OpenStreetMap Project for Data Wrangling with MongoDB

Roger Jennings

Udacity Data Analyst Nanodegree

Open Street Map Area

Data Overview[1]

The area chosen for this project is Austin, TX. The OpenStreetMap link is: https://www.openstreetmap.org/relation/113314. I chose this area because it is my home town.

File sizes:

OSM File: austin_texas.osm: 175,250KB

JSON File: austin_texas.osm.json: 259,927KB

Problems Encountered with the Map Data

An audit showed there were several types of problems with data. First, there are many variations of street names. The main ones were updated in the mapping list to account for them. Here are the updated expected names after auditing data for most common ones:

expected = ["Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane", "Road", "Trail", "Parkway", "Commons", "Crossing", "Highway", "Expressway", "Way", "Pass", "Speedway", "IH-35", "Loop", "Circle", "Cove", "620", "IH-35", "183"]

620, 183, and IH-35 were included because they are major highways with many addresses on these. This could cause some errors as there could be valid numbers at the end that are not related to highway numbers.

Here is the mapping that was used to update the addresses of other names before including in the database.

```
mapping = { "St": "Street",
        "St.": "Street",
        "street": "Street",
        "Ave": "Avenue",
        "Ave.": "Avenue",
        "Avene": "Avenue",
        "Rd.": "Road",
        "Rd": "Road",
        "RD": "Road",
        "Blvd": "Boulevard",
        "Blvd.": "Boulevard",
        "Cir": "Circle",
        "Ct": "Court",
        "Cv": "Cove",
        "Dr": "Drive",
        "Dr.": "Drive",
        "Expy": "Expressway",
        "Expwy": "Expressway",
```

```
"Hwy": "Highway",
"I35": "IH-35",
"IH35": "IH-35",
"Ln": "Lane",
"Pkwy": "Parkway",
"lane": "Lane"
}
```

Also, directions (W, West, N, North, etc.) were many times appended to the end of addresses. This requires looking before the designation to get street identity.

Also, suite numbers at the address location (123 Main St. Suite 400) were found to be quite common. Another variation is No. at the end of the street name.

