Wrangle OpenStreetMap Data

Student Notes Code Review Project Review Project History

Meets Specifications

Code Functionality

SPECIFICATION
Final project code functionality reflects the description in the project document.

MEETS SPECIFICATION

Reviewer Comments

Code submitted reflects work done in the submitted document PDF.

Code Readability

SPECIFICATION
Final project code follows an intuitive, easy-to-follow logical structure.

MEETS SPECIFICATION

Reviewer Comments

Distinct blocks separate distinct functions. There is no overlapping in function procedures. This makes the code extremely logical and easy to follow.

SPECIFICATION
Final project code that is not intuitively readable is well-documented with comments.

MEETS SPECIFICATION

Reviewer Comments

All critical areas are well commented.

SPECIFICATION

How satisfied are you with this feedback?

Student response shows understanding of the process of auditing, and ways to correct or standardize the data, including dealing with problems specific to the location, e.g. related to language or traditional ways of formatting.

MEETS SPECIFICATION

Reviewer Comments

Great work on the extensive cleaning procedures. Especially on such a large dataset. I would be interested to know the computation time.

List of Cleaning Procedures

1. Street names that don't share a common suffix (\1st Street vs. 1st St")
2. Inconsistent use of N vs North, S vs South, etc.
3. Address number included in the street names (\Adrian court, Suite A")
4. Intersections are used for some of the values (Pacheco Blvd @ Blum Rd N Of Sr 4)
5. Landmarks are used for some of the values (Tanforan Shopping Center).

SPECIFICATION

Some of the problems encountered during data audit are cleaned programmatically.

MEETS SPECIFICATION

Reviewer Comments

Procedures are carried out programmatically.

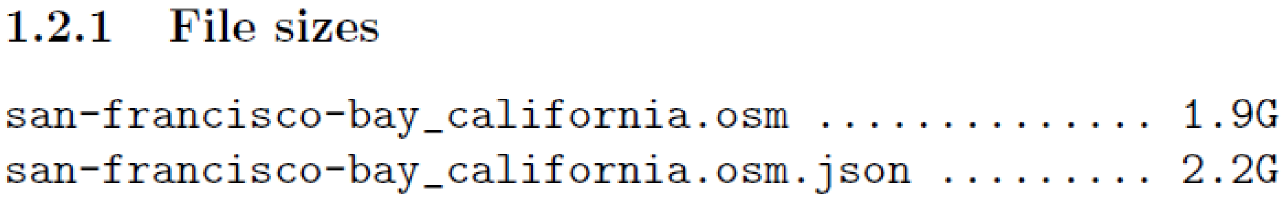
Overview of the data

SPECIFICATION

The dataset is at least 50 MB.

MEETS SPECIFICATION

Reviewer Comments



1.2.1 File sizes

```
san-francisco-bay_california.osm .............. 1.9G
san-francisco-bay_california.osm.json ......... 2.2G
```

Again, very impressed at the size of the OSM file.

