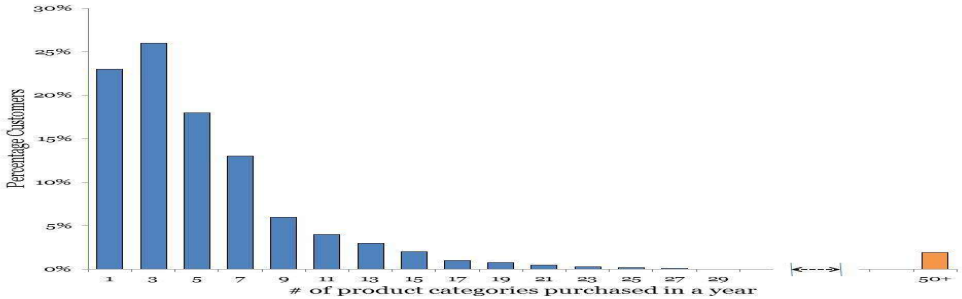
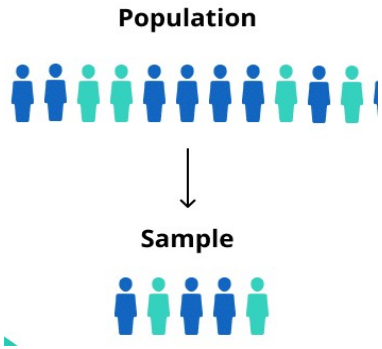
Exploratory Data Analysis

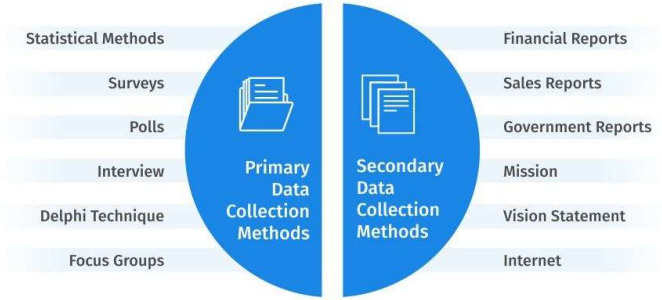
- **EDA: -** In statistics, exploratory data analysis is an approach of analyzing data sets to summarize their main characteristics, often using statistical graphics and other data visualization methods.



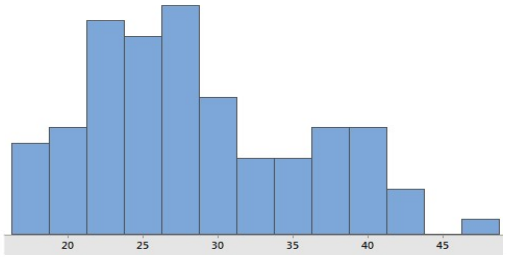
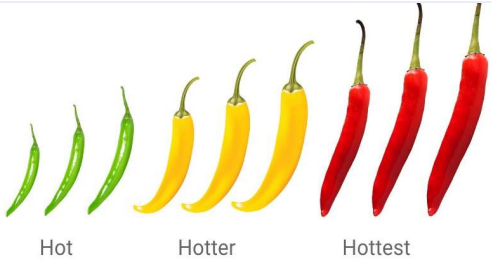
- **Data:** factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation
- **Population :** A **population** is the entire group that you want to draw conclusions about.
- **Sample** : A **sample** is the specific group that you will collect data from. The size of the sample is always less than the total size of the population.



- **Data Recourses :** sensor measurements, events, text, images, and videos. The *Internet of Things* (IoT) is spewing out streams of information.



