# Prosit : Descriptive statistics

### Animateur :

Elio RENZETTI

### Scribe :

Meriem ASSAM

### Gestionnaire :

Mohammed BENSALEM

### Secrétaire :

Joyline CHEPKOECH

## Key words:

* Exhaustive - including or considering all elements or aspects; fully comprehensive.
* Bike surfing – name of the company founded by CESI students
* Self service bicycles - users can rent a bicycle for a single trip, without having to return it to the initial terminal.
* Competitors - an organization or country engaged in commercial or economic competition with others.
* Virtual sensor network - a group of sensors that work together over the internet or a network, even if they are not physically connected or in the same location. These sensors can share information and coordinate their data, allowing them to measure things more efficiently as a team.
* CSV file - a text file that uses commas to separate values, and newlines to separate records and allows data to be saved in a table structured format.
* Monitoring data - observe and check the progress or quality of data over a period of time; keep under systematic review.
* Sensing mobility – perceiving motion

**Context :**

CESI students aim to launch Bikesurfing, a self-service bike rental company inspired by a successful Washington D.C. mode, With limited operating days due to city restrictions, they must choose the most profitable days to ensure success and convince city officials of its feasibility.

## Constraints :

* CSV files
* Limited data
* Limited operating days
* High competition

## Problem statement :

* To ensure the success of their business, they must choose the optimal days for operation and present a compelling case to city officials, demonstrating the feasibility and potential of their project.

## Delivrables :

* Deliver a convincing (exhaustive) schedule
* Open day

## Solution approaches:

* Regression
* Descriptive statistics
* Variable correlation
* XLS/Python

## Action Plan:

1. Study descriptive statistics
2. Analyse given data and visualize
3. Determine the data correlation
4. Detect anomalies
5. Regression
6. Prepare an exhaustive schedule/report

**PERSONAL NOTES** a t l

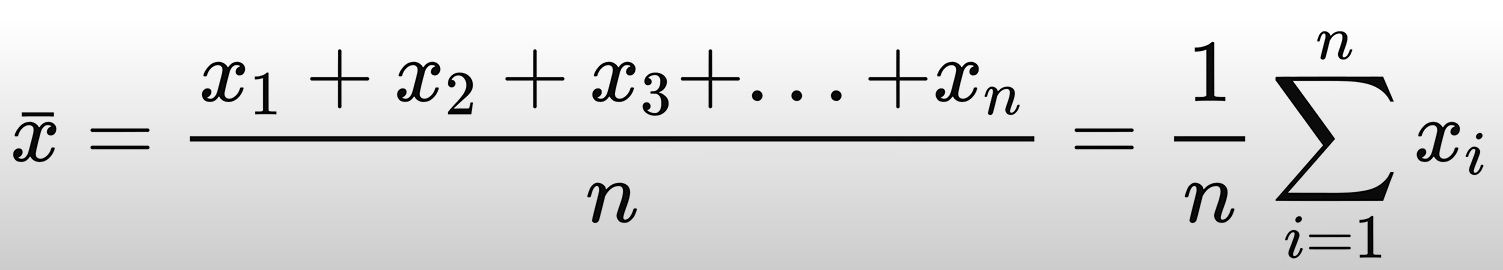
**DESCRIPTIVE STATISTICS**

Descriptive statistics describes the collected data without drawing conclusion about the population; in order to draw conclusions from a sample to population, we will use inferential statistics.

