Open Government Data (OGD) Platform India

http://data.gov.in

Ensuring Datasets Quality

- **Data Compositeness/Completeness/Consistency**
  - **Check for the constituent elements (variables) within the dataset**
  - **The dataset should be well explained in terms of the variable present therein the dataset through a descriptive metadata**
  - **The metadata should well describe the time-period, units, definitions, frequency, data source, jurisdiction and notes to special mention in the dataset**
  - **The time series data should be continuous in nature**

- **Data Coverage**
  - **Dataset should be made available at the lowest possible levels to allow users correctly describe the phenomena being measured**

- **Standard process of “data cleansing”**
  - **Dataset column headers should be self-explanatory and has to be as first row only**
  - **Assigning string, date, character and numbers to the required fields**
  - **Abbreviations and acronyms to be replaced by full forms.**
  - **No special characters and blank spaces (replaced with NA) in the matrix.**
  - **Dataset should be de-normalized without any merged column**
  - **No formula of calculated column should appear in dataset like Total or Average of available column or rows**
  - **Above all it must be in machine readable format viz. CSV, XML, JSON etc. If datasets are released as static CSV while uploading the file, it would convert those datasets to other open formats automatically**
  - **File name should not contain special character except _ and -; no blank space should not be present in file name.**
  - **Regular Expression for valid Character for header**

    ```
    [^A-Za-z0-9, "'./?;:&_ ()@$#*^%+=\[\]!-]
    ```