Udacity
Data Analyst nanodegree program
Project 2- Wrangling Open Street Map Data

Jamey McCabe – jm6819@att.com

Project completed: June 15, 2015

The project goal was to download a large data set of Geographic data from the opensource OpenStreetMap (OSM) facility and use techniques to understand, find issues and cleanup the data found. The purpose of doing this is;

- to practice new techniques and tools, notably the Xml ElementTree python library
- to contribute to a 'large data' project (gain experience as well as example work for a portfolio)
- to demonstrate skills with data analysis in order to give the instructors information about my skill level

The activities conducted during this project were the following:

1. Browsed and learned about the openStreet maps project largely using the OpenstreetMaps Wiki.
2. Researched, downloaded and used 2 Android apps (OSM Tracker and keypadMapper) who's purpose running on your mobile device is to capture Ways and Nodes as you move about .
3. Played with various techniques of downloading the currently available data in OSM for the Santa Fe NM metropolitan area.
4. Coded routines to investigate and summarize the name data of all Ways in the Santa fe metropolitan area, concentrating on handling a convention known as street-pretype
5. Submitted fixed data back into OSM
6. Analyzed the Santa Fe data in comparison to other cities:
    - Los Alamos, NM a nearby bu relatively recent city
    - Oak Park, IL a midwestern, non Spanish history city but with similar population
    - Los Angeles, CA – A significantly large dataset and also Spanish history city

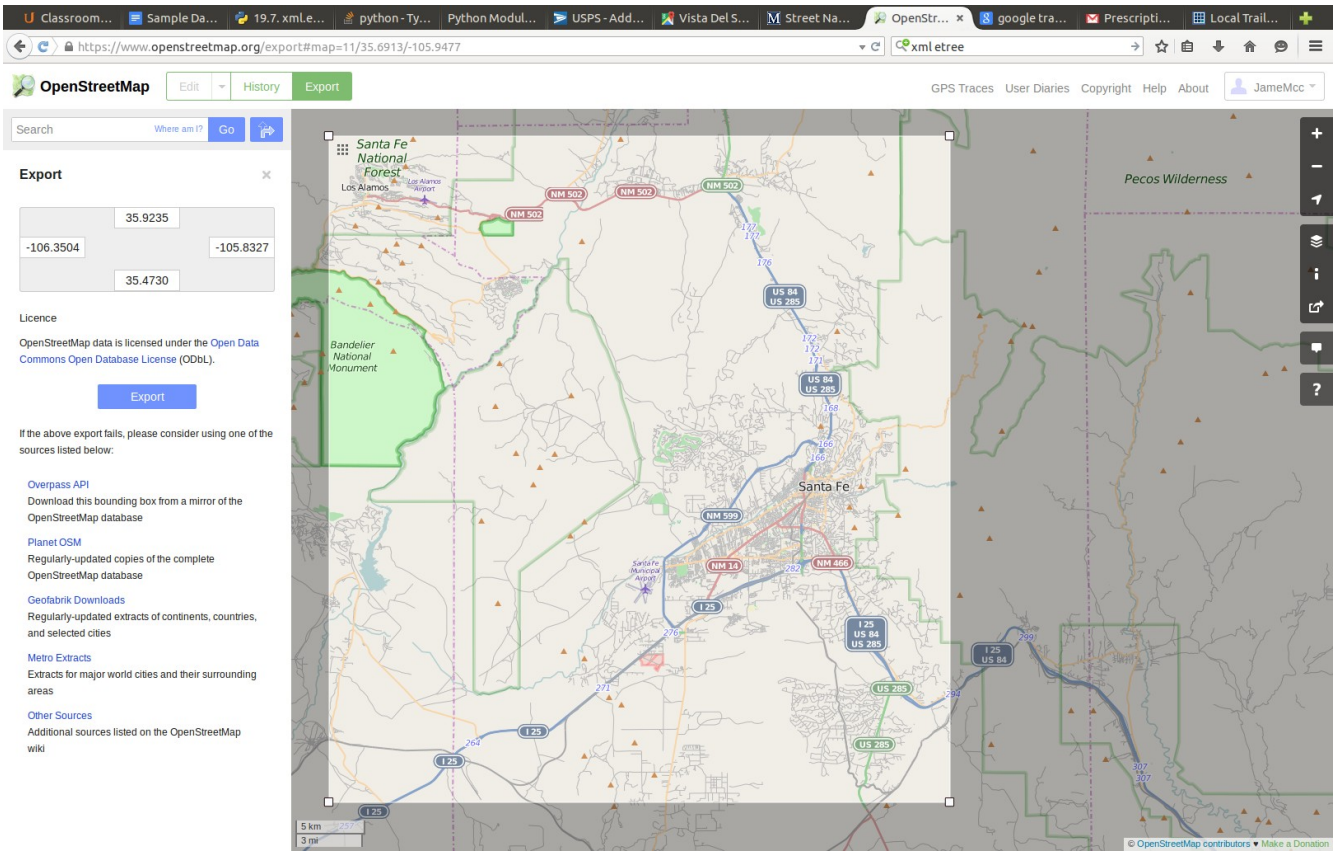
Table of Contents

Problems encountered in the map

1. Data size:

It was not possible to download the initial geographical desired dataset for Santa Fe and the Los Angeles data set was too large to easily processes it in memory. This was not really an issue with specific data, instead an issue with this sort of activity.

