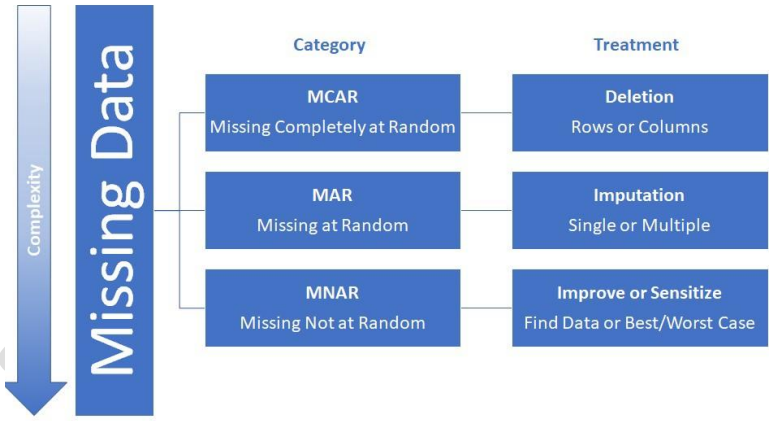
https://www.displayr.com/different-types-of-missing-data/

https://www.ncbi.nlm.nih.gov/books/NBK493614/

15. What do you mean by missing data and discuss about various types of missing data?

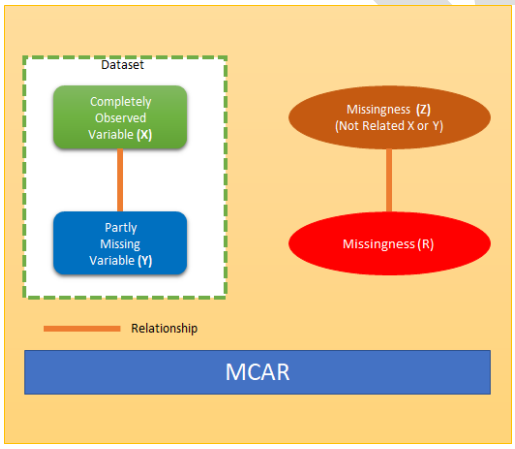
Missing data are defined as values that are not available and that would be meaningful if they are observed. Missing data can be anything from missing sequence, incomplete feature, files missing, information incomplete, data entry error etc. In order to understand what to do with missing values found in your dataset, firstly, you need to understand what type of missing values you have.

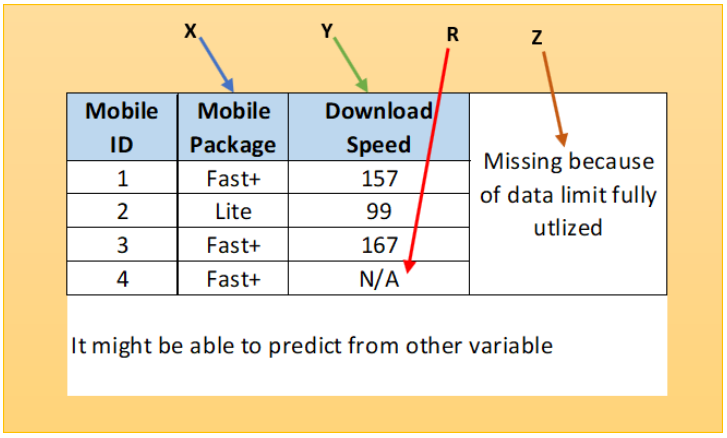


Three kinds of missing data:

- Missing at Random (MAR)
- Missing Completely at Random (MCAR)
- Missing Not at Random (MNAR)

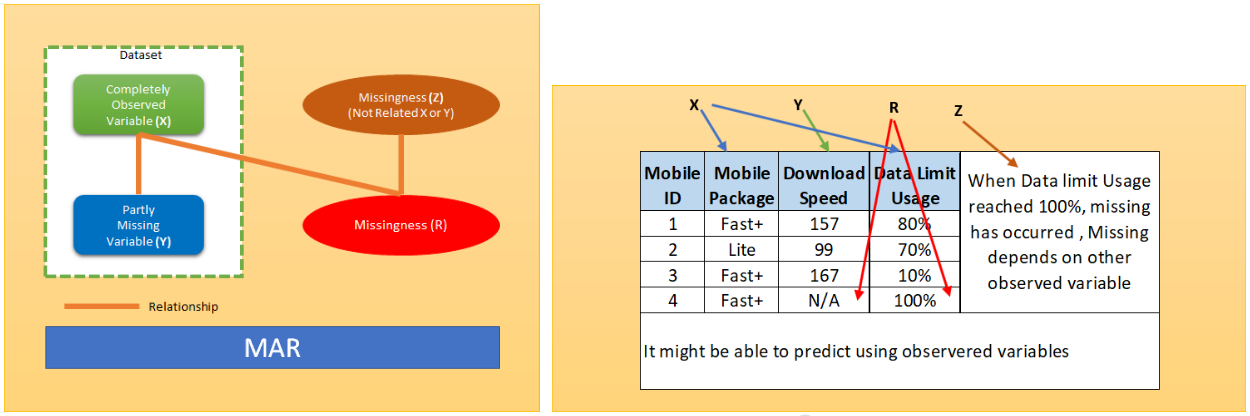
- **Missing completely at random (MCAR).** When we say data are missing completely at random, we mean that the missingness has nothing to do with the observation being studied



