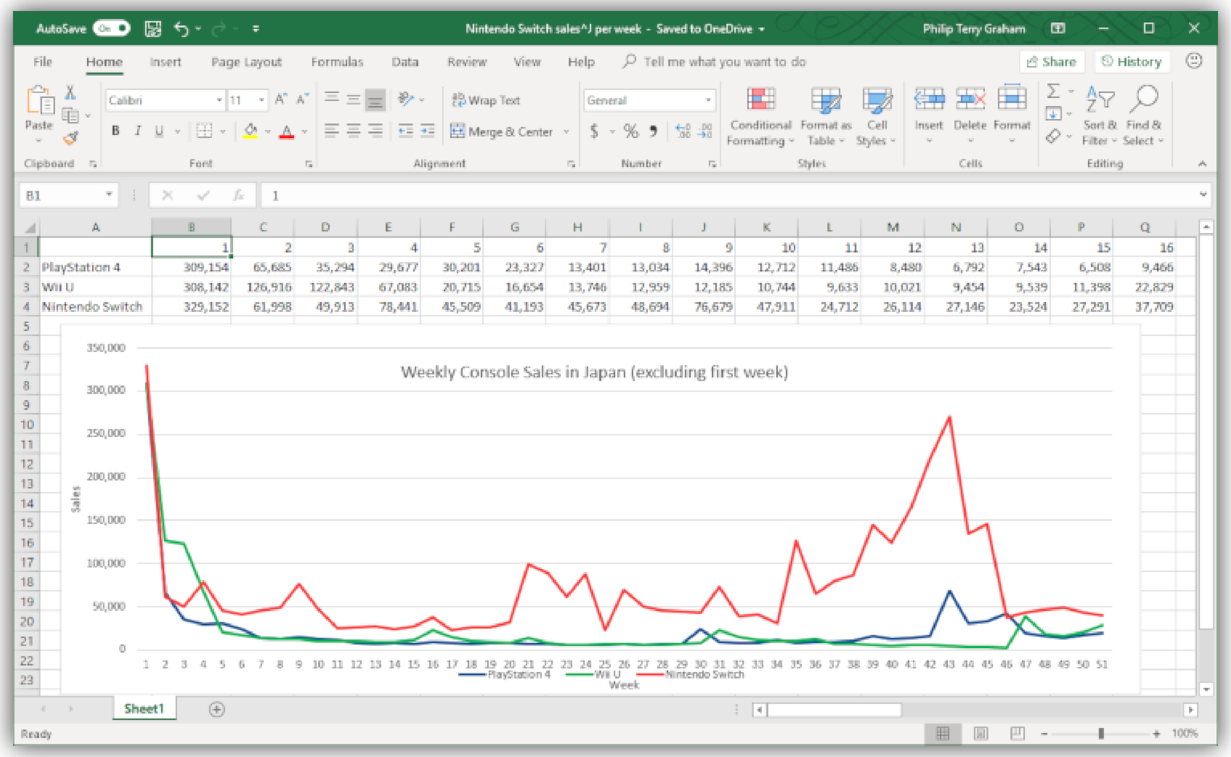
TRANSPORTATION DIVISION DATA ANALYTICS TOOLKIT

March 2022

Overview

- **Data Processing**
  - Excel
  - Python
  - R
- **Geospatial Data Processing**
  - ArcGIS
  - QGIS
- **Interactive Chart**
  - Tableau
  - Plotly
- **Interactive Map**
  - ArcGIS Online
  - Carto
  - Mapbox GL JS
- **Database**
  - MS SQL Server
  - SQLite / SpatiaLite
  - PostgreSQL / PostGIS
  - DBeaver
- **Other Tools**
  - GitHub
  - Visual Studio Code
  - Geosupport
  - DigitalOcean

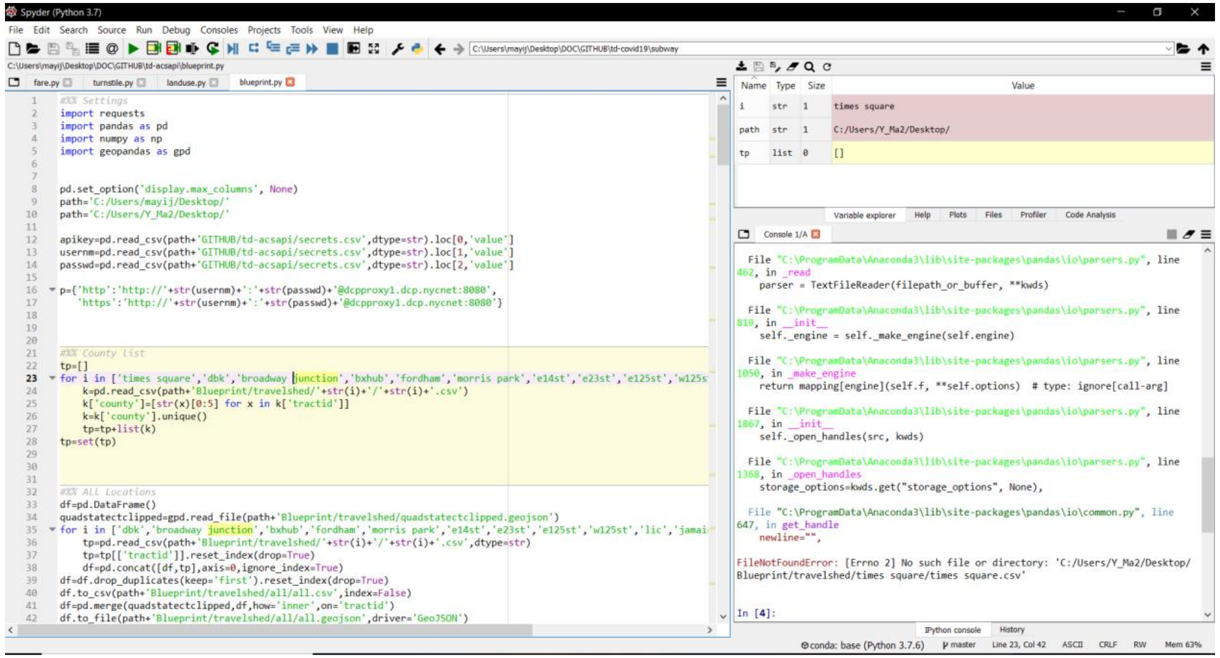
MS Excel



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
2	PlayStation 4	309,154	65,685	35,294	29,677	30,201	23,327	13,401	13,034	14,396	12,712	11,486	8,480	6,792	7,543	6,508	9,466
3	Wii U	308,142	126,916	122,843	67,083	20,715	16,654	13,746	12,959	12,185	10,744	9,633	10,021	9,454	9,539	11,398	22,829
4	Nintendo Switch	329,152	61,998	49,913	78,441	45,509	41,193	45,673	48,694	76,679	47,911	24,712	26,114	27,146	23,524	27,291	37,709

- Data analytics software by Microsoft
- Creating, reading, writing, manipulating, analyzing, and visualizing relatively small datasets
- Formulas (SUM, IF, VLOOKUP, etc.)
- Pivot table
- Static chart / map
- Data Analysis Tools
- Macros with VBA

Python


Spyder (Python 3.7)
File Edit Search Source Run Debug Consoles Projects Tools View Help
C:\Users\mayij\Desktop\DOC\GITHUB\td-covid19\subway
C:\Users\mayij\Desktop\DOC\GITHUB\td-acsapi\blueprint.py
fare.py
turnstile.py
landuse.py
blueprint.py
Variable explorer
Help
Plots
Files
Profiler
Code Analysis
Console 1/A
IPython console
History
conda: base (Python 3.7.6)
master
Line 23, Col 42
ASCII
CRLF
RW
Mem 63%

