# OpenStreetMap Project

Jill Andersen
Student ID: 001374500

## Map Area

Portland, Oregon, United States
https://www.openstreetmap.org/relation/186579

I lived in Portland and Hillsboro for several years. I thought it would be most helpful if I chose a location where I am familiar with the streets.

## Problems Encountered

I will discuss 2 problems that I encountered in the dataset

1. Inconsistent street names (i.e., St, St. Street, STREET). The updated names are written to the nodes_tag.csv file since these are both examples of nodes. The xml file remains the same.
    a. id: 4974324322, addr:street: NE 2nd Ave – Changed to NE 2nd Avenue
    b. id: 4894263410, addr:street: Southeast Main St. – Changed to Southeast Main Street

2. Inconsistent length of postcodes. Some are just the first 5 digits, whereas others are the complete 9 digits. Since I don't know the additional 4 digits, I thought it best to shorten it to include only the first 5 for uniformity.

## Overview of the Data

I used Microsoft Access, which is a SQL database to analyze the data files from the openstreetmap dataset I collected.

| | |
|---|---|
| map.osm | 57.7MB |
| nodes.csv | 21.3MB |
| nodes_tags.csv | 2.73MB |
| ways.csv | 1.88MB |
| way_nodes.csv | 6.4MB |
| ways_tags.csv | 5.51MB |
| # unique users | 348 |

## *Queries*

### ***Nodes_tags.csv queries:***

*Number of nodes_tags:* *69,593*

```
SELECT nodes_tags.key
FROM nodes_tags
ORDER BY nodes_tags.key DESC;
```

*Number of unique keys:* *344*

```
SELECT nodes_tags.key
FROM nodes_tags
GROUP BY nodes_tags.key;
```

*Number 'street' keys:* *4,061*

```
SELECT nodes_tags.key
FROM nodes_tags
WHERE (((nodes_tags.key)="street"))
ORDER BY nodes_tags.key DESC;
```

*Number of "postcode"keys:* *3.976*

```
SELECT nodes_tags.key
FROM nodes_tags
WHERE (((nodes_tags.key)="postcode"))
ORDER BY nodes_tags.key DESC;
```

## ***Nodes.csv queries:***

*Number of unique users:* *348*

```
SELECT nodes.uid
FROM nodes
GROUP BY nodes.uid;
```

*Number of nodes:* *227,670*

```
SELECT nodes.id
FROM nodes;
```

*Changes made:* *04/2008 – 11/2021*

```
SELECT nodes.timestamp
FROM nodes
ORDER BY nodes.timestamp DESC;
```

## ***Ways.csv queries:***

*Number of ways:* *29,127*

```
SELECT ways.id
FROM ways;
```

*Number of unique users:* *458*

```
SELECT ways.user
FROM ways
GROUP BY ways.user;
```

*Number of versions:* *62*

```
SELECT ways.version
FROM ways
GROUP BY ways.version;
```

*Changes made:* *06/2009 - 11/2021*

```
SELECT ways.timestamp
FROM ways
ORDER BY ways.timestamp DESC;
```

## ***Ways_nodes.csv queries:***

*Number of ways nodes:* *256,403*

```
SELECT ways_nodes.id
FROM ways_nodes;
```

*Number of unique ways_nodes positions:* *1,133*

```
SELECT ways_nodes.position
FROM ways_nodes
GROUP BY ways_nodes.position;
```

### ***Ways_tags.csv queries:***

*Number of ways tags:* *153,230*

```
SELECT ways_tags.id
FROM ways_tags;
```

*Number of unique keys:* *406*

```
SELECT ways_tags.key
FROM ways_tags
GROUP BY ways_tags.key;
```

*Number of unique values:* *13,232*

```
SELECT ways_tags.value
FROM ways_tags
GROUP BY ways_tags.value;
```

*Number of unique types:* *82*

```
SELECT ways_tags.type
FROM ways_tags
GROUP BY ways_tags.type;
```

*Number of ways amenities:* *58*

```
SELECT ways_tags.key, ways_tags.value, ways_tags.type
FROM ways_tags
GROUP BY ways_tags.key, ways_tags.value, ways_tags.type
HAVING (((ways_tags.key)="amenity") AND ((ways_tags.type)="regular"));
```

## Other Ideas about the Dataset

I audited inconsistent street names and postcodes, but there are a couple of other areas that are full of inconsistencies as well. For example, longitude and latitude, phone numbers and street numbers. Street numbers can be sorted to see if there are any outliers, for example, a street that contains houses such as: 300, 320, 340, 400, 900. You would need to research actual house numbers to determine if 400 or even 900 are correct for that particular street. There could also be additional audits to determine if any street numbers are included in the street address. (See example 1a. under Problems Encountered.)

The biggest inconsistencies though are the ‘key’, ‘value’ and ‘type’. They are generic tags for a wide variety of items. It would be beneficial to review them for similar items that can be cleaned.

### *Benefits*

The benefits of auditing the house numbers or building numbers would be especially helpful to all anyone who uses OSM for directions.  We’ve all experienced frustrations over incorrect directions!

### *Anticipated Problems*

However, auditing the house and building numbers would probably require someone local to physically got to each street with possible outliers to verify the veracity the house or building numbers or to confirm that it is an outlier. Otherwise, determining outliers would be a guessing game and could eliminate legitimate addresses.

### *Additional Data Exploration*

### ***Amenities:***

*Number of node amenities:* *75*

```
SELECT nodes_tags.key, nodes_tags.value, nodes_tags.type
FROM nodes_tags
GROUP BY nodes_tags.key, nodes_tags.value, nodes_tags.type
HAVING (((nodes_tags.key)="amenity"));
```

*Number of ways amenities: 58*

```
SELECT ways_tags.key, ways_tags.value, ways_tags.type
FROM ways_tags
GROUP BY ways_tags.key, ways_tags.value, ways_tags.type
HAVING (((ways_tags.key)="amenity") AND ((ways_tags.type)="regular"));
```

*Number of telephones still available: 23*

```
SELECT nodes_tags.key, nodes_tags.value, nodes_tags.type
FROM nodes_tags
WHERE (((nodes_tags.key)="amenity") AND ((nodes_tags.value)="telephone") AND
((nodes_tags.type)="regular"));
```

*Number of charging stations: 15*

```
SELECT nodes_tags.key, nodes_tags.value, nodes_tags.type
FROM nodes_tags
WHERE (((nodes_tags.key)="amenity") AND ((nodes_tags.value)="charging_station"));
```

*Number of fast food restaurants: 32*

```
SELECT ways_tags.key, ways_tags.value, ways_tags.type, Count(ways_tags.value) AS CountOfvalue
FROM ways_tags
GROUP BY ways_tags.key, ways_tags.value, ways_tags.type
HAVING (((ways_tags.key)="amenity") AND ((ways_tags.value)="fast_food"))
ORDER BY Count(ways_tags.value) DESC;
```

## Conclusion

I audited inconsistent street names and postcodes, but there are a couple of other areas that are full of inconsistencies as well. For example, longitude and latitude, phone numbers and street numbers. Street numbers can be sorted to see if there are any outliers, however, that could require additional research to check for accuracy.