Case Studies for Data Analytics

Assignment: 1

Date: February 6, 2019

Group Members:

Sai Krishna Lakshminarayanan (18230229)
Surya Balakrishnan Ramakrishnan (18231072)

Section 1: Satellite Image Classification

Classification of images:

Classification of LANDSAT 7 image from 2003

We start with loading all the obtained TIF files of the LANDSAT 7 images as a raster layer on top of each other. Once we have successfully loaded all the TIF the next step is to clip the data. In this case we are supposed to use bands 1 to 5 so we only select the TIF files for the corresponding bands required. This is done by selecting the clip multiple raster option. Once we have successfully clipped the raster layers of band 1 to 5 now, we preprocess the image by cropping the image for the given UL and LR Co-ordinates. The next step involves providing the meta data file which is the .MTL file as input. We obtained the following image in the 4-3-2 RGB format.

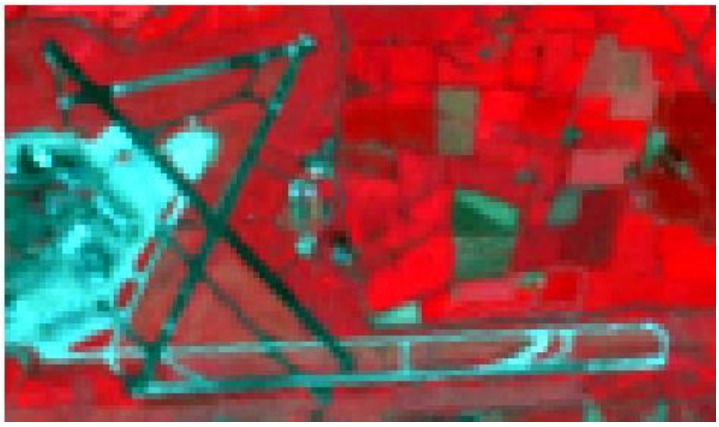


Fig 1 Image after clipping the required bands

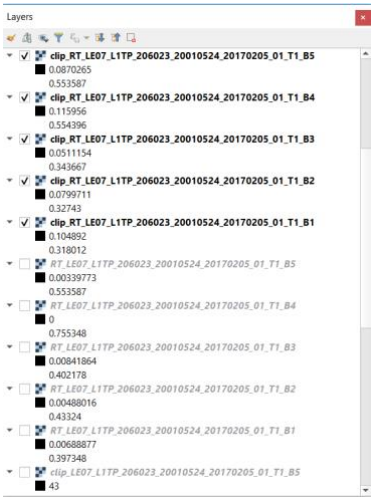


Fig 2 Corresponding clipped files

The next stage is to define band set and creating the training input file. We used the polygon selector tool to select training inputs for the image to be classified. We gave 7 sets of training data for each of the two classes to be classified. The classes which we chose was Built Up area and Non-Built up area. Now after defining the macro classification class we run the classification algorithm.

