Data Engineering: Pre-processing & Merging data from multiple sources

The identification of appropriate sources of data and its pre-processing is the integral part of any data analytics task. Data wrangling accounts for the most tedious and time-consuming activity in the analytics pipeline. Here, we try to collect data about playing pitches in area around Dublin, Ireland. The aim of this task is to create a dataset that exhibits following properties:

1. **Complete:** Should include all relevant fields from individual dataset
2. **Clean:** Free of discrepancies
3. **Unique:** No duplicate records
4. **Consistent:** Field values should be represented uniformly

The major tasks carried out in pre-processing this dataset are as follows:

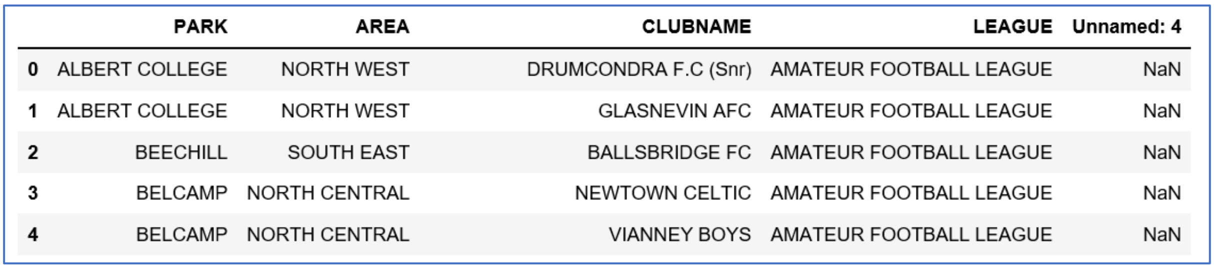
1. Exploratory analysis of data

Following sources are chosen for importing the datasets:
Data Source: [Open data from Irish government portal](https://data.gov.ie/)

No.	Dataset	Source
1.	Dublin City Council (DCC)	dccplayingpitchesp20120816-1550.csv
2.	Fingal County Council (FCC)	fccplayingpitchesp20111203-1424.xml
3.	Dún Laoghaire-Rathdown county council (DLR)	dlr-pitches.csv

1. Dublin City Council (DCC)

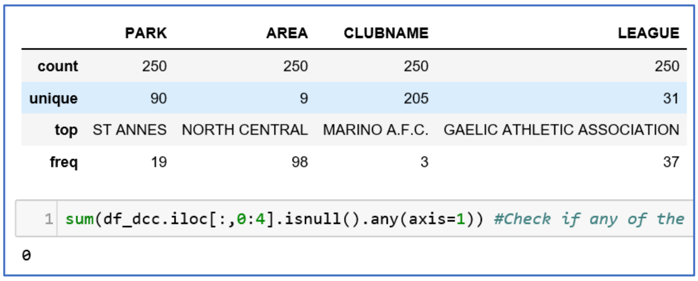
A. Observations:



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

Figure 1. DCC dataset

- Has 250 records, of which 90 are unique. This indicates most of the pitch names are repeated
- League name appears not to be of much relevance as we are concerned only about pitches in the region
- Some parks are associated with quite a few clubs and so have redundant observations



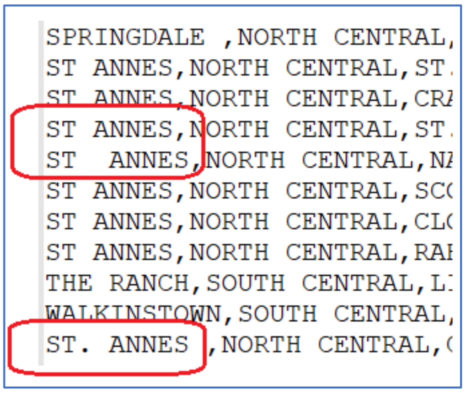
	PARK	AREA	CLUBNAME	LEAGUE
count	250	250	250	250
unique	90	9	205	31
top	ST ANNES	NORTH CENTRAL	MARINO A.F.C.	GAELIC ATHLETIC ASSOCIATION
freq	19	98	3	37

```
sum(df_dcc.iloc[:,0:4].isnull().any(axis=1)) #Check if any of the
```

```
0
```

