CSCI 556 Data Analysis & Visualization

Exploratory Data Analysis

Instructor: Dr. Jinoh Kim

Exploratory data analysis (EDA)

- Exploring the data - first step in data science
- Structured vs. unstructured data
- Numeric vs. categorical variables
- Mean, variation, distribution of data
- Binary and categorical data
- Correlation
- Multivariate analysis

Structured vs. unstructured

- Much of data is unstructured
  - Images: collection of pixels with each pixel containing RGB (red, green, blue) info
  - Texts: sequences of words and nonword characters
- To apply the statistical concepts, data should be in a structured form (like database table)

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

Data types

- Basic types of structured data: numeric and categorical
- Numeric: integer or real numbers
  - Continuous: Data that can take on any value in an interval
  - Discrete: Data that can take on only integer values
- Categorical: set of values
  - Also called “nominal”
  - Binary (dichotomy): Special case of categorical data with just two categories of values (0/1, true/false)
  - Ordinal: Categorical data that has an explicit ordering

Rectangular data (data frame)

- Two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables)
  - Typical frame of reference for an analysis in data science
- Record: A row in the table (aka sample, instance, example, observation)
- Feature: A column in the table (aka attributes, variable, predictor, input)
- Outcome: Many data science projects involve predicting an outcome (aka target, output, response)
  - Features are used to predict the outcome

Rectangular data (example)

ID	Sepal length	Sepal width	Petal length	Petal width	Label
1	5.1	3.5	1.4	0.2	Iris-setosa
2	4.9	3	1.4	0.2	Iris-setosa
3	7	3.2	4.7	1.4	Iris-versicolor
4	6.4	3.2	4.5	1.5	Iris-versicolor
5	6.9	3.1	4.9	1.5	Iris-versicolor
6	5.5	2.3	4	1.3	Iris-versicolor
7	6.1	3	4.9	1.8	Iris-virginica
8	6.4	2.8	5.6	2.1	Iris-virginica
9	7.2	3	5.8	1.6	Iris-virginica

Nonrectangular data structures

- Time series data: successive measurements of the same variable
- Spatial data: used in mapping and location analytics
- Graph data: used to represent physical, social, and abstract relationships
- Need specialized methodology in data science

Mean of data

- Mean: average value

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

- Trimmed mean: dropping a fixed number of sorted values at each end and then taking an average of the remaining values

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

- Weighted mean: multiplying each data value by a weight and dividing their sum by the sum of the weights

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

Median and outliers

- Median: the middle number on a sorted list of the data
- Outlier: any value that is very distant from the other values in a data set
  - Exact definition of an outlier is somewhat subjective
- Median is a robust estimate of location since it is not influenced by outliers that could skew the results
- Trimmed mean is a compromise between the median and the mean, widely used to avoid the influence of outliers

Variation of data

- Variability (dispersion) measures whether the data values are tightly clustered or spread out
- Variability metrics
  - Deviation: difference between the observed values and the estimate of location
  - Mean absolute deviation: mean of the absolute value of the deviations from the mean

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

  - Variance: sum of squared deviations from the mean divided by the number of data instances

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

  - Standard deviation: square root of the variance

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

Estimates based on percentiles

- Estimating dispersion by spread of data
  - Range: difference between the largest and the smallest value in a data set
  - Percentile (quantile): 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
  - Interquartile range (IQR): difference between the 75th percentile and the 25th percentile
  - Others like median absolute deviation, ranks, etc

Percentile (quantile)

- *P*th percentile: 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
- To find the 80th percentile
  - Sort the data
  - Starting with the smallest value, proceed 80 percent of the way to the largest value
- Example: data = <10, 3, 4, 7, 8, 2>
  - 0th percentile = 2
  - 100th percentile = 10
  - 50th percentile = 5.5
- Median = 50th percentile

Interquartile range (IQR)