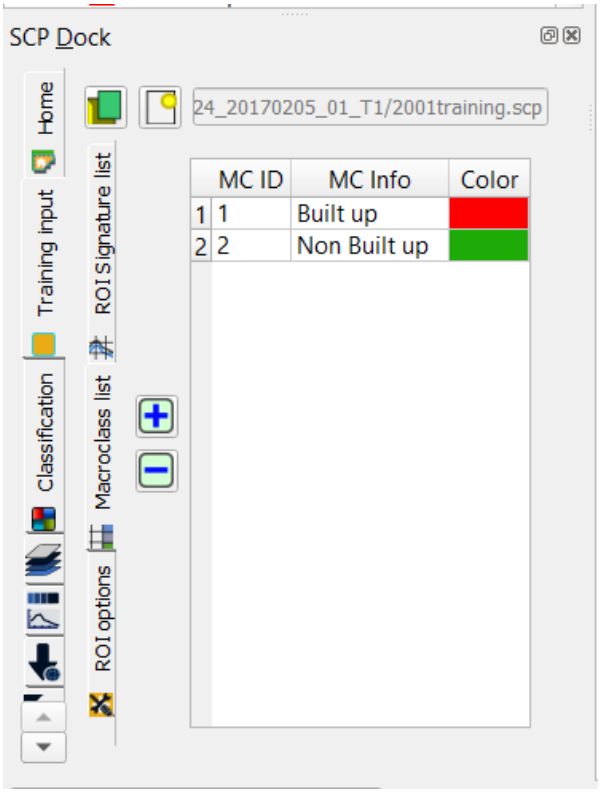


Fig 3 Classification of macro class

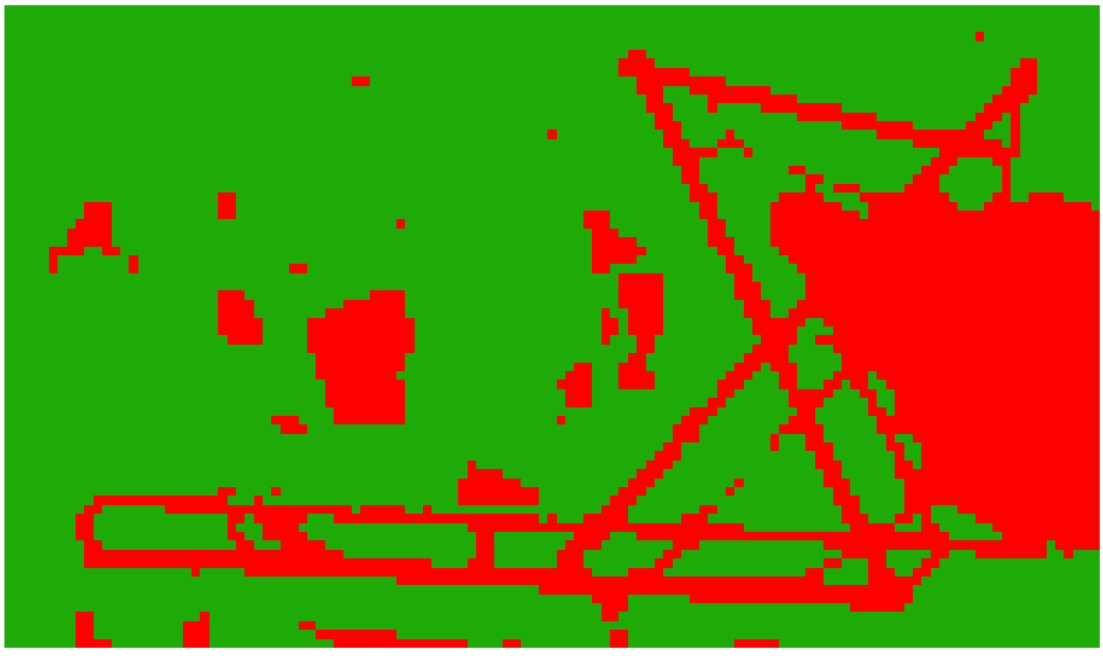


Fig 4 Classified LANDSAT 7 Image

Classification of LANDSAT 8 image from 2013

Repeating the same steps for the second image which is of the year 2013 captured from LANDSAT 8, but we choose bands 2 to 6. We get these following outputs after the above mentioned steps are repeated for this image.

