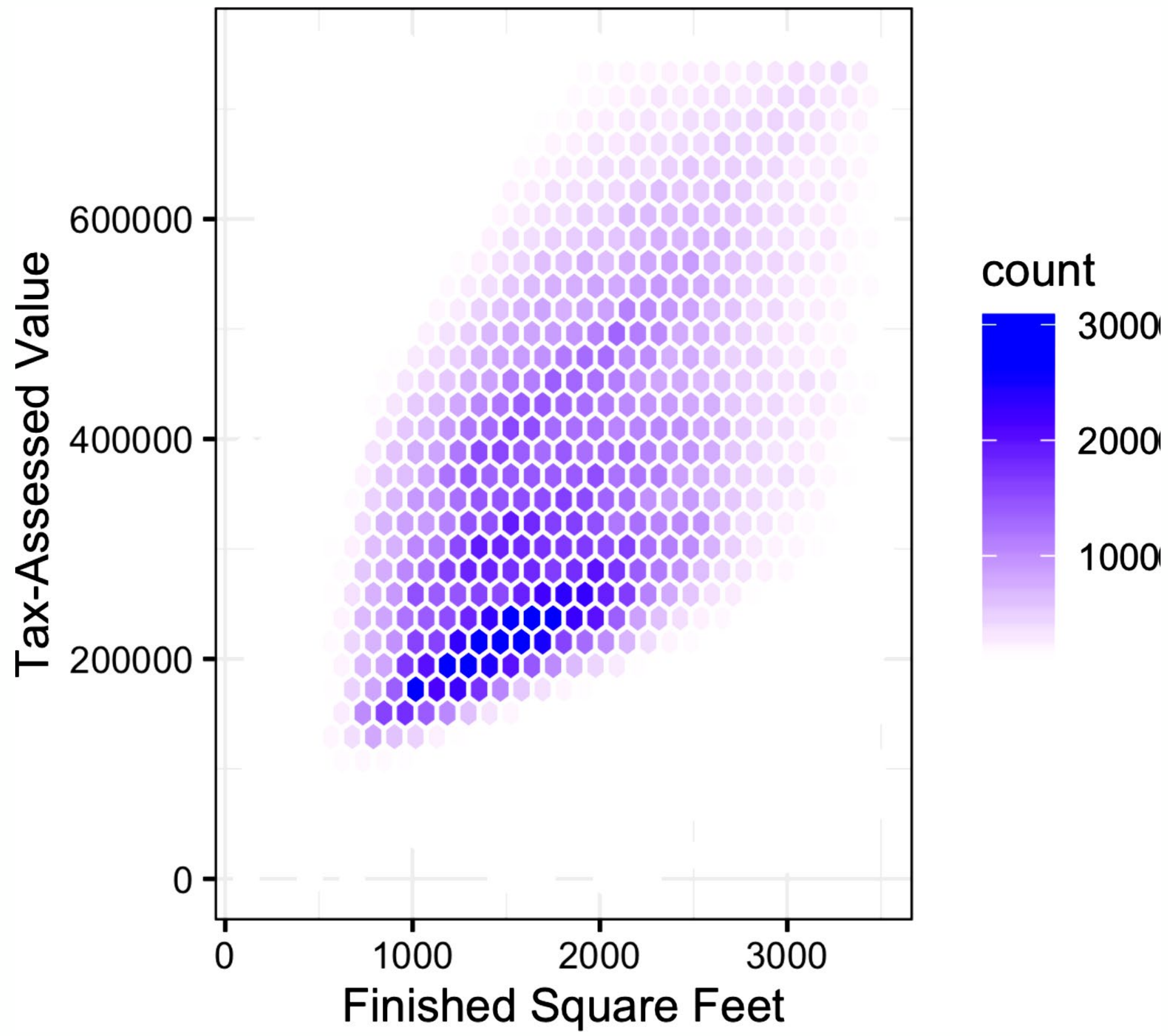


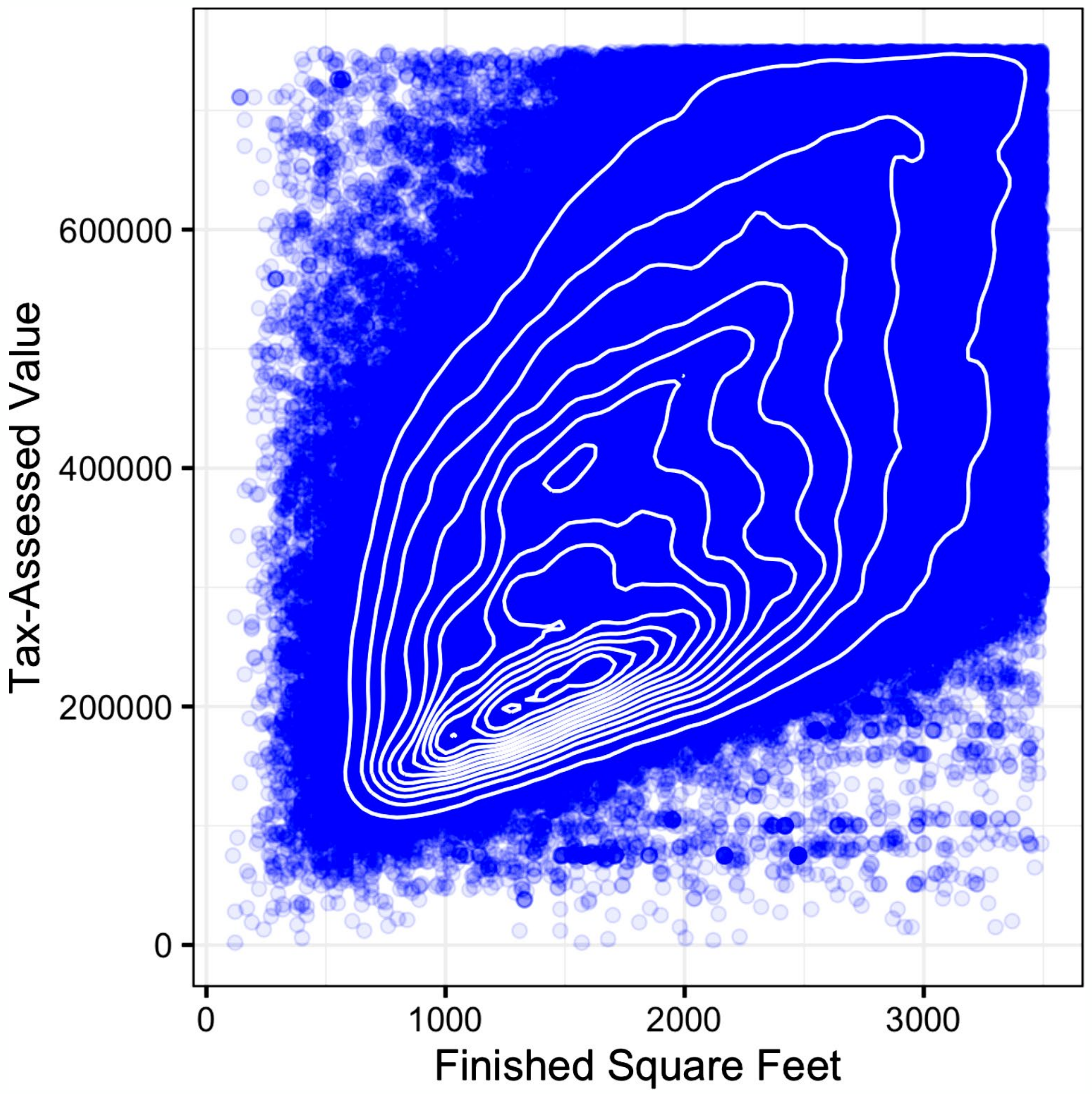
Multivariate Analysis

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)