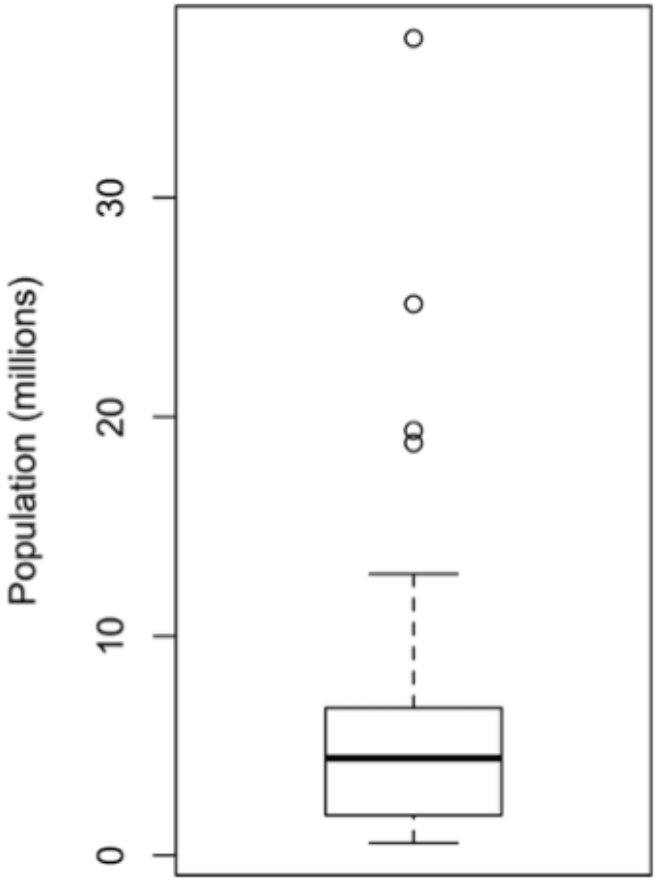
- IQR = difference between the 25th percentile (Q1) and the 75th percentile (Q3)
- Example: data=<3,1,5,3,6,7,2,9>
  - After sorting, data=<1,2,3,3,5,6,7,9>
  - 50th percentile (median) = 4
  - First half = <1,2,3,3>, second half=<5,6,7,9>
  - 25th percentile = median of first half = 2.5
  - 75th percentile = median of second half = 6.5
  - IQR = 6.5 – 2.5 = 4
- Rule for outliers = 1.5 * IQR
  - Suspected outlier if an observation falls more than 1.5*IQR (above Q3 or below Q1)

Data distribution

- Exploring how the data is distributed overall
  - Rather than giving a single number like mean or variation
- Tools:
  - Boxplot: visualizing the distribution of data based on percentile info
  - Frequency table: tally of the count of numeric data values that fall into a set of intervals (bins)
  - Histogram: plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y- axis
  - Density plot: smoothed version of the histogram

Boxplot

- Top and bottom of the box are the 75th and 25th percentiles, respectively
- Median is shown by the horizontal line in the box
- The dashed lines (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 (i.e., outliers meeting 1.5 * IQR rule)



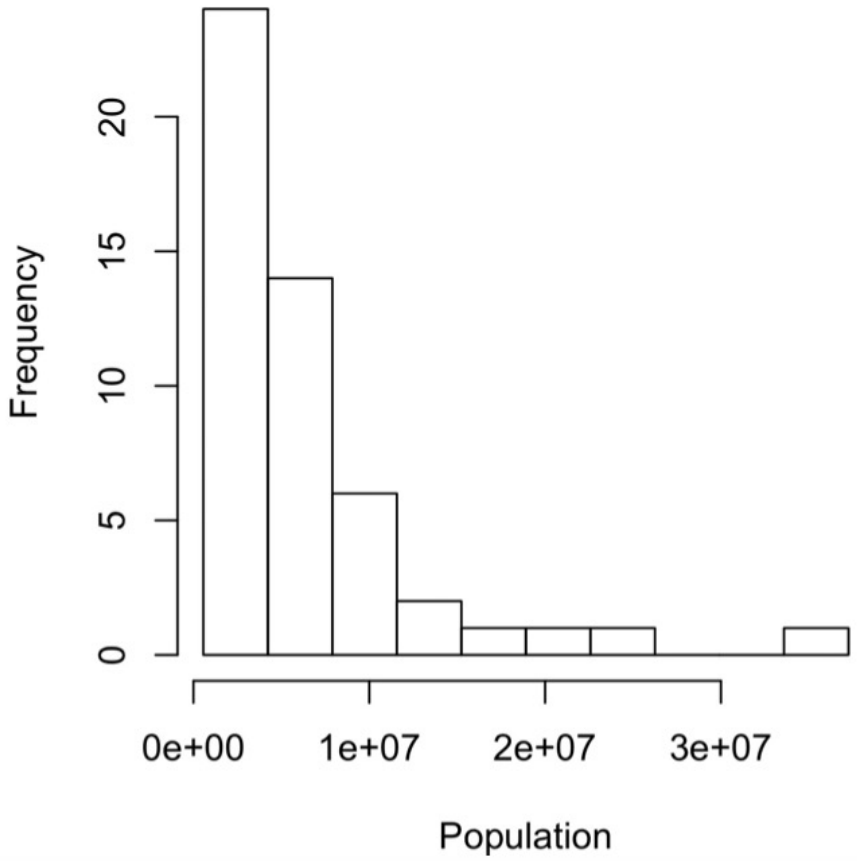
Frequency table

- Frequency table of a variable divides up the variable range into equally spaced segments, and tells us how many values fall in each segment

