# OpenStreetMap Data Case Study

Made By- Prakhar Gupta

The city which I have chosen is **San Jose** . The reason for me to chose San Jose is that its like the capital of Silicon Valley. Since i was a child i wanted to work in Silicon Valley ,hoping some day i will work there.

Reference link to Mapzen- https://mapzen.com/data/metro-extracts/metro/san-jose_california/

## Problems Faced

1. I hard to view such OSM files on a notepad(windows platform). All tags appear to be merged into each other so I turned to Sublime Text to view the files

2. I had no prior knowledge of  how naming is in US Streets so I ran multiple iterations on audit.py file and updated the mapping dictionary so that better data cleaning could be possible

3.Thanks to Sublime Text I found that some phone numbers were not in order and saw a chance to clear that part too. I found through googling that US uses this format

Reference=https://www.timeanddate.com/worldclock/dialingcodes.html?p1=1794&p2=283&number=

+1 (XXX)XXX-XXXX

Where +1 is the US Code and (408) is the code for San Jose

4. For postal codes apparently US use two type of formats

a.Statecode zip5

b.Statecode zip5-zip4

Reference =http://mentalfloss.com/article/53384/what%E2%80%99s-deal-those-last-4-digits-zip-codes

# Audit and Cleaning

- Lat_Long_check.py checks for incorrect entries in the Latitude and Longitude. Since the OSM file was downloaded from mapzen it would be wise to check for and incorrect entry and the last thing we want to end up with is node or way in the data which does not belong to the area which we are interested in
- audit.py file is used for cleaning the data . I started with a few values in Mapping and Expected dict and list and added more for better cleaning. It also cleans the phonenumber and places the Postcode in order

  Few examples of cleaning are

| Before | After |
|---|---|
| Palm Valley Blvd - | Palm Valley Boulevard |
| San Antonio Valley Rd - | San Antonio Valley Rd |
| Cherry Ave - | Cherry Avenue |
| 95014-0454 - | CA 95014-0454 |
| 95014 - | CA 95014 |
| +1-408-588-4045 - | +1(408)588-4045 |
| 1.408.245.5620 - | +1(408)245-5620 |
| +1 408 736 3726 - | +1(408)736-3726 |
| 4088718765 - | +1(408)871-8765 |

- The sanjose+csv.py file uses above audit file and converts it into 5 csv files

# File Size

san-jose_california.osm = 349MB

sample.osm = 3.53MB

nodes.csv = 135MB

nodes_tags.csv = 3.07MB

way.csv = 13.3 MB

ways_nodes.csv = 46.6MB

ways_tags.csv = 21MB

san_jose.db = 254MB

# Queries

## Number of nodes-

QUERY = '"SELECT count(*)as num from nodes;'"

*Number of Node (1685948,)*

## Number of ways-

QUERY = '"SELECT count(*)as num from ways;

*Number of ways (230633,)*

## Number of Unique Users

QUERY = '"SELECT COUNT(DISTINCT(user)) from (select user from nodes UNION ALL select user from ways);'"

*No of unique users (1373,)*

## Top 10 contributers

QUERY = '"SELECT user,count(user)  from (select user from nodes UNION ALL select user from ways)
group by user
order by  count(user) desc
limit 10;

Top 10 contributers

| | 0 | 1 |
|---|---|---|
| 0 | andygol | 295612 |
| 1 | nmixter | 284955 |
| 2 | mk408 | 147242 |
| 3 | Bike Mapper | 91106 |
| 4 | samely | 81084 |
| 5 | RichRico | 76205 |
| 6 | dannykath | 74460 |
| 7 | MustangBuyer | 65043 |
| 8 | karitotp | 63527 |
| 9 | Minh Nguyen | 53156 |

# More Data Explorations

## Top 5 amenities  available in San Jose

QUERY = '"SELECT value,count(*)as num from (select value,key from nodes_tags  UNION ALL select value,key from ways_tags)
where key='amenity'
group by value
order by num desc
limit 5;"'

```
Top 5 amenities  available in San Jose
                   0     1
0           parking  2108
1        restaurant  1049
2         fast_food   534
3            school   534
4  place_of_worship   354
```

## Common Fast Food Chains in San Jose

query= '"SELECT value,COUNT(*) FROM nodes_tags WHERE value='Starbucks'or value="McDonald's" or value='Taco Bell' or value='Subway' or value ='Burger King' group by value ;"'

```
Common Fast Food Chains in San Jose
             0   1
0  Burger King   8
1   McDonald's   9
2    Starbucks  94
3       Subway  52
4    Taco Bell  12
```

# Conclusion

Although the San Jose area  dataset is of fairly reasonable quality and I have cleaned the data which is required for this project

There are still many typos which are caused due to human input. Some are due to a lack of quality or no convection like-

query= '"SELECT value,COUNT(*) FROM nodes_tags WHERE value='Starbucks/ Target Cafe' ;"'

```
                          0  1
0  Starbucks/ Target Cafe  1
```

We can see its a valid starbucks but we missed it in our analysis because of improper naming convection of nodes and ways

The above limitations could be avoided by

- Developing a convection for naming while new entry is placed
- Program for error control
- Making use of manual entry through skilled user who can add entries in a designated manner
- Since Openstreet map is free it can use the google search api (free) version or google geocode api to compare the new inputs with the google once for better accuracy

  Limitations with above epproach
  - I was just a data of a city San Jose which was 349MB and imagining doing error control for ever city would be a mammoth task and require more funding
  - It will also require experienced professionals
  - Using google api have a problem Free apis have a limited access of around 500 per day entries above 500 would cause the  requesting program of API to crash or generate a no response