FIT5196-Data Wrangling
Assessment Task 1 Report

Exploring Raw Data

Submitted by
Ramprasath Karunakaran
(26994437)
on
10/08/2016

1. INTRODUCTION

The main objective of the task is to explore the given raw data and parse through the data and clean it to obtain further insights from the data. The given dataset is in the file 'data.dat'. To process and clean this file, we make use of the pandas and numpy library in Python Language (Version 2.7.12) and the environment needed is Jupyter Notebook(Anaconda version 4.1.1). The following sections will explain the steps involved in cleaning the given file and loading the data into a new CSV file.

2. EXPLORING RAW DATA

First step is to have a look at the file 'data.dat' which has around 14000 lines in them.

```
1  Start = 10061965,End = 10061965,Country:Japan,Location : nan,Type:Storm,Sub_Type:Tropical cyclone,Names:Dinah,Killed = 61.
   0,Affected = 30000.0,Cost = nan,ID = 1965-0036
2  Start = 08092004,End = 08092004,Country:Grenada,Location : nan,Type:Storm,Sub_Type:Tropical cyclone,Names:Ivan,Killed = 39.
   0,Affected = 60000.0,Cost = 889.0,ID = 2004-0462
3  Start = 15031995,End = 15031995,Country:Russia,Location : Kalmoukie:Daghestan (Mer ...,Type:Transport
   Accident,Sub_Type:Water,Names:nan,Killed = 52. 0,Affected = nan,Cost = nan,ID = 1995-0362
4  Start = 00021983,End = 00021983,Country:Cuba,Location : Santiago de Cuba:Pinar d ...,Type:Flood,Sub_Type:nan,Names:nan,Killed =
   15. 0,Affected = 164575.0,Cost = 60.0,ID = 1983-0042
5  Start = 00001983,End = 00001983,Country:Nepal,Location : nan,Type:Mass movement wet,Sub_Type:Landslide,Names:nan,Killed = 21.
   0,Affected = nan,Cost = nan,ID = 1983-0526
6  Start = 26071996,End = 26071996,Country:Russia,Location : Volgorgrad,Type:Industrial Accident,Sub_Type:Gas Leak,Names:Oil
   refinery,Killed = nan,Affected = nan,Cost = nan,ID = 1996-0209
7  Start = 19062003,End = 19062003,Country:Turkey,Location : Kayseri,Type:Miscellaneous
   accident,Sub_Type:Explosion,Names:Dormitory of a school,Killed = 10. 0,Affected = 13.0,Cost = nan,ID = 2003-0296
8  Start = 16022003,End = 22022003,Country:Pakistan,Location : Baluchistan:Sindh:North
   ...,Type:Storm,Sub_Type:nan,Names:nan,Killed = 51. 0,Affected = 2557.0,Cost = nan,ID = 2003-0086
9  Start = 00081986,End = 00081986,Country:Honduras,Location : Northeast Honduras/Mosqui
   ...,Type:Flood,Sub_Type:nan,Names:nan,Killed = nan,Affected = 30000.0,Cost = nan,ID = 1986-0093
10 Start = 13102007,End = 13102007,Country:Colombia,Location : Near Suarez (Cauca),Type:Industrial
   Accident,Sub_Type:Collapse,Names:Cold mine,Killed = 22. 0,Affected = 24.0,Cost = nan,ID = 2007-0500
11 Start = 26092002,End = 26092002,Country:Senegal,Location : Au large de la Gambie,Type:Transport
   Accident,Sub_Type:Water,Names:Ferry 'Joola',Killed = 1200. 0,Affected = nan,Cost = nan,ID = 2002-0622
12 Start = 22051997,End = 22051997,Country:Mexico,Location : Arteaga:Patzcuaro:Micho ...,Type:Earthquake (seismic
   activity),Sub_Type:Earthquake (ground shaking),Names:nan,Killed = nan,Affected = 12000.0,Cost = nan,ID = 1997-0116
13 Start = 28032004,End = 11042004,Country:Canada,Location : Manitoba,Type:Flood,Sub_Type:General flood,Names:nan,Killed =
   nan,Affected = 1000.0,Cost = nan,ID = 2004-0146
14 Start = 20021943,End = 20021943,Country:Mexico,Location : nan,Type:Volcano,Sub_Type:Volcanic eruption,Names:Paricutin,Killed =
   nan,Affected = 3000.0,Cost = nan,ID = 1943-0008
15 Start = 28051980,End = 28051980,Country:Canada,Location : Webb:Saskatchewan,Type:Transport
   Accident,Sub_Type:Road,Names:nan,Killed = 22. 0,Affected = 11.0,Cost = nan,ID = 1980-0212
16 Start = 09082003,End = 11082003,Country:Algeria,Location : Reggane (Adrar region):T ...,Type:Flood,Sub_Type:General
   flood,Names:nan,Killed = 13. 0,Affected = nan,Cost = nan,ID = 2003-0396
17 Start = 02082000,End = 30082000,Country:India,Locations : Gujarat:Andhra Pradesh: ...,Type:Flood,Sub_Type:General
   flood,Names:nan,Killed = 867. 0,Affected = 22000000.0,Cost = 43.0,ID = 2000-0445
18 Start = 25092008,End = 28092008,Country:Viet Nam,Location : Lang Son:Son La:Bac Gia ...,Type:Storm,Sub_Type:Tropical
   cyclone,Names:Typhoon 'Hagupit' (Nina),Killed = 46. 0,Affected = 58511.0,Cost = 63.0,ID = 2008-0426
```