- Awesome for ton of data



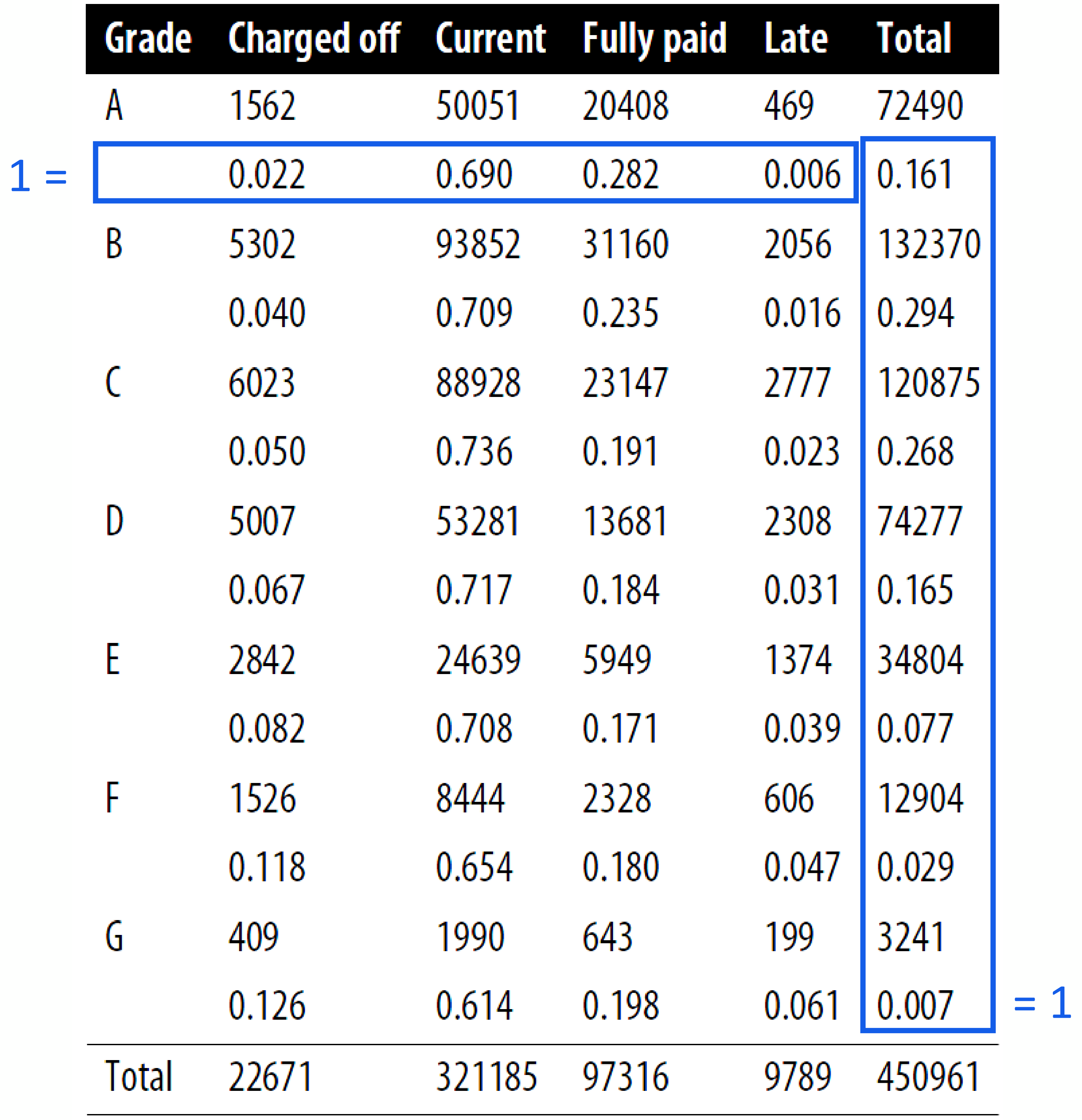


We can also use Heatmaps for this kind of analysis

Multivariate Analysis

Contingency Table (Two Categorical Variables)

- A table of counts by category
- Can also include column and total percentages.
- Pivot tables in Excel

