OPENSTREETMAP DATA CASE STUDY

Map Area

New York City (a piece of Manhattan) , NY, United States

```
<bounds minlat="40.7030000" minlon="-74.0145000" maxlat="40.7688000" maxlon="-73.9696000"/>
```

I choose this area because I work in financial markets and this area is the financial center of the United States. One day I wish to go there.

I used mongodb over SQL for learning purposes. I think it would be a good idea to learn mongo.

PROBLEMS ENCOUNTERED IN THE MAP

After initially downloading a small sample size of the map, I processed it in data.py and inserted in MongoDB to view a bit of the schema. After using the schema analyzer

- Some second level "k" tags separated by "."

```
<tag k="cityracks.housenum" v="295"/>
<tag k="cityracks.large" v="0"/>
<tag k="cityracks.rackid" v="2410"/>
<tag k="cityracks.small" v="1"/>
<tag k="cityracks.street" v="Canal St"/>
```

That was solved replacing "." For ":" over the tags

```
# replace second level tags containing '.' to ':'
def replace_dot(text):
    return text.replace(".",":")
```

- Second level name tags with multiple languages.

```
<tag k="name" v="Mott Street"/>
<tag k="name:en" v="Mott Street"/>
<tag k="name:zh" v="勿街"/>
```

When name is a dictionary instead of string, I consider the name[“en”] element

- Misplaced value inside second level country tag

```
<tag k="tiger:county" v="New York, NY"/>
```

Because we are only analyzing a map inside United States, the country property is not useful.

- Name abbreviations (ie: St, Ave, etc.)

```
<tag k="addr:street" v="E. 54th St."/>
```

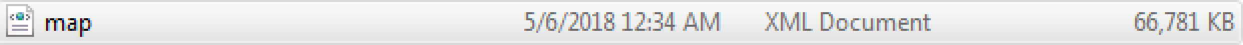
In this case, I used a function to replace for the cases found in street names auditing.

```
mapping = {"St": "Street",
           "St.": "Street",
           "Ave": "Avenue",
           "Ave.": "Avenue",
           "N.": "North",
           "W.": "West",
           "E": "East",
           "E.": "East",
           "Fdr": "Federal",
           "Streer": "Street",
           "Steet": "Street",
           "S": "South",
           "S.": "South",
           "Avene": "Avenue"
           }
```

```
# update name abbreviations
def update_name(name):
    for key in mapping:
        newname = re.sub(r'\b' + key + r'\b\.?', mapping[key], name)
    return newname
```

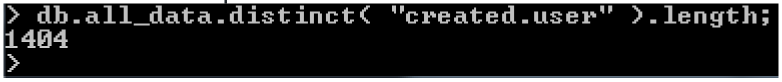
OVERVIEW OF THE DATA

- Size of the file:



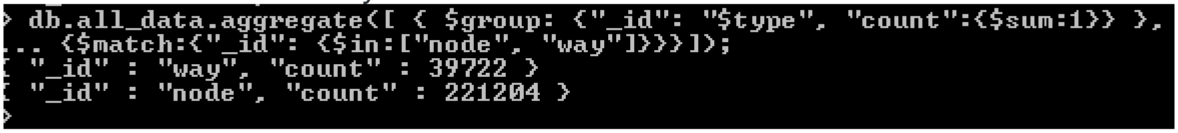
```
map                5/6/2018 12:34 AM      XML Document                66,781 KB
```

- Number of unique users:



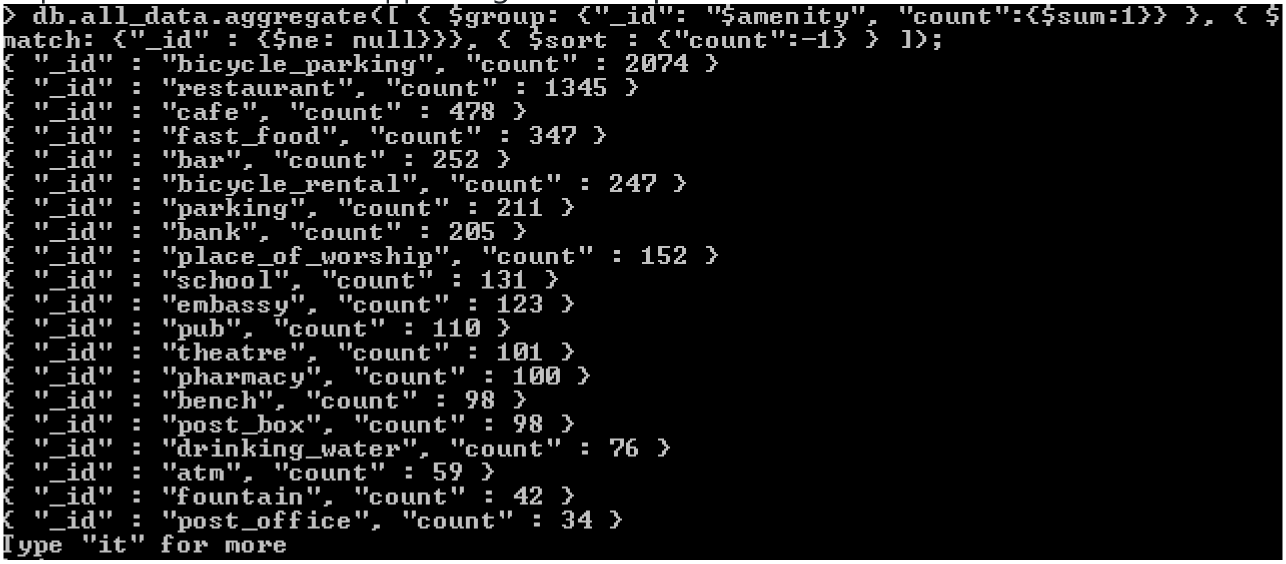
```
> db.all_data.distinct( "created.user" ).length;
1404
>
```

- Number of nodes and ways:

