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

Process: *Sending raw data to interim*

*Author: Ewan Nikhil Thomas*

Raw files with data on NREGA assests at Panchayat or local body level was shared by Hemanth on One Drive. They 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: *state\_interim\_data\_create.py*

**There are 7 functions which were created to cover for all states and these respond to the following issues:**

1. Check unique key for each asset or row of information using 'block\_name', 'panchayat\_name', 'work\_code','work\_started\_date'.
2. The column ‘work\_name’ is removed from the data and the data is further filtered to on the basis of column ‘work\_status’ to include only works which are either “Completed” or “Physically Completed”.
3. 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.
4. Strip new line trails existing in multiple columns.
5. Remove digits and special characters-based corruptions from string columns such as 'block\_name', 'panchayat\_name', ‘district’, ‘state’. Rajasthan requires special attention.
6. 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.
7. Rearrange the columns to bring ‘s.no’ as the first column.
8. A function called ‘special\_rows\_filter’ is used to handle state specific issues such as presence of unknown block\_name called “KURWAI” in the state of Kerala.

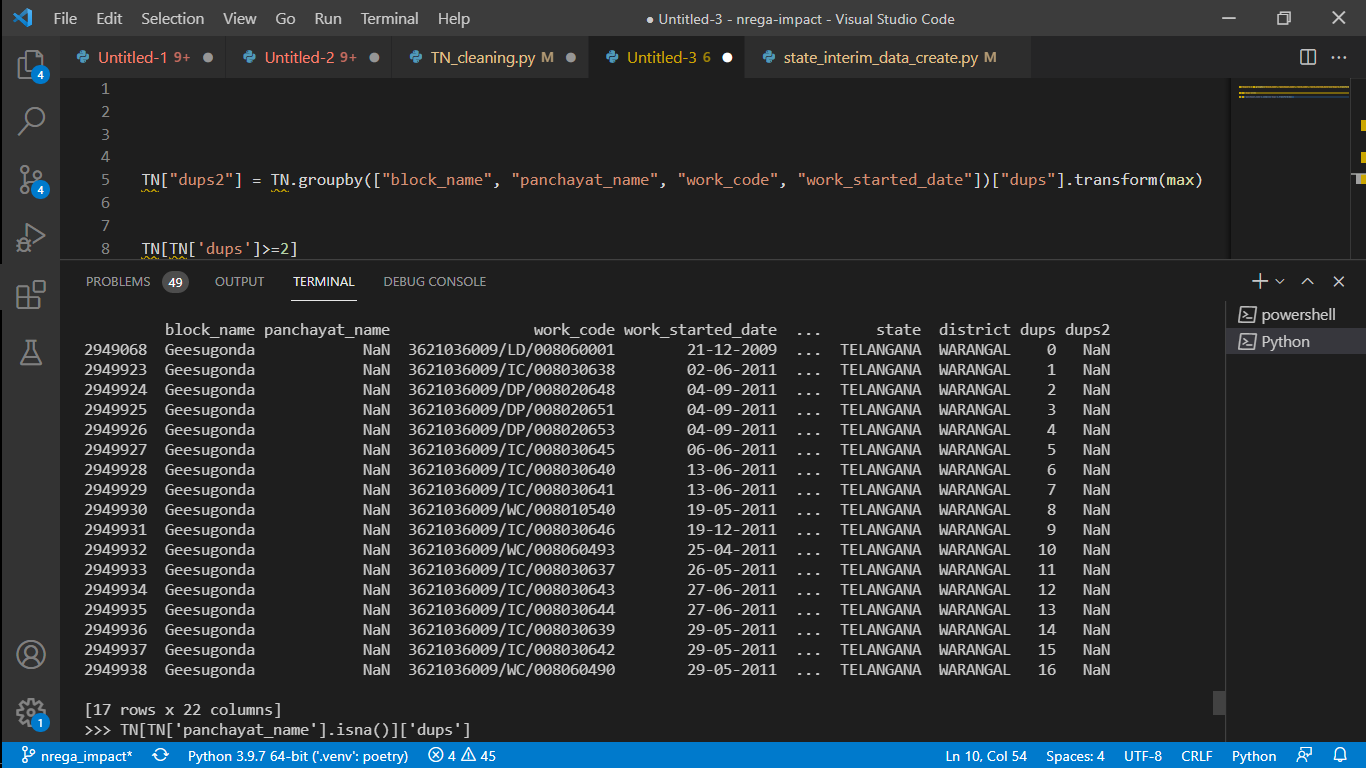
**The above corrections would be further referred as “Basics” and any state file specific correction would be addressed accordingly.**