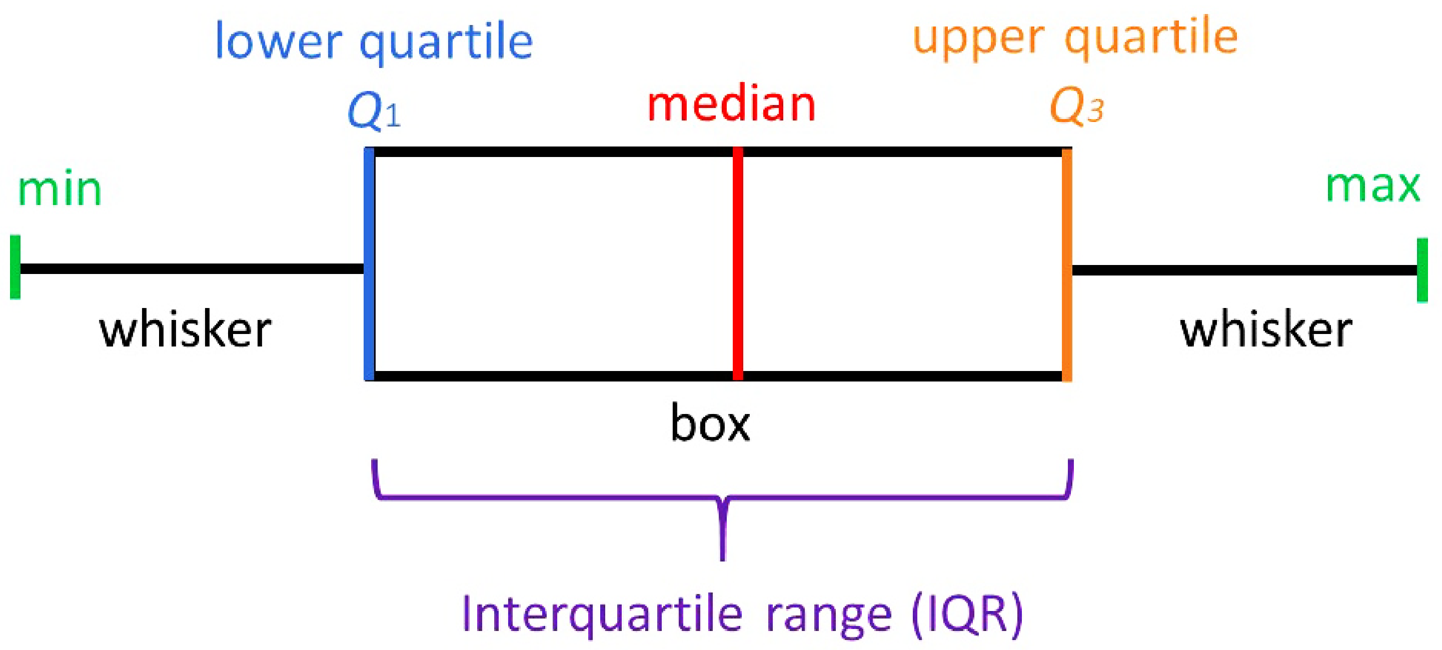


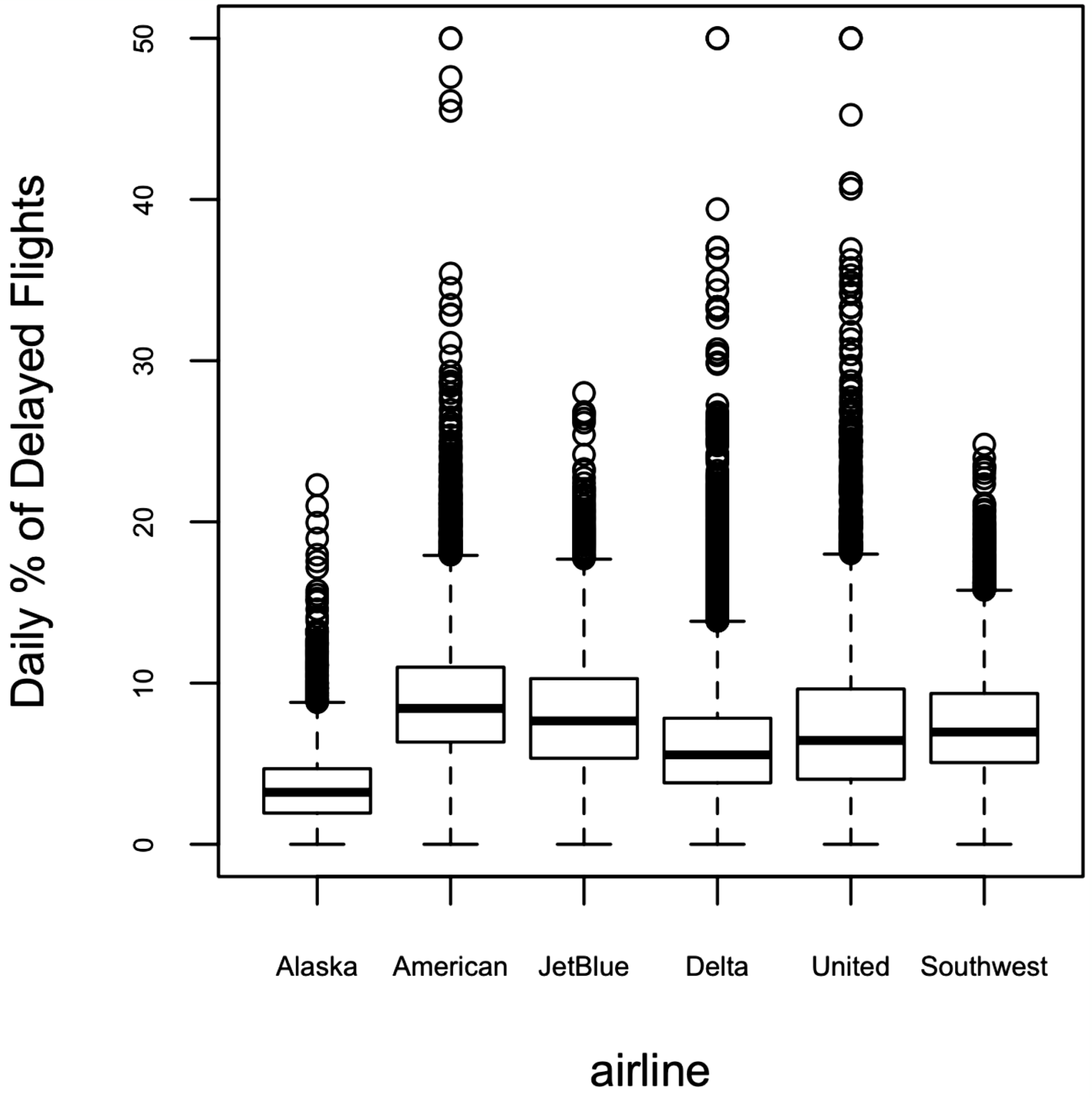
Grade	Charged off	Current	Fully paid	Late	Total
A	1562	50051	20408	469	72490
	0.022	0.690	0.282	0.006	0.161
B	5302	93852	31160	2056	132370
	0.040	0.709	0.235	0.016	0.294
C	6023	88928	23147	2777	120875
	0.050	0.736	0.191	0.023	0.268
D	5007	53281	13681	2308	74277
	0.067	0.717	0.184	0.031	0.165
E	2842	24639	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

Multivariate Analysis

Boxplots (Categorical and Numeric Data)