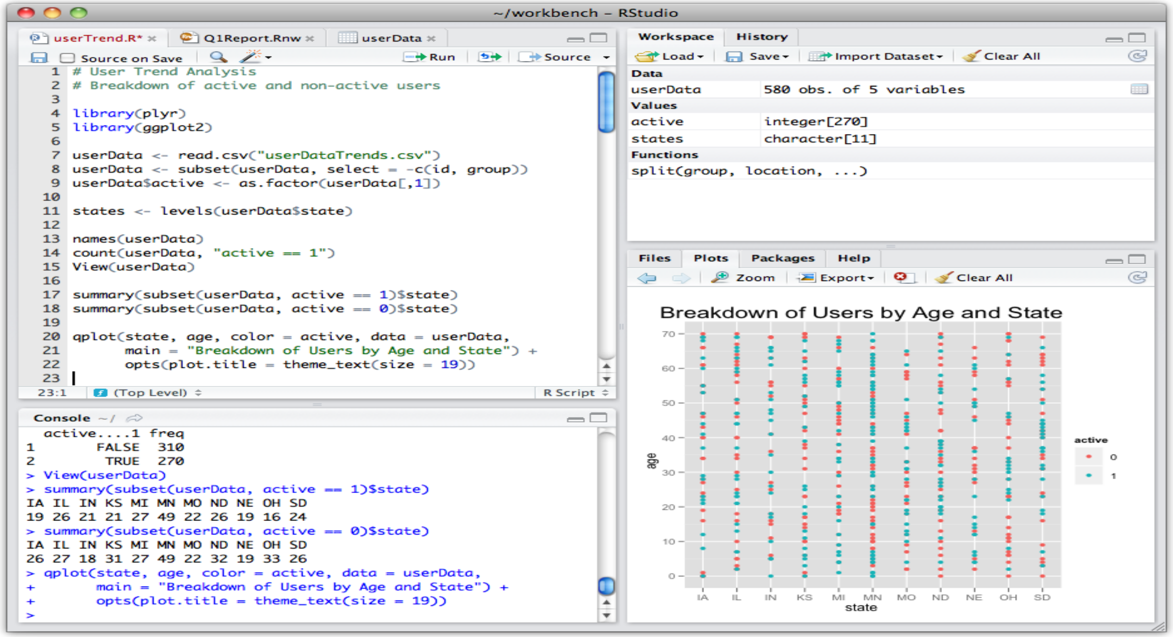
```python
#%% Settings
import requests
import pandas as pd
import numpy as np
import geopandas as gpd


pd.set_option('display.max_columns', None)
path='C:/Users/mayij/Desktop/'
path='C:/Users/Y_Ma2/Desktop/'

apikey=pd.read_csv(path+'GITHUB/td-acsapi/secrets.csv',dtype=str).loc[0,'value']
usernm=pd.read_csv(path+'GITHUB/td-acsapi/secrets.csv',dtype=str).loc[1,'value']
passwd=pd.read_csv(path+'GITHUB/td-acsapi/secrets.csv',dtype=str).loc[2,'value']

p={'http':'http://'+str(usernm)+':'+str(passwd)+'@dcpproxy1.dcp.nycnet:8080',
   'https':'http://'+str(usernm)+':'+str(passwd)+'@dcpproxy1.dcp.nycnet:8080'}



#%% County List
tp=[]
for i in ['times square','dbk','broadway junction','bxhub','fordham','morris park','e14st','e23st','e125st','w125s
    k=pd.read_csv(path+'Blueprint/travelshed/'+str(i)+'/'+str(i)+'.csv')
    k['county']=[str(x)[0:5] for x in k['tractid']]
    k=k['county'].unique()
    tp=tp+list(k)
tp=set(tp)



#%% All Locations
df=pd.DataFrame()
quadstatectclipped=gpd.read_file(path+'Blueprint/travelshed/quadstatectclipped.geojson')
for i in ['dbk','broadway junction','bxhub','fordham','morris park','e14st','e23st','e125st','w125st','lic','jamai
    tp=pd.read_csv(path+'Blueprint/travelshed/'+str(i)+'/'+str(i)+'.csv',dtype=str)
    tp=tp[['tractid']].reset_index(drop=True)
    df=pd.concat([df,tp],axis=0,ignore_index=True)
df=df.drop_duplicates(keep='first').reset_index(drop=True)
df.to_csv(path+'Blueprint/travelshed/all/all.csv',index=False)
df=pd.merge(quadstatectclipped,df,how='inner',on='tractid')
df.to_file(path+'Blueprint/travelshed/all/all.geojson',driver='GeoJSON')
```

Name	Type	Size	Value
i	str	1	times square
path	str	1	C:/Users/Y_Ma2/Desktop/
tp	list	0	[]

```
  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\parsers.py", line
462, in _read
    parser = TextFileReader(filepath_or_buffer, **kwds)

  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\parsers.py", line
819, in __init__
    self._engine = self._make_engine(self.engine)

  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\parsers.py", line
1050, in _make_engine
    return mapping[engine](self.f, **self.options)  # type: ignore[call-arg]

  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\parsers.py", line
1867, in __init__
    self._open_handles(src, kwds)

  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\parsers.py", line
1368, in _open_handles
    storage_options=kwds.get("storage_options", None),

  File "C:\ProgramData\Anaconda3\lib\site-packages\pandas\io\common.py", line
647, in get_handle
    newline="",

FileNotFoundError: [Errno 2] No such file or directory: 'C:/Users/Y_Ma2/Desktop/
Blueprint/travelshed/times square/times square.csv'


In [4]:
```

- Open-source general–purpose programming language
- Reading, writing, manipulating, analyzing, and visualizing relatively big data
- Reproducibility and transparency
- IDE: Spyder / Jupyter
- Data wrangling: Pandas
- Spatial data: GeoPandas
- Interactive visualization: Plotly
- Dashboard / App: Streamlit / Dash

R



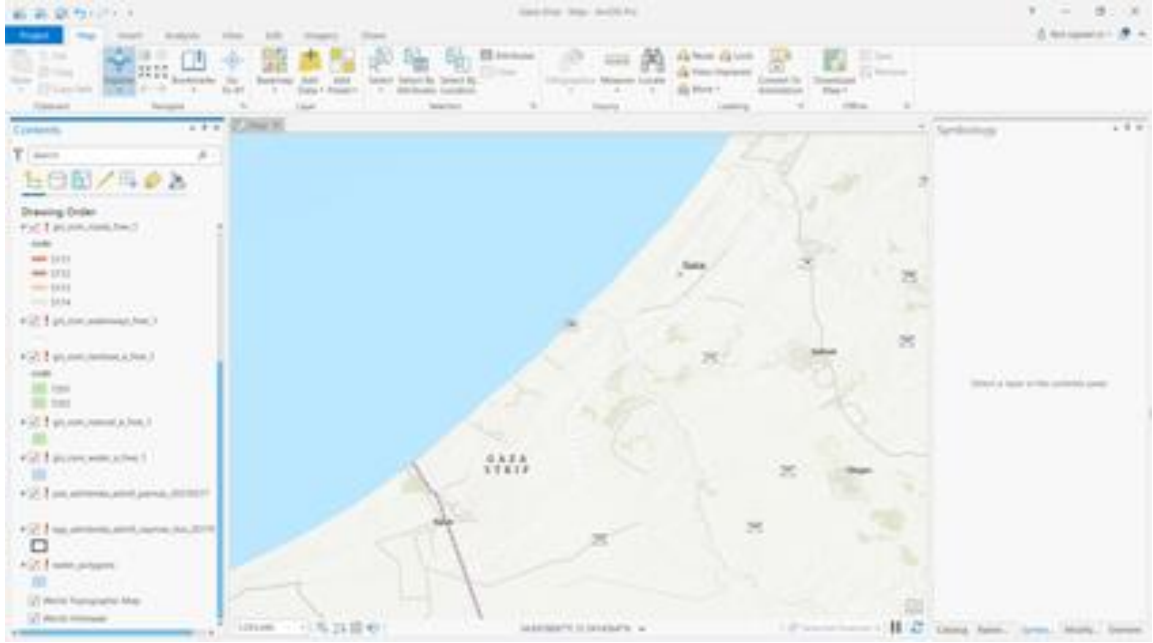
```r
# User Trend Analysis
# Breakdown of active and non-active users

library(plyr)
library(ggplot2)

userData <- read.csv("userDataTrends.csv")
userData <- subset(userData, select = -c(id, group))
userData$active <- as.factor(userData[,1])

states <- levels(userData$state)

names(userData)
count(userData, "active == 1")
View(userData)

summary(subset(userData, active == 1)$state)
summary(subset(userData, active == 0)$state)

qplot(state, age, color = active, data = userData,
      main = "Breakdown of Users by Age and State") +
      opts(plot.title = theme_text(size = 19))

