Data Exploration Part 1

Lesson 1

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Inter Quartile Range
and
Probability Density Functions

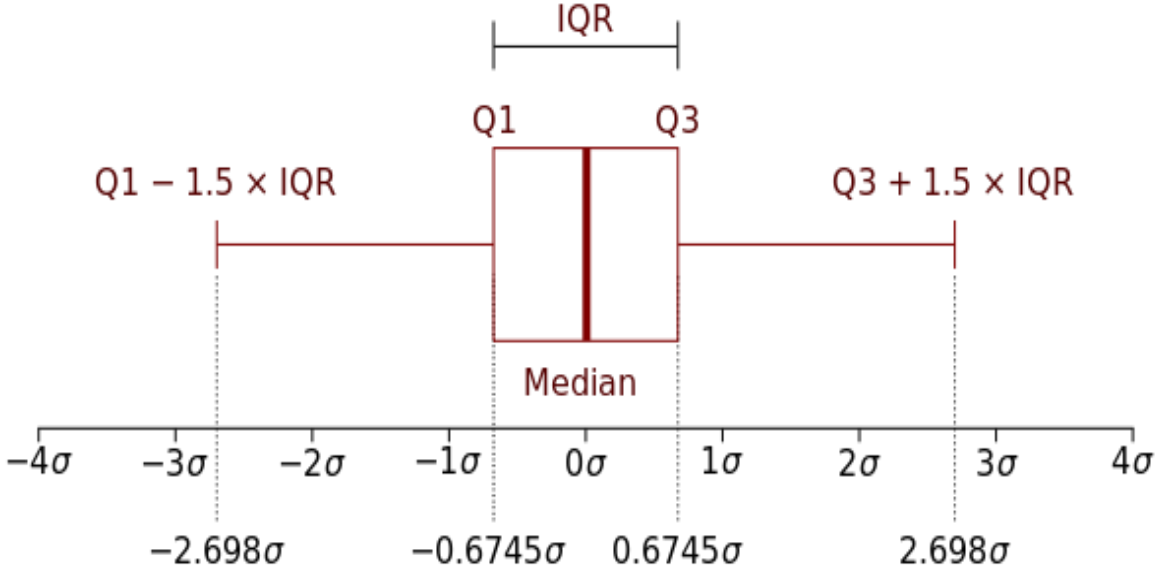
Data Exploration (Descriptive Statistics)

- What is it?
  - First look at your data
  - Summary Statistics

- Purpose: To gain a clear understanding of your data
  - What are the dimensions?
  - What columns are of interest?
  - Missing data?
  - Outliers?
  - Patterns?
  - Need to reformat?
  - Data types

Inter Quartile Range (Q3 – Q1)

- "Middle 50%" = 75% - 25th percentile
- Measures variability
- Identifies outliers
  - below Q1 − 1.5 IQR or above Q3 + 1.5 IQR



Cumulative Distribution Function

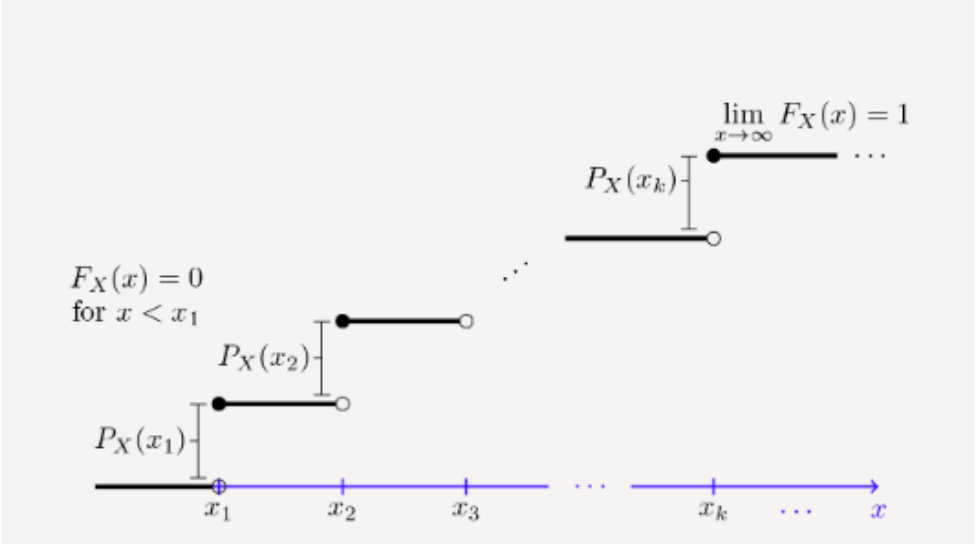
Probability that some random variable X will be less than or equal to a certain value:

- Probability, so 0 < x < 1
- Continuous and discrete variables
- PMF can only be used on discrete
  - Takes as input x, returns vector from [0,1] of probabilities "p"
  - Form of a staircase
- Jumps at each x(k)

F(x)=P(X≤x)

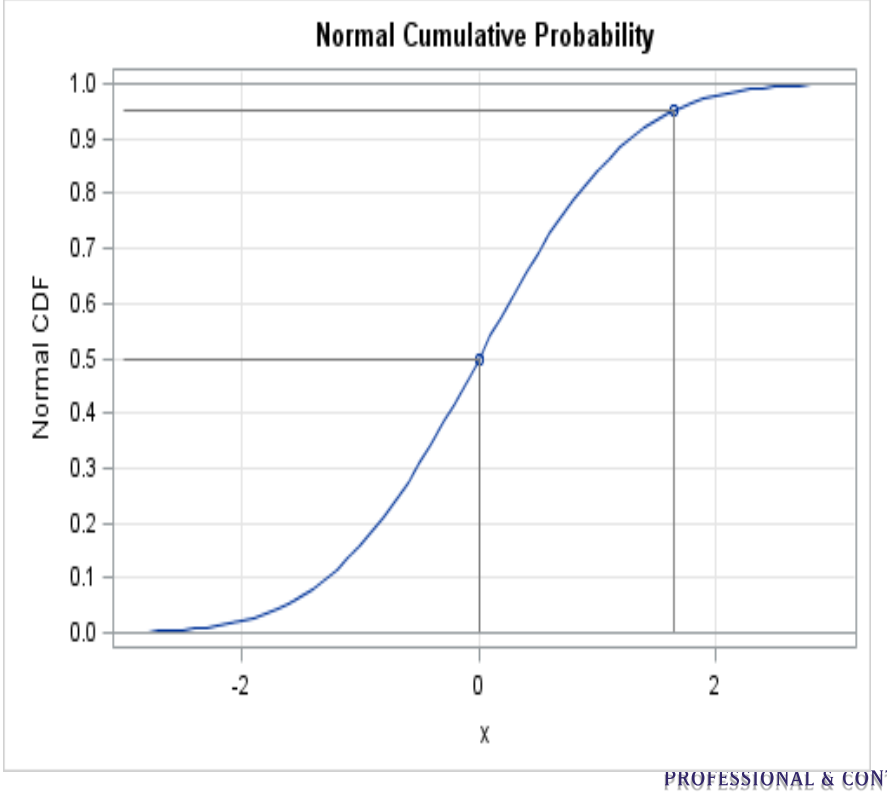
F(x)=P(X≤x)

F(x)=P(X≤x)

