I have chosen the location of Nashville, TN for my analysis. The reason for choosing Nashville is I go to Nashville regularly living in the South plus it's a very vibrant and a growing city.

The OSM file is 300 MB uncompressed.

Link to the map: - Nashville map

There are 3 main problems that I saw with the file which are: -

1) Street names: - There are a lot of street names which are abbreviated which can be fixed. Examples are "**Murfreesboro Rd", "N Woodland St"**
2) City Names: - City Names sometimes have state names or some county names associated with it. Examples are "**Antler, Tennessee**", "**Nashville-Davidson**".
3) Some postal codes have 9-digit numbers in it instead of 5. Examples are "**37174-7436**".

In the following sections I will explain how I fixed the Street and city names.

Street Names

To fix the street names I used the following function in the audit.py file.

```
def update_name(name, mapping):

    m = street_type_re.search(name)
    if m.group() in mapping.keys():
        if m not in expected:
            name = re.sub(m.group(), mapping[m.group()], name)

    return name
```

This function takes in all the street names that are abbreviated and fixes them using the following mapping dictionary.

```
mapping = {"St": "Street",
           "St.": "Street",
           "Rd.": "Road",
           "Rd":"Road",
           "Ave":"Avenue"
           }
```

City Names

The way I fix city names is using the following function in the audit.py file

```
def clean_c(city):
   if ',' in city:
      city = city.split(',')[0]
   elif '-' in city:
      city = city.split('-')[1]
   return(city.title())
```

Some general Statistics of the data

Number of unique users: 1072

Query: SELECT distinct user from nodes
UNION
SELECT distinct user from way

Number of nodes: 1341353

Query: SELECT count(*) from nodes

Number of ways: 138923

Query: SELECT count(*) from way

9-digit postal codes
I tried investigating how may postal codes where there that were 9 digits. There weren't that many actually. A total of only 22 were present which means that data largely has 5 digits for postal codes.

```
Query: SELECT tags.value, COUNT(*) as count
        FROM (SELECT * FROM node_tags
                    UNION ALL
              SELECT * FROM way_tags) tags
        WHERE tags.key='postcode'
        AND tags.value  like '%-%'
        GROUP BY tags.value
        ORDER BY count DESC;
```

Result:

Postal code	Count
37174-6120	8
37174-7436	6
37243-0468	2
37207-4405	1
37209-1057	1
37235-1826	1
37243-0470	1
37243-0471	1
38451-2074	1

Top 10 users by count of records contributed

Query: SELECT e.user, COUNT(*) as num
FROM (SELECT user FROM nodes UNION ALL SELECT user FROM way) e
GROUP BY e.user
ORDER BY num DESC
LIMIT 10

user	num
woodpeck_fixbot	276324
Shawn Noble	177953
st1974	96258
AndrewSnow	55783
Rub21	53782
TIGERcnl	52403
StevenTN	29582
darksurge	27550
dchiles	26679
42429	26652

Additional suggestions:

I think the data can be improved by adding better key values. If we look at the following results where the key is cuisine it doesn't really capture the actual cuisine type all the time. For example, "ice_cream" and "sandwhich" are not cuisines but types of food.

If we can improve this that will give us a better understanding what the actual cuisines are and separate cuisines from the food type.
Some problems that might come up is people often don't care that much when contributing to open source so to make them follow best practices would be harder than if they were given some kind of recognition as on a leaderboard where there achievements are celebrated based on the accuracy of their contribution.

Top 10 cuisines

Query:
```
select value, count(*) ct
    from node_tags
    where key ='cuisine'
    Group BY value
    ORDER BY ct desc
    LIMIT 10
```

cuisine	count
burger	42
mexican	39
sandwich	23
coffee_shop	21
pizza	17
american	16
chicken	12
ice_cream	12
regional	10
japanese	9