# Data Understanding

This section provides a concise overview of the location data provided from the ‘Postcode’ data and the data provided by Foursquare.com Website for analysis. We have looked at it under the following section providing a holistic view of the data provided.

## Initial data collection and Data description

### Post Code details

The first data source link (*https://www.doogal.co.uk/PostcodeDistrictsCSV.ashx*) consists of Post Code related data and has the following features:

1. ***Postcode:*** *This consist of the outer part of the postcode*
2. ***Latitude:*** *Latitude coordinates of the location*
3. ***Longitude:*** *Longitude coordinates of the location*
4. ***Town/Area:*** *The name of the location or areas*
5. ***Region :*** *This is the location in which the Town or area is within*
6. ***Population:*** *The estimated number of people within that location*
7. ***Nearby districts:*** *Other**postcodes locations within the proximity to this postcode.*

*All the above fields were chosen to provide a framework to enable the analysis and address the business problem. The seven fields were used, and ‘Population’ field was also included to understand the number of individuals within the locality that may patronise and have an impact on the business. Other data fields that were excluded for this exercise are Easting, Northing, Grid Reference,*  *Postcodes , Active postcodes and Households for this project .There are about 3,096 records, but the number of records used in the project is dependent upon which cities are chosen.*

### FourSquare.com

The following fields can be extracted from the ***FourSquare.com*** site to enable us extract and use the fields for data manipulation and analysis, with other data that would assist in project.

* ***venue name*** *:**The name of the location with the neighborhood.*
* ***location – (lat, lng)*** *: The geographical location of the venue.*
* ***Distance*** *:**The distance from the ‘venue’ to the point of reference.*
* *Postal Code : The postal code of the ‘venue name’.*
* ***Neighborhood*** *: the description of the location of the ‘venue name’.*
* *City : The city within which the ‘venue name’.*
* *State : The state within which the ‘venue name’.*
* *Country : The country within which the ‘venue name’.*
* ***formattedAddress*** *:**The full address of the ‘venue name’ including post code.*
* *name : A categorical description of the ‘venue location’.*
* *pluralName : A categorical description of the ‘venue location’ in plural.*
* *shortName : A categorical description of the ‘venue location’.*
* *icon : Link to a graphic icon.*
* *photos : Link to a graphic photo.*
* *referralId : Reference.*
* *items : A description of the ‘venue location’.*
* *Type : A high-level description of the ‘venue location’ eg like ‘general’.*

The items in bold indicate that it was used directly within this project.

## Data Exploration

The data from the different sources were explored to understand the respective features and other things related to the data.

Below is a snap-shot of the Postcode sample data:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Postcode** | **Town/Area** | **Region** | **Latitude** | **Longitude** | **Nearby districts** | **Population** |
| EC1A | St Bartholomew s Hospital | City of London | 51.5203 | -0.104446 | EC1M, EC3P, EC1N, EC2Y, EC4A, EC1R, EC4P, EC4M... | 878.0 |
| EC2M | Broadgate, Liverpool Street | City of London | 51.5184 | -0.086829 | EC2N, EC3V, EC2A, EC2R, EC3A, EC3M, EC3R, EC3B... | 99.0 |
| EC2N | Old Broad Street, Tower 42 | City of London | 51.5160 | -0.087486 | EC2M, EC3V, EC2R, EC3A, EC3M, EC2V, EC3R, EC3B... | 5.0 |

A snapshot of the regions in the

**Region Counts**

Highland 85

Liverpool 52

Argyllshire 46

Cornwall 46

Birmingham 42

..

Newcastle upon Tyne 1

Hertfordshire 1

Rothes 1

Torridge 1

Ramsgate 1

Below is an extract of data for a location in Manchester from the FourSquare.com website and is used within the project:



For example, in Manchester we have 100 venues were returned by Foursquare.com, with 11 Hotels for a specific location (53.48, -2.24) that was searched.

The fields associated with geographical location of the cities (‘Latitude’ and ‘Longitude’) was the primary means used to pull the data sources together. Data from FourSquare.com data was extracted from a ‘*json’* format into a python dataset for data manipulation and analysis.

So, the location is used to extract the data for the two independent data sources and then used for analysis, reporting and the various conclusion that will be reached.

## Integrity

The data from Foursquare.com is coming from a reliable source and has been vetted, and the data relating to the postcodes has been reviewed and tidied up where its been deemed not suitable for processing.

## Completeness

For the purpose of this project, the data was complete.

## Consistency

For the purpose of this project, the data was consistent

## Accuracy

The data was reviewed and after the removal the ‘*N/A’* values from the dataset, the data was deemed accurate.