Initial Audit:

To begin, I used ElementTree to gather the initial information about the Victoria, Canada dataset, counting the number of unique tags using parse_map.py. The results were as follows:

```
{'bounds': 1,
 'member': 31168,
 'nd': 1534552,
 'node': 1408866,
 'osm': 1,
 'relation': 4014,
 'tag': 609568,
 'way': 118469}
```

I then used tags.py to analyze the patterns of the k tags. Using regular expressions, I searched f or the number of tags in the following categories.

```
{'lower': 392751, 'lower_colon': 207715, 'other': 9101, 'problemchars': 1}
```

Lower represents tags that contain only lowercase letters and are valid

Lower_colon represents tags with a colon in their name that are nevertheless valid

Problemchars represents tags with a problematic character

Other is the catchall category for tags that did not match the first three categories

Problems and Challenges in the Data:

I started by auditing the data using Audit_Postal_Code.py and Audit_Street_Name.py to get a sense of the problems in the Postal Codes and Street names respectively.

This was a fairly consistent and accurate data set, but there were a few issues:

- Inconsistent Street Names: Most street names were formatted consistently, but there were a few key inconsistencies.
    1. St and St. were used for Street
    2. Dr was used for Drive
    3. Rd was used for Road

- Mistakes in Street Names: There were almost non-existent, but the data did have a few mistakes.
    1. MainStreet was used instead of Main Street
    2. Deerpath was used for Deerpath Road
- Zip Codes: This data set includes locations from two countries with two quite different styles of zip code. Five digits for the US and 6 characters, alternating numbers and letters, for Canada.
    1. There were a few mistakes or inconsistencies in the US zip codes. One zip code had the city name, Port Angeles instead of the zip code. Another had the state abbreviation WA at the beginning. There were also a few long-form zip codes
    2. The major difference with Canadian zip codes included differences between capitalized and lower case letters and whether or not the characters were separated by a space or a dash. To standardize this I chose capital letters with no space or dash in between – for example V8P1A1
- Additional Street Name Quirks
    1. Along with the problems listed above, there were a few additional quirks in the data. St. is used not just for street, but also as a (correctly formatted) abbreviation for Saint.
    2. Additionally, the data set involved some *very* obscure categories of street name that required additional research to verify, for example 'Wynd', which is apparently a narrow lane between houses and commonly used in Scotland and North England and quite rare elsewhere.

I used Clean_OSM_File.py to clean the data, fixing street names and zip code errors, which output a new, cleaned OSM file, which I next converted to CSV with Convert_To_CSV.py.

File Sizes:

The sizes of the files used are as follows:

- victoria_canada.osm: 284 MB
- nodes_csv: 113 MB
- nodes_tags.csv: 8.75 MB
- ways_csv: 6.86 MB
- ways_nodes.csv: 12.1 MB
- ways_tags.csv: 36.6 MB
- victoria.db: 205 MB

Query Results:

Once the auditing was finished and the data was converted from XML to CSV format I loaded the CSV files into an SQL database using Create_Database.py. I then used sqlite online (https://sqliteonline.com/) to run queries on the database. Since I did not use a savable py file, I include the code for my queries here.

Number of Nodes:

```
SELECT COUNT(*) FROM nodes
```

1408866

Number of Ways:

```
SELECT COUNT(*) FROM ways
```

118469

Number of Unique Users:

```
SELECT COUNT(DISTINCT(e.uid))
FROM (SELECT uid FROM nodes UNION ALL SELECT uid FROM ways) e;
```

973

Number of Users Contributing only Once:

```
SELECT COUNT(*)
FROM
   (SELECT e.user, COUNT(*) as num
    FROM (SELECT user FROM nodes UNION ALL SELECT user FROM ways) e
    GROUP BY e.user
    HAVING num=1) u;
```

171

Top 10 Users:

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

Alester	707485
Hai-Etlik	151137
Alester_imports	80735
Tylerritchie	52240
Glassman	36524
Oceanearth	30609
Petersfreeman	23858
Woodpeck_fixbot	19266
Madrona	17092
Kc12345	16854