Figure 2. DCC Dataset details

B. Issues with DCC dataset:



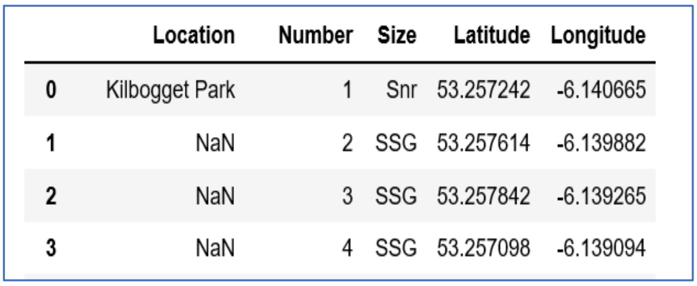
```
SPRINGDALE ,NORTH CENTRAL,
ST ANNES,NORTH CENTRAL,ST.
ST ANNES,NORTH CENTRAL,CRA
ST ANNES,NORTH CENTRAL,ST.
ST  ANNES,NORTH CENTRAL,NA
ST ANNES,NORTH CENTRAL,SCO
ST ANNES,NORTH CENTRAL,CLO
ST ANNES,NORTH CENTRAL,RAI
THE RANCH,SOUTH CENTRAL,LI
WALKINSTOWN,SOUTH CENTRAL,
ST. ANNES ,NORTH CENTRAL,(
```

Figure 3. Inconsistent names

1. Column Headers and park, club, league names in upper case
2. Discrepancies in Park names (spaces, periods, eg. "St. Annes", " St Annes "):
3. Extra column of null values due to discrepancy in CSV file. Need to import only required columns

2. Dún Laoghaire-Rathdown county council (DLR dataset)

A. Observations:

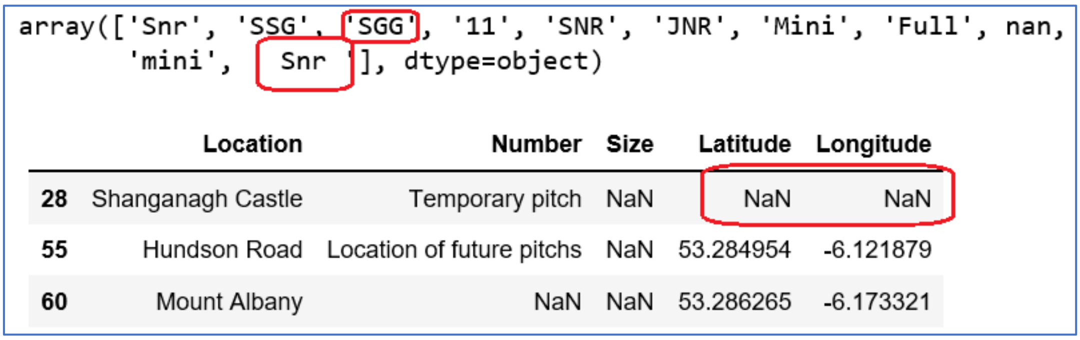


	Location	Number	Size	Latitude	Longitude
0	Kilbogget Park	1	Snr	53.257242	-6.140665
1	NaN	2	SSG	53.257614	-6.139882
2	NaN	3	SSG	53.257842	-6.139265
3	NaN	4	SSG	53.257098	-6.139094

Figure 4. DLR dataset

1.Has multiple redundant rows
2.Observations are at different granularity than earlier dataset. Individual pitches within the parks are recorded with different geographical coordinates

B. *Issues with DLR dataset:*



```
array(['Snr', 'SSG', 'SGG', '11', 'SNR', 'JNR', 'Mini', 'Full', nan,
       'mini', ' Snr '], dtype=object)
```

