Imports - Three Tips for Working With Open Data

Ben Coleman
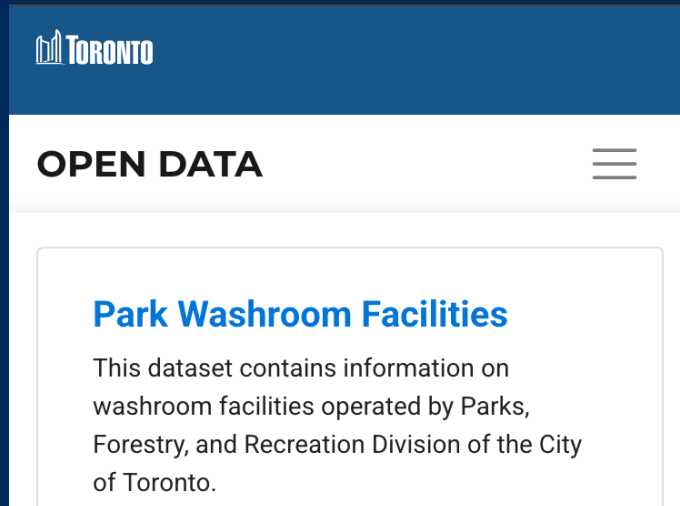
🇨🇦 Toronto, Canada

Open Data and Imports

Open Data

- Compatible License
  OpenStreetMap Foundation:
  License Compatibility

hours:
"9 a.m. to 10 p.m."

Imports

- Community norms and process:
  OpenStreetMap Import Guidelines

opening_hours:
"May-Oct 09:00-22:00"

Profiling

For each column/attribute:

- Purpose and meaning
- Data format and units
- Values
- Missing (nulls) and errors
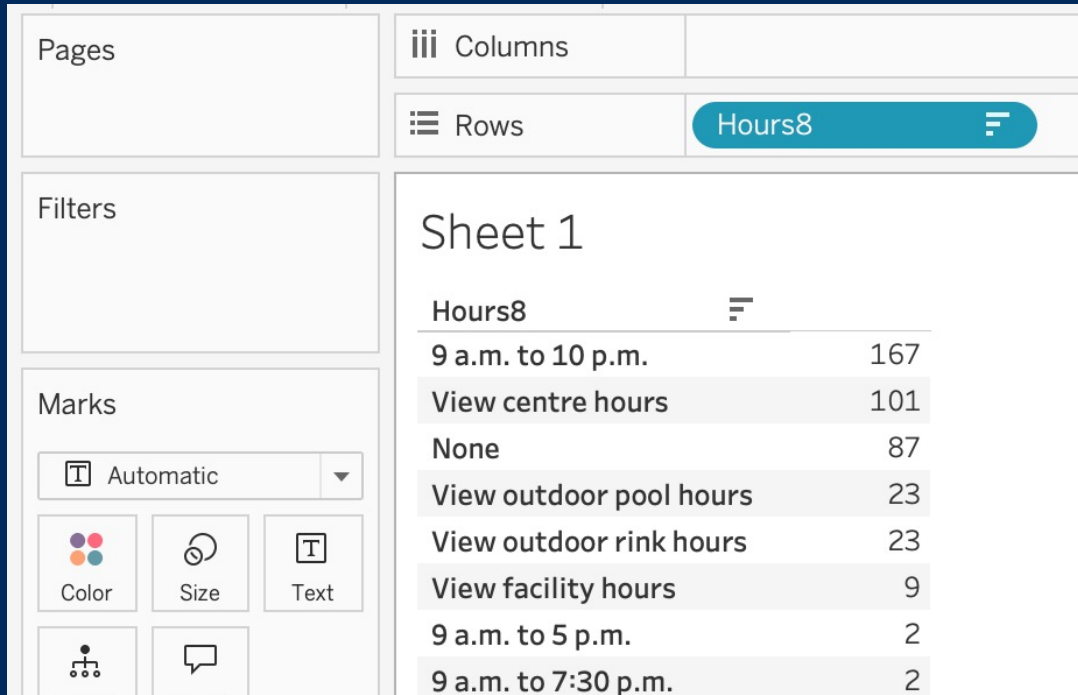
Save the original files and the date downloaded!