Example MongoDB Queries

Sample Street Addresses

This query displayed a sample of street addresses in the database.

```
pipeline = [{'$match': {'address.street':{'$exists':1}}}, {'$limit' : 5}]

sample_addresses  = db.austin.aggregate(pipeline)['result']

pprint.pprint(sample_addresses)
```

```
[{u'_id': ObjectId('552996448726e33844f210ef'),
 u'address': {u'city': u'Austin',
              u'housenumber': u'2911',
              u'postcode': u'78705',
              u'state': u'TX',
              u'street': u'San Jacinto Boulevard'},
 u'amenity': u'pub',
 u'created': {u'changeset': u'11808243',
              u'timestamp': u'2012-06-05T17:27:27Z',
              u'uid': u'190869',
              u'user': u'elbatrop',
              u'version': u'2'},
 u'id': u'280232008',
 u'name': u'Crown and Anchor',
 u'phone': u'+1 512 322-9168',
 u'pos': [30.2925357, -97.735633],
 u'type': u'node',
 u'website': u'http://crownandanchorpub.com'},
{u'_id': ObjectId('552996448726e33844f21196'),
 u'address': {u'housenumber': u'1607',
              u'postcode': u'78701',
              u'street': u'San Jacinto Boulevard'},
 u'amenity': u'pub',
 u'created': {u'changeset': u'12282918',
              u'timestamp': u'2012-07-18T13:14:09Z',
              u'uid': u'722137',
              u'user': u'OSMF Redaction Account',
              u'version': u'5'},
 u'id': u'281362888',
 u'name': u"Schultz's Beer Garden",
 u'pos': [30.2778151, -97.7363223],
 u'type': u'node',
 u'wheelchair': u'limited'},
```

```
{u'_id': ObjectId('552996458726e33844f2d675'),
 u'address': {u'housename': u'4401',
              u'postcode': u'78664',
              u'street': u'North IH 35'},
 u'created': {u'changeset': u'22071440',
              u'timestamp': u'2014-05-01T19:48:28Z',
              u'uid': u'1554107',
              u'user': u'jgpacker',
              u'version': u'3'},
 u'id': u'345201951',
 u'name': u'Round Rock Premium Outlet',
 u'pos': [30.5664621, -97.6901121],
 u'shop': u'clothes',
 u'type': u'node',
 u'website': u'premiumoutlets.com'},
{u'_id': ObjectId('552996458726e33844f2fddf'),
 u'address': {u'city': u'Austin',
              u'housenumber': u'161',
              u'postcode': u'78748',
              u'state': u'TX',
              u'street': u'West Slaughter Lane'},
 u'amenity': u'fast_food',
 u'created': {u'changeset': u'28264414',
              u'timestamp': u'2015-01-19T21:53:17Z',
              u'uid': u'703517',
              u'user': u'Iowa Kid',
              u'version': u'7'},
 u'id': u'354680878',
 u'name': u'Chik-Fil-A',
 u'pos': [30.1670069, -97.7926409],
 u'type': u'node'},
{u'_id': ObjectId('552996458726e33844f2fdf4'),
 u'address': {u'city': u'Austin',
              u'housenumber': u'9300',
              u'postcode': u'78748',
              u'street': u'South I-35 Service SB'},
 u'amenity': u'fast_food',
 u'created': {u'changeset': u'20649142',
              u'timestamp': u'2014-02-19T03:32:04Z',
              u'uid': u'703517',
              u'user': u'Iowa Kid',
              u'version': u'5'},
 u'cuisine': u'burger',
 u'id': u'354684318',
 u'name': u'Whataburger',
 u'pos': [30.1661438, -97.7879823],
 u'type': u'node'}]
```


Amenities

The following amenity query gave a wide variety of amenities. The results included car wash, pub, atm, post office, fire station, and police.

```
amenities = db.austin.aggregate([{'$match': {'amenity':{'$exists':1}}}, {'$limit' : 50}])['result']

pprint.pprint(amenities)
```


```
[{u'_id': ObjectId('552996428726e33844f0dbfd'),
 u'amenity': u'car_wash',
 u'created': {u'changeset': u'28338845',
              u'timestamp': u'2015-01-22T22:59:51Z',
              u'uid': u'1132286',
              u'user': u'Cam4rd98',
```

```
        u'version': u'3'},
 u'id': u'152713302',
 u'pos': [30.383292, -97.955601],
 u'type': u'node'},
{u'_id': ObjectId('552996448726e33844f210ef'),
 u'address': {u'city': u'Austin',
        u'housenumber': u'2911',
        u'postcode': u'78705',
        u'state': u'TX',
        u'street': u'San Jacinto Boulevard'},
 u'amenity': u'pub',
 u'created': {u'changeset': u'11808243',
        u'timestamp': u'2012-06-05T17:27:27Z',
        u'uid': u'190869',
        u'user': u'elbatrop',
        u'version': u'2'},
 u'id': u'280232008',
 u'name': u'Crown and Anchor',
 u'phone': u'+1 512 322-9168',
 u'pos': [30.2925357, -97.735633],
 u'type': u'node',
 u'website': u'http://crownandanchorpub.com'},
{u'_id': ObjectId('552996448726e33844f210f0'),
 u'amenity': u'atm',
 u'created': {u'changeset': u'11511739',
        u'timestamp': u'2012-05-05T20:58:24Z',
        u'uid': u'290680',
        u'user': u'wheelmap_visitor',
        u'version': u'2'},
 u'created_by': u'Potlatch 0.10',
 u'id': u'280232499',
 u'name': u'University Federal Credit Union',
 u'pos': [30.2915028, -97.734911],
 u'type': u'node',
 u'wheelchair': u'yes'},
{u'_id': ObjectId('552996448726e33844f21196'),
 u'address': {u'housenumber': u'1607',
        u'postcode': u'78701',
        u'street': u'San Jacinto Boulevard'},
 u'amenity': u'pub',
 u'created': {u'changeset': u'12282918',
        u'timestamp': u'2012-07-18T13:14:09Z',
        u'uid': u'722137',
        u'user': u'OSMF Redaction Account',
        u'version': u'5'},
 u'id': u'281362888',
 u'name': u"Schultz's Beer Garden",
 u'pos': [30.2778151, -97.7363223],
 u'type': u'node',
 u'wheelchair': u'limited'},
{u'_id': ObjectId('552996448726e33844f23655'),
 u'amenity': u'police',
 u'created': {u'changeset': u'10143546',
        u'timestamp': u'2011-12-18T03:22:14Z',
        u'uid': u'77990',
        u'user': u'varmint',
        u'version': u'3'},
 u'id': u'313248184',
 u'pos': [30.569543, -97.8469475],
 u'type': u'node'},
{u'_id': ObjectId('552996448726e33844f25038'),
 u'amenity': u'post_office',
 u'created': {u'changeset': u'22851483',
        u'timestamp': u'2014-06-10T14:22:28Z',
        u'uid': u'703517',
        u'user': u'Iowa Kid',
```

```
        u'version': u'5'},
 u'id': u'315295676',
 u'name': u'Post Office',
 u'pos': [30.141048, -97.8275752],
 u'type': u'node'},
{u'_id': ObjectId('552996448726e33844f251f7'),
 u'amenity': u'fire_station',
 u'created': {u'changeset': u'3360135',
        u'timestamp': u'2009-12-13T02:32:21Z',
        u'uid': u'77990',
        u'user': u'varmint',
        u'version': u'3'},
 u'id': u'316161859',
 u'name': u'Liberty Hill Fire Department',
 u'pos': [30.6653739, -97.9107804],
 u'type': u'node'},
```