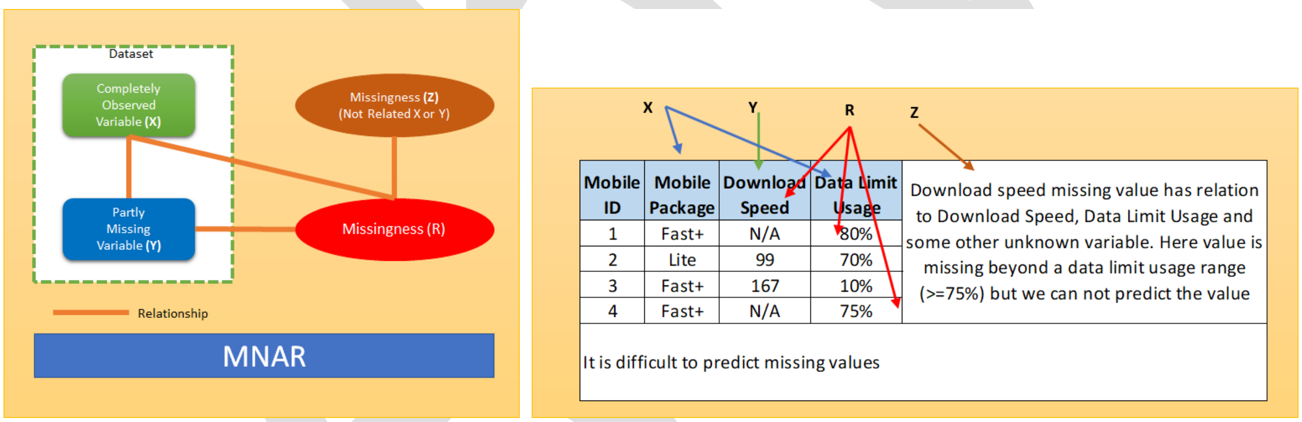


One example of mobile data. Here one sample has missing value which is not because of dataset variables but because of another reason.

- **Missing at random (MAR).** When we say data are missing at random, we mean that missing data on a partly missing variable (Y) is related to some other completely observed variables(X) in the analysis model but not to the values of Y itself.



- **Missing not at random (MNAR).** If the characters of the data do not meet those of MCAR or MAR, then they fall into the category of missing not at random (MNAR). When data are missing not at random, the missingness is specifically related to what is missing, e.g. a person does not attend a drug test because the person took drugs the night before, a person did not take English proficiency test due to his poor English language skill. The cases of MNAR data are problematic. The only way to obtain an unbiased estimate of the parameters in such a case is to model the missing data but that requires proper understanding and domain knowledge of the missing variable. The model may then be incorporated into a more complex one for estimating the missing values. A pictorial view of MNAR as below where missingness has direct relation to variable Y. It can have other relation as well



Reference: https://towardsdatascience.com/all-about-missing-data-handling-b94b8b5d2184

~~~~~~

16. How to handle missing data in data cleaning process?

1. Ignore the data row - This is usually done when the *class label* is missing or many attributes are missing from the row (not just one). However, you'll obviously get poor performance if the percentage of such rows is high.

For example, let's say we have a database of students enrolment data (age, SAT score, state of residence, etc.) and a column classifying their success in college to "Low", "Medium" and "High". Let's say our goal is to build a model predicting a student's success in college. Data rows who are missing the success column are not useful in predicting success so they could very well be ignored and removed before running the algorithm.

2. Use a global constant to fill in for missing values - Decide on a new global constant value, like "*unknown*", "*N/A*" or *minus infinity*, that will be used to fill all the missing values. This technique is used because sometimes it just doesn't make sense to try and predict the missing value.
~~~~~~

For example, let's look at the students enrollment database again. Assuming the *state of residence* attribute data is missing for some students. Filling it up with some state doesn't really makes sense as opposed to using something like "*N/A*".