Top 10 Postal Codes:

```
SELECT tags.value, COUNT(*) as count
FROM (SELECT * FROM nodes_tags
            UNION ALL
    SELECT * FROM ways_tags) tags
WHERE tags.key='postcode'
GROUP BY tags.value
ORDER BY count DESC
LIMIT 10
```

98368	47
V9B1L8	46
98250	44
98221	40
98362	29
V8K1V5	25
98239	22
98248	16
V8R1L6	13
98277	12

Top 10 Amenities:

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

Bench	1320
Waste Basket	511
Toilets	324
Restaurants	248
Café	188
Parking	176
Bicycle Parking	162
Fuel	118
Drinking Water	116
Fountain	97

Top 10 Cuisines:

```
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
  JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='restaurant') i
  ON nodes_tags.id=i.id
WHERE nodes_tags.key='cuisine'
GROUP BY nodes_tags.value
ORDER BY num DESC
LIMIT 10;
```

Regional	19
Thai	14
Chinese	12
Mexican	12
Pizza	12
Sushi	12
Fish and Chips	6
Italian	6
American	5
Asian	5

Top Religions:

```
SELECT nodes_tags.value, COUNT(*) as num
FROM nodes_tags
   JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='place_of_worship') i
   ON nodes_tags.id=i.id
WHERE nodes_tags.key='religion'
GROUP BY nodes_tags.value
```

Christian	29
Sikh	2
Eckankar	1

Conclusions and Suggestions for Future Analysis:

The data seemed to be fairly consistent with few mistakes in this extract. There were a few errors to be found, especially in the occasional inconsistent abbreviation of a word (rd, st, dr), or mistyped street name (MainStreet, Deerpath).

This data was made trickier by the fact that it included information from two countries with different conventions. This was most apparent with postal codes, which have very different formats depending on country. Both countries seemed to have the most editing needed to postal codes, with more mistakes or issues on the US side (the city name Port Angeles listed in place of the zip code being the most glaring example.)

I also noticed that within the top users Alester and Alester_Imports seem to be the same person. Upon checking the Open Street Map wiki, Alester_Imports seems to be the name Alester uses when he is importing data from somewhere else:

"I normally edit under the username alester, but I've created this account for the purpose of importing data."

http://wiki.openstreetmap.org/wiki/User:Alester_imports

Interestingly, when I analyzed the database using SQL, I found that the more common zip codes in the top ten were American. Only three out of ten were Canadian. When I ran the entire list without limiting the number, there were far more Canadian Postal Codes, most appearing only time. This makes sense, however, as Canadian postal codes cover far smaller areas than their

American cousins. All the same it would be interesting to have a metro extract that covered more Canadian territory. This may be problematic as Victoria is also close to another major Canadian metropolitan area, Vancouver.

I was also surprised to find that the top amenity was benches. Victoria is a coastal city with a harbor and there are many places to sit, but the sheer number, especially compared to other amenities seemed a bit odd. I also noted that while cafes were considered separate from restaurants, they were quite frequent, something I would definitely expect from this data.

The top cuisine was regional, with all the vagueness that term suggests. However, most of the other cuisines in the top ten are Asian in some way, a definite trend in this corner of the pacific. Victoria has a famous Chinatown, which would make me think the number of Chinese restaurants should be higher. Perhaps the data is incomplete. Victoria is also famous for its fish and chips made with fresh caught fish, which show up in the top ten. Personally, I would count cafes as restaurant, and while the data is organized differently from my expectations, it was done so systematically and consistently.

The top religion was not surprising, but what did surprise me a little was the general lack of other kinds of places of worship. The Pacific Northwest is not the most religious place, but I would have expected a few other kinds of religious building. I would guess that they do exist, but have not made it into the data yet. There is a sizable Sikh community in Victoria so the two Sikh gurdwaras weren't surprising. I did have to look up Eckankar, which seems to be a new religion, founded in Minnesota in the 1960s. Needless to say, this was the kind of unexpected information this sort of analysis brings up.

All in all this data seems fairly well organized and consistent, if not a bit incomplete in places. I would also like to see a map extract with less or no American data included. After all, a user interested in Victoria is unlikely to need to know about places in Port Angeles, Port Townsend or Anacortes. This may be a case of easier said than done, however.