

API-231 / GIS-PubPol

Meeting 11 (Lab Exercise + Problem Set 6)

Yuri M. Zhukov
Visiting Associate Professor of Public Policy
Harvard Kennedy School

February 29, 2024

Goal: geocode historical lynching locations in U.S.

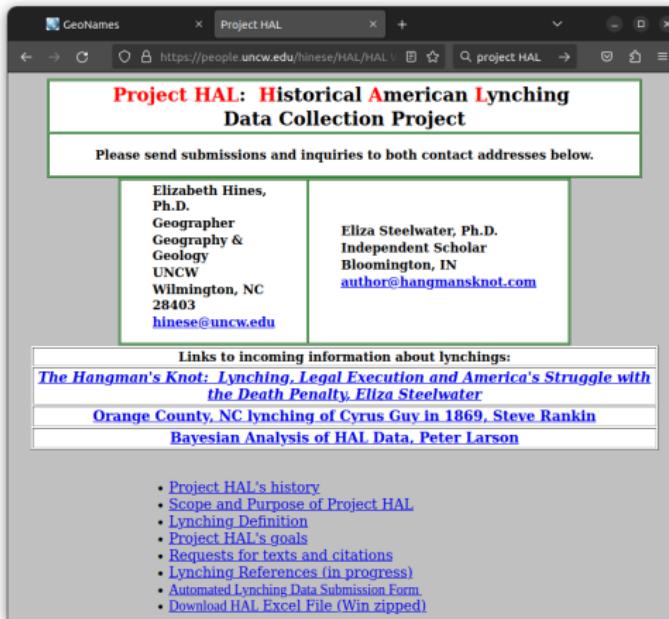


Figure 1: We will geocode these data

Overview of lab exercise and problem set

1. Lab exercise
 - a) Geocode lynching data sample
 - b) Point-in-polygon analysis (lynchings per county)
2. Problem set
 - a) Create a map and boxplot, showing relationship between 100 lynching locations and 1920 U.S. Presidential election results

We will use 2 methods to geocode these lynching locations:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	State	Year	Mo	Day	Victim	County	Race	Sex	Mob	Offense	Note	2nd Name	3rd Name
1861	MS	1891	6	28	Wm. Gates	Clay	Blk	Male		Attempted rape			
1862	MS	1901	5	21	Milt Calvert	Clay	Blk	Male		Attempted rape		Matt Calvert	
1863	MS	1901	5	21	Unnamed Negro	Clay	Blk	Male		Cohabitation			
1864	MS	1913	3	26	Henry Brown	Clay	Blk	Male		Murderous assault			
1865	MS	1915	6	27	Unnamed Negro	Clay	Blk	Male		Entered girl's room			
1866	MS	1915	12	30	Samuel Sykes	Clay	Blk	Male		Attempted murder			
1867	MS	1916	3	18	Jeff Brown	Clay	Blk	Male		Attempted assault (rape)			
1868	MS	1885	5	6	Unnamed Chinese	Coahoma	Other	Male		Assaulted girl			
1869	MS	1886	4	29	Unnamed White	Coahoma	Wht	Male		Cutting levee			
1870	MS	1886	4	29	Unnamed White	Coahoma	Wht	Male		Cutting levee			
1871	MS	1893	8	20	Charles Tart	Coahoma	Blk	Male		Assault	Charles Hart	Sam Wilborn	
1872	MS	1900	11	8	Lit Nabors	Coahoma	Blk	Male		Murder	Uncertain	Kit Nabors	
1873	MS	1902	5	11	Horace Muller	Coahoma	Blk	Male		Unknown		Horace Muller	
1874	MS	1905	11	22	David Sims	Coahoma	Blk	Male		Murder		Davis Simms	
1875	MS	1908	10	11	Jim Davis	Coahoma	Blk	Male		Murderous assault		Joseph Davis	
1876	MS	1908	10	11	Frank Davis	Coahoma	Blk	Male		Murderous assault			
1877	MS	1909	9	6	Hiram McDaniels	Coahoma	Blk	Male		Complicity in murder			

Figure 2: Project HAL raw data

Method 1: Geocode using web service/API (OpenStreetMap)

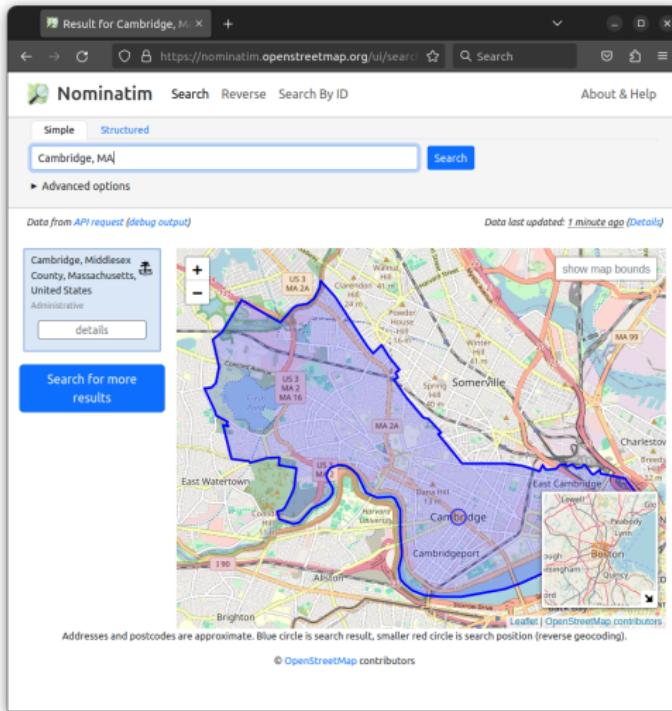


Figure 3: OSM Nominatum

Method 2: Geocode offline, with gazetteer data (GeoNames)

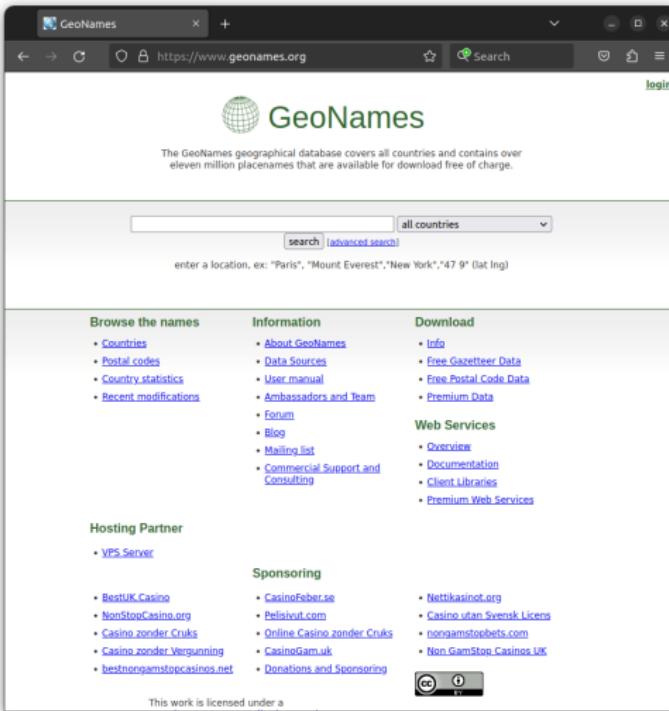


Figure 4: GeoNames

You will then do some *point-in-polygon* analysis to see how lynchings (as part of broader voter suppression efforts) may have impacted vote share in 1920

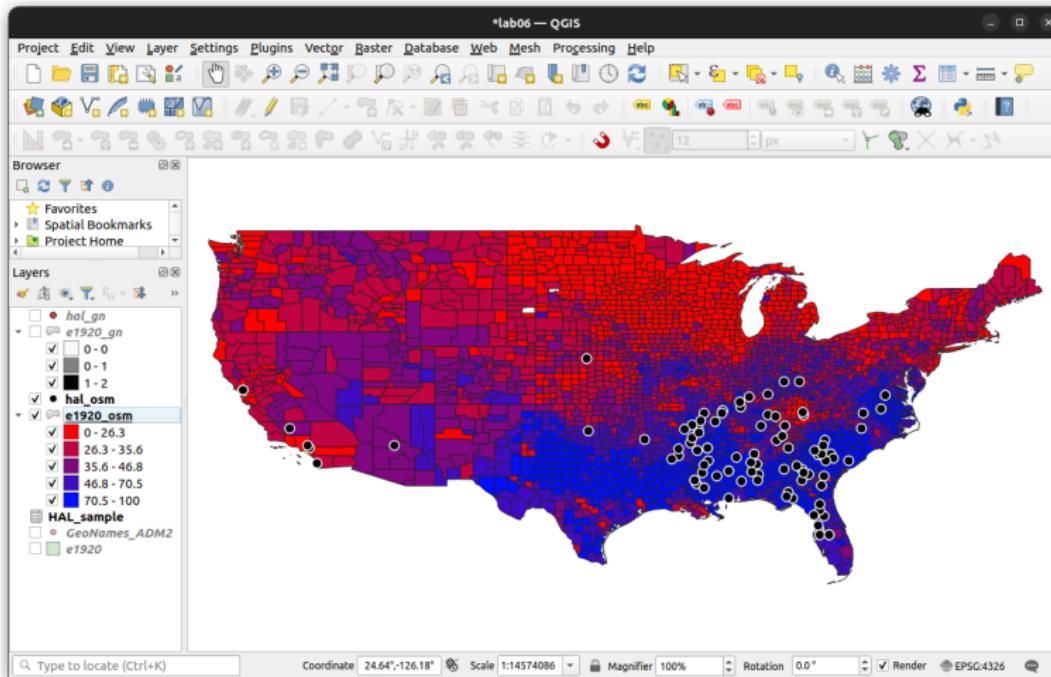


Figure 5: 1920 Presidential elections results

Your assignment: create (1) boxplot of vote share against lynchings, (2) map of geocoded lynching locations vs. election results

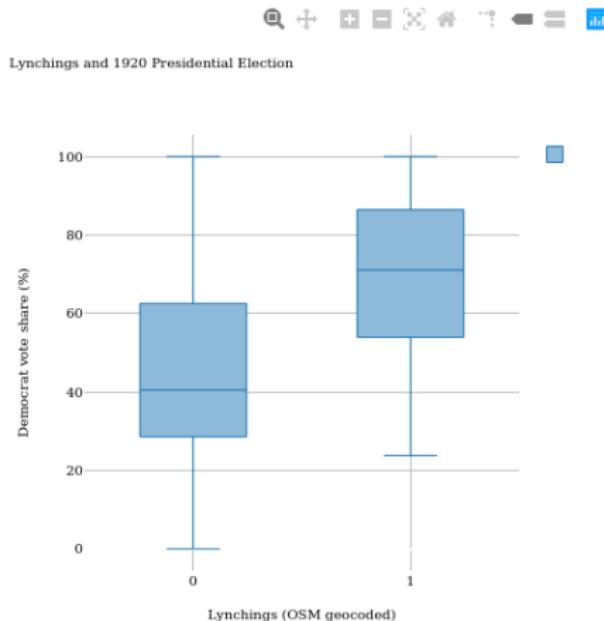


Figure 6: Boxplot

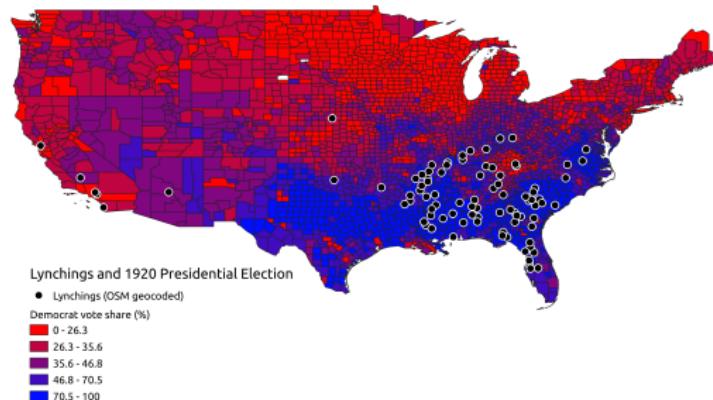


Figure 7: Map

You can make these plots in QGIS or in R. Instructions for both are below.