```

```
  active....1 freq
1       FALSE  310
2        TRUE  270
> View(userData)
> summary(subset(userData, active == 1)$state)
IA IL IN KS MI MN MO ND NE OH SD
19 26 21 21 27 49 22 26 19 16 24
> summary(subset(userData, active == 0)$state)
IA IL IN KS MI MN MO ND NE OH SD
26 27 18 31 27 49 22 32 19 33 26
> qplot(state, age, color = active, data = userData,
+       main = "Breakdown of Users by Age and State") +
+       opts(plot.title = theme_text(size = 19))
>
```

```
Data
userData        580 obs. of 5 variables
Values
active          integer[270]
states          character[11]
Functions
split(group, location, ...)
```

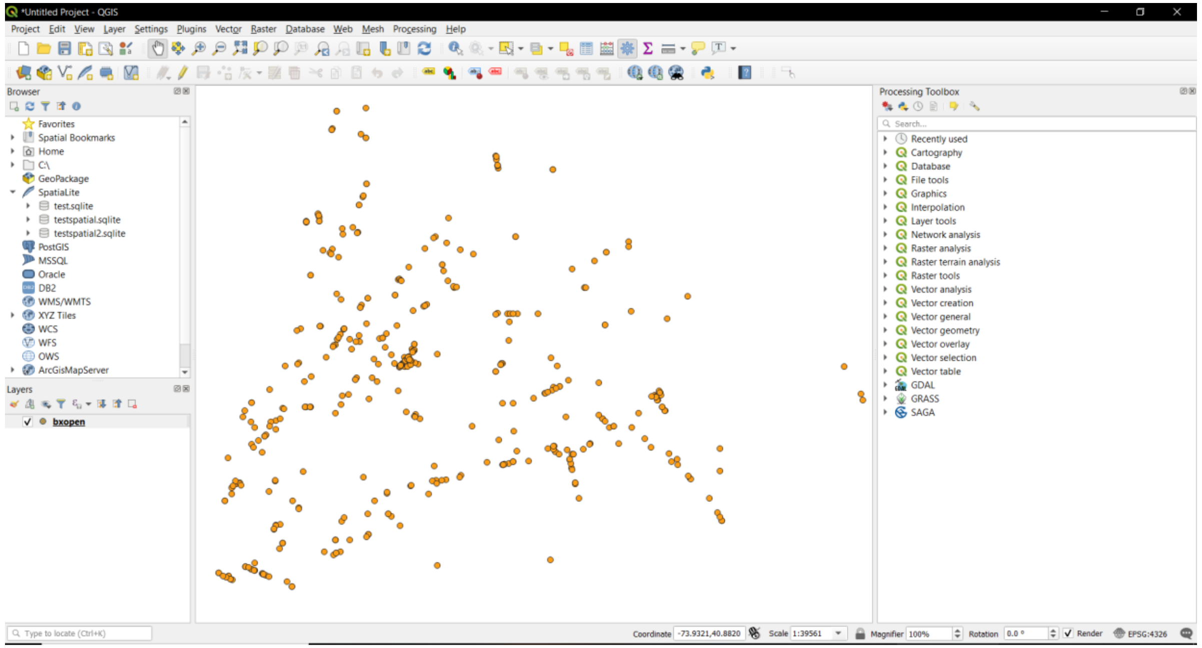
- Open-source programming language optimized for statisticians
- Reading, writing, manipulating, analyzing, and visualizing relatively big data
- IDE: RStudio
- Data wrangling: tidyverse (dplyr)
- Spatial data: sf
- Interactive visualization: Plotly
- Dashboard / App: R Markdown / R Shiny

ArcGIS Desktop



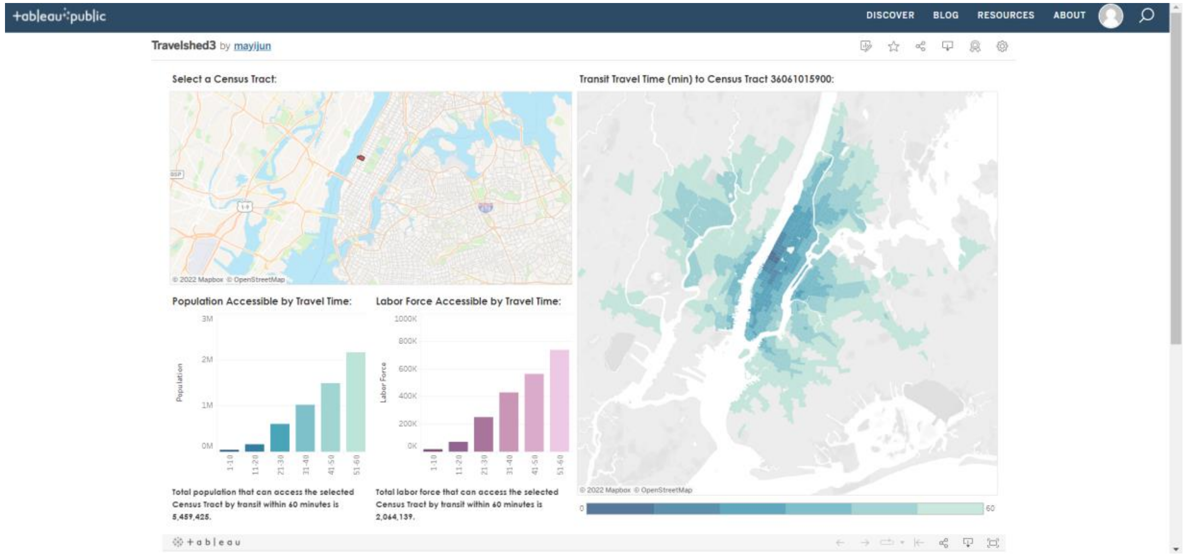
- Proprietary GIS software by ESRI
- Creating, reading, writing, manipulating, analyzing, and visualizing geospatial data
- Data format: Shapefile, Geodatabase, etc.
- ArcGIS Pro replacing ArcMap

QGIS



- Open-source GIS software
- Creating, reading, writing, manipulating, analyzing, and visualizing geospatial data
- Data format: Shapefile, GeoJSON, etc.
- Toolbox and plugins for geoprocessing

Tableau



Travelshed3 by mayijun

Select a Census Tract:

Transit Travel Time (min) to Census Tract 36061015900:

Population Accessible by Travel Time:

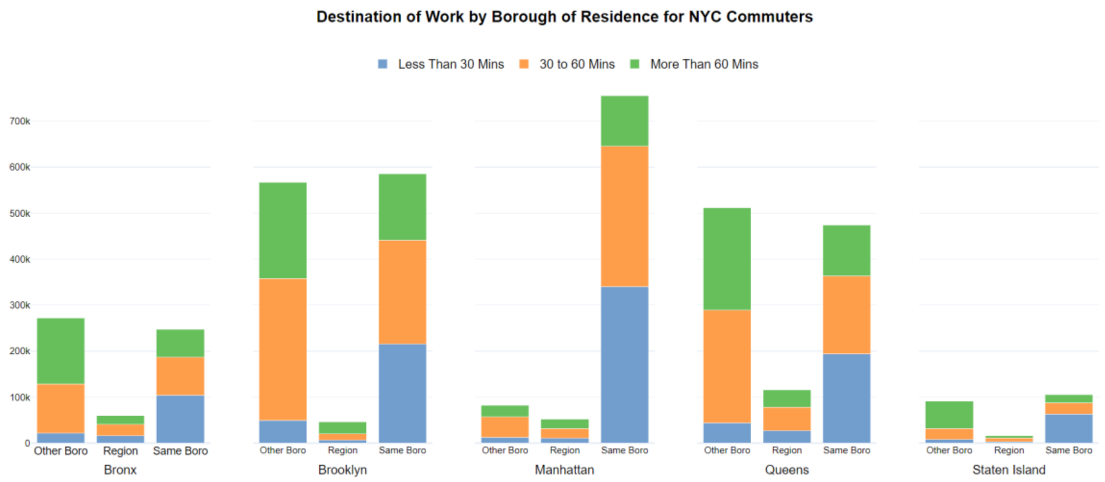
Labor Force Accessible by Travel Time:

Total population that can access the selected Census Tract by transit within 60 minutes is 5,459,425.

