**DATA CLEANING**

**Introduction:**

Data Cleaning is the process of correcting and deleting inaccurate records from a data base. It also consists of identifying and replacing incomplete, inaccurate, irrelevant data and records. The importance of Data Cleaning is to fix the poor quality data.

**Data Migration:**

It is the process of extracting data from one location and transferring to another.

**Benefits of Data Cleaning:**

* Streamlined business practices
* Increased productivity
* Faster sales cycle
* Better decisions

**Types of data issues:**

Duplicate data – two or more identical data

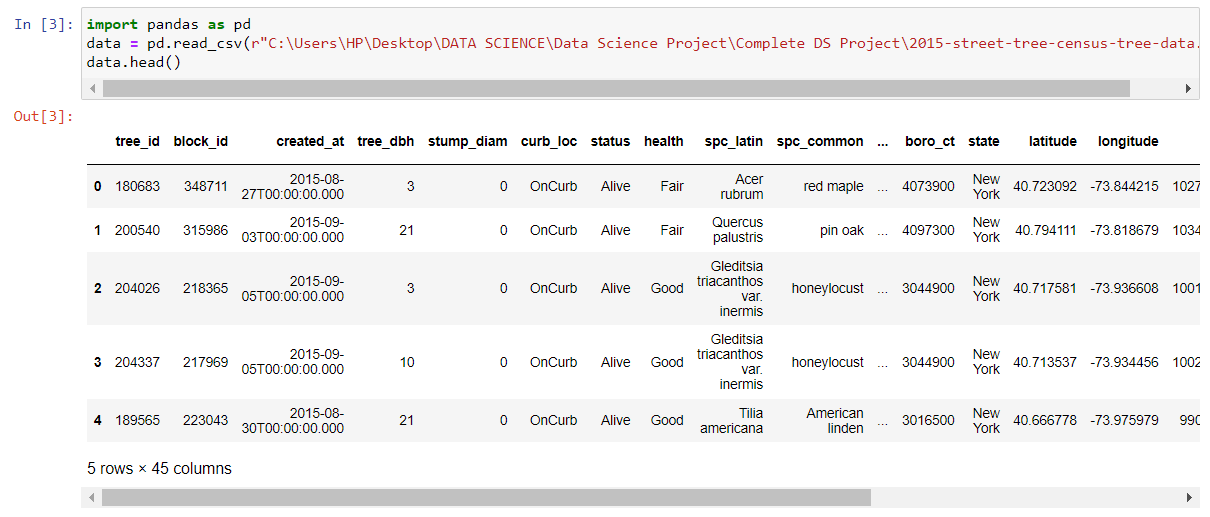
Conflicting data - When there are same records with different attributes, it means data is conflicting. For example, a company with different versions of addresses may cause delivery issues.

Incomplete data - The data that has missing attributes.