3. Use attribute mean - Replace missing values of an attribute with the mean (or median if its discrete) value for that attribute in the database.

(For example, in a database of US family incomes, if the average *income* of a US family is X you can use that value to replace missing income values.)

4. Use attribute mean for all samples belonging to the same class - Instead of using the mean (or median) of a certain attribute calculated by looking at all the rows in a database, we can limit the calculations to the relevant class to make the value more relevant to the row we're looking at.

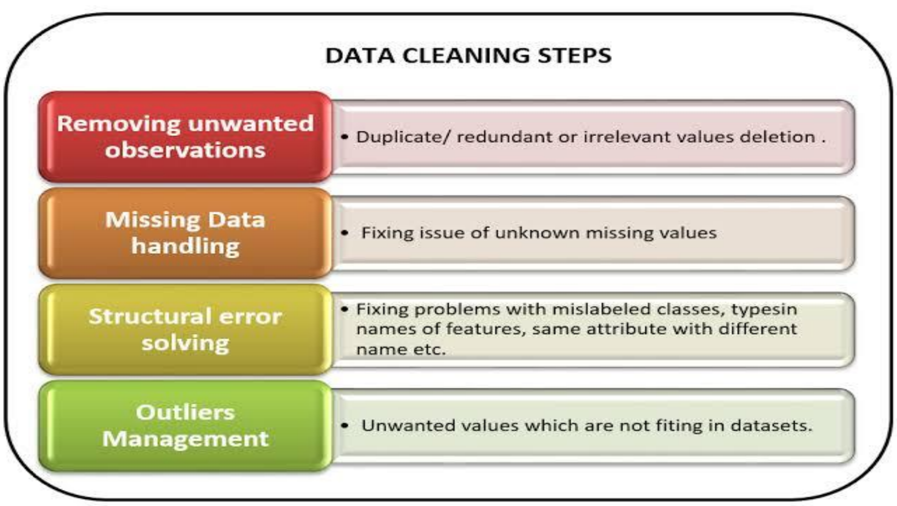
(Let's say you have a cars pricing database that, among other things, classifies cars to *"Luxury"* and *"Low budget"* and you're dealing with missing values in the cost field. Replacing missing cost of a luxury car with the average cost of all luxury cars is probably more accurate than the value you'd get if you factor in the low budget cars.)

5. Use a data mining algorithm to predict the most probable value -The value can be determined using regression, inference based tools using Bayesian formalism, decision trees, clustering algorithms (K-Mean\Median etc.).

(For example, we could use clustering algorithms to create clusters of rows which will then be used for calculating an attribute mean or median as specified in technique #3.
Another example could be using a decision tree to try and predict the probable value in the missing attribute, according to other attributes in the data.)

~~~~~~

17. Explain the methods in data cleaning?



Data Cleaning Methods: Understanding the what and why behind data cleaning is one, going ahead to implement it is another. Therefore, this section will be covering the steps involved in data cleaning, and further explanations on how each of these steps is carried out.

Removal of Unwanted Observations: Since one of the main goals of data cleansing is to make sure that the dataset is free of unwanted observations, this is classified as the first step to data cleaning. Unwanted observations in a dataset are of 2 types, namely; the duplicates and irrelevances.
~~~~~~

Duplicate Observations: A data is said to be a duplicate if it is repeated in a dataset, with it having more than one occurrence. This usually arises when the dataset is created as a result of combining data from two or more sources.

This can also occur in some other cases, including when a respondent makes more than one submission to a survey or error during data entry.

Irrelevant Observations: Irrelevant observations are those that don't actually fit the specific problem that you're trying to solve. Like having the price when you are only dealing with quantity.

For example, if you were building a model for prices of apartments in an estate, you don't need data showing the number of occupants of each house. Irrelevant observations mostly occur when data is generated by scraping from another data source.

Fix Data Structure: After removing unwanted observations, the next thing to do is to make sure that the wanted observations are well-structured. Structural errors may occur during data transfer due to a slight human mistake or incompetency of the data entry personnel.

Some of the things one should look out for when fixing data structure include; typographical errors, grammatical blunders, and so on. The data structure is mostly concerned with categorical data.

Filter-out Outliers:

- In order to improve the performance of your model, you should remove outliers. Outliers are data points that differ significantly from other observations in a data set.
- Outliers are very tricky, in the sense that they are of the same type with other observations, making them look wanted but hugely different from the others. For example, a particular data point may be numerical like other observations in the data set but may turn out to be a big 1000 with the rest between the ranges 1-10.
- Although problematic to some models, there should be a valid reason for removing an outlier. Outliers may arise from a measurement error that is unlikely to be real data, while it may also be as a result of scraping a bigger dataset.
- Outliers may give more insight into your model the way the other observations can't. Hence, you should be careful when removing outliers from your data.