Santa Fe download issue: The OSM data is extensive and likely as a result the OSM website and web enabled data extract routines do not allow a custom created download of the entire Santa Fe area. My goal was to outline and download an area that was a little bit larger than the Santa Fe area and would also include Los Alamos NM where I grew up as well as El Dorado, an area to the South East of Santa Fe where my sister lives. Below see the area that failed on numerous attempts to download due to data size.



A resolution to this issue was to download a pre-created extract of the Santa Fe, NM area from this location: https://mapzen.com/metro-extracts/ . Using this extract was valuable as the pure Santa Fe metro data provided insight and challenge that woul not have been as obvious in an larger area.

Los Angeles OSM Size: The raw OSM file for Los Angeles is 1.1 GB. When using the native Python based XMLETREE ITERPARSE function the whole file is read into memory which on my 8GB of memory UBUNTU system fills up all real memory and about 6GB of swap. The Python code then runs extremely slowly. Substituting the CELEMENTREE library did speed it up from over 8 hours ot about 5 minutes though all 8GB of real memory and swap file was still used. Key finding is to use the CELEMETREE library in Python.

2. street-pretypes and lack of addressed node data

2 issues emerged with running exploratory analysis. The initial exploratory analysis was to summarize the Addresses found in the node data by Street type e.g. Road, Avenue, Place. These 2 issues turned out to be exploratory findings and not actual dirty data but serve to illustate that new data will usually come with significant hurdles even before you can find dirty data and begin to clean it up.

The first issue is that Santa Fe has a significant amount of Spanish language content and Spanish naming convention for streets. From the class provided routine to summarize by street type, a large portion of street names in Santa Fe did not fall into the normal American Street Type categories. See extract below. I've highlighted the data which is not being usefully summarized:

```
{'Anasazi': set(['Camino Anasazi']),
 'Ancha': set(['750 Canada Ancha']),
 'Avenue': set(['Don Gaspar Avenue',
                'Lincoln Avenue',
                'Park Avenue',
                'West Palace Avenue']),
 'Chaco': set(['Placita Chaco']),
 'Chelly': set(['Camino de Chelly']),
 'Court': set(['Office Court']),
 'Hualapai': set(['Camino Hualapai']),
 'Lejo': set(['Camino Lejo']),
 'Madre': set(['Acequia Madre']),
 'Marcos': set(['Vuelta San Marcos']),
 'Oraibi': set(['Camino Oraibi']),
 'Palace': set(['Cliff Palace']),
 'Peralta': set(['Paseo de Peralta']),
 'Road': set(['Canyon Road',
              'Cerrillos Road',
              'Country Club Road',
              'East Barcelona Road',
              'West Barcelona Road',
              'West Cordova Road']),
 'Sol': set(['Paseo del Sol']),
 'Street': set(['Agua Fria Street',
                'Alto Street',
                'Baca Street',
                'East Booth Street',
                'East San Francisco Street',
                'East de Vargas Street',
                'Galisteo Street',
                'Hickox Street',
                'Market Street',
                'North Guadalupe Street',
                'Sandoval Street',
                'West Alameda Street',
                'West San Francisco Street',
                'West Water Street']),
 'Trail': set(['Old Santa Fe Trail'])}
```

The first issue and finding from the above is that Santa Fe is an older city and many of its streets were put in place while it was still a Spanish colony (1608-1850) as well as the recent cultural desire to be Santa Fe style. Because of this, most streets have names in Spanish and many have a Spanish style street-pretype. A street-pretype is when the street type preceeds the Street name as in ***Via*** Brisa (the street I live on) rather than follows it as in Fair Oaks ***Avenue***, (the street I lived on until recently in Oak Park Il). In a paper prepared in 2006 by the FGDC (Federal Geographic Data Comittee) this is referred to as a street-pretype and is a generic issue with geographic data. A an exmple of how this appears in the OSM data see the XML for a wellknown building in Santa Fe on ***Paseo De*** Peralta:

```
2121842-       <node id="357614810" lat="35.6916495" lon="-105.9364852" version
="2" timestamp="2015-04-25T20:22:14Z" changeset="30483946" uid="360392" user="ma
xerickson">
2121999-               <tag k="addr:city" v="Santa Fe"/>
2122035-               <tag k="addr:housenumber" v="463"/>
2122073-               <tag k="addr:postcode" v="87501"/>
2122110:               <tag k="addr:street" v="Paseo de Peralta"/>
2122156-               <tag k="amenity" v="social_centre"/>
2122195-               <tag k="gnis:feature_id" v="935960"/>
2122235-               <tag k="name" v="Scottish Rite Temple"/>
2122278-               <tag k="phone" v="505-982 4414"/>
2122314-               <tag k="website" v="http://www.nmscottishrite.org/"/>
2122370-               <tag k="wikipedia" v="en:Scottish Rite Temple (Santa Fe,
 New Mexico)"/>
2122444-       </node>
```

