Data Analysis of Atlanta Area OpenStreet Map

MAP AREA:

http://www.openstreetmap.org/export#map=5/42.618/-7.559
(minlat=32.844, minlon=-85.386, maxlat=34.618, maxlon=-83.269)

DATA WRANGLING PROCESS

My data wrangling process began with the XML file download of the given geographical map location above from openstreetmap.com. The download was over 2 gigabytes in size, quite large. I decided to begin by first analyzing and exploring the data contained before asking any questions and see what insights might be gleaned by exploring. The idea was to convert and store xml data in database in order to more easily explore for insights and trends. In order to accomplish this, my first step, audit the XML data. Auditing the data required I check for validity, accuracy, completeness, consistency and uniformity. To audit for validity, I set constraints for the different kinds of data contained in the dataset and verified that these constraints were met. There were a number of different constraints set for different attributes in the dataset. A list of the attributes and their constraints. id, version, and uid had an integer constraint. lat, lon, had a float constraint. Timestamp, user, k, v attributes had a string constraint. To audit (for validity) each piece of data in the dataset I used ElementTree python module to parse the XML dataset. Due to its enormous file size, I would not be able to manually sift through millions of records, that would require an inordinate amount of time. I wrote a script to sift through each record and verify that each attribute in the record met the constraint requirement. If an attribute did not meet a constraint requirement, it was written to separate file to be manually diagnosed, otherwise, it was written to a CSV file. Next, I audited for accuracy. To do this, I compared a sample XML file from the authoritative standard, osm wiki, to my dataset. I ensured all expected tags and attributes were present in my dataset. Auditing for completeness was a bit out of reach but I it's safe to say the dataset is never complete due to continuous change in the environment from new building erections, new shopping malls, and other structures the dataset is unceasingly morphing. My final audit step required I verify that each record was consistent with others like it. In order to verify this requirement, I programmed a solution to check each type of tag for certain attributes. If a tag was missing certain attributes, it was written to different file for manual diagnosis. Once the auditing was complete all that remained was to write the data to a database. For this last step, I made use of the database management software Sqlite and database API sqlite3 to import CSV files, created after the auditing process, to SQL database. That is how I prepared my dataset for exploration and analysis.

PROBLEMS ENCOUNTERED DURING ANALYSIS

Inconsistency with some zip codes:
  some node tags had a single zip code, others had a range.
Inconsistency with building designations:
  singular and plural forms of a given designation,
  lowercase and uppercase wording for other designations.
  some houses were given a zip code for city name.

GENERAL STATISTICS

Size of atlant_georgia.osm file: 2.4G

Number of nodes: 11739508
  select count() for nodes;
Number of ways: 847011
  select count() for ways;
Number of relations: 4241
  select count() for relations;
Number of node tags: 1802612
  select count() for node_tags;
Number of way tags: 4231798
  select count() for way_tags;
Number of way nodes: 12488808
  select count() for way_nodes;
Number of relation tags: 18333
  select count() for relation_tags;
Number of member tags: 31003
  select count() for member_tags;
Number of distinct node creators: 2,053
  select count() from (select distinct count() from nodes);
Number of distinct way creators: 1,617
  select count() from (select distinct count() from ways);
Number of distinct relation creators: 234
  select count() from (select distinct count() from relations);
User with most entries for nodes: Liber with 5,209,113 entries
  select user,count(user) from nodes group by user order by count(user) desc;
User with most entries for ways: Saikrishna_FultonCountyImport with 256,918 entries
  select user,count(user) from ways group by user order by count(user) desc;
user with most entries for relations: Liber with 1,012 entries
  select user,count(user) from relations group by user order by count(user) desc;

From the above statistics, the user, Liber, was a very active contributor to the dataset, responsible for almost half of node entries in the dataset. Saikrishna_FultonCountyImport responsible for a huge slice of way entries as well. It seems a few people are responsible for the bulk of information entered into the dataset.

OVERVIEW AND ADDITIONAL IDEAS

Some designations describing nodes:
  select distinct key from node_tags;
  source,addr,highway,name,ele,power,amenity,natural,religion,
  place,is_in,shop,waterway,traffic,aeroway,shelter,access
  building,wheelchair,bike route,public transport,historic,
  takeaway,delivery,ATM,subway,parking, and a lot more..

Some statistics:
  select key,value,count() from node_tags group by key,value order by count() desc;
  1,946 railway crossing
  15,641 power towers
  31,165 turning circles
  5,606 natural trees locations
  5,299 traffic signals

Top 10 zip codes with the most places tagged:
  select value, count() from node_tags where key='addr:postcode' group by value order by count() desc;
  30114 with 7,272 nodes

30188 with 7,110 nodes
30115 with 5,461 nodes
30132 with 5,197 nodes
30189 with 5,024 nodes
30349 with 4,333 nodes
30213 with 3,052 nodes
30107 with 2,943 nodes
30157 with 2,890 nodes
30102 with 2,440 nodes

Top 10 Amenities:
select value, count() from node_tags where key='amenity' group by value order by count() desc limit 10;
place of worship with 3,994
grave yards with 2,086
schools with 2,070
restaurants with 964
fast food with 533
fuel with 323
bench with 317
fire station with 237
post office with 214
ATM with 209

Religion with most locations:
select value,count() from node_tags where key='religion' group by value order by count() desc limit 10;
Christian with 3933 locations
Jewish with 7 locations
Muslim with 4 locations
Unitarian_Universalist with 2 locations
Hindu with 1 location

We see that the Atlanta area is overwhelmingly of the Christian faith which isn't a big surprise. Followed by the Jewish
faith, Muslim faith, Unitarian Universalist, and Hindu faith. I was surprised to see their are more graveyards
than schools, restaurants, and gas stations. Didn't realize there were that many graveyards in this city.