Example – accessible:

- Provides a list of accessible features, if available.
- Text, features are separated by commas, e.g.
  "Accessible Stall, Entrance at Grade"
  - Entrance at Grade [n=89]
  - Accessible Stall [n=79]
  - Child Change Table [n=33]
  - Entrance Access Ramp [n=19]
  - Automatic Door Opener [n=16]
  - Adult Change Table [n=1]
- 169 with values, 176 missing
- 1 accidental "hours" value

Desktop Tools

Tableau Public 🔗

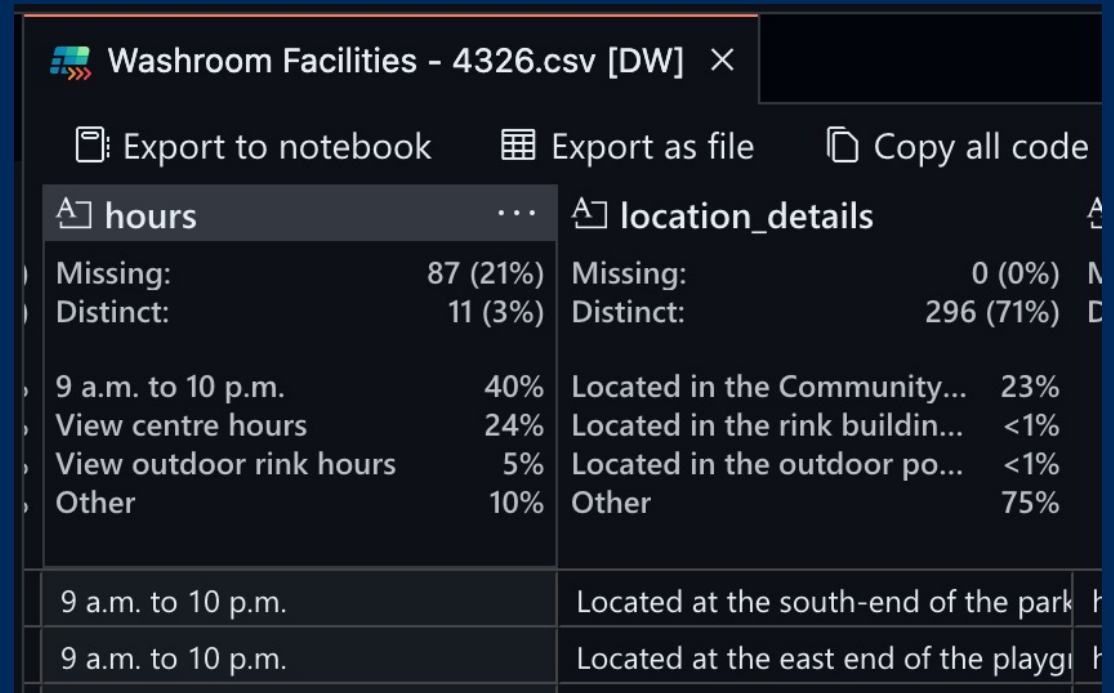- Supports most common open data filetypes including for spatial files

VS Code Data Wrangler 🔗

- Limited filetype support, but a great stepping stone to writing code

Python Coding Tools

GeoPandas

Jupyter Notebook

```
+ Code   + Markdown   | ▷ Run All   ↻ Restart   ≡ Clear All Outputs   | ⊞ View data   [x] Jupyter Variables   ≡ Outline   ⋯
```

```python
import pandas as pd
import geopandas as gpd
```
[1]   ✓ 4.0s

```
                     Follow link (cmd + click)
```

```python
gdf = gpd.read_file("https://ckan0.cf.opendata.inter.prod-toronto.ca/dataset/
washroom-facilities/resource/6d848f38-45a3-41e8-9783-804385ec5a16/download/
Washroom%20Facilities%20-%204326.geojson")
```
[2]   ✓ 1.4s

Outputs are collapsed ...

```python
gdf
```
[5]   ✓ 0.0s   ⊞ Open 'gdf' in Data Wrangler

⋯

	_id	id	asset_id	location	alternative_name	type	accessible	hours	locat
0	1	1431	79344	Grange Park	Grange Park North Washroom	Washroom Building	None	9 a.m. to 10 p.m.	Lo southl
1	2	263	79231	Bellevue Square Park	Bellevue Square Park Washroom	Washroom Building		9 a.m. to 10 p.m.	Lo eas
2	3	241	1000000059	Trinity Bellwoods Park	Trinity Bellwoods Playground Washroom	Washroom Building		9 a.m. to 10 p.m.	loc ou
				Marie	Marie Curtis Park	Washroom	Accessible Stall, Child	9 a.m.	Lo