A second issue is also apparent from the above. There is not much node data with addresses in the Santa Fe area, only 38 streets. In fact the first clue, my own street “Via Brisa” is not in the data set under node data.

Running the class provided routine to count the number of tags in this data set gives the following:

```
{'bounds': 1,
'member': 428,
'nd': 56348,
'node': 47892,
'osm': 1,
'relation': 30,
'tag': 42284,
'way': 6029}
```

The dataset is 11.8 Mb and has 47,892 node tags which would seem to indicate there would be significantly more addresses.

Further investigation revealed the following:

1. it takes Nodes to make up Ways
2. in Santa Fe most streets (and thus the Way data) are not straight, they are curvy
3. it takes more nodes (points) to make up a curvy street. It takes 2 nodes to make a straight way – a begin and end point but it takes a node for every section of a curvey street.

4. there are 6,029 ways

It was found that in Santa Fe all but 38 of the Node tags are documenting the shape of the ways not the actual addresses on the street. A somewhat non-technology oriented city like Santa Fe has not yet had significant Address-Nodes captured in the OpenStreetMap community.

3. Way Data Issues

In the ~ 6000 Ways in the Santa fe dataset there are approximately ~2400 unique way names. In that data via the technique of summarizing streets by their street type the following issues were discovered. All 57 were submitted as changes to OpenStreetMaps data, giving the author (Jamey McCabe) their first contribution to the OSM community:

General Type of issue	Fix	Count
Acronyms need to be spelled out		
	Change Ave to Avenue	1
	Change Blvd to Boulevard	1
	Change Cam to Camino	2
	Change Cii to Calle	1
	Change Cli to Calle	1
	Change Cll to Calle	1
	Change Ct to Court	9
	Change DR to Drive	1
	Change Dr to Drive	7
	Change Ln to Lane	1
	Change Rd to Road	18
	change St to Street	1
	Change vis to Vista	2
Mispelled		
	CAmino to Camino	1
	Caltamira to Calle Altimira Court	1
	Paso to Paseo	1
	Change Vis to Via	1
	Entrade to Entrada	1
Street is not accurrately named		
	change Cereza to Plaza Rojo	1
	change Ristra to Ristra Plaza	1
	Drop Circle from Via Janna Circle	1
	remove Curcle from Cinco Pintores Curcle	1
Other issues		
	Gwendloyn does not exist – not a	1
	Fix nodes that make up this way	1
	Total Result	**57**

The routine to summarize by street type in the Way data and handle spanish style street-pretypes includes the following additional logic to the class provided routine:

```
# reverse types used to indicate street-pretypes to use as street types
reverseTypes = ["Aquecia","Acequia","Arroyo","Avenida", "Caballo","Calle",
            "Callecita","Calleja","Callejon", "Camino",
            "Caminito", "Campo", "Canada","Casa","Corrida","Corte",
            "Entrade","Estrasa","Estrada",
            "Hacienda","La","Las","Loma","Monte",
            "Parque","Pasaje", "Paseo","Placita","Plaza","Plazuela","Pueblo",
            "Puerto","Ruta",
            "Senda","Sendero","Sentiero", "Sierra", "Tierra",
            "Valle","Vereda","Via","Viale","Viejo","Vis","Vista","Vuelta"]
# twoWord Types are street-pretypes which have 2 words
twoWordTypes = ["County Road","El Camino","State Route"]

def audit_street_type(street_types, street_name):
    nameWords = street_name.split()
    # if any direction is provided as the first word of a street name, remove it
    if nameWords[0] in ["E","East","N","North","S","South","W","West"]:
        del nameWords[0]
    # if any 2nd word is de or del use the first word as the street-type
    if len(nameWords)> 2 and nameWords[1] in ["de","del","a"]:
        street_type = nameWords[0] + ' ' + nameWords[1]
    # if the 1st word is found in the pre-streettype list use it as street type
    elif len(nameWords)> 0 and nameWords[0] in reverseTypes:
        street_type = nameWords[0]
    # if the first 2 words is in the 2 word list use it as the street type
    elif len(nameWords)> 2 and nameWords[0] + ' ' +nameWords[1] in twoWordTypes:
        street_type = nameWords[0] + ' ' + nameWords[1]
    # original code – find the post street type
    else:
        m = street_type_re.search(street_name)
        if m:
            street_type = m.group()
        #if street_type not in expected:
    street_types[street_type].add(street_name)
```