Some designations describing ways:
highways,
buildings,
waterways,
cycleways,
one ways,
bridges,
railways
...
Types of highways:
residential,
service,
foot way,
unclassified,
secondary,
tertiary,
primary,
paths,

motorway,
motorway_link
Types of buildings:
commercial,
tower,
civic,
industrial,
stadium,
retail,
office,
apartments,
public,
residential,
bank,
hotel
…

User with most entries for commercial buildings: maven149 with 1297 entries
select ways.user, count(ways.user) from ways,way_tags where ways.id=way_tags.way_id and way_tags.key='building' and way_tags.value='industrial' group by ways.user order by count(ways.user) desc limit 10;

User with most entries for industrial buildings: Jack the Ripper with 168 entries
select ways.user, count(ways.user) from ways,way_tags where ways.id=way_tags.way_id and way_tags.key='building' and way_tags.value='industrial' group by ways.user order by count(ways.user) desc limit 10;

Users who contributed to apartment entries:
SomeoneElse_Revert
select distinct ways.user from ways,way_tags where ways.id=way_tags.way_id and way_tags.key='building' and way_tags.value='apartment';

Users who contributed to hotel entries:
Sundance
mackerski
GoWestTravel
rjhale1971
maven149
mjn
tman0
nivardus
DeVietor
IanH
Jack the Ripper
cbyrne
Human BackPack
select distinct ways.user from ways,way_tags where ways.id=way_tags.way_id and way_tags.key='building' and way_tags.value='hotel';

Top 3 ways:
select key, count() from way_tags group by key order by count() desc;
buildings with 368,721 buildings
highways with 266,899 roads
waterways with 64,058 waterways

Top 10 type of roadways:
select value,count() from way_tags where key='highway' group by value order by

count() desc;
residential roads with 159,832 roads
service roads with 60,293 roads
foot way 9,011 roads
unclassified roads 5,881 roads
secondary with 5,851 roads
tertiary with 5,313 roads
primary with 3,731 roads
paths with 3,729 roads
motorway_link with 3,457 roads
motorway 2,265 roads

Most prevalent buildings:
select value,count() from way_tags where key='building' group by value order by count() desc;
houses with 201,829 buildings
commercial with 4,192 buildings
retail with 3,552 buildings
residential with 2,223 buildings
apartments with 1,228 buildings
school with 498 buildings
industrial with 445 buildings
church with 294 buildings
office with 172 buildings
university with 171 buildings

Top 10 cities where most houses:
select value,count() from way_tags where key='addr:city' group by value order by count() desc limit 10;
Atlanta with 112,845 houses
Roswell with 16,596 houses
Fairburn with 5,608 houses
Alpharetta with 5,031 houses
Marietta with 2,590 houses
Union City with 2,557 houses
Palmetto with 1,668 houses
Riverdale with 1,182 houses
Kennesaw with 650 houses
Smyrna with 299 houses

Most people in live Atlanta and Roswell which is not surprising. Having lived in a few of these cities. It's obvious that some are more populous than others. I am surprised that Kennesaw made the top 10. It's more populous, compared to other places, than I anticipated. One confusing aspect had to do with buildings designated as residential. Because there are other buildings with the tag, 'house', and 'apartment', the tag, 'residential' was a bit confusing as I had assumed a residential building was either house, apartment or gated community. Will dig further.

IDEAS FOR IMPROVEMENT

An idea that I believe would go a long way in improving the quality of the data and completeness of the data is along the lines of gamification. Quotas are set and for set quotas met by the volunteers they could be entered into a lottery or provided gift cards to certain places. Another idea, organize bi weekly or weekly meetings where volunteers get together and work on the project as a community. I believe that would greatly help in cultivating a sense of meaning and a sense of responsibility.

BENEFITS

The benefits from the implementation of the dataset improvement ideas, a more complete dataset would emerge and, by extension, a more accurate dataset. Also, a more reliable map would result.

ANTICIPATED PROBLEMS

I do see a few issues with my improvement ideas. To start, I'm not sure how motivated people will be given the allure of gift cards. I do believe some will definitely see worth in it but not very confident a significant number of people will.

Also, persuading other to give up their time is no easy feat and factoring for the unpredictability of life, schedules don't always work out as we'd like but I do believe, for the most part, that these issues can be mitigated.

CONCLUSION

I found some interesting information about the Atlanta area that were not readily apparent to me. Having only skimmed the surface of a 2.4 gigabyte file, there's much more information to be uncovered. Definitely found some inconsistencies. I cannot say I found all the inconsistencies but, so far in my exploration, I have not found any more discrepancies.