Handle Missing Data: You may end up with missing values in your data due to errors during data collection or non-response bias from respondents. You can avoid this by adding data validation to your survey.

However, now that you already have missing data, how do you handle them?

There are 2 common ways of handling missing data, which are; entirely removing the observations from the data set and imputing a new value based on other observations.

Drop Missing Values

- By dropping missing values, you drop information that may assist you in making better conclusions on the subject of study. It may deny you the opportunity of benefiting from the possible insights that can be gotten from the fact that a particular value is missing.
- For example, when collecting the scores of students in various exams, student A's score mag be missing in mathematics because he didn't sit for the exam. Assume that this happened because he was sick, and sick students are allowed to rest the exam at a later date.
- If the whole observation was deleted, we cannot detect that he was sick.

Inpute Missing Values

- Consider another student B who wrote the exam but his score was missing because the teacher forgot to enter her score. If the teacher imputes a random score for her, it may end up rendering the data incorrect.
- Student B may have scored higher or lower than the score the teacher randomly assigns to her.
- Therefore, if data is missing, you should always indicate it in your dataset. You can indicate missing values by simply creating a **Missing** category if the data is categorical, or flagging and filling with 0 if it is numerical.
- This way, the algorithm will be aware that there are missing values in the dataset.

~~~~~~

18.Compare and contrast the homogeneous and heterogeneous data?

<u>HOMOGENEOUS DATA</u>

A data set is homogeneous if it is made up of things (i.e. people, cells or traits) that are similar to each other. For example a data set made up of 20-year-old college students enrolled in Physics 101 is a homogeneous sample.

Homogeneous: Red, Green, Purple

<u>HETEROGENEOUS DATA</u>

Heterogeneous data are any data with high variability of data types and formats. They are possibly ambiguous and low quality due to missing values, high data redundancy, and untruthfulness. It is difficult to integrate heterogeneous data to meet the business information demands. For example, heterogeneous data are often generated from Internet of Things (IoT). Data generated from IoT often has the following four features

1. First, they are of heterogeneity. Because of the variety of data acquisition devices, the acquired data are also different in types with heterogeneity.
2. Second, they are at a large-scale. Massive data acquisition equipment is used and distributed, not only the currently acquired data, but also the historical data within a certain time frame should be stored.
3. Third, there is a strong correlation between time and space. Every data acquisition device is placed at a specific geographic location and every piece of data has a time stamp. The time and space correlation is an important property of data from IoT.
4. Fourth, effective data accounts for only a small portion of the big data. A great quantity of noises may be collected during the acquisition and transmission of data in IoT. Among datasets acquired by acquisition devices, only a small amount of data is valuable.

Heterogeneous: White, 1/2/2015, 424291.23

~~~~~~

19. Analyze the checking the importance of consistency checking in data cleaning? [DOUBTFUL ANSWER]

Consistency - You can measure consistency by comparing two similar systems. Or, you can check the data values within the same dataset to see if they are consistent or not. Consistency can be relational. **For example**, a customer's age might be 15, which is a valid value and could be accurate, but they might also be stated senior-citizen in the same system.
In such cases, you'll need to cross-check the data, similar to measuring accuracy, and see which value is true. Is the client a 15-year old? Or is the client a senior-citizen? Only one of these values could be true.

There are multiple ways to make your data consistent.

- *Check different systems*: You can take a look at another similar system to find whether the value you have is real or not. If two of your systems are contradicting each other, it might help to check the third one. In our previous

example, suppose you check the third system and find the age of the customer is 65. This shows that the second system, which said the customer is a senior citizen, would hold.

