Problems encountered in my map

・Name inconsistency(Starbucks Coffee,Starbucks)

When I try to count cafe by descending order in mongoDB, I found that there are two different name for Starbucks Coffee.(Result is below).This result is of sample part of the map,so the number is small.

```
>db.sf.aggregate([{"$match":{"cuisine":{"$exists":1},"amenity":"cafe"}},
{"$group":{"_id":"$name","count":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : "Starbucks Coffee", "count" : 4 }
{ "_id" : "Starbucks", "count" : 3 }
{ "_id" : "Espresso Roma", "count" : 1 }
{ "_id" : "Sun Maxim's", "count" : 1 }
...
```

The name inconsistency will result in the diffrent outcome, so I adjust "Starbucks" to "Starbucks Coffee". The result is below.

```
>db.sf_r.aggregate([{"$match":{"cuisine":{"$exists":1},"amenity":"cafe"}},
{"$group":{"_id":"$name","count":{"$sum":1}}},
{"$sort":{"count":-1}},{"$limit":10}] )

{ "_id" : "Starbucks Coffee", "count" : 66 }
{ "_id" : "Peet's Coffee & Tea", "count" : 13 }
{ "_id" : "Peet's Coffee and Tea", "count" : 5 }
{ "_id" : "Philz Coffee", "count" : 5 }
{ "_id" : "Peet's Coffee", "count" : 4 }
{ "_id" : "Quickly", "count" : 3 }
{ "_id" : "Beanery", "count" : 3 }
{ "_id" : "Highwire Coffee Roasters", "count" : 2 }
```

{ "_id" : "Tart to Tart", "count" : 2 }
{ "_id" : "Blue Bottle Coffee", "count" : 2 }

"Starbucks Coffee" and "Starbucks" are now only "Starbucks Coffee" and there's no more "Starbucks".

・Zip code inconsistency

Sorting the post code in descending order,I found that the post code includes Oakland that is not in SF.

>db.sf_r.aggregate([{"$match":{"address.postcode":{"$exists":1}}},{"$group":{"_id":"$address.postcode","count":{"$sum":1}}},{"$sort":{"count":-1}}])

{ "_id" : "94122", "count" : 4580 }
{ "_id" : "94611", "count" : 2982 }
{ "_id" : "94116", "count" : 2022 }
{ "_id" : "94610", "count" : 1349 }
{ "_id" : "94133", "count" : 1056 }
{ "_id" : "94117", "count" : 992 }
{ "_id" : "94127", "count" : 597 }
{ "_id" : "94103", "count" : 506 }
{ "_id" : "94109", "count" : 441 }
{ "_id" : "94063", "count" : 381 }
{ "_id" : "94587", "count" : 257 }
{ "_id" : "94114", "count" : 200 }
{ "_id" : "94061", "count" : 169 }
{ "_id" : "94110", "count" : 165 }
{ "_id" : "94102", "count" : 158 }
{ "_id" : "94123", "count" : 130 }
{ "_id" : "94108", "count" : 125 }

{ "_id" : "94131", "count" : 123 }
{ "_id" : "94105", "count" : 112 }
{ "_id" : "94501", "count" : 112 }

The post code starts with “941” is SF's post code. However, “946..” is Oakland's.
“940.." is Redwood City’s and “945..” is Alameda County's.
OpenStreetMap doesn’t classify the post codes perfectly.

Overview of the data

File Sizes
san-francisco_california.osm……952MB
san-francisco_california.osm.json……1.08 GB

#Number of documents
> db.sf_r.find().count()
4974618

#Number of unique users
> db.sf_r.distinct("created.user").length
2202

#Number of nodes
> db.sf_r.find({"type":"node"}).count()
4470070

#Number of ways
> db.sf_r.find({"type":"way"}).count()
504548

#Number of distinct amenities

```
> db.sf_r.distinct("amenity").length
170
```

Other ideas about the datasets

・About amenity

There are 170 kinds of amenities in SF.So, I'll look into more about amenity. To begin with, I sorted amenities in descending order and confined to top 10 amenities.

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

{ "_id" : "parking", "count" : 3992 }
{ "_id" : "restaurant", "count" : 2659 }
{ "_id" : "school", "count" : 1302 }
{ "_id" : "place_of_worship", "count" : 1120 }
{ "_id" : "bench", "count" : 978 }
{ "_id" : "cafe", "count" : 842 }
{ "_id" : "post_box", "count" : 646 }
{ "_id" : "fast_food", "count" : 585 }
{ "_id" : "bicycle_parking", "count" : 483 }
{ "_id" : "drinking_water", "count" : 442 }
```

It seems that the number of parkings are the highest.
I'm interested in what is the popular fast-food shop in SF, so I investigated the number of fast-food chain in SF.

```
> db.sf_r.aggregate([{"$match":{"cuisine":{"$exists":1},"amenity":"fast_food"}},
{"$group":{"_id":"$name","count":{"$sum":1}}},
{"$sort":{"count":-1}},{"$limit":10}] )
```

{ "_id" : "McDonald's", "count" : 48 }
{ "_id" : "Subway", "count" : 29 }
{ "_id" : "Burger King", "count" : 17 }
{ "_id" : "Wendy's", "count" : 12 }
{ "_id" : "Taco Bell", "count" : 10 }
{ "_id" : "Chipotle", "count" : 8 }
{ "_id" : "Jack in the Box", "count" : 7 }
{ "_id" : "KFC", "count" : 7 }
{ "_id" : "Noah's Bagels", "count" : 6 }
{ "_id" : "Jamba Juice", "count" : 5 }

McDonald's has the largest number in fast-food restaurants.

・Who contributed the most to make OpenStreetMap(OSM) in SF.

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

{ "_id" : "ediyes", "count" : 944015 }

"ediyes" has contributed the most to make OSM and the contribution is quite high (19.0%)

Top 3 users' contribution account for 41.7% of OSM in SF and
top 10 user's contribution account for 63.6%.

This shows that quite small number of users make the most of OSM.

Conclusion

OpenStreetMap are edited by human so it includes errors,typos and inconsistency.Therefore,we need to audit and develop plan for cleaning and then, we need to write some codes to clean when we plan to utilize this map.