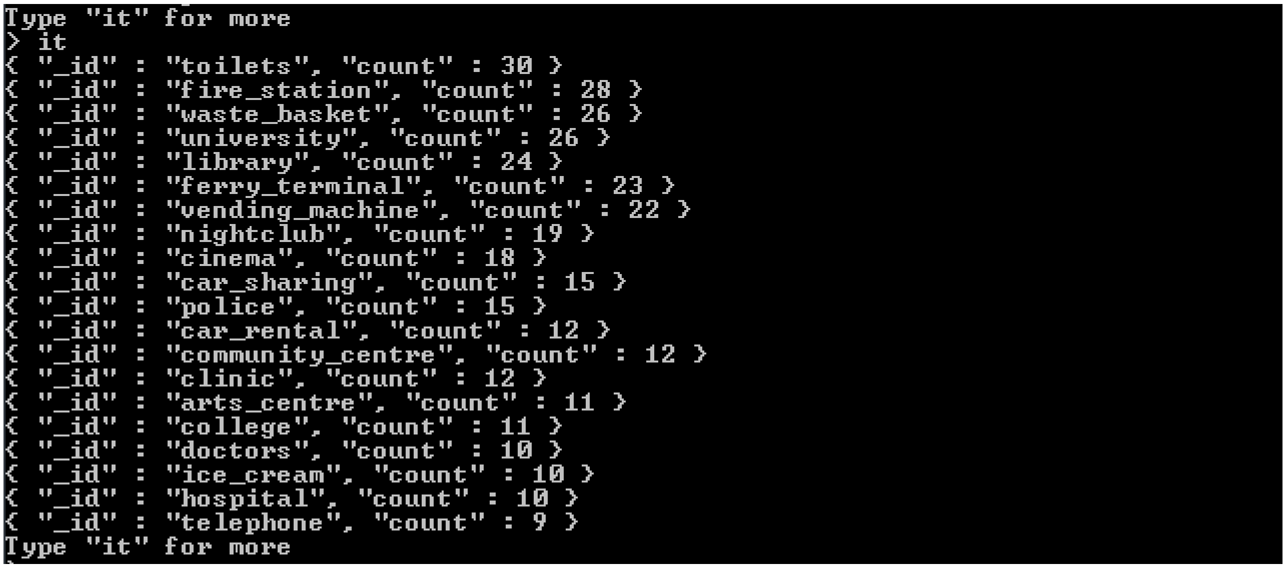


```
> db.all_data.aggregate([ { $group: {"_id": "$type", "count":{$sum:1}} },
... {$match:{"_id": {$in:["node", "way"]}}}]);
{ "_id" : "way", "count" : 39722 }
{ "_id" : "node", "count" : 221204 }
>
```

- Top number of amenities appearing in the map.



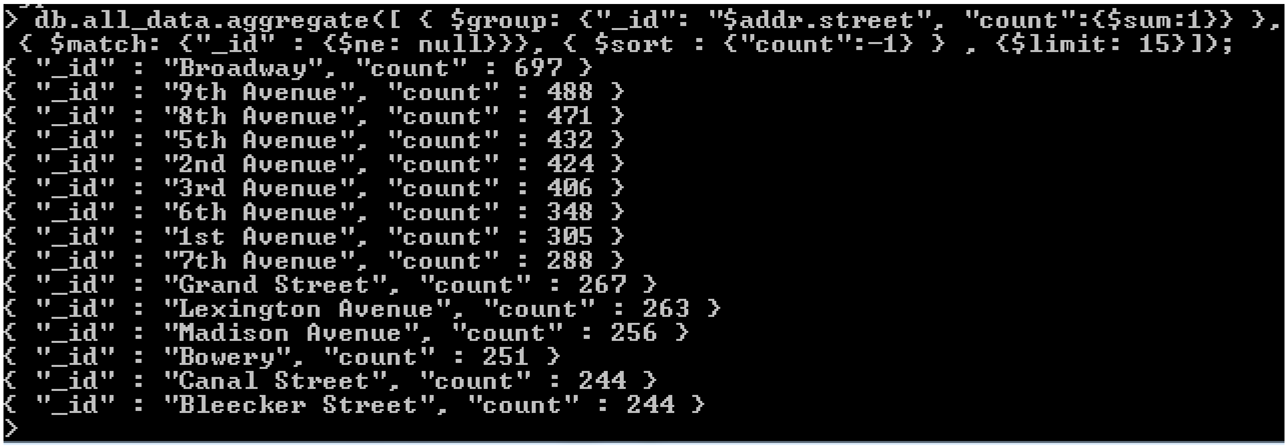
```
> db.all_data.aggregate([ { $group: {"_id": "$amenity", "count":{$sum:1}} }, { $
match: {"_id" : {$ne: null}}}, { $sort : {"count":-1} } ]);
{ "_id" : "bicycle_parking", "count" : 2074 }
{ "_id" : "restaurant", "count" : 1345 }
{ "_id" : "cafe", "count" : 478 }
{ "_id" : "fast_food", "count" : 347 }
{ "_id" : "bar", "count" : 252 }
{ "_id" : "bicycle_rental", "count" : 247 }
{ "_id" : "parking", "count" : 211 }
{ "_id" : "bank", "count" : 205 }
{ "_id" : "place_of_worship", "count" : 152 }
{ "_id" : "school", "count" : 131 }
{ "_id" : "embassy", "count" : 123 }
{ "_id" : "pub", "count" : 110 }
{ "_id" : "theatre", "count" : 101 }
{ "_id" : "pharmacy", "count" : 100 }
{ "_id" : "bench", "count" : 98 }
{ "_id" : "post_box", "count" : 98 }
{ "_id" : "drinking_water", "count" : 76 }
{ "_id" : "atm", "count" : 59 }
{ "_id" : "fountain", "count" : 42 }
{ "_id" : "post_office", "count" : 34 }
Type "it" for more
```



```
Type "it" for more
> it
{ "_id" : "toilets", "count" : 30 }
{ "_id" : "fire_station", "count" : 28 }
{ "_id" : "waste_basket", "count" : 26 }
{ "_id" : "university", "count" : 26 }
{ "_id" : "library", "count" : 24 }
{ "_id" : "ferry_terminal", "count" : 23 }
{ "_id" : "vending_machine", "count" : 22 }
{ "_id" : "nightclub", "count" : 19 }
{ "_id" : "cinema", "count" : 18 }
{ "_id" : "car_sharing", "count" : 15 }
{ "_id" : "police", "count" : 15 }
{ "_id" : "car_rental", "count" : 12 }
{ "_id" : "community_centre", "count" : 12 }
{ "_id" : "clinic", "count" : 12 }
{ "_id" : "arts_centre", "count" : 11 }
{ "_id" : "college", "count" : 11 }
{ "_id" : "doctors", "count" : 10 }
{ "_id" : "ice_cream", "count" : 10 }
{ "_id" : "hospital", "count" : 10 }
{ "_id" : "telephone", "count" : 9 }
Type "it" for more
```