SPECIFICATION

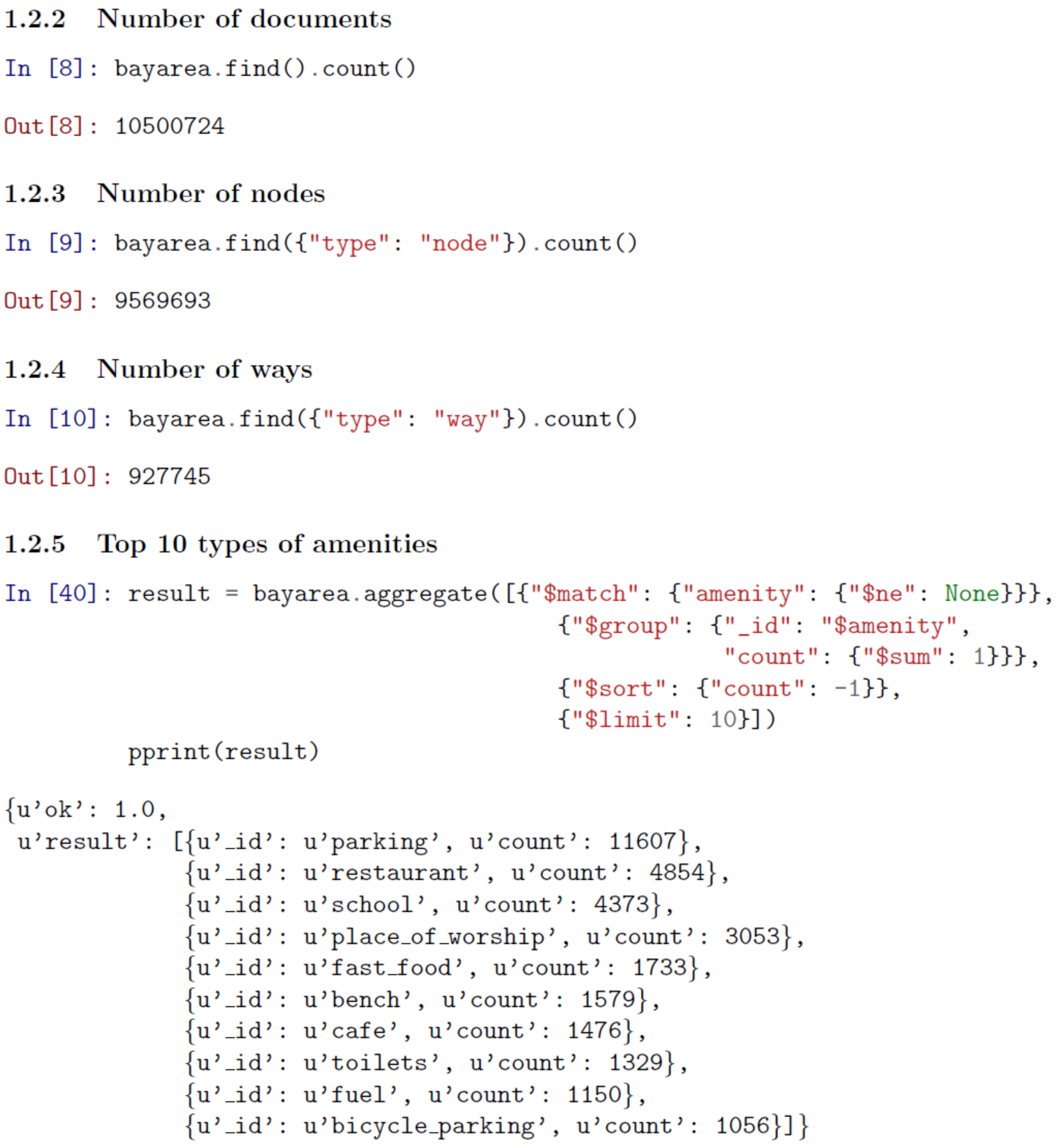
Student response provides a statistical overview of a dataset, like:

- size of the file

- number of unique users
- number of nodes and ways
- number of chosen type of nodes, like cafes, shops etc

MEETS SPECIFICATION

Reviewer Comments



1.2.2 Number of documents

```
In [8]: bayarea.find().count()

Out[8]: 10500724
```

1.2.3 Number of nodes

```
In [9]: bayarea.find({"type": "node"}).count()

Out[9]: 9569693
```

1.2.4 Number of ways

```
In [10]: bayarea.find({"type": "way"}).count()

Out[10]: 927745
```

1.2.5 Top 10 types of amenities

```
In [40]: result = bayarea.aggregate([{"$match": {"amenity": {"$ne": None}}},
                                     {"$group": {"_id": "$amenity",
                                                 "count": {"$sum": 1}}},
                                     {"$sort": {"count": -1}},
                                     {"$limit": 10}])
         pprint(result)

{u’ok’: 1.0,
 u’result’: [{u’_id’: u’parking’, u’count’: 11607},
             {u’_id’: u’restaurant’, u’count’: 4854},
             {u’_id’: u’school’, u’count’: 4373},
             {u’_id’: u’place_of_worship’, u’count’: 3053},
             {u’_id’: u’fast_food’, u’count’: 1733},
             {u’_id’: u’bench’, u’count’: 1579},
             {u’_id’: u’cafe’, u’count’: 1476},
             {u’_id’: u’toilets’, u’count’: 1329},
             {u’_id’: u’fuel’, u’count’: 1150},
             {u’_id’: u’bicycle_parking’, u’count’: 1056}]}
```

SPECIFICATION

Student response also includes the MongoDB queries used to obtain the statistics.

MEETS SPECIFICATION

Reviewer Comments

ADDITIONALLY (OPTIONAL)

Pipelines

It is recommended that coders create pipelines prior to database submission/query. This added step allows coders to review pipelines for errors that could potentially harm the database. Here is an example.

