地图区域

中国香港
数据来源：https://www.openstreetmap.org/search?query=Hongkong#map=11/22.2794/114.1628
选择香港区域是因为该地区人口密度大，经济发达，地图数据量较大，能从多维的角度分析，同时可以使我更了解这个城市。

地图中遇到的问题

在sample.osm(34.5MB)中，进行初步的处理，发现数据的问题如下：
1 value值中"港","香港 Hong Kong","HK","Hong Kongo"等多种书写形式，统一转成"Hong Kong"。
2 street中有"Rd","Lu","St.","S.","St"等，统一转成"Road"或"Street"。
3 phone中书写格式不规范，如：27898521，(852)25299280等，改成+852 2529 9280或者+86 750 595 9387等。

将四种Hong Kong的书写形式进行统一

```
def update_city(city):
    if city in["港","香港 Hong Kong","HK","Hong Kongo"]:
        city = "Hong Kong"
    return city
```

将各种不规范的字符串进行统一

```
mapping = { "St": "Street",
            "St.": "Street",
            "S.": "Street",
            "Rd.": "Road",
            "Lu": "Road",
            "Load": "Road"
            }
```

对phone的数据进行整理

```
def update_phone_num(phone_num):
    """
    Clean phone number for insertion into SQL database
    """
    # Check for valid phone number format
    m = PHONENUM.match(phone_num)
    if m is None:
        # Convert all dashes to spaces
        if "-" in phone_num:
            phone_num = re.sub("-", "", phone_num)
        # Remove all brackets
        if "(" in phone_num or ")" in phone_num:
            phone_num = re.sub("[()]", "", phone_num)
        # Remove all the spaces
        if " " in phone_num:
            phone_num = re.sub(" ", "", phone_num)
        # Space out 8 straight numbers
        if re.match(r'\d{8}', phone_num) is not None:
            phone_num = "+852"+" "+phone_num
        #大陆的电话处理

        if re.match(r'0\d{11}', phone_num) is not None:
            phone_num = phone_num[:4] + " " + phone_num[5:11]
        #手机号的86的处理
        if re.match(r'86\d{11}', phone_num) is not None:
            phone_num = "+" + phone_num
        #另一种手机号的86的处理
        if re.match(r'86\d{10}', phone_num) is not None:
            phone_num = "+" + phone_num
        # Ignore tag if no area code and local number (<8 digits)
        elif sum(c.isdigit() for c in phone_num) < 8:
            return None
    return phone_num
```

数据的大小

```
nodes.csv(44.7MB);
nodes_tags.csv(10.85MB);
ways.csv(20.7MB);
ways_nodes.csv(92.6MB);
ways_tags.csv(32.6MB);
osdb(170MB)
```

数据探索和分析

查看每张表中有多少条字段

```
select count(*) from nodes;
select count(*) from nodes_tags;
select count(*) from ways;
select count(*) from ways_nodes;
select count(*) from ways_tags;
```

525575; 574246; 357054; 690343; 3733528;

查看所有的用户数量

```
select count(a.uid) as num from
(select uid from nodes union all select uid from ways) a ;
```

882629

查看前5名user贡献的数量

```
select a.user,count(a.user) as num
from (select user from nodes union all select user from nodes) a
group by a.user
order by num desc
limit 5;
```

user	num
hlaw	146688
katpatuka	82082
mixmaxtw	64498
bTonyB	55562
Popolon	51518

香港里面island很多，对它们进行统计，前面一种查询中包含了island的命名的地方，如学校等，后面添加查询条件key='place'更为准确

```
select count(*) from ways_tags where value like '%Island%';
select count(*) from ways_tags where value like '%Island%'and key='place'
```

2052 ; 874

查看nodes和ways表中timestamp最早的的5条记录

```
SELECT timestamp FROM Nodes UNION SELECT timestamp From Ways
    ORDER BY timestamp LIMIT 5
```

2007-01-02T18:44:29Z | 2007-01-02T18:44:30Z | 2007-03-08T11:48:49Z | 2007-03-08T11:49:19Z | 2007-03-08T11:50:50Z

香港道路很多都实行单线行驶，在实际的数据中也反应了其比较严厉的执行

```
SELECT count(*) FROM ways_tags WHERE key='oneway';
SELECT count(*) FROM ways_tags WHERE key='oneway' AND value='yes'
```

152600 ; 147984

整理香港地区的前5名cuisine

```
SELECT nodes_tags.value, COUNT(*) as num
    FROM nodes_tags
        JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='restaurant') a
        ON nodes_tags.id=a.id
    WHERE nodes_tags.key='cuisine'
    GROUP BY nodes_tags.value
    ORDER BY num DESC
    LIMIT 5
```

value	num
chinese	195
japanese	56
noodle	34
thai	24
indian	22

前5名银行的数量,实际中除了四大行，后面的Hang Seng Bank,Bank of East Asia,Citibank和招商银行数量也很多

```
SELECT nodes_tags.value, COUNT(*) as num
    FROM nodes_tags
        JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='bank') a
        ON nodes_tags.id=a.id
    WHERE nodes_tags.key='name'
    GROUP BY nodes_tags.value
    ORDER BY num DESC
    LIMIT 5
```

value	num
Bank of China	139
Industrial and Commercial Bank of China	87
China Construction Bank	64
HSBC Bank	58
Agricultural Bank of China	40

查看clothes的商店数量

```
SELECT nodes_tags.value, COUNT(*) as num
    FROM nodes_tags
        JOIN (SELECT DISTINCT(id) FROM nodes_tags WHERE value='clothes') a
        ON nodes_tags.id=a.id
    WHERE nodes_tags.key='shop'
```

536

小结

在OpenStreetMap中的数据的写法中，没有统一的标准，有些用户的写法很随意，比如很多简写，在本地人可能一眼就看的出来，但是外地人就不是那么容易了，例如在bank分析中，有个银行书写的IBC,可能我就会认为是工商银行(ICBC),用户自己写错了，但是国内兴业银行的简称也是IBC,在国外IBC Bank是美国德克萨斯州的一家银行，而且香港作为国际金融中心，完全有可能有美帝的银行，花旗银行就是很好的说明。用户在OpenStreetMap上进行城市的标注时，很有必要有相关的专业人士按照统一的标准口径进行审核，并可以辅助修改，审核合格后，才能在上面进行发布。制定统一书写标准，规范用户书写也是很有必要，但是审核通过并辅助修改的效率似乎更高，当然这就需要人力资本的支出了。

香港作为国际金融中心，其数据集非常大，该项目中将里面的部分数据进行了清洗，并进行了数据分析，还可以从其它的维度进行分析。数据集非常有用，但是又非常杂乱，在该项目中完成了数据分析从数据获取，数据清洗，数据入库和数据分析的一整套过程，全面提高了对数据分析的认识。