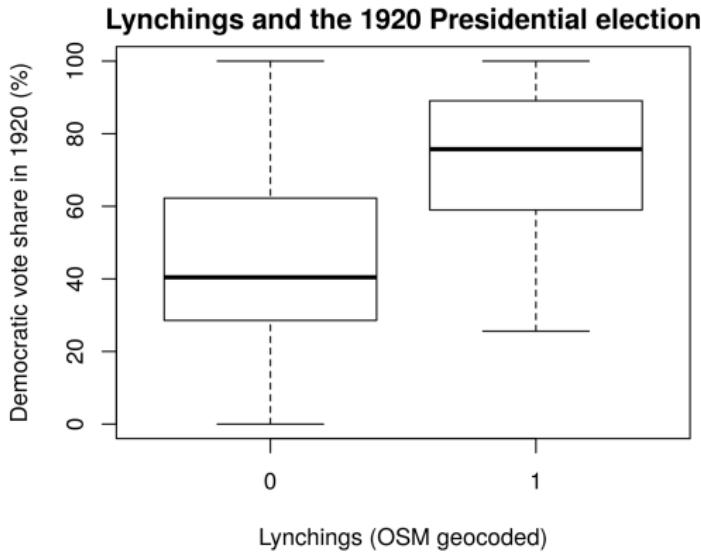


Figure 8: Boxplot in R

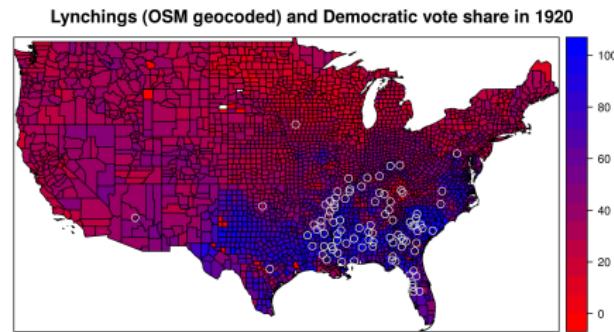


Figure 9: Map in R

We have three (and a half) sources of data:

Category	Type	Format	Data source
Lynchings (sample)	Table (non-geo)	.csv	Project HAL
County gazetteer	Table	.csv	GeoNames
1920 county borders + 1920 election results	Vector (polygons)	.geojson	Newberry Library CQ Voting/Elections

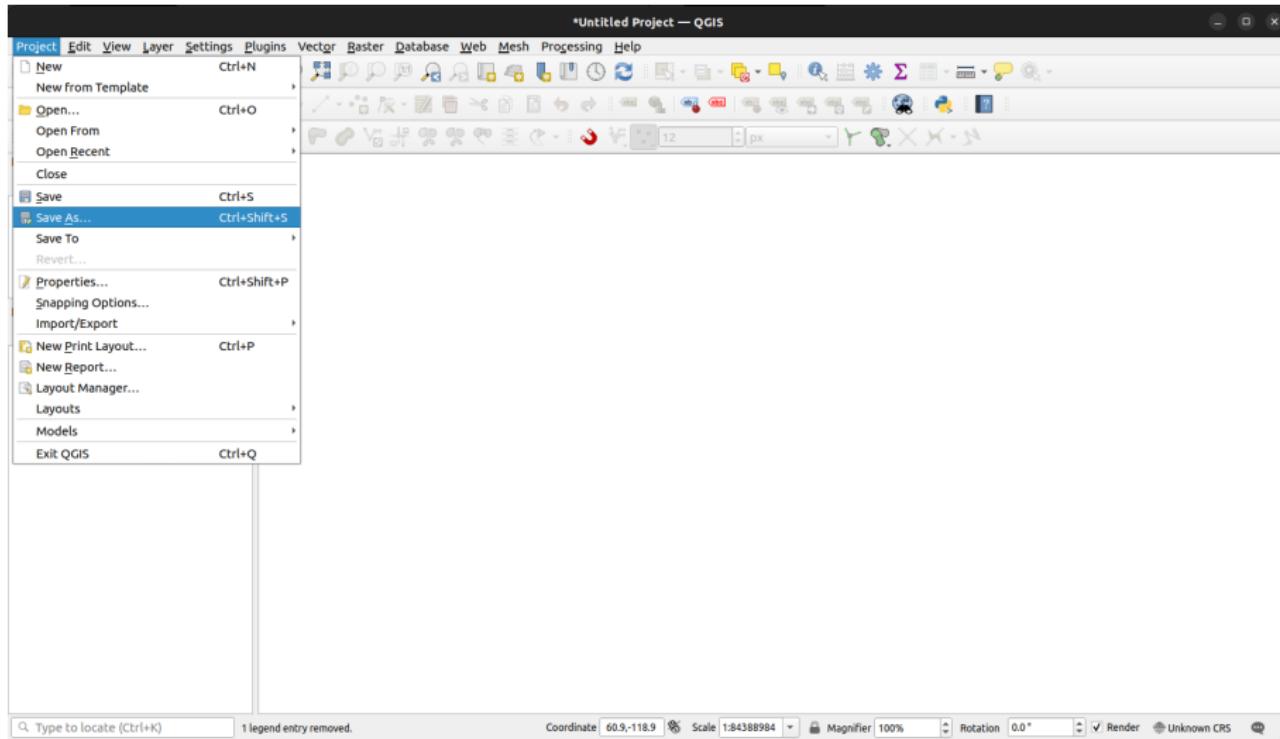
These are all in the PS06.zip file posted on Canvas.

Let's open QGIS...

QGIS

Always save your progress!

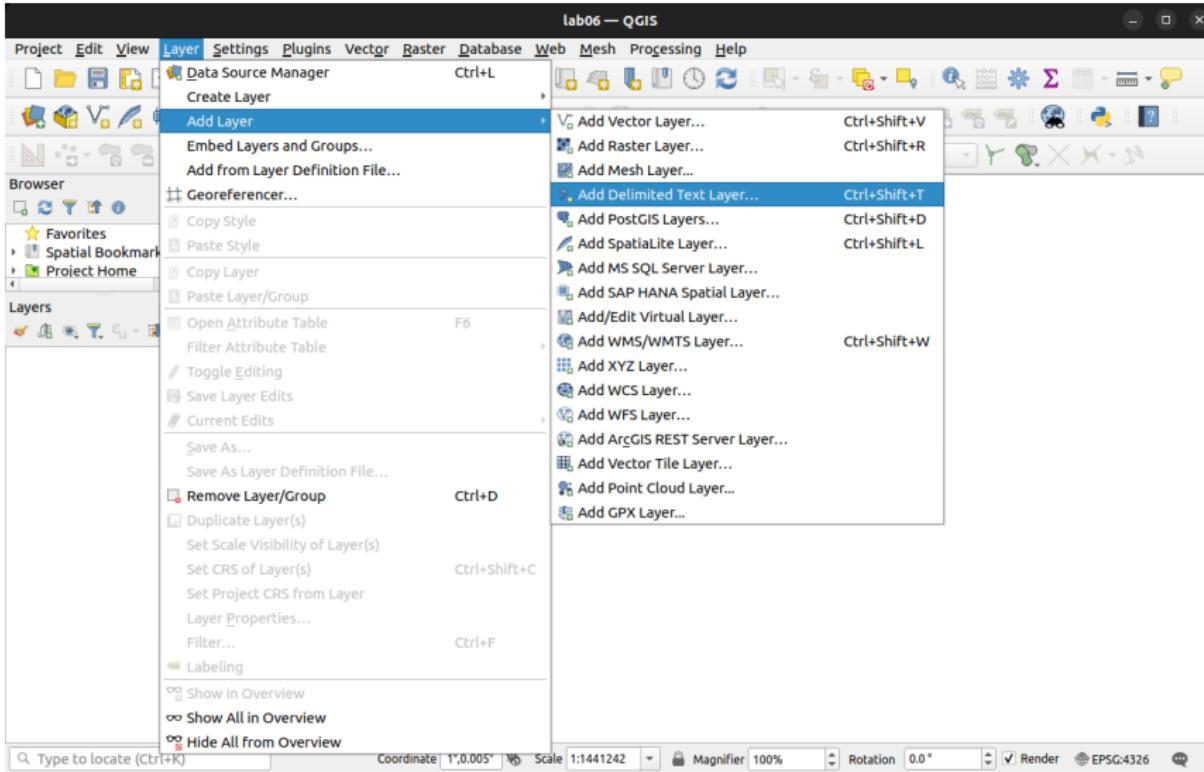
Go to Project → Save As...



Geocoding

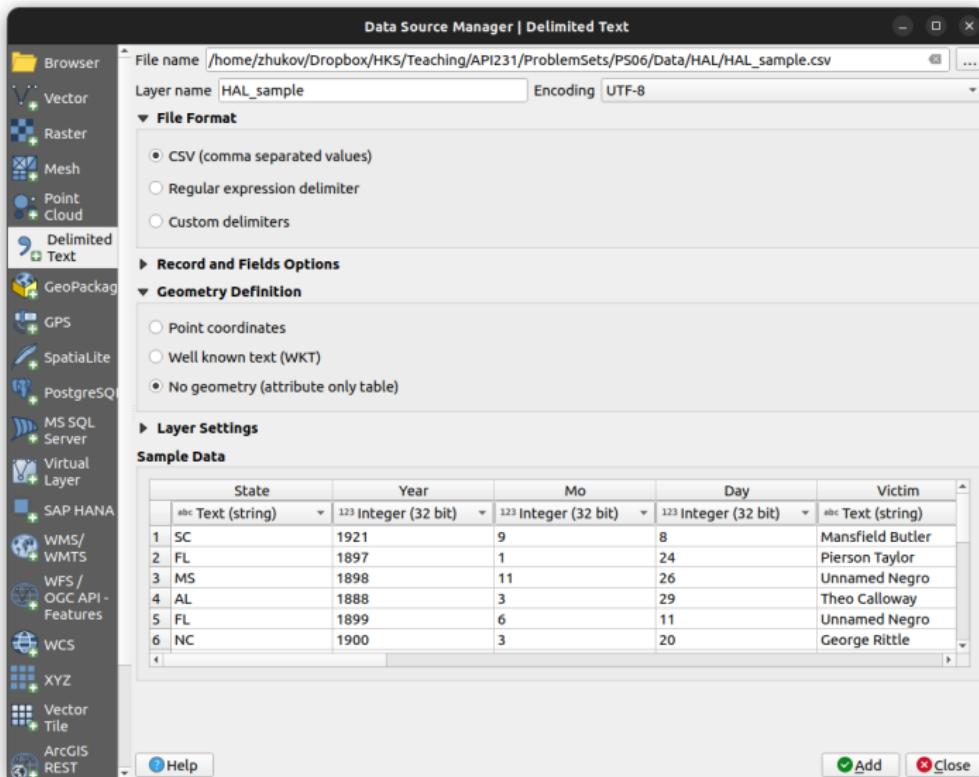
Load the Project HAL data:

Layer → Add Layer → Add Delimited Text Layer...

