FIT5192 – Data Wrangling – S2 2016
Assignment 3

Data Integration and Reshaping

Submitted By
Ramprasath Karunakaran
(26994437)
on 30-09-2016

INTRODUCTION

The following report discusses the data integration task performed using python language for two given datasets in Disasters.csv and Disasters_Recent.csv file. The python program discussed in this report imports the two datasets, resolves pre-merging conflicts, integrates the two dataset and again resolve the post-merge conflicts. After the integration of two datasets, a sample of the merged dataset satisfying the given query is plotted in a World Map using Matplotlib's Basemap.

Programming Environment: Python 2.7.12 and Jupyter Notebook Anaconda version 2
Python libraries used: pandas, re, datetime, datetime, matplotlib,numpy, Basemap
Input Data files: Disasters.csv, Disasters_Recent.csv
Output Data file: 26994437_data_integrated.csv

2. Task 1 – Data Integration

Two given datasets are loaded into two different dataframes - disaster_df and disaster_recentdf. Initially, the problem noted in disaster_recentdf is:

- Two column headers 'Country iso' and 'Country name' are merged into one column header 'Country iso;Country name' resulting in an unnamed column at the end of the dataframe.
- To resolve this, the column headers are renamed.

Before proceeding with data merge, two schemas are compared and the synonymous columns which can be mapped are listed.

- 'Type' and 'disaster type' columns holds info about disaster types.
- 'Sub_Type' and 'disaster subtype' columns contain information about disaster subtypes.
- 'Location' and 'region' can be merged because region information contains the exact location of disaster.
- 'Killed' and 'Total deaths' can be mapped as it has information about death toll.
- 'Affected' and 'Total affected' can be merged as it is the total of affected, homeless and injured.
- 'International support' in both the datasets can be mapped but the unit of those values are not found.

Conflicts before merging:

Date Conflicts:

Start and End columns in disaster_df will be empty after merging, as disaster_recentdf has only year of occurrence.

Solution: Add two new columns 'Start' and 'End' to disaster_recentdf. Since dates and months are not in the second dataset, set day and month to '01'.

Countries Conflict:

Comparing countries in two datasets, there are 3260 rows not matching. Two different naming conventions are followed in the datasets.

Solution:

- Choose the correct convention among the two datasets. The first dataset's source is not given. The recent disaster set is imported from EM-DAT - The International Disaster Database. Hence, convention followed in the second dataset is chosen
- A loop is written to find the closest match in the recent dataframe and replace it in the disaster_df.
- Some complete mismatches in countries such as C�te d�Ivoire which has special characters in between are to be replaced manually using a lookup table.

Country ISO Conflict:

There are no ISO country names for the first dataset. Only the recent disaster data has ISO names. This ISO name will be useful while plotting the countries in mash map because the naming convention of countries in shape files and merged dataset are different.

Solution:

- Make a list of countries with ISO names using the unique ISO names in the recent dataset.
- Map the list to the disaster_df.
- Remaining ISO names are filled by the loop to find the closest match in the country_info dataframe and replace it in the disaster_df.

Continents Conflict:

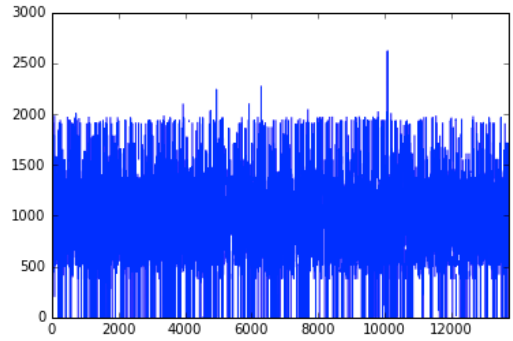
There are no continent names for the first dataset. Only the recent disaster data has ISO names. This continent name will be useful while analysing data by aggregating continent data.

Solution:

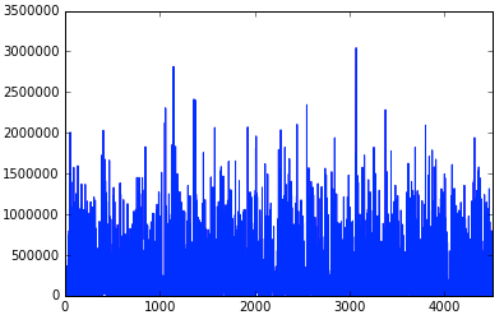
- Extract unique country ISO and continent from recent_df.
- Map them to country ISO in disaster_df.
- Country ISO having null values are extracted out and they are mapped using a manual look up dictionary.

Possible Conflict with International Support

Both the datasets have different ranges of values.



The above picture is the plot of disaster_df which has an average of 1014.79 whereas disaster_recentdf has a different plot with an average of 327835.5.



From the above statistics, range in the recent disaster set is quite high compared to the range in old disaster set. Since, there is no information regarding the unit of the values in that column from the source, some conversion in the units leading to difference in ranges is suspected. This problem can only be rectified if the source or unit of those values are made available.

