**1. Problems Encountered in the Map**

After downloading a small sample size of the Columbus area and running it against a python audit script, I noticed a number of problems with the data. The most significant of which were:

* Over­-abbreviated street names (“'S Court St'”)
* Incomplete Address Data
* Some postal codes logged Incorrectly ('postcode': 'Sunbury, OH 43074')

**Over-abbreviated Street Names**

As I was writing my shaping script in python, I would periodically test to make sure the code was returning the correct values. Through this process I realized that a large number of street names were overly abbreviated. I used a custom dictionary to update all the problematic address substrings, such that 'S Court St' became ‘South Court Street’.

**Postal Codes**

The same testing process as mentioned previously led me to discover another issue. Many postcode values were improperly logged. For instance, a number of values included street and state. Luckily this was a pretty easy fix with regex. The update process here, involved taking the entire sequence of numbers at the end of the problematic strings.

So

'postcode’: 'Sunbury, OH 43074'

Became

'postcode’: 43074

**Incomplete Address Data**

Unfortunately, there were many blank address values in the Columbus Dataset. To make matters slightly more frustrating, in many cases where an address value did exist, it was only partially filled out.

Ex:

{‘street’: 'Livingston Avenue', 'housenumber': '5000'}

{‘city’: 'Worthington', 'street': 'North High Street', 'housenumber': '7172'}

There wasn’t an easy fix to this issue, due to the fact that in order to complete the address data set, I would have to infer the non-existing values from the ones that were already imputed. As I believed this task outside the scope of this project I decided to leave it alone for now.

**2. Data Overview**

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

**File sizes**

Reduced\_columbus\_ohio.osm ………… 17.3 MB

Reduced\_columbus\_ohio.osm.json…….16.3 MB

Enter values from these Queries !!!

**#Number of documents**

>db.columbus.find().count()

**# Number of nodes**

>db.columbus.find({‘type’:’node’}).count()

**#Number of ways**

>db.columbus.find({"type":"way"}).count()

#**Number of unique users**

>db. columbus.distinct({"created.user"}).length

**#Top 1 contributing user**

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

**# Number of users appearing only once (having 1 post)**

>db.columbus.aggregate([{"$group":{"\_id":"$created.user", "count":{"$sum":1}}}, {"$group":{"\_id":"$count", "num\_users":{"$sum":1}}}, {"$sort":{"\_id":1}}, {"$limit":1}])

**3. Additional Ideas**

**Contributor statistics and gamification suggestion**

…..

….

….

….