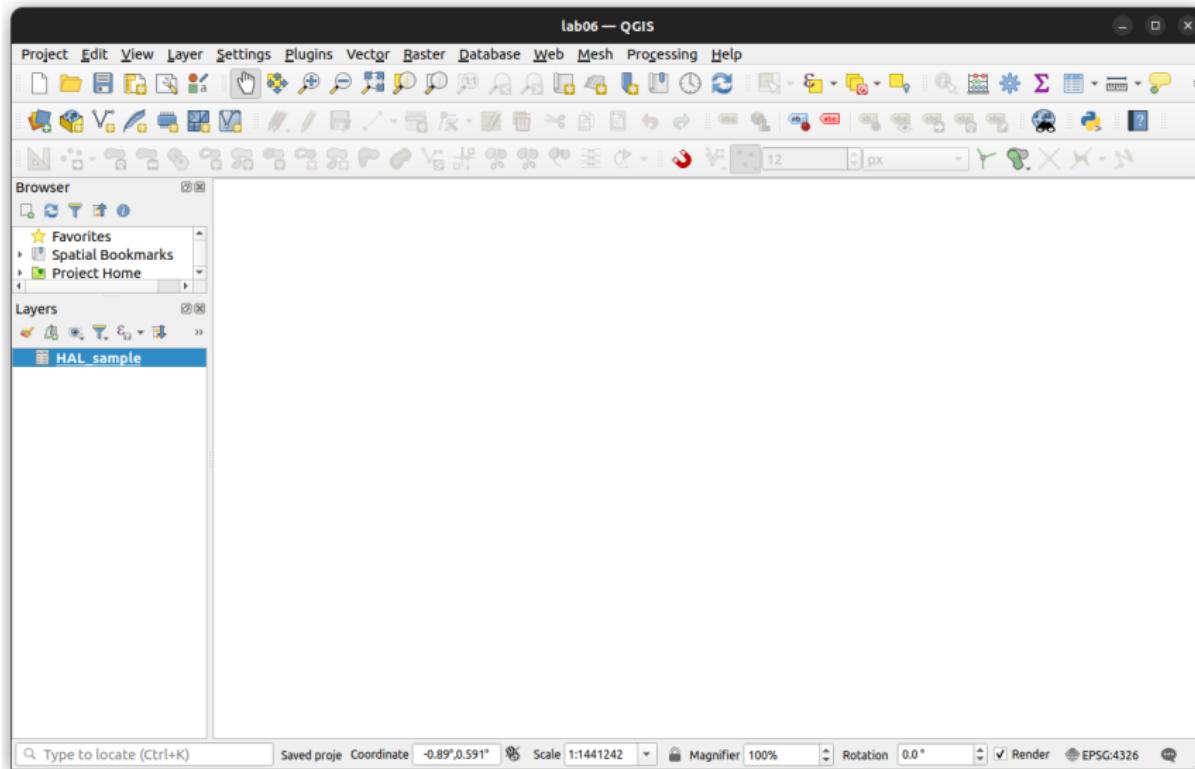


Open the file HAL_sample.csv in the HAL folder.

Geometry definition should be set to No geometry. Click Add '.



You should now see HAL_sample in your layer menu. But there are no points on map, because the data are not geocoded.



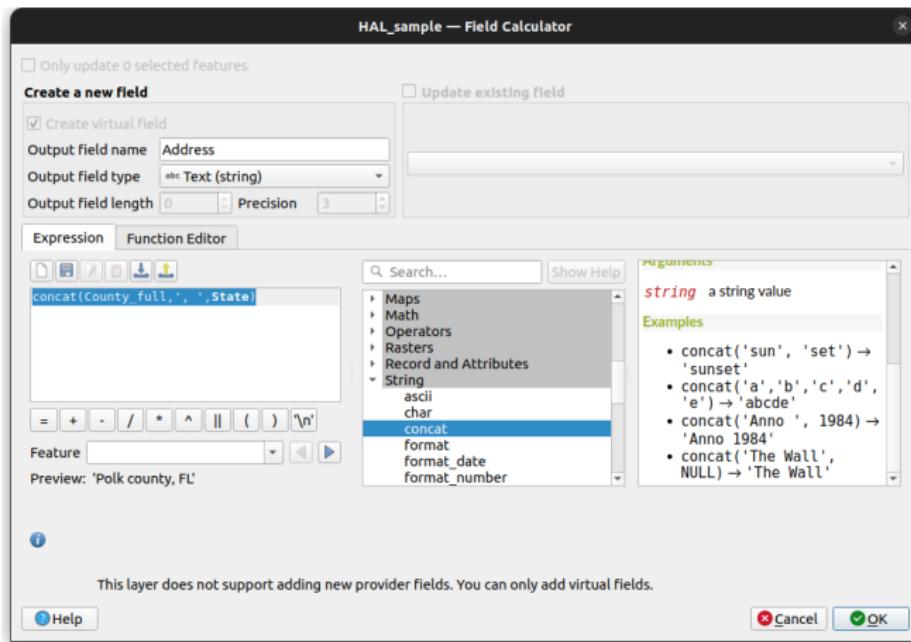
Open the Attribute Table for HAL_sample, navigate to the Field Calculator

HAL_sample — Features Total: 100, Filtered: 100, Selected: 0														
	State	Year	Mo	Day	Victim	County	Race	Sex	Mob	Offense	Note	2nd Name	3rd Name	County_full
1	SC	1921	9	8	Mansfield ...	Aiken	Blk	Male	NULL	Muderous ...	NULL	NULL	NULL	Aiken county
2	FL	1897	1	24	Pierson Tay...	Leon	Blk	Male	NULL	Attempted...	NULL	NULL	NULL	Leon county
3	MS	1898	11	26	Unnamed ...	Lauderdale	Blk	Male	NULL	Assault	Uncertain	NULL	NULL	Lauderdale...
4	AL	1888	3	29	Theo Callo...	Lowndes	Blk	Male	NULL	Murder	NULL	NULL	NULL	Lowndes c...
5	FL	1899	6	11	Unnamed ...	Marion	Blk	Male	Blk	Aided in ly...	NULL	NULL	NULL	Marion cou...
6	NC	1900	3	20	George Rittle	Moore	Blk	Male	NULL	Informer	NULL	George Ritter	NULL	Moore cou...
7	FL	1902	7	29	Alonzo Will...	Pasco	Blk	Male	NULL	Rape	NULL	NULL	NULL	Pasco county
8	AR	1887	12	29	Wm. Herring	Clay	Wht	Male	NULL	Murder	NULL	Wm. Herrig	NULL	Clay county
9	LA	1884	10	24	Unnamed ...	St. Tammany	Blk	Male	NULL	Murder	Uncertain	NULL	NULL	St. Tamma...

Create new field, Address, of type Text (string).

For the Expression, write concat(County_full, ', ', State).

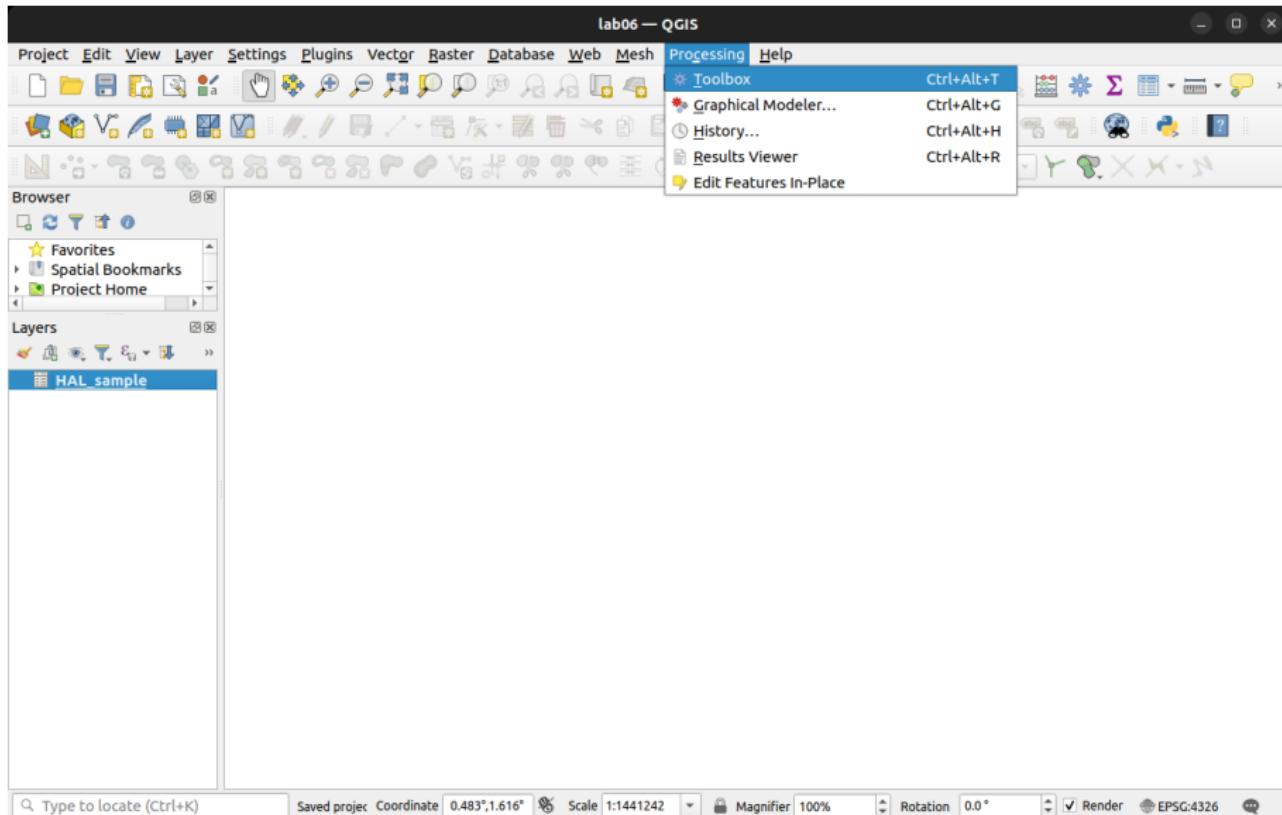
The Preview should show a county-state ID. Click OK