Conflicts after merging:

After the two dataframes are merged using the union of keys from both frames(outer join), lot of conflicts arise, each one of them will be discussed briefly.

Subgroup and Group Conflict:

Subgroup and Group for disasters are null for the first 13000 rows as they are present only in the recent disaster set. It would be a better option to drop the two columns since most of the data is missing. But the data can be rolled up further during analysis if group and subgroup for a disaster is present.

First, to find subgroups, group by ‘type’ is done over existing data. Since there are lot of missing values, subgroups for some types are missing.

Solution:

- The missing subgroups can be found from the classification document available in EM-DAT source.
- In the source classification, industrial accident, transport accident and miscellaneous accident are actually subgroups and the types are defined for each subgroups. But in the given data, types have been moved to the subtypes, leaving the subgroups empty. Instead of altering the original cleaned dataset, both types and subgroups will be same for these particular types.
- There is no type as 'Complex Disasters' in the classification. Hence, we make both subgroup and group as 'Complex Disasters' since there are only two rows for that type.
- The following table is used to assign subgroups and groups to missing values.

Type	Subgroup	Group
Animal Accident	Biological	Natural
Drought	Climatological	Natural
Earthquake	Geophysical	Natural
Epidemic	Biological	Natural
Extreme Temperature	Meteorological	Natural
Flood	Hydrological	Natural
Impact	Extraterrestrial	Natural
Industrial Accident		Technological
Insect Infestation	Biological	Natural
Landslide	Hydrological	Natural
Mass Movement	Geophysical	Natural
Miscellaneous Accident		Technological
Storm	Meteorological	
Transport Accident		Technological
Volcanic Activity	Geophysical	Natural
Wildfire	Climatological	Natural
Complex Disasters		Complex

Year Column Conflict:
Solution: From the start dates in the dataset, year can be extracted and imputed into the empty year values in the dataset.

Conflicts with Affected Column:
From the EM-DAT source, it is evident that the columns Affected, Injured and Homeless are the subset of Total Affected columns. All the three columns values are added to give total affected column.
But those three columns have only 4503 rows. Imputing those columns with a value like 0, will affect the analysis. Having them will also affect the analysis since there are lot of missing values. Therefore, it is decided to drop those three columns as the clubbed information is available on Affected column.

Conflicts with Total Damage and Occurrence columns:

- Total damage column is found only in recent dataset. There is no way to determine the damages cost of the disaster with only 4000 rows of data. Hence, it is not included in the global schema.
- There is no occurence column in disaster_df resulting in 13000 missing rows of occurence values. When tried to find how occurence can be derived by doing 'group by' on a set of columns, only 8440 rows were given.

- Hence, it is not sure of how this column is derived. Even the source did not have explanatory notes about this column. So it is not included in the global schema.

Deciding the global key:

First, a numeric primary key starting from 10000 is generated. To relate semantically with the data, country ISO name and year information can be combined with the key. Hence, unique identifier is assigned for the dataset.
The global schema is :
(ID, Start, End, Country, Location, Type, Sub_type, Names, Killed, Affected,International Support, Country iso, continent, year, disaster group, disaster subgroup)
The merged data is written to a file named 26994437_data_integrated.csv.

Task 2: Mash up map:

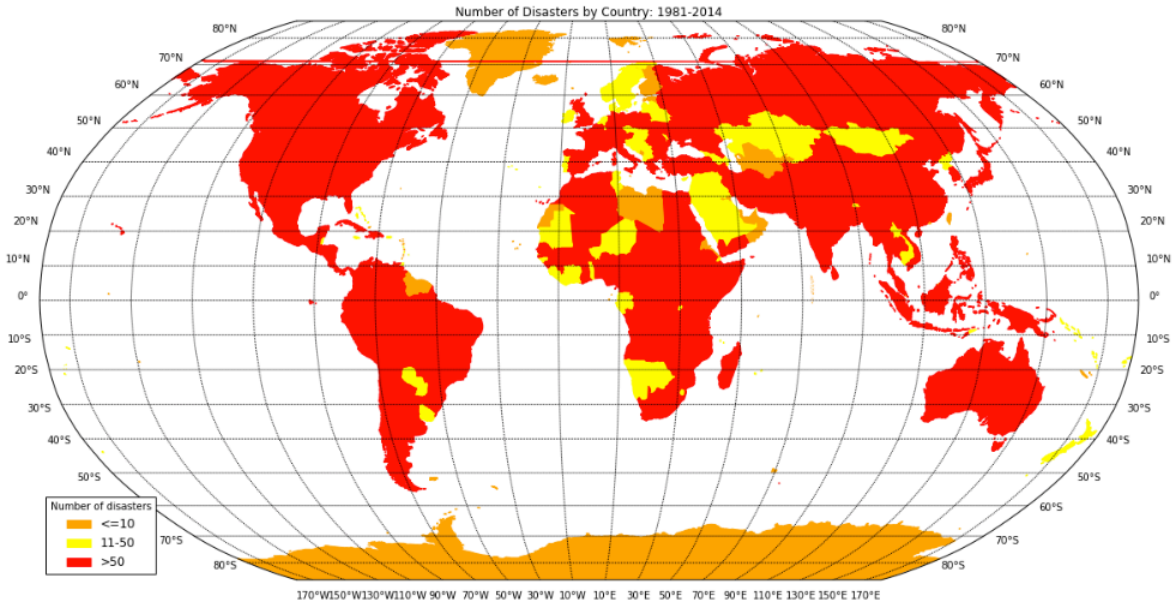
The query is to find the number of disasters between 1980-2015 in each country. Once the dataset to be plotted is extracted, ranges have to be decided to assign each range a colour for plotting.