#Note – Regression is the bridge between descriptive and inferential statistics

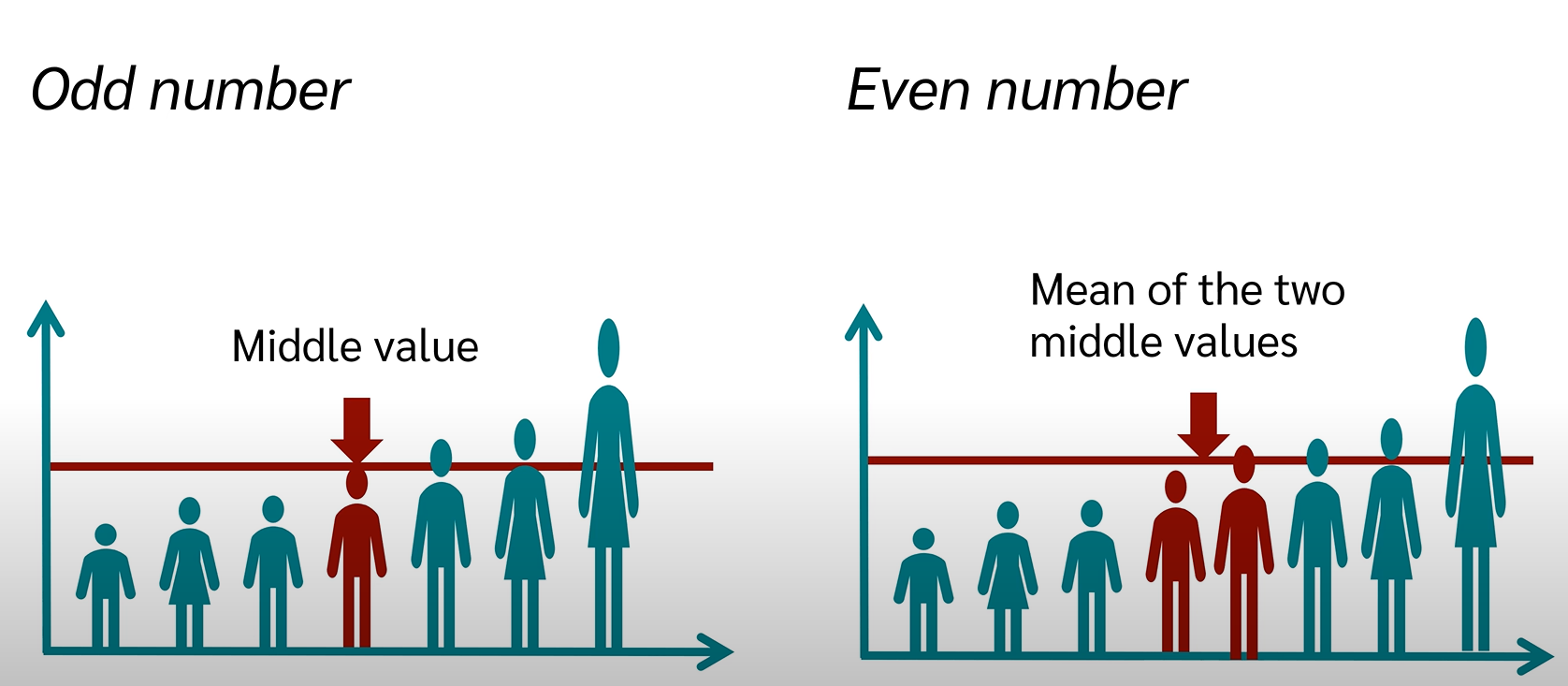
**KEY TERMS**

* Measures of Central Tendency
  + Arithmetic Mean – sum of observations divided by the total number of observations