An example of XML that was fixed is:

```
...
9037389-              <nd ref="144023476"/>
9037413:              <tag k="name" v="Cinco Pintores Curcle"/>
9037457-              <tag k="highway" v="residential"/>
9037494-              <tag k="tiger:cfcc" v="A41"/>
9037526-              <tag k="tiger:county" v="Santa Fe, NM"/>
9037569-              <tag k="tiger:reviewed" v="no"/>
9037604-              <tag k="tiger:zip_left" v="87506"/>
9037642:              <tag k="tiger:name_base" v="Cinco Pintores Curcle"/>
9037697-              <tag k="tiger:zip_right" v="87506"/>
9037736-      </way>
```

Overview of the Data

The following is an overview of Santa Fe OSM data. This was the OSM data pulled March 9, 2015.

1. Quality of K tags:

A way of looking at the quality of data which exists in any one geographical area is to look at the type and quantity of K tags for that area. The K tags are the key value pairs for any type of data captured about Nodes and Ways. A common belief is that “Quality is Relative”. In this case, quality is easier to understand when compared across different geographical areas. In that persuit, the following chart compares the Ktags from Santa Fe with that from a nearby but fairly young city, Los Alamos, NM, and to a much more typical Midwest City, Oak Park,IL.

	Tags with Problem Characters	Lower case with colon (likely tags with values)	Lower case letters (simple tags)	Other characters
Santa Fe, NM	1	24675	16192	1416
Los Alamos, NM	0	7151	9131	123
Oak Park, IL	0	6219	6772	2027

Santa Fe does have 1 tag with problem characters a space in the K value “credit union”:

```
11263685-      <way id="320111663" version="1" timestamp="2015-01-02T22:47:02Z"
 changeset="27875561" uid="1579244" user="modza">
11263800-              <nd ref="2653238409"/>
11263825-              <nd ref="3266091496"/>
11263850-              <nd ref="2653238409"/>
11263875-              <tag k="atm" v="yes"/>
11263900-              <tag k="name" v="Del Norte Credit Union"/>
11263945-              <tag k="phone" v="+1 505-988-3628"/>
```

```
11263984-             <tag k="source" v="http://www.dncu.org"/>
11264028-             <tag k="website" v="http://www.dncu.org"/>
11264073-             <tag k="building" v="yes"/>
11264103-             <tag k="addr:city" v="Santa Fe"/>
11264139-             <tag k="addr:state" v="NM"/>
11264170-             <tag k="wheelchair" v="no"/>
11264201-             <tag k="addr:street" v="Cerrillos Road"/>
11264245:             <tag k="credit union" v="bank"/>
11264280-             <tag k="addr:postcode" v="87507"/>
11264317-             <tag k="opening_hours" v="Monday - Thursday 9:00 a.m. to
 5:00 p.m. Friday 9:00 a.m. to 6:00 p.m., Sat. 9:00 am to 1:00 pm"/>
11264444-             <tag k="addr:housenumber" v="3286"/>
11264483-     </way>
```

In the XML above both the k tag value of “credit union” and “atm” are miscoded and would be better coded as k=”amenity” v=”bank” and k=”amenity” v=”atm”. Thee updates rquire an advanced OSM editor and the issue is not signifiant since it is in the Way data not in the Node data for the credit union.

A second observation the small # of tags containing other chacters in Los Alamos NM. Though difficult to be sure , people in Los Alamos are highly educated city these contributors have put more work put into it and it is cleaner. As regarding Santa Fe's relatively large number of Tags with other charcters this may indicate a need for more detailed cleaning.

Via these statistics it would seem Santa Fe is 3 to 4 times larger than either Los Alamos or Oak Park. As a cross check on their 2010 population sizes though, it seems datasets from Santa Fe and datasets from Oak Park should be more similar since their population size is similar.

https://en.wikipedia.org/wiki/Santa_Fe,_New_Mexico	67,947	100%
https://en.wikipedia.org/wiki/Los_Alamos_County,_New_Mexico	17,950	26%
https://en.wikipedia.org/wiki/Oak_Park,_Illinois	51,878	76%

Since our focus is Santa Fe, we might suspect the difference reveals something wrong with the Santa Fe data though, having too much data is not typically an issue. To try to find reasons for this it seems we would need to compare actual K tags. Here's a comparison of the K tags by count for the higher tag counts. In this comparision

1. the yellow highlighted items stand out as being too large or too small in comparision to the other cities.
2. the green highligted Analysis explains the discrepancy between Santa Fe and Oak Park.