Fig 5 Clipped Region of Landsat 8 image

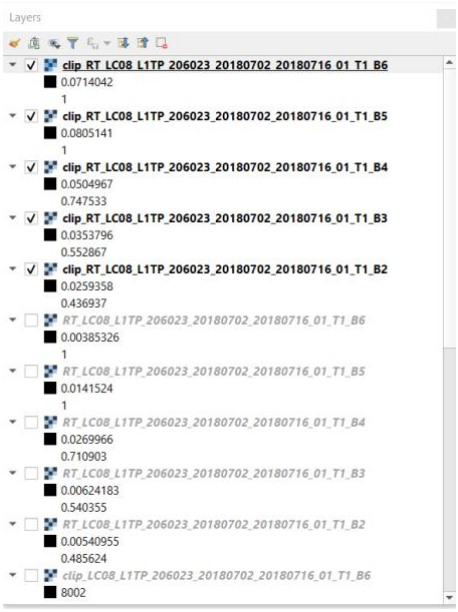


Fig 6 Files obtained after clipping

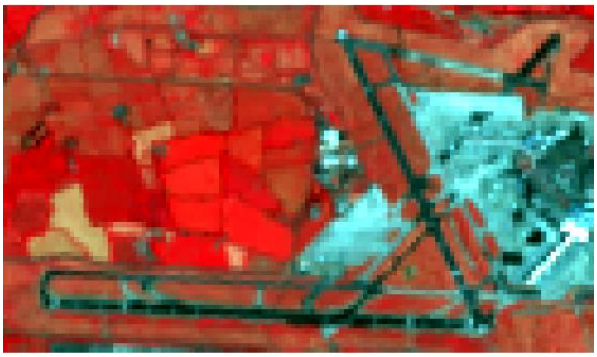


Fig 7 RGB 4-3-2 colour scheme of clipped image

