I worked with the OSM file for San Jose, CA, which is near where I live.

Here is the size of that file and the resulting cleaned files:

```
sanjose.osm ......... 287 MB
OSM.db .......... 808 MB
nodes.csv ............. 107 MB
nodes_tags.csv ........ 2.5 MB
ways.csv .............. 10   MB
ways_tags.csv ......... 21 MB
ways_nodes.cv ......... 35 MB
```

Some Queries and Results:

```
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 100;
```

```
Sunnyvale,37
"San Jose",8
"Morgan Hill",5
"Santa Clara",3
Saratoga,3
2,1
54,1
Cupertino,1
Milpitas,1
"San José",1
```

```
SELECT COUNT(*) FROM nodes;
7691549
```

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

1021757

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

1241

```
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;
```

```
nmixter,1730268
mk408,915108
"Bike Mapper",495300
samely,466572
dannykath,432264
RichRico,428760
karitotp,340860
n76,229458
matthieun,196626
"Minh Nguyen",196590
```

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

```
restaurant,3825
fast_food,1845
cafe,1120
place_of_worship,940
bicycle_parking,875
bench,870
school,760
toilets,735
```

```
fuel,615
bank,580
```

```
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;
```

```
chinese,305
vietnamese,295
mexican,255
pizza,245
japanese,195
indian,145
italian,130
thai,125
american,115
sushi,95
```

```
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;
```

```
95014,21023
95070,1365
94087,1142
94086,1050
95051,944
95037,933
95054,545
```

```
95127,538
95125,519
95050,476
```

Pulling out unexpected street types resulted in these:
Unexpected Street Types

6
Winchester
Ln
West
Rd
114
7.1
A
Hill
Way
ave
Circle
East
Alameda
Highway
Real
Expressway
20
Paviso
Boulvevard
1
Oro
Flores
Hamilton
Marino
Hwy
Volante
Dr
CA
robles
201
Esquela
0.1
Bellomy
Napoli
Sq
Plaza
Barcelona
St
Cir

Franklin
Palamos
Presada
Loop
Bascom
Sorrento
Portofino
Julian
Seville
Luna
street
Terrace
Madrid
Blvd
Ave
Row
Ct

And these explain what I needed to fix

6 > delete

<tag k="addr:city" v="Santa Clara"/>
<tag k="addr:housenumber" v="280"/>
<tag k="addr:postcode" v="95050"/>
<tag k="addr:state" v="CA"/>
<tag k="addr:street" v="Martin Avenue #6"/>

Winchester > Winchester Boulevard
<tag k="name" v="Toys "R" Us"/>
<tag k="shop" v="toys"/>
<tag k="source" v="www.toysrus.com"/>
<tag k="website" v="http://www.toysrus.com/"/>
<tag k="addr:street" v="Winchester"/>
<tag k="addr:postcode" v="95128"/>
<tag k="addr:housenumber" v="751"/>

Ln > Lane

West > OK
<tag k="height" v="6.7"/>
<tag k="building" v="yes"/>
<tag k="addr:street" v="Park Circle West"/
<tag k="addr:postcode" v="95014-1923"/>
<tag k="addr:housenumber" v="10227"/>

Rd > Road

114 > delete
<tag k="addr:street" v="West Evelyn Avenue Suite #114"/>

7.1 > Highway 17
<tag k="landuse" v="industrial"/>
<tag k="addr:county" v="Santa Clara"/>
<tag k="addr:street" v="Hwy 17 PM 7.1"/>
<tag k="attribution" v="Caltrans"/>
<tag k="description" v="District 4 - South Bay Region"/>
<tag k="caltrans:type" v="SL"/>
<tag k="caltrans:district" v="4"/>
<tag k="caltrans:dynsegpm" v="SCL 17 7.1"/>

A > couldn't find
Hill > OK
Way > OK
ave > Ave
Circle > OK
East > OK
Alameda > OK
<tag k="addr:street" v="The Alameda"/>
<tag k="addr:postcode" v="95050"/>
<tag k="internet_access" v="yes"/>
<tag k="addr:housenumber" v="2221"/>

Highway >OK
Real > OK
Expressway > OK
20 > delete
<tag k="addr:street" v="Coleman Ave #20"/>

Paviso > OK

<tag k="height" v="4.9"/>
<tag k="building" v="yes"/>
<tag k="addr:street" v="Via Paviso"/>
<tag k="addr:postcode" v="95014-6322"/>
<tag k="addr:housenumber" v="20415"/>

Boulvevard > Boulevard

1 > delete
<tag k="addr:street" v="Stewart Drive Suite #1"/>

<tag k="name" v="Cisco Systems Building 1"/>
<tag k="building" v="office"/>
<tag k="addr:city" v="San Jose"/>
<tag k="addr:street" v="Zanker Road"/>
<tag k="addr:postcode" v="95134"/>
<tag k="addr:housename" v="Cisco Systems Building 1"/>
<tag k="addr:housenumber" v="3850"/>