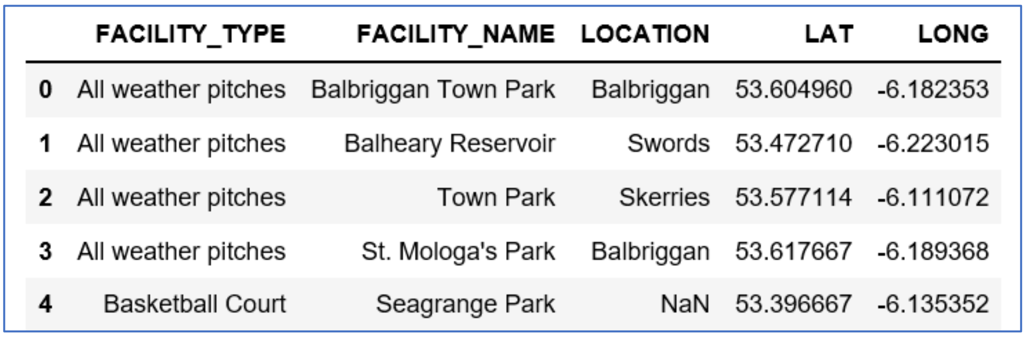
	Location	Number	Size	Latitude	Longitude
28	Shanganagh Castle	Temporary pitch	NaN	NaN	NaN
55	Hundson Road	Location of future pitchs	NaN	53.284954	-6.121879
60	Mount Albany	NaN	NaN	53.286265	-6.173321

Figure 5. Some of the issue with DLR dataset

1. Missing Location (Park) names:
2. Inconsistent pitch size values: some pitch sizes are abbreviated, some are not
3. Missing Latitude/ Longitude for one of the observations (Shanganagh Castle)
4. 'SGG' is an incorrectly recorded pitch type for a single observation. Should be changed to 'Single Sided Games'

3. Fingal county council (FCC dataset)

A. Observations:



	FACILITY_TYPE	FACILITY_NAME	LOCATION	LAT	LONG
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

Figure 6. DCC dataset

1. Includes facility type as different field from other dataset
2. Has location information as well
3. Already has latitude longitude information unlike DCC dataset

B. Issues with FCC dataset:

1. Missing Location names
2. All upper-case headers

2. Data Modelling

The modelling objective is to create a clean and complete dataset of playing pitches around Dublin. All the datasets have different set of features and we need to include only the relevant ones. In alignment to our objective, I identify the need to include the following features in the resultant dataset:

1. 'Park': Park or facility name, where the pitches are located
2. 'Location': Brief address of the park
3. 'Latitude': X location co-ordinate
4. 'Longitude': Y location co-ordinate
5. 'County Council': one of the 3 county councils in consideration
6. 'Additional Information': Miscellaneous information, that is unique in each dataset. eg. Pitch Number, Size, Clubs, type of pitch

Resultant dataset would have the following schema:

Park	**Location**	**Latitude**	**Longitude**	**County Council**	**Additional Information**

Tidy schema:

However, depending upon the analytical task at hand, a following schema looks much tidier but may have many more null values:

Park	**Location**	**Latitude**	**Longitude**	**County Council**	**Pitch Number**	**Size**	**Facility**	**Clubs**

3. Data Quality Enhancement:

To achieve the above resultant data model, each of the 3 datasets must be wrangled individually.

Enhancing quality with geospatial data.:

Some of the datasets are missing the geographical coordinates, while others are missing the addresses. For datasets to be complete, this missing but critical information must be procured.

The process of geocoding and reverse geocoding has been employed to enhance the data quality of all three datasets. A combination of two APIs is used to fetch the geospatial data:

1. *Nominatim with geopy:*
   - Used to search Open Street Maps data
   - OpenStreetMap® is open data, licensed under the Open Data Commons Open Database License (ODbL) by the OpenStreetMap Foundation (OSMF).
   - Fairly accurate and updated cartography data
   - users are free to copy, distribute, transmit and adapt the data (unlike some other proprietary services), if you credit OpenStreetMap and its contributors
2. *ArcGIS for Developers API with ArcGIS python package:*
   - provided by ESRI (Environmental Systems Research Institute) which is an international supplier of geographic information system (GIS) software, web GIS and geodatabase management applications
   - High degree of accuracy over smaller area of interest

##ArcGIS and Nominatim both have been availed following their terms and conditions of use.