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,AI
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

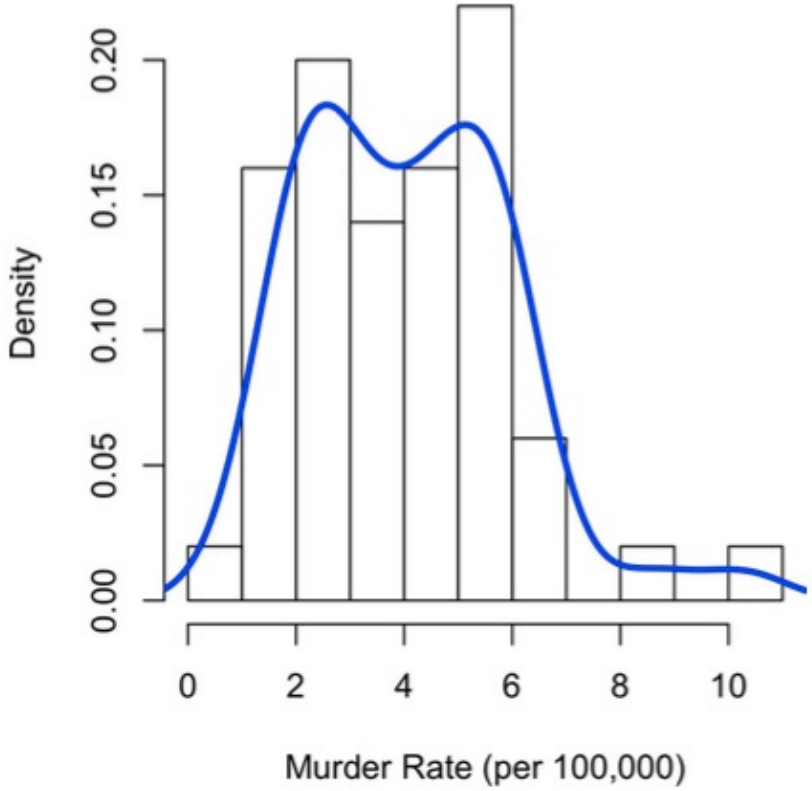
Histogram

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



Density plot

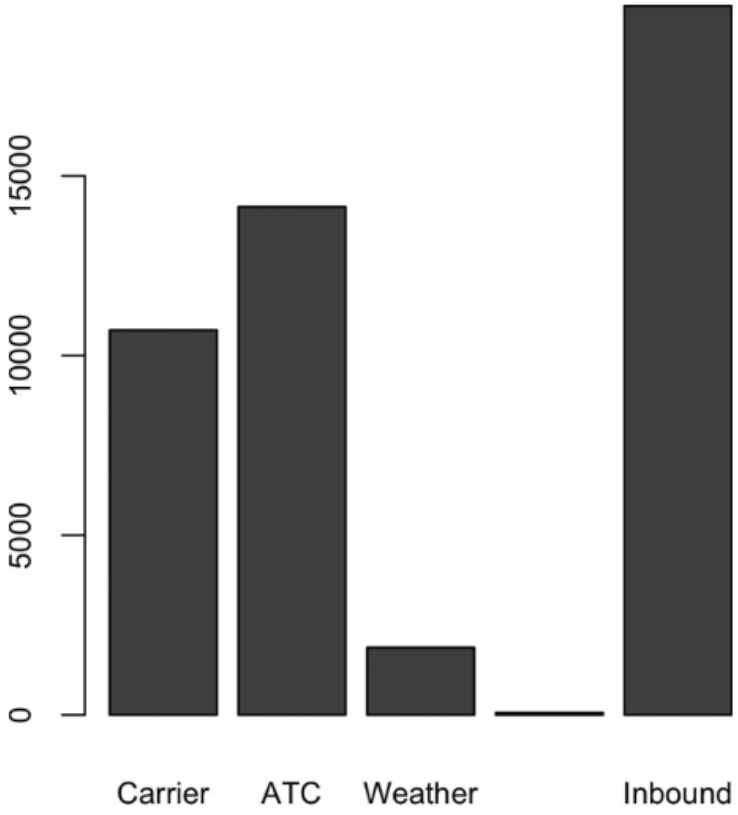
- Shows the distribution of data values as a continuous line (i.e., smoothed histogram)
- Typically computed directly from the data through a kernal density estimate



Categorical data (including binary)

- For categorical data, simple proportions or percentages tell the story of the data
  - Can be explored using bar charts, pie charts, etc.

