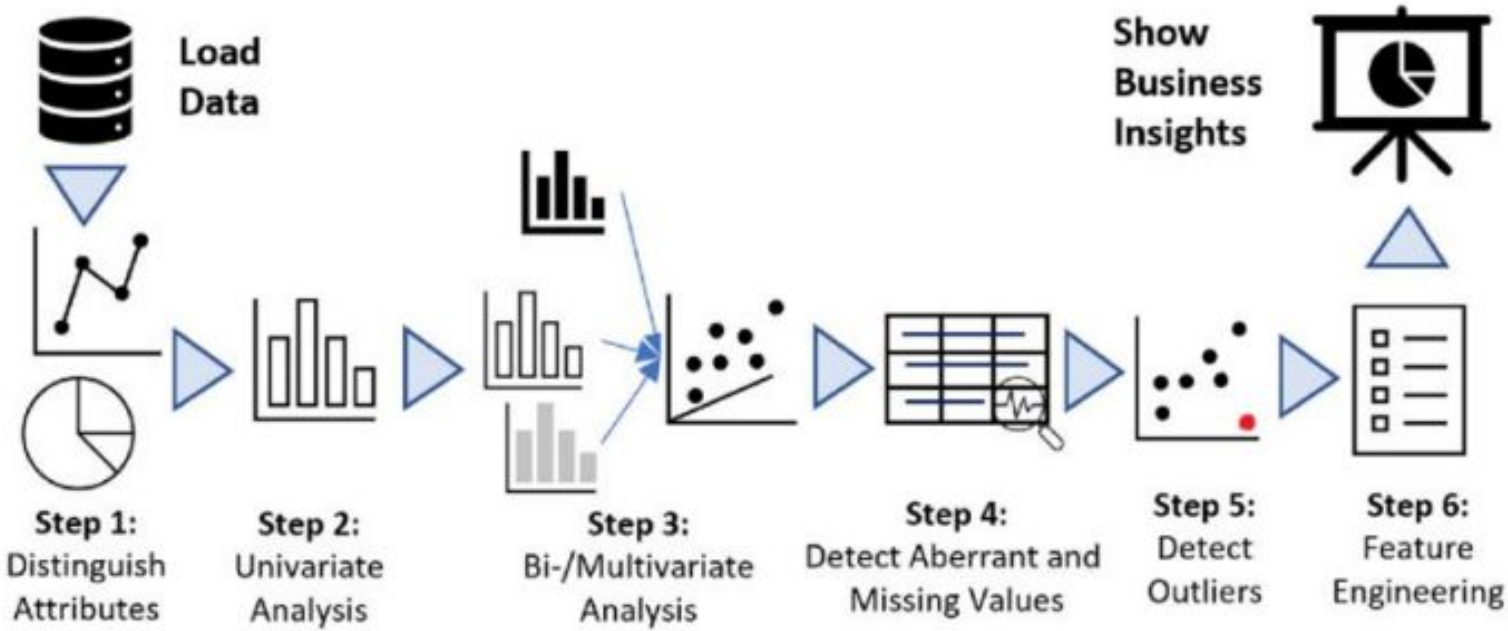


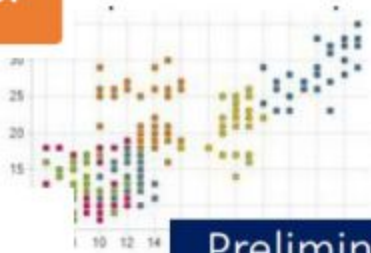

# Exploratory Data Analysis (EDA)

https://www.researchgate.net/figure/Exploratory-Data-Analysis-EDA-steps_fig2_342282008

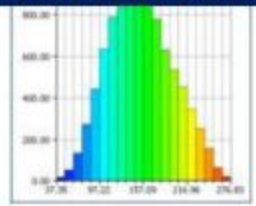
# Data Exploration

- Not always sure what we are looking for(until we find it)
  - Does not mean we need to find a needle in a haystack.
  - We need to start with hypothesis (questions).
  - Questions can change. :Might find something didn't expect.