1. DCC Dataset:

(A) Data Cleaning:

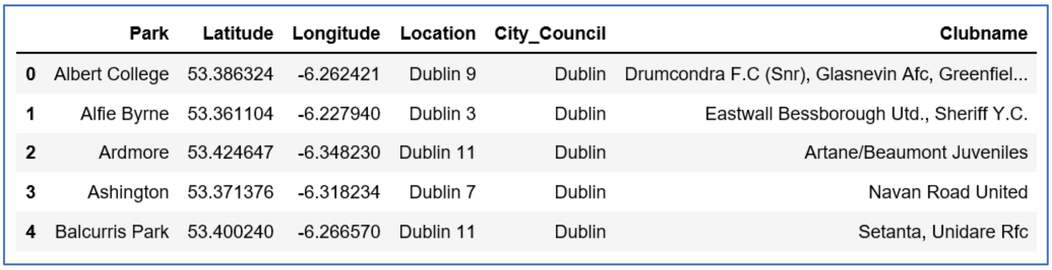
1. Update all headers and column values to title case
2. Strip spaces trailing and following the park names. Replace extra spaces in between the names with single space. Remove periods in names like Fr., St.

(B) Data Enhancement:

1. Retrieve park location, **Latitude and Longitude co-ordinates** using Park names and geocoding
2. **Group** records by park X, Y location co-ordinates and aggregate club names in a single comma separated value, so that duplicate par records are excluded. A park associated with multiple clubs had multiple redundant entries in the dataset. Now, all clubs associated with single park would be combined together

Note: Data cleansing and data enhancement is elegantly achieved by constructing a pipeline of those operations using pandas method chaining

Resulting dataset:



	Park	Latitude	Longitude	Location	City_Council	Clubname
0	Albert College	53.386324	-6.262421	Dublin 9	Dublin	Drumcondra F.C (Snr), Glasnevin Afc, Greenfiel...
1	Alfie Byrne	53.361104	-6.227940	Dublin 3	Dublin	Eastwall Bessborough Utd., Sheriff Y.C.
2	Ardmore	53.424647	-6.348230	Dublin 11	Dublin	Artane/Beaumont Juveniles
3	Ashington	53.371376	-6.318234	Dublin 7	Dublin	Navan Road United
4	Balcurris Park	53.400240	-6.266570	Dublin 11	Dublin	Setanta, Unidare Rfc

2. FCC Dataset:

(A) Imported data from XML source:

- Fingal county council pitch data is imported from XML source using xml package

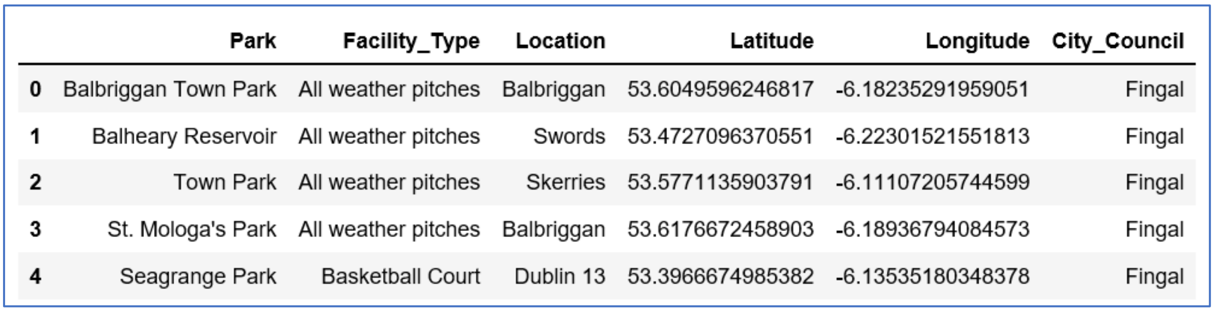
(B) Data Cleaning:

1. Update all headers to title case
2. Modify column headers 'LAT' to 'Latitude' and 'LONG' to Longitude
3. Locations in some of the observations are missing. The locations would be fetched from latitude/ longitude coordinates using reverse geocoding.