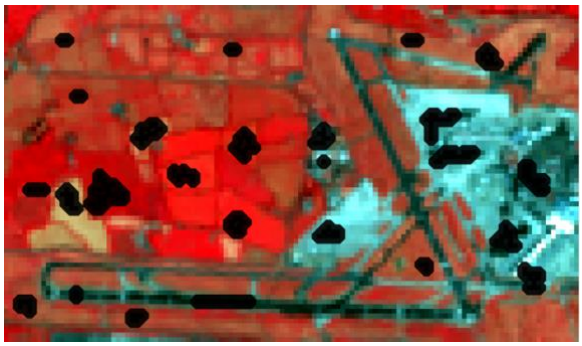


Fig 8 Training Inputs provided for classification

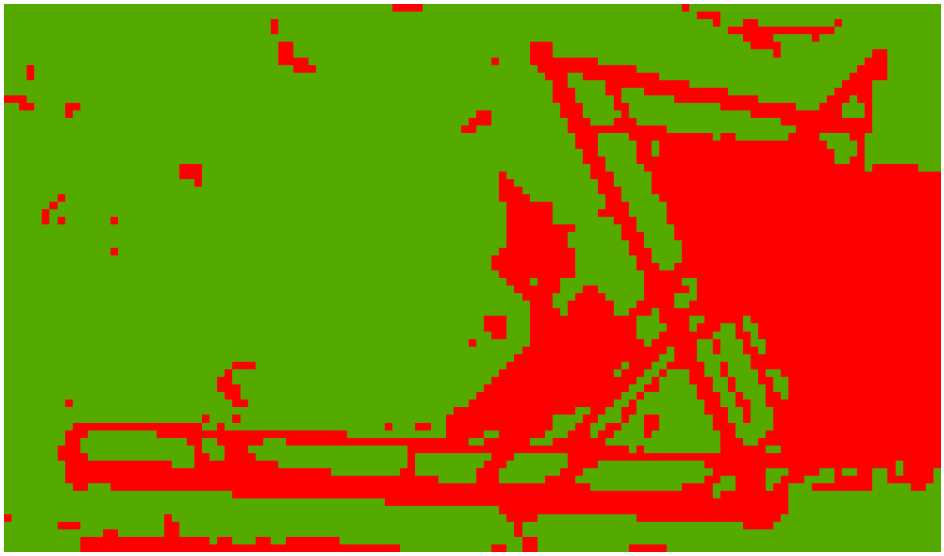
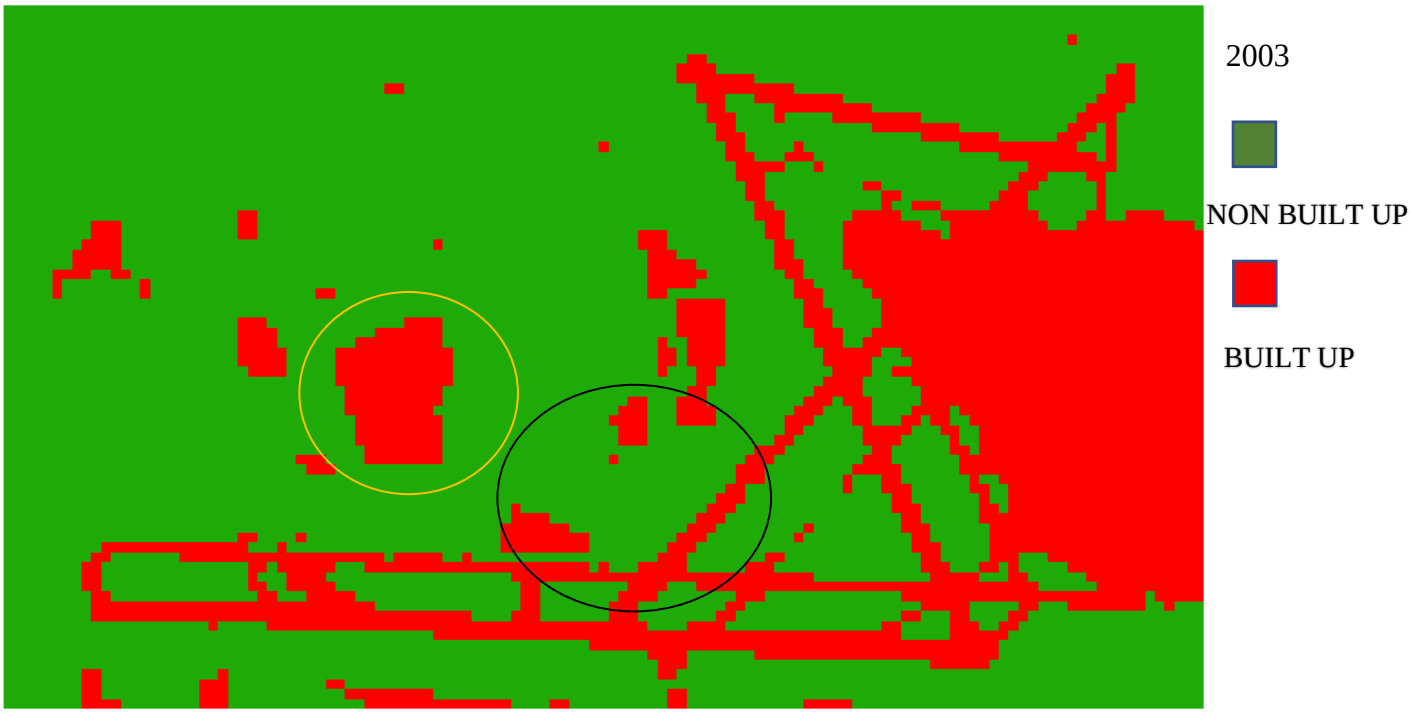
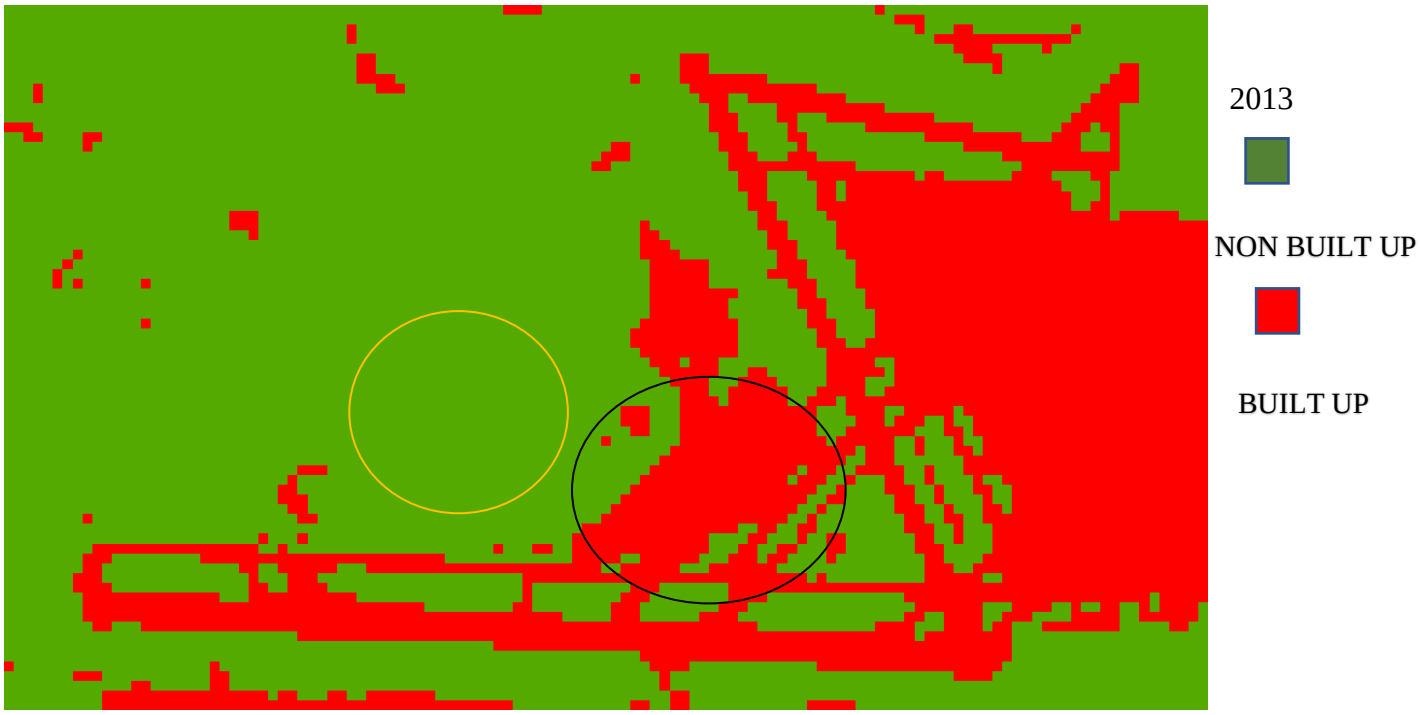


