**Common Data problems**

High-quality data is the absolute greatest driver of revenue for a modern business. Good data can lead to a drastic boost in lead conversion rates, account-based success, and closed won deals. On the other hand, poor data quality can drastically reduce the ROI of a company’s CRM and marketing automation investment. Below are the common data problems;

1) **Incorrect Data**

Data decays at a rate of 2.2% per month. Therefore, it’s almost definitely going to be the case that some of your data is outdated. It’s a massive issue, as anywhere from 10 to 25% of existing data in your system has errors within it.

2) **Poor Organization**

If you’re not able to easily search through your data, you’ll find that it becomes significantly more difficult to make use of. Through different organizational methods and procedures, there are dozens of ways that data can be represented.

3) **Too Much Data**

40% of people reported that there’s often too much data to properly work off of inside a database.

4) **Inconsistent Data**

When dealing with multiple data sources, inconsistency is a big indicator that there’s a data quality problem.

5) **Poor Data Security**

20% of people say that they would never consider doing business again with a company that failed to handle their data in a professional and secure manner.

6) **Poorly Defined Data**

Oftentimes data is poorly defined, which causes great confusion around the proper methodology for management. For example, data that’s sectioned into the wrong category, like a company account being filed as a single person’s contact, is going to really mess things up in your database and make the whole thing more difficult to understand and sort through.

7) **Hidden data**

**How to overcome common Data problems**

1. Many times, if data has not been entered correctly in the system, or some files may have been corrupted, the remaining data has many missing variables. For example, if the address does not contain a zip code at all, the remaining details might be of little interest, because it will be challenging to determine the geographical dimension. With a data integration tool, you can help convert unstructured data to structured data. And also, move data from various formats into one consistent form.
2. Finally, there's no point in running big data analytics or making contact with customers based on data that is just plain wrong. Data can quickly become inaccurate. By not gathering all the hidden data, your data is not complete and limits you from making decisions based on complete and accurate data sets. The more obvious way for inaccurate data is data in systems filled with human mistakes, like a type or wrong information provided by the customer or inputting details in the wrong field. These can be among the toughest data quality issues to be found, mainly if the encoding is still appropriate- for example, entering an inaccurate, but legitimate, social security number can go unnoticed by a database that only checks in isolation the veracity of the information. There’s no cure for human error, but ensuring you have clear procedures that are followed consistently is a good start. Automation tools to reduce the amount of manual work when moving data between systems is also hugely useful in reducing the risk of mistakes by tired or bored workers.
3. Multiple copies of the same records take a toll on computing and storing, but may also produce skewed or incorrect insights when undetected. One of the critical problems could be human error — someone simply entering data multiple times by accident — or an algorithm that went wrong.The solution suggested for this problem is called "data deduplication." It is a combination of human intuition, data analysis, and algorithms to detect possible duplicates based on chance scores and common sense to determine where records look like a near match.

**What is Record Linkage**

**Record linkage**" is the term used by statisticians, epidemiologists, and historians, among others, to describe the process of joining records from one data source with another that describe the same entity. There are two main types of linkage algorithms: deterministic and probabilistic.

1. **Deterministic Linkage Methods**

Deterministic algorithms determine whether record pairs agree or disagree on a given set of identifiers, where agreement on a given identifier is assessed as a discrete—“all-or-nothing”—outcome. Match status can be assessed in a single step or in multiple steps. In a single-step strategy, records are compared all at once on the full set of identifiers. A record pair is classified as a match if the two records agree, character for character, on all identifiers and the record pair is uniquely identified (no other record pair matched on the same set of values). A record pair is classified as a no match if the two records disagree on any of the identifiers or if the record pair is not uniquely identified. In a multiple-step strategy (also referred to as an iterative or stepwise strategy), records are matched in a series of progressively less restrictive steps in which record pairs that do not meet a first round of match criteria are passed to a second round of match criteria for further comparison.

1. **Probabilistic Linkage Methods**

The deterministic approach ignores the fact that certain identifiers or certain values have more discriminatory power than others do. Probabilistic strategies have been developed to assess (1) the discriminatory power of each identifier and (2) the likelihood that two records are a true match based on whether they agree or disagree on the various identifiers.

**Summary of Steps for Probabilistic Record Linkage**

* Estimate the m and u probabilities for each linking variable using the observed frequency of agreement and disagreement patterns among all pairs, commonly generated using the EM (expectation-maximization) algorithm described by Fellegi-Sunter.
* Calculate agreement and disagreement weights using the m and u probabilities.
* Calculate a total linking weight for each pair by summing the individual linking weights for each linkage variable.
* Compare the total linkage weight to a threshold above which pairs are considered a link. The threshold is set using information generated in step 1.

Text and Categorical data problems.

**Categorical data** is a data that represents discrete values which belong to specific finite set of categories. These categorical data can either be a text or numeric in nature or even unstructured data like images! There are two major classes of categorical data**, nominal** and **ordinal.**

**Common problems**

Too much data.

Duplicate data.

Poor data recovery.

Incorrect data.

Poorly defined data.