Total labor force that can access the selected Census Tract by transit within 60 minutes is 2,064,139.

- Data visualization software owned by Salesforce
- Creating dashboards with interactive charts and maps
- Tableau Public (free) for completely public data visualization
- Tableau Desktop / Online (require license) for non-public data

Plotly



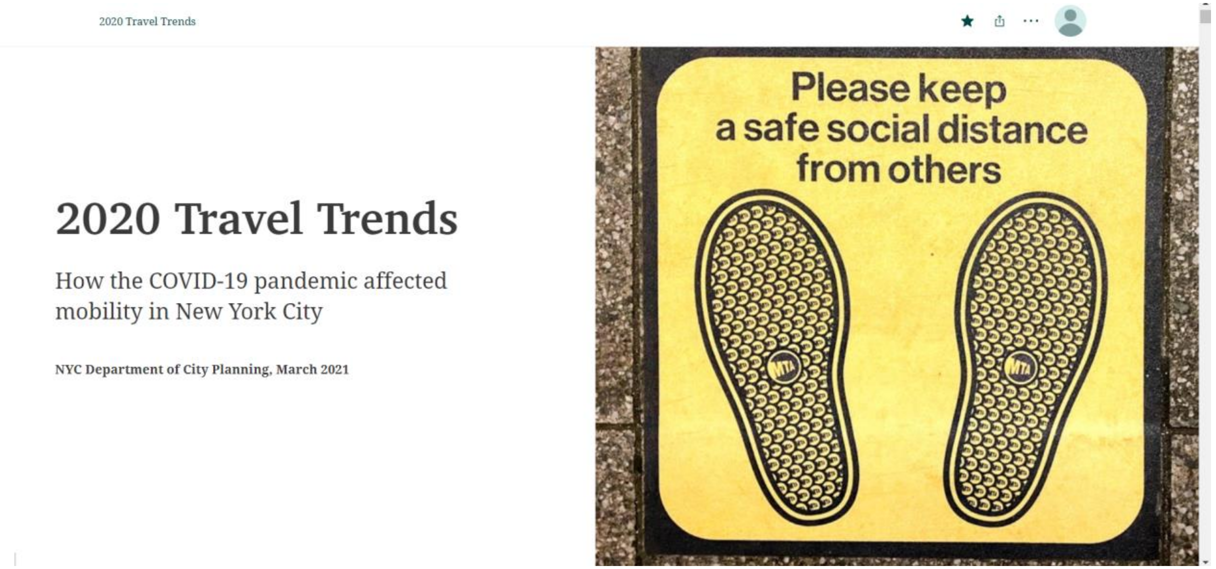
Destination of Work by Borough of Residence for NYC Commuters

- Open-source data visualization tools
- Creating interactive charts and maps
- Multi-platform graphing libraries: Python, R, JavaScript, etc.
- Dash for creating dashboards / apps
- Hosting platform: GitHub Pages

ArcGIS Online

ArcGIS Online

Mapping and analysis: location intelligence for everyone



2020 Travel Trends

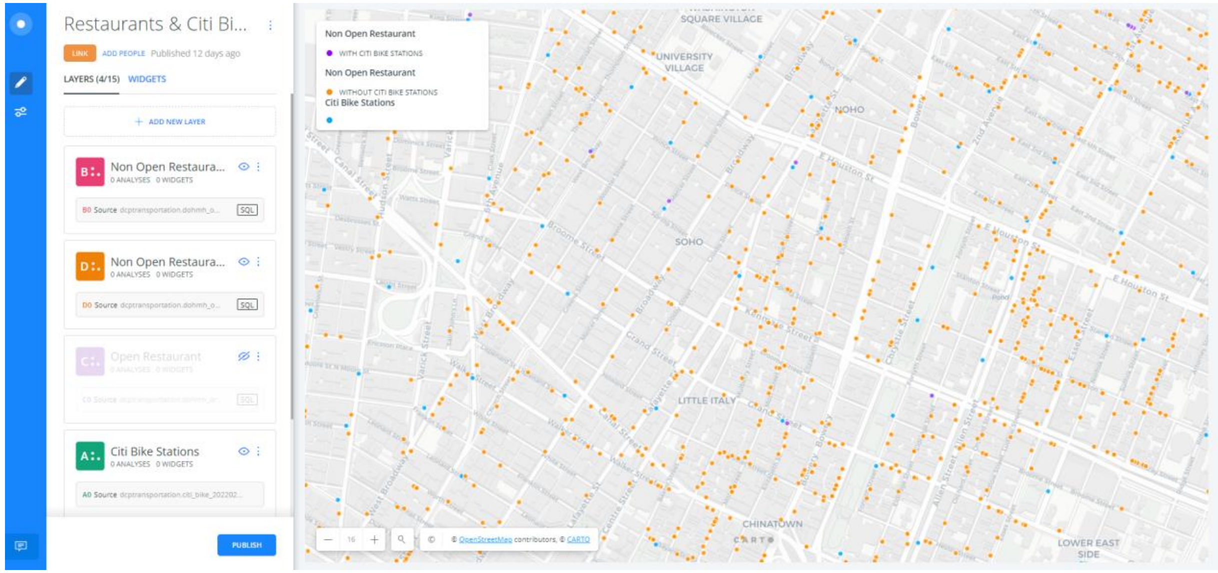
2020 Travel Trends

How the COVID-19 pandemic affected mobility in New York City

NYC Department of City Planning, March 2021

- Web-based mapping software by ESRI
- Creating interactive charts, maps, dashboards, story maps, etc.
- Using StoryMaps to organize final report with interactive charts / maps from other platforms embedded

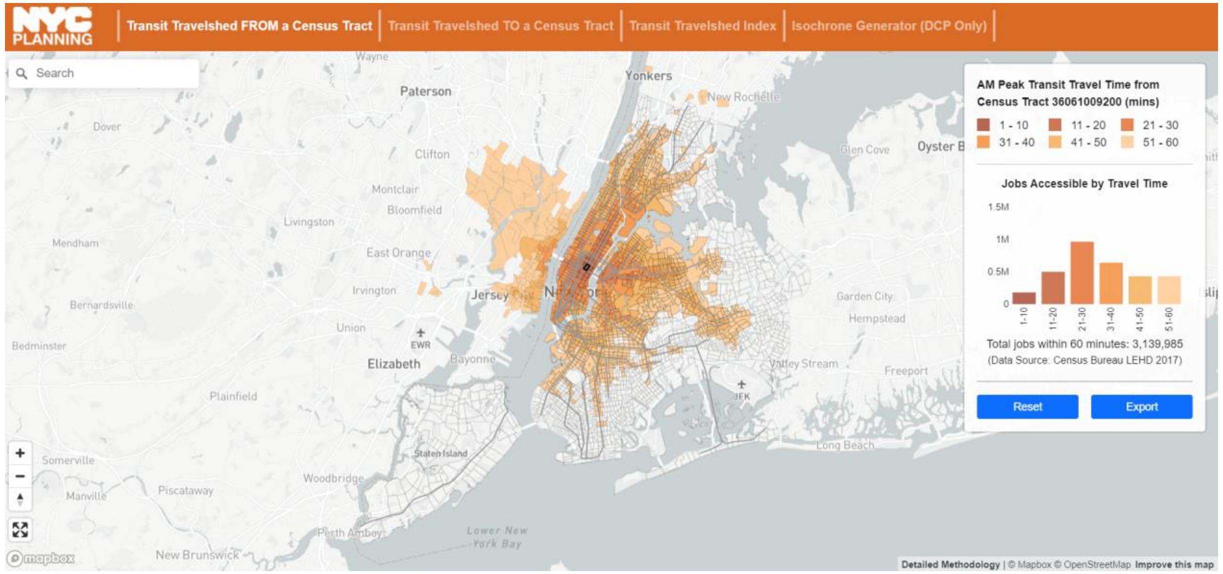
Carto



- Web-based mapping platform
- Storing geospatial data and creating interactive maps
- Customization: SQL for data processing, CSS for styling, and HTML for legend editing

