**Documentation on the Data Cleaning Associated with NREGA Assets (State-wise)**

Raw files with data on NREGA assests at Panchayat or local body level was shared by Hemanth on One Drive. The are stored in the “data/raw” subdirectory.

**Data Overview**

The files are named according to the corresponding States and each file contains 20 columns which includes local body information, assets codes, start date and end date of construction, amount sanctioned for the asset etc.

**Data Cleaning**

Script to be used: xxxx

**There are 5 functions which were created to cover for all states and these respond to the following errors:**

1. Check unique key for each asset or row of information using 'block\_name', 'panchayat\_name', 'work\_code','work\_started\_date'.
2. Some string columns which contained dates (especially 'work\_physically\_completed\_date') had a string strip ‘01-01-1900’. This bit was present in multiple files and has been identified and removed.
3. Strip new line trails existing in multiple columns.
4. Remove digits and special characters-based corruptions from string columns such as 'block\_name', 'panchayat\_name', ‘district’, ‘state’.
5. To convert 'work\_started\_date', 'work\_physically\_completed\_date' into date64 format. Before converting to date64 format, the above columns would be stripped of any alphabets that might be present.
6. Rearrange the columns to bring ‘s.no’ as the first column.

**The above corrections would be further referred as “Basics” and any state file specific correction would be addressed accordingly.**

|  |  |  |
| --- | --- | --- |
| **State** | **Issues Rectified** | **Status** |
| Goa | Basics | Cleaned |
| Tripura | Basics | Cleaned |
| Manipur | Basics | Cleaned |
| Sikkim | Basics | Cleaned |
| Punjab | Basics | Cleaned |
| Haryana | Basics | Cleaned |
| Arunachal Pradesh |  | Processing for duplicates |
| Nagaland | Basics |  |
| Uttarakhand | Download again |  |
| Uttar Pradesh |  |  |
| Telangana |  |  |
| Tamil Nadu |  |  |
| Madhya Pradesh |  |  |
| Kerala |  |  |
| Lakshadweep | Basics | Cleaned |
| Karnataka |  |  |
| Meghalaya | Basics | Cleaned |
| Mizoram | Basics | Cleaned |
| Rajasthan |  |  |
| Jharkhand |  |  |
| Odisha |  |  |
| Puducherry | Basics | Processing for duplicates |
| J&K | Has issues with date line trails. I suspect that some entries of both columns are clubbed in one.  Download again |  |
| Himachal Pradesh | Basics | Cleaned |
| Gujarat |  |  |
| Dadra & Nagar Haveli | Download again |  |
| West Bengal |  |  |
| Andhra Pradesh |  |  |
| Assam | Basics | Cleaned |
| Andaman & Nicobar | Basics | Cleaned |
|  |  |  |

After cleaning, each state file would be rendered as a new cleaned file bearing its old name and saved in the sub directory “data/interim/NREGA\_assets”.

Note: Apart from the above, while cleaning the files, the script also takes a count of cases of issues with date columns. For example, count of missing values, cases where begin > closing date etc and these results for each state would be sent to an excel file, “finish\_date.xlsx”, available in the sub directory “data/interim”.