• Author: Pragya Jha

# OpenStreetMap Data Case Study

# Map Area

The area of map is the Austin TX area.

The data file is downloaded from mapzen.

The link to the file is <https://drive.google.com/drive/u/2/folders/1ZQEZfbolcAhrkPDCFmZ8scc5mfx-QYdM>

The date I downloaded the file is Dec 22, 2017

The reason I choose this area of map is that I live right now in Texas and it is my favorite place in this state.



The data file is in Openstreet XML format.

([http://wiki.openstreetmap.org/wiki/OSM\_XML#Example\_OSM\_XML\_file](http://wiki.openstreetmap.org/wiki/OSM_XML%23Example_OSM_XML_file))

# Problems Encountered in the Map

I created a small sample size of the Austin area and ran it against a samle.py file but it did not show all the inconsistencies. So, I had to run the full file and I observed the following issues:

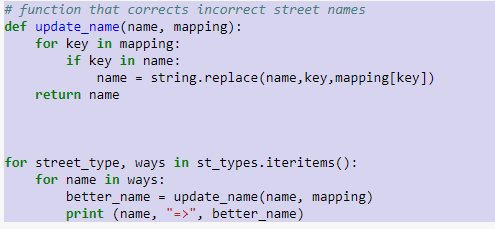
* Over abbreviated street names *(“Heritage Centre Cir”, “Round Rock Ave”)*
* Inconsistent postal codes *(“TX78745-1”, “tx 1”)*
* “Incorrect” postal codes (Austin area zip codes all begin with “78” however only 1 postal code was incorrect, that is, it did not start with 78)
* Street names in second ­level “k” tags pulled from Tiger GPS data and divided into segments, in the following format:

<tag k="tiger:name\_base\_1" v="Rooster Springs" />

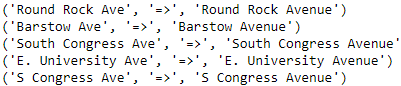
<tag k="tiger:name\_type\_1" v="Rd" />

<tag k="tiger:name\_direction\_prefix\_1" v="N" />

**Over abbreviated Street Names**



So, Ave became Avenue

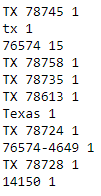


St became Street



**Zip Codes**

The zip code of Austin area should start with *78. Surprisingly, there were very few incorrect codes*



# Data Overview and Additional Ideas

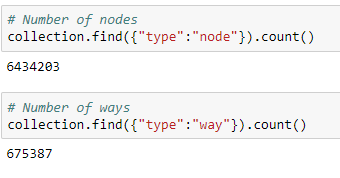
This section contains basic statistics about the dataset, the MongoDB queries used to gather them, and some additional ideas about the data in context.

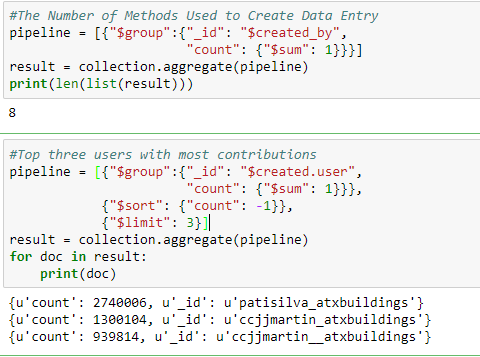
## Data file size

* austin\_texas.osm(The original downloaded OpenStreetMap in xml format): 1391 MB
* sample.osm (The sample file in xml format): 28 MB
* austin\_texas.osm.json(The processed OpenStreetMap in json format): 2135 MB

## Summary statistics of the dataset

* Number of documents: 7109607
* Number of unique users: 1427





# Other ideas about the datasets

I looked at users contributed to Austin OpenStreetMap most frequently. The top three users are:

Which basically are just two users wil slight variation in the names for the last two users.

{u'count': 2740006, u'\_id': u'patisilva\_atxbuildings'

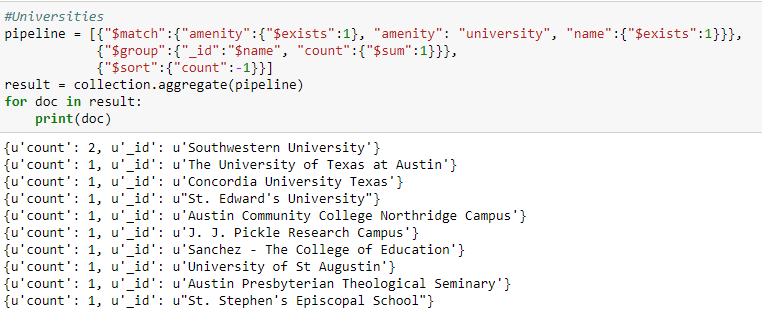
{u'count': 1300104, u'\_id': u'ccjjmartin\_atxbuildings'