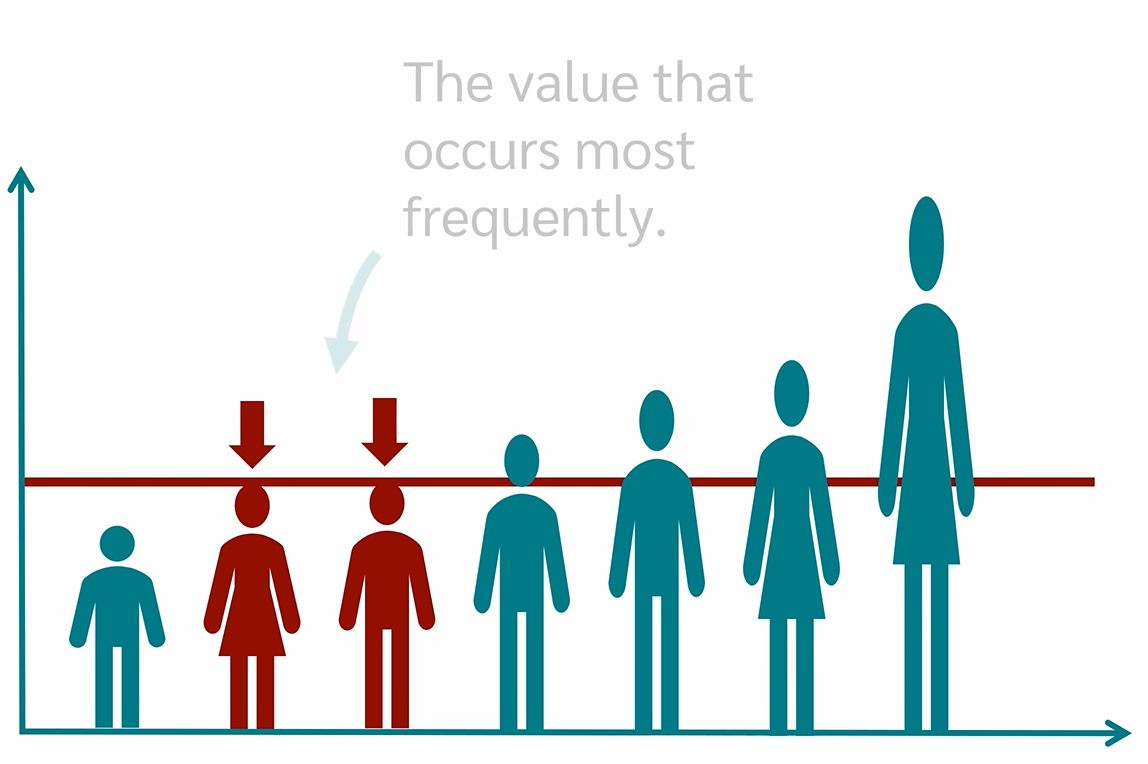
Not resistant to outliers

* + Median – if elements are arranged in ascending order, then

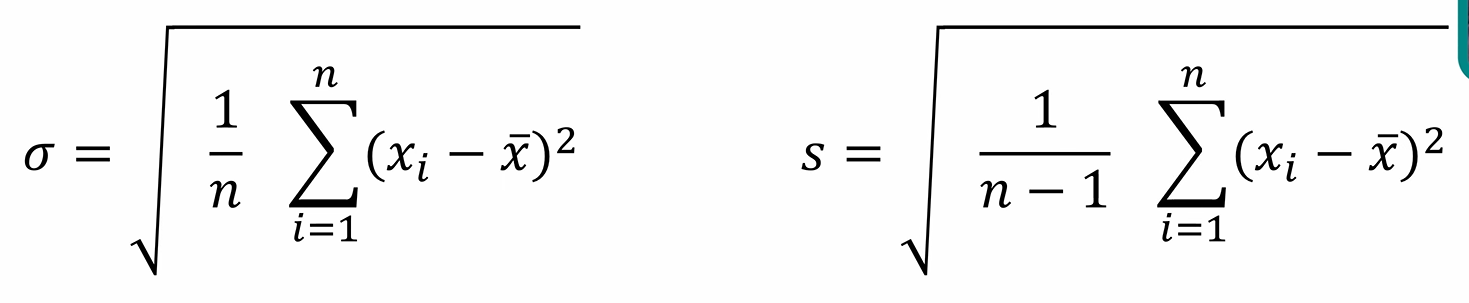


It is resistant to extreme values or outliers

* + Mode – the value(s) that occur most frequently



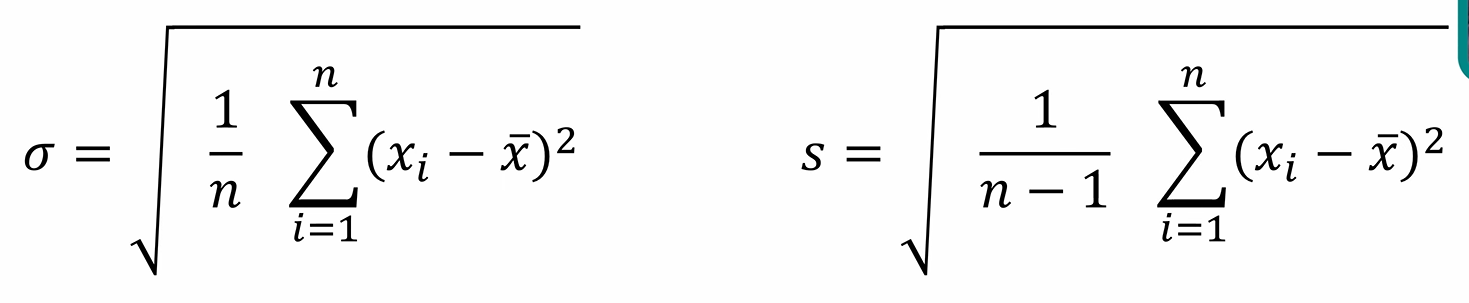
* Measures of Dispersion – describes how spread out the values in a dataset are
  + Standard Deviation – quadratic distance of the distance from the mean