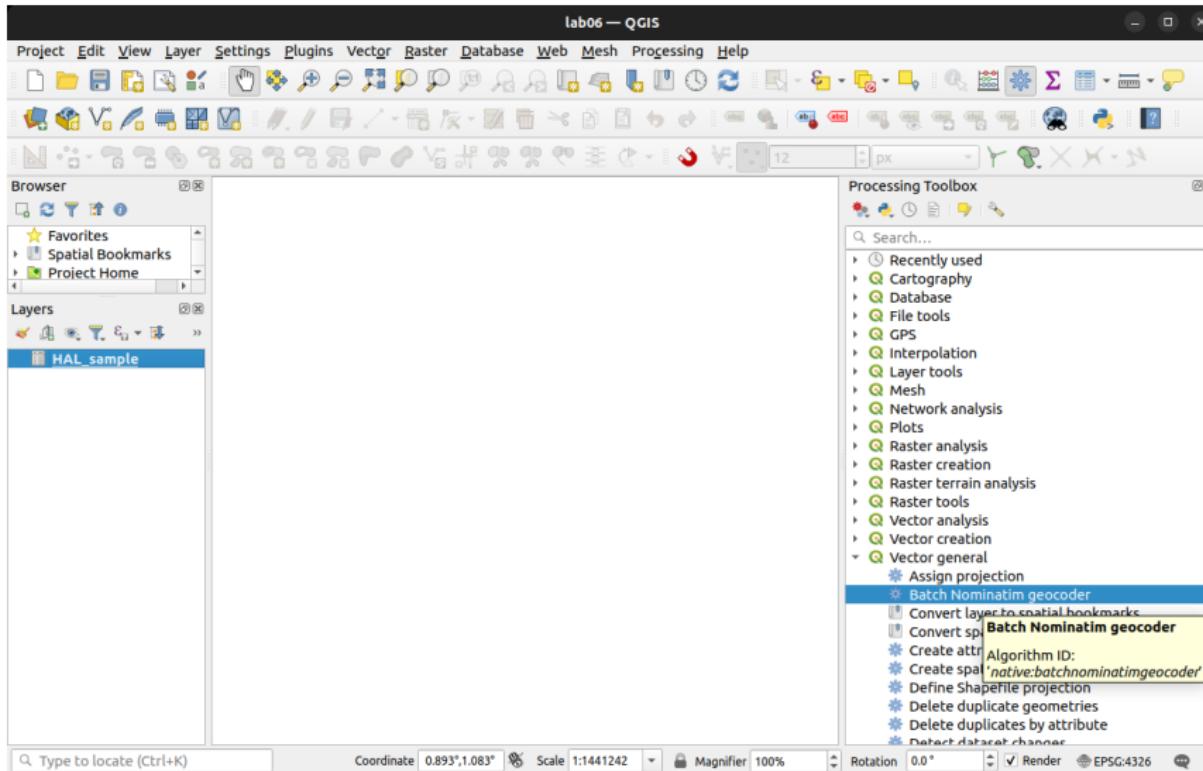
The new field should appear in the Attribute Table

HAL_sample — Features Total: 100, Filtered: 100, Selected: 0															
	State	Year	Mo	Day	Victim	County	Race	Sex	Mob	Offense	Note	2nd Name	3rd Name	County_full	Address
1	SC	1921	9	8	Mansfield ...	Aiken	Blk	Male	NULL	Muderous ...	NULL	NULL	NULL	Aiken county	Aiken coun...
2	FL	1897	1	24	Pierson Tay...	Leon	Blk	Male	NULL	Attempted...	NULL	NULL	NULL	Leon county	Leon count...
3	MS	1898	11	26	Unnamed ...	Lauderdale	Blk	Male	NULL	Assault	Uncertain	NULL	NULL	Lauderdale...	Lauderdale...
4	AL	1888	3	29	Theo Callo...	Lowndes	Blk	Male	NULL	Murder	NULL	NULL	NULL	Lowndes c...	Lowndes c...
5	FL	1899	6	11	Unnamed ...	Marion	Blk	Male	Blk	Aided in ly...	NULL	NULL	NULL	Marion cou...	Marion cou...
6	NC	1900	3	20	George Rittle	Moore	Blk	Male	NULL	Informer	NULL	George Ritter	NULL	Moore cou...	Moore cou...
7	FL	1902	7	29	Alonzo Will...	Pasco	Blk	Male	NULL	Rape	NULL	NULL	NULL	Pasco county	Pasco coun...
8	AR	1887	12	29	Wm. Herring	Clay	Wht	Male	NULL	Murder	NULL	Wm. Herrig	NULL	Clay county	Clay count...
9	LA	1884	10	24	Unnamed ...	St. Tammany	Blk	Male	NULL	Murder	Uncertain	NULL	NULL	St. Tamma...	St. Tamma...

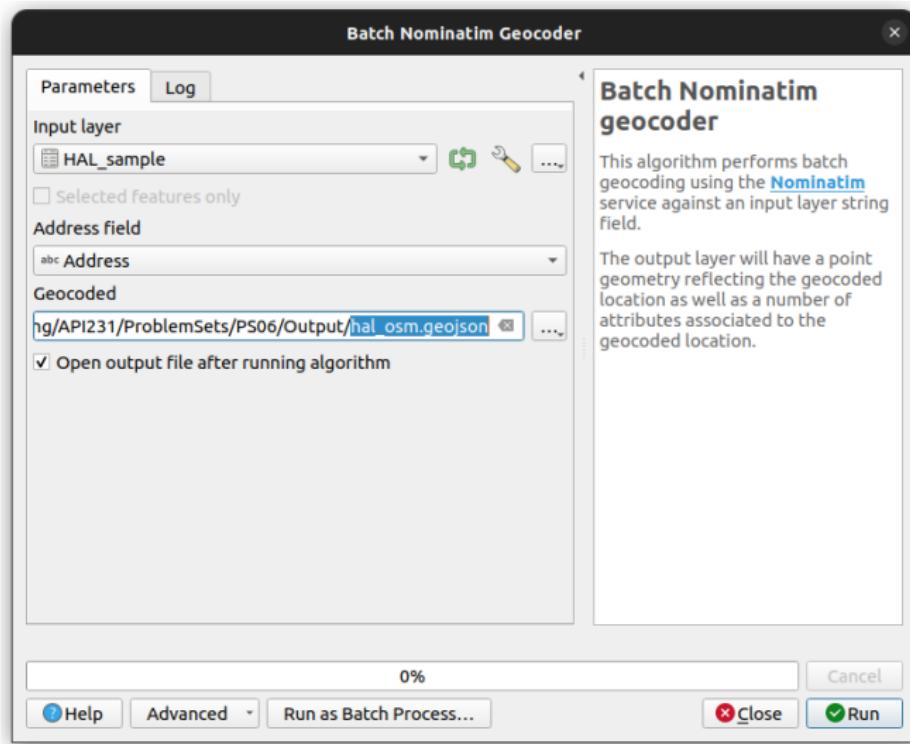
To access the OSM/Nominatum geocoder, go to Processing → Toolbox



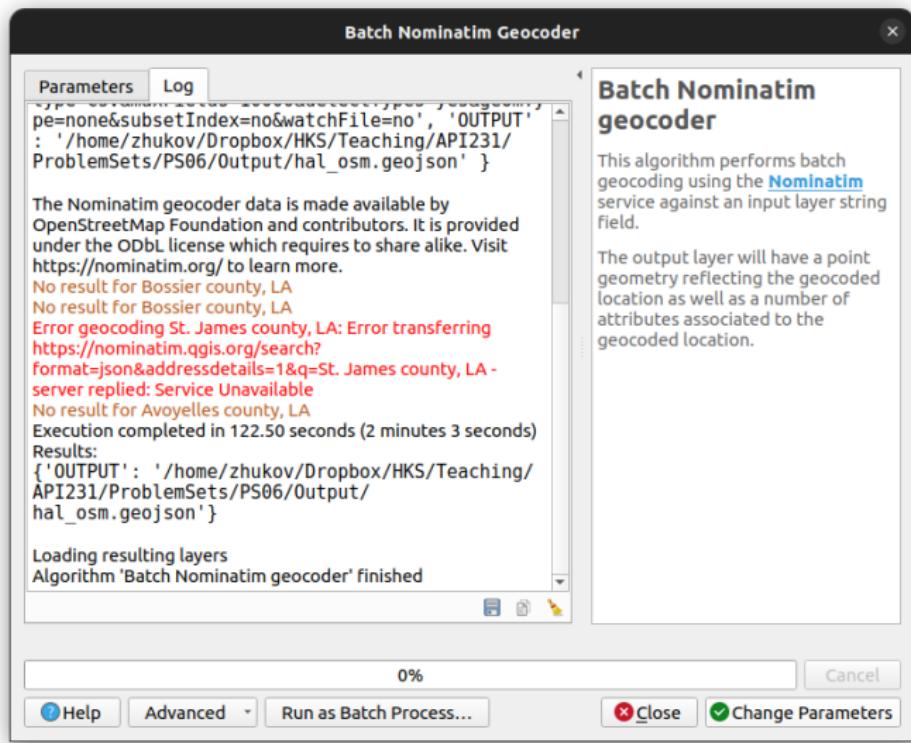
In the Processing Toolbox panel, go to Vector general menu → Batch Nominatum geocoder



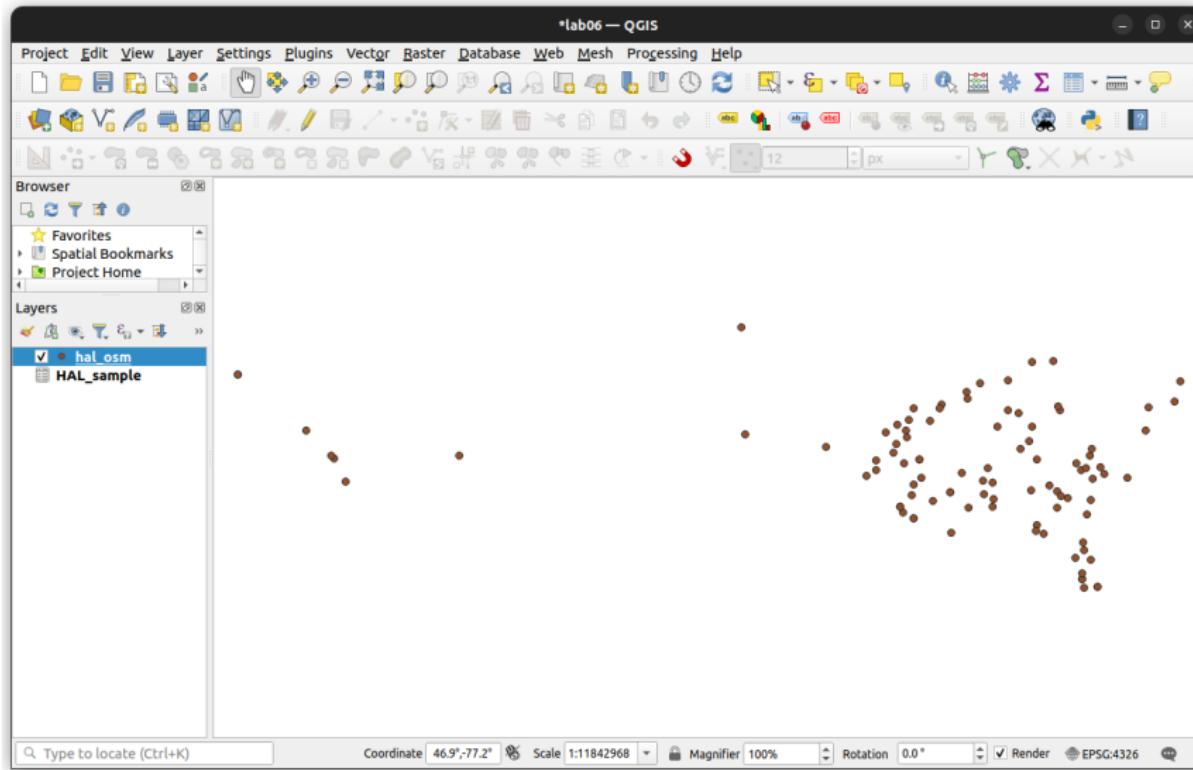
Set Input layer = HAL_sample and Address field = Address.
Save the output to a file called hal_osm.geojson. Click Run