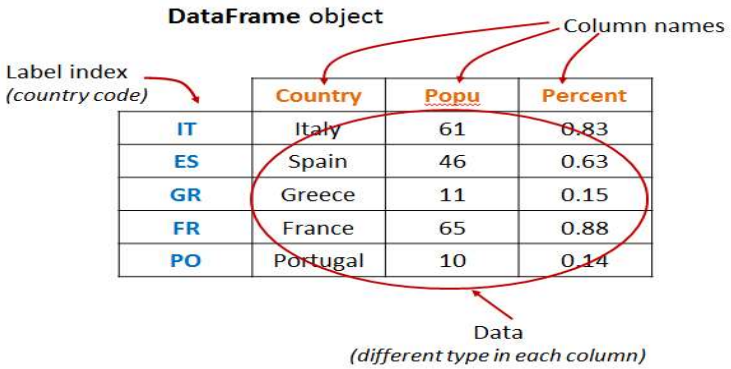
Types of Data

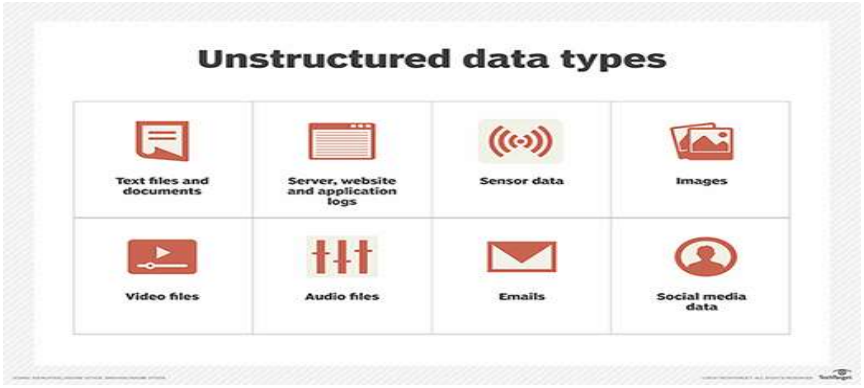
Numerical		Categorical	
Continuous	***Discrete***	***Binary***	***Ordinal***
• Continuous Data can take any value (within a range). • **Discrete data** is counted. Ex: A dog's weight, The length of a leaf, Lots more!	• Discrete Data can only take certain values. • **Continuous data** is measured Ex: Number of students in class, result of rolling dice	**Binary data** is **data** whose unit can take on only two possible states, traditionally labeled as 0 and 1 in accordance with the **binary** numeral system and Boolean algebra.	**Ordinal** data is a categorical, statistical data type where the variables have natural, ordered categories and the distances between the categories is not known.
			

Everything is data.

Data Structures

Rectangular	Non- rectangular
• Object like DataFrame, spreadsheet, database table. • Matrix of data **Terminology** **1. Feature:** Columns in table 2. Target : Output variable of data. 3. Records: Number of rows within the data. **Tools:** **Excel, Database table & DataFrame. (In python & R)**	• Data other than rectangular structure like text, image, spital data. • Graphs, Network diagram.





Central Tendency

an estimate of where most of the data is located (i.e., its central tendency).

Mean: sum of all values divided by the number of values.

Weighted Mean: sum of all values times a weight divided by the sum of the weights.

Median: value such that one-half of the data lies above and below

Percentile: value such that *P* percent of the data lies below

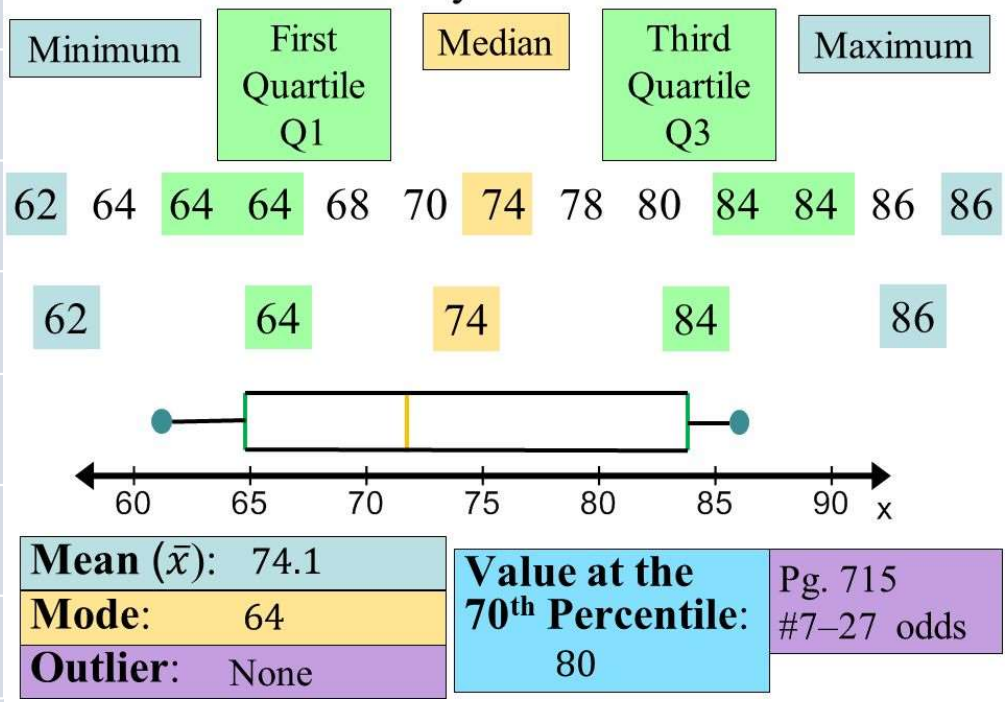
Weighted median: The value such that one-half of the sum of the weights lies above and below the sorted data

