OpenStreetMap Project-Data Wrangling with SQL

Choice of City:

I have chosen Chennai city in India,because it is my hometown.I want to find interesting insights about the city which I care the most.
https://mapzen.com/data/metro-extracts/metro/chennai_india/

Problems Encountered in the dataset:

- Abbreviated Street Names(Ex:S Mada Street,3rd St, Ragavan Colony, Mettuppalayam, Ashok Nagar, Chennai)
- Postal code of Chennai should be of 6 digits in length and without a space between them.
- Some of the tag k values representing area names are written in local language Tamil.

Let us explore each problem in detail.

1.Abbreviated Street Names:

Street Names containing abbreviated characters like S incase of S Mada Street and St incase of 3rd St, Ragavan Colony, Mettuppalayam, Ashok Nagar, Chennai appears in the middle of the address.Also certain records contain the whole address in place of Street name.I have modified the update_name function in audit.py file such that the entire street name is searched for common abbreviated characters like S for South,St for Street,Rd for Road etc and the values are replaced.

```
def update_name(name, mapping):
   words = name.split()
   for w in range(len(words)):
      if words[w] in mapping:
         words[w] = mapping[words[w]]
   name = " ".join(words)
   return name
```

2.Incorrect Postal Code:

Postal code of certain records contain 7 characters like 6000042,2 characters incase of 16 which doesn't follow the norm of 6 digits.Such records are skipped.11 records contains a space between 3 digits in case of 600 003.Space between digits are trimmed and the resulting postal code is stored in database.

```
if(tag_values["key"] == "postcode"):
                if(len(tag_values["key"])!=6):
                   break
                else:
                   v_value=v_value.replace(" ","")
                   tag_values["value"]=v_value
```

Data Overview:

File Sizes:

chennai_india.osm---390MB(uncompressed)
nodes.csv----------147MB
nodes_tags.csv----------904KB
ways.csv----------24.2MB
ways_nodes.csv----------54.1MB
ways_tags.csv----------14.1MB

relations.csv----------56.7KB
relation_members----------96.1KB
relation_tags----------63.3KB
Openstreet_Chennai.db----237MB

Number of Nodes:

```
SELECT COUNT(*) FROM nodes;
```
1833084

Number of Ways:

```
SELECT COUNT(*) FROM ways;
```
410990

Number of Relations:

```
SELECT COUNT(*) FROM relations;
```
975

Number of Unique Users:

```
SELECT COUNT(DISTINCT(e.uid)) FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
```
1091

INSIGHTS FROM DATA:

Names in Tamil:

I want to find the number of area names which are captured in Tamil in the dataset and compare it with the total area names present in the dataset.

```
select count(*) from (select * from node_tags union all select * from way_tags)e where e.key='ta';
```
546
```
select count(*) from (select * from node_tags union all select * from way_tags)e where e.key='name';
```
15782

Approximately 3% of the area names are captured in Tamil.

Most Frequent shops:

A number of shops of different categories are present in Chennai.Let us find the top 10 categories of the shops along with their count.

```
SELECT e.value, COUNT(*) as num FROM (SELECT * FROM node_tags UNION ALL SELECT * FROM way_tags) e where e.key='shop' GROUP BY e.value ORDER BY num DESC LIMIT 10;
```

supermarket|123
convenience|52
bakery|49
clothes|38
car|24
hairdresser|21
department_store|20
electronics|17
greengrocer|16
car_repair|15

SuperMarkets are found frequently in Chennai followed by Convenience store.

Popular Sports Centre in Chennai:

Sports centre is a place where aspiring sports persons undertake training to improve skills.Let us look at the query to find the top 10 popular sports centres in Chennai.

```
select e.value,count(*) as num from (select * from node_tags union all select * from way_tags)e where e.key='sport'
group by e.value order by num desc limit 10;
tennis|24
basketball|10
swimming|9
cricket|8
soccer|8
multi|7
athletics|4
motor|4
football|3
hockey|3
```

Tennis takes the top position with lions share of the total centres followed distantly by basketball.

Summary Statistics on Sports Centres:

23.5% of sports centres belongs to Tennis

Top 6 centres contribute to 55% of the total sports centres in Chennai

Sports having only one centre

```
 select u.value from (select e.value,count(*) as num from (select * from node_tags union all select * from way_tags)e
where e.key='sport' group by e.value having num=1)u;
archery
badminton;table_tennis;weightlifting
baseball
beachvolleyball
cricket_nets
field_hockey
horse_racing
netball
roller_skating
running
skateboard
skating
table_tennis
```

Here table_tennis appears twice.So skipping that from count.

```
select * from way_tags where value='badminton';
96687172|sport|badminton|regular
```

Badminton is also played in another sports centre.Weighlifting is played in only one centre.There are 12 sports which are played only in one centre.