- Mode: most commonly occurring category or value in a data set
  - Example: the mode of the cause of delay at DFW airport is "late inbound"



Categorical data (cont'd)

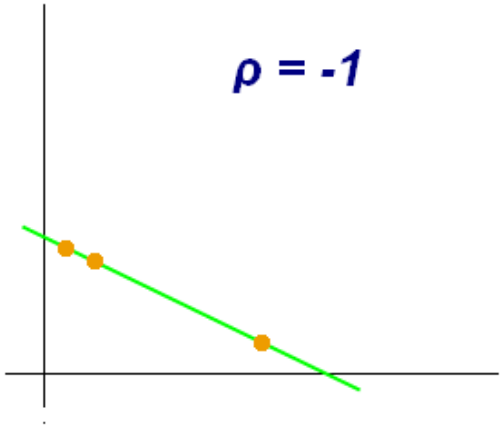
- Expected value: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence
- Example: 5% of the attendees will sign up for the $300 service, 15% for the $50 service, and 80% will not sign up for anything

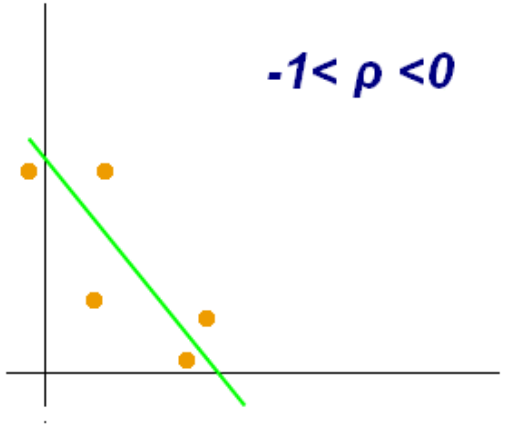
$$EV = (0.05)(300) + (0.15)(50) + (0.80)(0) = 22.5$$

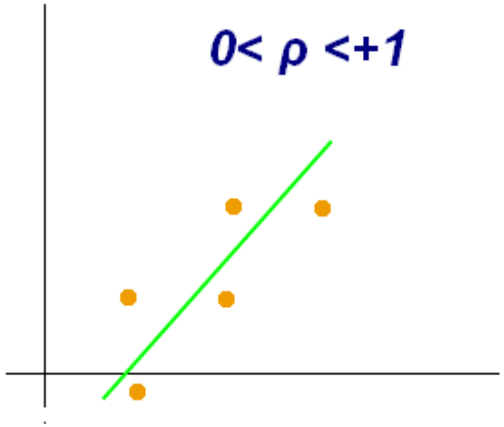
Correlation

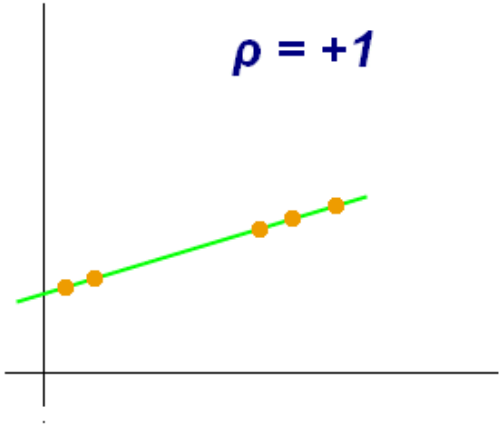
- Variables X and Y (each with measured data) are said to be positively correlated if high values of X go with high values of Y, and low values of X go with low values of Y.
- If high values of X go with low values of Y, and vice versa, the variables are negatively correlated.
- Correlation coefficient (denoted as *r* or ρ): metric that measures the extent to which numeric variables are associated with one another (ranges from –1 to +1)
  - positive correlation → +1, negative correlation → -1, less correlation → 0