# Exploratory Data Analysis (EDA)

Understand nature of data



## Step 2-A: Explore

Preliminary analysis

## Step 2-B: Pre-process

Clean

Integrate

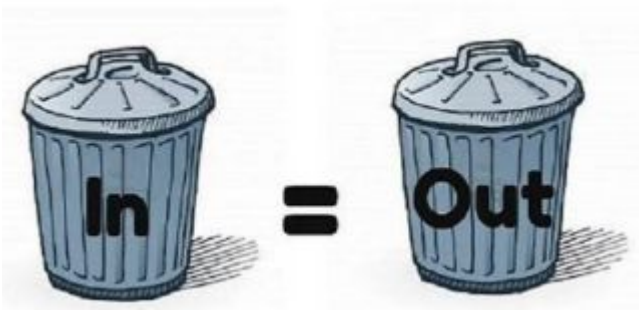
Package

# What is EDA?

- The analysis of datasets based on various numerical methods and graphical tools.
- Exploring data for patterns, trends, underlying structure, deviations from the trend, anomalies and strange structures.
- It facilitates discovering unexpected as well as confirming the expected.
- Another definition: An approach/philosophy for data analysis that employs a variety of techniques (mostly graphical).

# Why examine data?

- Why do we need to examine the data (or variables) before starting the analysis
  - Data tell stories
  - To catch mistakes
  - To see patterns in the data
  - To find violations of statistical assumptions
- Garbage in, garbage out
  - Bad data leads to bad models
  - Best algorithm cannot help if the data are too dirty
- Every measurement has 2 parts:
  - The True Score (the actual state of things in the world) and
  - ERROR! (mistakes, bad measurement, report bias, context effects, etc.)
  - Data = True X + error



# Aim of EDA

- Maximize insight into a data set
- Uncover underlying structure
- Extract important variables
- Detect outliers and anomalies
- Test underlying assumptions
- Develop parsimonious models
- Determine optimal parameter settings.

# What do we examine?

Before examining, we need to have some knowledge in the data domain

- Distributions (symmetric, normal, skewed, exponential etc..)
- Identify useful raw data & transforms (e.g. log(x))
- Data quality problems
- Outliers  (e.g. assess data quality)
- Correlations and inter-relationships
- Subsets of interest
- Suggest functional relationships

- Think interactive and visual:  Pandas build to support this!!
  - Humans are the best pattern recognizers

# Classification of EDA

- Exploratory data analysis is generally cross-classified in two ways. First, each method is either non-graphical or graphical. Second, each method is either univariate or multivariate (usually bivariate)
  - **Non-graphical methods** generally involve calculation of summary statistics, while **graphical methods** obviously summarize the data in a diagrammatic or pictorial way.
  - **Univariate methods** look at one variable (data column) at a time, while **multivariate** methods look at two or more variables at a time to explore relationships. Usually, our multivariate EDA will be biariate (looking at exactly two variables), but occasionally it will involve three or more variables
- It is always a good idea to perform univariate EDA on each of the components of a multivariate EDA before performing the multivariate EDA.

*Seltman, H.J. (2015). Experimental Design and Analysis. http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

# Basic EDA Workflow

- **Build** a DataFrame from the data (ideally, put all data in this object)
- **Clean** the DataFrame. It should have the following properties:
  - Each row describes a single object
  - Each column describes a property of that object
  - Columns are numeric whenever appropriate
  - Columns contain atomic properties that cannot be further decomposed
- Explore **global properties**. Use histograms, scatter plots, and aggregation functions to summarize the data.
- Explore **group properties**. Use groupby and small multiples to compare subsets of the data.

Pavlos Protopapas, Kevin Rader, Rahul Dave, Margo Levine. CS 109A/STAT 121A/AC 209A. Harvard University, 2017.

# EDA Questions

- Some common questions that exploratory data analysis is used to answer are:
  - What's the data/data type?
  - Are there outliers in the data?
  - What are the central tendency (mean, median, mode)?
  - What are spread/skewness/kurtosis?
  - Are there any relationships among features?

# Data Dictionary Example

- A Data Dictionary is a collection of names, definitions, and attributes about data elements that are being used or captured in a database
- A Data Dictionary also provides metadata about data elements

| Variable | Description | Variable type | Data type |
|---|---|---|---|
| Area | Total area including parking lot | Continuous | float |
| Floor Type | 1 - Tile, 2 - Concrete | Categorical | String |
| Number of Floor | Including basement | Ordinal | Integer |
| Number of lot | Count Space inside | Discrete | integer |

# Data Types and Measurement Scales

- Variables may be one of several types, and have a defined set of valid values.
- Two main classes of variables are:
  - **Categorical Variables: (Discrete, qualitative)**
    - Nominal or ordinal variables
  - **Continuous Variables: (Quantitative, numeric)**
    - Continuous data can be rounded or binned to create categorical data.
    - Ex. Height, Weight

# Categorical Data Summary

- Best way to examine categorical variables is by checking their frequencies
- Make assumption about the data and examine it. For example:
  - Occurance of each categories equally?
  - All text are spelling correctly
  - No duplicate field with the same info
    - Field with male vs. field with 0
  - Contain all English words
  - All images are grayscale