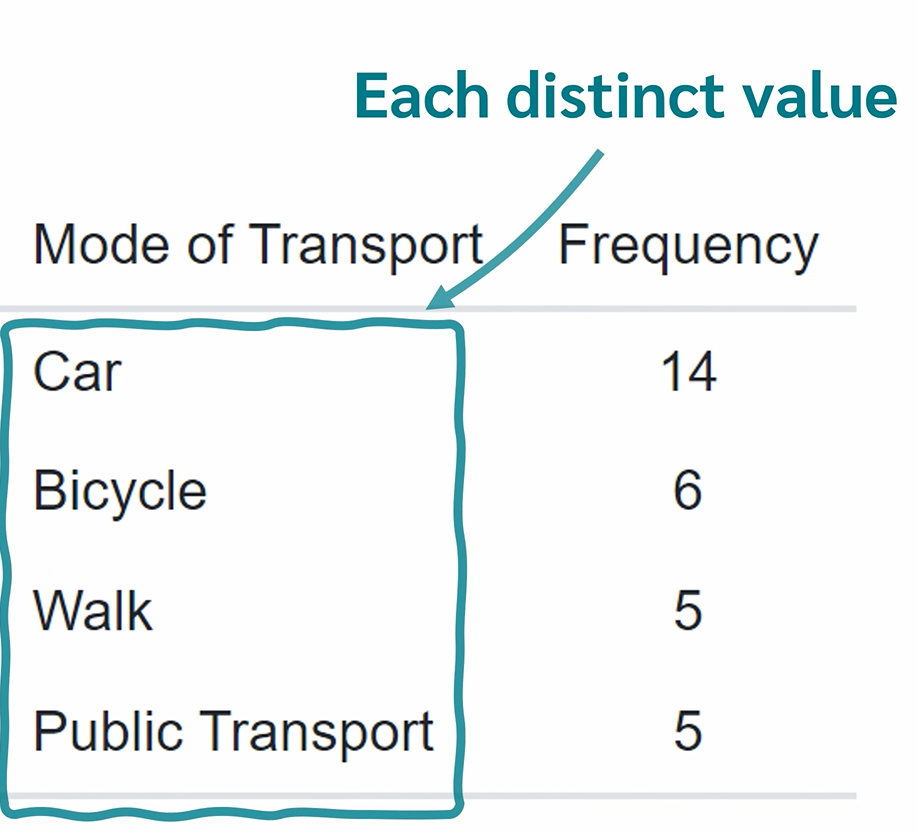
We will use this for our prosit as our survey does not cover the whole population