Oro > OK

<node id="3911854957" lat="37.232446" lon="-121.779851" version="3" timestamp="2015-12-27T03:05:52Z" changeset="36187711" uid="2786613" user="chris@zontine">
<tag k="name" v="California Sports Center"/>
<tag k="leisure" v="dance"/>
<tag k="addr:city" v="San José"/>
<tag k="addr:street" v="Via del Oro"/>
<tag k="description" v="Gymnastics and Dance"/>
<tag k="addr:postcode" v="95119"/>
<tag k="addr:housenumber" v="150"/>

Flores > OK

<tag k="name" v="Woodlands Healthcare Center"/>
<tag k="amenity" v="nursing_home"/>
<tag k="website" v="http://woodlandshealth.net/"/>
<tag k="addr:city" v="Los Gatos"/>
<tag k="addr:state" v="CA"/>
<tag k="addr:street" v="Terreno De Flores"/>
<tag k="addr:postcode" v="95032"/>
<tag k="addr:housenumber" v="14966"/>

<nd ref="65542394"/>
<nd ref="65542397"/>
<nd ref="65542399"/>
<nd ref="65542401"/>
<tag k="name" v="Avenida de Las Flores"/>
<tag k="highway" v="residential"/>
<tag k="tiger:cfcc" v="A41"/>
<tag k="tiger:tlid" v="122958252"/>
<tag k="tiger:county" v="Santa Clara, CA"/>
<tag k="tiger:source" v="tiger_import_dch_v0.6_20070809"/>
<tag k="tiger:name_base" v="Avenida de Las Flores"/>
<tag k="tiger:separated" v="no"/>
<tag k="tiger:upload_uuid" v="bulk_upload.pl-9f300d22-5de3-4867-

bd5e-8c2a200c22ad"/>

Hamilton > Mt Hamilton Rd

<node id="1211580167" lat="37.365065" lon="-121.763415" version="1" timestamp="2011-03-21T06:45:44Z" changeset="7624353" uid="55774" user="nmixter">
<tag k="addr:city" v="Mt Hamilton"/>
<tag k="addr:full" v="94 Mount Hamilton"/>
<tag k="addr:state" v="CA"/>
<tag k="addr:street" v="Mount Hamilton"/>
<tag k="addr:country" v="US"/>
<tag k="addr:housenumber" v="94"/>
</node>

Marino > OK

<way id="377385920" version="1" timestamp="2015-10-28T22:54:11Z" changeset="34938668" uid="3343986" user="n76_cupertino_import">
<nd ref="3807594515"/>
<nd ref="3807594473"/>
<nd ref="3807594474"/>
<nd ref="3807594480"/>
<nd ref="3807594479"/>
<nd ref="3807594490"/>
<nd ref="3807594487"/>
<nd ref="3807594517"/>
<nd ref="3807594515"/>
<tag k="height" v="7.9"/>
<tag k="building" v="yes"/>
<tag k="addr:street" v="Via San Marino"/>
<tag k="addr:postcode" v="95014-6333"/>
<tag k="addr:housenumber" v="10822"/>

Hwy > Highway

Volante > OK

<tag k="addr:street" v="Via Volante"/>
<tag k="addr:postcode" v="95014-6315"/>
<tag k="addr:housenumber" v="20326"/>

Dr > Drive

CA > too many probably CA state mistake delete?

robles > Rio Robles misc

<tag k="addr:street" v="rio robles"/>
<tag k="addr:postcode" v="95134"/>
<tag k="addr:housename" v="oak"/>
<tag k="addr:housenumber" v="110"/>

201 > Great America Parkway

<tag k="name" v="Nook"/>
<tag k="office" v="company"/>
<tag k="addr:city" v="Santa Clara"/>
<tag k="addr:state" v="CA"/>
<tag k="addr:street" v="Great America Pkwy Ste 201"/>
<tag k="addr:postcode" v="95054"/>
<tag k="addr:housenumber" v="4555"/>

Esquela> OK

<nd ref="462096697"/>
<nd ref="462096698"/>
<nd ref="65418224"/>
<nd ref="462096699"/>
<nd ref="462096700"/>
<nd ref="462096697"/>
<tag k="name" v="Challenger School"/>
<tag k="amenity" v="school"/>
<tag k="addr:street" v="Camina Esquela"/>
<tag k="addr:housenumber" v="730"/>

0.1 > Unknown

<tag k="landuse" v="industrial"/>
<tag k="addr:county" v="Alameda"/>
<tag k="addr:street" v="Ala 680 PM 0.1"/>
<tag k="attribution" v="Caltrans"/>
<tag k="description" v="District 4 - East Bay Region"/>
<tag k="caltrans:type" v="SL"/>
<tag k="caltrans:district" v="4"/>
<tag k="caltrans:dynsegpm" v="ALA 680 M0.1"/>
</way>

Bellomy > Bellomy Street

<tag k="name" v="Gravitron"/>
<tag k="building" v="house"/>
<tag k="addr:street" v="Bellomy"/>
<tag k="addr:postcode" v="95050"/>
<tag k="addr:housename" v="Gravitron"/>
<tag k="addr:housenumber" v="842 "/>