Script + Terminal 😎

```
(Pdb) gdf.columns
Index(['_id', 'parent_id', 'asset_id', 'location', 'alternative_name', 'type',
       'accessible', 'hours', 'location_details', 'url', 'address',
       'PostedDate', 'AssetName', 'Reason', 'Comments', 'Status', 'geometry'],
      dtype='object')
(Pdb) gdf["hours"].describe()
count                      332
unique                      11
top         9 a.m. to 10 p.m.
freq                       167
Name: hours, dtype: object
(Pdb) gdf["hours"].value_counts(dropna=False)
hours
9 a.m. to 10 p.m.          167
View centre hours          101
<NA>                        87
View outdoor rink hours     23
View outdoor pool hours     23
View facility hours          9
9 a.m. to 7:30 p.m.          2
9 a.m. to 5 p.m.             2
View centre hours.           2
View arena hours             1
6:30 a.m. to 11:30 p.m.      1
View the facility hours      1
Name: count, dtype: Int64
(Pdb) █
```

Add `breakpoint()` to your script

```
.describe()
.value_counts(
    dropna=False)
.dtype
.sample(n)
```

Validation

```
    ),
    "accessible": pa.Column(
        str,
        nullable=True,
        required=True,
        checks=pa.Check(check_accessible),
    ),
```

```python
# validate city data and throw errors if input assumptions have changed
def check_accessible(accessible: pd.Series):
    known_values = [
        pd.NA,
        "Entrance at Grade",
        "Accessible Stall",
        "Child Change Table",
        "Entrance Access Ramp",
        "Automatic Door Opener",
        "Adult Change Table",
        # known error; ignored by the get_wheelchair_description function:
        "9 a.m. to 10 p.m.",
    ]
    accessible_lists = accessible.replace("", pd.NA).str.split(", ")
    return accessible_lists.apply(lambda x: pd.Series(x).isin(known_values).all())
```

Know when data changes break your assumptions!

PANDERA

Pydantic

Advocacy and Impact

Open Data usage is generally **anonymous**

Who to talk to:
- Data Owners
- Politicians

What to say:
- Thank you!
- How the data was used
- **Tangible impact**

Saturday, January 11, 2025 at 13:54:46 Eastern Standard Time

Subject: Impact of open data - import of Toronto public washrooms into OpenStreetMap
Date: Friday, September 27, 2024 at 7:22:14 PM Eastern Daylight Saving Time
From: Benjamin Coleman
To: pfrdatagovernance@toronto.ca <pfrdatagovernance@toronto.ca>

Hi there,

I wanted to share with you that myself and another OpenStreetMap user in Toronto were able to use the Park Washroom Facilities dataset made available on the City of Toronto open data portal to do an import of all the locations into OpenStreetMap. I thought it would be useful to write to you as the data owner to (a) thank you for making the data available, (b) share some information about the import and its impact, and (c) share a couple comments about the dataset.

First of all, the thank you – I really appreciate that this dataset was made available, and appreciate all the work that goes into keeping it detailed, well-documented, and up to date. Please share our gratitude with the staff at PF&R whose hard work have made this valuable information easy to access.

Second, about the import – as you likely know, OpenStreetMap is a free and open-source map database. It provides its own web map with basic features, but it is also a key data source for services like Apple Maps/Garmin/Mapbox, fitness apps like Strava, and for many researchers.

Our import of the Park Washroom Facilities dataset has added around 200 washrooms that were not in OpenStreetMap before and added additional detail to around 100 washrooms that had previously been added by community contributors. This means that e.g. for a Strava user on a run or bike ride, it would be much easier now for them to find a washroom when they need one. That benefit is multiplied throughout all the services that use OpenStreetMap data.

OpenStreetMap requires data imports to have detailed plans to ensure that additions are "good data" and thoughtfully merged with existing features added by the community. If you are curious, you can read our plan here: https://wiki.openstreetmap.org/wiki/Import/Toronto_Public_Washroom_Import. I'll also note that we added a field to the OpenStreetMap entries that includes the asset ID of each washroom, which enables easy cross-referencing between the City data and the OpenStreetMap data.

Thank you!

GitHub

[tallcoleman](#)

[toronto-osm-washroom-import](#)