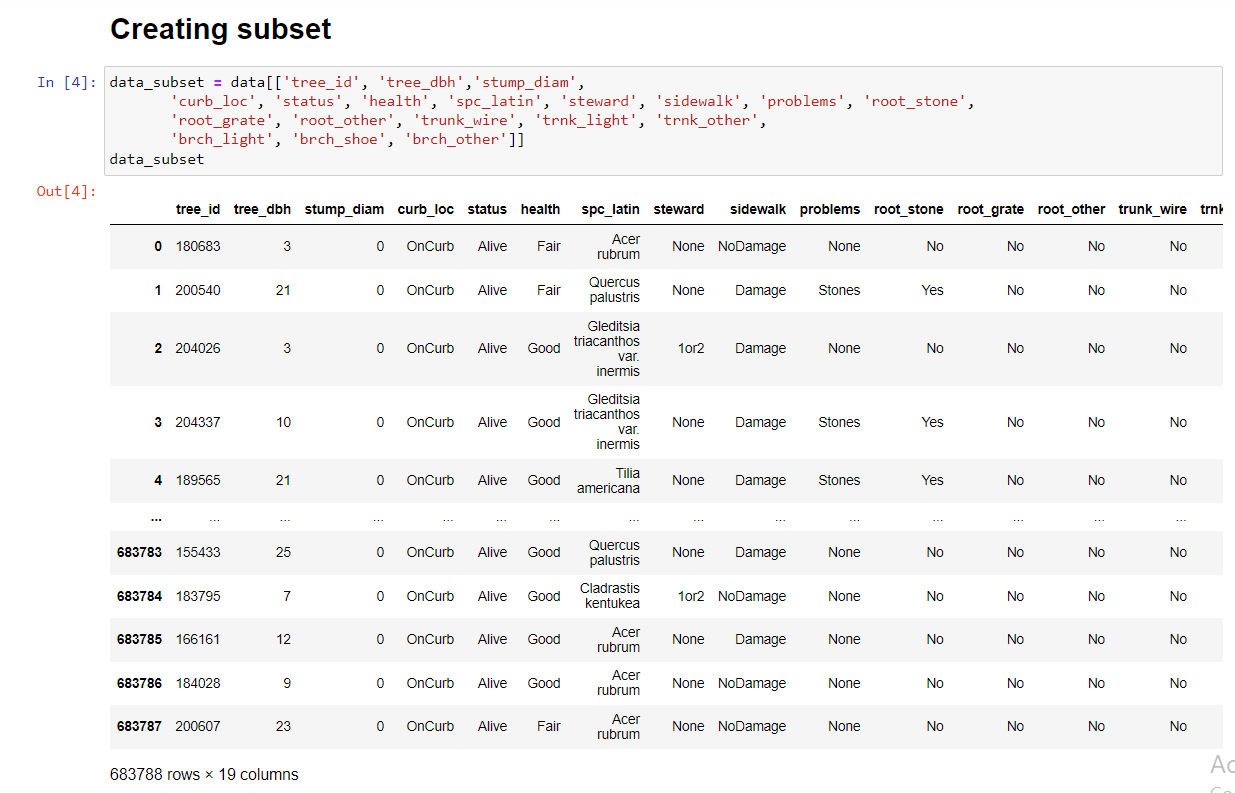
Invalid data - Data attributes are not conforming to standardization. For example, 9 digits phone number records rather than 10 digits.

**Steps under Data cleaning process:**

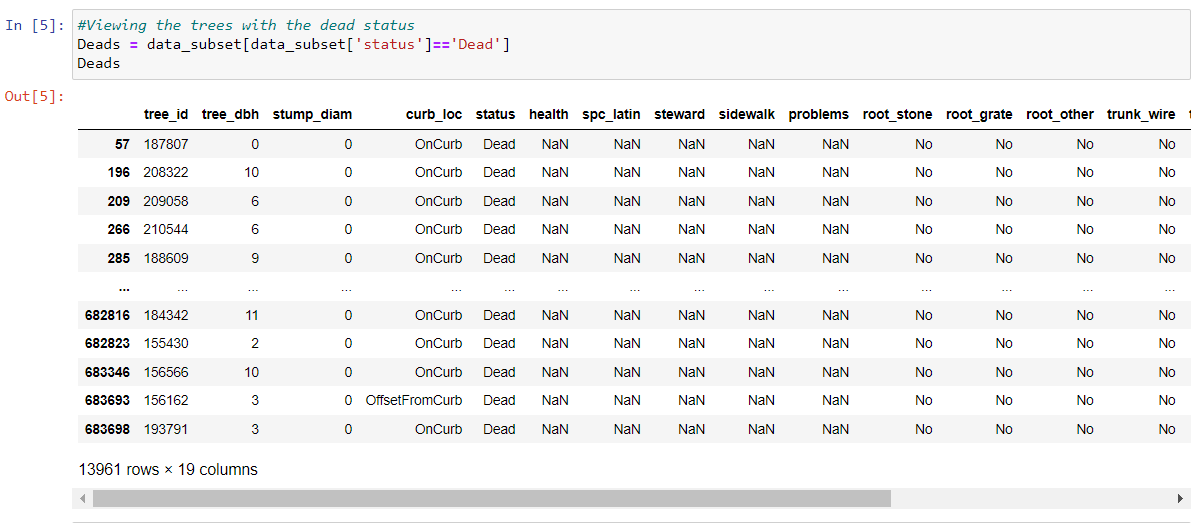
**Step1:** Import the pandas library and import the csv file