# pandas.Categorical

```
class pandas.Categorical(values, categories=None, ordered=None,
dtype=None, fastpath=_NoDefault.no_default, copy=True)
```

[source]

- Represent a categorical variable
- Categoricals can only take on a limited, and usually fixed, number of possible values (categories). In contrast to statistical categorical variables, a Categorical might have an order, but numerical operations (additions, divisions) are not possible.
- All values of the Categorical are either in categories or np.nan. Assigning values outside of categories will raise a ValueError. Order is defined by the order of the categories
- Fast and memory efficient

```
>>> pd.Categorical(['a', 'b', 'c', 'a', 'b', 'c'])
['a', 'b', 'c', 'a', 'b', 'c']
Categories (3, object): ['a', 'b', 'c']
```

```
>>> c = pd.Categorical([1, 2, 3, 1, 2, 3, np.nan])
>>> c
[1, 2, 3, 1, 2, 3, NaN]
Categories (3, int64): [1, 2, 3]
```

Missing values are not included as a category.

```
>>> c = pd.Categorical(['a', 'b', 'c', 'a', 'b', 'c'], ordered=True,
...                    categories=['c', 'b', 'a'])
>>> c
['a', 'b', 'c', 'a', 'b', 'c']
Categories (3, object): ['c' < 'b' < 'a']
>>> c.min()
'c'
```

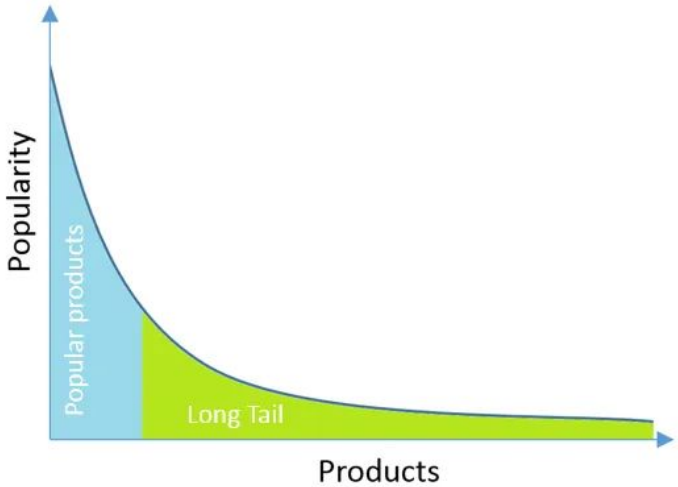
Ordered Categoricals can be sorted according to the custom order of the categories and can have a min and max value.

# Numerical Variable Summaries

- **Range**:
  - Max, Min
- **Central tendency:**
  - Mean – the average value
  - Median – the middle value
  - Mode – the most frequent value
- **Variability:**
  - Variance – the spread around the mean
  - Standard deviation : Standard error of the mean (estimate)
- Means and variances are ways to describe a distribution of the numerical data.
- Distributions is one of the best ways to understand the data

# What are we looking for?

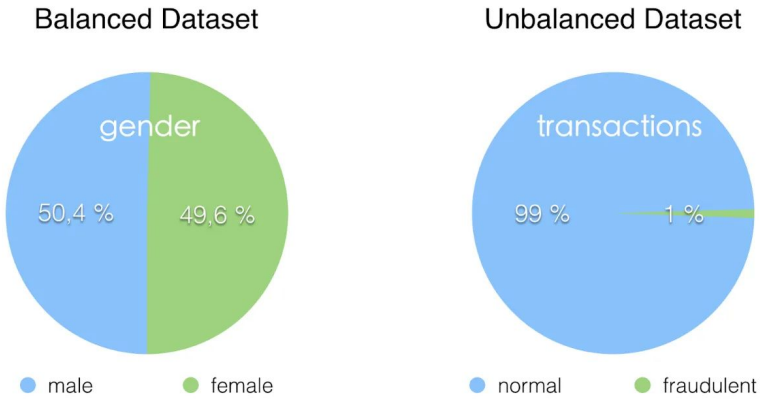
- Data in real life are not usable
  - Missing data
  - Mistake:  e.g. age = 210,  misspelling text
  - Mismatch in unit, scale: e.g. 2543 vs 2000,  1/1/2540
  - Require merging or combine filed
    - Building 1 has meter number 0, 1, 2, 3
    - Building 2 has meter number 0,1,2,...
- Imbalance data
- Unusual distribution
  - Recommendation movies:  long tail
- Unusable data
  - Emoji in text




a **long tail** of some distributions of numbers is the portion of the distribution having many occurrences far from the "head" or central part of the distribution.
https://en.wikipedia.org/wiki/Long_tail

