OpenStreetMap Project: Data Wrangling with SQL

By Fahzy Abdul-Rahman

Map Area: Irving, Texas, United States

For this project, I've selected a portion of Irving city, located in the state of Texas (TX) of the United States of America. The actual area selected is actually the area which I am considering for a house purchase.

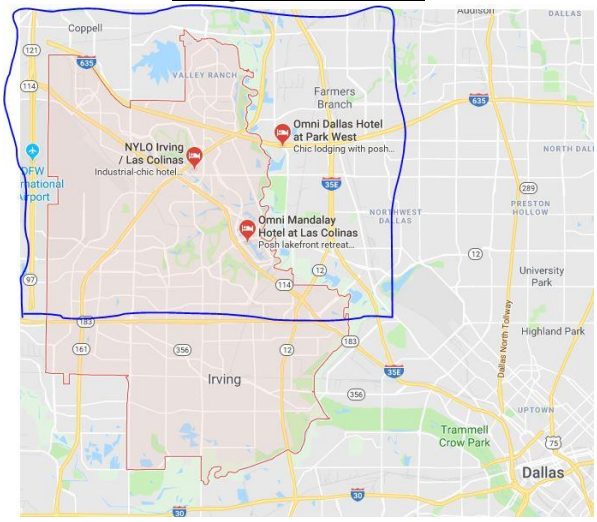
Irving, TX, USA: https://en.wikipedia.org/wiki/Irving,_Texas

The goals of this project are:

1. To extract data from an XML file;
2. To identify data inconsistencies and rectify them, if possible;
3. To investigate interesting findings based on queries.

As indicated on the map below, Irving city is located North West of Dallas downtown. West of Irving is the Dallas Fort Worth airport. The square area in blue is as estimated area selected for this project.

Irving & Dallas, Texas



Irving, TX, Located Close to Mexico Border

Overview of the Data

The OSM file for Irving city is quite clean for the most part. While I found slight issue with street name standardization, the zipcodes look legit. This may be due to a large portion of the data being inputted by a nearby city's economic development analyst.

```
The full OpenStreetMap file is 50.1 MB
while its sample file size is 1.0 MB ,
for records systematically selected every
50-th (top) row.

Nodes.csv file size: 19.9 MB
Nodes_tags.csv file size: 211.3 KB
Ways.csv file size: 1.5 MB
Ways_nodes.csv file size: 6.2 MB
Ways_tags.csv file size: 3.4 MB
```

```
Number of nodes: 231703
Number of nodes tags: 5651
Number of ways: 23602
Number of ways nodes: 264778
Number of ways tags: 100012

Number of ways tags: 100012
```

Data table size, users (distinct) count, top 10 contributors, top amenities

We have 329 contributors for the nodes and ways files. The most productive contributor is *Andrew Matheny*, who I found to be an Analyst at Allen Economic Development Corporation. Note: Allen and Irving are cities in the Dallas-Fort Worth Metropolitan area. Andrew's OSM Profile: http://www.openstreetmap.org/user/Andrew%20Matheny Andrew's LinkedIn Profile: https://www.linkedin.com/in/andrewmatheny

Under his LinkedIn Volunteer Experience, Andrew noted his experience as: "Mapping areas in OpenStreetMap where no digital maps exist, specifically in areas with humanitarian crises, a high risk for disaster, or poverty. These maps help NGOs, relief organizations, governments, and local communities distribute aid, navigate their communities and operate more intelligently. Contributed over 3,162 miles of roads".

```
Unique contributors, contributions: (329, 255305)

# Top 5 Contributors
                     user  NumRec
0  Andrew Matheny_import  100007
1        woodpeck_fixbot   29984
2             Stephen214   25295
3                Zachy_P   14464
4                dwh1985    8617
```

Wow! Andrew's edit actually represents almost two-fifths (39.2% [100,007/255,305]) of all nodes and ways edits. Numbers two and three on the list are woodpeck_fixbot (29984,11.7%) and Stephen214 (25295, 9.9%). This means that three users made three-fifth (60.8%) edits for the selected area's nodes and ways. I don't think this is how Open Source is supposed to work.

Amenities

It's not surprising that place of worship actually ranked as one of the top amenities in this region given its relatively religious background and the area is composed of mainly residential area. Food amenities with *burger* as their food types has the most amenity (9), followed by sandwich (6). This is consistent with I expected from a Dallas suburb city.

```
                value  NumAm
0   place_of_worship     51
1          fast_food     49
2         restaurant     44
```

```
                 value  num
0               burger    9
1             sandwich    6
2              chinese    4
```

3	post_box	24	3	coffee_shop	4
4	fountain	17	4	mexican	4
5	parking	14	5	pizza	4
6	school	12			
7	car_rental	10			
8	fuel	10			

Problems encountered in your map

As noted above, I found issues with street name standardization but not zipcode. When correcting abbreviations such as "Dr" and "St", extra steps need to be taken in order to about corrected results of "DrDrive" and "StStreet".