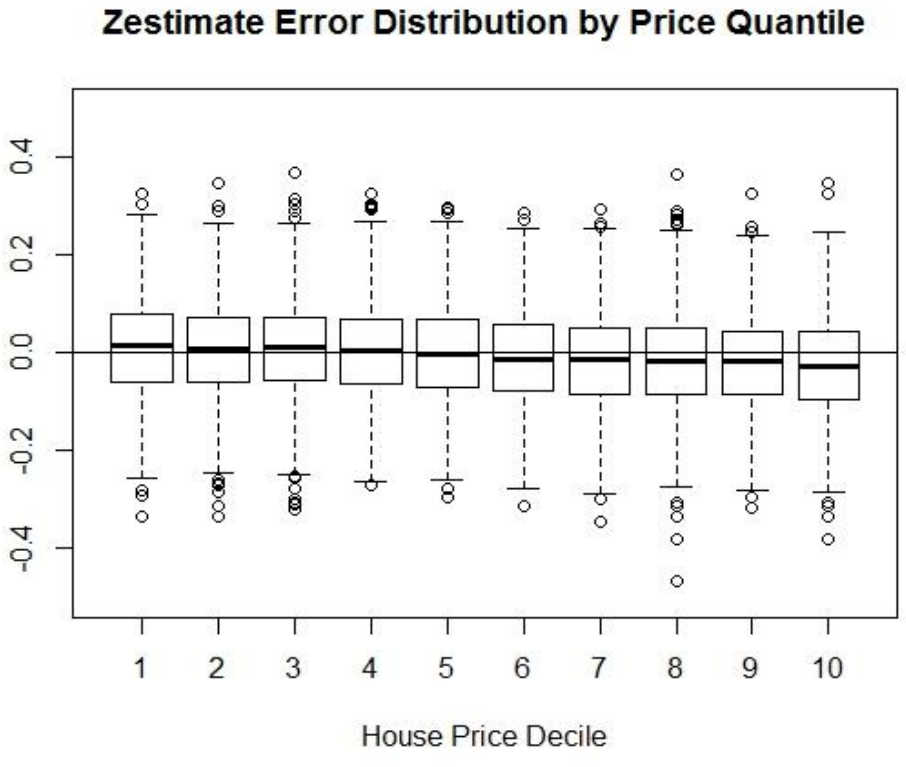


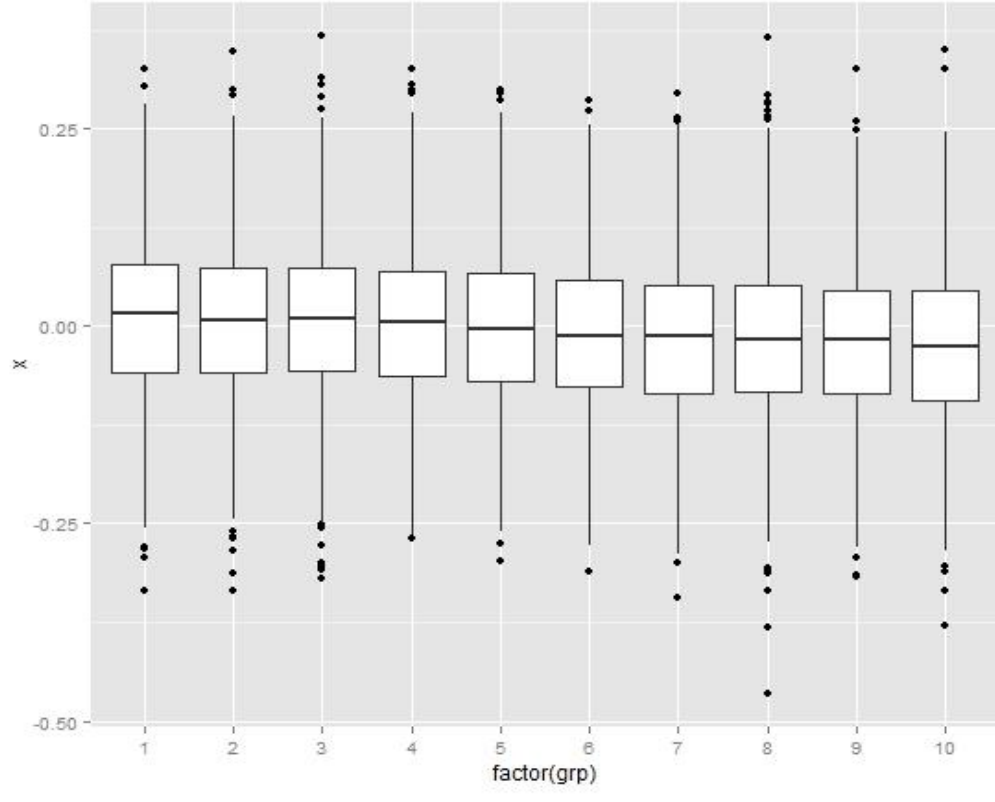
Quantiles of Numerical Vectors

- Quantiles are inverse values of the CDF (cumulative distribution function).
- Inverse tells you what value of x would make F(x) return a value "p"
- Standard Normal: (shown in figure)
  - Quantile(0.5) = 0, means at x=0, 50% of the distribution lies to the left. (This is also the median)
  - Quantile(0.95) = 1.65