# Example of imbalance data

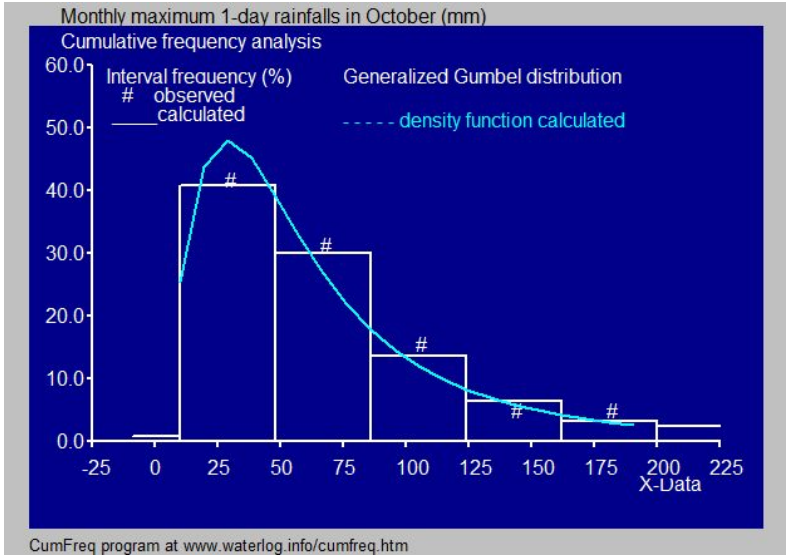
- A situation, primarily in classification machine learning, where one target class represents a significant proportion of observations\
- Imbalanced datasets are those where there is a severe skew in the class distribution, such as 1:100 or 1:1000 examples in the minority class to the majority class.
- **Example**: fraud detection, disease screening, churn, predictive maintenance
- Approaches:
  - Upsampling the minority class
  - Undersampling the majority class
  - Change the loss function
  - Combinations of the above methods




https://medium.com/strands-tech-corner/unbalanced-datasets-what-to-do-144e0552d9cd

# Exploration of 1 variable

- The techniques such as mean, mode, max, min are common to summarize from 1 variable data
- Histogram is an approximate representation of the distribution of numerical data
- Distribution is visual representation of continuous variable data
- There are different distribution types
  - Uniform
  - Normal
  - Poisson
  - Exponential




Src: https://en.wikipedia.org/wiki/Histogram

# Data distribution

- A data distribution is a function or a listing which shows all the possible values (or intervals) of the data. It show how often each value occurs.
- Density functions are functions that describe how the proportion of data or likelihood of the proportion of observations change over the range of the distribution, commonly refer to continuous random variable
  - Probability Density function: calculates the probability of observing a given value.
  - Cumulative Density function: calculates the probability of an observation equal or less than a value.
- Probability mass function (pmf) is a function that gives the probability that a discrete random variable is exactly equal to some value.

# Normal distribution (Medium-tailed)

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

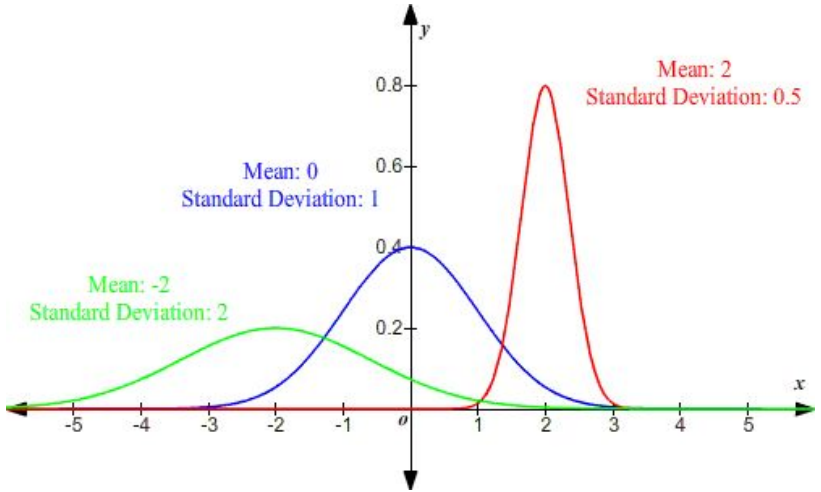
- A Normal (aka Gaussian) distribution is the most common assumption of statistics, thus it is often important to check if your data are normally distributed.
  - Law of large number (With enough measurements (data), most variables are distributed normally)
- Standard deviation (SD) (sqrt variance) shows the distribution of scores around the mean.
  - High SDs (relative to the mean) indicate the scores are spread out
  - Low SDs tell you that most scores are very near the mean.