Fig 9 Classified LANDSAT 8 image

Comparison of both the images side by side:

Image 1 is LANDSAT 7 image from 2003 and Image 2 is LANDSAT 8 image from 2013





The yellow circle in the images above shows that in the year 2003 there was a region which either was built up and then taken down before 2013 or was wrongly classified. However, the black circle clearly shows that there was nothing built up in the year 2003 which in 2013 we can clearly see as classified as a built-up area.

Explanation for figures:

- Figure 1 is obtained once all the raster bands are clipped together.
- Figure 2 is the raster bands which represent the clipped image
- Figure 3 is the macro classification classes given to classify the clipped image
- Figure 4 is the obtained classified LANDSAT 7 image from the year 2003
- Figure 5 represents the raster image obtained after clipping in the RT data.
- Figure 6 is the raster bands which represent the clipped image
- Figure 7 Clipped image represented as RGB 4-3-2 colour scheme
- Figure 8 Training data input for classification.
- Figure 9 is the obtained classified LANDSAT 8 image from the year 2013.

Identification of the Region:

The clipped region of both the Landsat images is of **Dublin Airport**. Since the image is geo-tagged one can easily locate the geographic location using its latitude and longitude. For the purpose of cross verifying the same we used google earth.

Section 2: Multi Source Open Data Quality

Data Observation:

- Firstly, we load the DCC csv file using pandas
- Column 4 of the csv file has some junk value, so we remove it
- We use the head function to get the gist of the csv file

```
             PARK           AREA              CLUBNAME  \
0  ALBERT COLLEGE     NORTH WEST  DRUMCONDRA F.C (Snr)
1  ALBERT COLLEGE     NORTH WEST         GLASNEVIN AFC
2       BEECHILL      SOUTH EAST        BALLSBRIDGE FC
3         BELCAMP  NORTH CENTRAL        NEWTOWN CELTIC
4         BELCAMP  NORTH CENTRAL          VIANNEY BOYS

                    LEAGUE  Unnamed: 4
0  AMATEUR FOOTBALL LEAGUE         NaN
1  AMATEUR FOOTBALL LEAGUE         NaN
2  AMATEUR FOOTBALL LEAGUE         NaN
3  AMATEUR FOOTBALL LEAGUE         NaN
4  AMATEUR FOOTBALL LEAGUE         NaN


             PARK           AREA              CLUBNAME  \
0  ALBERT COLLEGE     NORTH WEST  DRUMCONDRA F.C (Snr)
1  ALBERT COLLEGE     NORTH WEST         GLASNEVIN AFC
2       BEECHILL      SOUTH EAST        BALLSBRIDGE FC
3         BELCAMP  NORTH CENTRAL        NEWTOWN CELTIC
4         BELCAMP  NORTH CENTRAL          VIANNEY BOYS

                    LEAGUE
0  AMATEUR FOOTBALL LEAGUE
1  AMATEUR FOOTBALL LEAGUE
2  AMATEUR FOOTBALL LEAGUE
3  AMATEUR FOOTBALL LEAGUE
4  AMATEUR FOOTBALL LEAGUE
```

Fig 1: Loading DCC csv and getting a gist of it.

- Now we load the DLR csv file which is the second csv file and get a gist of it

```
        Location Number Size   Latitude  Longitude
0  Kilbogget Park      1  Snr  53.257242  -6.140665
1             NaN      2  SSG  53.257614  -6.139882
2             NaN      3  SSG  53.257842  -6.139265
3             NaN      4  SSG  53.257098  -6.139094
4             NaN      5  SGG  53.256674  -6.140134
```

Fig 2: Loading DLR csv and getting a gist of it.