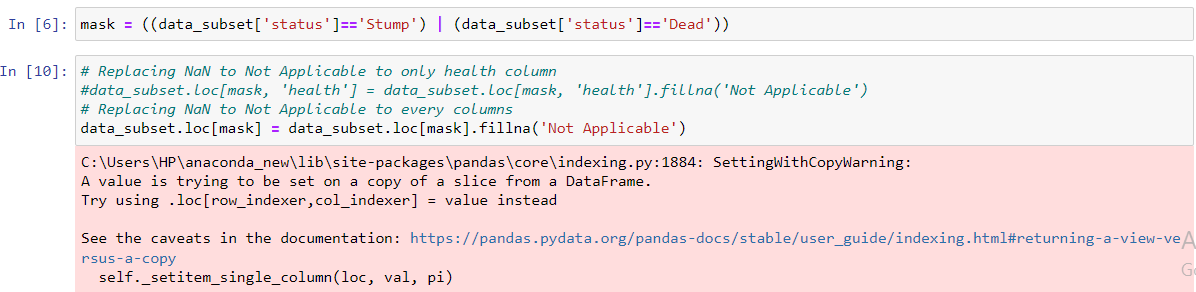
**Step 2:** Create a subset and extract the necessary columns to be worked on



**Step 3:** View the trees available in the data set with the status as ‘Dead’

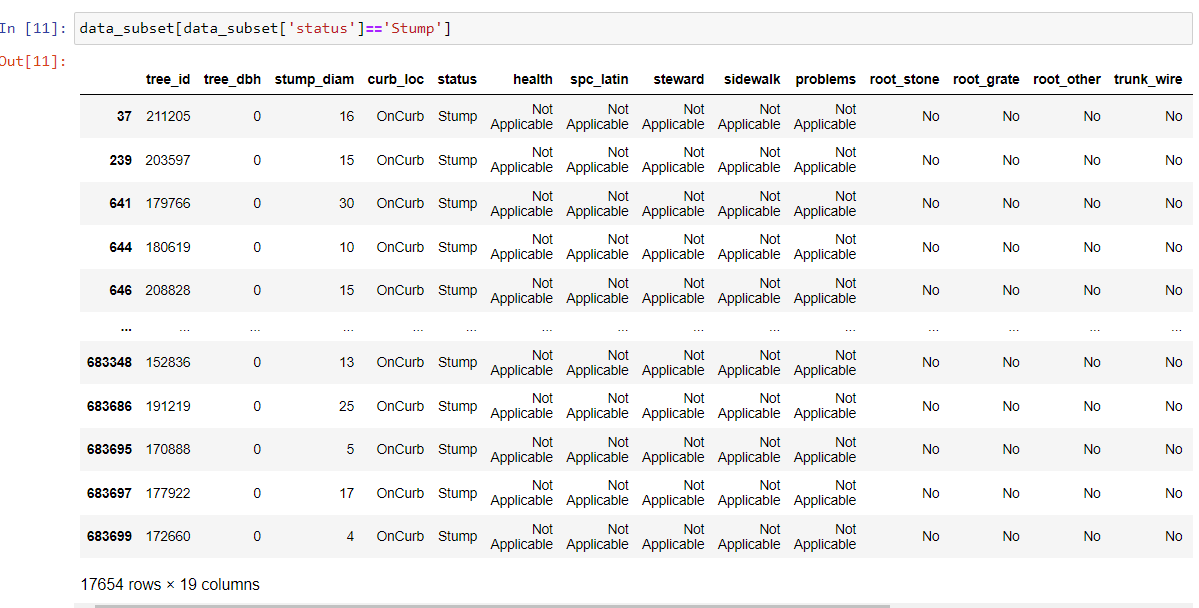


**Step 4:** Mask ‘Stump’ and ‘Dead’ values of status columns and fill NaN values to Not Applicable to every columns

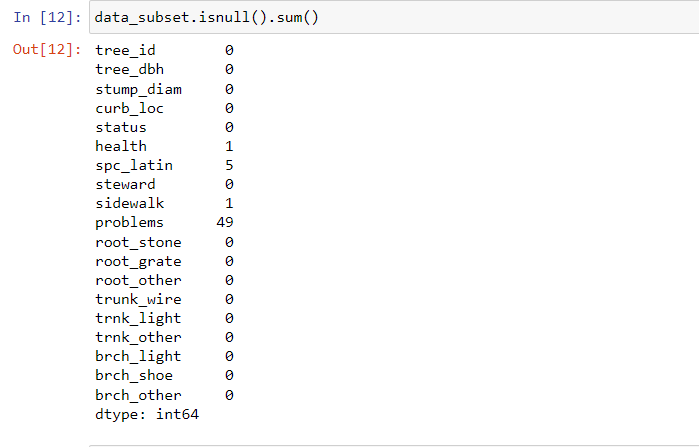