|  |  |  |
| --- | --- | --- |
| **State** | **Issues Rectified** | **Status** |
|  |  |  |
| Andaman & Nicobar | Basics | Cleaned |
| Andhra Pradesh | Basics | Cleaned |
| Arunachal Pradesh | Basics | Cleaned |
| Assam | Basics | Cleaned |
| Dadra & Nagar Haveli | Doesn’t have any entries where work\_status is “Completed” or “Physically Completed”. | Download again |
| Goa | Basics | Cleaned |
| Gujarat | Basics | Cleaned |
| Haryana | Basics | Cleaned |
| Himachal Pradesh | Basics | Cleaned |
| J&K | Has issues with date line trails. Some entries of both date columns are clubbed in one. However, after filtering for work\_status such rows have automatically been eliminated. | Cleaned |
| Jharkhand | Basics | Cleaned |
| Karnataka | Basics | Cleaned |
| Kerala | Has a district Bagalkote which doesn’t belong to Kerala. I believe this can also happen in other states however, can’t verify all. That district and its block is in the Karnataka file. The corresponding block Kurwai is in MP and only rows with this block name is having dups. This specific block\_name has been filtered out in the specific\_rows\_filter. | Cleaned |
| Lakshadweep | Basics | Cleaned |
| Madhya Pradesh | Missing values in district column for block “BABAI CHICHLI”. This belongs to the NARSINGHPUR district which has another set of observations including the above block and I suspect that these are duplicates of the existing ones and also without enough info. The row with missing observations for the specific block\_name has been filtered out in the specific\_rows\_filter. | Cleaned |
| Manipur | Basics | Cleaned |
| Meghalaya | Basics | Cleaned |
| Mizoram | Basics | Cleaned |
| Nagaland | Basics | Cleaned |
| Odisha | Basics | Cleaned |
| Puducherry | Basics | Cleaned |
| Punjab | Basics | Cleaned |
| Rajasthan | Panchayat\_name contains Unicode. The Unicode based on Devanagiri script has been transliterated to English using adequate measures (Scraping and external script and data dependencies are involved). Refer Unicode Dictionary Documentation.docx available in the reports sub directory. | Cleaned |
| Sikkim | Basics | Cleaned |
| Tamil Nadu | Basics | Cleaned |
| Telangana | Panchayat\_name is missing for 17 observations in the district of Geesugonda. Process for dups2 failing at this point. However, dups is working. Check screen snip at end of this document. These 17 rows have been removed using the special\_rows\_filter. | Cleaned |
| Tripura | Basics | Cleaned |
| Uttar Pradesh | Basics | Cleaned |
| Uttarakhand | Basics | Cleaned |
| West Bengal | 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”.

**Screen snip for Telangana panchayat\_name==’NaN’:**



Process: *Sending interim data to processed*

**Data Overview**

The files are named according to the corresponding States and cleaned from the above process.

**Data Cleaning**

Script to be used: *state\_interim\_to\_processed.py*

There are two steps involved in this stage of cleaning:

1. Converting the following columns from string to integers: master\_work\_category\_name, work\_category\_name, work\_status, finished\_when, is\_secure.
2. Removing the columns work\_type and agency\_name.

The category codes for the converted variable are available in the Codes Document.docx available in the reports sub directory.