After running, the log may tell you that some addresses could not be geocoded.
Click Close

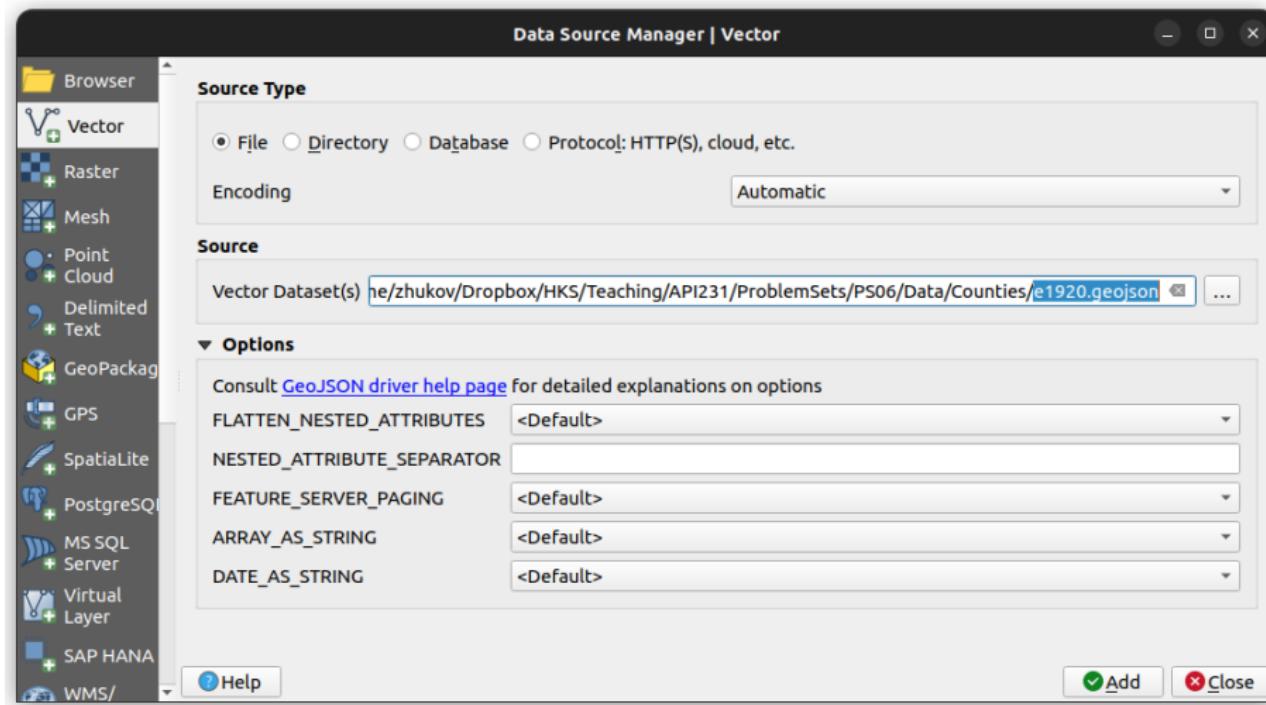


You should see the geocoded points in the main project window, and a new layer, `hal_osm`. Let's plot it against an historical county boundary shapefile.

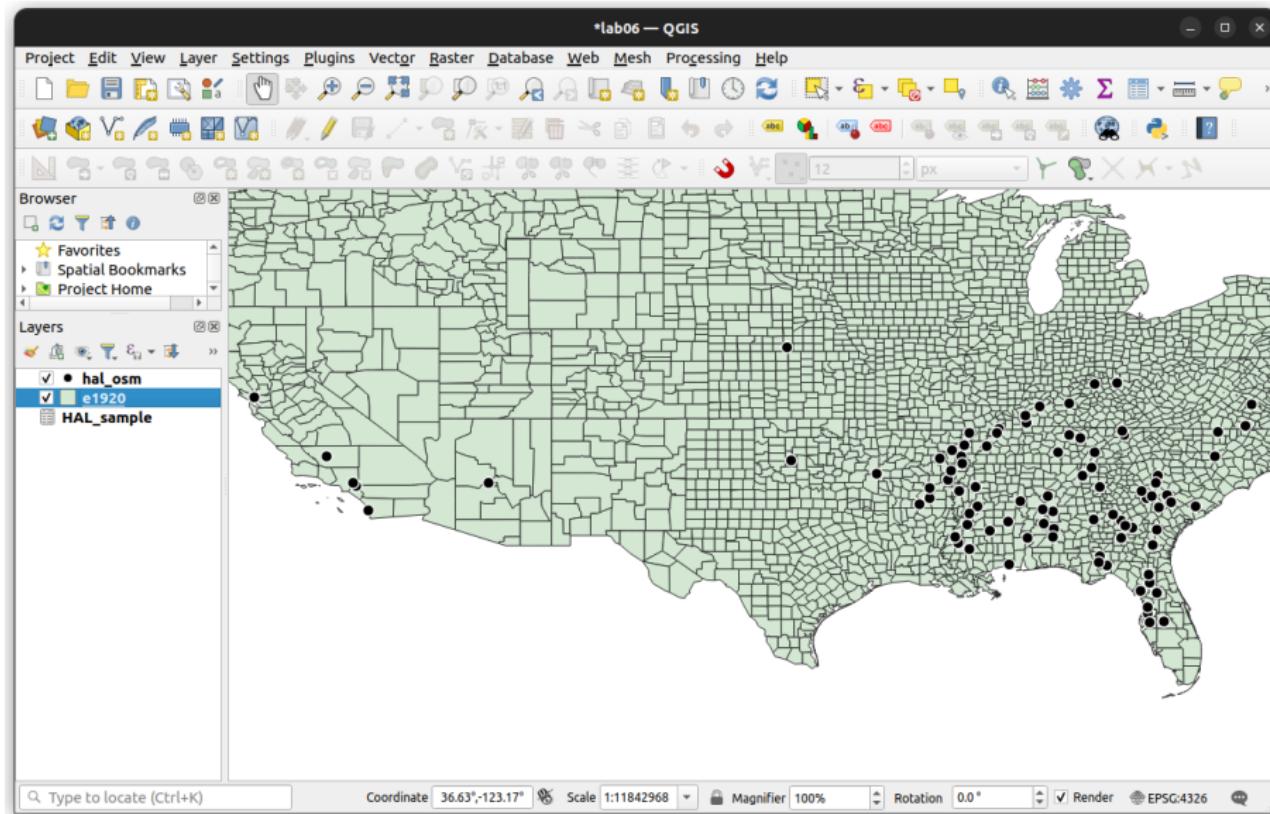


Go to Layer → Add Layer → Add Vector Layer....

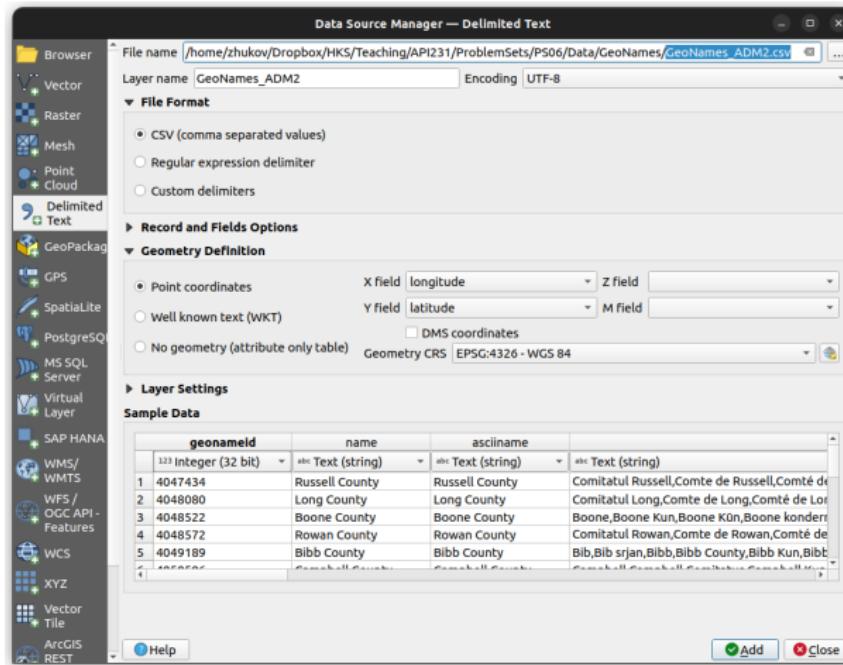
Navigate to e1920.geojson file in Data/Counties folder. Add it to the map.



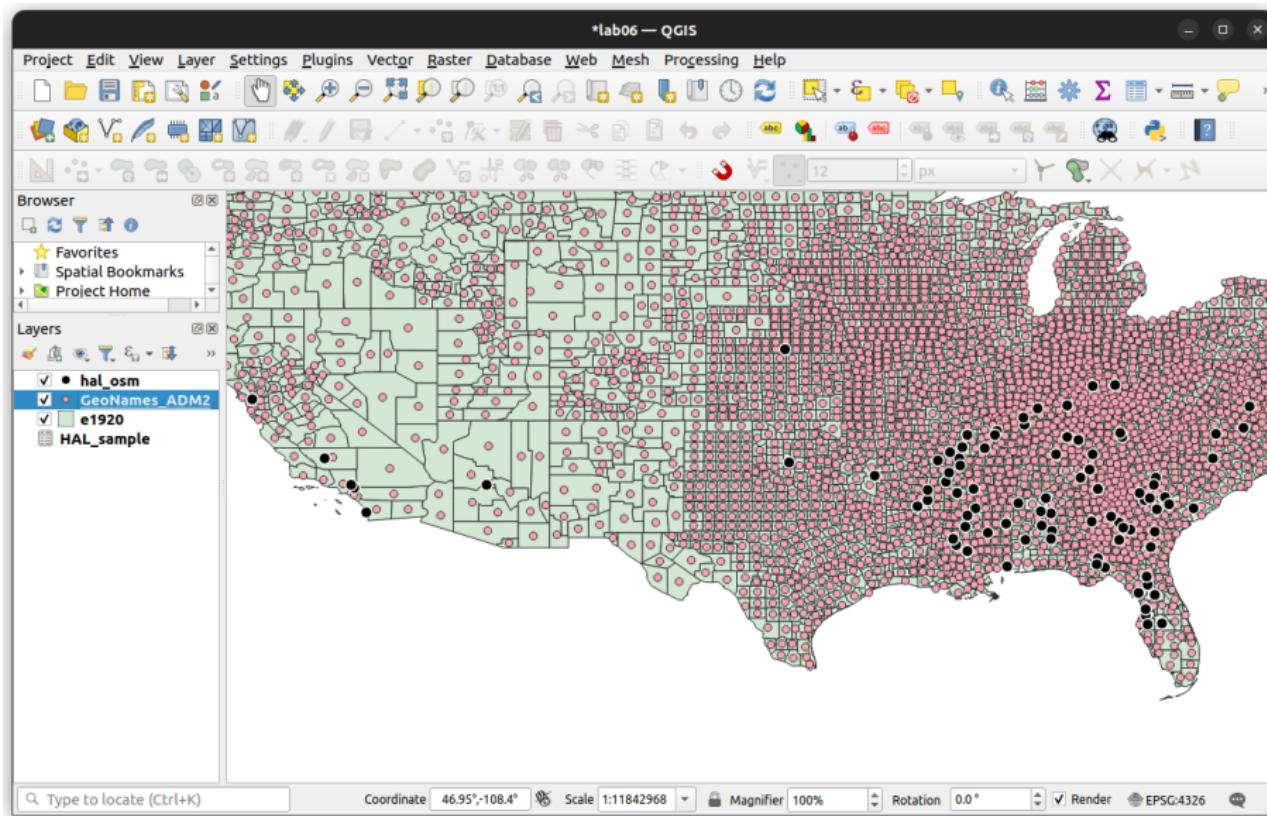
Looks about right. Now let's try geocoding with gazetteer data.