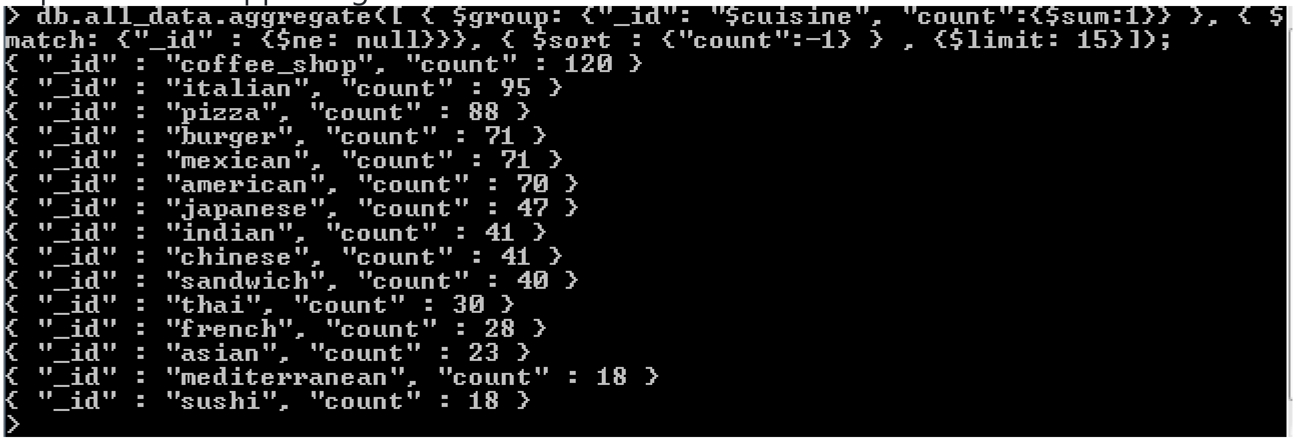
ADDITIONAL IDEAS

- Top 15 number of streets appearing in result :



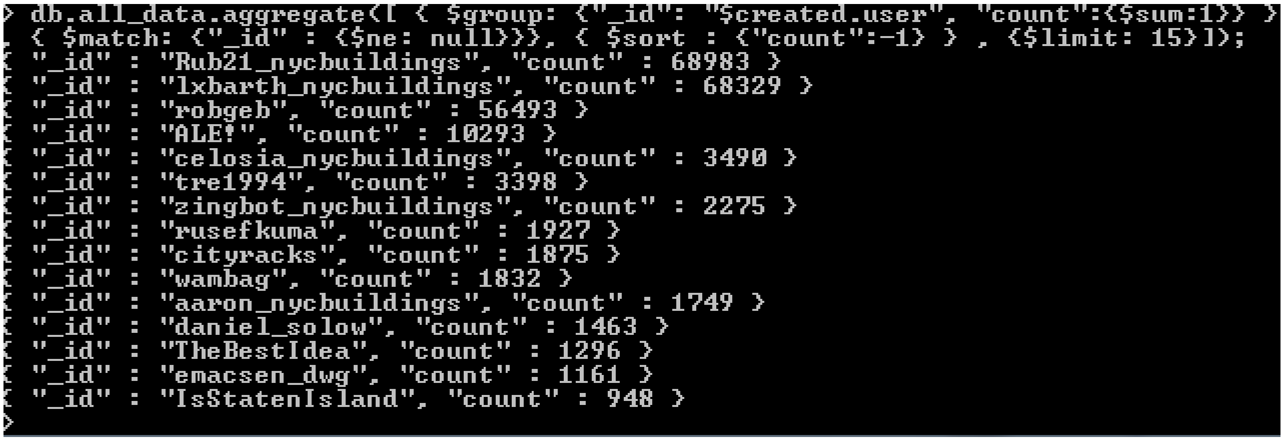
```
> db.all_data.aggregate([ { $group: {"_id": "$addr.street", "count":{$sum:1}} },
 { $match: {"_id" : {$ne: null}}}, { $sort : {"count":-1} } , {$limit: 15}]);
{ "_id" : "Broadway", "count" : 697 }
{ "_id" : "9th Avenue", "count" : 488 }
{ "_id" : "8th Avenue", "count" : 471 }
{ "_id" : "5th Avenue", "count" : 432 }
{ "_id" : "2nd Avenue", "count" : 424 }
{ "_id" : "3rd Avenue", "count" : 406 }
{ "_id" : "6th Avenue", "count" : 348 }
{ "_id" : "1st Avenue", "count" : 305 }
{ "_id" : "7th Avenue", "count" : 288 }
{ "_id" : "Grand Street", "count" : 267 }
{ "_id" : "Lexington Avenue", "count" : 263 }
{ "_id" : "Madison Avenue", "count" : 256 }
{ "_id" : "Bowery", "count" : 251 }
{ "_id" : "Canal Street", "count" : 244 }
{ "_id" : "Bleecker Street", "count" : 244 }
>
```

- Top 15 cousine appearing in result :



