Data Handling:

Once you source the data, it is essential to get rid of the irregularities in the data and fix it to improve its quality. One can encounter different kinds of issues in a dataset. Irregularities may appear in the form of missing values, anomalies/outliers, incorrect format and inconsistent spelling, etc. These irregularities may propagate further and affect the assumptions and analysis based on that dataset and may hamper the further process of machine learning model building. Hence, data cleaning is a very important step in EDA.

In general, any given data set is expected to have different types of data. Following are some examples with their data types.

|  |  |  |
| --- | --- | --- |
| Example | Variable Type | Data Type |
| Height, weight, age, temperature | Numerical variable | Int, float |
| Size of clothes, months, type of jobs, blood group. | Categorical variable | Object |
| Grades in exam, education level, months, integer ratings | Ordinal categorical type | Object, int, float |
| Date, time, timestamp | Date and time variable | Date and time |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Data quality issue | Examples | How to resolve |
| Fix rows and columns | Incorrect rows | Header rows, footer rows | Delete |
| Summary rows | Total, subtotal rows | Delete |
| Extra rows | Column numbers, indicators, blank rows | Delete |
| Missing Column Names | Column names as blanks, NA, XX etc. | Add the column names |
| Inconsistent column names | X1, X2,C4 which give no information about the column | Add column names that give some information about the data |
| Unnecessary columns | Unidentified columns, irrelevant columns, blank columns | Delete |
| Columns containing Multiple data values | E.g. address columns containing city, state, country | Split columns into components |
| No Unique Identifier | E.g. Multiple cities with same name in a column | Combine columns to create unique identifiers e.g. combine City with the State |
| Misaligned columns | Shifted columns | Align these columns |
|  |  |  |  |
| Missing Values | Disguised Missing values | blank strings, "NA", "XX", "999" etc | Set values as missing values |
| Significant number of Missing values in a row/column |  | Delete rows, columns |
| Partial missing values | Missing time zone, century etc | Fill the missing values with the correct value |
|  |  |  |  |
| Standardise Numbers | Non-standard units | Convert lbs to kgs, miles/hr to km/hr | Standardise the observations so all of them have the same consistent units |
| Values with varying Scales | A column containing marks in subjects, with some subject marks out of 50 and others out of 100 | Make the scale common. E.g. a percentage scale |
| Over-precision | 4.5312341 kgs, 9.323252 meters | Standardise precision for better presentation of data. 4.5312341 kgs couldbe presented as 4.53 kgs |
| Remove outliers | Abnormally High and Low values | Correct if by mistake else Remove |
| Standardise Text | Extra characters | Common prefix/suffix, leading/trailing/multiple spaces | Remove the extra characters |
| Different cases of same words | Uppercase, lowercase, Title Case, Sentence case, etc | Standadise the case/bring to a common case |
| Non-standard formats | 23/10/16 to 2016/10/20 “Modi, Narendra" to “Narendra Modi" | Correct the format/Standardise format for better readability in R |
|  |  |  |  |
| Fix Invalid Values | Encoding Issues | CP1252 instead of UTF-8 | Encode unicode properly |
| Incorrect data types | Number stored as a string: "12,300" | Convert to Correct data type |
| Date stored as a string: "2013-Aug" |
| String stored as a number: PIN Code "110001" stored as 110001 |
| Correct values not in list | Non-existent country, PIN code | Delete the invalid values, treat as Missing |
| Wrong structure | Phone number with over 10 digits |
| Correct values beyond range | Temperature less than -273° C (0° K) |
| Validate internal rules | Gross sales > Net sales |
| Date of delivery > Date of ordering |
| If Title is "Mr" then Gender is "M" |
|  |  |  |  |
| Filter Data | Duplicate data | Identical rows, rows where some columns are identical | Deduplicate Data/ Remove duplicated data |
| Extra/Unnecessary rows | Rows that are not required in the analysis. E.g if observations before or after a particular date only are required for analysis, other rows become unnecessary | Filter rows to keep only the relevant data. |
| Columns not relevant to analysis | Columns that are not needed for analysis e.g. Personal Detail columns such as Address, phone column in a dataset for | Filter columns-Pick columns relevant to analysis |
| Dispersed data | Parts of data required for analysis stored in different files or part of different datasets | Bring the data together, Group by required keys, aggregate the rest |

**Types of missing values:**

* MCAR: It stands for Missing completely at random. The reason behind the missing value is not dependent on any other features.
* MAR: It stands for Missing at random. The reason behind the missing value may be associated with some other features.
* MNAR: It stands for Missing not at random. There is a specific reason behind the missing value.

**Imputation on categorical/numeric columns:**

Categorical column: Impute the most popular category. Imputation can be done using logistic regression techniques.

Numerical column: Impute the missing value with mean/median/mode. The other methods to impute the missing values involve the use of interpolation, linear regression. These methods are useful for continuous numerical variables.

**Handling Missing Values:**

1. Check % of missing values. (use 🡪 100**\*(**new**.**OCCUPATION\_TYPE**.**isnull()**.**sum()**/**new**.**shape[0]))
2. Drop if % is more.
3. If required fill with 0.
4. Else fill numerical columns with mean (when Normal distribution) or Median (when Skewed distribution) and categorical columns with Mode.
5. If required perform fwdfill or bwdfill (as price do not change in one day)
6. KNN, Naïve-Bayes