- *Check the latest data:* Another way to improve the consistency of your data is to check the more recent value. It can be more beneficial to you in specific scenarios. You might have two different contact numbers for a customer in your record. The most recent one would probably be more reliable because it's possible that the customer switched numbers.
- *Check the source*: The most fool-proof way to check the reliability of the data is to contact the source simply. In our example of the customer's age, you can opt to contact the customer directly and ask them their age. However, it's not possible in every scenario and directly contacting the source can be highly tricky. Maybe the customer doesn't respond, or their contact information isn't available.
- *Uniformity*: You should ensure that all the values you've entered in your dataset are in the same units. If you're entering SI units for measurements, you can't use the Imperial system in some places. On the other hand, if at one place you've entered the time in seconds, then you should enter it in this format all across the dataset.

~~~~~~

20. Explain the methods in visualizing relationships between variables?

VISUALIZING RELATIONSHIPS BETWEEN VARIABLES

• Scatter plots –

- Scatterplots can be used to identify whether a relationship exists between two continuous variables measured on the ratio or interval scales.
- The two variables are plotted on the x-and y-axis.
- Each point displayed on the scatterplot is a single observation.
- The position of the point is determined by the value of the two variables.
- Scatterplots allow you to see the type of relationship that may exist between two variables. – positive relationship – Negative relationship
- nature of the relationships
  - – Linear
  - – nonlinear

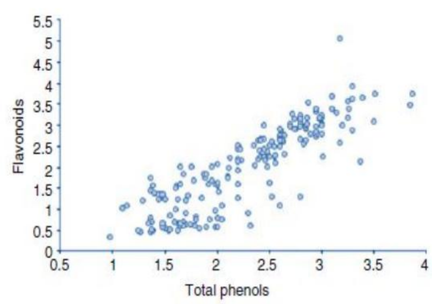


FIGURE 1: Example of a scatterplot where each point corresponds to an observation.

• Summary Tables and Charts

- Simple summary table is a common way of understanding the relationship between two variables where at least one of the variables is discrete.
- Summary tables can also be used to show the relationship between ordinal variables and another variable.
~~~~~~

Class	Count	Minimum (petal width (cm))	Maximum (petal width (cm))	Mean (petal width (cm))	Median (petal width (cm))	Standard deviation (petal width (cm))
Iris-setosa	50	0.1	0.6	0.244	0.2	0.107
Iris-versicolor	50	1	1.8	1.33	1.3	0.198
Iris-virginica	50	1.4	2.5	2.03	2	0.275

FIGURE 6: Example of a summary table.

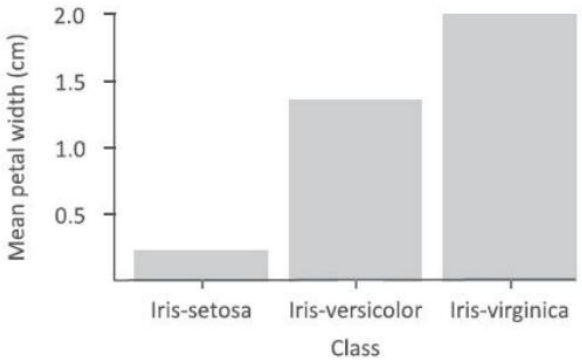


FIGURE 7: Example of a summary graph.

• Cross-Classification Tables

- **Cross-classification tables or contingency tables provide insight into the relationship between two categorical variables (or non-categorical variables transformed to categorical variables).**

		Infection Class		
		Infection negative	Infection positive	Totals
Test Results	Blood test negative	17	10	27
	Blood test positive	6	22	28
	Totals	23	32	55

FIGURE13 Contingency table showing the relationship between two dichotomous variables.

		Gender		
		Male	Female	Totals
Age-group	10–19	847	810	1657
	20–29	4878	3176	8054
	30–39	6037	2576	8613
	40–49	5014	2161	7175
	50–59	3191	1227	4418
	60–69	1403	612	2015
	70–79	337	171	508
	80–89	54	24	78
	90–99	29	14	43
	Total	21,790	10,771	32,561

FIGURE 14 Contingency table illustrating the number of females and males in each age-group.

~~~~~~

21. Discuss how to calculate metrics about relationships between variables?

Relationship between variables means correlation. So Karl pearson correlation.

Coefficient of Correlation- A coefficient of correlation is generally applied in statistics to calculate a relationship between two variables. The correlation shows a specific value of the degree of a linear relationship between the X and Y variables, say X and Y. There are various types of correlation coefficients. However, Pearson's correlation (also known as Pearson's R) is the correlation coefficient that is frequently used in linear regression.

Pearson's Coefficient Correlation- Karl Pearson's coefficient of correlation is an extensively used mathematical method in which the numerical representation is applied to measure the level of relation between linearly related variables. The coefficient of correlation is expressed by "r".

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Pearson correlation example

1. When a correlation coefficient is (1), that means for every increase in one variable, there is a positive increase in the other fixed proportion. For example, shoe sizes change according to the length of the feet and are perfect (almost) correlations.
~~~~~~

2. When a correlation coefficient is (-1), that means for every positive increase in one variable, there is a negative decrease in the other fixed proportion. For example, the decrease in the quantity of gas in a gas tank shows a perfect (almost) inverse correlation with speed.

3. When a correlation coefficient is (0) for every increase, that means there is no positive or negative increase, and the two variables are not related.