Mapbox GL JS

mapbox



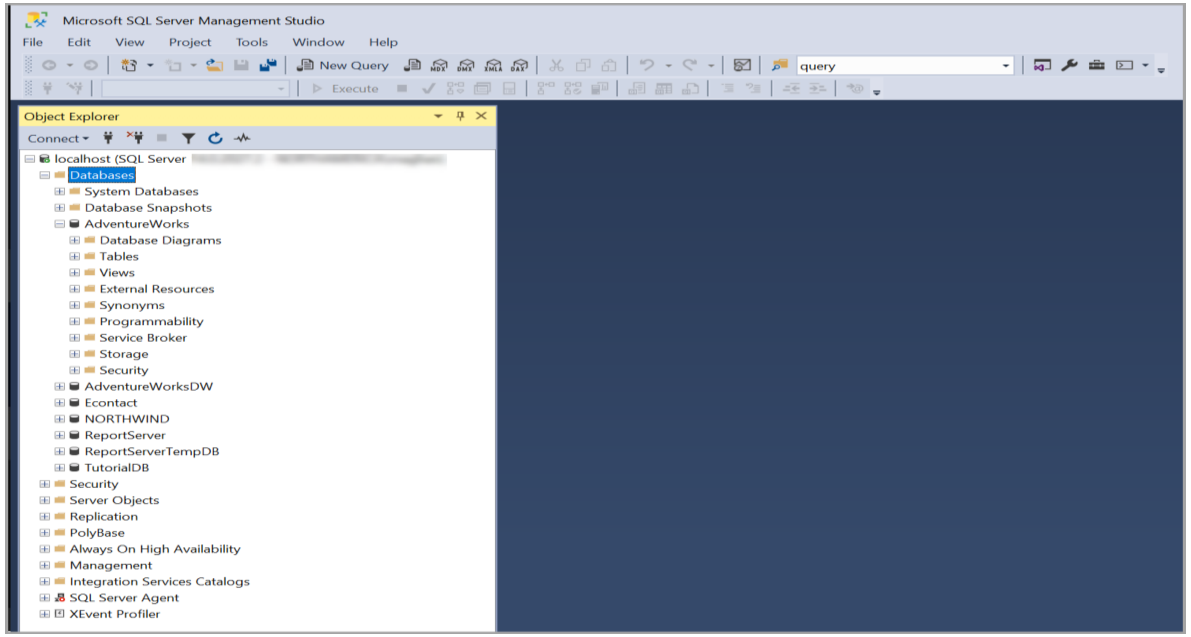
AM Peak Transit Travel Time from Census Tract 36061009200 (mins)

Jobs Accessible by Travel Time

Total jobs within 60 minutes: 3,139,985
(Data Source: Census Bureau LEHD 2017)

- JavaScript mapping library
- Creating interactive maps and highly customized web mapping apps
- Editing tool: Visual Studio Code
- Languages used: HTML, CSS, and JavaScript
- Hosting platform: GitHub Pages

MS SQL Server

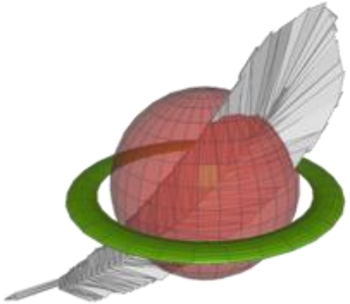


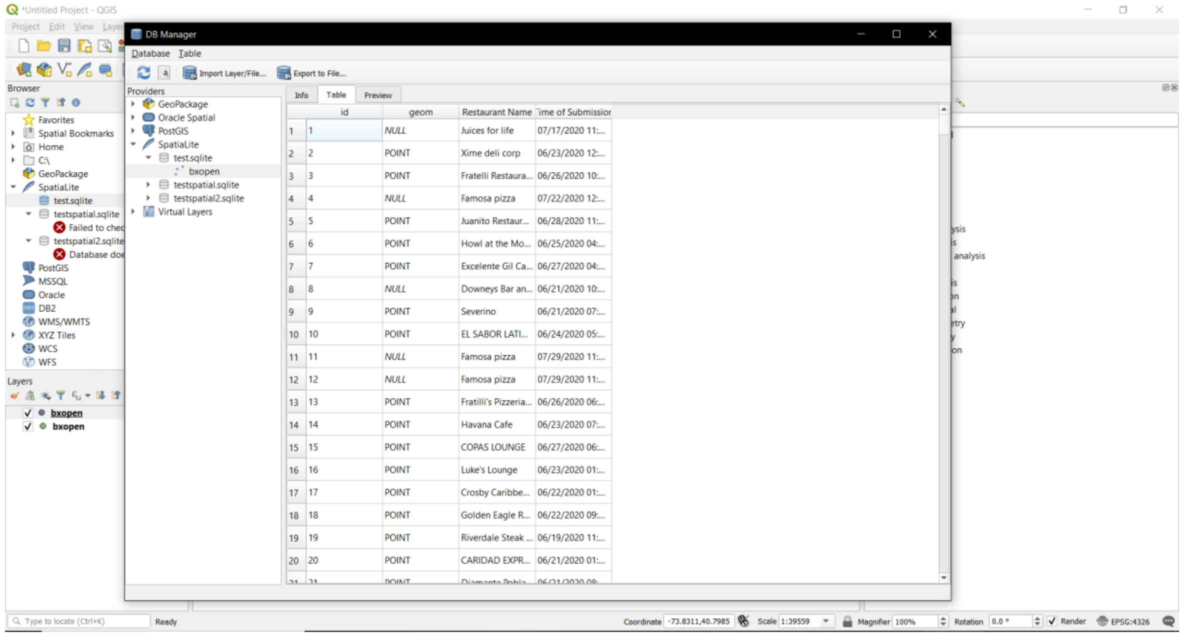
Microsoft SQL Server Management Studio

- localhost (SQL Server [redacted]
  - Databases
    - System Databases
    - Database Snapshots
    - AdventureWorks
      - Database Diagrams
      - Tables
      - Views
      - External Resources
      - Synonyms
      - Programmability
      - Service Broker
      - Storage
      - Security
    - AdventureWorksDW
    - Econtact
    - NORTHWIND
    - ReportServer
    - ReportServerTempDB
    - TutorialDB
  - Security
  - Server Objects
  - Replication
  - PolyBase
  - Always On High Availability
  - Management
  - Integration Services Catalogs
  - SQL Server Agent
  - XEvent Profiler

- Commercial client/server relational database management system by Microsoft
- Deployed and managed by ITD
- Storing confidential datasets (TLC data)
- Interface: SQL Server Management Studio / DBeaver