	Santa Fe	Los Alamos	Oak Park	Analysis
----- Problem	1	0	0	

Characters in Tags 'credit union'				
----- Seemingly Valid Compound Tags	24675	7151	6219	
'tiger:county'	3286	1004	964	
'tiger:cfcc'	3284	1004	357	Oak park is a inner city suburb with much fewer types of features compared to unique cities of Santa fe and Los Alamos
'tiger:name_base'	2675	497	357	
'tiger:reviewed'	2166	963	296	Oak Park without high tech community help as not had a lot of work Reviewing it's streets against the Tiger (us census) street and highway data.
'tiger:zip_left'	2089	339	351	
'tiger:zip_right'	2001	312	344	
'tiger:name_type'	1590	454	342	
'tiger:source'	1477	676	60	same
'tiger:tlid'	1474	673	57	same
'tiger:separated'	1300	548	52	same
'tiger:upload_uuid'	875	358	57	same
'gnis:feature_id'	328	40	116	Los Alamos has basically 1 important feature, the National Labratory, whereas Santa Fe as a very old and historic city has many.
'gnis:created'	210	37	91	same
'gnis:state_id'	205	35	91	same
'gnis:county_id'	205	35	91	same
'addr:street'	175	22	614	Oak Park has many more houses loaded as nodes. It is a grid based city so much easier to create house addressed nodes. Cant explain low number of addr:street for Los Alamos. Was going on theory it had good community support as a “geek” town.
----- Normal Tags found	16192	9131	6772	
Counter({'highway'	5851	1837	1083	
'name'	3787	823	852	
'oneway'	870	118	116	
'source'	639	16	88	Data in Oak Park created from very few sources whereas Santa fe data from a large #.
'service'	541	151	291	
'building'	465	812	#N/A	Los Alamos data consist of buildings whereas Santa fe has many fewer (especially in comparison to their relative sizes and Oak Park has none.
'lanes'	343	21	94	
'amenity'	336	237	199	
'ele'	327	48	119	

'surface'	302	96	81	
'ref'	270	39	61	
'access'	228	286	13	Los alamos as a scientifuc lab has many areas (likely ways) with special access rules as does a tourist oriented city like Santa Fe.
'power'	210	10	2	Santa Fe as the State Capital and a much larger geogrphic area has much more “utility power “ ways and nodes.
'maxspeed'	180	1	28	Not sure why no one has put work into Max speeds for street in Los Alamos like they have for Santa Fe.
'bicycle'	126	27	10	Not sure why no one has put work into Bicycle paths for street in Oak Park like they have for Santa Fe.
----- Other (strange) tags found are:1416	1416	123	2027	Many fewer non alphabetic containing tags in Los Alamos. Unclear why.

The finding is that inner cities will have a significantly lower density of tags than standalone and high tech cities but as expected there is no obvious flaw in the Santa Fe dataset.

The custom code to find and count the Ktag values as shown in the above table is:

```
"""
Counting the number of each type of K tag within a quality classification for th e K tag values
"""
lower = re.compile(r'^([a-z]|_)*$')
lower_colon = re.compile(r'^([a-z]|_)*:([a-z]|_)*$')
problemchars = re.compile(r'[=\+/&<>;\'"\?%#$@\,\. \t\r\n]')
# setup counters objects for each quality type
lower_colonKcount= Counter()
lowerKcount = Counter()
problemKcount = Counter()
otherKcount = Counter()

def key_type(element, keys):
    if element.tag == "tag":
        # Look for k tags and if found look at it's value
        if 'k' in element.attrib:
            # using regular expression for problem characters search the keys
             if problemchars.search( element.attrib['k']) is not None:
                keys['problemchars'] +=1;
                # document and count the # of these keys
                problemKcount[element.attrib['k']] += 1;
            elif lower_colon.match( element.attrib['k']) is not None:
                keys['lower_colon'] +=1
                # document and count the # of these keys
                lower_colonKcount[element.attrib['k']] += 1;
            elif lower.match( element.attrib['k']) is not None:
                keys['lower'] +=1;
                # document and count the # of these keys
```

```
                lowerKcount[element.attrib['k']] += 1;
            else:
                keys['other'] +=1
                #otherKcount[element.attrib['k']] += 1;
    return keys
def process_map(filename):
    keys = {"lower": 0, "lower_colon": 0, "problemchars": 0, "other": 0}
    for _, element in ET.iterparse(filename):
        keys = key_type(element, keys)

def test(filename):
    keys = process_map(filename)
    print "===== {} ====================".format(filename)
    print "----- Problem Characters in Tags found are:{}".format(keys['problemch ars'])
    pprint.pprint(problemKcount)
    print "----- Seemingly Valid Compound Tags found are:{}".format(keys['lower_ colon'])
    pprint.pprint(lower_colonKcount)
    print "----- Normal Tags found are:{}".format(keys['lower'])
    pprint.pprint(lowerKcount)
    print "----- Other (strange) tags found are:{}".format(keys['other'])
    pprint.pprint(otherKcount)

if __name__ == "__main__":
    test("santafe.osm")
    lowerKcount = Counter() ;lower_colonKcount= Counter();problemKcount = Counter() ;otherKcount =
Counter()
    test("LosAlamos.osm")
    lowerKcount = Counter()
    lower_colonKcount= Counter()
    problemKcount = Counter()
    otherKcount = Counter()
    test("OakPark.osm")
```