Here the cmd **data\_subset.loc[mask, 'health'] = data\_subset.loc[mask, 'health'].fillna('Not Applicable')** replaces NaN to Not Applicable to only health column, whereas data**\_subset.loc[mask] = data\_subset.loc[mask].fillna('Not Applicable')** replaces NaN to every columns

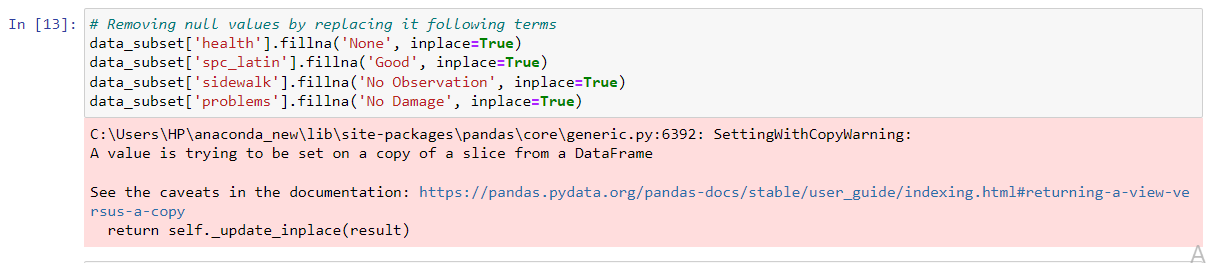
**Step 5:** Now check again the data columns whether the NaN values are replaced to Not Applicable.



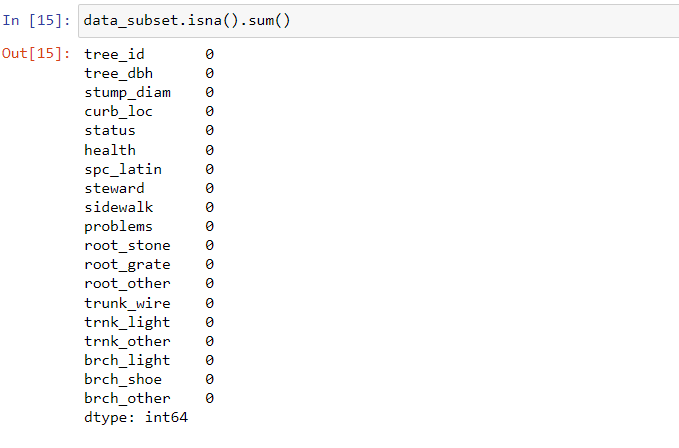
**Step 6:** Check for null values after replacing NaN values.