Reference: *https://www.simplypsychology.org/correlation.html*

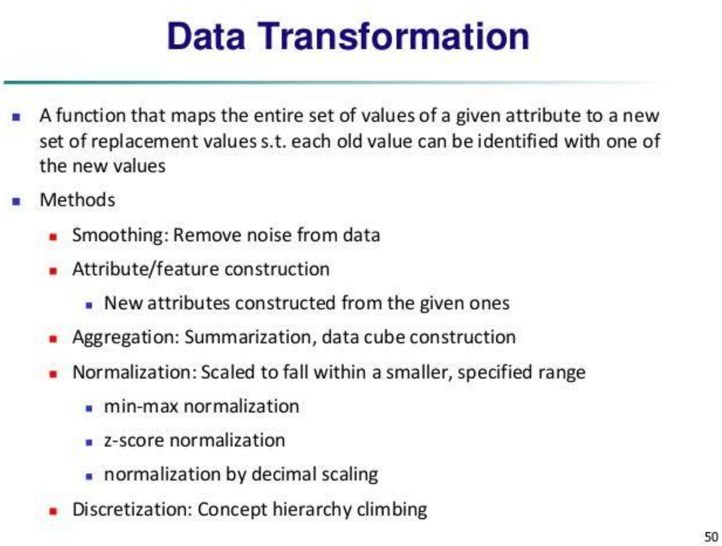
https://byjus.com/commerce/karl-pearson-coefficient-of-correlation/#:~:text=Karl%20Pearson's%20Coefficient%20of%20Correlation%20is%20an%20extensively%20used%20mathematical,is%20expressed%20by%20%E2%80%9Cr%E2%80%9D.

~~~~~~

22. Discuss about data transformation and its techniques?

Data Transformation

- ✓ Data transformation is a technique used to **convert** the raw data into a suitable format that eases data mining in retrieving the strategic information efficiently and fastly.
- ✓ Raw data is difficult to trace or understand that's why it needs to be preprocessed before retrieving any information from it.
- ✓ Data transformation includes data cleaning techniques as well as a data reduction technique to convert the data into the appropriate form.
- ✓ Data transformation is one of the essential data preprocessing techniques that must be performed on the data before data mining in order to provide patterns that are easier to understand. Knowing data transformation let's move toward the strategies involved in data transformation.



Data Transformation Strategies

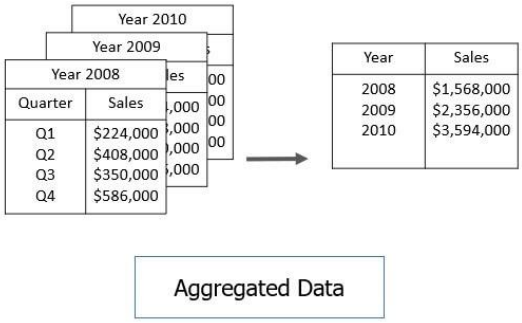
1. Data Smoothing

Smoothing the data means removing noise from the considered data set. The noise is removed from the data using the techniques such as binning, regression, clustering.
~~~~~~

- ✓ **Binning:** This method splits the sorted data into the number of bins and smoothens the data values in each bin considering the neighbourhood values around it.
- ✓ **Regression:** This method identifies the relation among two dependent attributes so that if we have one attribute it can be used to predict the other attribute.
- ✓ **Clustering:** This method groups, similar data values and form a cluster. The values that lie outside a cluster are termed as outliers.

2. Data Aggregation

Data aggregation transforms a large set of data to a smaller volume by implementing aggregation operation on the data set.



Aggregated Data

4. Data Normalization

Normalizing the data refers to **scaling the data values to a much smaller range like such as [-1, 1] or [0.0, 1.0]**. There are different methods to normalize the data as discussed below.

For our further discussion consider that we have a numeric attribute A and we have n number of observed values for attribute A that are U1, U_2, U_3, ….U_n.

a. Min-max normalization

The first technique we will cover is min-max normalization. It is the linear transformation of the ------>original unstructured data. It scales the data from 0 to 1. It is calculated by the following formula:

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

where is the current value of feature *F*.

b. Z-score normalization

The next technique is z-score normalization. It is also called zero-mean normalization. The essence of this technique is the data transformation by the values conversation to a common scale where an average number equals zero and a standard deviation is one. A value is normalized to ′ under the formula:

$$v' = \frac{v - \mu_F}{\sigma_F},$$

Here is the mean and is the standard deviation of feature *F*.

c. Decimal Scaling

And now we finally will move on to the decimal scaling normalization technique. It involves the data transformation by dragging the decimal points of values of feature *F*. The movement of decimals is very dependent on the absolute value of the maximum. A value of feature *F* is transformed to by calculating:

$$v' = \frac{v}{10^j},$$

In this formula, *j* is the lowest integer while *Max* (|) < 1.