Trimmed mean: The average of all values after dropping a fixed number of extreme values.

Robust: Not sensitive to extreme values.

Outlier: A data value that is very different from most of the data.

Five-number Summary & Box-and-Whisker Plot



Mean ($\bar{x}$): 74.1

Mode: 64

Outlier: None

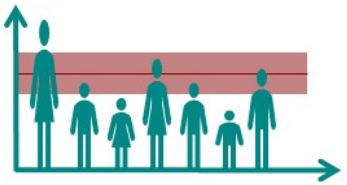
Value at the 70th Percentile: 80

Pg. 715 #7–27 odds

Data Variability

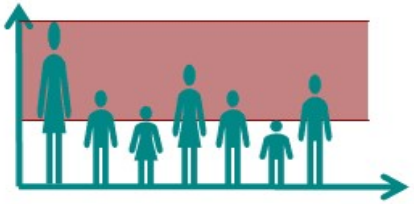
variability, also referred to as *dispersion*, measures whether the data values are tightly clustered or spread out.
Deviations: The difference between the observed values and the estimate of location.
Variance: The sum of squared deviations from the mean divided by $n - 1$ where n is the number of data values.
Standard deviation: The square root of the variance.
Mean absolute deviation: The mean of the absolute values of the deviations from the mean.
Median absolute deviation from the median:The median of the absolute values of the deviations from the median.
Range: The difference between the largest and the smallest value in a data set.
Order statistics: Metrics based on the data values sorted from smallest to biggest.
Percentile: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: The difference between the 75th percentile and the 25th percentile.

Standard deviation / variance



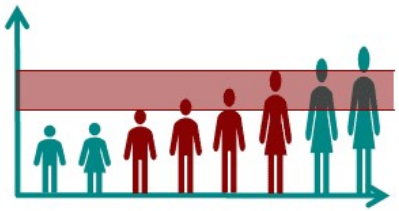
Average distance of all measured values from the mean value

Range



Distance between lowest and highest value of a distribution

Quantile distance

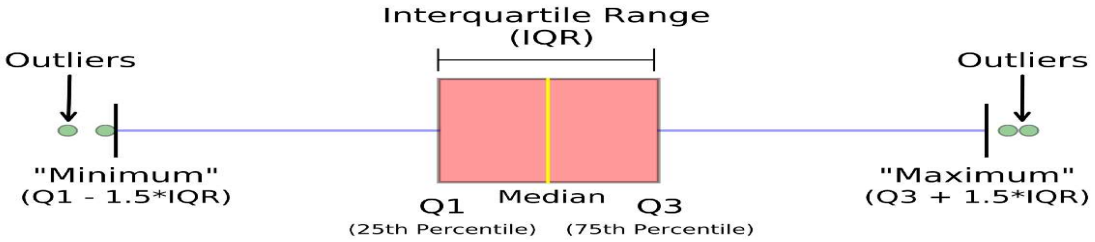


Spectrum in which the middle 50% of the values lie.

Difference between the first and the third quartile

Exploring Data Distribution

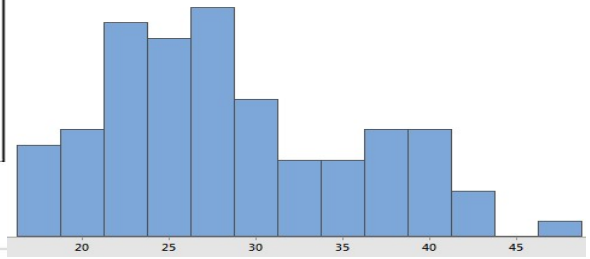
Box Plot: A plot introduced by Tukey as a quick way to visualize the distribution of data.