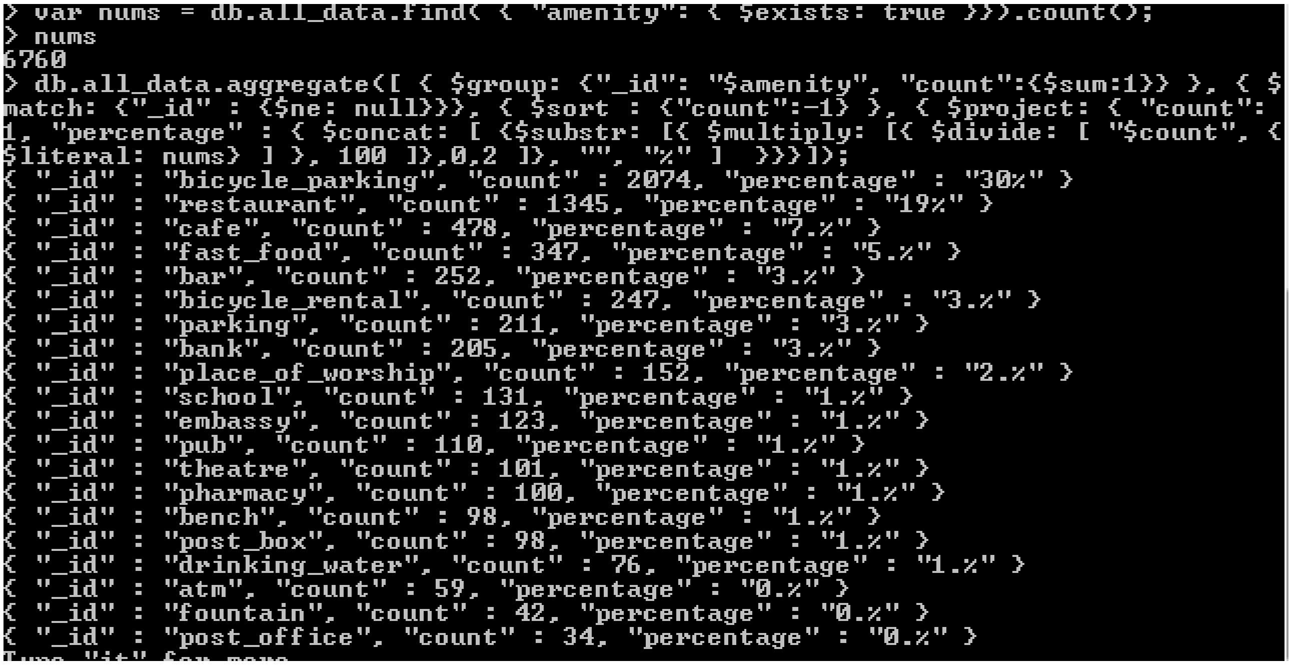
```
> db.all_data.aggregate([ { $group: {"_id": "$cuisine", "count":{$sum:1}} }, { $
match: {"_id" : {$ne: null}}}, { $sort : {"count":-1} } , {$limit: 15}]);
{ "_id" : "coffee_shop", "count" : 120 }
{ "_id" : "italian", "count" : 95 }
{ "_id" : "pizza", "count" : 88 }
{ "_id" : "burger", "count" : 71 }
{ "_id" : "mexican", "count" : 71 }
{ "_id" : "american", "count" : 70 }
{ "_id" : "japanese", "count" : 47 }
{ "_id" : "indian", "count" : 41 }
{ "_id" : "chinese", "count" : 41 }
{ "_id" : "sandwich", "count" : 40 }
{ "_id" : "thai", "count" : 30 }
{ "_id" : "french", "count" : 28 }
{ "_id" : "asian", "count" : 23 }
{ "_id" : "mediterranean", "count" : 18 }
{ "_id" : "sushi", "count" : 18 }
>
```

- Top 15 users contribution :



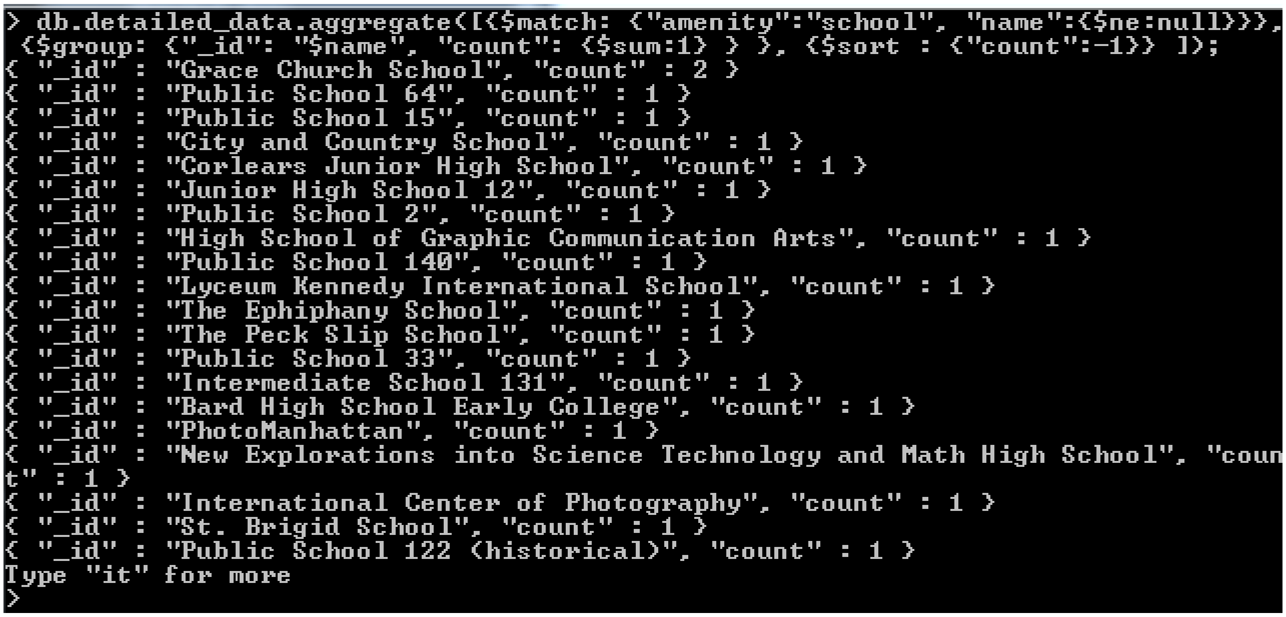
```
> db.all_data.aggregate([ { $group: {"_id": "$created.user", "count":{$sum:1}} }
, { $match: {"_id" : {$ne: null}}}, { $sort : {"count":-1} } , {$limit: 15}]);
{ "_id" : "Rub21_nycbuildings", "count" : 68983 }
{ "_id" : "lxbarth_nycbuildings", "count" : 68329 }
{ "_id" : "robgeb", "count" : 56493 }
{ "_id" : "ALE!", "count" : 10293 }
{ "_id" : "celosia_nycbuildings", "count" : 3490 }
{ "_id" : "tre1994", "count" : 3398 }
{ "_id" : "zingbot_nycbuildings", "count" : 2275 }
{ "_id" : "rusefkuma", "count" : 1927 }
{ "_id" : "cityracks", "count" : 1875 }
{ "_id" : "wambag", "count" : 1832 }
{ "_id" : "aaron_nycbuildings", "count" : 1749 }
{ "_id" : "daniel_solow", "count" : 1463 }
{ "_id" : "TheBestIdea", "count" : 1296 }
{ "_id" : "emacsen_dwg", "count" : 1161 }
{ "_id" : "IsStatenIsland", "count" : 948 }
>
```

