Project: OpenStreetMap Data Wrangling with SQL

Name: Niraj Patel

Map area: Jersey City

OpenStreetMap URL: https://www.openstreetmap.org/relation/170953

The reason I have chosen this area is that currently in this area and love it. I have extracted OSM file for the surrounding Jersey City area.

Unique Tags

{'bounds': 1,

 'member': 63356,

 'meta': 1,

 'nd': 441443,

 'node': 342901,

 'note': 1,

 'osm': 1,

 'relation': 595,

 'tag': 153597,

 'way': 57672}

Patterns in the Tags

{'lower': 101578, 'lower_colon': 45920, 'other': 5413, 'problemchars': 686}

Problems in the Map Data

Through my exploratory analysis, there were inconsistencies in the following areas:

1. Street name
2. Zip code

In street names, there were so much inconsistencies as in few cases abbreviations were used and In some cases, full names were given. In case of zip codes, there were also some discrepancies such as had the extra four-digit number or even in one instance state name was written before the zip code.

'Adams': {'Adams St'},

'Amity': {'Amity Street'},

'Baldwin': {'Baldwin Avenue'},

'Barrow': {'Barrow Street'},

Data Overview

- Map.osm   80.1 MB
- map.db   76 MB
- nodes.csv   39.1 MB
- nodes_tags.csv   1.156 MB
- ways.csv   4.7 MB
- ways_nodes.csv   12.08 MB
- ways_tags.csv   5.9 MB
- sample.osm   1.5 MB

Number of nodes:

sqlite> SELECT COUNT(*) FROM nodes;

685802

Number of ways:

sqlite> SELECT COUNT(*) FROM ways;

115344

Number of unique user:

sqlite> SELECT COUNT(DISTINCT(a.uid)) FROM (SELECT DISTINCT(uid) FROM ways UNION ALL SELECT DISTINCT(uid) FROM nodes) as a;

711

Top 10 contributing users

sqlite> 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;

b'minewman'|52181

b'Valustaides'|13730

b'robgeb'|9964

b'peace2'|7750

b'OceanVortex'|7166

b'ingalls'|7146

b'3yoda'|6286

b'choess'|5900

Number of Starbucks Coffee Shops:

sqlite> SELECT COUNT(*) FROM nodes_tags WHERE value LIKE '%Starbucks%';

18

Restaurants

sqlite> SELECT COUNT(*) FROM nodes_tags WHERE value LIKE '%Restaurant%';

189

Additional Ideas

From my analysis, there is a lot of data for Jersey City that is missing. There are several more locations of restaurants, cafes, and shops which are not included in the OSM.

The problem can be solved in two ways:

1. Encouraging the local people to post their business and surroundings on the OSM. They should also add the data for tourist spots or locations. Currently, it does not have any data to address this.

    Sqlite>SELECT COUNT(*) FROM nodes_tags WHERE value LIKE '%tourist spots %';

    0

    When I ran this query, there were no results. So by adding more local data, the incoming tourists can be benefited from this.

2. Tourists should also be encouraged to enter the data as they explore the city.

3. The data should be checked voluntarily and periodically to ensure that the data entered is accurate and complete.

Benefits and Anticipated Problems in Implementing the Improvement

By implementing this, we can say that if more complete and accurate data will be published on the OpenStreetMap then more people will start using it.

If more people start using it, the contribution and publishing the data will also increase. Through this, we can ensure that the learning effect is properly utilized to improve the OpenStreetMap.

The Problems can be that not many people are aware of OpenStreetMap. Common people mostly use services like Google or Apple maps which are highly accurate and popular. This lack of awareness leads to less contribution because of which it suffers the most.