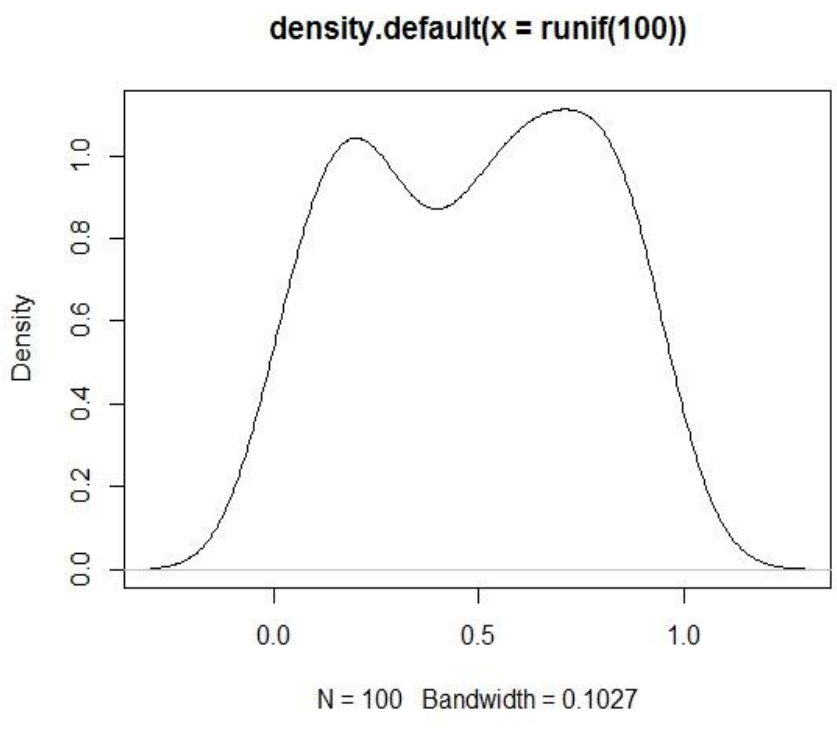


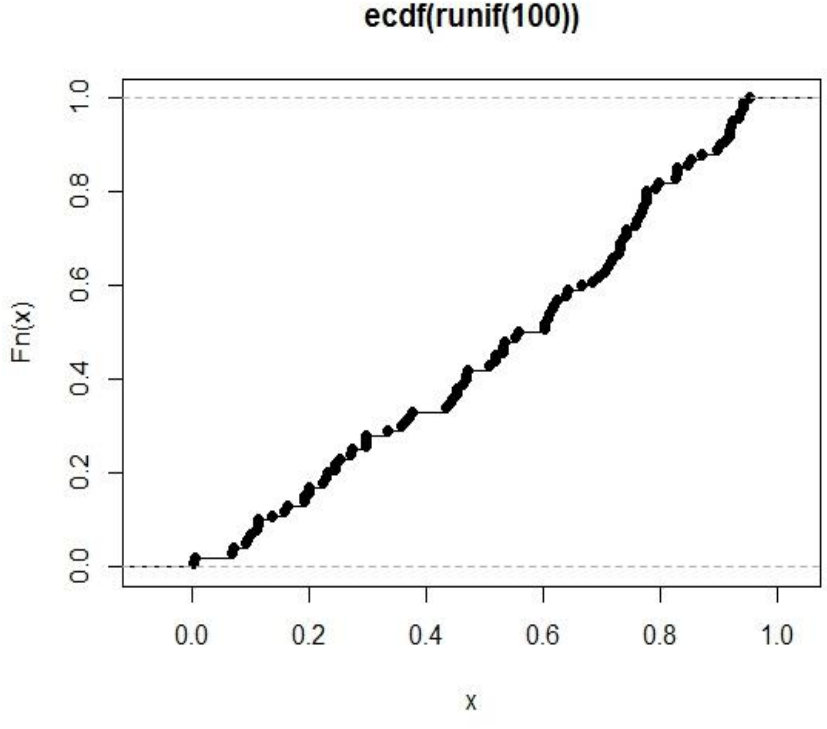
Visualizing IQR: Boxplots





Visualizing Densities/CDFs





Covariance

- Expected value of the differences between x and y and their corresponding mean.
- E.g. if x is above it's mean when y is also above it's mean, then they will have a high covariance.
- Highly interpretable, but not bounded.
- Measures strength and direction of relationship

$$Cov(X,Y) = \frac{\sum (X_i - \bar{X}) * (Y_i - \bar{Y})}{n}$$

Xi = some element in the sample X
Xbar = sample mean for x
N = number of elements in both samples