$\mu =$ Mean

$\sigma =$ Standard Deviation

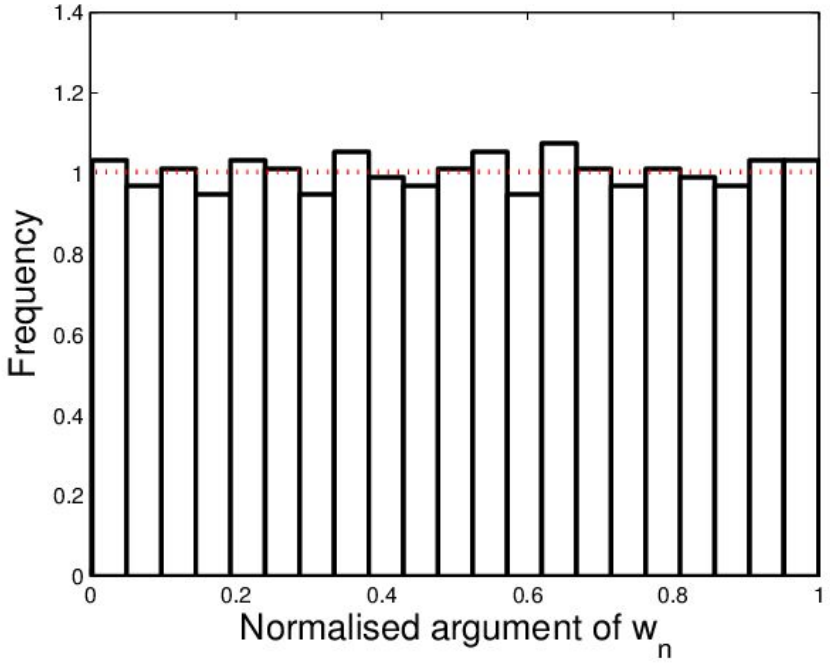
$\pi \approx 3.14159\cdots$

$e \approx 2.71828\cdots$




https://www.varsitytutors.com/hotmath/hotmath_help/topics/normal-distribution-of-data

# Uniform distribution (Thin-tailed)




- Roll a dice
- Lottery/gambling

The probability density function of the continuous uniform distribution is

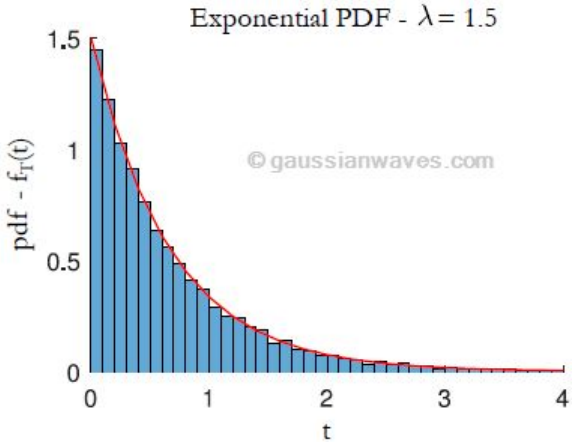
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{for } x < a \text{ or } x > b. \end{cases}$$

Src: https://www.researchgate.net/figure/Histogram-of-argwn-p-2p-vs-uniform-density-on-0-1_fig2_265915054

# Exponential distribution (fat-tailed)

$X$ is said to have an **exponential distribution** with parameter $\lambda(>0)$ if its pdf is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$




- Represent time between events: equipment failure, process/message arrivals
- Memoryless property: $P(X > a+b \mid X > b) = P(X > a)$
- You're waiting for a bus. You have already waited for 30 minutes, but the bus has not arrived. What's the probability that you will have to wait additional 10 minutes?

# Simulation

- The process of using a computer to mimic a physical experiment (lifetime of equipment, customer arrival, user preference).
- In this class, those experiments will almost invariably involve chance.
- Simulation process
  a. What to simulate (
  b. Simulating one value
  c. Number of repetitions
  d. Simulating multiple values
     - Create an empty array in which to collect all the simulated values
     - Create a “repetitions sequence (n repetitions with np.arange(n)
     - Create a loop

# Simulation example

1. What to simulate: equipment lifetime according to the exponential distribution
2. Simulating one value: `x = np.random.exponential(1/Lambda)`
3. Number of repetitions: NN = 1000
4. Simulating multiple values
   a. Collect results:

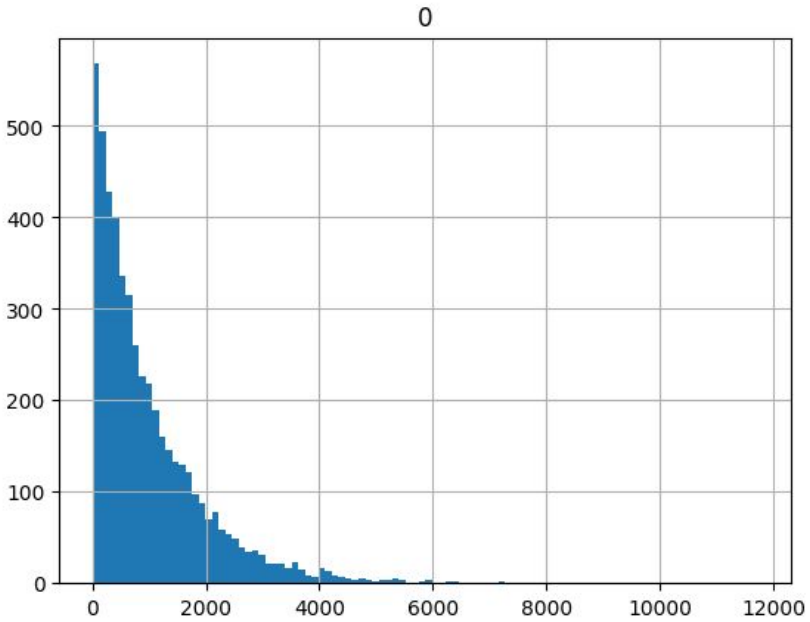
```
R_interval[i] = ss - TT[0]
I_interval[i] = x
A_interval[i] = TT[0] - (ss - x)
X_inverval[i] = ss
```

```
for i in range(NN):
  ss = 0
  TT = -1000*np.log(1-np.random.rand(1)) #Arrival time of the observer

  while (ss<TT):
    x = np.random.exponential(1/Lambda) #Exponential RV
    ss = ss + x
    count += 1

  R_interval[i] = ss - TT[0]
  I_interval[i] = x
  A_interval[i] = TT[0] - (ss - x)
  X_inverval[i] = ss
```

# Simulation example

```
xx = pd.DataFrame(X_inverval)
xx.hist(bins = 100)
```

Kurtosis = 8.36
Skewness = 2.17



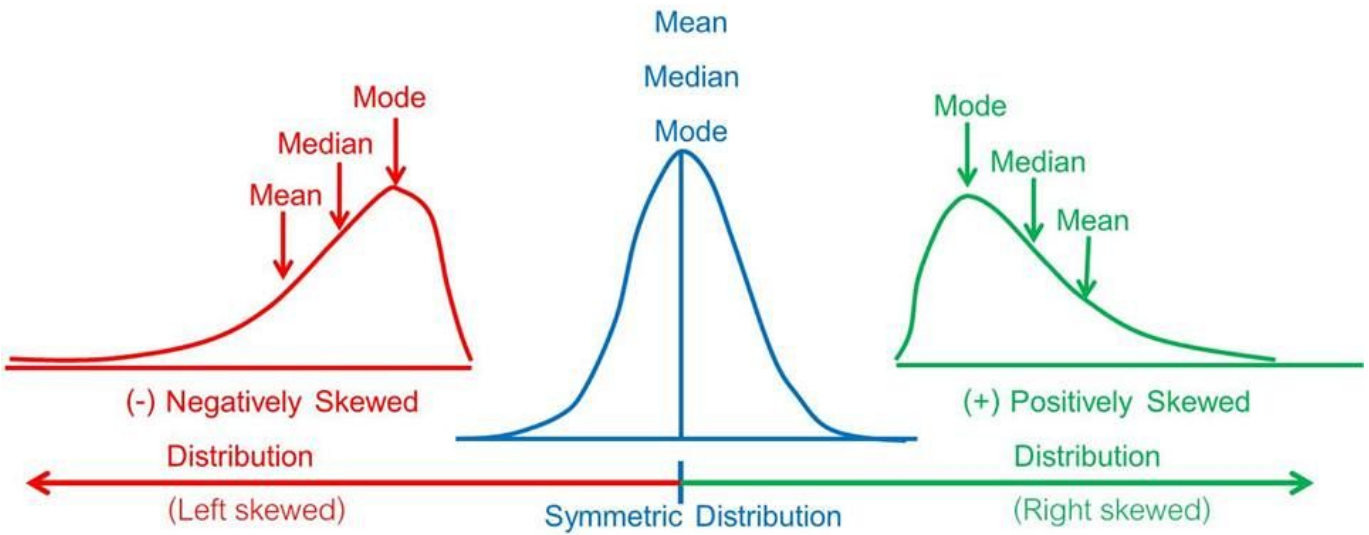
# Equipment lifetime

A company orders an equipment whose average lifetime is 1 hour. A manager arrives everyday to check the equipment and expect to wait 30 minutes before the equipment breaks down. As the manager waits, he carefully records the equipment lifetime and finds that the average wait time is longer than 30 minutes. He tells his colleagues that this indicates that the lifetime of the equipment is longer than 1 hour, but data from the technician who monitors the equipment lifetime do not give such indications. Other than that the manager may have had some bad luck, is there a logical explanation?