5. Data Discretization: *Data discretization transforms numeric data by mapping values to interval or concept labels. Discretization techniques include the following:*

- **Data discretization by binning**: This is a top-down unsupervised splitting technique based on a specified number of bins.
- **Data discretization by histogram analysis**: In this technique, a histogram partitions the values of an attribute into disjoint ranges called buckets or bins. It is also an unsupervised method.
- **Data discretization by cluster analysis**: In this technique, a clustering algorithm can be applied to discretize a numerical attribute by partitioning the values of that attribute into clusters or groups.
- **Data discretization by decision tree analysis**: Here, a decision tree employs a top-down splitting approach; it is a supervised method. To discretize a numeric attribute, the method selects the value of the attribute that has minimum entropy as a split-point, and recursively partitions the resulting intervals to arrive at a hierarchical discretization.
- **Data discretization by correlation analysis**: This employs a bottom-up approach by finding the best neighboring intervals and then merging them to form larger intervals, recursively. It is supervised method.

~~~~~~

23.Write a short notes on association analysis ?

Association analysis is the task of finding interesting relationships in large datasets. These interesting relationships can take two forms: frequent item sets or **association** rules. Frequent item sets are a collection of items that frequently occur together.

Refer this: https://chih-ling-hsu.github.io/2017/03/25/Data-Mining-Association-Analysis

~~~~~~

24.Write about comparative statistics with examples?

comparative analysis as comparison analysis: Use comparison analysis to measure the financial relationships between variables over two or more reporting periods. Businesses use comparative analysis as a way to identify their competitive positions and operating results over a defined period. Larger organizations may often comprise the resources to perform financial comparative analysis monthly or quarterly, but it is recommended to perform an annual financial comparison analysis at a minimum.
Reference: https://www.sjsu.edu/faculty/watkins/compstat.htm

~~~~~~

25.What is clustering? Explain in detail the applications of clustering?

What is Clustering?
~~~~~~

In clustering, a group of different data objects is classified as similar objects. One group means a cluster of data. Data sets are divided into different groups in the cluster analysis, which is based on the similarity of the data. After the classification of data into various groups, a label is assigned to the group. It helps in adapting to the changes by doing the classification.

Applications of Cluster Analysis

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
- Clustering also helps in classifying documents on the web for information discovery.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.
- As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

~~~~~~

26. Discuss about k-means clustering with an example ?

The K-means clustering technique is simple, and we begin with a description of the basic algorithm. We first choose K initial centroids, where K is a user- specified parameter, namely, the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same.

Basic K-means algorithm.

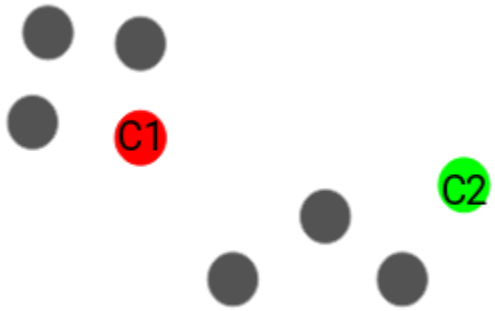
We have these 8 points and we want to apply k-means to create clusters for these points. Here’s how we can do it.

Step 1: Choose the number of clusters *k*

The first step in k-means is to pick the number of clusters, k.

Step 2: Select k random points from the data as centroids

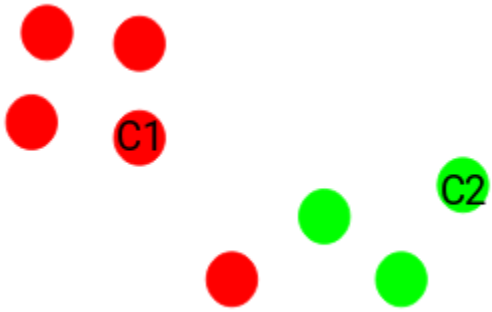
Next, we randomly select the centroid for each cluster. Let’s say we want to have 2 clusters, so k is equal to 2 here. We then randomly select the centroid:
~~~~~~



Here, the red and green circles represent the centroid for these clusters.

Step 3: Assign all the points to the closest cluster centroid

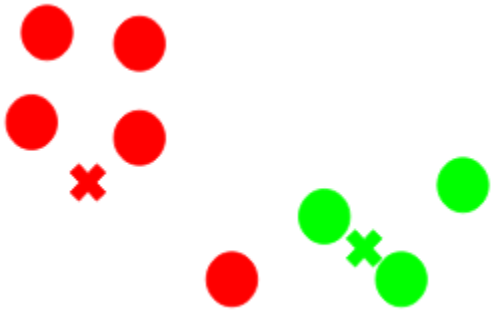
Once we have initialized the centroids, we assign each point to the closest cluster centroid:



Here you can see that the points which are closer to the red point are assigned to the red cluster whereas the points which are closer to the green point are assigned to the green cluster.