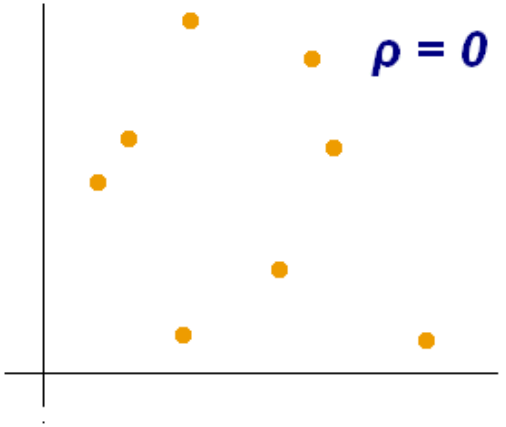
Correlation examples











Source: Wiki

Pearson's correlation coefficient

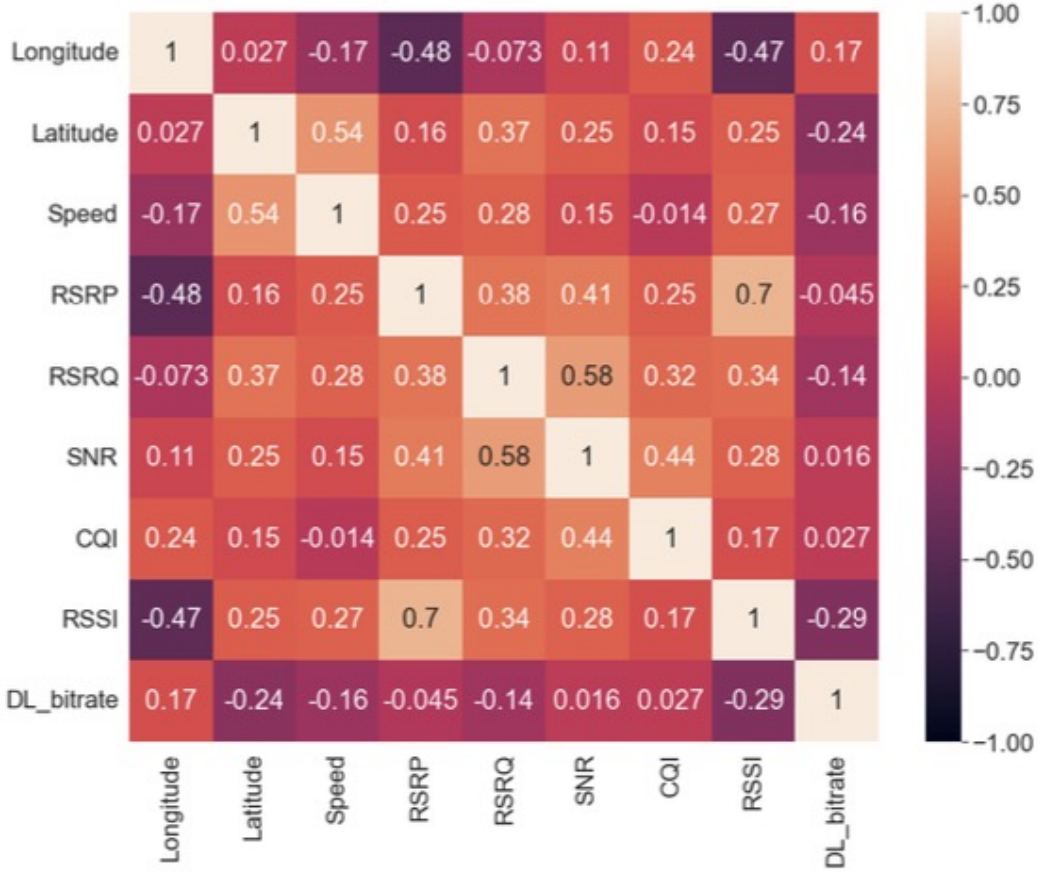
$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{(N-1)s_x s_y}$$

- N: number of samples
- $\bar{x}$, $\bar{y}$: mean of x and y
- s_x, s_y: standard deviation of x and y