Top Users

This query displayed the top 10 users in the database.

```
pipeline = [{'$match': {'created.user':{'$exists':1}}}, {'$group': {'_id':'$created.user', 'count':{'$sum':1}}},
{'$sort': {'count':-1}}, {'$limit' : 10}]

result  = db.austin.aggregate(pipeline)['result']

pprint.pprint(result)
```

```
[{u'_id': u'woodpeck_fixbot', u'count': 3143997},
 {u'_id': u'varmint', u'count': 501284},
 {u'_id': u'richlv', u'count': 484288},
 {u'_id': u'Iowa Kid', u'count': 453700},
 {u'_id': u'Clorox', u'count': 440944},
 {u'_id': u'HJD', u'count': 376942},
 {u'_id': u'Cam4rd98', u'count': 356970},
 {u'_id': u'afdreher', u'count': 329944},
 {u'_id': u'Chris Lawrence', u'count': 242982},
 {u'_id': u'TexasNHD', u'count': 234943}]
```

Postcode Count

This query displayed the number of postal codes for each one in the area.

```
pipeline = [{"$match":{"address.postcode":{"$exists":1}}},
{"$group":{"_id":"$address.postcode","count":{"$sum":1}}}, {"$sort":{"count":-1}}]

postcode = db.austin.aggregate(pipeline)['result']

pprint.pprint(postcode)
```

```
[{u'_id': u'78704', u'count': 776},
 {u'_id': u'78705', u'count': 662},
 {u'_id': u'78757', u'count': 598},
 {u'_id': u'78759', u'count': 572},
 {u'_id': u'78681', u'count': 557},
 {u'_id': u'78640', u'count': 530},
 {u'_id': u'78702', u'count': 480},
 {u'_id': u'78746', u'count': 469},
 {u'_id': u'78701', u'count': 465},
 {u'_id': u'78745', u'count': 398},
 {u'_id': u'78751', u'count': 352},
```