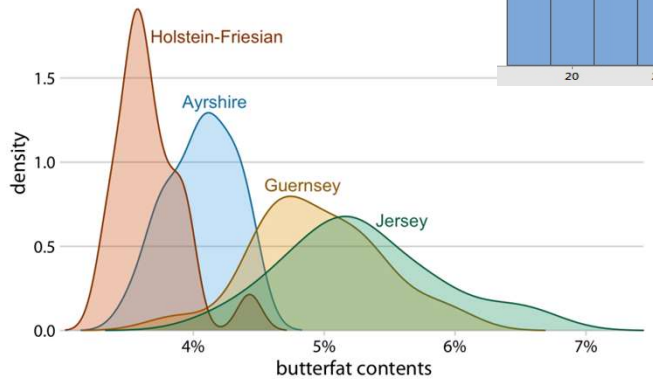
Frequency table: A tally of the count of numeric data values that fall into a set of intervals (bins).

Mark	Tally	Frequency
4	II	2
5	II	2
6	IIII	4
7	𝍸	5
8	IIII	4
9	II	2
10	I	1

Histogram A plot of the frequency table with the bins on the x-axis and the count (or proportion) on the y-axis. While visually similar, bar charts should not be confused with histograms.

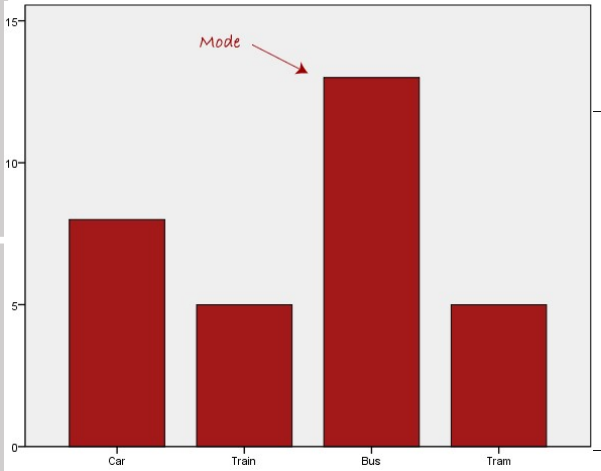


Density plot: A smoothed version of the histogram, often based on a *kernel density estimate*.

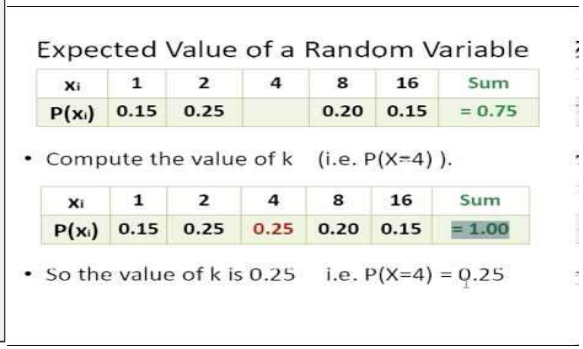


Binary & Categorical Data

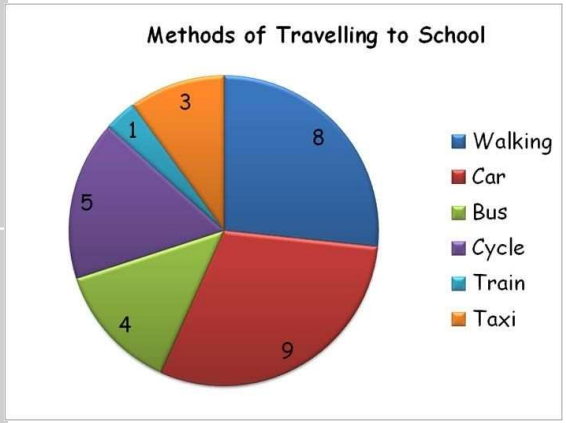
Mode: The most commonly occurring category or value in a data set.`



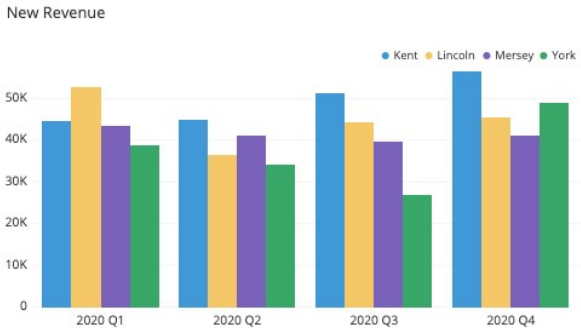
Expected value: When the categories can be associated with a numeric value, this gives an average value based on a category's probability of occurrence.