Napoli > OK

<tag k="addr:street" v="Via Napoli"/>
<tag k="addr:postcode" v="95014"/>
<tag k="addr:housenumber" v="20238"/>

Sq > Square
Plaza > OK
Barcelona > OK

<tag k="addr:street" v="Calle de Barcelona"/>
<tag k="addr:postcode" v="95014-3373"/>
<tag k="addr:housenumber" v="19475"/>

St > Street
Cir > Circle

Franklin > Franklin Street
<tag k="addr:street" v="Franklin"/>
<tag k="addr:postcode" v="95050"/>
<tag k="outdoor_seating" v="yes"/>
<tag k="addr:housenumber" v="1595"/>

Palamos > OK

<tag k="addr:street" v="Via Palamos"/>
<tag k="addr:postcode" v="95014-6331"/>
<tag k="addr:housenumber" v="20325"/>

Presada > OK
<tag k="addr:street" v="Paseo Presada"/>
<tag k="addr:postcode" v="95070"/>
<tag k="addr:housenumber" v="13249"/>
Loop > OK

Bascom > Bascom Rd
<tag k="name" v="San Jose Water Company"/>
<tag k="landuse" v="industrial"/>
<tag k="addr:city" v="San Jose"/>
<tag k="addr:street" v="S. Bascom"/>
<tag k="addr:housenumber" v="1221"/>

Sorrento > OK
<tag k="addr:street" v="Via Sorrento"/>
<tag k="addr:postcode" v="95014-6313"/>
<tag k="addr:housenumber" v="10860"/>

Portofino > OK
<tag k="addr:street" v="Via Portofino"/>
<tag k="addr:postcode" v="95014-6310"/>
<tag k="addr:housenumber" v="20337"/>
</way>

Julian > Julian Street
<tag k="addr:street" v="West Julian"/>
<tag k="addr:postcode" v="95110"/>
<tag k="addr:housenumber" v="350"/>

Seville > OK
<tag k="addr:street" v="Corte de Seville"/>
<tag k="addr:postcode" v="95014-3407"/>
<tag k="addr:housenumber" v="10420"/>

Luna > OK
<tag k="addr:city" v="Santa Clara"/>
<tag k="addr:street" v="Calle de Luna"/>
<tag k="addr:housenumber" v="2281"/>

street > Street
Terrace > OK
Madrid > OK
<tag k="addr:street" v="Corte de Madrid"/>
<tag k="addr:postcode" v="95014-3406"/>
<tag k="addr:housenumber" v="10420"/>

Blvd > Boulevard
Ave > Avenue
Row > OK
Ct > Court

This the resulting mapping dictionary I created:

```
mapping = { "St": "Street",
            "St.": "Street",
            "Ave":"Avenue",
            "Rd.":"Road",
            "#6":"",
            "Winchester":"Winchester Boulevard",
            "Ln":"Lane",
            "Rd.":"Road",
            "114":"",
            "A":"",
            "ave":"Avenue",
            "20":"",
```

```
            "Boulvevard":"Boulevard",
            "1":"",
            "Hamilton":"Hamilton Road",
            "Hwy":"Highway",
            "Dr":"Drive",
            "CA":"",
            "0.1":"",
            "Bellomy":"Bellomy Street",
            "Cir": "Circle",
            "Franklin":"Franklin Street",
            "Bascom":"Bascom Rd",
            "Julian":"Julian Street",
            "street":"Street",
            "Blvd":"Boulevard",
            "Ct":"Court",
            }
```

I also truncated all ZIP and Postcodes to 5 digits:

```
def fix_zip(zip):
    new_zip = zip[:5]
    print new_zip
    return new_zip
```

And created a function to make one off changes:

```
def fix_misc(snippet):
    if snippet == "Great American Pkwy Ste 201":
        new_snippet == "Great American Parkway"
        print "**"
        print new_snippet
        return new_snippet
    elif snippet == "rio robles":
        new_snippet = "Rio Robles Drive"
        print "**"
        return new_snippet
    elif snippet == "Rio Robles":
        new_snippet = "Rio Robles Drive"
        print "**"
        return new_snippet
    elif snippet == "Zanker Road, San Jose,":
        print "**"
        new_snippet = "Zanker Road"
        return new_snippet
    elif snippet == "Zanker Road, San Jose,":
        print "**"
        new_snippet = "Zanker Road"
        return new_snippet
```

```
    elif snippet == "wilcox Avenue":
        print "**"
        new_snippet = "Wilcox Avenue"
        return new_snippet
    elif snippet == "Ala 680 PM":
        print "**"
        new_snippet = "Unknown"
        return new_snippet
    elif snippet == "Hwy 17 PM 7.1":
        print "**"
        new_snippet = "Highway 17"
        return new_snippet
    elif snippet.endswith("#"):
        print "**"
        new_snippet = snippet[:-2]
        return new_snippet
    else:
        new_snippet = snippet
        return snippet
```

My code

osm_csv.py has the shape_element() function
write_change.py has the helper functions to clean data