* + Variance – squared standard deviation

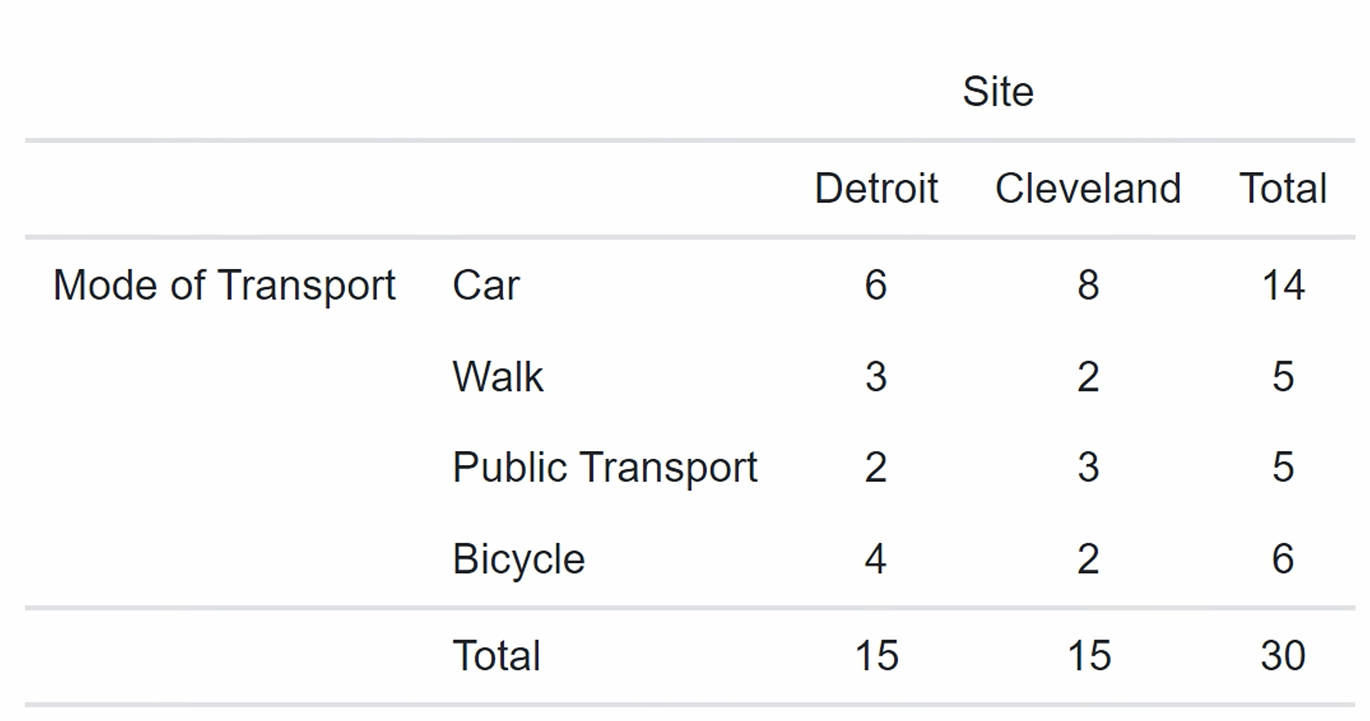




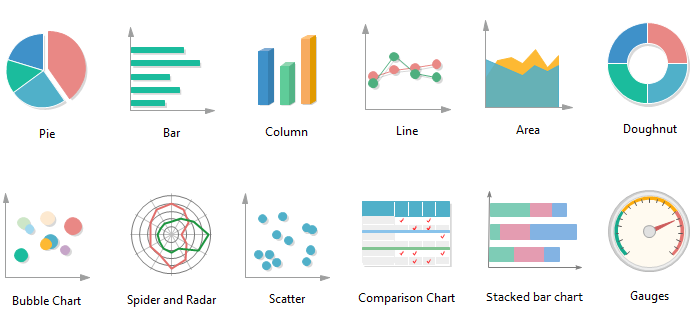
* + Range – difference between min and max value
  + Interquartile Range – represents middle 50% of the data, i.e. difference between the third quartile Q3 and the first quartile Q1
* Table
  + Frequency