Add the GeoNames gazetteer to the project, using Add Delimited Text Layer.... Load GeoNames_ADM2.csv from the Data/GeoNames folder. Set X field = longitude and Y field = latitude



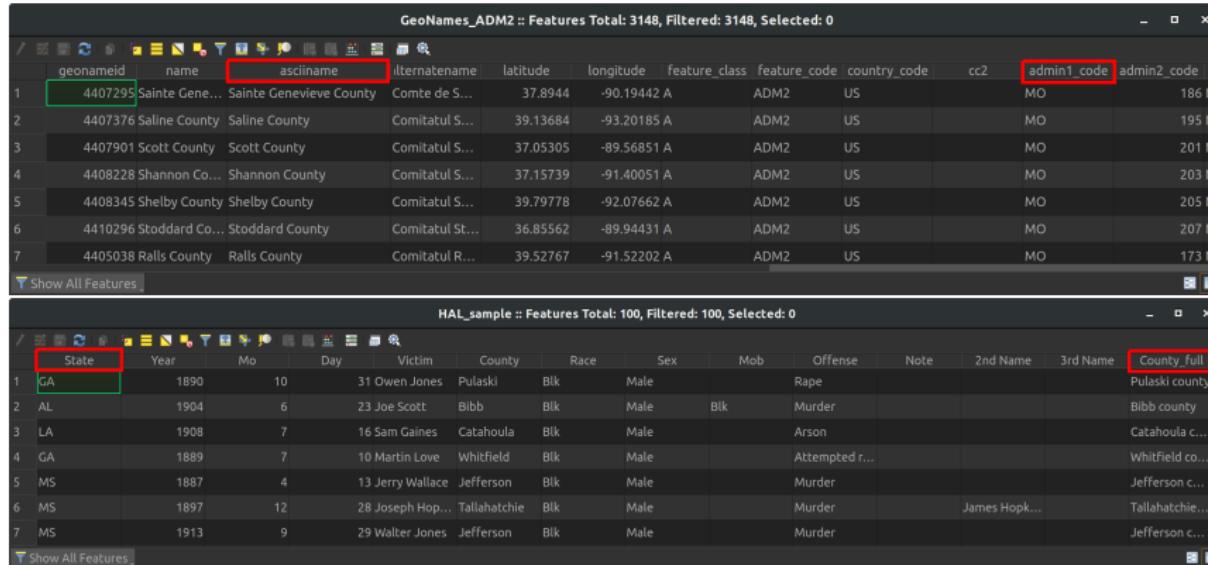
The centroids for (contemporary) counties should appear in the project window.



In the Attribute Table, we see that

- asciiname field in GeoNames_ADMIN2 corresponds to County_full in HAL_sample
- admin1_code corresponds to State

We can use these to create a unique key to match on



The image shows two QGIS attribute tables side-by-side. The top table is titled "GeoNames_ADMIN2 :: Features Total: 3148, Filtered: 3148, Selected: 0". It has columns: geonameid, name, asciiname, alternatename, latitude, longitude, feature_class, feature_code, country_code, cc2, admin1_code, admin2_code, and iso3166_2_code. The "asciiname" column is highlighted with a red border. The bottom table is titled "HAL_sample :: Features Total: 100, Filtered: 100, Selected: 0". It has columns: State, Year, Mo, Day, Victim, County, Race, Sex, Mob, Offense, Note, 2nd Name, 3rd Name, and County_full. The "State" and "County_full" columns are highlighted with red borders.

geonameid	name	asciiname	alternatename	latitude	longitude	feature_class	feature_code	country_code	cc2	admin1_code	admin2_code	iso3166_2_code
1	4407295	Sainte Genevieve County	Comte de S...	37.8944	-90.19442 A	ADM2	US	MO		186 N		
2	4407376	Saline County	Comitatul S...	39.13684	-93.20185 A	ADM2	US	MO		195 N		
3	4407901	Scott County	Comitatul S...	37.05305	-89.56851 A	ADM2	US	MO		201 N		
4	4408228	Shannon Co...	Shannon County	37.15739	-91.40051 A	ADM2	US	MO		203 N		
5	4408345	Shelby County	Shelby County	39.79778	-92.07662 A	ADM2	US	MO		205 N		
6	4410296	Stoddard Co...	Stoddard County	36.85562	-89.94431 A	ADM2	US	MO		207 N		
7	4405038	Ralls County	Comitatul R...	39.52767	-91.52202 A	ADM2	US	MO		173 N		

Show All Features

State	Year	Mo	Day	Victim	County	Race	Sex	Mob	Offense	Note	2nd Name	3rd Name	County_full
1	GA	1890	10	31	Owen Jones	Pulaski	Blk	Male	Rape				Pulaski county
2	AL	1904	6	23	Joe Scott	Bibb	Blk	Male	Murder				Bibb county
3	LA	1908	7	16	Sam Gaines	Catahoula	Blk	Male	Arson				Catahoula c...
4	GA	1889	7	10	Martin Love	Whitfield	Blk	Male	Attempted r...				Whitfield co...
5	MS	1887	4	13	Jerry Wallace	Jefferson	Blk	Male	Murder				Jefferson c...
6	MS	1897	12	28	Joseph Hop...	Tallahatchie	Blk	Male	Murder	James Hop...			Tallahatchie c...
7	MS	1913	9	29	Walter Jones	Jefferson	Blk	Male	Murder				Jefferson c...

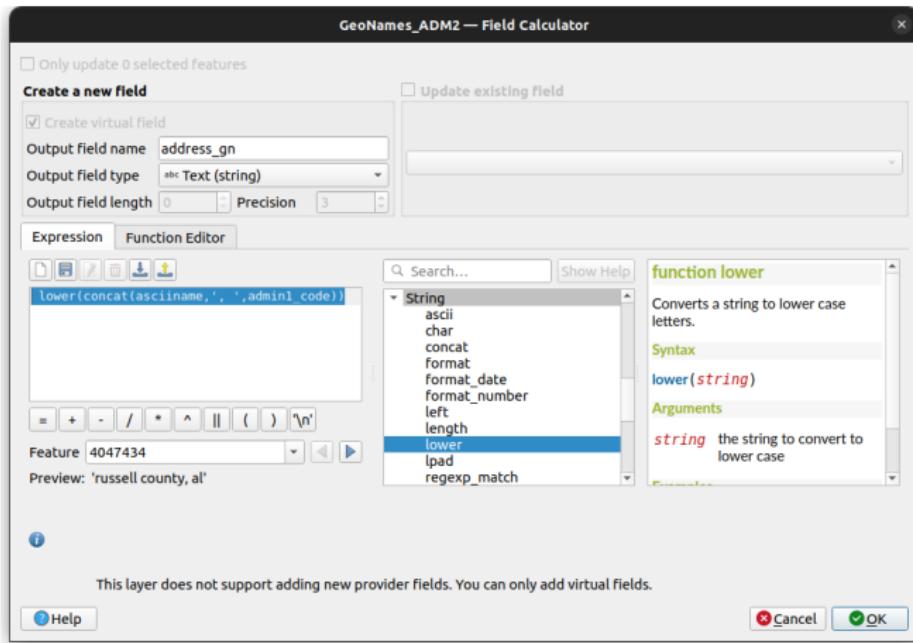
Show All Features

Open the Field Calculator for GeoNames_ADM2.

Create a new variable called address_gn of type Text (string).

Set Expression to lower(concat(asciiname, ', ', admin1_code)).

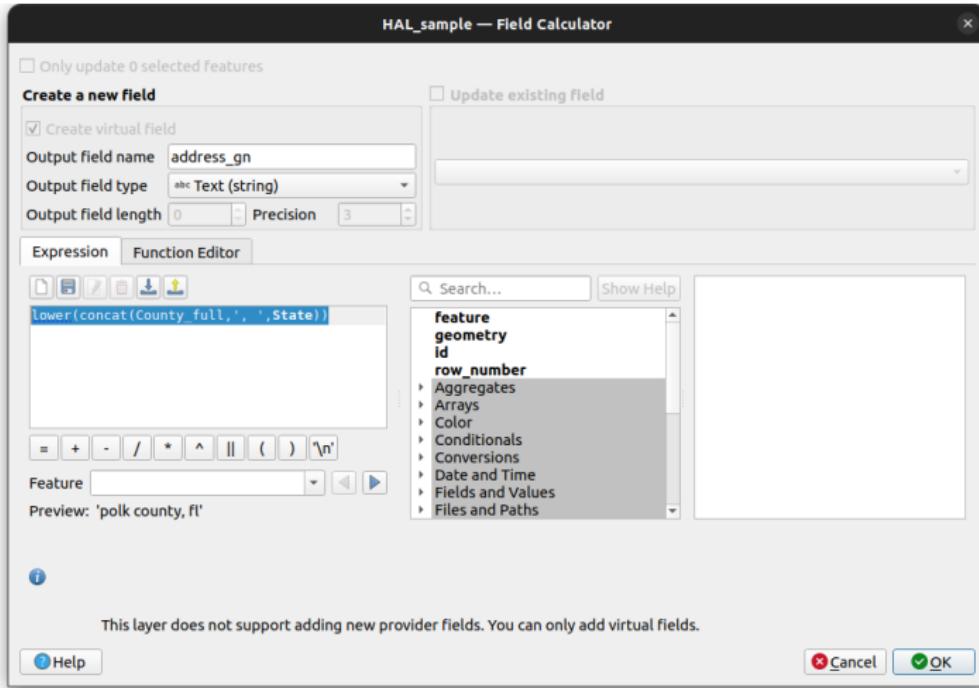
The Preview should show a *lower case* county-state ID. Click OK



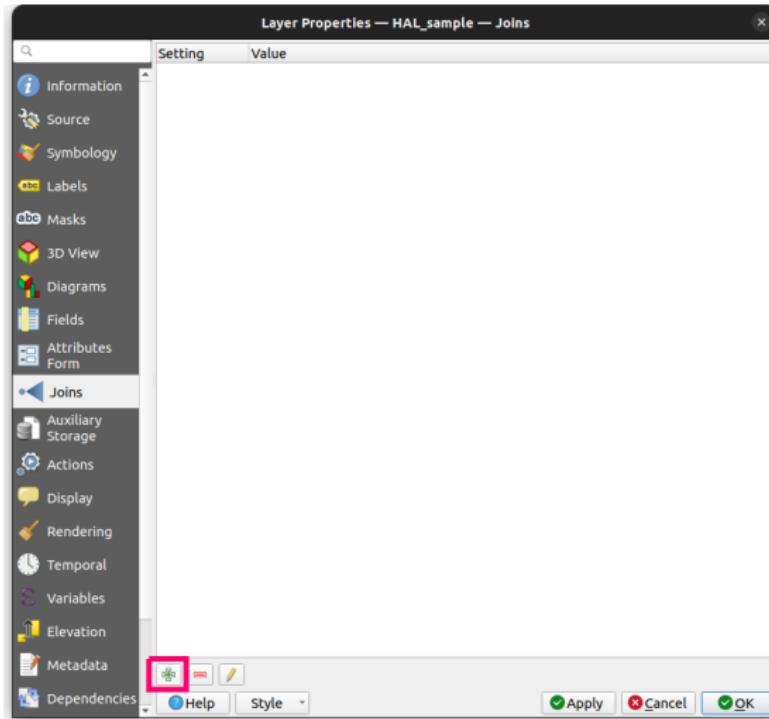
Now open Field Calculator for HAL_sample.

