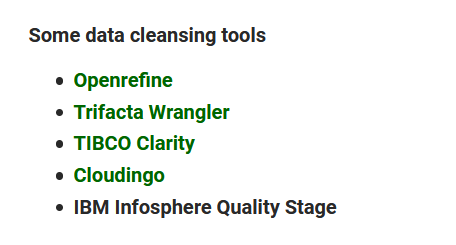


* Dropping observations with missing values.
* Imputing the missing values from past observations.
* Outliers can cause problems with certain types of models. For example, linear regression models are less robust to outliers than decision tree models.
* The errors that arise during measurement, transfer of data or other similar situations are called structural errors.

Example : America/America, red-yellow ->(red,-, yellow)

***Data Cleaning Methods :***

1. **Removal of Unwanted Observations**
2. **Duplicate Observations**
3. **Irrelevant Observations**
4. **Fix Data Structure**
5. **Filter-out Outliers**
6. **Handle Missing Data**
7. **Drop Missing Values**
8. **Inpute Missing Values**



**Mention what are the missing patterns that are generally observed?**

The missing patterns that are generally observed are

* Missing completely at random
* Missing at random
* Missing that depends on the missing value itself
* Missing that depends on unobserved input variable

**what is KNN imputation method?**

* In KNN imputation, the missing attribute values are imputed by using the attributes value that are most similar to the attribute whose values are missing.
* By using a distance function, the similarity of two attributes is determined.

**Data validation methods used by data analyst?**

Usually, methods used by data analyst for data validation are

* Data screening
* Data verification

**tools used in Big Data?**

Tools used in Big Data includes

* Hadoop
* Hive
* Pig
* Flume
* Mahout
* Sqoop

**What is Map Reduce?**

Map-reduce is a framework to process large data sets, splitting them into subsets, processing each subset on a different server and then blending results obtained on each.

**Clustering**? **What are the properties for clustering algorithms?**

Clustering is a classification method that is applied to data. Clustering algorithm divides a data set into natural groups or clusters.

Properties for clustering algorithm are

* Hierarchical or flat
* Iterative
* Hard and soft
* Disjunctive

**Statistical methods that are useful for data-analyst?**

Statistical methods that are useful for data scientist are

* Bayesian method
* Markov process
* Spatial and cluster processes
* Rank statistics, percentile, outliers detection
* Imputation techniques, etc.
* Simplex algorithm
* Mathematical optimization

**Correlogram analysis?**

* A correlogram analysis is the common form of spatial analysis in geography.
* It consists of a series of estimated autocorrelation coefficients calculated for a different spatial relationship.
* It can be used to construct a correlogram for distance-based data, when the raw data is expressed as distance rather than values at individual points.

**Hash table?**

* In computing, a hash table is a map of keys to values. It is a data structure used to implement an associative array.
* It uses a hash function to compute an index into an array of slots, from which desired value can be fetched.

**hash table collisions? How is it avoided?**

A hash table collision happens when two different keys hash to the same value.  Two data cannot be stored in the same slot in array.

To avoid hash table collision there are many techniques, here we list out two

* **Separate Chaining**:

It uses the data structure to store multiple items that hash to the same slot.

* **Open addressing**:

It searches for other slots using a second function and store item in first empty slot that is found

**what is imputation? List out different types of imputation techniques?**

During imputation we replace missing data with substituted values.  The types of imputation techniques involve are

* **Single Imputation**
* Hot-deck imputation: A missing value is imputed from a randomly selected similar record by the help of punch card
* Cold deck imputation: It works same as hot deck imputation, but it is more advanced and selects donors from another datasets
* Mean imputation: It involves replacing missing value with the mean of that variable for all other cases
* Regression imputation: It involves replacing missing value with the predicted values of a variable based on other variables
* Stochastic regression: It is same as regression imputation, but it adds the average regression variance to regression imputation
* **Multiple Imputation**
* Unlike single imputation, multiple imputation estimates the values multiple times

**N-gram:**

* An n-gram is a contiguous sequence of n items from a given sequence of text or speech.
* It is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n-1).

|  |  |
| --- | --- |
| **Data Mining** | **Data Profiling** |
| * Data Mining refers to the analysis of data with respect to finding relations that have not been discovered earlier. * It mainly focuses on the detection of unusual records, dependencies and cluster analysis. | * Data Profiling refers to the process of analyzing individual attributes of data. * It mainly focuses on providing valuable information on data attributes such as data type, frequency etc. |

**Important steps in the data validation process :**

* **Data Screening:** Different kinds of algorithms are used in this step to screen the entire data to find out any inaccurate values.
* **Data Verification:** Each and every suspected value is evaluated on various use-cases, and then a final decision is taken on whether the value has to be included in the data or not.

### ****waterfall chart :****

* The waterfall chart shows both positive and negative values which lead to the final result value.
* For example, if you are analyzing a company’s net income, then you can have all the cost values in this chart. With such kind of a chart, you can visually, see how the value from revenue to the net income is obtained when all the costs are deducted.

### ****Pivot Table, and what are the different sections of a Pivot Table :****

* A Pivot Table is a simple feature in Microsoft Excel which allows you to quickly summarize huge datasets.
* It is really easy to use as it requires dragging and dropping rows/columns headers to create reports.
* we can create one Pivot Table from multiple different tables when there is a connection between these tables.

A Pivot table is made up of four different sections:

* **Values Area**: Values are reported in this area
* **Rows Area**: The headings which are present on the left of the values.
* **Column Area:**The headings at the top of the values area makes the columns area.
* **Filter Area:** This is an optional filter used to drill down in the data set.