By looking at the data, we understand the following things:

1. Each row has values separated by commas along with their respective keys.
2. The first two values Start and End are the dates which is actually not in a date format. We confirm this in one of the following sections.
3. Another ambiguous column found here is the Location. Each Location has multiple values separated by colon.
4. Even though they are in key value pairs. The assignment symbols are different such as ':' and '='.

3. PROCESSING DATA

The next step would be to load the data into a data frame in Pandas. We do this because Data Frames take the entire set as an object and allows us to parse it more efficiently. When we try to load the data into a data frame, we get an error saying that the parser saw 12 fields instead of the 11 fields. By looking at those lines mentioned, we find that there are more commas separating the location attribute. We can rectify this issue by skipping the 22 lines which has more fields. 22 lines out of 14,328 lines is 0.15% percent of the data set. Hence, skipping them will not be affecting our decision making process at the end.

Once the data has been loaded, we can use the replace function in the dataframe to replace all the ':' with '=' for making the notations uniform in the entire dataset. After this step, the dataframe will look as follows.

	0	1	2	3	4	5	6	7	8
0	Start = 10061965	End = 10061965	Country=Japan	Location = nan	Type=Storm	Sub_Type=Tropical cyclone	Names=Dinah	Killed = 61.0	Affected = 30000.0
1	Start = 08092004	End = 08092004	Country=Grenada	Location = nan	Type=Storm	Sub_Type=Tropical cyclone	Names=Ivan	Killed = 39.0	Affected = 60000.0
2	Start = 15031995	End = 15031995	Country=Russia	Location = Kalmoukie=Daghestan (Mer ...	Type=Transport Accident	Sub_Type=Water	Names=nan	Killed = 52.0	Affected = nan
3	Start = 00021983	End = 00021983	Country=Cuba	Location = Santiago de Cuba=Pinar d ...	Type=Flood	Sub_Type=nan	Names=nan	Killed = 15.0	Affected = 164575.0
	Start =	End =			Type=Mass			Killed	Affected =

Now we start to clean the columns one by one. The location column is loaded in a dataframe and again a replace method is used to replace the '=' with ','. After that, using 'str' method we perform lstrip and rstrip to remove 'Location =' and '…' from each value. The cleaned column would look like below.

```
Out[6]: 0
1
2                Kalmoukie,Daghestan (Mer
3                Santiago de Cuba,Pinar d
4
5                              Volgorgrad
6                                 Kayseri
7              Baluchistan,Sindh,North
8            Northeast Honduras/Mosqui
9                    Near Suarez (Cauca)
10                 Au large de la Gambie
11             Arteaga,Patzcuaro,Micho
12                               Manitoba
13
14                     Webb,Saskatchewan
15             Reggane (Adrar region),T
16          s , Gujarat,Andhra Pradesh,
17               g Son,Son La,Bac Gia
18                                 North
19                             South,East
20                               eningrad
21             United Republic of Rukwa
22                  Cajamarca department
23                                   Fez
24           Abricots region (Grand'An
25                                uisiana
26             Franklin,Jefferson count
27               Buriganga river (North)
28                          Near Acapulco
29                           Huize County
                        ...
14298                s , Northwest,Bihar
14299            Cancun,Puerto Maderos,
14300                        Basse-Kotto
14301             Mererani (Arusha region)
14302                          Aden Gulf
14303        Qijiang district (Chongqi
```