SQLite / SpatiaLite

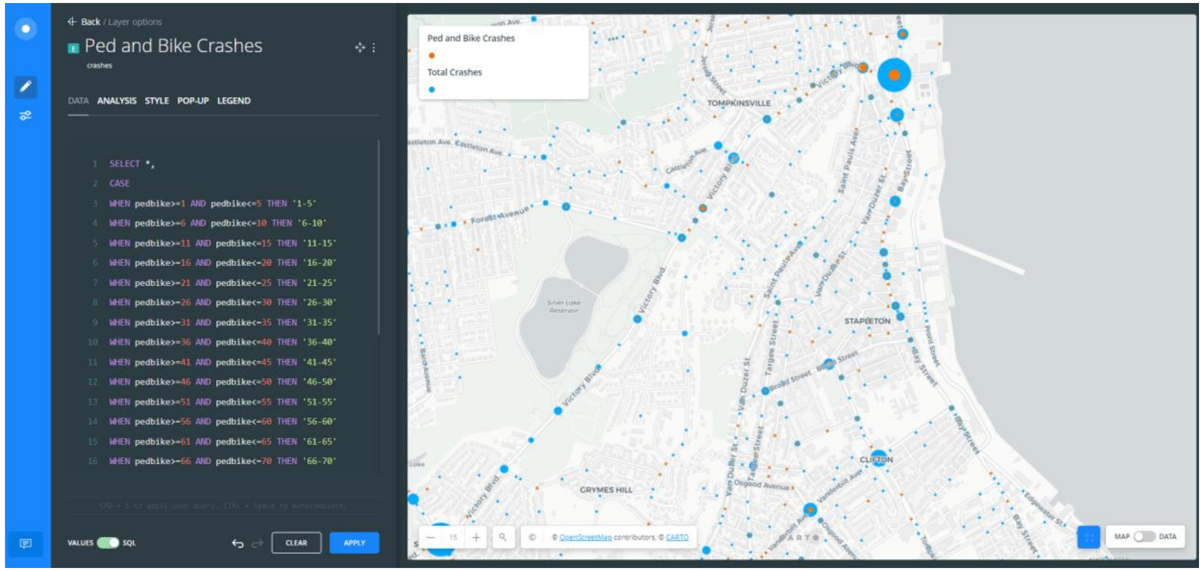




	id	geom	Restaurant Name	'ime of Submissior
1	1	*NULL*	Juices for life	07/17/2020 11:...
2	2	POINT	Xime deli corp	06/23/2020 12:...
3	3	POINT	Fratelli Restaura...	06/26/2020 10:...
4	4	*NULL*	Famosa pizza	07/22/2020 12:...
5	5	POINT	Juanito Restaur...	06/28/2020 11:...
6	6	POINT	Howl at the Mo...	06/25/2020 04:...
7	7	POINT	Excelente Gil Ca...	06/27/2020 04:...
8	8	*NULL*	Downeys Bar an...	06/21/2020 10:...
9	9	POINT	Severino	06/21/2020 07:...
10	10	POINT	EL SABOR LATI...	06/24/2020 05:...
11	11	*NULL*	Famosa pizza	07/29/2020 11:...
12	12	*NULL*	Famosa pizza	07/29/2020 11:...
13	13	POINT	Fratilli's Pizzeria...	06/26/2020 06:...
14	14	POINT	Havana Cafe	06/23/2020 07:...
15	15	POINT	COPAS LOUNGE	06/27/2020 06:...
16	16	POINT	Luke's Lounge	06/23/2020 01:...
17	17	POINT	Crosby Caribbe...	06/22/2020 01:...
18	18	POINT	Golden Eagle R...	06/22/2020 09:...
19	19	POINT	Riverdale Steak ...	06/19/2020 11:...
20	20	POINT	CARIDAD EXPR...	06/21/2020 01:...
21	21	POINT	Diamante Pobla...	06/21/2020 08:...

- SQLite: open-source file-based relational database management system
- SpatiaLite: spatial extension to SQLite
- Storing large public datasets into single database
- Interface: DBeaver / QGIS

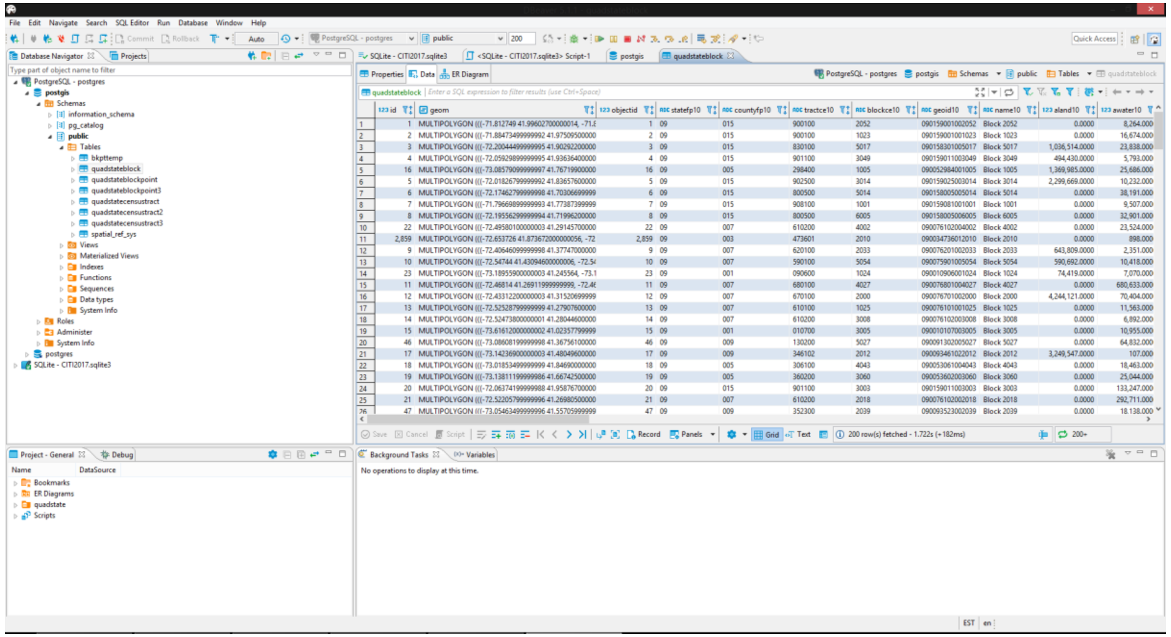
PostgreSQL / PostGIS



```
SELECT *,
CASE
WHEN pedbike>=1 AND pedbike<=5 THEN '1-5'
WHEN pedbike>=6 AND pedbike<=10 THEN '6-10'
WHEN pedbike>=11 AND pedbike<=15 THEN '11-15'
WHEN pedbike>=16 AND pedbike<=20 THEN '16-20'
WHEN pedbike>=21 AND pedbike<=25 THEN '21-25'
WHEN pedbike>=26 AND pedbike<=30 THEN '26-30'
WHEN pedbike>=31 AND pedbike<=35 THEN '31-35'
WHEN pedbike>=36 AND pedbike<=40 THEN '36-40'
WHEN pedbike>=41 AND pedbike<=45 THEN '41-45'
WHEN pedbike>=46 AND pedbike<=50 THEN '46-50'
WHEN pedbike>=51 AND pedbike<=55 THEN '51-55'
WHEN pedbike>=56 AND pedbike<=60 THEN '56-60'
WHEN pedbike>=61 AND pedbike<=65 THEN '61-65'
WHEN pedbike>=66 AND pedbike<=70 THEN '66-70'
```

- PostgreSQL: open-source client/server relational database management system
- PostGIS: spatial extension to PostgreSQL
- Storing spatial datasets
- Interface: Carto