* + Contigency

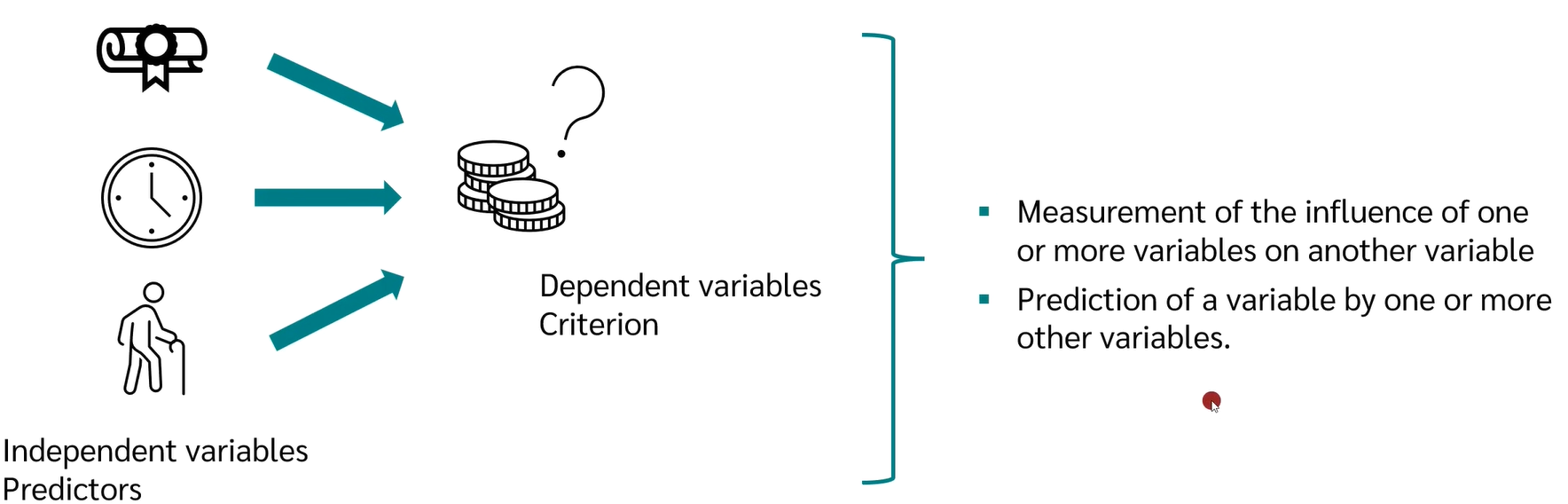


* Charts



**REGRESSION ANALYSIS**

Regression analysis makes it possible to infer or predict a variable on the basis of one or more other variables



* Simple Linear – one independent variable, dependent variable is metric
  + Usually showed by a scatter plot
* Multiple Linear – multiple independent variables, dependent variable is metric
* Logistic – categorical [nominal/ordinal] dependent variable

**RANDOM PROCESS**

* Flipping a coin
* Rolling a dice
* Measuring the rain that might fall tomorrow

**RANDOM VARIABLE**

Mapping outcomes of random processes to numbers, i.e., quantifying the outcomes



* Discrete random variables
  + Distinct or separate values
* Continuous random variables
  + Any value in an interval (can be infinite)

**PROBABILITY DISTRIBUTION**

Mathematical function that can be thought of as providing the probabilities of occurrence of different possible outcomes in an experiment. For instance, if the random variable X is used to denote the outcome of a coin toss (“the experiment”), then the probability distribution of X would take the value 0.5 for X = heads, and 0.5 for X = tails (assuming the coin is fair). Probability distributions are divided into two classes –

* **Discrete Probability Distribution –** If the probabilities are defined on a discrete random variable, one which can only take a discrete set of values, then the distribution is said to be a discrete probability distribution. For example, the event of rolling a die can be represented by a discrete random variable with the probability distribution being such that each event has a probability of 1661​.
* **Continuous Probability Distribution –** If the probabilities are defined on a continuous random variable, one which can take any value between two numbers, then the distribution is said to be a continuous probability distribution. For example, the temperature throughout a given day can be represented by a continuous random variable and the corresponding probability distribution is said to be continuous.

**INTERPOLATION**

Estimate an unknown value or set of values that fall in between known data points.

**EXTRAPOLATION**

Estimates unknown values that extend beyond the known data.

**CORRELATION**

* Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).
* It’s a common tool for describing simple relationships without making a statement about cause and effect.
* The sample correlation coefficient, r, quantifies the strength of the relationship. Correlations are also tested for statistical significance.
* Correlation can’t look at the presence or effect of other variables outside of the two being explored.
* It doesn’t tell us about cause and effect.
* It also cannot accurately describe curvilinear relationships.

**SHAPE OF DISTRIBUTION**

When graphed, the data in a set is arranged to show how the points are distributed throughout the set. These distributions show the spread (dispersion, variability, scatter) of the data. The spread may be stretched (covering a wider range) or squeezed (covering a narrower range).

The shape of a distribution is described by its number of peaks and by its possession of symmetry, its tendency to skew, or its uniformity. (Distributions that are skewed have more points plotted on one side of the graph than on the other.)

**FREQUENCY OF DISTRIBUTION**

* A frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.
* The frequency is how often a value occurs in an interval, while the distribution is the pattern of frequency of the variable.
* Typically used within a statistical context when dealing with quantitative data, such as demographical or population data collected by a census, or financial figures in the context of stock trading.
* Generally, frequency distributions can be associated with the charting of a normal distribution.
* a frequency distribution is a representation that displays the number of observations within a given interval.
* The representation of a frequency distribution can be graphical or tabular.
* Frequency distributions are particularly useful for normal distributions, which show the observations of probabilities divided among standard deviations.

**PERCENTILES AND QUARTILES**

* Quartiles are values that separate the data into four equal parts.
* Percentiles are values that separate the data into 100 equal parts.

**3 SIGMA RULE**