Create the same new field, address_gn, of type Text (string).

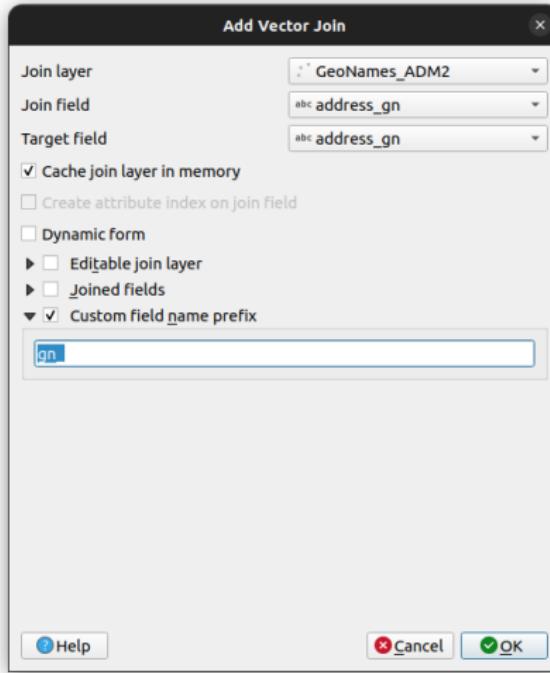
Set Expression to lower(concat(County_full, ', ', State)). Click OK



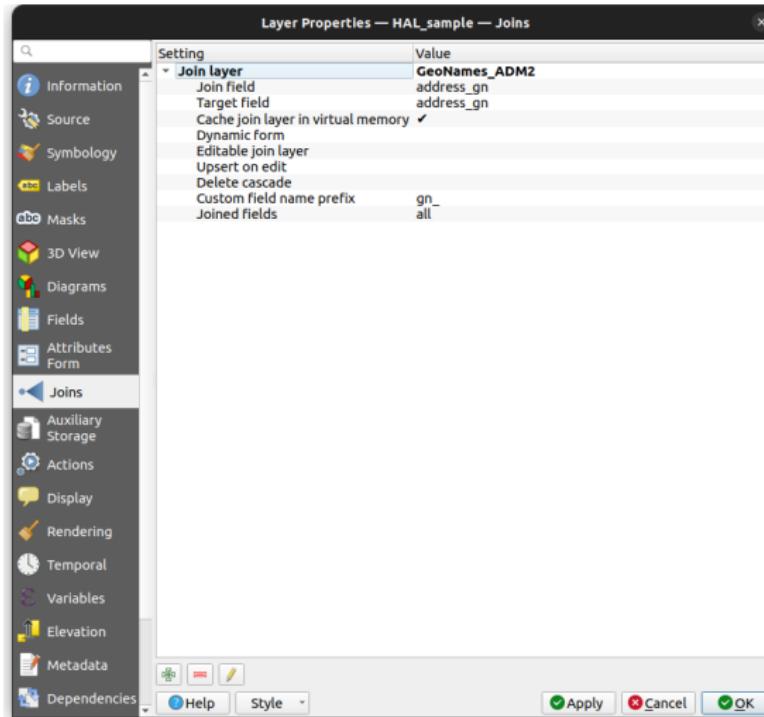
Now we are ready to join these layers. Double-click HAL_sample layer to bring up the Properties window. Open the Joins tab, and click the + button.



Set Join layer = GeoNames_ADM2, with address_gn as the join and target field.
Check the box next to Custom Field Name Prefix and enter gn_ in the box.
Click OK.



The new join should appear in the Joins tab



If you click on the Fields tab, you should see the new fields appended to HAL_sample

The screenshot shows the 'Layer Properties — HAL_sample — Fields' dialog in QGIS. The left sidebar contains tabs for Information, Source, Symbology, Labels, Masks, 3D View, Diagrams, Fields (which is selected), Attributes, Form, Joins, Auxiliary Storage, Actions, Display, Rendering, Temporal, Variables, Elevation, Metadata, and Dependencies. The main area is a table titled 'Fields' with columns: Id, Name, Alias, Type, Type name, Length, and Precision. The table lists 21 fields:

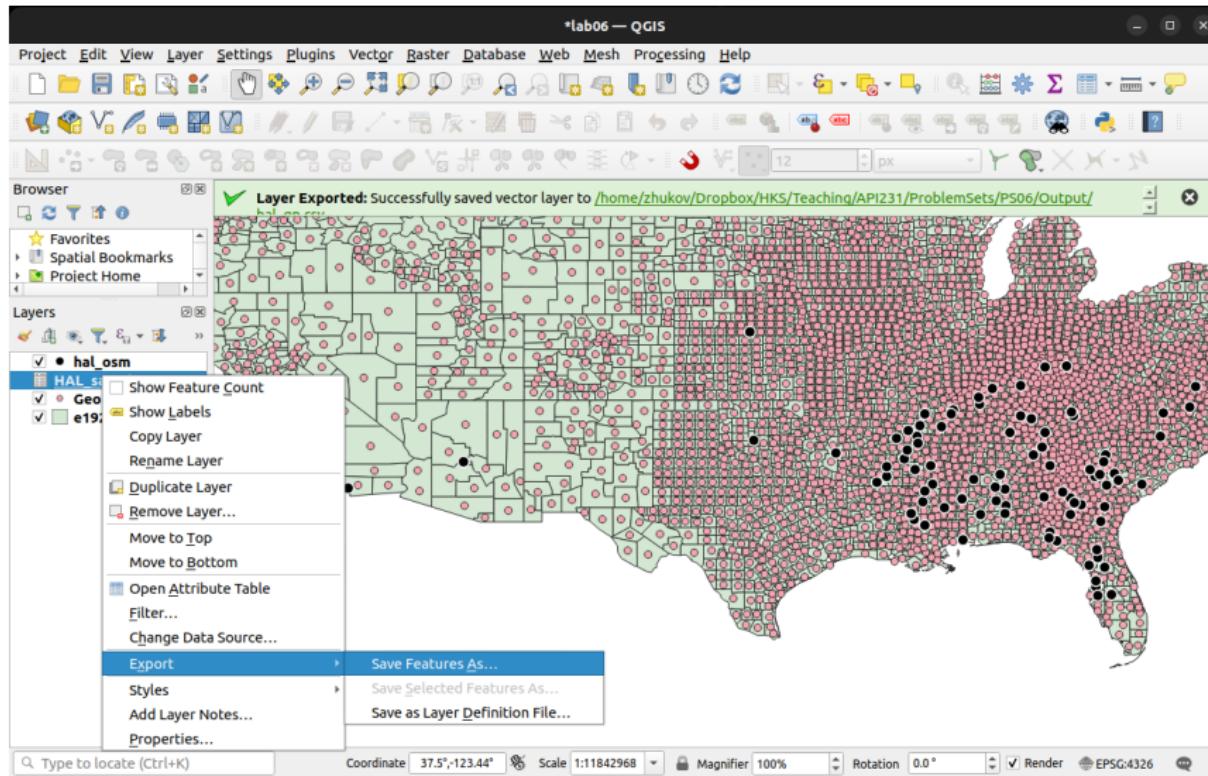
Id	Name	Alias	Type	Type name	Length	Precision
0	State		Text (string)	text	0	0
1	Year		Integer (32 bit)	integer	0	0
2	Mo		Integer (32 bit)	integer	0	0
3	Day		Integer (32 bit)	integer	0	0
4	Victim		Text (string)	text	0	0
5	County		Text (string)	text	0	0
6	Race		Text (string)	text	0	0
7	Sex		Text (string)	text	0	0
8	Mob		Text (string)	text	0	0
9	Offense		Text (string)	text	0	0
10	Note		Text (string)	text	0	0
11	2nd Name		Text (string)	text	0	0
12	3rd Name		Text (string)	text	0	0
13	County_full		Text (string)	text	0	0
14	gn_geonameid		Integer (32 bit)	integer	0	0
15	gn_name		Text (string)	text	0	0
16	gn_ascliname		Text (string)	text	0	0
17	gn_alternatenames		Text (string)	text	0	0
18	gn_latitude		Decimal (double)	double	0	0
19	gn_longitude		Decimal (double)	double	0	0
20	gn_feature_class		Text (string)	text	0	0

At the bottom are buttons for Help, Style, Apply, Cancel, and OK.

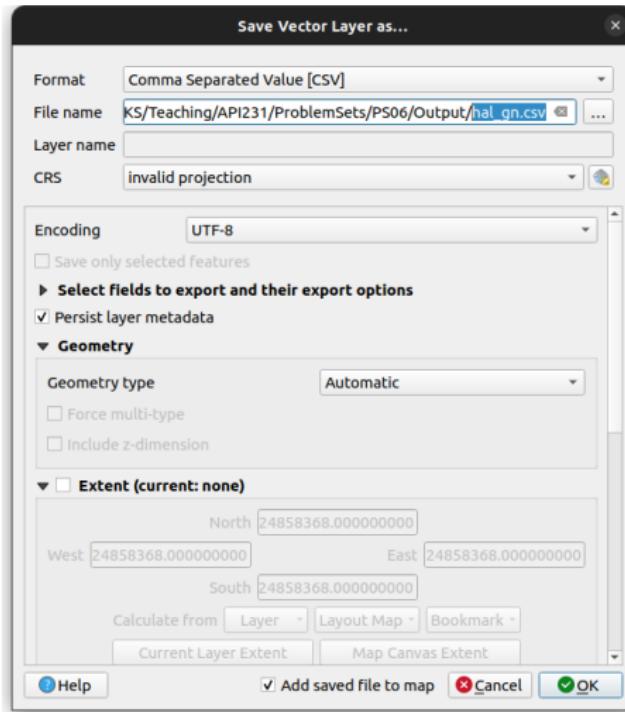
You can also open the Attribute Table for HAL_sample. Do you see the coordinates?