DBeaver



	id	geom	objectid	statefp10	countyfp10	tractce10	blockce10	geoid10	name10	aland10	awater10
1	1	MULTIPOLYGON (((-71.812749 41.99602700000014, -71.8	1	09	015	900100	2052	090159001002052	Block 2052	0.0000	8,264.000
2	2	MULTIPOLYGON (((-71.88473499999992 41.97509500000	2	09	015	900100	1023	090159001001023	Block 1023	0.0000	16,674.000
3	3	MULTIPOLYGON (((-72.20044499999995 41.90292200000	3	09	015	830100	5017	090158301005017	Block 5017	1,036,514.0000	23,838.000
4	4	MULTIPOLYGON (((-72.05929899999995 41.93636400000	4	09	015	901100	3049	090159011003049	Block 3049	494,430.0000	5,793.000
5	16	MULTIPOLYGON (((-73.08579099999997 41.76719900000	16	09	005	298400	1005	090052984001005	Block 1005	1,369,985.0000	25,686.000
6	5	MULTIPOLYGON (((-72.01826799999992 41.83657600000	5	09	015	902500	3014	090159025003014	Block 3014	2,299,669.0000	10,232.000
7	6	MULTIPOLYGON (((-72.17462799999998 41.70306699999	6	09	015	800500	5014	090158005005014	Block 5014	0.0000	38,191.000
8	7	MULTIPOLYGON (((-71.79669899999993 41.77387399999	7	09	015	908100	1001	090159081001001	Block 1001	0.0000	9,507.000
9	8	MULTIPOLYGON (((-72.19556299999994 41.71996200000	8	09	015	800500	6005	090158005006005	Block 6005	0.0000	32,901.000
10	22	MULTIPOLYGON (((-72.49580100000003 41.29145700000	22	09	007	610200	4002	090076102004002	Block 4002	0.0000	23,524.000
11	2,859	MULTIPOLYGON (((-72.653726 41.87367200000056, -72	2,859	09	003	473601	2010	090034736012010	Block 2010	0.0000	898.000
12	9	MULTIPOLYGON (((-72.40646099999998 41.37747000000	9	09	007	620100	2033	090076201002033	Block 2033	643,809.0000	2,351.000
13	10	MULTIPOLYGON (((-72.54744 41.43094600000006, -72.54	10	09	007	590100	5054	090075901005054	Block 5054	590,692.0000	10,418.000
14	23	MULTIPOLYGON (((-73.18955900000003 41.245564, -73.1	23	09	001	090600	1024	090010906001024	Block 1024	74,419.0000	7,070.000
15	11	MULTIPOLYGON (((-72.46814 41.26911999999999, -72.46	11	09	007	680100	4027	090076801004027	Block 4027	0.0000	680,633.000
16	12	MULTIPOLYGON (((-72.43312200000003 41.31520699999	12	09	007	670100	2000	090076701002000	Block 2000	4,244,121.0000	70,404.000
17	13	MULTIPOLYGON (((-72.52528799999999 41.27907600000	13	09	007	610100	1025	090076101001025	Block 1025	0.0000	11,563.000
18	14	MULTIPOLYGON (((-72.52473800000001 41.28044600000	14	09	007	610200	3008	090076102003008	Block 3008	0.0000	6,892.000
19	15	MULTIPOLYGON (((-73.61612000000002 41.02357799999	15	09	001	010700	3005	090010107003005	Block 3005	0.0000	10,955.000
20	46	MULTIPOLYGON (((-73.08608199999998 41.36756100000	46	09	009	130200	5027	090091302005027	Block 5027	0.0000	64,832.000
21	17	MULTIPOLYGON (((-73.14236900000003 41.48049600000	17	09	009	346102	2012	090093461022012	Block 2012	3,249,547.0000	107.000
22	18	MULTIPOLYGON (((-73.01853499999999 41.84690000000	18	09	005	306100	4043	090053061004043	Block 4043	0.0000	18,463.000
23	19	MULTIPOLYGON (((-73.13811199999986 41.66742500000	19	09	005	360200	3060	090053602003060	Block 3060	0.0000	25,044.000
24	20	MULTIPOLYGON (((-72.06374199999988 41.95876700000	20	09	015	901100	3003	090159011003003	Block 3003	0.0000	133,247.000
25	21	MULTIPOLYGON (((-72.52205799999996 41.26980500000	21	09	007	610200	2018	090076102002018	Block 2018	0.0000	292,711.000
26	47	MULTIPOLYGON (((-73.05463499999996 41.55705999999	47	09	009	352300	2039	090093523002039	Block 2039	0.0000	18.138.000

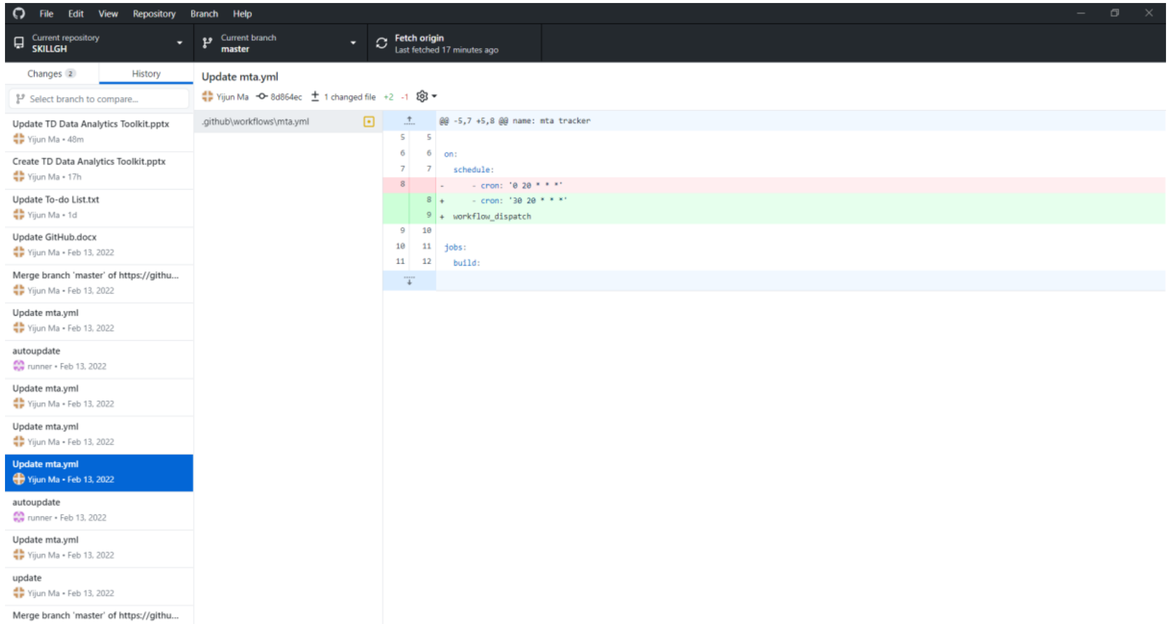
200 row(s) fetched - 1.722s (+182ms)

No operations to display at this time.

- Open-source database administration tool

- Supports CSV, MS SQL Server, SQLite, PostgreSQL, MySQL, Oracle, IBM DB2, MS Access, etc.

GitHub



Update TD Data Analytics Toolkit.pptx
Yijun Ma • 48m

Create TD Data Analytics Toolkit.pptx
Yijun Ma • 17h

Update To-do List.txt
Yijun Ma • 1d

Update GitHub.docx
Yijun Ma • Feb 13, 2022

Merge branch 'master' of https://githu...
Yijun Ma • Feb 13, 2022

Update mta.yml
Yijun Ma • Feb 13, 2022

autoupdate
runner • Feb 13, 2022

Update mta.yml
Yijun Ma • Feb 13, 2022

Update mta.yml
Yijun Ma • Feb 13, 2022

Update mta.yml
Yijun Ma • Feb 13, 2022

autoupdate
runner • Feb 13, 2022

Update mta.yml
Yijun Ma • Feb 13, 2022

update
Yijun Ma • Feb 13, 2022

Merge branch 'master' of https://githu...

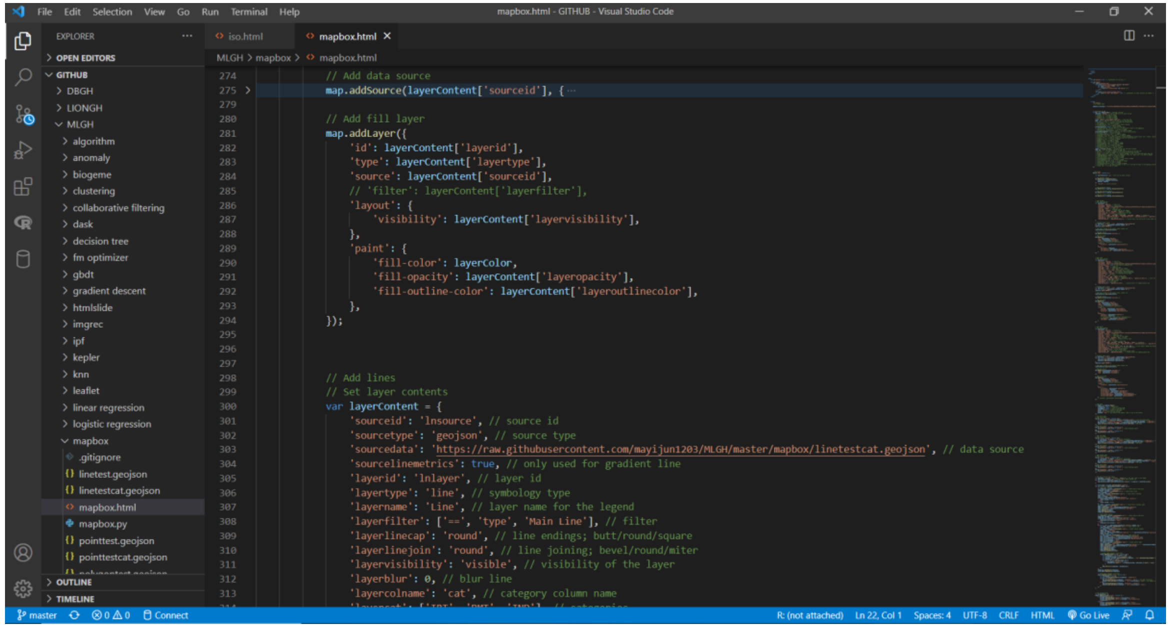
Update mta.yml
Yijun Ma 8d864ec 1 changed file +2 -1

