Introduction

I used the street map data from my hometown Bellevue, Washington (https://www.openstreetmap.org/export#map=13/47.6065/-122.1647) which is a 69.7 MB OSM file.

I sought to analyze the following:
- The number of food establishments
- The most proliferate establishment (fast food or resstaurant)
- The cuisines with the most number of establishments
- The number of zip codes contained in the map area
- The top contributors to the map

Problems Encountered in the Map

In analyzing the map, I encountered some problems:

1. Abbreviated street names

The road names lacked consistency as can be seen below with this sample set of addresses:

```
'NE': set(['105th Avenue NE',
           '106th Ave NE',
           '107th Avene NE',
           '108th Ave NE',
           '108th Avenue NE',
           '110th Ave NE',
           '110th Avenue NE',
           '111th Ave NE',
           '111th Avenue NE',
           '112TH AVE NE',
           '112th Av NE',
```

To clean this, I did the following:

- Audit the map to inspect the titles of the roads
- Create a dictionary with the titles mapped to a consistent road title, e.g. Ave. or ave -> "Avenue"
- Write a function so that when I process the map to create a JSON file, I search the street names for any words that appear in the dictionary and replace them with the consistent naming convention
- One difficulty I had to overcome was the issue that the street names do not appear in a consistent place in the address line: sometimes the street name was the last word and other times it was a word in the middle of the string. So, I had to create a method that accommodated this and didn't rely on the placement of the word

This resulted in the following consistent data:

'NE': set(['105th Avenue NE',
'106th Avenue NE',
'107th Avenue NE',
'108th Avenue NE',
'108th Avenue NE',
'110th Avenue NE',
'110th Avenue NE',
'111th Avenue NE',
'111th Avenue NE',
'112TH Avenue NE',
'112th Avenue NE',

2. Invalid tag data

I inspected the map tags to make sure the tags contained the right data or the right format that I would expect but I found that 5 tags contained "testing point" as its value, which was not what was expected. So, I ignored importing this tag into the JSON file.

3. Inconsistent zip codes

The zip code or postcode information lacked consistency as can be seen below from auditing the data:

98039
WA
98004-5002
98036
98033
99004
98118
NE
98004-5983
98033-7722
98004-4452
98059
98004-5903
…

Some values are strings, some are in the format of “xxxxx-xxxx,” and others are in the format “xxxxx.”

Additionally, some zip codes appear under another tag: “tiger:zip_left” or “tiger:zip_right,” which indicate the boundaries of a labeled “way.” Sometimes, these ways would span multiple zip codes and hence required consolidating the information to a single zip code for consistency purposes.

Therefore, to clean the data, I had to do the following:

- Account for the different tagging for Way and Nodes: zip_left, zip_right vs. “addr:postcode”

- Find the Way postcodes and take the first five digits to match the format of "xxxxx"
- Find the Node postcodes and do the following:
  - Take the zipcode as is if it matched the format "xxxxx"
  - Take the first five digits to match the format "xxxxx" and delete the remaining digits
  - Delete the data if it contained letters

This resulted in zip code data being cleaned to be consistent:

98039
98004
98036
98033
99004
98118
98004
98033
98004
98059
98004

Overview of the Data

Once in JSON format, the data becomes easier to read. Each document contains information about points on the map.
For instance, in the below document, we can see that this point represents a Burger King that user "goldfndr" created.
These documents then determine the points of interest or the information that goes on the map.


```
{
 "cuisine": "burger",
 "amenity": "fast_food",
 "name": "Burger King",
 "created": {
  "changeset": "814784",
  "user": "goldfndr",
  "version": "3",
  "uid": "7247",
  "timestamp": "2009-01-20T03:03:12Z"

 "pos":
  47.616762,
  -122.1833799
 ],
 "created_by": "Potlatch 0.10f",
 "source": "knowledge",
 "type": "node",
 "id": "333392511"
}
```


Analysis of the Data

I analyzed the map to look at the following information:

- The number of food establishments
- The most proliferate establishment (fast food or resstaurant)
- The cuisines with the most number of establishments
- The number of zip codes contained in the map area
- The top contributors to the map

My results are as follows:

- The number of food establishments
  67 Total fast food establishments
  246 Total restaurants
  313 Total number of food establishments

- The most proliferate establishment (fast food or restaurant):
  Subway with 10 establishments
  McDonald's with 5 establishments
  Red Robin with 4 establishments
  Burger King with 3 establishments
  Jimmy John's with 3 establishments

- The cuisines with the most number of establishments:
  burger with 18 establishments
  mexican with 16 establishments
  pizza with 14 establishments
  sandwich with 13 establishments
  148 establishments not labeled

- The number of zip codes contained in the map area:
  There are 19 zip codes:
    None,
    98004
    98005
    98006
    98007
    98008
    98027
    98033
    98036
    98039
    98040
    98052
    98053
    98056
    98059
    98105
    98115
    98118
    99004

- The top contributors to the map:
  Glassman_Import – 55,175 contributions
  zephyr – 44,502 contributions
  sctrojan79-import – 32,251 contributions

Glassman – 31,432 contributions
STBrenden – 25,690 contributions

Other Ideas About the Dataset

Additional analysis
With more time and ability, I'd be curious to inspect the following:

- Information about the top contributors, e.g. are they knowledgeable about a certain area (e.g. zipcode 98006) or a specific type of information (e.g. highways or restaurants)
- Correlation between number of fast food establishments and schools
- The density of religious institutions and schools

With even more time, I would be interesting to compare some of these characteristics against the same characteristics in other map areas

Preventing inconsistent data
To prevent the type of data cleaning I had to do, we could have put in place methods to ensure data consistency from the contributors. For example:

- Writing a policy or standard formatting data sheet
- Employing rules to check for data consistency, e.g. street names matched dictionary values or zip codes followed the same format

I realize this may require additional work or burden for the contributors and will require balancing the ease of use with the need for consistent data. Also, doing this will require some central administrative effort, e.g. to create and maintain the rules, and may be more than the sponsoring organization will bear.

Filling in additional data
We could also use another data source to help fill in missing information. For example, we could overlay the postcode latitude and longitude data to fill-in the postcode information. This would be especially helpful if we wanted to the extended format "xxxxx-xxxx" since many contributors will not know the latter four digits of a postcode.

Again, this requires some central administration to create and to maintain and may not be what the sponsoring organization wants to invest in. Additionally, this will assume that the other data source is accurate to the latitude and longitude, which it may not be.