Udacity_OpenStreetMap

Project

This is a project where I needed to choose an area of the world in the
    OpenStreetMap and use data munging techiques, such as assessing the
    quality of the data for validity, accuracy, completeness, consistency
    and uniformity, to clean the OpenStreetMap data for a part of the world
    that you care about. Choose to learn SQL or MongoDB and apply your
    chosen schema to the project.

Code Organization:

There are two folders one for the code I used to analyze my area and other
    for the Use Case example.

Map Area

São Paulo, SP, Brazil

- [https://www.openstreetmap.org/relation/298285#map=10/-23.6824/-46.5957]
    (link for Sao Paulo Map on OpenStreetMap)
- [https://mapzen.com/data/metro-extracts/metro/sao-paulo_brazil/](link for
    Sao Paulo OSM)

This area is my hometown, so I'm more interesetd to see what database
    querying reveals and I'd like to contribute to it's improment o
    OpenStreetMap.org.

The small area I whose their coordinates are: min lat: -23.6246 , max lat:
    -23.6162 , min lng: -46.6421 max lng: -46.6193.

Problems Encountered in the map:

   ## Challenges Encountered:
       - Street Names:
           - I found that some street names they were written in caps
             letters and other wasn't. So I tried to uniformalize those.
             Example: "RUA","rUa" when the correct and most commom used
             is "Rua".

           - I also found that some street names there wasn't a type on
             them. Example: "Angelo Meneguesso", "Oito". So we cannot
             know if it a street, avenue or a road.

           Example of Problems:
               - Misstyping street names: "RUA","rUa" instead of "Rua"
               - Missing type of street: "Oito" (so it would be "Rua Oito",
                 "Avenida Oito". I don't know.).

       - PostCode:
           - Post Codes from São Paulo City is 5 numbers and than be follow
             by a hiffen and more three numbers and must have: less than
             0600 or in between 0800 and 0850.
           - I found that some places were from anothers cities like:
             "Santos", "Santana de Parnaíba". "São Caetano"
           - I also found that some postcodes were incomplete like: from

            "Rua Bresser": "1194"

        Example of Problems:
            - Others cities in dataset: Mauá,Jundiai, Cotia, Santos.
            - Some postcodes are incomplete.

        Query:

```sql
Select nt.value,COUNT(nt.value) from nodes_tags nt, (Select
    id,value from nodes_tags where key='postcode' and ((value >
    '0600' and value < '0800') or (value > '0850'))) ps where
    ps.id = nt.id and nt.key = 'city' group by nt.value Having
    COUNT(nt.value)> 1 order by COUNT(nt.value) DESC;
```

        nt.value   COUNT(nt.value)
        ---------------------- ---------------
        São Bernardo do Campo   117
        Guarulhos   81
        Mairiporã   66
        Mauá   58
        Santo André   37
        Osasco   36
        Santos   33
        Suzano   30
        São Caetano do Sul   28
        Diadema   11
        São José dos Campos   10
        Barueri   7
        Arujá   6
        Cotia   6
        São Roque   6
        Franco da Rocha   5
        Jundiaí   5
        Taboão da Serra   5
        Várzea Paulista   5
        Itupeva   4
        São Paulo   4
        Ferraz de Vasconcelos   3
        Guarujá   3
        ITARIRI   3
        Jundiai   3
        Araçariguama   2
        Carapicuíba   2
        Itapevi   2
        Salto   2
        Santo Andre   2

        I also noticed that, were 4 points in this query there which the
            city is allocated to 'Sao Paulo'.
        After a investigation n those points (doing some more queries
            and using Google Maps), I noticed that some points actually
            are located in other cities and others in the postcode are
            incomplete.

    ### Problems are Cleaned Programmaticaly:

    - I solved the uniformatly probem for street names and removing from

dataset those with no mapping correspondent and those with wrong
Postcodes.

Overview of the data:

OSM XML Size:

```
sao-paulo_brazil.osm ......... 907 MB
database_OSM.db.........658.3 MB
nodes.csv ............. 346.6 MB
nodes_tags.csv ........ 9.5 MB
ways.csv .............. 34.5 MB
ways_tags.csv ......... 51.1 MB
ways_nodes.cv ........ 132 MB
```

Oversize statistics:

- number of unique users:
```sql
sqlite> SELECT COUNT(DISTINCT(e.uid))
        FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways)
            e;
```

There are 2178 users.

- number of nodes and ways:

    - Number of Nodes:

        ```sql
        sqlite> SELECT COUNT(*) FROM nodes;
        ```

        4003764 Nodes

    - Number of ways:
        '''sql
        sqlite> SELECT COUNT(*) FROM ways;
        '''
        553876 ways

- number of chosen type of nodes, like cafes, shops etc.

Top 10 contributing users
```sql
sqlite> SELECT SUM(num)
FROM(
    SELECT e.user, COUNT(*) as num
    FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
    GROUP BY e.user
    ORDER BY num DESC
    LIMIT 10);
```

```
Bonix-Mapper,2345049
AjBelnuovo,262639
cxs,191624
MCPicoli,106375
"O Fim",105778
johnmogi,95857
```

```
        ygorre,91096
        patodiez,85006
        naoliv,84366
        "Roberto Costa",65525
```

Responsable for 3433315 points, which is 75% of data.

Number of user appering only once

```sql
sqlite> SELECT COUNT(*) FROM (SELECT e.user, COUNT(*) as num
...> FROM (Select user from nodes union all select user from ways) e
...> Group by e.user
...> Having num = 1)
```

There are 476 users with only 1 post.

```sql
SELECT COUNT(*) FROM (SELECT e.user, COUNT(*) as num
FROM (Select user from nodes union all select user from ways) e
Group by e.user
Having num < 5) ;
```

There are 847 users with less than 5 posts.

Contributor statistics and gamification suggestion

The contributions of users seems incredibly skewed, possibly due to
    automated versus manual map editing (the word "bot" appears in some
    usernames). Here are some user percentage statistics:

Top user contribution percentage ("Bonix-Mapper") 51.45%
Combined top 2 users' contribution ("Bonix-Mapper" and "AjBelnuovo")
    57.21%
Combined Top 10 users contribution 75.3%
There were 0.38% (847) of users that did less than 5 posts.

Additional Data Exploration

Top 10 appearing amenities

```sql
sqlite> SELECT value, COUNT(*) as num
FROM nodes_tags
WHERE key='amenity'
GROUP BY value
ORDER BY num DESC
LIMIT 10;
```

restaurant|1132
bank|685
fuel|615
fast_food|477
school|473
parking|427
pharmacy|410

```
pub|379
place_of_worship|359
bicycle_rental|274
```

Ideas to Improve the dataset:

- As showed above the dataset contained some points from others cities
  (like Santos) it will be nice it had some way to verify the new
  points when they are created. Google Maps could be used in this
  process. But this may also cause a legal issue or something like
  that since it is using data from another source

- Due to the low number of users contributions to Open Street Map I
  think that another idea would be if it had an app with gamifications
  to incentivate differents users to contribute to the system. Alhough
  it can improve the participation of new users the developer that
  creates the app may have collected the data and use to another task.

References:
- [https://www.openstreetmap.org/relation/298285#map=10/-23.6824/-46.5957]
  (link for Sao Paulo Map on OpenStreetMap)
- [https://mapzen.com/data/metro-extracts/metro/sao-paulo_brazil/](link for
  Sao Paulo OSM)
- [https://thiagorodrigo.com.br/artigo/cep-sao-paulo-lista-de-cep-por-
  bairro-e-cidade-da-grande-sao-paulo/](Sao Paulo PostCodes)