2. Types of Tags:

As with the K tag analysis it is useful to compare Santa Fe types of tags to other cities as shown below

	Santa Fe	Los Alamos	Oak Park
bounds':	1	1	1
'member':	428	535	4301
'nd':	56348	39350	19030
'node':	47892	35480	15937
'osm':	1	1	1
'relation':	30	19	167
'tag':	42284	15405	15018
'way':	6029	2761	2287
Total	153013	93552	56742
Nodes/way	8	13	7

This analysis reveals:

1. As shown in green – Santa Fe's data has more node's and ways, especially as compared to Oak park Il. As Santa Fe covers much more land area than Oak Park it has more streets and room for other defined areas (ways). In a second check of Wikipedia, Santa Fe cover 37.4 square miles whereas Oak Park only 4.7 square miles. As Santa Fe is not laid out in a grid and Oak Park is it needs many more nodes per way to describe it's ways.

2. Los Alamos though ¼ the population of Santa Fe does have almost as many Nodes. This indicates that Los Alamos has some special properties but does not seem to be a special reflection on Santa Fe's data. Though area explained the difference between Santa fe and Oak Park – it does not explain why Los Alamos with an area of 109 square Miles, almost 3 times as large as Santa Fe has 30% less Nodes and Ways.

3. Contributors:

As OpenStreetMaps data is community sourced and maintained, an interesting element of the data is who it came from. Looking at the contributors to the Santa Fe NM data as compared to who contributed for Los Alamos NM and Oak Park IL we see:

Found 141 user in santafe.osm	# tags	Found 82 user in LosAlamos.osm	# tags	Found 80 user in OakPark.osm	# tags
None	99062	None	56292	None	38351
cluening'	6318	Dschwen'	13223	chicago-buildings'	8091
Wolfram Sobotta'	6307	woodpeck_fixbot'	8068	NE2'	1541
Timothy Smith'	5234	cbkiyanda'	4355	korky99_04'	1397
kriscarle'	4500	cluening'	4132	daniel_erik'	1126
Zephrys'	4099	techlady'	1474	Umbugbene'	1020
woodpeck_fixbot'	3544	TIGERcnl'	1456	bbmiller'	812
JDub'	2746	Data411'	1132	SednaBoo'	562
Latze'	2614	monemmer'	919	ediyes'	456
anjbe'	1895	Seldom'	619	Chris Lawrence'	440
n76'	1038	AndyAyre'	478	Zol87'	393
Thomas8122'	856	jcsom'	453	RichRico'	308
PHerison'	836	Other	1950	Other	1704
t_u_b_o'	652				
triplemultiplex'	639				
TIGERcnl'	620				
vonvonvon'	609				
NE2'	580				
the Sandinator'	538				
DaveHansenTiger'	518				
jonesydesign'	505				
Dschwen'	500				
maxerickson'	409				
SamatJain'	408				
42429'	314				
Other	7670				

Analysis of this data shows:

1. Santa Fe has significantly better contribution given the following:

- 141 contributors total – almost twice as many as in Oak Park. That said – Los Alamos with ¼ the population does have a more dense contributor base, reflecting it's “geek” heritage.
- 8 users contributed more than 2000 tags each in Santa Fe versus none in Oak Park

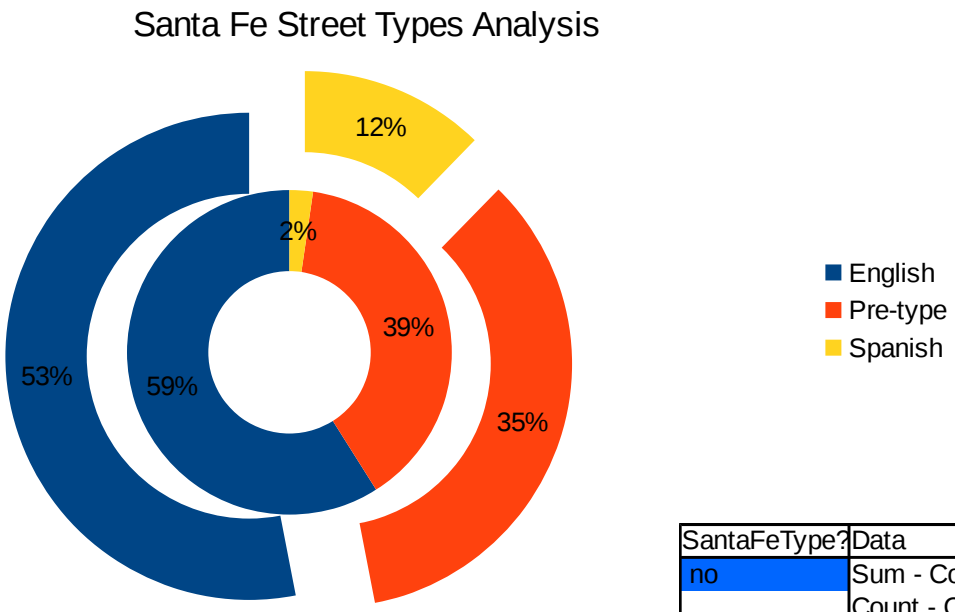
assuming that user 'chicago-buildings' is more of a automated submission and not a real contributor. Again though, with 4 entities in Los Alamos contributing more than 2000 elements Loas Alamos is getting more attention than Santa Fe.

