P3: Wrangle OpenStreetMap Data Medellin Colombia

I chose to work on Medellin, Colombia because it is my hometown and I was curious how accurate the information currently is and how it has changed overtime.

Challenges Encountered :

- Abbreviation of street names: Some streets were abbreviated and in order to be consistent, these were changed as follow:
    - "calle": "Calle"
    - "CL":'Calle'
    - "Cl":'Calle'
    - "CALLE":'Calle'
    - "CARRERA":'Carrera'
    - "carrera":"Carrera"
- Some entries also had the full address under street name. I changed these by removing the information provided starting with a '#' sign, which indicated the housenumber. Ex:

    <tag k="addr:street" v="Calle 30 #46-30"

- The city and country were also abbreviated in some entries and these were changed to Medellin and Colombia, respectively.
- Some street names had extra information (usually cross street) followed by a semicolon. The extra information was removed. Example:

    <tag k="name" v="Calle 53;Maracaibo"

- Other problems encountered were street names with other information other than the street name. Some of these were updated in the database instead of programmatically since they each had a unique mistake. For example:
    - 54 B 06 Interior 201 --> Carrera 5 este
    - N 55 – 240 --> Carrera 10

During the audit I noticed that cities surrounding Medellin were part of the data. If we want to consider these part of the map then the name and description of the data should be Medellin Metropolitan Area. However, if this is the case then a lot more places should be added to provide concise information of the map.

```
pd.read_sql_query('''select distinct value, count(*)
                     from (select key, value from nodes_tags
                     union all
                     select key, value from ways_tags)
                     where key='city' and value not like 'Med%'
                     group by value;''', con)
```

	value	count(*)
0	Angelopolis	1
1	Bello	3
2	Comuna 8	2
3	Copacabana	4
4	El Carmen De Vibora	1
5	El Carmen De Viboral	3
6	El Carmen de Viboral	5
7	El Poblado, Medellín	1
8	El Retiro	1
9	El poblado	1
10	Envigado	28
11	Girardota	5
12	Itagüi	2
13	Itagüí	5
14	La Ceja	14
15	La Ceja del Tambo	12
16	Marinilla	2
17	Rionegro	13
18	Rionegro, Antioquia	1
19	Sabaneta	4
20	Sabaneta Antioquia	1
21	Sabaneta, Antioquia	1
22	Vereda Pontezuela Rionegro Antioquia Colombia	1
23	el carmen de Viboral	1
24	itagüí	1

As we can see even the cities that are outside of Medellin have different format and we would need to clean up that information as well.

Analysis:

Size of the files:

- Size of the entire osm file is: 67018416
- Size of sample osm file is: 13553023
- Number of unique users: 645
- Number of nodes: 320837
- Number of ways: 35147

Top 10 users:

```
pd.read_sql_query('''
                    select user, count(*) as "Total Number of Contributions"
                    from (select user from nodes
                    union all
                    select user from ways)
                    group by user
                    order by count(*) desc
                    limit 10;''',con)
```

	user	Total Number of Contributions
0	carciofo	111145
1	harrierco	25521
2	Argos	24738
3	JosClag	13849
4	Kleper	12160
5	cris_1994	11243
6	humano	9690
7	Antares_alf	9401
8	mono11	7952
9	harriercoold	7405

Number of schools:

```
pd.read_sql_query('''select count(value) as 'Total Number of Schools'
                     from (select key, value from nodes_tags
                     union all
                     select key, value from ways_tags)
                     where key='amenity' and value='school';''', con)
```

	Total Number of Schools
0	200

Number of Hospitals:

```
pd.read_sql_query('''select count(value) as 'Total Number of Hospitals'
                    from (select key, value from nodes_tags
                    union all
                    select key, value from ways_tags)
                    where key='amenity' and value='hospital';''', con)
```

	Total Number of Hospitals
0	85

Opportunities:

- The information should be standardized and give the user rules on format to follow when entering/editing information. For example when entering street name, only the street name should be allowed without giving extra information such as cross street or housenumber.
- In Colombia we use the neighborhood names a lot to narrow down location, it would be good to have an entry for neighborhood name.
- I see that the top user is carciofo who appears to have contributed 111145. One opportunity to explore is how to get people more involved to contribute on improving the data., perhaps through a game or an app that could be easily accessible and user friendly.