Data Wrangling with Mongo DB

Analyzing OpenStreetMaps data

By Nikita Barsukov

Map Area: Kharkiv, Ukraine

# Introduction

This text is a grading paper for a Udacity’s course “Data Wrangling with MongoDB”. For this paper it was required to download an OpenStreetMaps dataset representing any part of the world, and perform a brief analysis of this dataset, according to given specifications. Minimum required size of an uncompressed dataset is 50 Mb.

I chose an area around Kharkiv, a large the city on the East of Ukraine, where I grew up. The dump is available on http://1drv.ms/1SnZXM3

I visualized some aspects of this dataset, made an overview of the dataset, tried to identify and, if possible, fix inconsistencies, and suggested ways to improve the quality of dataset. When I identified problems and directions for improvement, I also dataset for Kharkiv with two other data dumps for Stockholm and Copenhagen.

# Data Overview

File size of original data dump is 179,5 Mb. It contains 901.917 elements, earliest one was created in April 2007. Plot on Figure 1 below shows number of created elements in this dataset by month.

We can see that rate of adding elements to map of Kharkiv are is uneven. It was almost negligible during first three years, but it quickly started expanding after autumn 2010. We can also see a sharp spike in mid 2012 (August 2012 to be precise). A bit less than 230.000 elements were created during this month, more than a quarter of all the elements in the dataset.

When we look at creation times at a different angle and break it down just by month and day of week, we will also see an interesting picture (see figure 2 below).

The breakdown by month shows us that most of entries were made in August, which corroborates with the previous line chart. However, breakdown by weekday is more or less even, there is no spike on that chart.

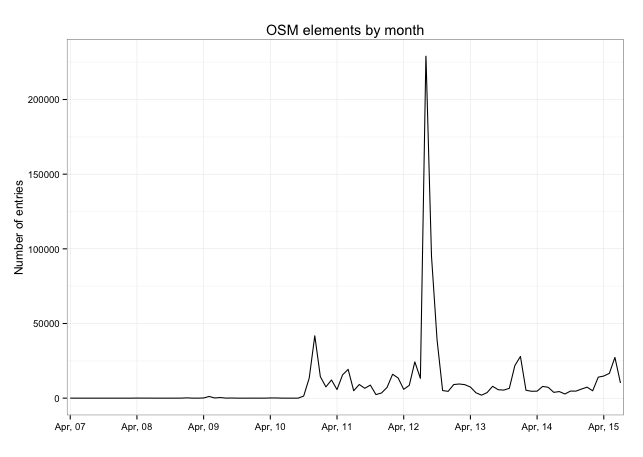
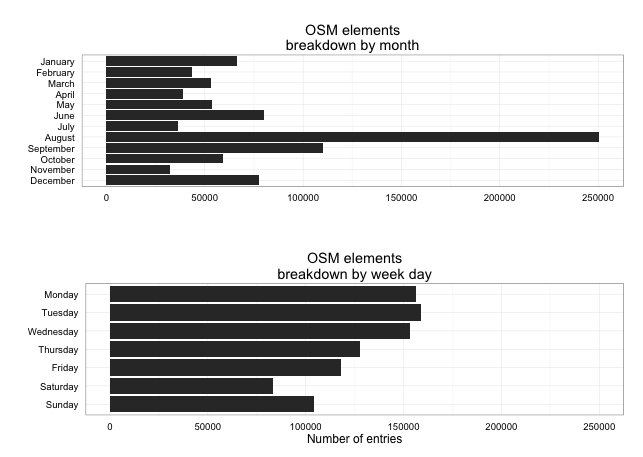


Figure 1



Figure