Correlation matrix

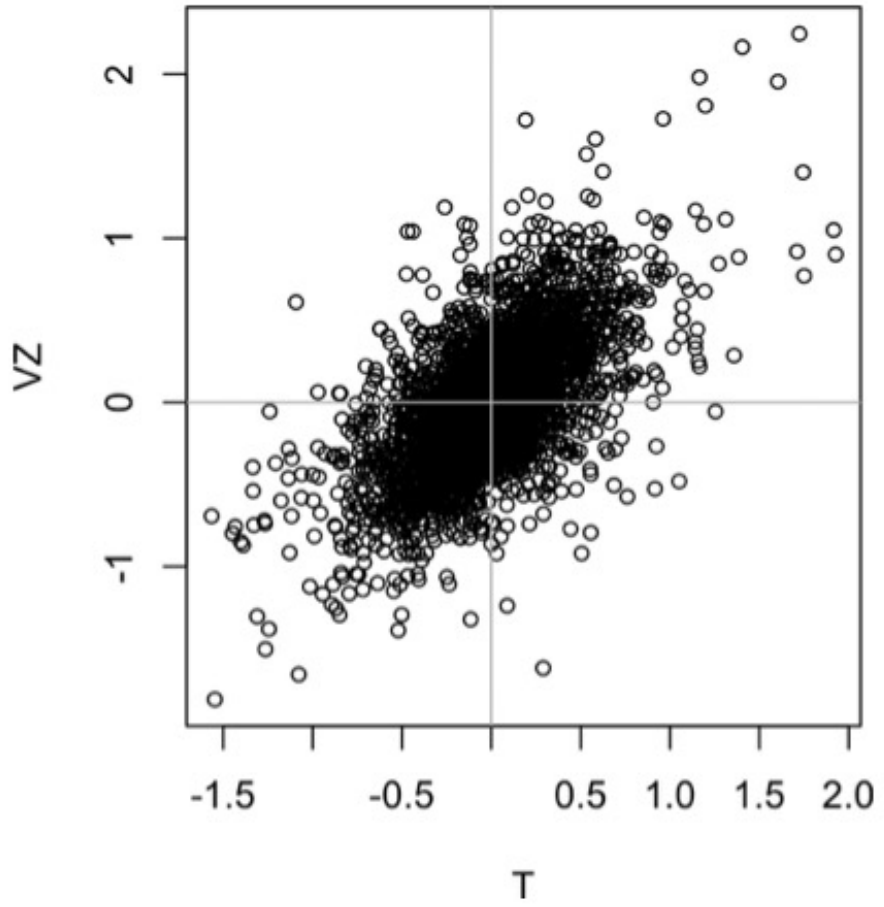
- Correlation matrix: table where the variables are shown on both rows and columns, and the cell values are the correlations between the variables

Example: correlation matrix for nine variables



Scatterplots

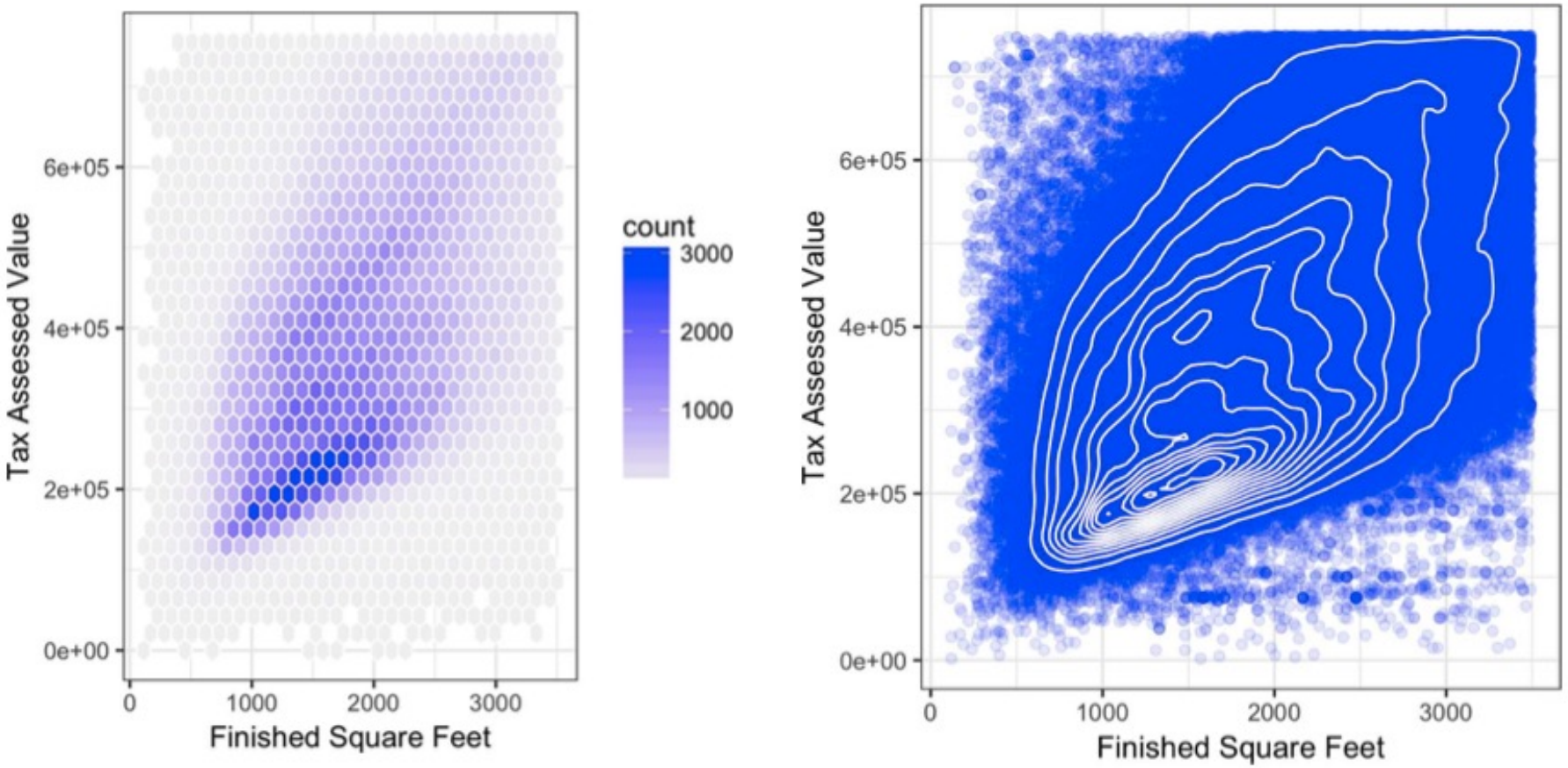
- Simple way to visualize the relationship between two measured data variables