(C) Data Enhancement:

1. A column identifying the county council that the pitches belong to, is added.

Resulting FCC dataset:



	Park	Facility_Type	Location	Latitude	Longitude	City_Council
0	Balbriggan Town Park	All weather pitches	Balbriggan	53.6049596246817	-6.18235291959051	Fingal
1	Balheary Reservoir	All weather pitches	Swords	53.4727096370551	-6.22301521551813	Fingal
2	Town Park	All weather pitches	Skerries	53.5771135903791	-6.11107205744599	Fingal
3	St. Mologa's Park	All weather pitches	Balbriggan	53.6176672458903	-6.18936794084573	Fingal
4	Seagrange Park	Basketball Court	Dublin 13	53.3966674985382	-6.13535180348378	Fingal

Figure 7. Enhanced FCC dataset

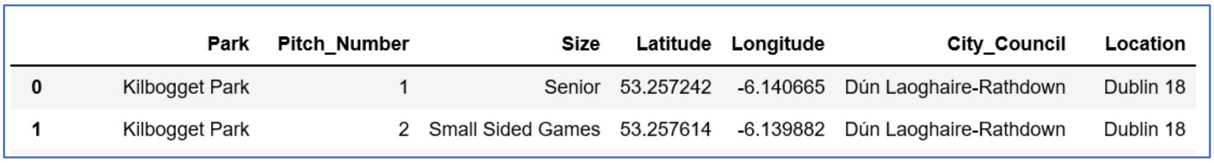
3. DLR Dataset:

(A) Data Cleaning:

1. Rename 'Location' column to 'Park' in order to be consistent with proposed data model
2. Fill missing park names with previous values
3. Replace pitch size abbreviations with actual sizes. eg. SNR with 'Senior', 11 with '11-a-side'
4. Fetch missing Latitude/ Longitude for one of the parks (Shanganagh Castle)
5. 'SGG' is an incorrectly recorded pitch type for a single observation. It is changed to 'Single Sided Games'

(B) Data Enhancement:

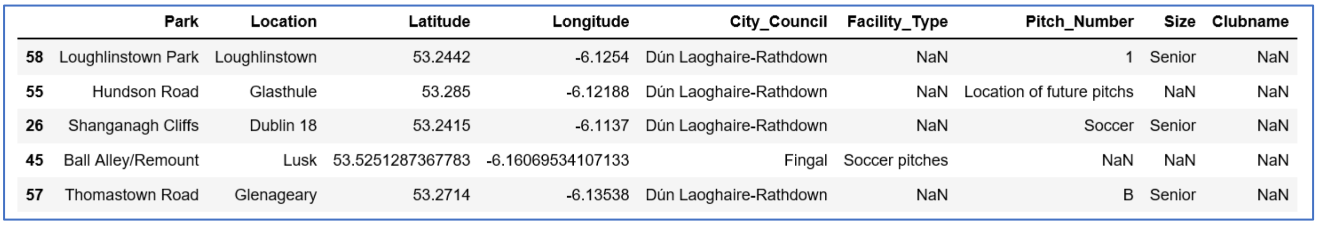
1. A column identifying the *county council* that the pitches belong to, is to be added.
2. A column for park *location* is to be added which would be helpful for completeness. The locations would be fetched from latitude/ longitude coordinates using reverse geocoding.