- Top amenity appears as bicycle parking with 30%, followed by restaurant with 19% :



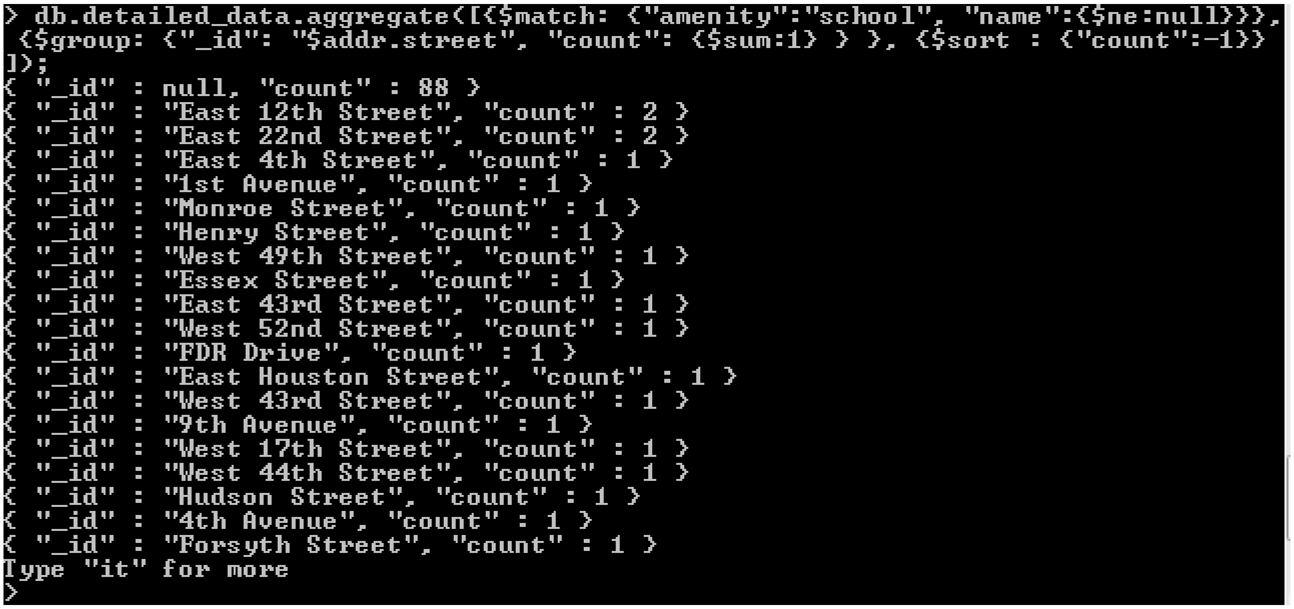
```
> var nums = db.all_data.find( { "amenity": { $exists: true }}).count();
> nums
6760
> db.all_data.aggregate([ { $group: {"_id": "$amenity", "count":{$sum:1}} }, { $
match: {"_id" : {$ne: null}}}, { $sort : {"count":-1} }, { $project: { "count":
1, "percentage" : { $concat: [ {$substr: [{ $multiply: [{ $divide: [ "$count", {
$literal: nums} ] }, 100 ]},0,2 ]}, "", "%" ]  }}}]);
{ "_id" : "bicycle_parking", "count" : 2074, "percentage" : "30%" }
{ "_id" : "restaurant", "count" : 1345, "percentage" : "19%" }
{ "_id" : "cafe", "count" : 478, "percentage" : "7.%" }
{ "_id" : "fast_food", "count" : 347, "percentage" : "5.%" }
{ "_id" : "bar", "count" : 252, "percentage" : "3.%" }
{ "_id" : "bicycle_rental", "count" : 247, "percentage" : "3.%" }
{ "_id" : "parking", "count" : 211, "percentage" : "3.%" }
{ "_id" : "bank", "count" : 205, "percentage" : "3.%" }
{ "_id" : "place_of_worship", "count" : 152, "percentage" : "2.%" }
{ "_id" : "school", "count" : 131, "percentage" : "1.%" }
{ "_id" : "embassy", "count" : 123, "percentage" : "1.%" }
{ "_id" : "pub", "count" : 110, "percentage" : "1.%" }
{ "_id" : "theatre", "count" : 101, "percentage" : "1.%" }
{ "_id" : "pharmacy", "count" : 100, "percentage" : "1.%" }
{ "_id" : "bench", "count" : 98, "percentage" : "1.%" }
{ "_id" : "post_box", "count" : 98, "percentage" : "1.%" }
{ "_id" : "drinking_water", "count" : 76, "percentage" : "1.%" }
{ "_id" : "atm", "count" : 59, "percentage" : "0.%" }
{ "_id" : "fountain", "count" : 42, "percentage" : "0.%" }
{ "_id" : "post_office", "count" : 34, "percentage" : "0.%" }
Type "it" for more
```

When looking into schools, we found that there are 131 documents with “amenity”=”school” and Grace Curch School appearing two times.