**Hint**: the lifetime is average 60 minutes. So, to simplify things, assume that the intervals alternate between 40 and 80 minutes.

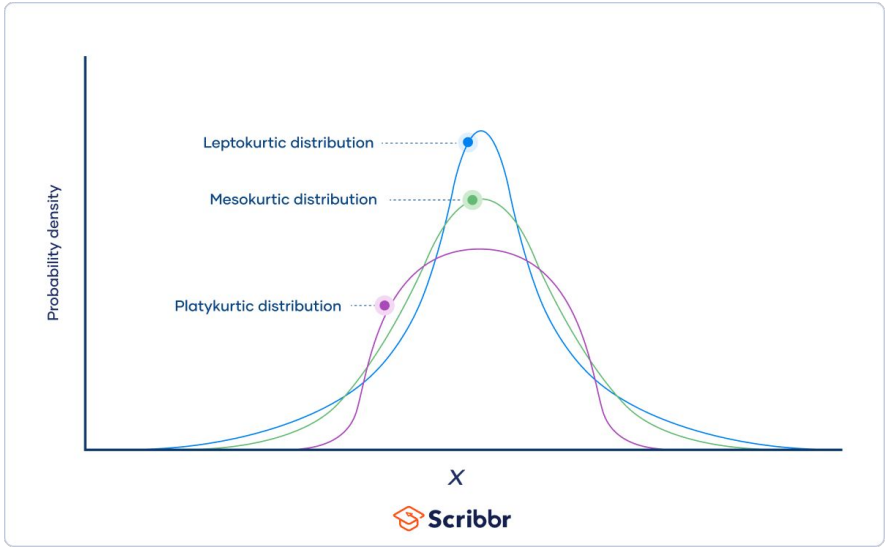
# Skewness

- Skewness is a measure of the lack of symmetry. A distribution is symmetric if it looks the same to the left and right of the center point.



# Kurtosis

- Kurtosis is a measure of tailedness (heavy-tailed or light-tailed relative to a normal distribution)
  - high kurtosis (leptokurtic) (>3) tend to have heavy tails, or outliers (e.g. Exponential distribution)
  - low kurtosis (platykurtic ) (<3) tend to have light tails, or lack of outliers
  - A uniform distribution would be the extreme case of platykurtic (no outlier).
  - Mesokurtic (~3): Normal distribution

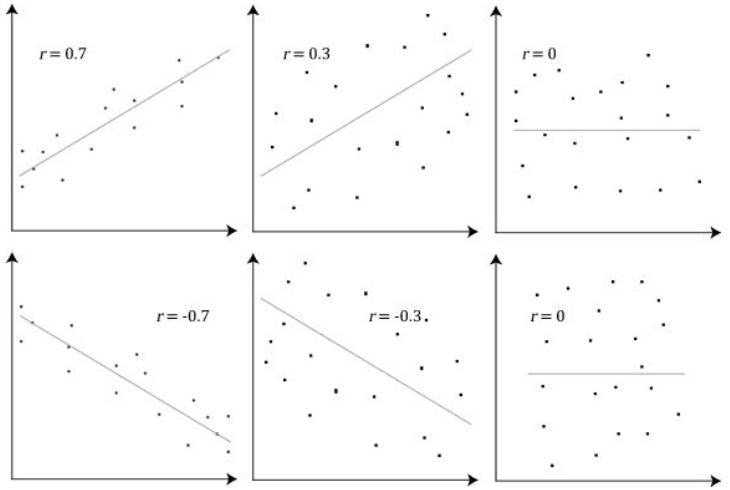





Src: https://www.researchgate.net/figure/The-distribution-patterns-Skewness-and-Kurtosis_fig1_253332732