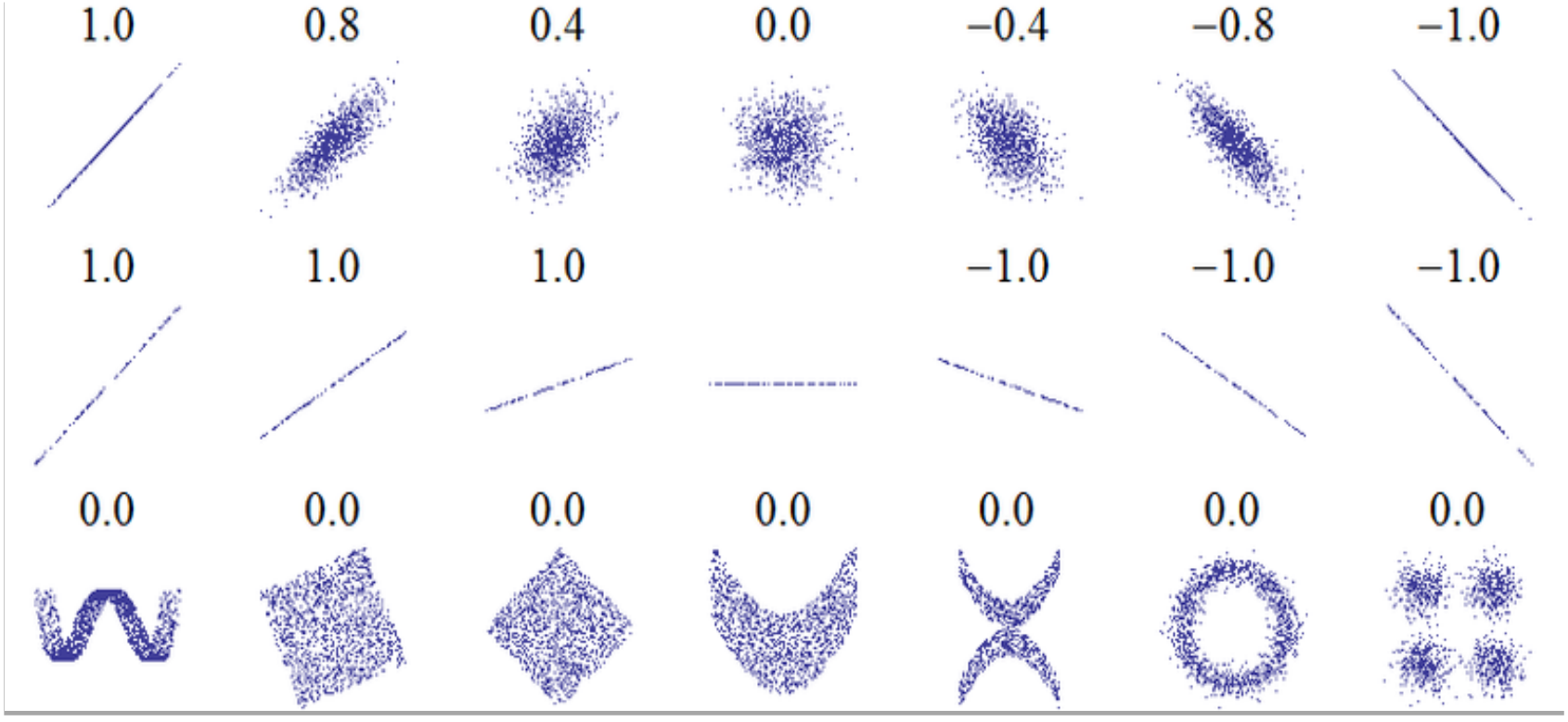
Correlation

- Correlations (pearsons) = scaled covariance
  - Bounded between 0 and 1.
  - Not as interpretable.

$$r = r_{xy} = \frac{\mathrm{Cov}(x, y)}{S_x \times S_y}$$

Sx = std dev

Visualizing Relationships: Scatterplots



Frequency: Counts

- Numerical and categorical variables