.github\workflows\mta.yml

```
@@ -5,7 +5,8 @@ name: mta tracker

 on:
   schedule:
-      - cron: '0 20 * * *'
+      - cron: '30 20 * * *'
+  workflow_dispatch

 jobs:
   build:
```

- Version control and source code management platform
- Storing scripts, hosting web apps, automating updates, etc.
- GitHub Desktop: simple tool to pull and push scripts to the cloud
- GitHub Pages: place to host interactive charts / maps and web apps for public sharing
- GitHub Actions: tool to automate the workflows and update the data regularly

Visual Studio Code

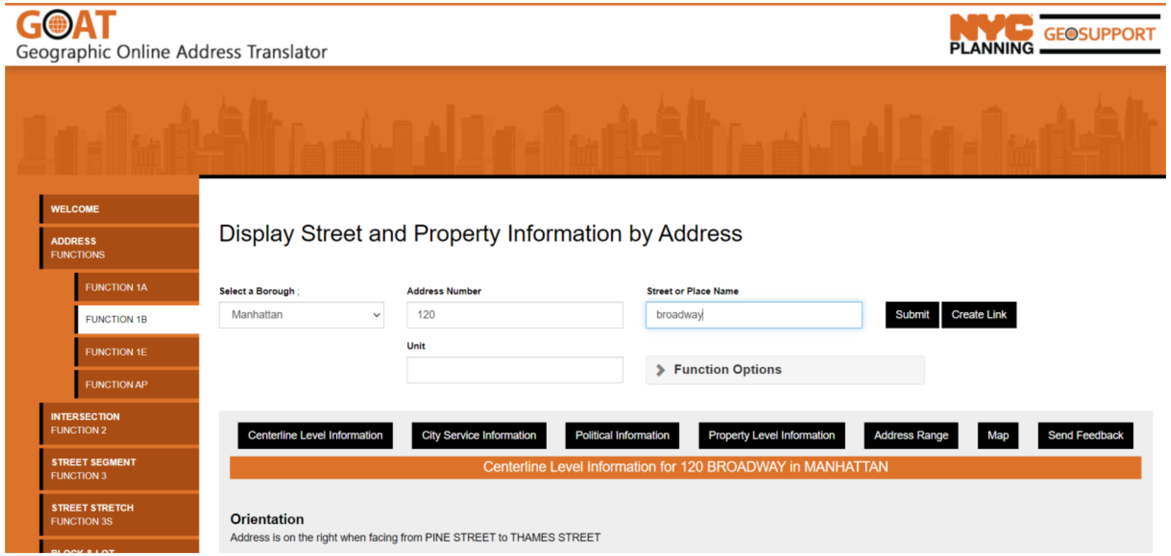


```
            // Add data source
            map.addSource(layerContent['sourceid'], {···

            // Add fill layer
            map.addLayer({
                'id': layerContent['layerid'],
                'type': layerContent['layertype'],
                'source': layerContent['sourceid'],
                // 'filter': layerContent['layerfilter'],
                'layout': {
                    'visibility': layerContent['layervisibility'],
                },
                'paint': {
                    'fill-color': layerColor,
                    'fill-opacity': layerContent['layeropacity'],
                    'fill-outline-color': layerContent['layeroutlinecolor'],
                },
            });



            // Add lines
            // Set layer contents
            var layerContent = {
                'sourceid': 'lnsource', // source id
                'sourcetype': 'geojson', // source type
                'sourcedata': 'https://raw.githubusercontent.com/mayijun1203/MLGH/master/mapbox/linetestcat.geojson', // data source
                'sourcelinemetrics': true, // only used for gradient line
                'layerid': 'lnlayer', // layer id
                'layertype': 'line', // symbology type
                'layername': 'Line', // layer name for the legend
                'layerfilter': ['==', 'type', 'Main Line'], // filter
                'layerlinecap': 'round', // line endings; butt/round/square
                'layerlinejoin': 'round', // line joining; bevel/round/miter
                'layervisibility': 'visible', // visibility of the layer
                'layerblur': 0, // blur line
                'layercolname': 'cat', // category column name
```

- Source code editor developed by Microsoft
- Editing scripts for creating web apps and testing the apps locally
- Supported languages: HTML, CSS, JavaScript, SQL, Python, R, etc.

Geosupport

GEOSUPPORT™
SYSTEM



Display Street and Property Information by Address

Select a Borough : Manhattan

Address Number: 120

Street or Place Name: broadway

Unit:

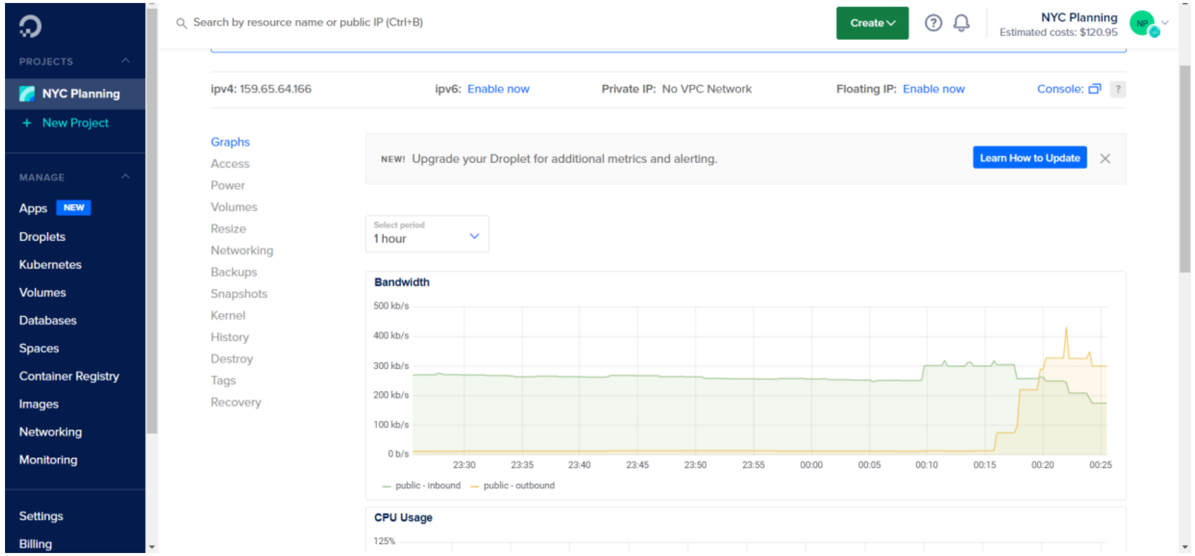
Centerline Level Information for 120 BROADWAY in MANHATTAN

Orientation

Address is on the right when facing from PINE STREET to THAMES STREET

- Geocoding package for NYC addresses
- GOAT: web app for geocoding individual address
- GBAT: batch geocoding multiple addresses
- Geoclient (DOITT) / Geoservice (DCP): geocoding API
- Python-Geosupport: python package to geocode large datasets (especially confidential data)

DigitalOcean



Estimated costs: $120.95

ipv4: 159.65.64.166 ipv6: Enable now Private IP: No VPC Network Floating IP: Enable now Console:

NEW! Upgrade your Droplet for additional metrics and alerting.

Select period: 1 hour

- Cloud computing service provider
- Streaming data, processing massive datasets, hosting web service, etc.
- Connection tool: PuTTY / SSH
- Data transfer tool: WinSCP / Cyberduck