I did not find the dataset to be useful when it comes to residence unit. My hope in using this dataset for information for house hunting was crushed. Users are very interested in inputting and altering public amenities than residence units. I found more evidence on contributors not being too concerned about residential units. For instance, North Macarthur Boulevard is definitely a major road with many housing units. The housing units should be way more than 120 let alone 12! Other interesting observations:

- fast food, restaurants, and cafe are separately categorized
- fountain and waste basket categorizations
- 10 fuel? That cannot be right.

I also noticed that the hierarchical categorization could be improved. So for eating out places, I had to include expand the value selection to include 'restaurant', 'fast_food', and 'cafe'.

> i.e. WHERE value in ('restaurant','fast_food','cafe')

Data input for the OpenStreetMap should be taken with caution since it is an open-sourced information. Like Wikipedia, incorrect or intentionally false information may creep into the site. For instance, *Iowa Kid* has the most residence street input but does not have reputable score on the OpenStreetMap site.

I am not used to SQL queries producing output in tagging format. So, for most queries, I have utilized pandas.read_sql_query command to obtain output in table format.

The .db file (mine: irvingtx.db) costed me 20 hours after I forgot to include coding lines on data updating, update_name() . I found its solutions but couldn't see the corrected street name updates being reflected in the output. It turned out that I have to manually delete the .db file or create another .db. Otherwise, the .db will keep the old data from shape_element().

Other ideas about the datasets

I found that Andrew Matheny was involved in close to 40% of all nodes and ways records, but he is not concerned about residence units. As a researcher myself, i would be interested in seeing how Andrew utilized this map information. Note that Andrew is an analyst of a new, small, but booming city of Allen. My hypotheses are that:

- his team is using this information to figure out what relates to a successful city and
- his team merges this map with other maps and economic data for city planning.

If I had more time, I would attempt to clean and figure out a convention for street type not found in the expected street suffix list. Some of the unexpected street suffix are: 'Camilla', 'Cima', 'Clemente', 'Deseo', 'Middlefork', 'Nest', 'Redondo' and 'Rio'. I suspect many of these names are apartment complex names and subdivision names while a few others are potentially Spanish street suffix such as "Rio" (River) and "Lago" (Lake). The Spanish suffix street names are expected given that Texas borders Mexico and Texas' historical Spain and Mexico connections

Although I didn't find the dataset to be useful for house hunting, I see great potentials in this dataset utilization. For instance:

- overlaying this data with house prices and their trend, an analyst may look into factors related to house price trends;
- the same can be done on economic development to identify what factor drives stimulate and maintain economic growth; and
- the city government may predict future needs in the city and build the appropriate public facilities.

Reference

28 Jupyter Notebook tips, tricks, and shortcuts: https://www.dataquest.io/blog/jupyter-notebook-tips-tricks-shortcuts/

Andrew's LinkedIn Profile: https://www.linkedin.com/in/andrewmatheny

Andrew's OSM Profile: http://www.openstreetmap.org/user/Andrew%20Matheny

Commiting and Pushing a Jupyter Notebook on github: https://stackoverflow.com/questions/48003022/jupyter-notebook-and-github

CSV File Reading and Writing: https://docs.python.org/2/library/csv.html

How do I check in SQLite whether a table exists?: https://stackoverflow.com/questions/1601151/how-do-i-check-in-sqlite-whether-a-table-exists?rq=1

How to check file size in python?: http://stackoverflow.com/questions/2104080/how-to-check-file-size-in-python
How to Include image or picture in jupyter notebook: https://stackoverflow.com/questions/32370281/how-to-include-image-or-picture-in-jupyter-notebook
Installing a pip package from within a Jupyter Notebook not working: https://stackoverflow.com/questions/38368318/installing-a-pip-package-from-within-a-jupyter-notebook-not-workinghttps://stackoverflow.com/questions/38368318/installing-a-pip-package-from-within-a-jupyter-notebook-not-working
Installing Packages: https://packaging.python.org/tutorials/installing-packages/
Keyboard shortcut to paste clipboard content into command prompt window (Win XP): https://stackoverflow.com/questions/131955/keyboard-shortcut-to-paste-clipboard-content-into-command-prompt-window-win-xp
OSM XML: https://wiki.openstreetmap.org/wiki/OSM_XML
Printing a properly formatted SQLite table in Python: https://stackoverflow.com/questions/37051516/printing-a-properly-formatted-sqlite-table-in-python
Problem with SCHEMA: https://discussions.udacity.com/t/problem-with-schema/319234/4
Python BeautifulSoup give multiple tags to findAll: https://stackoverflow.com/questions/20648660/python-beautifulsoup-give-multiple-tags-to-findall
Rubric: https://review.udacity.com/#!/rubrics/25/view
Shortcuts: https://shortcutworld.com/Jupyter-Notebook/win/Jupyter-Notebook_Shortcuts
Street suffix: https://en.wikipedia.org/wiki/Street_suffix
To parse XML: https://pythonprogramming.net/tables-xml-scraping-parsing-beautiful-soup-tutorial/
Updating postal code: https://discussions.udacity.com/t/updating-postal-code/245757/8
UTF-8: Inserting newlines in xml file generated via xml.etree.ElementTree in python: http://stackoverflow.com/questions/3095434/inserting-newlines-in-xml-file-generated-via-xml-etree-elementtree-in-python