- Number of occurrences for an event in a fixed period
  - Ex. Number of times a gene is expressed after a medical treatment
- Modeled using Poisson distribution
  - Assume events are random and uniformly distributed

Poisson Distribution Formula

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where

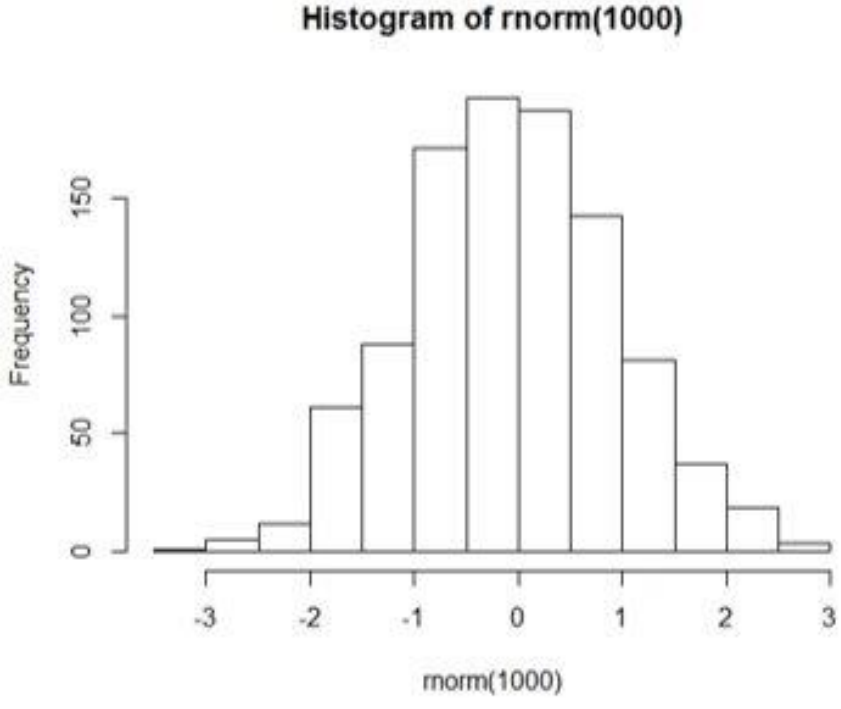
x = 0, 1, 2, 3, ...

