1. Problems Encountered in the Map

Once I downloaded the sample size of Singapore map. I ran agaist audit.py to check for common errors in the street name. Interesting enough, I found some of them listed below.

- Abbreviated street names
- Mix of street names in english and malay
- Some addresses are not in singapore

Abbreviated street names

Names such as “Bedok North Ave 3” can be further clean up to be become “Bedok North Avenue 3”, or “Orchard Rd” can become “Orchard Road”

Mix of street name in english and malay

Because singapore is in a southeast asia county. Its a mix of chinese, english and malay culture. Some of the street names might consist of malay words such as “Jalan” which stands for Road in english. Also, in malay language the word road is normally being placed at the front. These names needs to be thoroughly check to make sure that they are not abbreviated as well. One such case would be “Jl. Todak'” which can be shown as “Jalan Todak” or “Lor Telok” becomes “Lorong Telok”

Some addresses are not in singapore

Singapore’s postal code should be a 6 digit [2]. While checking on it I discover that some of the data also consist of near by cities such as `Johor Bahru in Malaysia or Batam in Indonesia.`

`By doing a query in db for top cities`

```
> db.sg.aggregate([{"$match":{"address.city":{"$exists":1}}},
{"$group":{"_id":"$address.city", "count":{"$sum":1}}},
{"$sort":{"count": -1}}, {"$limit":10}])

{ "_id" : "Singapore", "count" : 5694 }
{ "_id" : "Johor Bahru", "count" : 70 }
{ "_id" : "SKUDAI", "count" : 35 }
{ "_id" : "Batam", "count" : 13 }
{ "_id" : "Masai", "count" : 13 }
```

```
{ "_id" : "Nusajaya", "count" : 4 }
{ "_id" : "Kulai", "count" : 3 }
```

or by doing a query in db for top countries

```
> db.sg.aggregate([{"$match":{"address.country":{"$exists":1}}},
{"$group":{"_id":"$address.country", "count":{"$sum":1}}},
{"$sort":{"count": -1}}, {"$limit":10}])

{ "_id" : "SG", "count" : 7295 }
{ "_id" : "MY", "count" : 18 }
{ "_id" : "ID", "count" : 8 }
{ "_id" : "Indonesia", "count" : 1 }
```

2. Data Overview

File sizes

```
singapore.osm ......... 183.6 MB
singapore.osm.json .... 202.1 MB
```

Number of documents

```
> db.sg.find().count()

924795
```

Number of nodes

```
> db.sg.find({"type":"node"}).count()

804531
```

Number of ways

```
> db.sg.find({"type":"way"}).count()

120221
```

Number of unique users

```
> db.sg.distinct("created.user").length

954
```

Top 1 contributing user

```
> db.sg.aggregate([{"$group": {"_id": "$created.user", "count":
{"$sum": 1}}}, {"$sort": {"count": -1}}, {"$limit":1}])

{ "_id" : "JaLooNz", "count" : 267916 }
```

No of chosen type of nodes

```
> db.sg.aggregate([{"$match":{"amenity":{"$exists":1}, "type":
"node"}}, {"$group":{"_id":"$amenity", "count":{"$sum":1}}},
{"$group":{"_id":"_id", "total_count":{"$sum":1}}}])

{ "_id" : "_id", "total_count" : 77 }
```

3. Additional Ideas

The overall participation of the user is quite low.

Top user (JaLooNz) only contributed around 28%
Top 5 users contributed around 50%

To increase the user participation, I would suggest an easier way for user to check in a location through a mobile app rather than submitting the information through web. While checking in to that location, user can submit a new location easily. This also increases the accuracy of the data because the mobile app can capture the GPS data along with it.

Additional data exploration using MongoDB queries

Top places for religion worship

```
> db.sg.aggregate([{"$match":{"amenity":{"$exists":1},
"amenity":"place_of_worship", "religion": {"$exists":1}}},
```

```
{"$group":{"_id":"$religion", "count":{"$sum":1}}},
{"$sort":{"count": -1}}])

{ "_id" : "muslim", "count" : 430 }
{ "_id" : "christian", "count" : 166 }
{ "_id" : null, "count" : 90 }
{ "_id" : "buddhist", "count" : 60 }
{ "_id" : "hindu", "count" : 13 }
{ "_id" : "taoist", "count" : 7 }
{ "_id" : "jewish", "count" : 4 }
{ "_id" : "Sikh", "count" : 1 }
```

Top cuisine around town

```
> db.sg.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":
"restaurant", "cuisine":{"$exists":1}}},
{"$group":{"_id":"$cuisine", "count":{"$sum":1}}},
{"$sort":{"count": -1}}, {"$limit":5}])

{ "_id" : "chinese", "count" : 25 }
{ "_id" : "asian", "count" : 15 }
{ "_id" : "seafood", "count" : 13 }
{ "_id" : "japanese", "count" : 13 }
{ "_id" : "italian", "count" : 10 }
```

Conclusion

While working through Singapore data, I find that some of the data which the user contributed is not standardise and some of the data might be out of date. Through this exercise of cleaning and auditing the data using data.py or audit.py. I hope its well cleaned after this process. Also, since a lot of the students in this class is also picking a location to do audit and clean up the data from OpenStreetMap. I think we can contribute back a cleaner data back to OpenStreetMap.

Reference

[1] http://en.wikipedia.org/wiki/List_of_roads_in_Kuala_Lumpur

[2] http://en.wikipedia.org/wiki/Postal_codes_in_Singapore