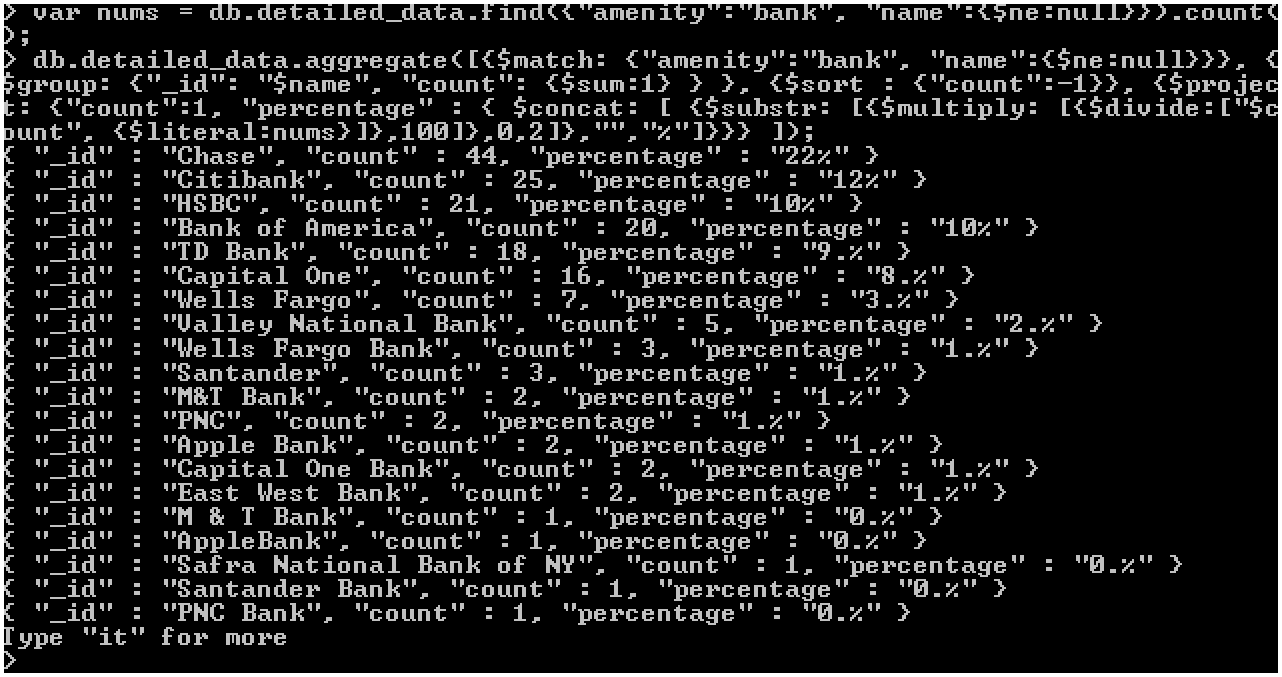
```
> db.detailed_data.aggregate([{$match: {"amenity":"school", "name":{$ne:null}}},
 {$group: {"_id": "$name", "count": {$sum:1} } }, {$sort : {"count":-1}} ]);
{ "_id" : "Grace Church School", "count" : 2 }
{ "_id" : "Public School 64", "count" : 1 }
{ "_id" : "Public School 15", "count" : 1 }
{ "_id" : "City and Country School", "count" : 1 }
{ "_id" : "Corlears Junior High School", "count" : 1 }
{ "_id" : "Junior High School 12", "count" : 1 }
{ "_id" : "Public School 2", "count" : 1 }
{ "_id" : "High School of Graphic Communication Arts", "count" : 1 }
{ "_id" : "Public School 140", "count" : 1 }
{ "_id" : "Lyceum Kennedy International School", "count" : 1 }
{ "_id" : "The Ephiphany School", "count" : 1 }
{ "_id" : "The Peck Slip School", "count" : 1 }
{ "_id" : "Public School 33", "count" : 1 }
{ "_id" : "Intermediate School 131", "count" : 1 }
{ "_id" : "Bard High School Early College", "count" : 1 }
{ "_id" : "PhotoManhattan", "count" : 1 }
{ "_id" : "New Explorations into Science Technology and Math High School", "coun
t" : 1 }
{ "_id" : "International Center of Photography", "count" : 1 }
{ "_id" : "St. Brigid School", "count" : 1 }
{ "_id" : "Public School 122 (historical)", "count" : 1 }
Type "it" for more
>
```

Grouping by street we found that there are two schools in “East 12th Street” and “East 22nd Street”, with 88 documents that not contain “addr:street” field.