Multivariate analysis

- Exploring two or more variables
  - Univariate analysis explores a single variable only
- Correlation analysis is an important method that compares two variables (*bivariate analysis*)
- Tools:
  - Hexagonal binning: plot of two numeric variables with the records binned into hexagon
  - Contour plots: plot showing the density of two numeric variables like a topographical map
  - Violin plots: Similar to a boxplot but showing the density estimate

Hexagonal binning and contours

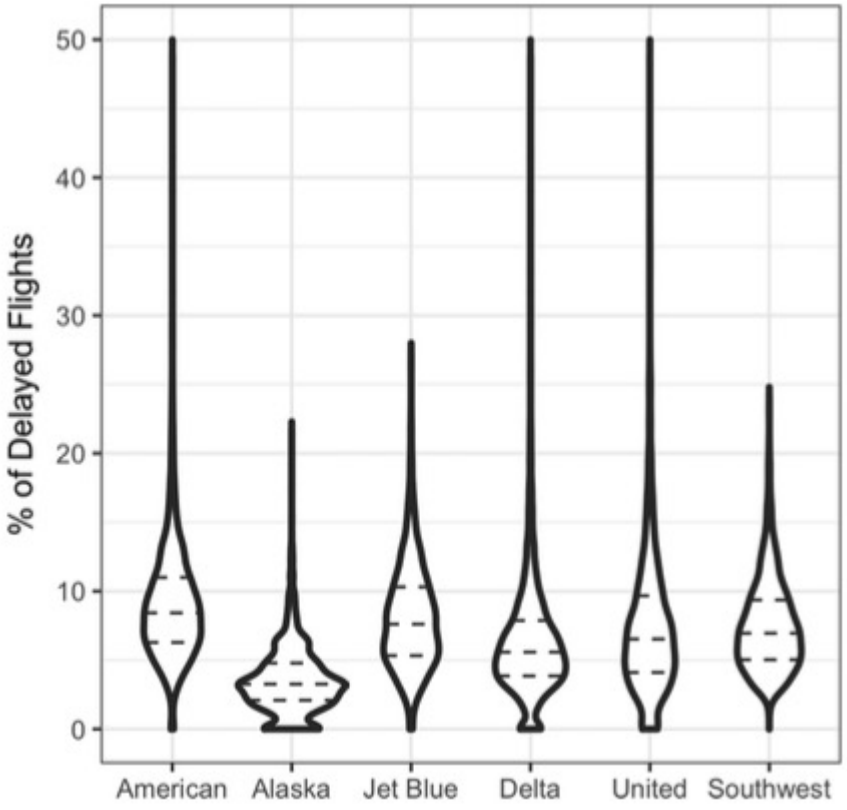


- Hexagonal binning (left) and corresponding contour plot (right)

