Udacity: Data Wrangling with MongoDB Final Project

P.S. Aravind

Overview of the data

The data selected for this project is from Openstreetmap data set which is freely available XML data for download. One of the primary reasons for choosing this data set is that its human edited which is prone to errors and has lots of opportunities for cleaning.

To start I selected the data set for my city where I grew up "Chennai, India", after spending couple of days analyzing the data I found that the file size for this city did not satisfy the project requirements of 50 MB. So I selected data set for another city, New Delhi from India, file size for this data set is around 85 MB. This data set has around 486 unique users, 397,069 Nodes and 68,188 ways.

Output from "mapparser.py":

```
{'bounds': 1,
 'member': 14525,
 'nd': 525619,
 'node': 397069,
 'osm': 1,
 'relation': 373,
 'tag': 143726,
 'way': 68188}
```

Output from "users.py":

```
{'1012178',
 '102407',
 '1028662',
. . .
 '989011',
 '989651'}
Unique user count: 486
```

The OSM data for New Delhi city was parsed and converted into XML format and then this XML data was loaded into MongoDB for analysis. Prior to loading, the data set was cleansed and reformatted to a structure that is suitable for analysis. Following section details various statistics compiled from MongoDB queries on this data:

1. Query top users, count of entries created, min and max dates

User 'Oberaffe' is the most active user creating over 224,536 entries from Dec 2012 to Nov

2013.

```
[{'_id': 'Oberaffe',
  'count': 224536,
  'max': '2013-11-23T19:00:27Z',
  'min': '2008-12-01T19:02:47Z'},
 {'_id': 'Edolis',
  'count': 52664,
  'max': '2013-10-28T12:21:52Z',
  'min': '2011-05-09T10:20:59Z'},
 {'_id': 'marek kleciak',
  'count': 16010,
  'max': '2013-06-20T13:05:36Z',
  'min': '2011-03-07T16:51:13Z'},
 {'_id': 'PlaneMad',
  'count': 15905,
  'max': '2013-11-05T16:38:52Z',
  'min': '2009-10-27T06:27:32Z'},
...
 {'_id': 'thevikas',
  'count': 11837,
  'max': '2013-12-15T11:57:37Z',
  'min': '2008-05-22T18:37:35Z'}]
```

MongoDB Query:

```
pipeline = [ { "$group" : {"_id" : "$created.user",
                           "count" : { "$sum" : 1 },
                           "min" : {"$min" : "$created.timestamp"},
                           "max" : {"$max" : "$created.timestamp"}} },
                  { "$sort" : { "count" : -1}},
                  { "$limit" : 10 }]
```

2. Query earliest entry timestamp

The file has the earliest entry for New Delhi on 23 Sep, 2007.

```
[{'_id': ObjectId('534eaf7775917475529e85b0'),
  'timestamp': '2007-09-23T02:35:38Z'}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "timestamp" : "$created.timestamp"}},
                     { "$sort" : { "timestamp" : 1}},
                     { "$limit" : 1 }]
```

3. Query latest entry timestamp

The file has the latest entry for New Delhi on 17 Dec, 2013.

```
[{'_id': ObjectId('534eafaf7591747552a494ae'), 'timestamp':
'2013-12-17T18:02:10Z'}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "timestamp" : "$created.timestamp"}},
                     { "$sort" : { "timestamp" : -1}},
                     { "$limit" : 1 }]
```

4. Query count of entries by Year and Month, sorted by count

July 2007 and Mar 2013 were the active months when the New Delhi entries were created.

```
[{'_id': '2012-07', 'count': 97021},
 {'_id': '2013-03', 'count': 70619},
 {'_id': '2012-11', 'count': 34187},
 {'_id': '2010-09', 'count': 22989},
 {'_id': '2012-01', 'count': 22204},
….
 {'_id': '2008-02', 'count': 11},
 {'_id': '2009-04', 'count': 10},
 {'_id': '2008-04', 'count': 1}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "datetime" :
                          { "$substr": ["$created.timestamp", 0, 7 ]}}},
                     { "$group" : {"_id" : "$datetime",
                                   "count" : { "$sum" : 1 }}},
                     { "$sort" : { "count" : -1}}]
```