Pie charts: The frequency or proportion for each category plotted as wedges in a pie.

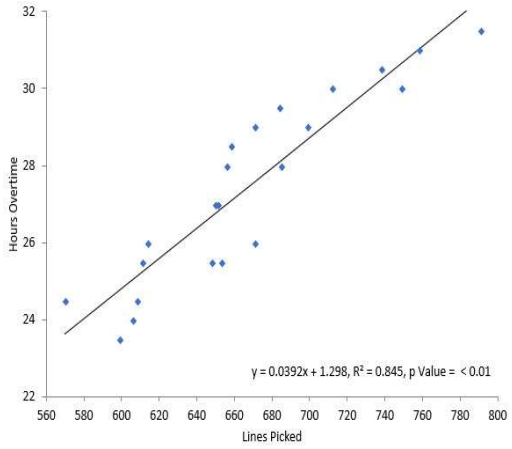


Bar charts: The frequency or proportion for each category plotted as bars.



Correlation

In statistics, correlation or dependence is any statistical relationship, whether causal or not, between two random variables or bivariate data.

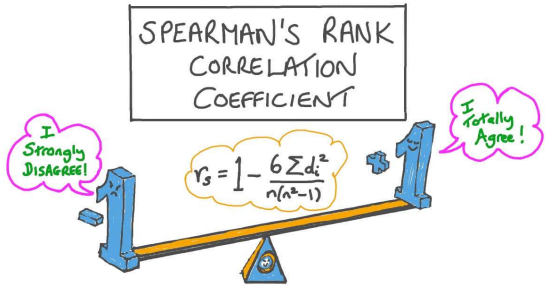


Correlation Coefficient Value (r)	Direction and Strength of Correlation
-1	Perfectly negative
-0.8	Strongly negative
-0.5	Moderately negative
-0.2	Weakly negative
0	No association
0.2	Weakly positive
0.5	Moderately positive
0.8	Strongly positive
1	Perfectly positive

Correlation can be positive, negative, no, perfect, linear, nonlinear, single, multiple & partial.

Methods used to determine correlations:

1. Scatter Plot
2. K-Pearson Coefficient Correlation
3. Spearman Rank



Exploring Two / more than two variables

Contingency table: A tally of counts between two or more categorical variables.

Also Called Crosstab.

Hexagonal binning:A plot of two numeric variables with the records binned into hexagons.

Contour plot
A plot showing the density of two numeric variables like a topographical map.

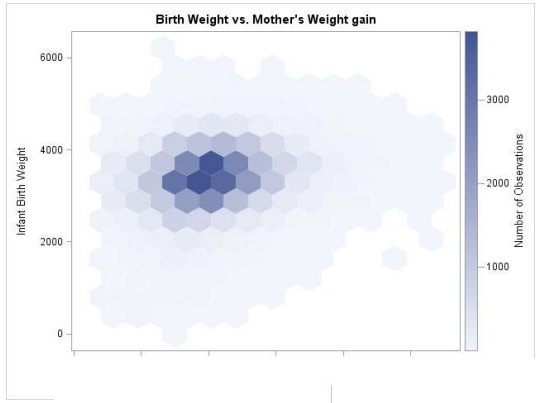
Violin plot
Similar to a boxplot but showing the density estimate.

Example of Contingency Table

A simple 2 x 2 Contingency Table

Groups	Dogs	Cats	Total
Males	42	10	52
Females	9	39	48
Total	51	49	100

Marginal Totals
Variables involved in the experiment and tabulated in contingency table are called marginal totals.



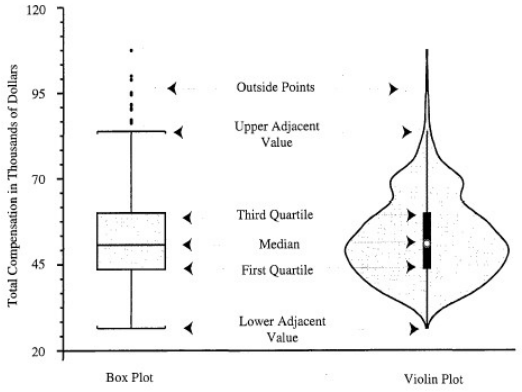


Figure 1. Common Components of Box Plot and Violin Plot. Total compensation for all academic ranks.