## Project 3(Wrangling OpenStreetMap data)

**OpenStreetMap** (**OSM**) is a collaborative project to create a free editable map of the world. The creation and growth of OSM has been motivated by restrictions on use or availability of map information across much of the world, and the advent of inexpensive portable satellite navigation devices. OSM is considered a prominent example of volunteered geographic information.

In this project, I downloaded the OSM file(size:54.8 MB) for the city of Ann Arbor in Michigan and used data munging techniques, such as assessing the quality of the data for validity, accuracy, completeness, consistency and uniformity, to clean the OpenStreetMap data for that part. Then I imported the data to SQLite database and did some data exploration on it.

For starting this project, I just looped through "way" and "node" tags to get the secondary tags that were included in these two tags. In the first and second cells of the notebook, I got the distinct child tags for "way" and "node" tags, and printed out the keys and their values for these two tags.

I tried to figure out how to clean the data by looking at these child tags and select the appropriate tags to clean.

In the third cell, I displayed the value of some of the child tags, one by one, to see whether they need to be cleaned or not.

I examined "gnis:date", "soure:date" and "survey:date" and they all had consistent date format:

| field | Date format |
|---|---|
| gnis:date | MM/DD/YYYY |
| Source:date | YYYY-MM-DD |
| Survey:date | YYYY-MM-DD |

As a result, I didn't make any changes in these fields.

**House number:**

I also figured out that in "addr:housenumber" all of the housenumber contain just the digits, except several of them which contain letters, and one which was a range. By investigating more in OpenStreetmap website , we can figure out that having a range of housenumber is reasonable since some buildings are with multiple house numbers. Also having housenumbers like 340A is not far from our expectations so housenumber field is fine and doesn't need cleaning.

**Phone number:**

Another field that I was interested in was "contact:phone" which needed to be cleaned.

I figured out that "contact:phone" was entered in the following formats:

+1 *** ***-****

+1 *** *** ****

So, I just decided to make all of them the same by eliminating the hyphen(-) and putting space instead of hyphen. This part is done in the section “preparing for database”

**Problematic street names:**

I also worked on problematic street names and make them consistent. This is done is cell 4 of the notebook.

**Post code:**

Another field was “postcode”, I figured out that most of the postcodes just contain 5 digits, a few of them contain hyphen and four digits after that (*****-****) and also two of them contain the state name (MI). So I made all of them consistent in a way that they just contain 5 digits.

**Address field:**

I also read in Udacity forum that sometimes for address there are multiple entries with the same information (for example, 'address' keys with the complete address as one value, as well as the address broken down into 'addr:street' in the same element), by further investigation I figured out that there is no filed starting with “addr:” other than “addr:street”, so the address field was fine.

**Zipcode:**

Some of the zipcodes contain the abbreviation of the state which are unnecessary and also some contain four digits after the main post code which I want to exclude in order to have more consistency.

Possible zipcodes for Ann Arbor: (48103,48104,48105, 48106, 48107, 48108, 48109)

## Data exploration using SQLite :

### Number of nodes and ways:

```
number of nodes:248675
number of ways:35683
```

### Number of unique users:

'''

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

'''

```
Output: 328
```

### Number of shops:

'''

```
SELECT COUNT(*),key FROM nodes_tags WHERE key=="shop";
```

'''

```
number of shops:398
```

**Number of sports:**

'''

```
SELECT COUNT(*),key FROM nodes_tags WHERE key=="sport";
```

'''

```
number of sports:380
```

**Number of cafes:**

'''

```
SELECT COUNT(*),key FROM nodes_tags WHERE value=="cafe";
```

'''

```
number of cafes:48
```

**Different types of amenities:**

'''

```
SELECT DISTINCT value FROM nodes_tags WHERE key=="amenity";
```

'''

```
[(u'library',), (u'childcare',), (u'fuel',), (u'hardware_store',), (u'pub',),
(u'bicycle_parking',), (u'school',), (u'place_of_worship',),
(u'grave_yard',), (u'hospital',), (u'doctors',), (u'bench',), (u'cafe',),
(u'atm',), (u'parking',), (u'restaurant',), (u'police',), (u'bar',),
(u'fire_station',), (u'fast_food',), (u'vending_machine',), (u'prison',),
(u'toilets',), (u'post_box',), (u'fountain',), (u'car_wash',), (u'bank',),
(u'marketplace',), (u'parking_entrance',), (u'post_office',), (u'shelter',),
(u'drinking_water',), (u'nightclub',), (u'pharmacy',), (u'parking_space',),
(u'waste_basket',), (u'bus_station',), (u'cinema',), (u'theatre',),
(u'car_rental',), (u'dentist',), (u'charging_station',), (u'ice_cream',),
(u'recycling',), (u'brokerage',), (u'clinic',), (u'social_facility',),
(u'bicycle_repair_station',), (u'studio',), (u'food_court',),
(u'healthcare',), (u'compressed_air',), (u'office',), (u'community_centre',),
(u'university',), (u'storage',), (u'veterinary',), (u'dojo',),
(u'financial_advice',), (u'Speech Therapy',), (u'arts_centre',),
(u'public_building',), (u'townhall',), (u'concert_hall',), (u'college',),
(u'kindergarten',), (u'trailer_park',), (u'labour_union',),
(u'animal_boarding',)]
```

**Top 5 popular sports:**

'''

SELECT nodes_tags.value, COUNT(*) as num FROM nodes_tags WHERE key=="sport" GROUP BY nodes_tags.value ORDER BY num DESC LIMIT 5;

'''

```
[(u'tennis', 111), (u'baseball', 90), (u'soccer', 66), (u'basketball', 48),
(u'american_football', 15)]
```