OpenStreetMap Case Study

Map Area

- http://www.openstreetmap.org/relation/112298
- https://mapzen.com/data/metro-extracts/metro/denver-boulder_colorado/

Boulder/Denver, CO, United States

The reason I chose this area is because I am from the city of Boulder and this case study presents a unique opportunity to learn more about the city and the surrounding region. The city of Boulder is located just outside of Denver and is often grouped regionally with Denver, hence the dataset featuring both cities.

Auditing the Data

For the most part, this dataset required fairly minimal wrangling. The most pressing issue is the non standardized abbreviations such "st." or "st". Along the same lines, "CO" was changed to "Colorado" to maintain consistency. I expanded these values out using the `update_name` function:

```
def update_name(name, mapping):

    for k,v in mapping.iteritems():
        if re.search(v,name):
            return name
        else:
            name = re.sub(k,v,name,1)

    return name
```

Additionally, there were some redundancies in some records such as county where the state was included as well e.g. `county='Jefferson,CO'` which needed to be changed like so:

```
if re.search(", CO", value):
  value = value.split(",")[0]
```

Some attribute values had some odd key values such as `k='type'` which is confusing given the code to add the attribute type to the table.

Dataset Overview

After auditing and cleaning the data, they were loaded into a SQLite3 database. Some brief overview statistics are as follows:

File Sizes

```
denver-boulder_colorado.osm ....... 853.2 MB
boulder_denver.db ................. 600.7 MB
```

```
nodes.csv ......................... 323.3 MB
ways_nodes.csv .................... 104.2 MB
ways_tags.csv ...................... 61.1 MB
ways.csv ........................... 25.8 MB
nodes_tags.csv ..................... 18.8 MB
```

Unique Users

Users could contribute to this dataset either by adding nodes and/or ways.

```
sqlite> SELECT COUNT(DISTINCT a.uid)
   ...> FROM  (SELECT uid FROM nodes UNION ALL SELECT uid from ways) a;
2001
```

2001 unique users, that's a lot!

Note: attempting this with a join for a database this massive is unfeasible.

Total Number of Nodes and Ways

The total number of nodes can be gathered using a simple query:

```
sqlite> SELECT COUNT(*) FROM nodes;
3772108
```

There are 3,772,108 nodes! Given the amount of metadata associated with each node, this is truly amazing for what amounts to a relatively small metro area compared to others in the United States.

Similarly, for ways we have:

```
sqlite> SELECT COUNT(*) FROM ways;
421926
```

As expected, there is a smaller amount of ways in comparison to nodes since ways are essentially abstractions of multiple nodes.

Examining Types of Nodes

Looking at the node_tags values for nodes can bring up some interesting statistics for describing the Boulder/Denver region. Let's take a look at some:

Amenities

First an overview of the types of amenities could be useful before honing in on some certain types.

```
sqlite> SELECT value, COUNT(*) as num
   ...> FROM nodes_tags
   ...> WHERE key='amenity'
   ...> GROUP BY value
   ...> ORDER BY num
```

```
   ...> DESC
   ...> LIMIT 20;
```

```
restaurant ........ 1677
bicycle_parking .... 841
fast_food .......... 757
school ............. 752
bench .............. 724
place_of_worship ... 632
fuel ............... 498
parking ............ 482
cafe ............... 427
bank ............... 318
toilets ............ 307
parking_space ...... 287
post_box ........... 272
fire_station ....... 259
pub ................ 199
fountain ........... 196
bar ................ 187
atm ................ 178
swimming_pool ...... 130
waste_basket ....... 121
```

The results are interesting to say the least. Restaurants are by far the most popular amenity which is unsurprising (we have great food!) but what stands out the most is bicycle_parking in second place. This amenity is fairly vague; does it refer to designated bike rack areas? How many bike racks are needed to comprise a bicycle parking space? One? Three?

Regardless of how its defined, this result is plausible given how active the Boulder community is. Commuting via bike is very popular among college students and Boulderites which is in part why the city commonly receives the distinction of "Healthiest City in the United States".

Restaurant is also distinct from fast food which is interesting as I would expect fast food to outnumber classic, sit-down restaurants. Although it is not clear what the distinction between fast food and a restaurant is, looking at the most popular restaurants could give a rough idea of what the distinction may be.

Popular Restaurants

```
sqlite> SELECT nodes_tags.value, COUNT(*) as num
   ...> FROM nodes_tags,
   ...> (SELECT DISTINCT id FROM nodes_tags WHERE value='restaurant') as r
   ...> ON nodes_tags.id = r.id
   ...> WHERE nodes_tags.key='name'
   ...> GROUP BY nodes_tags.value
   ...> ORDER BY num
   ...> DESC
   ...> LIMIT 10;
```

```
Chipotle ............... 15
"Panera Bread" ......... 15
"Village Inn" .......... 14
```

```
"Tokyo Joe's" .......... 13
"Noodles & Company" .... 10
"Chili's" ............... 7
"Pizza Hut" ............. 7
Qdoba ................... 7
"Qdoba Mexican Grill" ... 7
Smashburger ............. 6
```

The results of popular restaurants reveals an oversight in the auditing process. In my samples, names did not have quotation marks around them but as we can see here, they can be duplicated if not taken out. Here, it's unknown how many Qdoba there actually are. Further inspection of the full list of restaurants revealed that outside of Qdoba, the top 10 listed is still indicative of the relative popularity of restaurants. However, this information is good to know in the event that I need to audit again for a more serious issue.

Back to the original question, it seems like "fast-casual" restaurants are distinct from fast food restaurants. This is to be expected as a number of these nationwide fast-casual chains actually started in the Boulder/Denver area, including: Chipotle, Village Inn, Noodles & Company, Qdoba, Tokyo Joe's, and Smashburger. That's 6 out of the top 10 listed! With this in mind, the earlier result of the amount of restaurants compared to fast food makes more sense.

Suggestions For Improvement

There are several things that could potentially alleviate some of the issues with analyzing this database. One is the amount of hard to utilize data from tiger GPS which is presumably added in bulk automatically via various methods. While potentially useful, its entries were numerous in ways_tags but was not very understandable compared to the human entries.

That being said, greater standardization of human entered nodes and ways could go a long way in reducing the amount of auditing necessary prior to actual use of these data. Perhaps a regional documentation guideline could be useful for OSM contributors who can follow a standard set by other local users so as to use a format that best suites the region. Common stylistic choices such as whether or not to wrap quotation marks around data values or how to handle abbreviations should be handled could all be standardized. The reason this would be beneficial at the regional/city level is that certain style choices might make more sense for certain regions and by having some granularity of differences, users can adapt to existing OSM data style trends wihch minimizes the amount of refactoring needed for old data.

This of course is not a perfect solution nor do I pretend it to be, however, it does seem flexible in that it does not dictate *what* data there is, rather, *how* to input it. The big issue with any standardization effort is the inabilty to get everyone to agree on one thing but I would hope that a smaller community of users might be able to do so. As it turns out, the majority of contributions are done by very few people so it could potentially cause problems if these few users all use different style choices and can't agree on what to use.