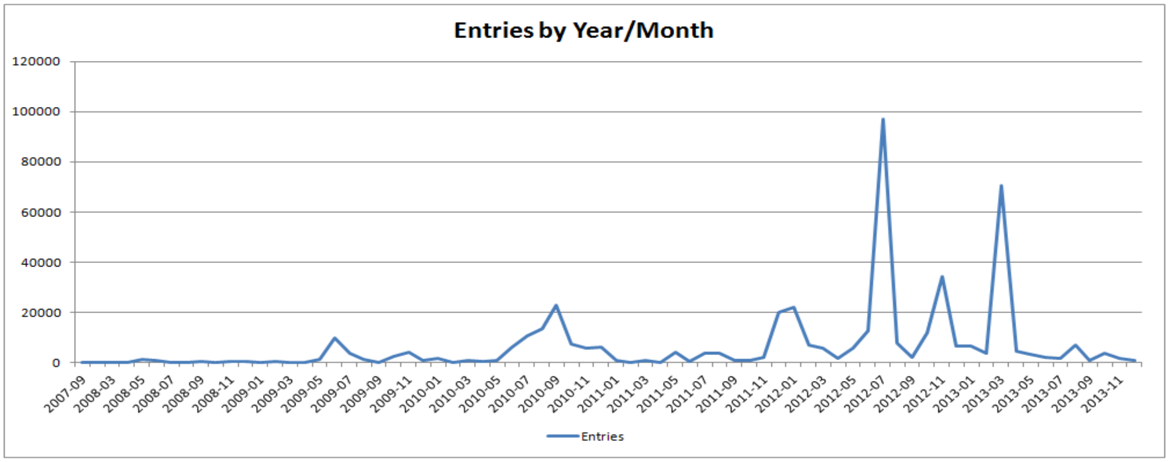
5. Query count of entries by Year and Month, sorted by Year/Month

Chart shows the entries by month from Sep 2007 to Dec 2013 showing the two peaks on July 2007 and Mar 2013.

```
[{'_id': '2007-09', 'count': 57},
 {'_id': '2008-02', 'count': 11},
 {'_id': '2008-03', 'count': 122},
 {'_id': '2008-04', 'count': 1},
 {'_id': '2008-05', 'count': 1182},
…
 {'_id': '2012-07', 'count': 97021},
…
 {'_id': '2013-03', 'count': 70619},
...
 {'_id': '2013-10', 'count': 3551},
 {'_id': '2013-11', 'count': 1902},
 {'_id': '2013-12', 'count': 858}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "datetime" : { "$substr": ["$created.timestamp",
0, 7 ]}}},
                    { "$group" : {"_id" : "$datetime",
                                  "count" : { "$sum" : 1 }}},
                     { "$sort" : { "_id" : 1}}]
```



6. Query count Place of Worship

There are 154 places identified as place of Worship in New Delhi.

```
[{'_id': 'Hindu', 'count': 1},
 {'_id': 'bahai', 'count': 1},
 {'_id': 'Sikh', 'count': 2},
 {'_id': 'jain', 'count': 2},
 {'_id': 'buddhist', 'count': 3},
 {'_id': 'sikh', 'count': 16},
 {'_id': 'christian', 'count': 22},
 {'_id': 'muslim', 'count': 28},
 {'_id': 'hindu', 'count': 79},
 {'_id': None, 'count': 465103}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "amenity" : { "$eq" : ["$amenity",
"place_of_worship" ]},
                                      "religion" : "$religion" }},
                     { "$group" : { "_id" : "$religion",
                                    "count" : { "$sum" : 1 }}},
                     { "$sort" : { "count" : 1}}]
```

7. Query count different Amenities

Following query shows different amenities in New Delhi.

```
[{'_id': 'crematorium', 'count': 1},
 {'_id': 'community_centre', 'count': 1},
 {'_id': 'community_hall', 'count': 1},
…
 {'_id': 'atm', 'count': 116},
 {'_id': 'hospital', 'count': 130},
 {'_id': 'fuel', 'count': 161},
 {'_id': 'place_of_worship', 'count': 196},
 {'_id': 'parking', 'count': 219},
 {'_id': 'school', 'count': 529},
 {'_id': None, 'count': 462977}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "amenity" : "$amenity" }},
                     { "$group" : { "_id" : "$amenity",
                                    "count" : { "$sum" : 1 }}},
                     { "$sort" : { "count" : 1}}]
```

8. Query count of different places of tourism

Following query shows different places of tourism in New Delhi.

```
[{'_id': 'theme_park', 'count': 1},
 {'_id': 'zoo', 'count': 1},
 {'_id': 'artwork', 'count': 2},
 {'_id': 'motel', 'count': 2},
 {'_id': 'viewpoint', 'count': 5},
 {'_id': 'museum', 'count': 7},
 {'_id': 'hostel', 'count': 9},
 {'_id': 'information', 'count': 13},
 {'_id': 'guest_house', 'count': 14},
 {'_id': 'attraction', 'count': 66},
 {'_id': 'hotel', 'count': 118},
 {'_id': None, 'count': 465019}]
```

MongoDB Query:

```
pipeline = [ { "$project" : { "tourism" : "$tourism" }},
                     { "$group" : { "_id" : "$tourism",
                                    "count" : { "$sum" : 1 }}},
                     { "$sort" : { "count" : 1}}]
```

Other ideas about the datasets

Considering the size of the city and 9.8 Million population, number of users: 486 who have contributed to this data set is very small. The data shows that there are 154 places identified as “Place of Worship”, considering the religious nature of the Indian population, shows that the New Delhi data is not complete and need to be updated with more details of the places of interest.