Each centre is dedicated to one sport.But the first value returned **"badminton;table_tennis;weightlifting"** shows that 3 sports are practised in the same centre.Let us find the centre name.

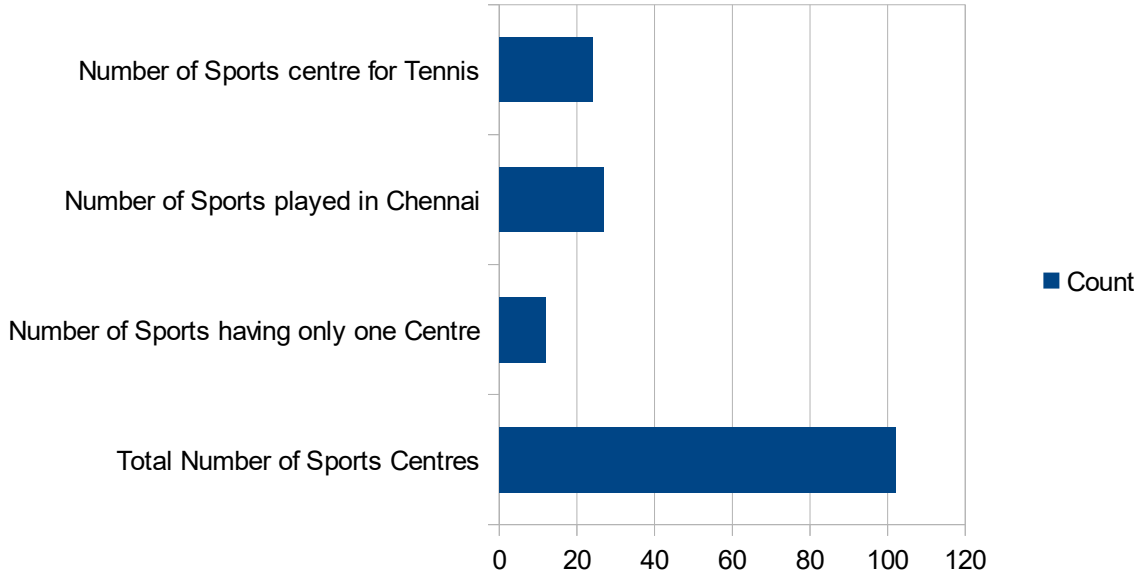
```
select way_tags.value from way_tags join(select * from way_tags where
value='badminton;table_tennis;weightlifting')u on way_tags.id=u.id where way_tags.key='name';
```

Sports Complex

Distinct Sports Played in Chennai:

Let us find the count of distict sports played in Chennai.

```
 select count(*) from (select distinct(e.value) from (select * from node_tags union all select * from way_tags)e where e.key='sport')u;
27
```



Areas where Internet Access is available:

Let us look at the places where internet access is available.Some places have mentioned wlan is available and for some places just internet access is available is provided.

```
SELECT i.value as num FROM node_tags  JOIN (SELECT id,value FROM node_tags WHERE key='name') i ON node_tags.id=i.id WHERE node_tags.key='internet_access' and node_tags.value !='no' GROUP BY i.value ORDER BY num DEsc;
thilak cafe
recreation centre
Vistana
Sree kamakshi silk house
Red Lollipop Hostel
Pandian Computers Chennai
Mylapore KANI GL
Farida's
CTS
Absolem
ARAVIND Traveler's Coffee
```

Other Ideas about dataset:

User who adds details about an area can use ta attribute to specify area names in Tamil.This will motivate other users from the Chennai region to specify details in their mother tongue.In most of the street names,along with the street name,area name is also included(Ex:Lloyd's Colony, Triplicane).Here Llyod's colony is the street name and Triplicane is the area name.A new attribute called area can be used or clubbed with exisiting name attribute.

Conclusion:

After reviewing Chennai data I find that more refining while inputting data would help in better details.Our own data.py can be extended to check all the elements present in the xml file.I have cleaned the data for Chennai in the best possible manner.I am happy to contribute to OpenStreetMap.org for the city I love.