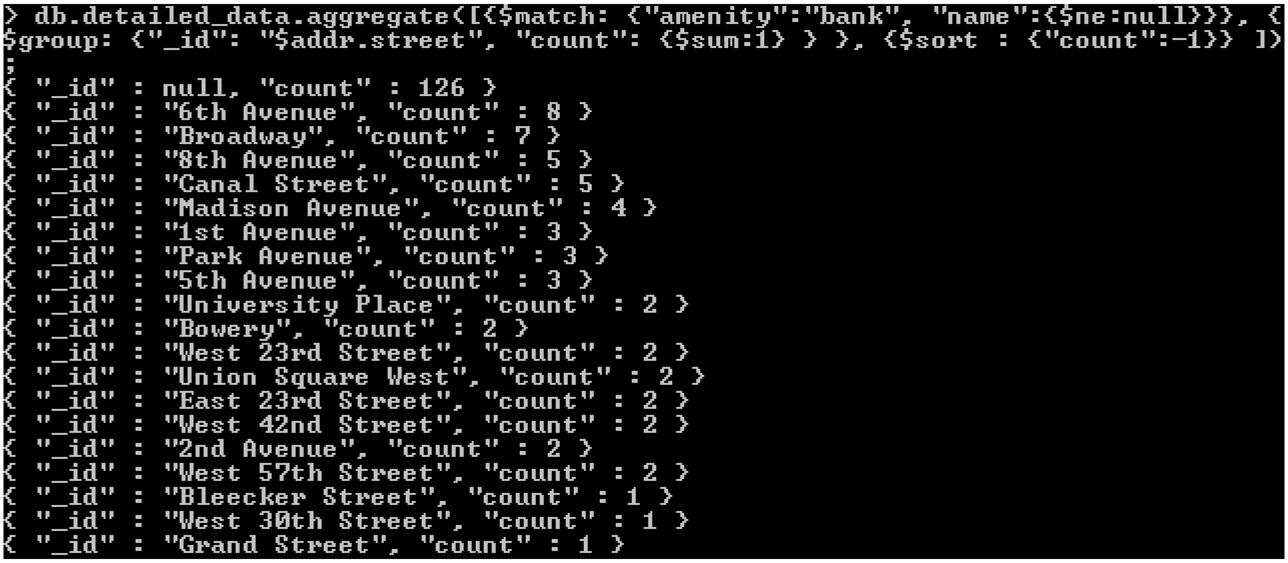
```
> db.detailed_data.aggregate([{$match: {"amenity":"school", "name":{$ne:null}}},
 {$group: {"_id": "$addr.street", "count": {$sum:1} } }, {$sort : {"count":-1}}
]);
{ "_id" : null, "count" : 88 }
{ "_id" : "East 12th Street", "count" : 2 }
{ "_id" : "East 22nd Street", "count" : 2 }
{ "_id" : "East 4th Street", "count" : 1 }
{ "_id" : "1st Avenue", "count" : 1 }
{ "_id" : "Monroe Street", "count" : 1 }
{ "_id" : "Henry Street", "count" : 1 }
{ "_id" : "West 49th Street", "count" : 1 }
{ "_id" : "Essex Street", "count" : 1 }
{ "_id" : "East 43rd Street", "count" : 1 }
{ "_id" : "West 52nd Street", "count" : 1 }
{ "_id" : "FDR Drive", "count" : 1 }
{ "_id" : "East Houston Street", "count" : 1 }
{ "_id" : "West 43rd Street", "count" : 1 }
{ "_id" : "9th Avenue", "count" : 1 }
{ "_id" : "West 17th Street", "count" : 1 }
{ "_id" : "West 44th Street", "count" : 1 }
{ "_id" : "Hudson Street", "count" : 1 }
{ "_id" : "4th Avenue", "count" : 1 }
{ "_id" : "Forsyth Street", "count" : 1 }
Type "it" for more
>
```

Chase is the top appearing bank, with 22% of time (44 occurrences)



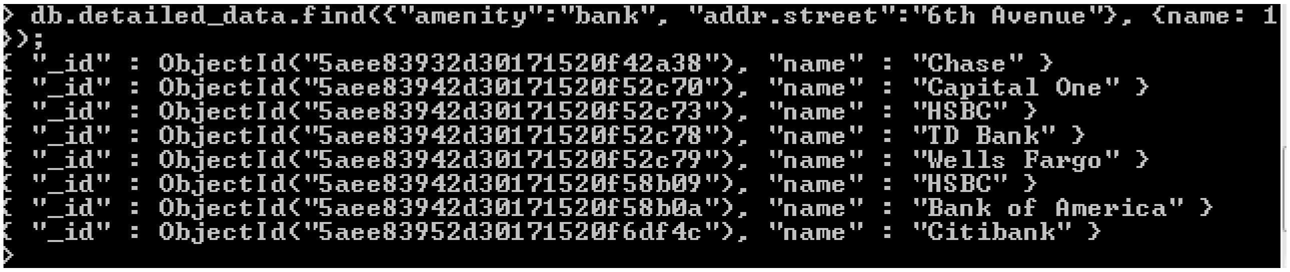
```
> var nums = db.detailed_data.find({"amenity":"bank", "name":{$ne:null}}).count(
);
> db.detailed_data.aggregate([{$match: {"amenity":"bank", "name":{$ne:null}}}, {
$group: {"_id": "$name", "count": {$sum:1} } }, {$sort : {"count":-1}}, {$projec
t: {"count":1, "percentage" : { $concat: [ {$substr: [{$multiply: [{$divide:["$c
ount", {$literal:nums}]},100]},0,2]},"","%"]}}} ]);
{ "_id" : "Chase", "count" : 44, "percentage" : "22%" }
{ "_id" : "Citibank", "count" : 25, "percentage" : "12%" }
{ "_id" : "HSBC", "count" : 21, "percentage" : "10%" }
{ "_id" : "Bank of America", "count" : 20, "percentage" : "10%" }
{ "_id" : "TD Bank", "count" : 18, "percentage" : "9.%" }
{ "_id" : "Capital One", "count" : 16, "percentage" : "8.%" }
{ "_id" : "Wells Fargo", "count" : 7, "percentage" : "3.%" }
{ "_id" : "Valley National Bank", "count" : 5, "percentage" : "2.%" }
{ "_id" : "Wells Fargo Bank", "count" : 3, "percentage" : "1.%" }
{ "_id" : "Santander", "count" : 3, "percentage" : "1.%" }
{ "_id" : "M&T Bank", "count" : 2, "percentage" : "1.%" }
{ "_id" : "PNC", "count" : 2, "percentage" : "1.%" }
{ "_id" : "Apple Bank", "count" : 2, "percentage" : "1.%" }
{ "_id" : "Capital One Bank", "count" : 2, "percentage" : "1.%" }
{ "_id" : "East West Bank", "count" : 2, "percentage" : "1.%" }
{ "_id" : "M & T Bank", "count" : 1, "percentage" : "0.%" }
{ "_id" : "AppleBank", "count" : 1, "percentage" : "0.%" }
{ "_id" : "Safra National Bank of NY", "count" : 1, "percentage" : "0.%" }
{ "_id" : "Santander Bank", "count" : 1, "percentage" : "0.%" }
{ "_id" : "PNC Bank", "count" : 1, "percentage" : "0.%" }
Type "it" for more
>
```

We have 126 documents with null field in “addr:street” for “amenity”=”bank”. Counted total of 8 banks in 6th Avenue.