- We use the itertools to open the FCC xml file.
- To extract the data from xml file we extract the information from the xml file and parse the file to convert it into a columns of data frame.

- We use pandas to convert the xml file to csv format.

```
<?xml version="1.0" encoding="UTF-8"?>
<xml-tables>
<Playing_Pitches-table>
<Playing_Pitches>
<FACILITY_TYPE>All weather pitches</FACILITY_TYPE>
<FACILITY_NAME>Balbriggan Town Park</FACILITY_NAME>
<LOCATION>Balbriggan</LOCATION>
<LAT>53.6049596246817</LAT>
<LONG>-6.18235291959051</LONG>
</Playing_Pitches>
<Playing_Pitches>
<FACILITY_TYPE>All weather pitches</FACILITY_TYPE>
<FACILITY_NAME>Balheary Reservoir</FACILITY_NAME>
<LOCATION>Swords</LOCATION>
<LAT>53.4727096370551</LAT>
<LONG>-6.22301521551813</LONG>
</Playing_Pitches>
<Playing_Pitches>
<FACILITY_TYPE>All weather pitches</FACILITY_TYPE>
<FACILITY_NAME>Town Park</FACILITY_NAME>
```

	FACILITY_TYPE	FACILITY_NAME	Location	Latitude	Longitude
0	All weather pitches	Balbriggan Town Park	Balbriggan	53.604960	-6.182353
1	All weather pitches	Balheary Reservoir	Swords	53.472710	-6.223015
2	All weather pitches	Town Park	Skerries	53.577114	-6.111072
3	All weather pitches	St. Mologa's Park	Balbriggan	53.617667	-6.189368
4	Basketball Court	Seagrange Park	NaN	53.396667	-6.135352

Fig 3 Loading xml file parsing it to convert it to data frame for FCC

- Observing for any null values in all the three datasets.

```
PARK        0
AREA        0
click to expand output; double click to hide output
LEAGUE      0
dtype: int64
0
Location     46
Number        1
Size          3
Latitude      1
Longitude     1
dtype: int64
52
FACILITY_TYPE     0
FACILITY_NAME     0
Location         23
Latitude          0
Longitude         0
dtype: int64
23
```

Fig 4 Checking for null values in all the data sets.

Data Quality Enhancement

- For the DCC dataset we observe that there is no column with data representing latitude and longitude of the area.

- We create a temporary column named address and combine the values of park, club and area. Then we use geocoder to obtain the latitude and longitude of the addresses.
- Even after this step we observed a few missing values, additionally we gave league information to extract the required information.
- Finally, we obtain the desired DCC with latitude and longitude without any null value.

```
                PARK            AREA               CLUBNAME  \
0  ALBERT COLLEGE      NORTH WEST  DRUMCONDRA F.C (Snr)
1  ALBERT COLLEGE      NORTH WEST          GLASNEVIN AFC
2       BEECHILL       SOUTH EAST         BALLSBRIDGE FC
3         BELCAMP  NORTH CENTRAL         NEWTOWN CELTIC
4         BELCAMP  NORTH CENTRAL           VIANNEY BOYS

                     LEAGUE Latitude Longitude
0  AMATEUR FOOTBALL LEAGUE  53.3855  -6.26063
1  AMATEUR FOOTBALL LEAGUE  53.3832   -6.2634
2  AMATEUR FOOTBALL LEAGUE  53.3153  -6.23165
3  AMATEUR FOOTBALL LEAGUE  53.4016  -6.18566
4  AMATEUR FOOTBALL LEAGUE  53.1424  -7.69205
```

```
In [293]: DCC.isnull().sum()#cross checking for null values if any

Out[293]: PARK         0
          AREA         0
          CLUBNAME     0
          LEAGUE       0
          Latitude     0
          Longitude    0
          dtype: int64
```

Fig 5 Quality enhancement for DCC data set.