Now the first two columns are left stripped to remove 'Start = ' and 'End = '.We use to_datetime to convert the column into date type. Now we get an error when we try to convert. This is because more than 300 rows have '00' as a date which is invalid. Skipping 300 lines(2%) would affect the analysis process. Hence we replace those dates with empty values(Pandas read them as NaT).

Now we concatenate the three cleaned columns to the original dataframe and drop the ambiguous columns. Now the dataframe will look like this.

	0	1	2	3	4	5	6	7	8
0	1965-06-10	1965-06-10	Country=Japan		Type=Storm	Sub_Type=Tropical cyclone	Names=Dinah	Killed = 61.0	Affec 3000
1	2004-09-08	2004-09-08	Country=Grenada		Type=Storm	Sub_Type=Tropical cyclone	Names=Ivan	Killed = 39.0	Affec 6000
2	1995-03-15	1995-03-15	Country=Russia	Kalmoukie,Daghestan (Mer	Type=Transport Accident	Sub_Type=Water	Names=nan	Killed = 52.0	Affec nan
3	NaT	NaT	Country=Cuba	Santiago de Cuba,Pinar d	Type=Flood	Sub_Type=nan	Names=nan	Killed = 15.0	Affec 1645
4	NaT	NaT	Country=Nepal		Type=Mass movement wet	Sub_Type=Landslide	Names=nan	Killed = 21.0	Affec nan
5	1996-07-26	1996-07-26	Country=Russia	Volgorgrad	Type=Industrial Accident	Sub_Type=Gas Leak	Names=Oil refinery	Killed = nan	Affec nan
6	2003-06-19	2003-06-19	Country=Turkey	Kayseri	Type=Miscellaneous accident	Sub_Type=Explosion	Names=Dormitory of a school	Killed = 10.0	Affec 13.0
								Killed	

Instead of using lsplit on all the remaining columns, we can remove by using a function which gives us the value after the pattern '= '. To do this, we move the data to a temporary .txt file and pull it back to a dataframe but this time we call the function in the converters argument of the read_csv method to remove all the key values. At the end, we name all the columns in the dataframe.

	Start	End	Country	Location	Type	Subtype	Names	Killed	Affected	Cost	ID
0	1965-06-10	1965-06-10	Japan		Storm	Tropical cyclone	Dinah	61.0	30000.0	nan	1965-0036
1	2004-09-08	2004-09-08	Grenada		Storm	Tropical cyclone	Ivan	39.0	60000.0	889.0	2004-0462
2	1995-03-15	1995-03-15	Russia	Kalmoukie,Daghestan (Mer	Transport Accident	Water	nan	52.0	nan	nan	1995-0362
3			Cuba	Santiago de Cuba,Pinar d	Flood	nan	nan	15.0	164575.0	60.0	1983-0042
4			Nepal		Mass movement wet	Landslide	nan	21.0	nan	nan	1983-0526
5	1996-07-26	1996-07-26	Russia	Volgorgrad	Industrial Accident	Gas Leak	Oil refinery	nan	nan	nan	1996-0209
6	2003-06-19	2003-06-19	Turkey	Kayseri	Miscellaneous accident	Explosion	Dormitory of a school	10.0	13.0	nan	2003-0296
7	2003-02-16	2003-02-22	Pakistan	Baluchistan,Sindh,North	Storm	nan	nan	51.0	2557.0	nan	2003-0086
8			Honduras	Northeast Honduras/Mosqui	Flood	nan	nan	nan	30000.0	nan	1986-0093
9	2007-10-13	2007-10-13	Colombia	Near Suarez (Cauca)	Industrial Accident	Collapse	Cold mine	22.0	24.0	nan	2007-0500
10	2002-09-26	2002-09-26	Senegal	Au large de la Gambie	Transport Accident	Water	Ferry 'Joola'	1200.0	nan	nan	2002-0622
11	1997-05-22	1997-05-22	Mexico	Arteaga,Patzcuaro,Micho	Earthquake (seismic activity)	Earthquake (ground shaking)	nan	nan	12000.0	nan	1997-0116