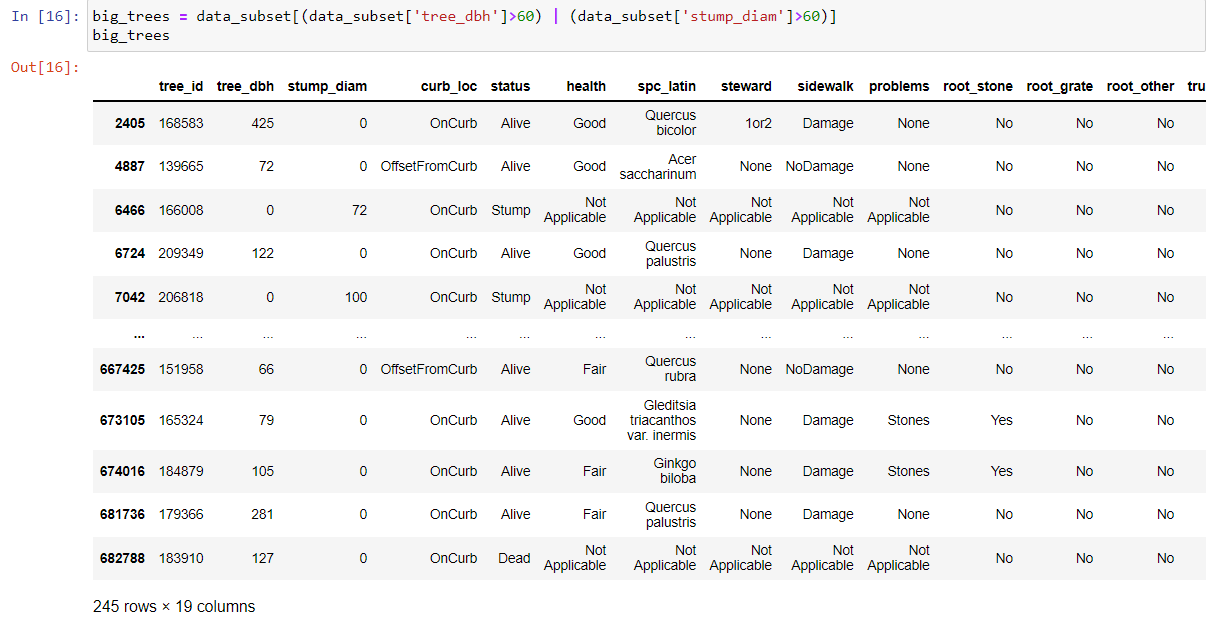
**Step 7:** Remove the null values by replacing it with valid values as follows



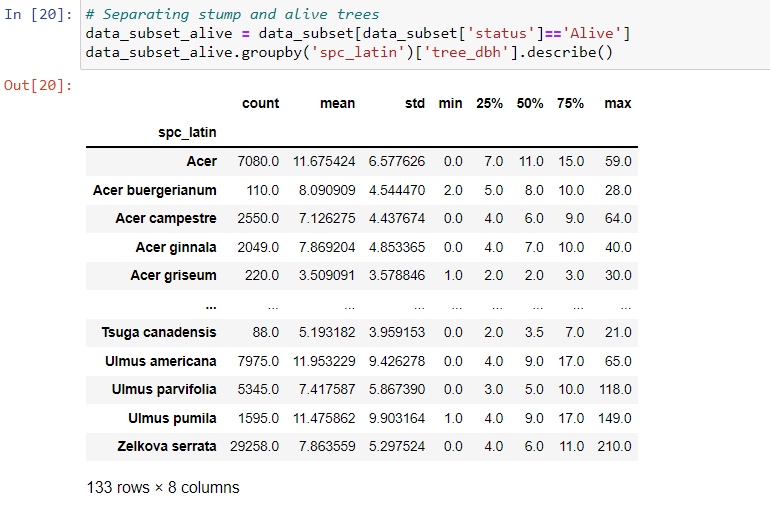
**Step 8:** Now check for the null values and it should not show the presence of null values



**Step 9:** After ensuring that there are no null values anymore we’ll get the absolute data with duplicates, null values and irrelevant data. This helps in reducing the data to be deal with. Now extract the values from the data through some expressions and comparisons.



**Step 10:** Separate stump and alive trees from the data base



Hence the process of Data Cleaning results in removing irrelevant columns, replacing null values with valid values and finally results in subset with necessary and valid values in the dataset.