# OpenStreetMap Sample Project

### Data Wrangling with MongoDB

Dustin Kopp

Map area: Norman, OK, United States

Data source: https://overpass-api.de/api/map?bbox=-97.7303,34.8358,-96.9942,35.6562

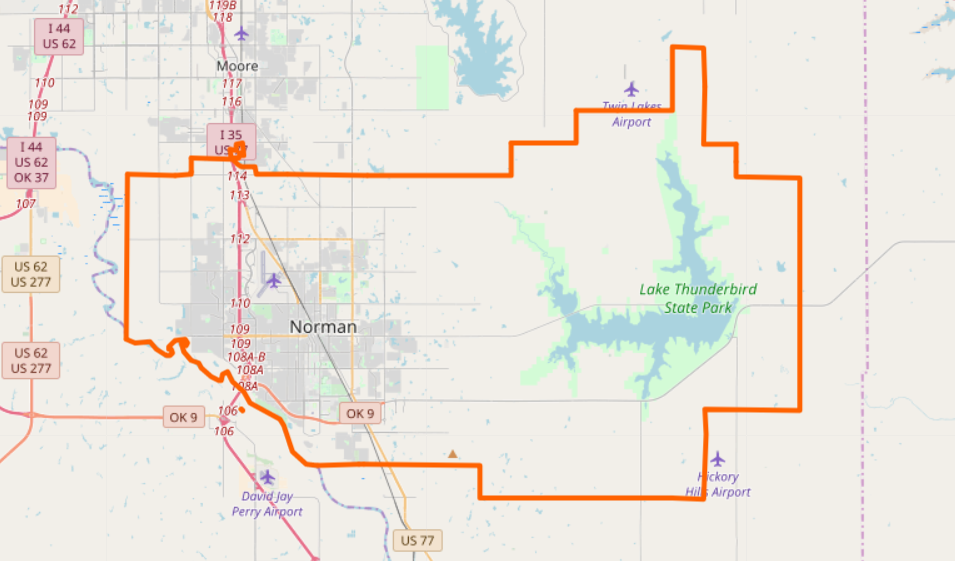


Fig 1. Norman, OK Selected area on OpenStreetMap

# Data Wrangling Process

In this project, I wrangled data obtained from OpenStreetMap about my home town of Norman, OK. I repeated a cycle of auditing and cleaning the data before beginning analysis using MongoDB.

Unfortunately, GitHub does not allow files over 100 MB, and the osm file I received for my area was ~126MB so I could not include it in the repository. It is accessible from the data source link above, and I have provided it here as well: [https:\\dustinkopp.com\norman.osm](https://dustinkopp.com/norman.osm). Once that was retrieved, I proceeded as follows:

First we audit the data. I executed “audit.py” which gathers information from the original osm file. It returns counts of all primary tags in the document, counts of all k tags in the document, as well as any irregular street names from it. It does not modify the file, that happens in a later process.

Next we parse out the document into JSON. I executed “parse.py” which updates all irregular street names and generates JSON objects from the xml data. These objects are written out to a norman\_parsed.json file.

After it was parsed we can do some further cleanup on the data.