# Assignment2

1. Difference between Quality and Tidiness issues, and rules for classifying them.

Data assessment involves examining:

* Data Quality and
* Data Tidiness.

|  |  |  |
| --- | --- | --- |
| S/N | Data Quality | Data Tidiness |
| 1 | Quality issues pertain to the content of data. | Tidiness issues pertain to data structure (its physical layout) with its semantics (its meaning). |
| 2 | Low quality data is also known as dirty data. | Untidy data is also known as messy data. |
| 3 | Data quality relates to its accuracy, completeness, consistency, and validity. | The requirements/rules for tidy data are:  Each variable forms a column.  Each observation forms a row.  Every cell is a single value  Each type of observational unit forms a table. |
| 3 | Symptoms of a dirty data include:  Incorrect data, inaccurate data, misleading data, duplicate data, unintegrated data, business rules violating data, inconsistent data, outdated data, unethical data, spell error, too much data and incomplete data | symptoms of messy data include:  Column headers are values, not variable names.  Multiple variables are stored in one column.  Variables are stored in both rows and columns.  Multiple types of observational units are stored in the same table.  A single observational unit is stored in multiple tables. |
| 4 | The importance of good quality data is that decisions based on it are accurate.  Ensuring a sufficient data quality (DQ) is essential for data-driven tasks like analysis, modeling, and reporting. | The importance of tidy data is that with a standardized framework for how your data should look, you spend less time on data cleaning and wrangling, and more time to focus on answering the problem at hand. |
| 5 |  | Causes of messy data include people, processes, organization, governance, enterprise-wide impact. |
| 6 |  | Assessment rules for data tidiness include |

Causes of dirty data include people, processes, organization, governance, enterprise-wide impact.

Assesment Rules for data quality include column analysis, cross domain analysis (functional dependency analysis), data verification, domain analysis, lexical analysis, primary key and foreign key analysis, sematic profiling

1. Visual assessment of data for Quality Issues
2. Programmatic assessment of data for Quality Issues
3. Visual assessment of data for Tidiness Issues
4. Programmatic assessment of data for Tidiness Issues
5. write down notable data issues on a dataset of your choice.