2. Los Alamos has 2 contributors with more than 8000 tags apiece which is higher than either Santa Fe or Oak Park (again eliminating 'chicago-buildings'). Given the usual goal and "benefit" that open source projects attracts and benefits from more contributors this may not be as successful a statistic for Los Alamos though and instead could be seen as another success for the Santa Fe dataset with more smaller contributors.

4. Way names as cultural indicator

During this cleanup and examination of Santa Fe NM Way data, it became clear that Santa Fe had a statistically signifant amount of streets with Street Pre Type in the Spanish style of street names. A theory is postulated that this might be a technique for analyzing the older culture of cities comparing them to each other based on cultural naming conventions. The subsequent data is an analysis of the amount of Spanish naming convention in Santa Fe and comparing it to other cities.

4.1 Santa Fe Spanish naming density: At the summary level – Santa Fe has 39% to 47% Spanish Naming conventions depending on how you categorize:



SantaFeType?	Data	
no	Sum - Count	1448
	Count - Count	190
yes	Sum - Count	951
	Count - Count	124
SpType	Sum - Count	57
	Count - Count	44
Total Sum of streets - Coun		**2456**
Total Count of street types		**358**

In this pictorial

- The 39% orange section is calculated from summing the #s of streets with a Pre-Street Type which is 941 streets of the 2496 streets in Santa Fe.
- 47% is calculated from adding streets with Spanish Type names (12%-44 street types) to the pre-streetypes (35%-124 street ypes). Note that this statistic is uses the unique types versus the sum of how many streets use that type.
    - A sub-finding is that there are many streets in Santa fe which have neither a spanish style street-pretype or a typical American post street type. In essence there is some other standard for naming streets in Santa Fe that does not seem to involve street types. It's possible this is a issue with the Santa Fe data (dirty data missing street types), but in a informal survey, personally looking around, there are many streets with no street type shown in the street signs.

4.2 Spanish Naming comparison between Santa Fe, NM and Oak Park, IL - As we did before it is useful to compare Santa Fe NM and it's street types to a more typical "english-only" city, Oak Park Illinois which has no "street-pretype".

Santa Fe New Mexico					Oak Park Illinois			
Street Type	**Count**	**SantaFeType?**			**Street Type**	**Count**	**SantaFeType?**	
Road	242	no	9.85%		Avenue	66	no	27.16%
Street	240	no	9.77%		Street	46	no	18.93%
Calle	196	yes	7.98%		Park	19	no	7.82%
Lane	172	no	7.00%		Boulevard	11	no	4.53%
Camino	152	yes	6.19%		House	10	no	4.12%
Drive	118	no	4.80%		Court	6	no	2.47%
Court	109	no	4.44%		School	6	no	2.47%
Circle	90	no	3.66%		Playground	4	no	1.65%
Trail	47	no	1.91%		Center	3	no	1.23%
Place	45	no	1.83%		Church	3	no	1.23%
Avenue	44	no	1.79%		Branch	2	no	0.82%
Plaza	39	yes	1.59%		Cleaners	2	no	0.82%
La	39	yes	1.59%		Club	2	no	0.82%
Via	33	yes	1.34%		Place	2	no	0.82%
Way	33	no	1.34%		Shop	2	no	0.82%
Camino de	28	no	1.14%		Square	2	no	0.82%
Calle de	27	no	1.10%		Subdivision	2	no	0.82%
Avenida	23	yes	0.94%		2	1	no	0.41%

The following are observations about the comparision of the 2 cities street types:

- Santa Fe's 2 top street types "road" and "street" are english like post street types adding to 20% of the total streets. However, Santa Fe has 2 spanish style street-pretype in the top 5 "Calle"

and “Camino” which add up to 15% of streets in Santa Fe.