{u'count': 939814, u'\_id': u'ccjjmartin\_\_atxbuildings'

I also looked at the top cuisines:



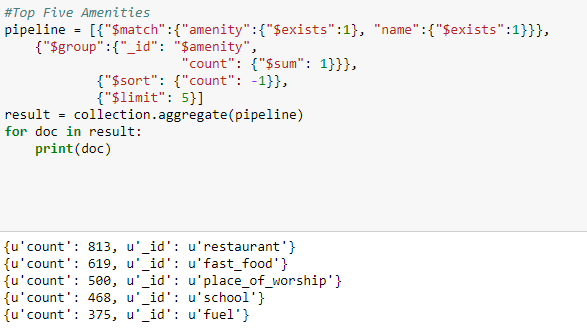
Austin is famous for the many universities in the city.

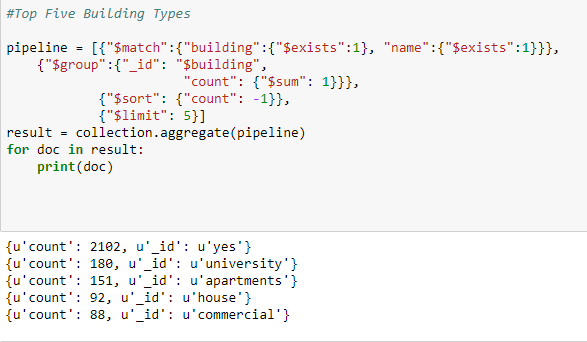


The most popular religion is Christianity

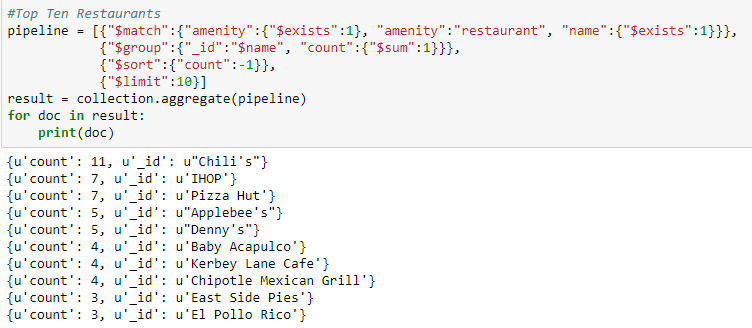


Top Five Amenities

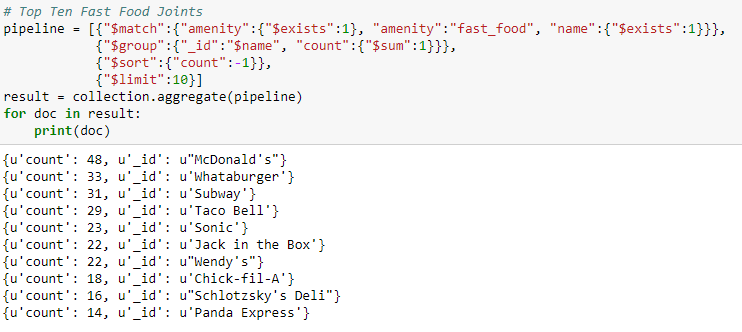




This Top ten is based on the number of branches of restaurants. The more branches it has, the more popular it is to be considered in this context.



Similarly, top ten fast-food places:



**Ideas to improve data quality of OSM**

Reviewing the data, the data seemed fairly clean and mainly the changes done in the audit was to deal with inconsistencies between contributors having different shorthand for street labels (street, drive, etc) as can be seen as Street names.

Most of the zipcodes were correct so there was not much to do in that context.

There needs to be a system in place to allow contributors to be benefitted for contributing entries to OpenStreetMap so there isn't that large of a skew in contributors. As in the case of Austin, the second and third contributors appear to be same.

**Potential cost of the implementation**

The majority of the recommendations (mainly for street names) can be done relatively easily as it would require more strict wording guidelines at time of submitting an entry and a script can be run to correct existing problems with street names.

When looking at the religion there was None category that does show up and this needs to be addressed in two ways: One not allowing a None category on submission (No blank entry) and secondly scanning alternative parts of the entry (scanning key words) to try to determine what this entry could be in terms of religion.