Udacity Data Analyst Nanodegree P3: Wrangle OpenStreetMap Data

Based on the course "Data Wrangling with MongoDB"

Umang Agarwal
Map Area: Kolkata, West Bengal, India

https://www.openstreetmap.org/node/245707150
https://mapzen.com/data/metro-extracts
https://en.wikipedia.org/wiki/Kolkata

Problems encountered:

1) Inconsistent abbreviation of street names. For example, "Hana Para Rd" (Here, "Rd" stands for road)

   -- Street names cleaned while inserting using the update_name python function written for Lesson 6 exercises. Please refer to the code in "Lesson6_ImprovingStreetNames.py" and "Lesson6_PreparingForDatabase.py".

2) Inconsistent and sometimes incorrect formatting used for phone numbers. For example, "+91-033-22837161" and "03323353029"

   -- Proper use of country and city dial-in codes was enforced before importing the dataset by calling the update_phone function from "phoneNumber_cleaner.py" in "Lesson6_PreparingForDatabase.py".

3) There are about 1005 contributions (node and way tags) created by a particular user named "FuckOSMFandODbL" As the name suggests, and as reconfirmed by manual inspection, the contributions made by the user are incorrect and spurious.

   -- Contributions by this user were removed after importing the dataset into MongoDB. The following query was used in the mongo shell:

```
> db.osm.remove({"created.user":"FuckOSMFandODbL"})
WriteResult({ "nRemoved" : 1005 })
```

Overview of the Data:

Some exploration of the data (using the python script in "zipcode_checker.py") revealed that there are only 30 zip codes in the dataset, all of which are valid and correctly formatted.

The size of the uncompressed "kolkata_india.osm" file is 651.6 MB and the "kolkata_india.osm.json" file is 767 MB.

High-level overview of the data points using basic queries in the mongo shell (after removal of contributions by one user):

Number of documents:

```
> db.osm.find().count()
3581294
```

Number of nodes:

```
> db.osm.find({"type":"node"}).count()
2960601
```

Number of ways:

```
> db.osm.find({"type":"way"}).count()
620692
```

Number of unique users:

```
> db.osm.distinct("created.uid").length
346
```

Number of cafes:

```
>  db.osm.find({"amenity":"cafe"}).count()
15
```

Number of universities:

```
> b = db.osm.find({"amenity":"university"}).count()
29
```

Number of colleges:

```
> a = db.osm.find({"amenity":"college"}).count()
71
```

Possible improvements:

Incomplete amenities data:

The data is still incomplete. The number of entries for amenities is very low. For the ones that are there in the dataset, the metadata is either incomplete or incorrect. For example, cuisine information about restaurants is still incomplete.

Perhaps, bots and screen scraping of popular Indian restaurant search websites (like zomato.com) could be used to add many more restaurants and rich metadata about these restaurants.

Incomplete Public Transport data:

Basic public transport data, like metro (subway) stations is also incomplete.

```
> db.osm.find({"railway":"subway_entrance"}).count()
8
```

Such data could easily be obtained from sources like Wikipedia. However, data about other modes of transport like buses and local intra-city railway may not be easy to find, as there is no source of reliable and structured data about these services.

Additional data exploration using the mongo shell:

Average number of constituent colleges per university:

(Universities in Kolkata are modeled on the collegiate university system where each university has one or more constituent colleges.)

```
> a/b
2.4482758620689653
```

Top 5 Amenities:

```
> db.osm.aggregate([{"$match":{"amenity":{"$exists":
1}}},{"$group"
:{"_id":"$amenity","count":{"$sum":1}}},{"$sort":{"count": -
1}},{"$limit":5}])
{ "_id" : "school", "count" : 153 }
{ "_id" : "hospital", "count" : 86 }
{ "_id" : "college", "count" : 71 }
{ "_id" : "restaurant", "count" : 66 }
{ "_id" : "place_of_worship", "count" : 61 }
```

Top 5 cuisines:

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

```
{ "_id" : "indian", "count" : 16 }
{ "_id" : "international", "count" : 4 }
{ "_id" : "multicuisine", "count" : 3 }
{ "_id" : "regional", "count" : 2 }
{ "_id" : "chinese", "count" : 2 }
```

Biggest religions:

```
db.osm.aggregate([{"$match":{"amenity":{"$exists":1},"amenity"
:"place_of_worship","religion":{"$exists":1}}},{"$group":{"_id
":"$religion","count":{"$sum":1}}},{"$sort":{"count":-
1}},{"$limit":3}])
{ "_id" : "hindu", "count" : 24 }
{ "_id" : "christian", "count" : 11 }
{ "_id" : "muslim", "count" : 6 }
```

Banks with most ATMs:

```
>
db.osm.aggregate([{"$match":{"amenity":"atm","name":{"$exists"
:1}}},{"$group":{"_id":"$name","count":{"$sum":1}}},{"$sort":{
"count" :-1}},{"$limit": 5}])
{ "_id" : "Axis Bank ATM", "count" : 6 }
{ "_id" : "SBI ATM", "count" : 4 }
{ "_id" : "State Bank of India", "count" : 3 }
{ "_id" : "State Bank of India ATM", "count" : 3 }
{ "_id" : "HDFC Bank ATM", "count" : 3 }
```

Number of Missionary Schools in Kolkata:

(Many schools in Kolkata are operated by Christian missionaries, and hence contain “St” (for Saint) in their name)

```
>
db.osm.aggregate([{"$match":{"amenity":"school","name":/.*St.*
/}},{"$group":{"_id":null,"count":{"$sum":1}}}])
{ "_id" : null, "count" : 15 }
```

A few examples of missionary schools:

```
>
db.osm.aggregate([{"$match":{"amenity":"school","name":/.*St.*
/}},{"$group":{"_id":"$name","count":{"$sum":1}}},{"$limit":3}
])
{ "_id" : "St. Thomas' School", "count" : 1 }
{ "_id" : "St. Joseph's Convent", "count" : 1 }
{ "_id" : "St Augustines Day School", "count" : 1 }
```

Number of Missionary Colleges in Kolkata:

```
>
db.osm.aggregate([{"$match":{"amenity":"college","name":/.*St.
*/}},{"$group":{"_id":null,"count":{"$sum":1}}}])
{ "_id" : null, "count" : 5 }
```