Data Wrangle Openstreetmaps Data

Udacity Nanodegree Project 2

Coursework:

- SQL for Data Analysis
- Intro to Data Wrangling

TABLE OF CONTENTS

DATA WRANGLING OPENSTREETMAPS DATA

1. Relevant Links

Github - Project Scripts

QGIS - Hosted map

---

2. Problems Encountered in the Map

Problems encountered with my map

Up until auditing my data I was still pretty naive about what “dirty data” meant. It seems pretty ridiculous but I had the idea that “it couldn’t happen to me” like it was a sickness or something. I did the audit and what I think is probably a relatively small amount of problems came up within the data. I did a few things to tidy it up.

- **Add expected results**

The first thing I did was add more expected results like “Alley”, “Circle”, “Pike”, “Terrace”, and “Ridge”.

```
expected = ["Street", "Avenue", "Boulevard", "Drive", "Court", "Place", "Square", "Lane",
"Road", "Trail", "Parkway", "Commons", "Alley", "Circle", "Pike", "Terrace", "Ridge"]
```

- **Remap variables**

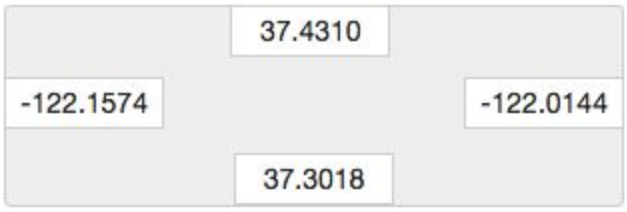
I remapped a few variables like Blvd and Blvd. to Boulevard, and Rd to Road.

```
mapping = { "St": "Street",
            "St.": "Street",
            "street": "Street",
            "Ave": "Avenue",
            "Ave." : "Avenue",
            "Rd": "Road",
            "Rd.": "Road",
            "Blvd" : "Boulevard",
            "Blvd." : "Boulevard"}
```

---

3. Data Overview

* Map Area



* File Size

51 MB

* Tables

```
sqlite> .tables
ElementaryGeometries                vector_layers
SpatialIndex                        vector_layers_auth
geom_cols_ref_sys                   vector_layers_field_infos
geometry_columns                    vector_layers_statistics
geometry_columns_auth               views_geometry_columns
geometry_columns_field_infos        views_geometry_columns_auth
geometry_columns_statistics         views_geometry_columns_field_infos
geometry_columns_time               views_geometry_columns_statistics
nodes                               virts_geometry_columns
nodes_tags                          virts_geometry_columns_auth
spatial_ref_sys                     virts_geometry_columns_field_infos
spatial_ref_sys_all                 virts_geometry_columns_statistics
spatial_ref_sys_aux                 ways
spatialite_history                  ways_nodes
sql_statements_log                  ways_tags
sqlite> count.tables
   ...> ;
Error: near "count": syntax error
sqlite> .tables
ElementaryGeometries                vector_layers
SpatialIndex                        vector_layers_auth
geom_cols_ref_sys                   vector_layers_field_infos
geometry_columns                    vector_layers_statistics
geometry_columns_auth               views_geometry_columns
geometry_columns_field_infos        views_geometry_columns_auth
geometry_columns_statistics         views_geometry_columns_field_infos
geometry_columns_time               views_geometry_columns_statistics
nodes                               virts_geometry_columns
nodes_tags                          virts_geometry_columns_auth
spatial_ref_sys                     virts_geometry_columns_field_infos
spatial_ref_sys_all                 virts_geometry_columns_statistics
spatial_ref_sys_aux                 ways
spatialite_history                  ways_nodes
sql_statements_log                  ways_tags
```

* Schema of nodes_tags and ways_tags tables

```
sqlite> .schema ways_tags
CREATE TABLE ways_tags ( id INTEGER, k TEXT, v TEXT );
CREATE INDEX ways_tags_idx ON ways_tags(id);
sqlite> .schema nodes_tags
CREATE TABLE nodes_tags ( id INTEGER, k TEXT, v TEXT );
CREATE INDEX nodes_tags_idx ON nodes_tags(id);
```

* Number of nodes_tags

```
sqlite> SELECT COUNT(*) FROM nodes_tags;
36715
```

* Number of ways_tags

```
sqlite> SELECT COUNT(*) FROM ways_tags;
220033
```

Number of Chosen Node Types

* Number of Highways

```
sqlite> SELECT COUNT(k) FROM nodes_tags WHERE k='highway';
4164
```

* Religions

```
sqlite> SELECT DISTINCT v FROM ways_tags WHERE k='religion'
christian
buddhist
yogic
scientologist

sqlite> SELECT COUNT(k) FROM ways_tags WHERE v='christian';
48
```

*** Popular way_tags**

```
sqlite> SELECT k FROM ways_tags GROUP BY k ORDER BY COUNT(*) DESC LIMIT 5;
building
source
addr:housenumber
addr:street
highway
```

* Types of Cuisine

```
sqlite> SELECT DISTINCT(v) FROM ways_tags WHERE k='cuisine';
burger
pizza
mexican
sandwich
american
barbecue
italian
```

---

4. Building the Map

QGIS - [Hosted map]

Overview

Layers Used

Layer 1:

Metadata of Layer "mountainview_clean_polygons"

This layer is queryable: yes

Display-Field: name

Attributes / Fields

Name	Type	Comment	Length	Precision
id_0	int		-1	0
id	qlonglong		-1	0
building	QString		-1	-1
source	QString		-1	-1
addr:housenumber	QString		-1	-1
addr:street	QString		-1	-1
highway	QString		-1	-1
paloalto_ca:id	QString		-1	-1
addr:city	QString		-1	-1

Layer 2:

Metadata of Layer "mountainview_shop"

This layer is queryable: yes

Display-Field: shop

Attributes / Fields

Name	Type	Comment	Length	Precision
id_0	int		-1	0
id	qlonglong		-1	0
shop	QString		-1	-1

---

5. Additional Ideas

It would be interesting if open street maps was an app that people could allow to access their location. The collected data could be used to build real time traffic reports based on the accumulation of data derived from vehicle speeds of people using the app. I haven't really looked into how google does it but they're probably doing this already. I think

being an open sourced non commercial entity would give people more faith in allowing the app to access their information.

A cross between https://photosynth.net/ and openstreetmaps that could access user photos by and upload them based on the lat/long they were taken could be some sort of open sourced google earth.

---

6. Conclusion

Very awesome Project.