Single Command Method

```
db.locations.aggregate([{"$group": {"_id": "$created.user", "count":
{"$sum":1}}}, {"$sort": {"count": -1}}, {"$limit": 5}])
```

Safer Method

```
pipeline = ([{"$group": {"_id": "$created.user", "count":
{"$sum":1}}},
{"$sort": {"count": -1}},
{"$limit": 5}]
db.locations.aggregate(pipeline)
```

This is just an example. This is particularly important when one is using `db.update()` or `db.insert()` to make database changes. Just something to think about and is optional.

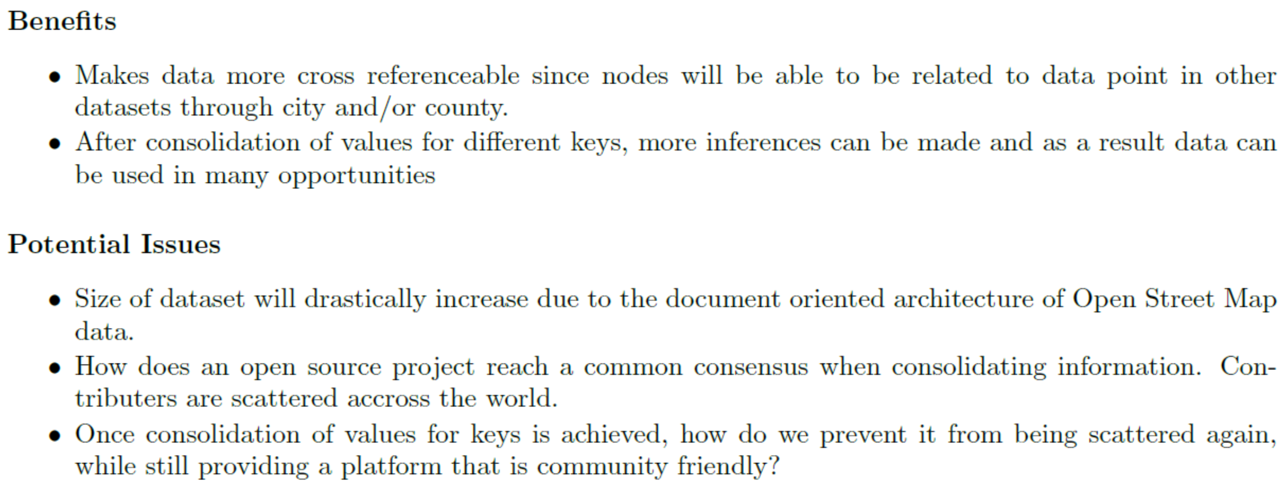
Other ideas about the datasets

SPECIFICATION

Student proposes one or more additional ways of improving and analyzing the data and gives thoughtful discussion about the benefits and anticipated problems in implementing the improvement.

MEETS SPECIFICATION

Reviewer Comments



Benefits

- Makes data more cross referenceable since nodes will be able to be related to data point in other datasets through city and/or county.
- After consolidation of values for different keys, more inferences can be made and as a result data can be used in many opportunities

Potential Issues

- Size of dataset will drastically increase due to the document oriented architecture of Open Street Map data.
- How does an open source project reach a common consensus when consolidating information. Contributers are scattered accross the world.
- Once consolidation of values for keys is achieved, how do we prevent it from being scattered again, while still providing a platform that is community friendly?

Outstanding extension of the discussion. This is exactly the type of deep and thoughtful contemplation we were looking for when thinking of ideas to improve OSM.

Thoroughness and Succinctness of Submission

SPECIFICATION

Student submission is long enough to thoroughly answer the questions asked without giving unnecessary detail. A good general guideline is that your question responses should take about 3-6 pages.

MEETS SPECIFICATION

Additional Reviewer Comments

Next, Using Data Visualization for Analytics

Congratulations on completing the course. Next we will be going into the world of R Statistical Language - a very powerful deep analytical and visualization tool that is gaining popularity in the Data Science Field in an exponential rate. It is one of my favorite languages and I hope you enjoy it as well. Good luck!

Download project

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

NANODEGREE PROGRAMS

Front-End Web Developer
Full Stack Web Developer
Data Analyst
iOS Developer
Android Developer
Intro to Programming
Tech Entrepreneur

STUDENT RESOURCES

Blog
Help & FAQ
Catalog
Veteran Programs

PARTNERS & EMPLOYERS

Georgia Tech Program
Udacity for Business
Hire Nanodegree Graduates