- The top street type in Oak Park is “Avenue,” and is used in 27% of the steets in Oak Park. “Avenue” is only used in 2% of the streets in Santa Fe. This is a significant difference and highly indicative of a geographic cultural difference.
- The second most used street type in Oak Park is “Street”. It is used in 19% of the streets in Oak Park whereas that same Street Type is used in 10% of the streets in Santa Fe. This is a lot more similar than Blvd and perhaps expected since the word street is such a common word used to describe the superset.
- The 3rd most frequently found “last word” in a way name in Oak Park ironically is “Park”. It is not a street type but an area title referring to the 19 public parks found in Oak Park. Public parks make up 8% of the ways in all of Oak Park's data. In Santa Fe OSM data, “Park” (though not shown in the above since it is so far down the list) occurs only 8 times which is 0.33% of the ways.
- In similar fashion the 4th most frequently occuring in Oak Park is a street type “Boulevard” at 5% or 11 times and occurs in Santa Fe only 3 times which is 0.11%.

4.3 Spanish Pre-Type as a cross city cultural indicator - The following is a little wider comparision of Spanish Pre-types across the mapping data. In order add a large data source as well as provide another dimension we added the Los Angles Way data to the mix. The theory being tested is that the Spanish pre-type can indicate the degre to which Spanish Culture permeates a geographic area. The table below lists the cities (actually counties) comparing the degree to which street pre-types appear in their Way data.

	SpanishInd			
County	FALSE	TRUE	Total Result	Spanish Culural Way %
Cook, IL	349		**349**	0.0%
Los Alamos, NM	488		**488**	0.0%
Los Angeles, CA	64434	2371	**66805**	3.5%
Orange, CA	34800	3828	**38628**	9.9%
Riverside, CA	21608	1995	**23603**	8.5%
San Bernardino, CA	30014	810	**30824**	2.6%
San Diego, CA	1776	191	**1967**	9.7%
Santa Fe, NM	1754	929	**2683**	34.6%
Ventura, CA	13641	994	**14635**	6.8%
Total Result	**168864**	**11118**	**179982**	

The theory seems well supported by the data above. Sanata Fe as the oldest Spain colonized city in North America has a the clearly highest density of Spanish naming. Other parts of California's Los Angeles metropolitan area – notably San Diego and Riverside also show a significantly significant Spanish cultural influence as begats their Spanish Colonial heritage as well.

Other ideas about the datasets

In the exploration of the Santa Fe (and Los Alamos and Oak Park) data the following were observed but not futher studied:

1. High Use of the key Bicycle in Santa Fe when this is not a common tag. In researching where the key Bicycle would be used in OSM it was not documented other than as a sub key value for a limited scenario “Road (UK) or path (USA, Canada) dedicated to cyclists on separate right of way.” See: http://wiki.openstreetmap.org/wiki/Bicycle . Instead of Bicycle= the tag used ot indicate Bicycle paths is Highway=Cycleway. In my knowledge of Santa Fe there are no or at least very few separated Bicycle paths. Further investigation of these and perhaps recoding if they are coded incorrectly is warranted.
2. Why would Los Alamos with the smallest population have lots of buildings in the OSM data and Oak Park has none? Los Alamos had 812 keys for Building and Oak Park had none. Investigation as to Los Alamos and what buildings are captured is warranted. It also seems to indicate room for a valuable contribution to Oak park for their numerous and famous Frank Lloyd Wright buildings.
3. Why does Los Alamos have so many Node's compared to ways? Los Alamos has 13 nodes per way as compared to Santa fe which has 8 nodes per way and Oak Park which has 7 nodes/way. This seems to call for some further investigation to see if this is an issue for Los Alamos or some technique which could be used in Santa Fe and Oak Park.
4. Explore cultural diversity of an area based on Street Types or other attributes of street names beyond street pretype as an indicator of Spanish culture.. From the street-pretype analysis it is clear that street names can be a metric of sorts to reflect age and culture of a city. Applying this across cities and some sort of US heat map would be interesting. Developing the criteria for cultural indicators could be challenging as street pretype was fairly simple but other cultural indicators (e.g. Chinese) would likely be less clear (e.g. perhaps use Chinese names such as Wei or Han) and more difficult to computationally derive (e.g. being sure that Wei was a the chinese name and not a mispelling).
5. Low “other” character count in Los Alamos values. From the comparison of Character quality between Santa Fe NM, Los Alamos NM and Oak Park, IL it was discovered that Los Alamos had a very low count of “other” characters in it's population of key-value pair – values. “other” is when there is a character found which is not “a-z” or in the problem character set “[=\ +/&<>;\'"\?%#$@”. It is usually dificult to answer such a question “why aren't there any?'. The hope though is by categorizing what the “other' characters are in Santa Fe and Oak Park data sets we might discover a practice that had been employed in Los Alamos data that makes it cleaner.