λ = mean number of occurrences in the interval

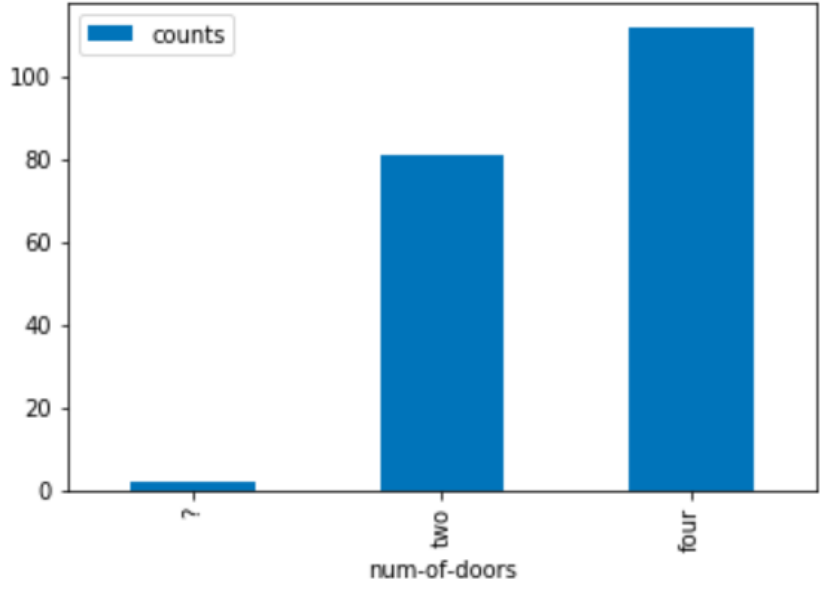
e = Euler's constant $\approx$ 2.71828

Visualizing Counts

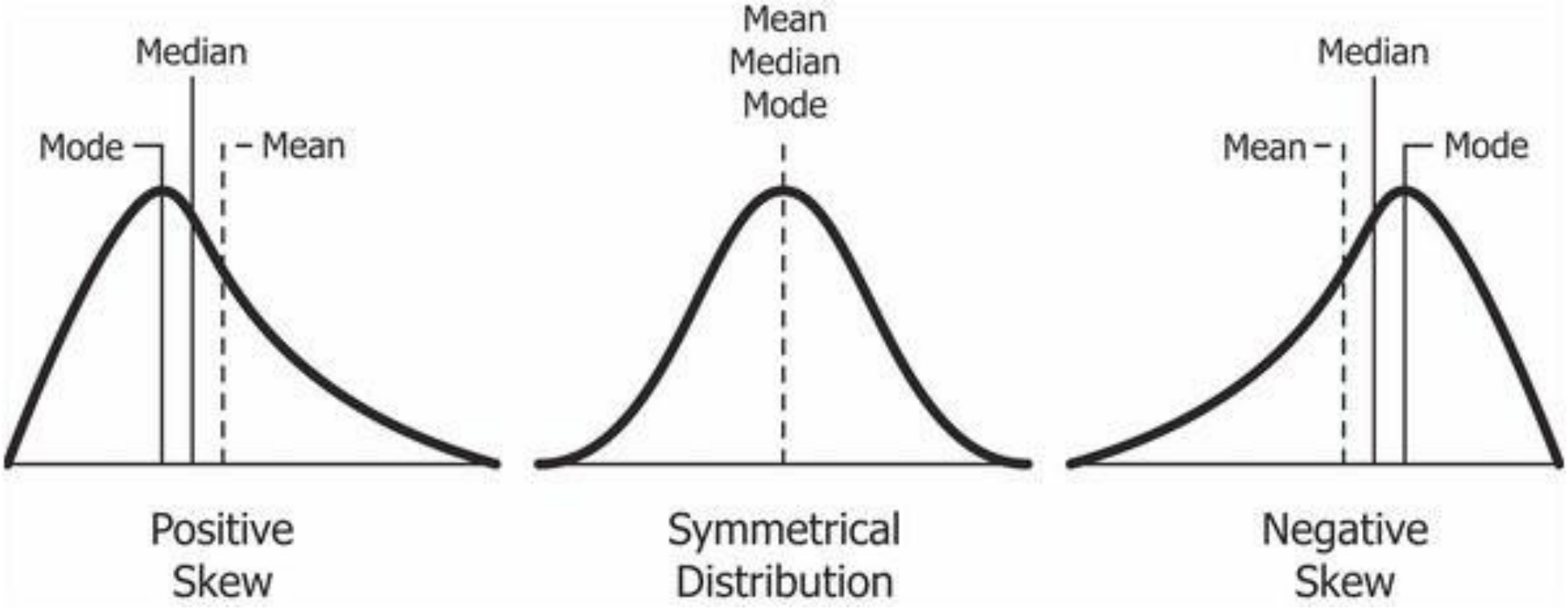
Histogram:
Number of values in bin



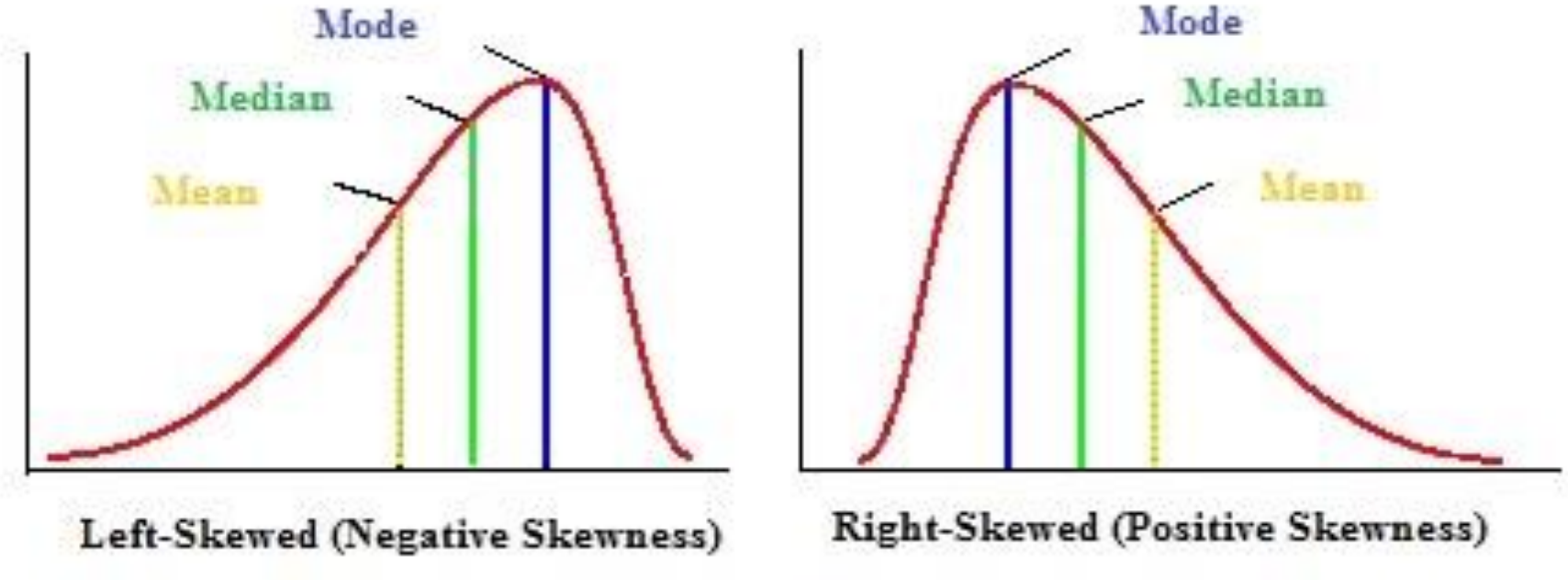
Bar Plot:
Count of Categorical Variables



Skew


Median
Mode
Mean
Positive Skew
Mean
Median
Mode
Symmetrical Distribution
Median
Mean
Mode
Negative Skew

Skew


Mode
Median
Mean
Left-Skewed (Negative Skewness)
Mode
Median
Mean
Right-Skewed (Positive Skewness)

Data Exploration Lab

Introduce Homework

https://canvas.uw.edu/courses/1247402/assignments/4548604?module_item_id=8995174