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



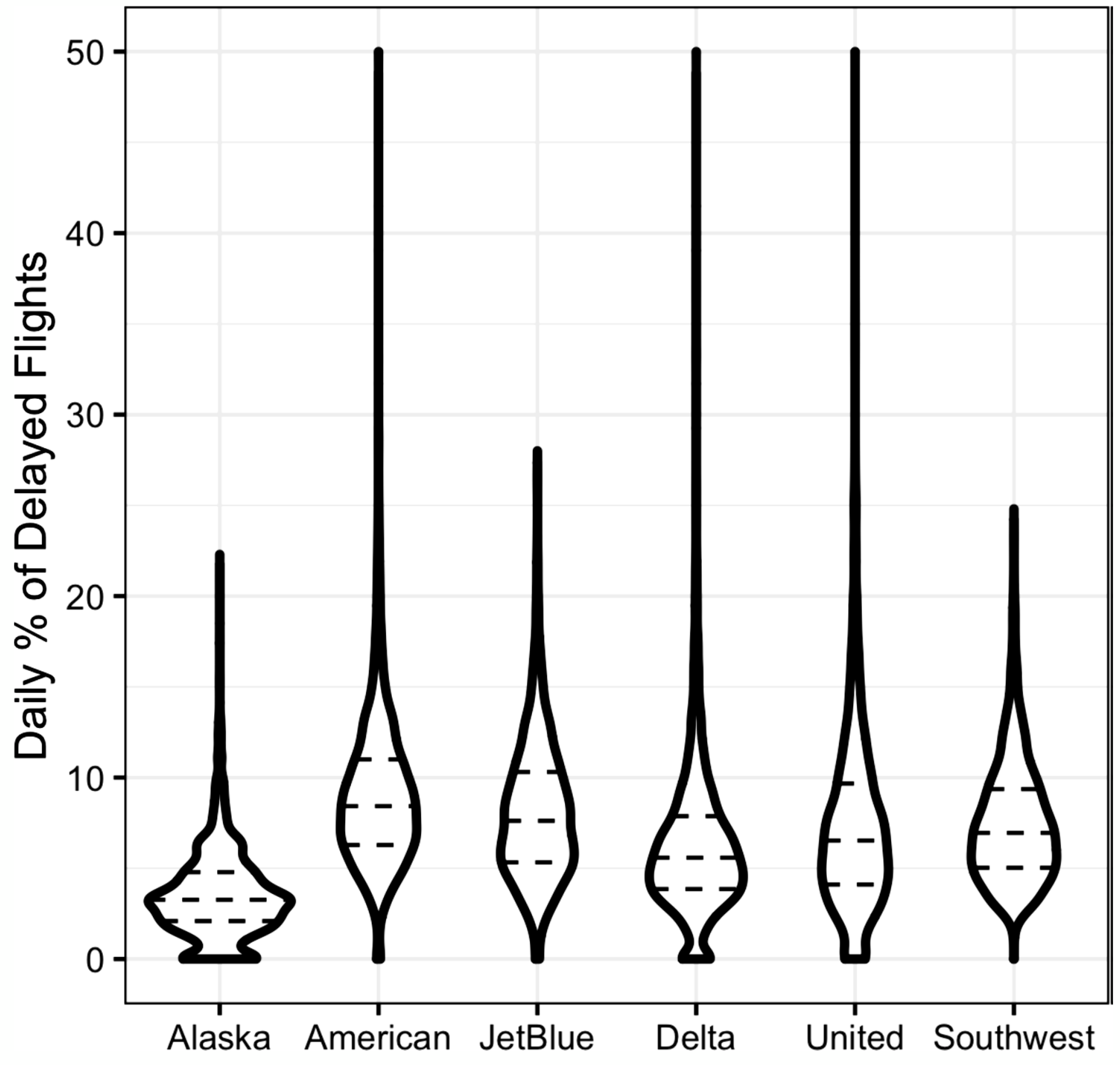


Multivariate Analysis

Violin Plot (Categorical and Numeric Data)

- An enhancement to the boxplot
- Plots the density estimate with the density on the y-axis
- Can show nuances in the distribution that aren't perceptible in a boxplot
- The boxplot more clearly shows the outliers in the data

