Essential Questions for Data Cleaning

- Revan Markad

1. **Are there any missing values (NULL or blank entries)?**
   1. Identify columns with missing data.
   2. Decide whether to fill in missing values, remove rows, or use imputation techniques.

2. **Are there any duplicate records?**

   1. Check for duplicate rows based on key columns.
   2. Decide whether to remove duplicates and retain the first or last occurrence.

3. **Are there any inconsistencies in formatting or data types?**

   1. Check if data is in the correct format (e.g., numerical, date).
   2. Standardize inconsistent entries, such as phone numbers, dates, or addresses.

4. **Are there outliers or invalid values in numerical fields?**

   1. Identify any unusually high or low values that don't make sense (e.g., negative ages, unrealistic purchase amounts).
   2. Decide whether to cap, remove, or correct outliers.

5. **Are there empty strings in text fields?**

    1. Identify any empty string ('') values in text fields.
    2. Decide whether to replace them with NULL values or some default value.

6. **Are there invalid or negative values in numeric fields?**

    1. Check if any numeric fields contain negative or logically incorrect values (e.g., negative salaries, charges, or quantities).

7. **Are the date fields within a valid range or correct format?**

    1. Verify that dates fall within acceptable ranges (e.g., birthdate should not be in the future).
    2. Ensure date formats are consistent (e.g., YYYY-MM-DD).

8. **Are there invalid category or text field values?**

    1. Check categorical columns for unexpected values (e.g., "other" categories, misspelled entries).
    2. Standardize these categories where applicable.

9. **Are there any relationships or dependencies between columns that should be checked?**

    1. For example, if age and senior_citizen are related, ensure consistency.
    2. If customer_id is in multiple tables, ensure it is consistent across them.

10. **Are there any incorrect or missing relationships between tables?**

    1. Ensure that foreign keys match the primary keys in related tables.
    2. Check for orphaned records (records in one table without corresponding entries in related tables).

11. **Is the dataset free of irrelevant or unnecessary data?**

    1. Identify any columns that are not useful for analysis and remove them.

12. **Are there any issues with encoding or special characters in text fields?**

    1. Ensure that text data is properly encoded (e.g., UTF-8) and clean any special or non-printable characters.