PROBLEMS ENCOUNTERED IN THE MAP

1) Inconsistent Street names. For example, boulevard is spelled in two ways.
Los Gatos Blvd
Stevens Creek Boulevard

2) Multiple format phone numbers
+1 408 736 6859
4087382177
+1-408-957-9215

3) Incorrect postcodes and city names in the file. The data file is of San Jose but there are zipcodes and city data in the file that belongs to other cities around San Jose

<node id="343601217" lat="37.377713" lon="-122.0322128" version="4" timestamp="2013-04-02T16:12:50Z" changeset="15583562" uid="318696" user="n76">
<tag k="name" v="US Post Office"/>
<tag k="amenity" v="post_office"/>
<tag k="addr:city" v="Sunnyvale"/>
<tag k="addr:state" v="CA"/>
<tag k="addr:street" v="South Taaffe Street"/>
<tag k="addr:country" v="US"/>
<tag k="addr:postcode" v="94086"/>
<tag k="addr:housenumber" v="141"/>

4) Multiple formats of San Jose in the data. "San jose","San Jose","San José","san jose"

5) Street names in the “k” tags divided into the following format:

<tag k="tiger:cfcc" v="A15"/>
<tag k="tiger:county" v="Santa Clara, CA"/>
<tag k="tiger:name_base" v="I-680"/>
<tag k="tiger:name_base_1" v="Sinclair"/>
<tag k="tiger:name_type_1" v="Fwy"/>

The first 4 problems have been cleaned programmatically in the project.

OVERVIEW OF THE DATA:

San Jose OSM XML file size (uncompressed) - 380.7 mb
osm.db ..... 242.8 MB
nodes_sanjose.csv ..... 146 MB
nodes_tags_sanjose.csv ..... 2.4 MB
ways.csv_sanjose ..... 13.4 MB
ways_tags_sanjose.csv ..... 20.6 MB
ways_nodes_sanjose.cv ..... 46.2 MB

1. Query to find distinct users

select count(e.user) from (select distinct(user) from nodes_sanjose union

select distinct(user) from ways_sanjose) e

1354

2. Query to find number of ways

select count(id) from ways_sanjose

226858

3. Query to find number of nodes

select count(id) from nodes_sanjose

1747498

4. Top 10 cuisine in San Jose

select value, count(*) from nodes_tags_sanjose where id in (select distinct(id) from nodes_tags_sanjose where key='amenity' and value='restaurant') and key='cuisine' group by value order by count(*) desc limit 10

(vietnamese', 73),
('mexican', 54),
('chinese', 52),
('pizza', 46),
('japanese', 36),
('italian', 26),
('indian', 25),
('american', 23),
('thai', 19),
('sushi', 16)

5. All the car rental companies in San Jose

select distinct(value) from nodes_tags_sanjose where id in (select id from nodes_tags_sanjose where key='amenity' and value='car_rental') and key='name';

('Hertz Car Rental',),
('Hertz Rent A Car',),
('Silicon Valley Auto Rental',),
('Hertz',),
('Enterprise Rent-A-Car',),
('United Rentals',),
('Enterprise Car Rental',),
('Service Rent-A-Car',),
('Advantage',),
('Avis',),
('Budget',),
('Dollar',),
('Thrifty',),
('Fox',),
('Enterprise',),

('National',),
('Alamo',),
('Payless',),
('Rental car return',)

6. Top 10 amenities in San Jose

```
select value , count(*) from nodes_tags_sanjose where key='amenity' group by value
order by count(*) desc limit 10
```

('restaurant', 761),
('fast_food', 397),
('bench', 313),
('cafe', 225),
('bicycle_parking', 202),
('place_of_worship', 167),
('toilets', 160),
('school', 138),
('parking_space', 128),
('bank', 120)

7. Biggest Religion in San Jose

```
select value, count(*) from nodes_tags_sanjose where id in (
select distinct(id) from nodes_tags_sanjose where value="place_of_worship")
and key="religion" group by value order by count(*) desc limit 1
```

('christian', 161)

8. After cleaning the data for postcode , the below query confirms that all the postcodes/ cityname combination in the database are only for San Jose. This is true only for nodes/ways that had both of the below tags
addr:city
addr:postcode

```
select value, id from nodes_tags_sanjose where id in (
select id from nodes_tags_sanjose where key='city' and value not in ('San Jose','San José','san Jose'))
and key='postcode'
```

[]

9. Query to find top 10 contributing users

```
select totalusers.user, count(*) from (select user from nodes_sanjose union all select user from ways_sanjose) totalusers
group by totalusers.user
Order by count(*) desc
Limit 10
```

andygol|295308

```
nmixter|283819
mk408|142708
Bike Mapper|90970
samely|80830
RichRico|75951
dannykath|73855
MustangBuyer|64993
karitotp|62385
Minh Nguyen|52076
```

OTHER IDEAS ABOUT DATASET

To minimize the number of errors in the data, there can be a golden rule document that states what kind of format to be used for street names and phone numbers. So if the document states that Avenue should be formatted as “Avenue” before the user puts it in the data, that may reduce the inconsistency in the street names. Similary, if there is a golden rule formatting document for phone number format for each country, that will also minimize the data inconsistencies.As can be seen by below query, the numb er of users contributing to the data is pretty high. As all of them may have their own format of writing data, the golden rule document can definitely help in reducing the inconsistencies

```
Select count(distinct(user.uid)) from (select uid from nodes_sanjose union all select uid from ways_sanjose) user;
```

Ofcourse implementing the above approach will be expensive and time consuming initially and the golden rule documents will need to be modified frequently to capture all the street names. But over time once majority of the data is captured, this will result in a higher ease of use by the people who want to to work on the OSM data for data wrangling