Step 4: Recompute the centroids of newly formed clusters

Now, once we have assigned all of the points to either cluster, the next step is to compute the centroids of newly formed clusters:



Here, the red and green crosses are the new centroids.

Step 5: Repeat steps 3 and 4

We then repeat steps 3 and 4:

The step of computing the centroid and assigning all the points to the cluster based on their distance from the centroid is a single iteration. But wait – when should we stop this process? It can't run till eternity, right?

K-means Clustering – Example 1:

A pizza chain wants to open its delivery centres across a city. What do you think would be the possible challenges?

- They need to analyse the areas from where the pizza is being ordered frequently.
- They need to understand as to how many pizza stores has to be opened to cover delivery in the area.
- They need to figure out the locations for the pizza stores within all these areas in order to keep the distance between the store and delivery points minimum.

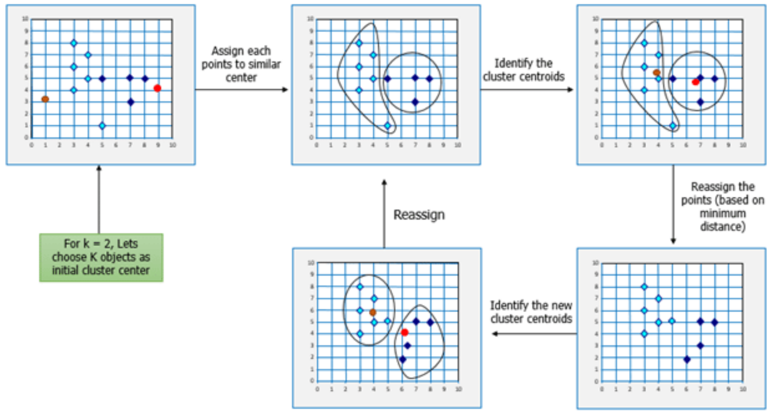
Resolving these challenges includes a lot of analysis and mathematics. We would now learn about how clustering can provide a meaningful and easy method of sorting out such real life challenges. Before that let's see what clustering is.

K-means Clustering Method:

If k is given, the K-means algorithm can be executed in the following steps:

- Partition of objects into k non-empty subsets
- Identifying the cluster centroids (mean point) of the current partition.
- Assigning each point to a specific cluster
- Compute the distances from each point and allot points to the cluster where the distance from the centroid is minimum.
- After re-allotting the points, find the centroid of the new cluster formed.

The step by step process:



Now, let's consider the problem in Example 1 and see how we can help the pizza chain to come up with centres based on K-means algorithm.

Similarly, for opening Hospital Care Wards:

K-means Clustering will group these locations of maximum prone areas into clusters and define a cluster center for each cluster, which will be the locations where the Emergency Units will open. These Clusters centers are the centroids of each cluster and are at a minimum distance from all the points of a particular cluster, henceforth, the Emergency Units will be at minimum distance from all the accident prone areas within a cluster.

~~~~~~

27. What is uni-variate analysis and explain its characteristics? [READ IT ATLEAST ONCE]

Univariate analysis is the simplest form of analyzing data. "Uni" means "one", so in other words your data has only one variable. It doesn't deal with causes or relationships (unlike regression) and it's major purpose is to describe; It takes data, summarizes that data and finds patterns in the data. Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- **THE DISTRIBUTION**
- **THE CENTRAL TENDENCY**
- **THE DISPERSION**

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution: The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But what do we do for a variable like income or GPA? With these variables there can be a large number of possible values, with relatively few people having each one. In this case, we group the raw scores into categories according to ranges of values. For instance, we might look at GPA according to the letter grade ranges. Or, we might group income into four or five ranges of income values.
~~~~~~

Frequency distribution table. -One of the most common ways to describe a single variable is with a frequency distribution. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the value are grouped into ranges and the frequencies determined.). Frequency distributions can be depicted in two ways, as a table or as a grap

Frequency distribution bar chart- Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- ✓ percentage of people in different income levels
- ✓ percentage of people in different age ranges
- ✓ percentage of people in different ranges of standardized test scores

Central Tendency: The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The Mean or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

The Median is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median. If we order the 8 scores shown above, we would get:

15, 15,15,20,20,21,25,36

There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median. The mode is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently. Notice that for the same set of 8 scores we got three different values -- 20.875, 20, and 15 -- for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.

Dispersion: Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The range is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

The Standard Deviation is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. Again let‘s take the set of scores:

15, 20,21,20,36,15,25,15

to compute the standard deviation, we first find the distance between each value and the mean.