Deciding ranges:

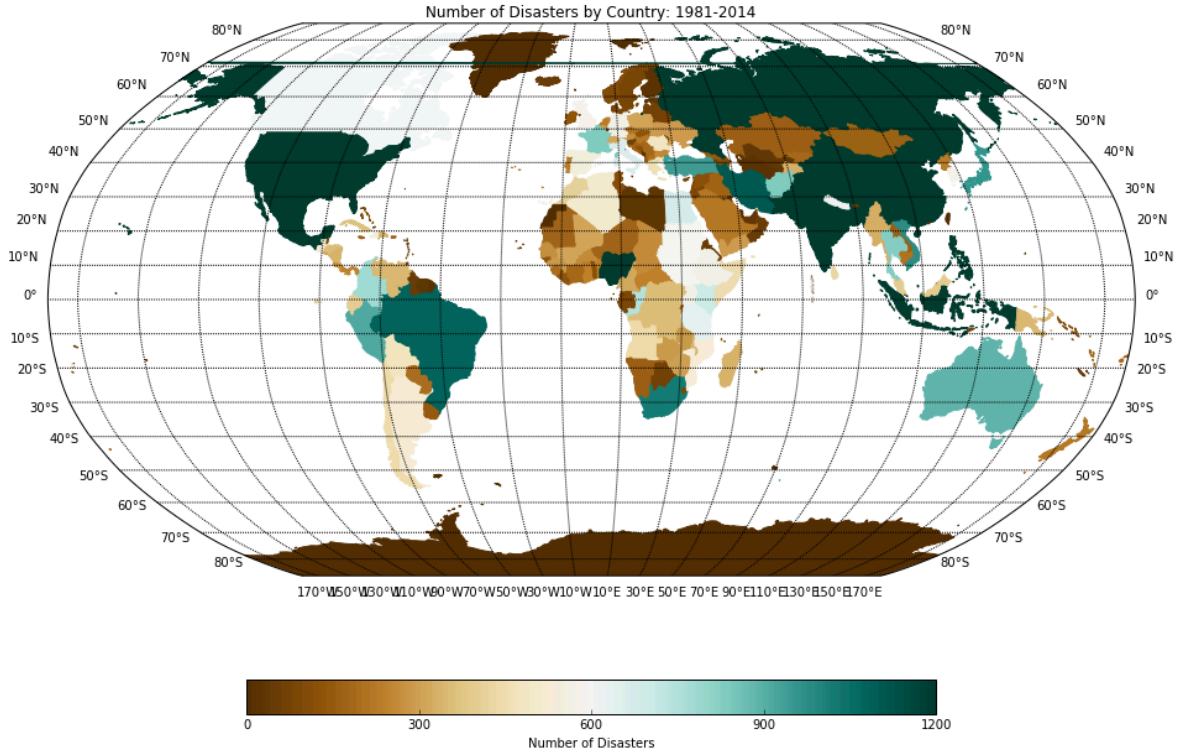
Using multiple queries, we get the following facts.

- Total number of disasters matching the query - 217
- Minimum number of disaster in the set - 1
- Maximum number of disaster in the set - 1152
- Number of disasters less than or equal to 10 - 52
- Number of disasters between 10 and 50 - 78
- Number of disasters greater than 50 – 87

Three equal sized bins are found and is ready to be plotted. Less than 10 is assigned ‘orange’, greater than 50 is assigned ‘yellow’ and in-between values are assigned ‘red’. Using Basemap in matplotlib, the following plot is obtained.



With sequential coloring and normalising the data between 0 and 1, the following plot is obtained.



From the above plot, lot of insights can be made than the first plot since there is a sequential spread of colours. Some of the insights found are:

- Very high number of disasters have occurred in most parts of Asia, United States and parts of Australasia.
- Very low number of disasters near the poles.
- In Africa, all ranges of disasters have occurred since there are multiple colors wide across the range of color bar.
- Other than the Australasian Islands, remaining islands have predominantly less number of disasters.

Conclusion:

It is very essential to know the source for two datasets while merging. Without a source or meta-data dictionary, it becomes very difficult to resolve conflicts arising while merging. Because of this problem, many columns like occurrence and total damage were dropped.