{u'_id': u'78664', u'count': 324},
{u'_id': u'78750', u'count': 312},
{u'_id': u'78758', u'count': 299},
{u'_id': u'78613', u'count': 286},
{u'_id': u'78703', u'count': 286},
{u'_id': u'78741', u'count': 266},
{u'_id': u'78712', u'count': 264},
{u'_id': u'78748', u'count': 262},
{u'_id': u'78723', u'count': 238},
{u'_id': u'78753', u'count': 237},
{u'_id': u'78731', u'count': 225},
{u'_id': u'78735', u'count': 222},
{u'_id': u'78660', u'count': 213},
{u'_id': u'78749', u'count': 209},
{u'_id': u'78610', u'count': 196},
{u'_id': u'78620', u'count': 185},
{u'_id': u'78722', u'count': 177},
{u'_id': u'78744', u'count': 171},
{u'_id': u'78737', u'count': 153},
{u'_id': u'78727', u'count': 150},
{u'_id': u'78752', u'count': 136},
{u'_id': u'78756', u'count': 135},
{u'_id': u'78738', u'count': 125},
{u'_id': u'78717', u'count': 120},
{u'_id': u'78729', u'count': 114},
{u'_id': u'78621', u'count': 111},
{u'_id': u'78726', u'count': 100},
{u'_id': u'78734', u'count': 100},
{u'_id': u'76574', u'count': 99},
{u'_id': u'78602', u'count': 98},
{u'_id': u'78665', u'count': 88},
{u'_id': u'78641', u'count': 74},
{u'_id': u'78652', u'count': 74},
{u'_id': u'78617', u'count': 74},
{u'_id': u'78721', u'count': 74},
{u'_id': u'78669', u'count': 65},
{u'_id': u'78732', u'count': 64},
{u'_id': u'78728', u'count': 63},
{u'_id': u'78739', u'count': 61},
{u'_id': u'78724-1199', u'count': 60},
{u'_id': u'78645', u'count': 51},
{u'_id': u'78626', u'count': 50},
{u'_id': u'78724', u'count': 49},
{u'_id': u'78628', u'count': 37},
{u'_id': u'78705-5609', u'count': 26},
{u'_id': u'TX 78758', u'count': 26},
{u'_id': u'78736', u'count': 25},
{u'_id': u'78640-6137', u'count': 25},
{u'_id': u'78733', u'count': 25},
{u'_id': u'78612', u'count': 25},
{u'_id': u'TX 78724', u'count': 13},
{u'_id': u'78680', u'count': 13},
{u'_id': u'TX 78745', u'count': 13},
{u'_id': u'78640-4520', u'count': 13},
{u'_id': u'78656', u'count': 13},
{u'_id': u'TX 78613', u'count': 13},
{u'_id': u'76574-4649', u'count': 13},
{u'_id': u'78646', u'count': 13},
{u'_id': u'14150', u'count': 13},
{u'_id': u'78691', u'count': 13},
{u'_id': u'78728-1275', u'count': 13},
{u'_id': u'78747', u'count': 13},
{u'_id': u'78753-4150', u'count': 13},
{u'_id': u'78758-7013', u'count': 13},
{u'_id': u'78704-5639', u'count': 13},
{u'_id': u'78676', u'count': 13},

{u'_id': u'78957', u'count': 13},
{u'_id': u'78754', u'count': 13},
{u'_id': u'78758-7008', u'count': 13},
{u'_id': u'78704-7205', u'count': 13},
{u'_id': u'TX 78728', u'count': 13},
{u'_id': u'78730', u'count': 13},
{u'_id': u'78666', u'count': 12},
{u'_id': u'Texas', u'count': 12},
{u'_id': u'78634', u'count': 12},
{u'_id': u'78619', u'count': 12},
{u'_id': u'TX 78759-3504', u'count': 12},
{u'_id': u'78682', u'count': 12},
{u'_id': u'78653', u'count': 12}]

Summary

There are still many more opportunities to clean and update the data than undertaken in this project. However, the data included in the Open Street Map database for the Austin area is quite sufficient for the objective. It does seem that Open Street Map could be made much more robust programmatically.