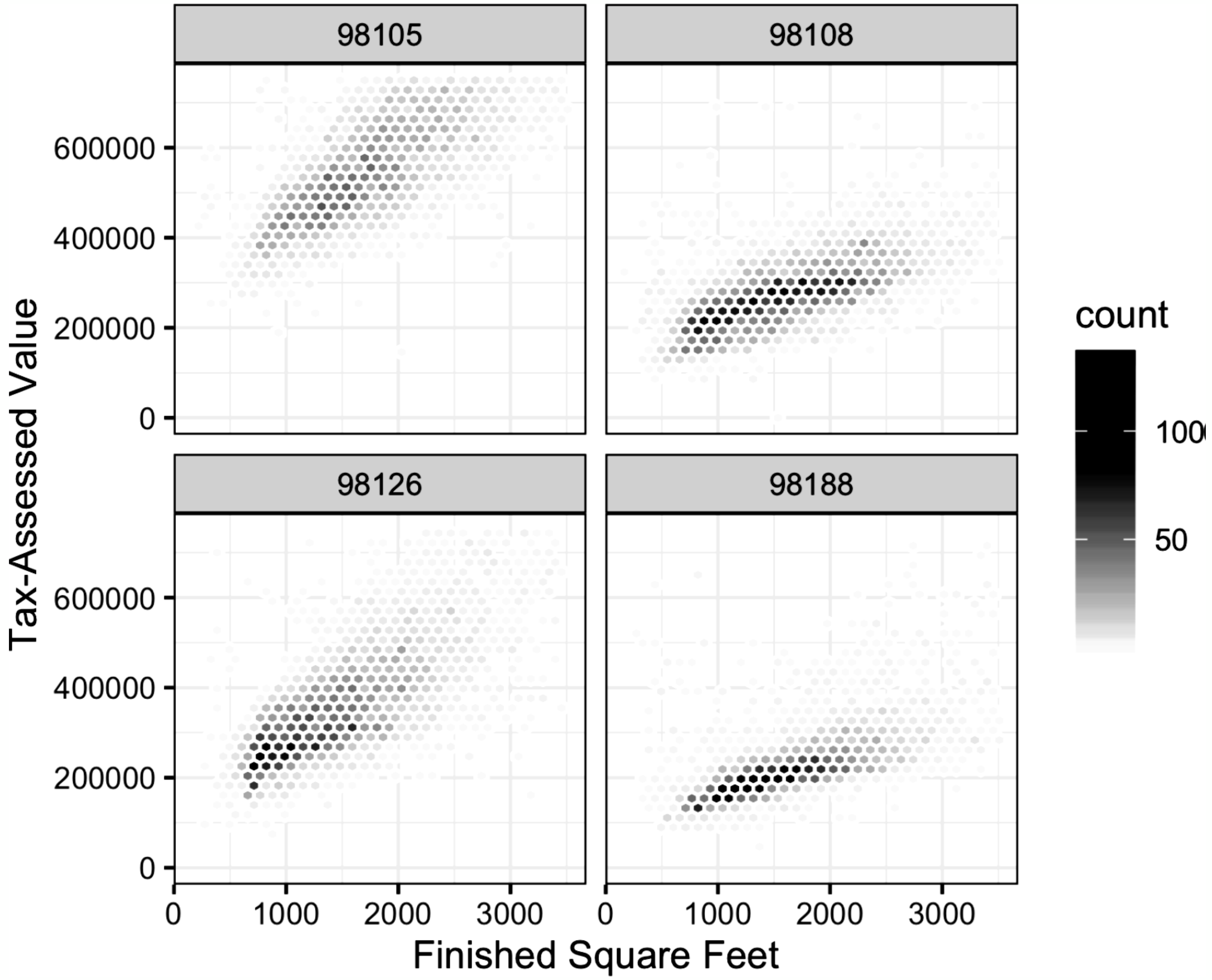
- You can combine a violin plot with a boxplot



Multivariate Analysis

Visualizing Multiple Variables

- All Types of Bivariate plots are expandable to Multivariate



THANK YOU TO ALL!

Any Comment or Question ?

Reza Barahmand

rbarahmand

rbarahmand