We notice some empty cells in the dataframe which can be replaced by NaN values in the numpy library. And also the datatypes of all the columns are objects. So we convert the columns Killed, Affected and Cost to numeric to gain some insights. This is done using

to_numeric method in pandas. But we get an error saying Killed column cannot be parsed. This is because each value in the column has a space after the decimal value. Hence this column is cleaned by using split and replace methods. Now all the values are in the integer format. Now we can convert the required columns to float64 type. After this conversion, the describe method in dataframe would return the following.

	Killed	Affected	Cost
count	1.129000e+04	8.869000e+03	3018.000000
mean	2.772850e+03	5.881095e+05	481.833042
std	7.754650e+04	7.228792e+06	3323.496327
min	9.999990e-01	1.000000e+00	0.003000
25%	NaN	NaN	NaN
50%	NaN	NaN	NaN
75%	NaN	NaN	NaN
max	5.000000e+06	3.000000e+08	125000.000000

Finally, the dataframe would be in the following format.

	Start	End	Country	Location	Type	Subtype	Names	Killed	Affected	Cost	ID
0	1965-06-10	1965-06-10	Japan	NaN	Storm	Tropical cyclone	Dinah	61.0	30000.0	nan	1965-0036
1	2004-09-08	2004-09-08	Grenada	NaN	Storm	Tropical cyclone	Ivan	39.0	60000.0	889.0	2004-0462
2	1995-03-15	1995-03-15	Russia	Kalmoukie,Daghestan (Mer	Transport Accident	Water	nan	52.0	nan	nan	1995-0362
3	NaN	NaN	Cuba	Santiago de Cuba,Pinar d	Flood	nan	nan	15.0	164575.0	60.0	1983-0042
4	NaN	NaN	Nepal	NaN	Mass movement wet	Landslide	nan	21.0	nan	nan	1983-0526
5	1996-07-26	1996-07-26	Russia	Volgorgrad	Industrial Accident	Gas Leak	Oil refinery	nan	nan	nan	1996-0209