Distributions of a numeric variable grouped by a categorical variable

- Violin plot is an enhancement to the boxplot and plots the density estimate with the density on the y-axis

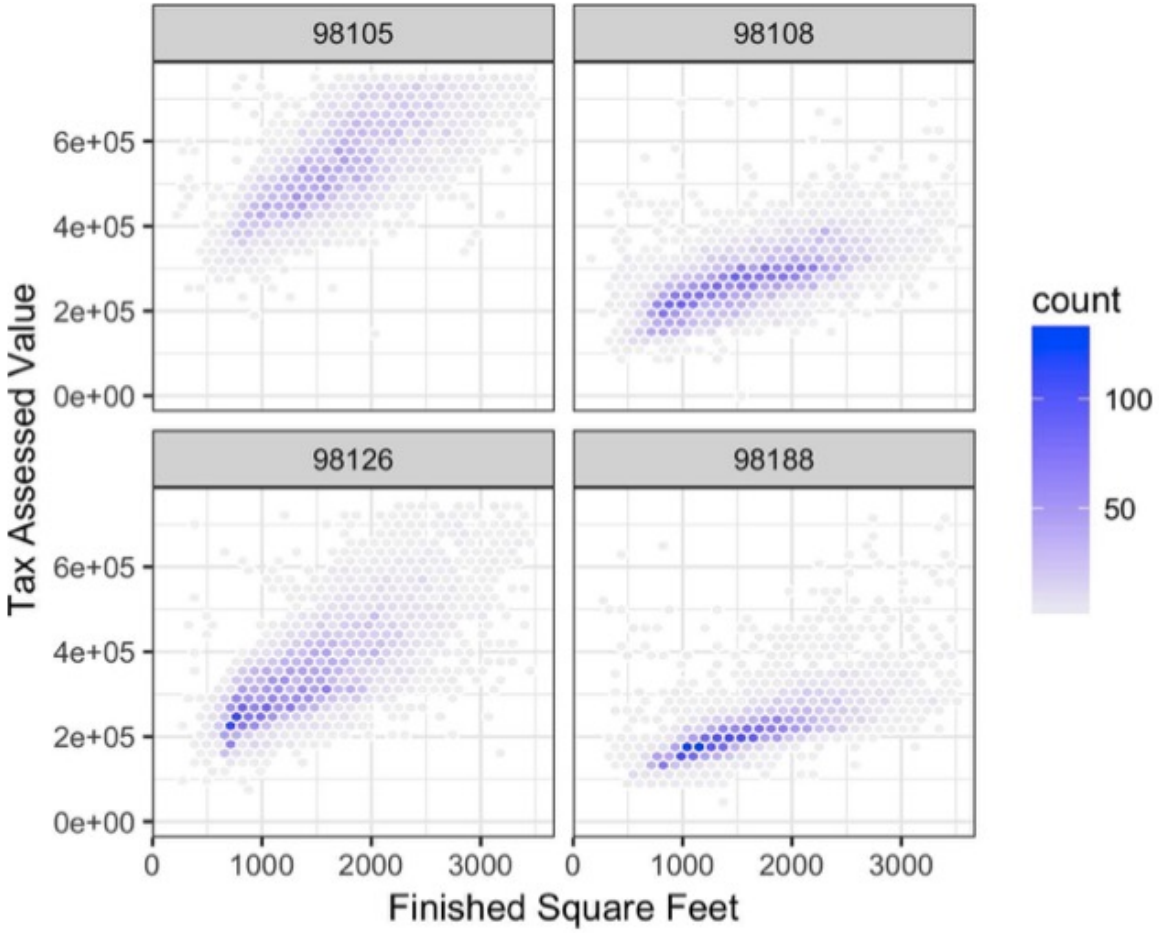
Example: percentage of flight delays varies across airlines



Visualizing multiple variables

- Charts comparing two variables are readily extended to more variables through the notion of conditioning

Example: relationship between homes' finished square feet and tax-assessed values by zip codes (conditioning variable)



Summary

- Structured vs. unstructured data
- Numeric vs. categorical variables
- Mean, variation, distribution of data
- Binary and categorical data
- Correlation
- Multivariate analysis