# Two Continuous Variables

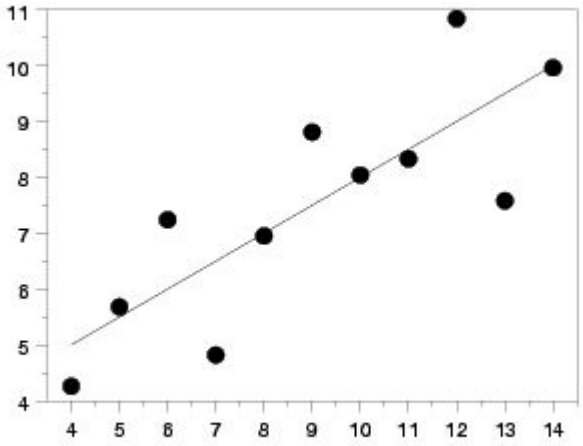
- For two numeric variables, the scatterplot is the obvious choice
  - standard tool to display relation between 2 variables
    - e.g. y-axis = response, x-axis = suspected indicator
  - useful to answer:
  - Any correlation between x,y?
    - Linear
    - Quadratic
    - other
  - outliers present?
  - You can use more than 2 dimensions!
  - x,y,z, space, color, time….



# Exploration of 2 variables

- Relationship between variables can be useful
- Correlation is any statistical relationship, whether causal or not, between two random variables or bivariate data

| Simple Stats |
| --- |
| N = 11<br>Mean of X = 9.0<br>Mean of Y = 7.5<br>Intercept = 3<br>Slope = 0.5<br>Correlation = 0.816 |




| X | Y |
| --- | --- |
| 10.00 | 8.04 |
| 8.00 | 6.95 |
| 13.00 | 7.58 |
| 9.00 | 8.81 |
| 11.00 | 8.33 |
| 14.00 | 9.96 |
| 6.00 | 7.24 |
| 4.00 | 4.26 |
| 12.00 | 10.84 |
| 7.00 | 4.82 |
| 5.00 | 5.68 |

| Observation |
| --- |
| - a linear curve with some scatter;<br>- there are no outliers;<br>- the vertical spread of the data appears to be of equal height |

# Summarizing Variables

- Categorical variables
  - Frequency tables - how many observations in each category?
  - Relative frequency table - percent in each category.
  - Bar chart and other plots.
- Continuous variables
  - Bin the observations (create categories .e.g., (0-10), (11-20), etc.) then, treat as ordered categorical.
  - Plots specific to Continuous variables.

The goal for both categorical and continuous data is data reduction while preserving/extracting key information about the process under investigation.

# EDA Methods

*Descriptive statistics*

Central tendency of a distribution

| | |
|---|---|
| Mean | $\frac{\text{Sum of all values}}{\text{Total number of values}}$ |
| Median | Middle value(when data are arranged in order |
| Mode | Most common value |

Measure of Variation

| | |
|---|---|
| **Variance** | **how far a set of numbers are spread out from mean** |
| Interquartile range | divides a data set into quartiles. |
| Standard deviation | dispersion of a set of data from mean |

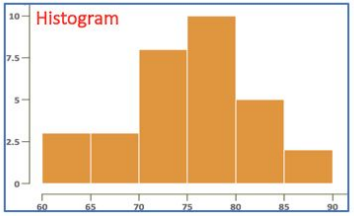
Skewness & Kurtosis

| | |
|---|---|
| **Skewness** | **Measure of symmetry** |
| Kurtosis | **Kurtosis is a measure of "peakedness" relative to a Gaussian shape** |

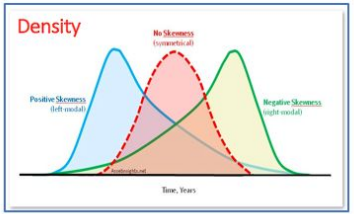
Visualizations

1-dimension

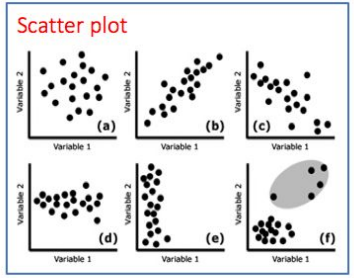
*Few data points*



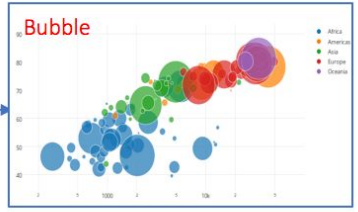

*Many data points*




2-dimension




3-dimension



# References


- Exploratory Data Analysis (Tukey, 1997)
- Visualizing Data (Cleveland, 1993)
- Visual Display of Quantitative Information (Tufte, 1983)
- https://www.itl.nist.gov/div898/handbook/eda/eda.htm
- https://web.archive.org/web/20180720062852id_/http://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf
- https://pandas.pydata.org/