OpenStreetMap Data Wrangling | Case Study - Querying

Answering questions

Who are the top 3 users who brought the more modifications?

```
SELECT DISTINCT nodes.user, COUNT(*)
FROM nodes
GROUP BY nodes.uid
ORDER BY COUNT(*) DESC
LIMIT 3;
```

```
[('crschmidt', 1198858), ('jremillard-massgis', 198009), ('OceanVortex', 84754)]
```

The 3 top users are:

1. `crschmidt`
2. `jremillard-massgis`
3. `OceanVortex`

Which proportion of the database did the top 10 users contribute to build?

```
SELECT DISTINCT nodes.user, COUNT(*) * 100.0 / (SELECT COUNT(*) FROM nodes)
FROM nodes
GROUP BY nodes.uid
ORDER BY (COUNT(*) * 100.0 / (SELECT COUNT(*) FROM nodes)) DESC
LIMIT 10;
```

```
[('crschmidt', 61.80088170765906), ('jremillard-massgis', 10.207322957391003),
('OceanVortex', 4.369051153890566), ('wambag', 4.077021576681348), ('morganwahl',
3.4048638260668747), ('MassGIS Import', 2.8826128734267003), ('ryebread',
2.6775477969680472), ('ingalls_imports', 1.4554053050922948), ('Ahlzen',
1.3437484535062107), ('mapper999', 0.6402999785552861)]
```

These are the top 10 contributors with the percentage of the total dataset they modified:

1. crschmidt - 61.80%
2. jremillard-massgis - 10.21%
3. OceanVortex - 4.37%
4. wambag - 4.08%
5. morganwahl - 3.40%
6. MassGIS Import - 2.88%
7. ryebread - 2.68%
8. ingalls_imports - 1.46%
9. Ahlzen - 1.34%
10. mapper999 - 0.64%

The top 10% contributors account for 92.86% of all the modifications.

What are the top 10 streets that contain the more nodes?

```
SELECT ways_tags.value, COUNT(*)
FROM ways_tags
WHERE ways_tags.key = 'name'
AND ways_tags.type = 'regular'
GROUP BY ways_tags.value
ORDER BY COUNT(*) DESC
LIMIT 10;
```

```
[('Washington Street', 253), ('Massachusetts Avenue', 157), ('Centre Street',
136), ('Broadway', 118), ('Beacon Street', 116), ('Cambridge Street', 97),
('Boylston Street', 87), ('Adams Street', 86), ('Blue Hill Avenue', 84),
('Northeast Corridor', 83)]
```

The top 10 streets containing the more nodes are:

1. Washington Street
2. Massachusetts Avenue
3. Centre Street

4. Broadway
5. Beacon Street
6. Cambridge Street
7. Boylston Street
8. Adams Street
9. Blue Hill Avenue
10. Northeast Corridor

On average, how many nodes does a way contain?

```
SELECT AVG(Count)
FROM
    (SELECT COUNT(*) as Count
    FROM ways
    JOIN ways_nodes
    ON ways.id = ways_nodes.id
    GROUP BY ways.id);
```

```
[(7.5277535169279854,)]
```

A way contained an average of 7.53 nodes.

What are the top 10 amenities in Boston?

```
SELECT value, COUNT(*) as Count
FROM nodes_tags
WHERE key='amenity'
GROUP BY value
ORDER BY Count DESC
LIMIT 10;
```

```
[('bench', 1069), ('restaurant', 678), ('school', 499), ('bicycle_parking', 318),
```

```
('library', 276), ('place_of_worship', 275), ('cafe', 270), ('fast_food', 195),
('bicycle_rental', 139), ('post_box', 124)]
```

The top 10 amenities in Boston are:

1. Benches
2. Restaurants
3. Schools
4. Bicycle_parking
5. Libraries
6. Place of worship
7. Cafe
8. Fast food
9. Bicycle rental
10. Post box

There's definitely room to seat in Boston!

What's the religious landscape like in Boston?

```
SELECT nodes_tags.value, COUNT(*) as Count
FROM nodes_tags
JOIN
    (SELECT DISTINCT(id)
    FROM nodes_tags
    WHERE value='place_of_worship') as Sub
ON nodes_tags.id=Sub.id
WHERE nodes_tags.key='religion'
GROUP BY nodes_tags.value
ORDER BY Count DESC;
```

```
[('christian', 245), ('jewish', 8), ('unitarian_universalist', 2), ('buddhist',
1), ('muslim', 1)]
```

Places of worship in Boston are christian, jewish, unitarian universalist, buddhist, and muslim.

Irish descendants enjoying a strong presence in Boston, it would have been great to have a distinction between catholics and protestants.

Further Analysis

- I focused this analysis on identifying irregular street names, cleaning them, and reformatting them so that one type corresponds to one term (Avenue, Street), getting rid of abbreviations.
- To improve the cleaning result, the process should also include the insertion of tags. Some street names contained valuable information about floors, for example, or suite numbers, that deserve to be preserved. This process would enable us to present organized information at their specific place, preserving both the street names and the Floor or Suite numbers.
- This data wrangling process should probably be pushed a step further in order to verify and clean zip codes as well, before being sent back to OpenStreetMaps. Zip codes could present similar issues than the ones encountered with street names:
    - Too many or too few numbers
    - Zip codes at the wring place (not in the right tag)
    - Zip codes containing irregular characters
    - Zip codes corresponding to another information that should be preserved but given elsewhere