OpenStreetMap­ Project  
Data Wrangling With MongoDB

Data Analyst NanoDegree

Mark Ayzenshtadt

Map Area: Moscow (South-Western part), Russia.

1. Problems Encountered in the Map

The main problems encountered in the map are:

* Inconsistent house numbers
* Inconsistent phone numbers

Inconsistent house numbers.

The house numbers can sometimes have letters in it, e.g. 11a.

Also, if a house has several buildings, the building type and number or letter is added.

The building type can be “корпус” (corpus) or “строение” (stroyenie), depending on whether or not it has a separate entry from the street, and it can also be a “владение” (vladeniye), if it is a custom-built structure or even not a building (e.g. construction site).

In some cases, a house can have multiple building types and numbers: “14 к2с1”

If a house stands on street intersection, it gets numbers from both streets, divided by a slash symbol, e.g. 5/16.

If a house was built in place of several demolished houses, its number is a range with a dash symbol: 7-11.

Everything above is only valid for Moscow, and in other cities houses are numbered differently.

Inconsistent phone numbers.

The phone numbers can have various formats, for example: +7 495 1234567, +7-495-123-45-67, +7(495)123-45-67.

The external code of Russia (+7), can be substituted for 8, the internal code for calling outside your phone zone.

Objects can have several phone numbers, separated by either a dot or semicolon.

1. Data Overview.

The basic statistics of the dataset and the MongoDB Queries.

Size:

Moscow - SW.osm – 55.2 MB

data.json – 60.3 MB

data\_cleaned.json – 57.8 MB

Number of documents: 261608

db.main.count()

Number of nodes: 217574

db.main.find({'type':'node'}).count()

Number of ways: 41391

db.main.find({'type':'way'}).count()

user\_data = db.main.aggregate([\

{'$project': {'user':1, '\_id':0}}, \

{'$group' : {'\_id' : '$user', 'count':{'$sum':1}}}, \

{'$sort' : {'count': -1}}, \

])

Number of unique users: 729

Top contributing user: name - Павел Гетманцев, entries - 17928.

Users, who made at least 1% of entries each: 22.

These users combined made 64% of the entries.

So, users, who made less than 1% of entries (less than 2616), together made 35% of the entries.

The 423 least contributing users combined made 1% of entries. Each of them made 25 entries or less.

1. Additional ideas.

The dataset can be used in any project that requires using