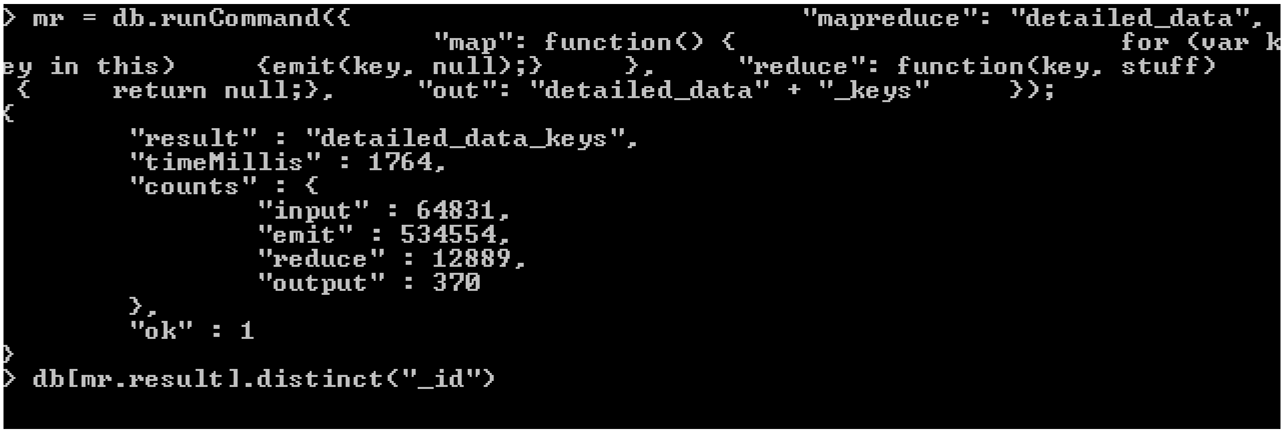
```
> db.detailed_data.aggregate([{$match: {"amenity":"bank", "name":{$ne:null}}}, {
$group: {"_id": "$addr.street", "count": {$sum:1} } }, {$sort : {"count":-1}} ])
;
{ "_id" : null, "count" : 126 }
{ "_id" : "6th Avenue", "count" : 8 }
{ "_id" : "Broadway", "count" : 7 }
{ "_id" : "8th Avenue", "count" : 5 }
{ "_id" : "Canal Street", "count" : 5 }
{ "_id" : "Madison Avenue", "count" : 4 }
{ "_id" : "1st Avenue", "count" : 3 }
{ "_id" : "Park Avenue", "count" : 3 }
{ "_id" : "5th Avenue", "count" : 3 }
{ "_id" : "University Place", "count" : 2 }
{ "_id" : "Bowery", "count" : 2 }
{ "_id" : "West 23rd Street", "count" : 2 }
{ "_id" : "Union Square West", "count" : 2 }
{ "_id" : "East 23rd Street", "count" : 2 }
{ "_id" : "West 42nd Street", "count" : 2 }
{ "_id" : "2nd Avenue", "count" : 2 }
{ "_id" : "West 57th Street", "count" : 2 }
{ "_id" : "Bleecker Street", "count" : 1 }
{ "_id" : "West 30th Street", "count" : 1 }
{ "_id" : "Grand Street", "count" : 1 }
```

Banks in 6th Avenue:



```
> db.detailed_data.find({"amenity":"bank", "addr.street":"6th Avenue"}, {name: 1
});
{ "_id" : ObjectId("5aee83932d30171520f42a38"), "name" : "Chase" }
{ "_id" : ObjectId("5aee83942d30171520f52c70"), "name" : "Capital One" }
{ "_id" : ObjectId("5aee83942d30171520f52c73"), "name" : "HSBC" }
{ "_id" : ObjectId("5aee83942d30171520f52c78"), "name" : "TD Bank" }
{ "_id" : ObjectId("5aee83942d30171520f52c79"), "name" : "Wells Fargo" }
{ "_id" : ObjectId("5aee83942d30171520f58b09"), "name" : "HSBC" }
{ "_id" : ObjectId("5aee83942d30171520f58b0a"), "name" : "Bank of America" }
{ "_id" : ObjectId("5aee83952d30171520f6df4c"), "name" : "Citibank" }
>
```

As suggested by the reviewer, it really would be good if there was a rating key on the data, so It could suggest the service quality on the specified amenity. This rating could be integrated by something like Foursquare app. After running the following command to check for all available fields in the collection we see that there are not any fields that describe the amenity rating (thanks to user Kristina @ https://stackoverflow.com/questions/2298870/get-names-of-all-keys-in-the-collection)



```
> mr = db.runCommand({                              "mapreduce": "detailed_data",
                          "map": function() {                         for (var k
ey in this)     {emit(key, null);}      },      "reduce": function(key, stuff)
 {     return null;},      "out": "detailed_data" + "_keys"      });
{
        "result" : "detailed_data_keys",
        "timeMillis" : 1764,
        "counts" : {
                "input" : 64831,
                "emit" : 534554,
                "reduce" : 12889,
                "output" : 370
        },
        "ok" : 1
}
> db[mr.result].distinct("_id")
```

CONCLUSION

Although I think that by analyzing a map I could know better the region, the process is more complex than I imagined because of the amount of data contained in the map. There are many fields that should be standardized when registering data in the system. I believe this would help us to have a slightly cleaner database.