This cleaned dataframe has to be loaded into a new csv file named ‘xx_parsed_data.csv’ and it’s file content would be something like below.

```
0,1965-06-10,1965-06-10,Japan,,Storm,Tropical cyclone,Dinah,61.0,30000.0,, 1965-0036
1,2004-09-08,2004-09-08,Grenada,,Storm,Tropical cyclone,Ivan,39.0,60000.0,889.0, 2004-0462
2,1995-03-15,1995-03-15,Russia,"Kalmoukie,Daghestan (Mer ",Transport Accident,Water,nan,52.0,,, 1995-0362
3,,,Cuba,"Santiago de Cuba,Pinar d ",Flood,nan,nan,15.0,164575.0,60.0, 1983-0042
4,,,Nepal,,Mass movement wet,Landslide,nan,21.0,,, 1983-0526
5,1996-07-26,1996-07-26,Russia,Volgorgrad,Industrial Accident,Gas Leak,Oil refinery,,,, 1996-0209
6,2003-06-19,2003-06-19,Turkey,Kayseri,Miscellaneous accident,Explosion,Dormitory of a school,10.0,13.0,, 2003-0296
7,2003-02-16,2003-02-22,Pakistan,"Baluchistan,Sindh,North ",Storm,nan,nan,51.0,2557.0,, 2003-0086
8,,,Honduras,Northeast Honduras/Mosqui ,Flood,nan,nan,,30000.0,, 1986-0093
9,2007-10-13,2007-10-13,Colombia,Near Suarez (Cauca),Industrial Accident,Collapse,Cold mine,22.0,24.0,, 2007-0500
10,2002-09-26,2002-09-26,Senegal,Au large de la Gambie,Transport Accident,Water,Ferry 'Joola',1200.0,,, 2002-0622
11,1997-05-22,1997-05-22,Mexico,"Arteaga,Patzcuaro,Micho ",Earthquake (seismic activity),Earthquake (ground
shaking),nan,,12000.0,, 1997-0116
12,2004-03-28,2004-04-11,Canada,Manitoba,Flood,General flood,nan,,1000.0,, 2004-0146
13,1943-02-20,1943-02-20,Mexico,,Volcano,Volcanic eruption,Paricutin,,3000.0,, 1943-0008
14,1980-05-28,1980-05-28,Canada,"Webb,Saskatchewan",Transport Accident,Road,nan,22.0,11.0,, 1980-0212
15,2003-08-09,2003-08-11,Algeria,"Reggane (Adrar region),T ",Flood,General flood,nan,13.0,,, 2003-0396
16,2000-08-02,2000-08-30,India,"s , Gujarat,Andhra Pradesh, ",Flood,General flood,nan,867.0,22000000.0,43.0, 2000-0445
17,2008-09-25,2008-09-28,Viet Nam,"g Son,Son La,Bac Gia ",Storm,Tropical cyclone,Typhoon 'Hagupit' (Nina),46.0,58511.0,63.0,
2008-0426
18,1994-05-18,1994-05-18,Bangladesh,North,Storm,nan,nan,15.0,100.0,, 1994-0122
19,2004-12-04,2004-12-04,Taiwan (China),"South,East",Storm,Tropical cyclone,Nanmadol (Yoyong/30W),3.0,,, 2004-0618
20,1991-02-23,1991-02-23,Soviet Union,eningrad,Miscellaneous accident,Fire,Hotel,17.0,,, 1991-0042
21,1910-10-13,1910-10-13,Tanzania Uni Rep,United Republic of Rukwa,Earthquake (seismic activity),Earthquake (ground
shaking),nan,,,, 1910-0018
22,,,Peru,Cajamarca department,Drought,Drought,nan,,,, 2004-9063
23,,,Morocco,Fez,Mass movement wet,Landslide,Cliff,1.0,12216.0,, 1982-0045
24,,,Haiti,Abricots region (Grand'An ,Flood,nan,nan,12.0,1200.0,, 2000-0797
25,1909-09-22,1909-09-22,United States,uisiana,Flood,Storm surge/coastal flood,nan,72.0,,, 1909-0017
26,2000-05-06,2000-05-09,USA,"Franklin,Jefferson count ",Flood,Flash flood,nan,2.0,2000.0,, 2000-0232
27,2003-04-14,2003-04-14,Bangladesh,Buriganga river (North),Transport Accident,Water,MV Mitali,130.0,,, 2003-0173
28,1974-06-17,1974-06-17,Mexico,Near Acapulco,Storm,Tropical cyclone,nan,13.0,,, 1974-0030
29,2005-05-08,2005-05-08,China,Huize County,Earthquake (seismic activity),Earthquake (ground shaking),nan,,18509.0,, 2005-0740
30,2005-02-28,2005-03-23,Nigeria,"Adamawa,Kano,Jigawa,Ba ",Epidemic,Viral Infectious Diseases,nan,561.0,23575.0,, 2005-0134
31,2001-08-15,2001-11-19,Viet Nam,"g An,Dong Thap,An Gi ",Flood,General flood,nan,310.0,1570270.0,84.0, 2001-0492
32,1987-10-14,1987-10-14,Venezuela,Near San Pedro river (Wes ,Flood,nan,nan,40.0,29.0,, 1987-0190
33,2005-04-17,2005-04-17,Switzerland,Valais,Transport Accident,Road,nan,12.0,,, 2005-0201
```

4. Conclusion

Though the cleaned data is in a proper table format, there are lot more things to be dealt with the dataset. The dataset is damaged so that there are many inconsistencies. For example, when we count the number of values in the start and end columns, they are of different values and they will be around 300 values short of the other columns' count. This means there are lots of missing dates. This can be avoided by keeping only the month and the year in the dataset but that would be like modifying the given data. And also huge amount of data is missing in the numeric columns and replacing them with appropriate values(mean, zero, etc.,) would definitely tamper with the analysis. Also, the locations have not been specified in the right format. Multiple locations would be difficult to parse when queries are written over the huge set of data. So they have to be split into separate columns. Once all these problems are properly dealt without damaging the originality, the data is fit for analysis.