Case Study: OpenStreetMap Data

Map Area

Austin, Texas, US (https://www.openstreetmap.org/relation/113314)

Austin has been my home for the last six years and it has been one of the fastest growing cities in America for several years. I'm interested to see how this grow is reflected in the data and try to draw conclusions on how the city is being affected by the increase of population.

Problems Encountered In The Map

I first downloaded a medium sample size of the data in the Austin area and ran it against a data.py file, which turns the data into csv files with the desired SQL schema. Next I created a SQL database, AustinTexasOsm.db, with the desired tables and did some basic queries to explore the data. The following problems where found:

- Abbreviated street names
- Inconsistent postcodes ("TX 78758" and "78704-7205")
- "FIXME" keys and types in the nodes_tags and ways_tags tables
- GNIS data is not relevant

This problems and my solution to them are discussed in further detail.

Abbreviated Street Names

Altough the street name data from the Austin map was surprisingly clean there was still some work to be done. The data was first audited using the audit_and_mapping.py script to determine the different types of abbreviated street names and map them to their complete name. Then the function shown below was used in the data.py file to unabbreviate the names.

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

    return name
```

Inconsistent Postcodes

Postcodes were introduced into the database in three main formats:

1. 78705
2. TX 78705
3. 78705-7564

I have decided to turn all postcodes into the five digit format, as it is the most widely used by far. After making this decision, cleaning up the data was easily done by dropping the string values before and after the five digits of interest using the following python function.

```
def update_postcode(postcode):
    if postcode[:2]=="TX":
        postcode = postcode[3:]
    elif postcode[5:6]=="-":
        postcode = postcode[:5]

    return name
```

Once the consistent values were reintroduced into the SQL database the following query brought out a few discrepant postcodes ("tx", "Texas", "14150"), which were deleted. Otherwise, the results of the query were as expected.

```
SELECT value, count(*) as postcode FROM
(SELECT value, key FROM nodes_tags
    UNION ALL
    SELECT value, key FROM ways_tags)
WHERE key='postcode'
group by value
order by count(*) desc;

DELETE FROM ways_tags
WHERE key='postcode'
    AND (value='tx' OR value='Texas' OR value= '14150')

DELETE FROM nodes_tags
WHERE key='postcode'
    AND (value='tx' OR value='Texas' OR value= '14150')
```

"FIXME" Keys and Types

The "FIXME" key allows users to leave a note to themselves or other contributors on how their entry could be improved as well as noting errors in the existing data. Although these notes are important to the functioning of Open Street Map they do not give much insight into our Austin data, except for giving a very general idea of data quality or consistent user interaction. Before deleting the "FIXME" rows a count of such rows is made.

```
SELECT key, count(*) as FIXME FROM
(SELECT value, key, type FROM nodes_tags
    UNION ALL
    SELECT value, key, type FROM ways_tags)
WHERE key='FIXME' OR type='FIXME';

FIXME|126
```

Out of more than two million rows, only 126 "FIXME" values are found, which suggest than Open Street Map users in Austin are active and the data is cleaned constantly. However, this assumption is completely speculative and much more research would need to be done to find if lack of "FIXME" notes indicates data cleanliness and high activity.

GNIS Data

The USGS Geographic Names Information System (GNIS) is a database for geographic features in the US and Antartica. The issue with GNIS is that it is a database of names rather than features and it was uploaded into OSM without verifying the names in the GNIS database still existed. Now, there is no way to know if the feature being referenced exists making the data irrelevant and unreliable. The data is deleted inside SQL to avoid mix ups.

Overview of the Data

This secton contains statistics about the database and files. SQL queries are also presented when appropiate.

Size of Files

```
austin_texas.osm_______1.41 GB
austin_texas.db________799.5 MB
nodes.csv______________600.6 MB
nodes_tags.csv_________11.4 MB
ways.csv_______________48.1 MB
ways_tags.csv__________68.8 MB
ways_node.csv__________169.5 MB
```

Number of Unique Users

```
SELECT count(*) as unique_users FROM
(SELECT uid, user FROM nodes
    UNION
    SELECT uid, user FROM ways);
```

```
1251 unique users
```

Top 10 Contributors

```
SELECT user, count(*) as top_users FROM
(SELECT user FROM nodes
    UNION ALL
    SELECT user FROM ways)
group by user
order by count(*) desc
limit 10;
```

```
patisilva_atxbuildings    2742450
ccjjmartin_atxbuildings   1300427
ccjjmartin__atxbuildings  940002
wilsaj_atxbuildings       358804
jseppi_atxbuildings       300854
woodpeck_fixbot           221391
kkt_atxbuildings          157844
lyzidiamond_atxbuildings  156357
richlv                    49800
johnclary_axtbuildings    48227
```

Total contributions by the 10 most active users accounts for 89% of all data.

Number of Nodes

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

```
6387905 nodes
```

Number of Ways

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

```
669630 ways
```

Top 10 Amenities

```
SELECT value, count(*) as amenities FROM
(SELECT key, value FROM nodes_tags
    UNION ALL
    SELECT key, value FROM ways_tags)
WHERE key='amenity'
group by value
order by count(*) desc
limit 10;
```

```
parking            2198
restaurant         807
waste_basket       603
fast_food          597
school             559
place_of_worship   516
fuel               443
bench              360
shelter            241
bank               185
```

Cafes vs. Bars

```
SELECT value, count(*) as amenities FROM
(SELECT key, value FROM nodes_tags
    UNION ALL
    SELECT key, value FROM ways_tags)
WHERE key='amenity' AND (value='cafe' OR value='bar')
group by value
```

```
bar    151
cafe   140
```

Phew! barely more bars than cafes.

Wikipedia's Guide to Austin

```
SELECT value as wiki FROM
(SELECT key, value FROM nodes_tags
    UNION ALL
    SELECT key, value FROM ways_tags)
WHERE key='wikipedia'
group by value
limit 15;
```

```
A. J. Jernigan House
Allens Boots
Anderson High School (Austin, Texas)
Austin Community College District
Austin High School (Austin, Texas)
Austin Independent School District#Elementary schools
Austin, Texas
Austin–Bergstrom International Airport
Barton Springs Pool
Bastrop, Texas
Big Lots
Blanton Museum of Art
BookPeople
Bowie High School (Austin, Texas)
Breakaway Airport
```

This can actually be a cool new way to explore new cities!

Additional Ideas

One of the main issues I see with the OSM data for Austin is that is seem outdated as most data is from before the year 2016 began.

```
SELECT timestamp , count(*) as amenities FROM
(SELECT timestamp FROM nodes
    UNION ALL
    SELECT timestamp FROM ways)
WHERE timestamp LIKE '2016%'
```

There were only 99,439 entries in 2016 and 24,072 in 2017 so far.

Given that when the OSM project started they imported several databases and the entry rate was much higher because the maps were at their infancy, my suspicion does not hold much ground. However, a higher quality data entry rate would result in better maps/osm databases and could be done by incentivizing apps that use OSM to upload some of the data they gather. This would give OSM a much quicker reaction time to an evolving city such as Austin. This of course could be a privacy issue but I believe most people would like the idea of having contributed towards mapping their city.