	Park	Pitch_Number	Size	Latitude	Longitude	City_Council	Location
0	Kilbogget Park	1	Senior	53.257242	-6.140665	Dún Laoghaire-Rathdown	Dublin 18
1	Kilbogget Park	2	Small Sided Games	53.257614	-6.139882	Dún Laoghaire-Rathdown	Dublin 18

4. Data Aggregation:

Next, I appended all datasets to form a single dataset as below:



	Park	Location	Latitude	Longitude	City_Council	Facility_Type	Pitch_Number	Size	Clubname
58	Loughlinstown Park	Loughlinstown	53.2442	-6.1254	Dún Laoghaire-Rathdown	NaN	1	Senior	NaN
55	Hundson Road	Glasthule	53.285	-6.12188	Dún Laoghaire-Rathdown	NaN	Location of future pitchs	NaN	NaN
26	Shanganagh Cliffs	Dublin 18	53.2415	-6.1137	Dún Laoghaire-Rathdown	NaN	Soccer	Senior	NaN
45	Ball Alley/Remount	Lusk	53.5251287367783	-6.16069534107133	Fingal	Soccer pitches	NaN	NaN	NaN
57	Thomastown Road	Glenageary	53.2714	-6.13538	Dún Laoghaire-Rathdown	NaN	B	Senior	NaN

Figure 8. Incomplete but tidy dataset

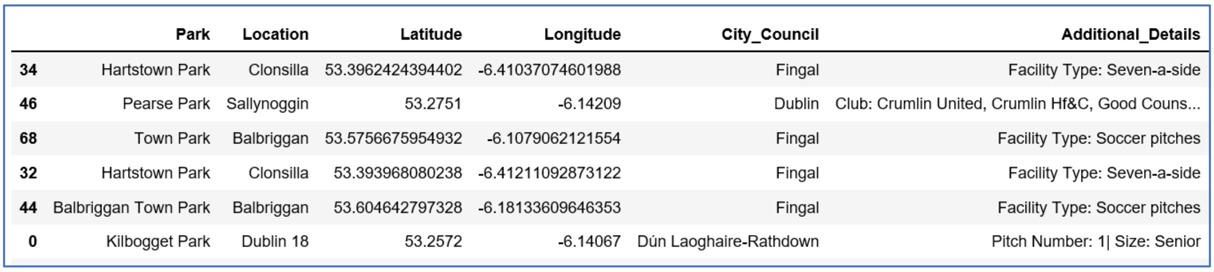
The combined dataset obtained above, exhibits the following properties:

1. Accurate: As accurate as the underlying open data sources are
2. Unique: No duplicate records
3. Consistent: All variables are consistently represented

However, the dataset is not *complete*. In the sense, it misses values that are characteristic to only individual datasets, like clubnames in Dublin city council, facility types in DLR dataset or pitch number level granularity in Fingal dataset. But, I have included all these fields as they are peculiar of typical playing pitches.
Even as there appear to be a lot of missing values in columns like club names, facility type, etc. the dataset is **tidy**. Being a tidy set, it is easier and faster to carry out any further analysis using this structure. As only a single variable is described by a single column and every row has unique observation.

- ➢ Loosing *tidiness* but achieving *completeness*:

- We can choose to compromise on tidiness of data and still not lose out on any information by combining non-common column into a single column such as 'Additional Information'



	Park	Location	Latitude	Longitude	City_Council	Additional_Details
34	Hartstown Park	Clonsilla	53.3962424394402	-6.41037074601988	Fingal	Facility Type: Seven-a-side
46	Pearse Park	Sallynoggin	53.2751	-6.14209	Dublin	Club: Crumlin United, Crumlin Hf&C, Good Couns...
68	Town Park	Balbriggan	53.5756675954932	-6.1079062121554	Fingal	Facility Type: Soccer pitches
32	Hartstown Park	Clonsilla	53.393968080238	-6.41211092873122	Fingal	Facility Type: Seven-a-side
44	Balbriggan Town Park	Balbriggan	53.604642797328	-6.18133609646353	Fingal	Facility Type: Soccer pitches
0	Kilbogget Park	Dublin 18	53.2572	-6.14067	Dún Laoghaire-Rathdown	Pitch Number: 1| Size: Senior

Figure 9. A complete but untidy dataset

Conclusion:

The above dataset is complete and includes a column of additional details which has peculiar attributes from all 3 datasets. Pitch number and size have been separated by '|'. Analysis over this dataset would be slightly more difficult than the tidy dataset but this achieves a trade-off for the completeness and tidiness.
Either of the two datasets can be used, depending on the analytical task at the hand!