OpenStreetMap Data Case Study by Burcu Kurtaran

Map Area

İstanbul, Turkey https://mapzen.com/data/metro-extracts/metro/istanbul_turkey/

İstanbul is the biggest city in Turkey. Also it is my hometown and where I am live. Because of these reasons, I prefer to explore İstanbul.

Problems Encountered in the Map

The data in Open Street Map project is built with people. It causes some problems. Some users prefer to use abbreviations, other ones prefer to use long names. I made some changes in data to provide consistency.

- Street address inconsistencies

```
"istanbul" : "Istanbul"
"İst" : "Istanbul"
"Ist" : "Istanbul"
"ist" : "Istanbul"
```

- Abbreviations

```
"sk." : "Street"
"Sk." : "Street"
"Sk" : "Street"
"sk" : "Street"
"Sok." : "Street"
"Cd" : "Avenue"
"Cd,": "Avenue"
"cd" : "Avenue"
"Cd." : "Avenue"
"cd." : "Avenue"
"Cad." : "Avenue"
"Cad" : "Avenue"
"Blv." : "Boulevard"
"Bulv.": "Boulevard"
"Mh" : "Neighborhood"
"mh": "Neighborhood"
"Mah.": "Neighborhood"
"Mh.," : "Neighborhood"
"Mah," : "Neighborhood"
```

- Misspelling

```
"Şirinyer" : "Sirinyer"
```

- Turkish names

```
"Sokak" : "Street"
"sokak" : "Street"
"Sokağı" : "Street"
"Sokak," : "Street"
"Caddesi" : "Avenue"
"caddesi" : "Avenue"
"Havalimanı" : "Airport"
"Havalımanı" : "Airport"
"havalimanı" : "Airport"
"Liman" : "Port"
"liman": "Port"
"Bulvar" : "Boulevard"
"Bulvari" : "Boulevard"
"bulvarı": "Boulevard"
"Bulvarı": "Boulevard"
"mahallesi": "Neighborhood"
"Mahallesi": "Neighborhood"
"Yerleşkesi" : "Campus"
"İzkent" : "Izkent"
"İzkent," : "Izkent"
"Meydanı" : "Square"
"Meydan" : "Square"
"Şirinkapı" : "Sirinkapi"
"İstikbal" : "Istikbal"
"Gaziosmanpaşa" : "Gaziosmanpasa"
"İstiklal" : "Istiklal"
"sahil" : "Coast"
"sahili" : "Coast"
"sahil" : "Coast"
"İskele" : "Port"
"İskelesi" : "Port"
"Alışveriş Merkezi": "Shopping Center"
"Paşa" : "Pasa"
"Şehitleri" : "Sehitleri"
"Çevre Yolu" : "Highway"
"Üniversite" : "University"
```

Above is the old name corrected with the better name. Using clean_data.py, the names in data are updated.

File sizes:

```
osm/istanbul_turkey.osm :   261 MB
nodes_csv               :   99 MB
nodes_tags.csv          :   3 MB
ways_csv                :   12 MB
ways_nodes.csv          :   38 MB
ways_tags.csv           :   11 MB
OpenStreetMap.db        :   143 MB
```

Number of nodes

```
SELECT COUNT(*) FROM nodes;
```

```
Number of nodes:  ['1219676']
```

Number of ways

```
SELECT COUNT(*) FROM ways;
```

```
Number of ways:  ['202857']
```

Number of unique users:

```
SELECT COUNT(DISTINCT(a.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) a;
```

```
Number of unique users:  2407
```

Top 10 contributing users

```
SELECT a.user, COUNT(*) as num
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) a
GROUP BY a.user
ORDER BY COUNT(*) DESC
LIMIT 10;
```

```
Nesim       117664
bigalxyz123 85980
Cicerone    61896
katpatuka   48717
Ckurdoglu   48180
JeLuF       47899
EC95        38317
```

```
canTurgay   36444
Sakthi20    27004
turankaya74 25278
```

Number of users appearing only once (having 1 post)

```
SELECT COUNT(*)
FROM
    (SELECT a.user, COUNT(*) as num
     FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) a
     GROUP BY a.user
     HAVING num=1)  b;
```

```
Number of users appearing only once:  ['679']
```

First contribition date

```
 SELECT timestamp FROM Nodes UNION SELECT timestamp From Ways
        ORDER BY timestamp desc
        LIMIT 1;
```

```
First contribution :  ['2007-03-09T15:50:46Z']
```

Additional Ideas

Additional Data Exploration

Top 10 appearing amenities

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

```
Pharmacy            2614
Restaurant          1062
Cafe                843
Bank                509
Fuel                409
Fast_Food           335
```

```
Parking             313
Atm                 284
Place_Of_Worship    270
School              201
```

Biggest religion (no surprise here)

```
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
    JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='Place_Of_Worship') i
    ON nodes_tags.id=i.id
WHERE nodes_tags.key='religion'
GROUP BY nodes_tags.value
ORDER BY num DESC
LIMIT 1;
```

```
Biggest religion :  ['Muslim 187']
```

"' # Conclusion

The İstanbul OpenStreetMap dataset has many typo errors which caused by human. The volume of data is huge. Considering there are many of contributors for this map, there is a great numbers of human errors in this project. It makes hard to find and clean all possiblities. With these project, only few data could be cleaned, but it satisfies the expectations to learn how to explore data set.

In OpenStreetMap, some prevention could be implemented to decrease the human errors. Also, new features could be added to increase the interaction with the places.

- For consistency, one global language could be selected and users only use the selected language to input data. It could be translated in their native language when they use data. However, it does not avoid the typo errors. The contribition of users could be decreased too. Users who do not know the selected global language as we expected could be given up to make contribition or make typo errors more than now.
- Competitions could be organized to encourage users. The user who makes the most contributions and at least misspellings could be rewarded with badges, gifts or travels.
- Especially in touristic cities, people who come to visit want to hear some advices from local people. The most popular cafe, restaurant, avenue could be voted people who live the city and people who visit the city seperately. These votes may give different results, but this system could be increased the contribition and interaction.

İstanbul is the most crowded area in our country. The dataset could not contain all informations. When the contribition is increased, the data shows more places that I do not go than now. But this data set gives a few new things about my hometown with using SQL.

Files

- `README.md` : this file
- `OpenStreetMapReport.pdf` : pdf format of md file
- `nodes.csv` : nodes csv
- `nodes_tags.csv` : nodes tags csv
- `ways.csv` : ways csv
- `ways_tags.csv` : ways tags csv
- `ways_nodes.csv` : ways nodes csv
- `OpenStreetMap.db` : Database
- `istanbul_turkey.osm` : sample data of the OSM file
- `osm_clean.py` : Incorrect street, city name find and update their names
- `csv_create.py` : build CSV files from OSM
- `import_data.py` : create database of the CSV files and import data to related tables from CSV files
- `query_executer.py` : execute given sql file
- `sql/select_way_size.sql` : query of Number of ways
- `sql/select_nodes_size.sql` : query of Number of nodes
- `sql/unique_users.sql` : query of Number of unique users
- `sql/contributing_user.sql` : query of Top 10 contributing users
- `sql/users_appearing_only_once.sql` : query of Number of users appearing only once
- `sql/biggest_religion.sql` : query of Biggest religion
- `sql/first_contribution.sql` : query of First contribution date
- `sql/amenity.sql` : query of Most popular amenity