- In the FCC dataset location is not available for some of the values.
- So, we obtain the list of indexes where location is null.
- Now using for loop, we obtain the values of locations using google geocoder and geopy by the method of reverse geocoding.
- Finally, the resulting FCC data set is obtained without any missing or null value.

```
FACILITY_TYPE    0
FACILITY_NAME    0
Location         0
Latitude         0
Longitude        0
dtype: int64
```

```
print(FCC.head())#desired FCC final dataset
```

```
         FACILITY_TYPE         FACILITY_NAME    Location   Latitude  Longitude
0  All weather pitches  Balbriggan Town Park  Balbriggan  53.604960  -6.182353
1  All weather pitches    Balheary Reservoir      Swords  53.472710  -6.223015
2  All weather pitches             Town Park    Skerries  53.577114  -6.111072
3  All weather pitches     St. Mologa's Park  Balbriggan  53.617667  -6.189368
4     Basketball Court        Seagrange Park   Dublin 13  53.396667  -6.135352
```

Fig 6 Quality enhancement for FCC data set.

- In DLR we check for null values of location in the data set and obtain the indexes where null values are present.
- The latitude and longitude values are obtained using reverse geocoding.
- Further we observe that still there are null values for location in the data set, so we use geocoder to obtain the latitude and longitude values were missing.
- After this step for one instance where the latitude and longitudes are not available, we use geocoder to find them.
- Since the final dataset requires location, latitude and longitude we disregard clearing the null values for size and number columns
- Finally, we get the resulting DLR data set without any missing or null values.

```
          Location Number Size   Latitude  Longitude
0  Kilbogget Park      1  Snr  53.257242  -6.140665
1      Cabinteely      2  SSG  53.257614  -6.139882
2          Dublin      3  SSG  53.257842  -6.139265
3      Ballybrack      4  SSG  53.257098  -6.139094
4      Cabinteely      5  SGG  53.256674  -6.140134
```

Fig 7 Quality enhancement for DLR data set.

Data Modelling

- DCC location contains the combination of park and area.
- FCC location contains the combination of facility name and location.
- We create 3 data frames with names df 1, df 2, df 3
- df 1 contains location latitude and longitude values of DCC
- df 2 contains location latitude and longitude values of DLR

- df 3 contains location latitude and longitude values of FCC
- Now we concatenate the 3 data frames df 1, df 2, df 3 by ignoring the indexes.
- If the indexes are not ignored, then the individual data set indexes will be considered.

400	St. Catherines,Unnamed Road, Coldblow, Co. Dub...	53.3661	-6.4642
401	Seagrange Park,Dublin 13	53.3924	-6.13628
402	St. Catherines,St. Catherines Demesne, Lucan	53.3711	-6.4675
403	Hazelbury Park,Castaheany	53.4011	-6.4284
404	Porterstown Park,Porterstown	53.3701	-6.40426
405	Robswall,Malahide	53.4419	-6.13573
406	Balheary Reservoir,Swords	53.472	-6.22303

407 rows × 3 columns

Fig 8 Before Removing Duplicate Values

- In the next step we the duplicate function to remove all redundant values.

400	St. Catherines,Unnamed Road, Coldblow, Co. Dub...	53.3661	-6.4642
401	Seagrange Park,Dublin 13	53.3924	-6.13628
402	St. Catherines,St. Catherines Demesne, Lucan	53.3711	-6.4675
403	Hazelbury Park,Castaheany	53.4011	-6.4284
404	Porterstown Park,Porterstown	53.3701	-6.40426
405	Robswall,Malahide	53.4419	-6.13573
406	Balheary Reservoir,Swords	53.472	-6.22303

342 rows × 3 columns

- Finally, we save the this in a csv file which is the final enhanced data set.

	LOCATION	Latitude	Longitude
0	ALBERT COLLEGE,NORTH WEST	53.3855	-6.26063
1	ALBERT COLLEGE,NORTH WEST	53.3832	-6.2634
2	BEECHILL ,SOUTH EAST	53.3153	-6.23165
3	BELCAMP,NORTH CENTRAL	53.4016	-6.18566
4	BELCAMP,NORTH CENTRAL	53.1424	-7.69205

References

- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.drop_duplicates.html
- https://www.datacamp.com/community/tutorials/joining-dataframes-pandas
- https://geocoder.readthedocs.io/api.html#reverse-geocoding
- https://gis.stackexchange.com/questions/189312/python-trying-to-figure-out-geopy-in-python-for-reverse-look-ups
- https://stackoverflow.com/questions/45336763/using-my-google-geocoding-api-key-with-python-geocoder
- https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b
- https://stackoverflow.com/questions/49898661/xml-to-csv-python