HAL_sample — Features Total: 100, Filtered: 100, Selected: 0															
Note	2nd Name	3rd Name	County_full	address_gn	gn_geonameid	gn_name	gn_asciiname	alternatenam	gn_latitude	gn_longitude	n_feature	clsn	feature	codn	coun
1	NULL	NULL	NULL	Aiken county	aiken coun...	4569073	Aiken County	Aiken County	Aiken,Aike...	33.54437	-81.63474	A		ADM2	US
2	NULL	NULL	NULL	Leon county	leon count...	4161831	Leon County	Leon County	Comitatul ...	30.45804	-84.27788	A		ADM2	US
3	Uncertain	NULL	NULL	Lauderdale...	lauderdale ...	4433028	Lauderdale...	Lauderdale...	Comitatul ...	32.40429	-88.66254	A		ADM2	US
4	NULL	NULL	NULL	Lowndes c...	lowndes co...	4073885	Lowndes C...	Lowndes C...	Comitatul ...	32.15475	-86.65011	A		ADM2	US
5	NULL	NULL	NULL	Marion cou...	marion cou...	4163456	Marion Cou...	Marion Cou...	Comitatul ...	29.2102	-82.05668	A		ADM2	US
6	NULL	George Ritter	NULL	Moore cou...	moore cou...	4480053	Moore Cou...	Moore Cou...	Comitatul ...	35.31072	-79.48131	A		ADM2	US
7	NULL	NULL	NULL	Pasco county	pasco coun...	4167895	Pasco County	Pasco County	Comitatul ...	28.30674	-82.43887	A		ADM2	US
8	NULL	Wm. Herrig	NULL	Clay county	clay county...	4105899	Clay County	Clay County	Clay,Clay C...	36.36839	-90.41738	A		ADM2	US
9	Uncertain	NULL	NULL	St. Tammany	st. tamman...		NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL

To display the geocoded locations, we need to export the joined file and re-import it.
Right-click on HAL_sample and go to Export → Save Features As....



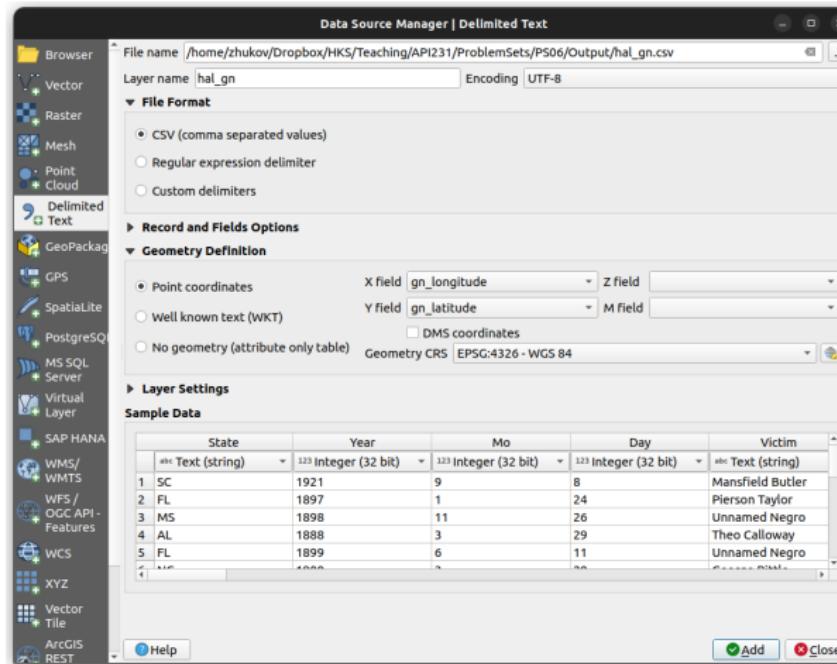
Save the file as `hal_gn.csv` (comma separated values)



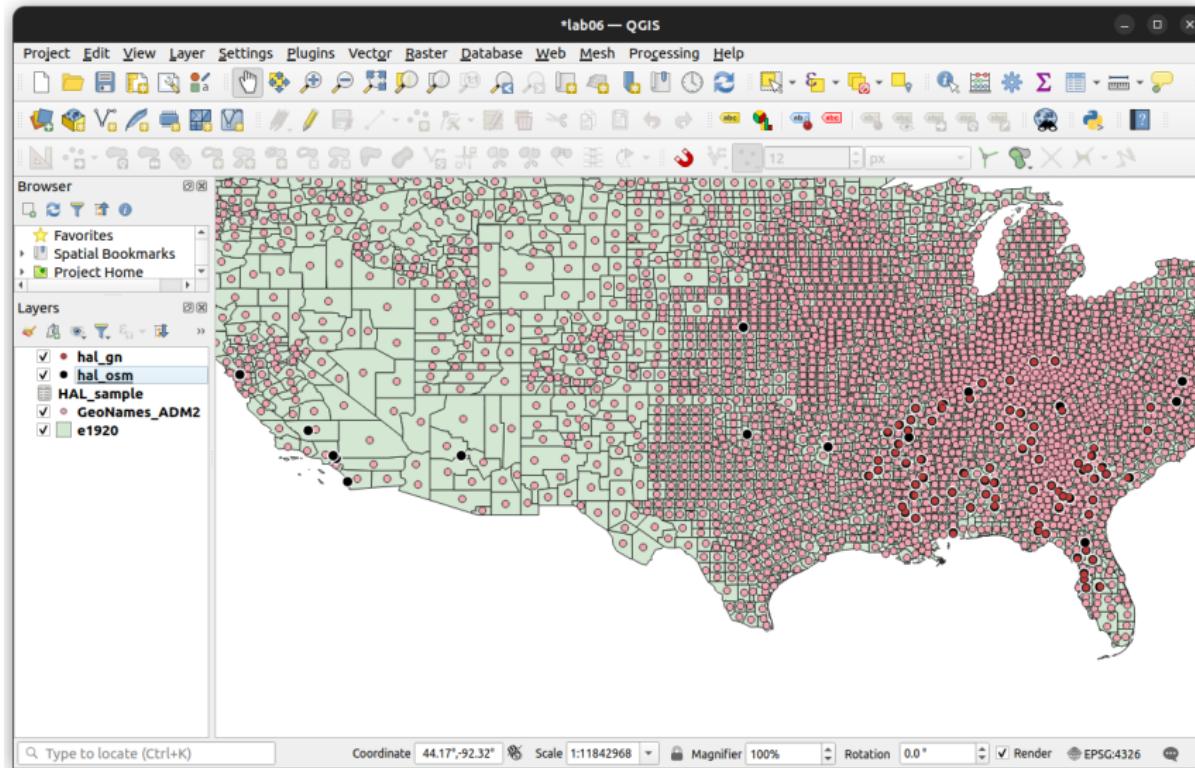
(Re-)import hal_gn.csv, using Add Delimited Text Layer....

Set X field = gn_longitude and Y field = gn_latitude.

Set CRS = EPSG:4326

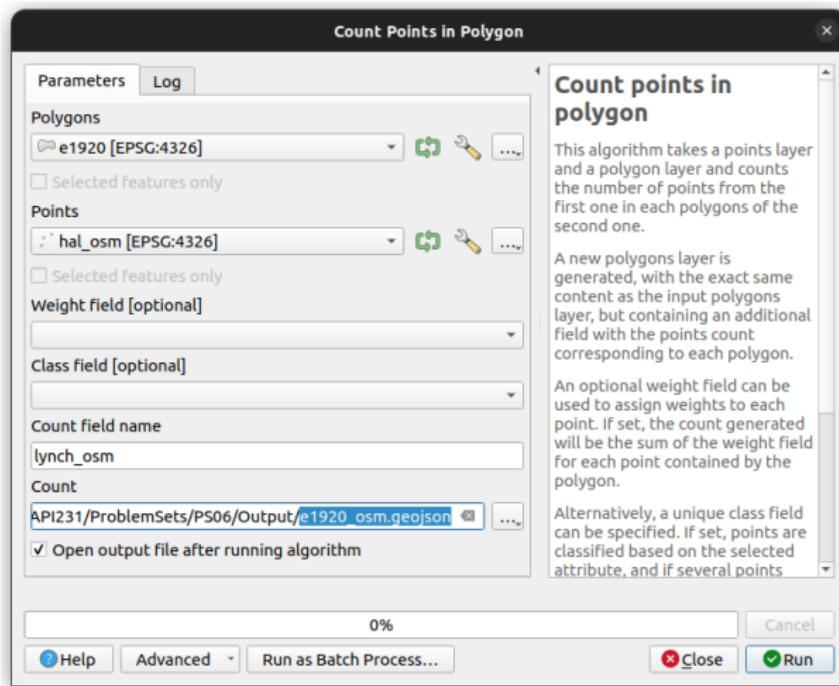


The new geocoded locations should appear on the project window.
Now let's do some point-in-polygon analysis!

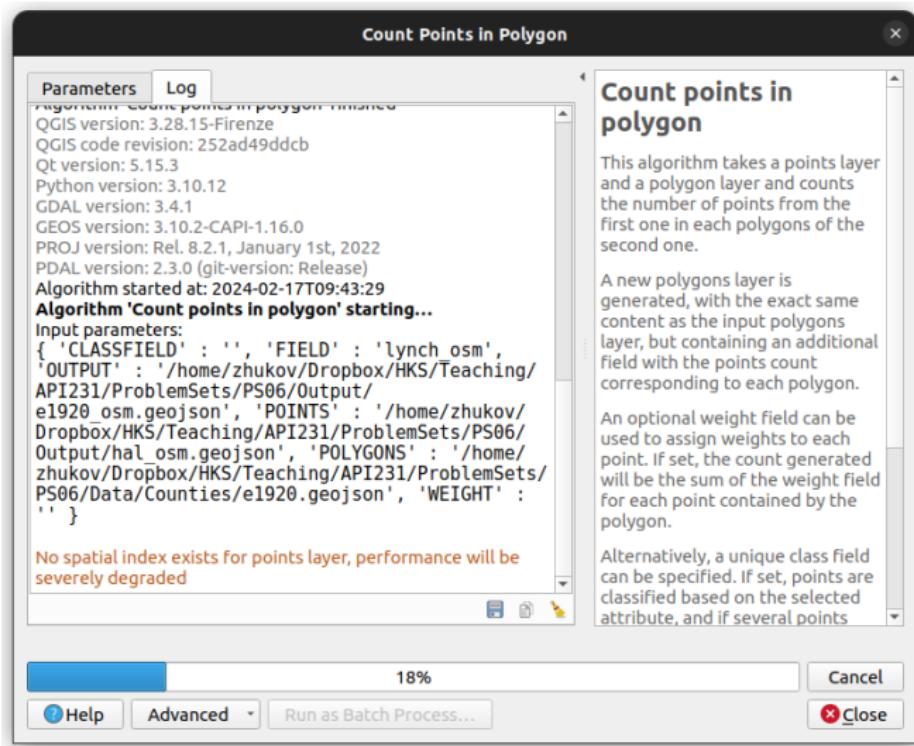


Boxplot

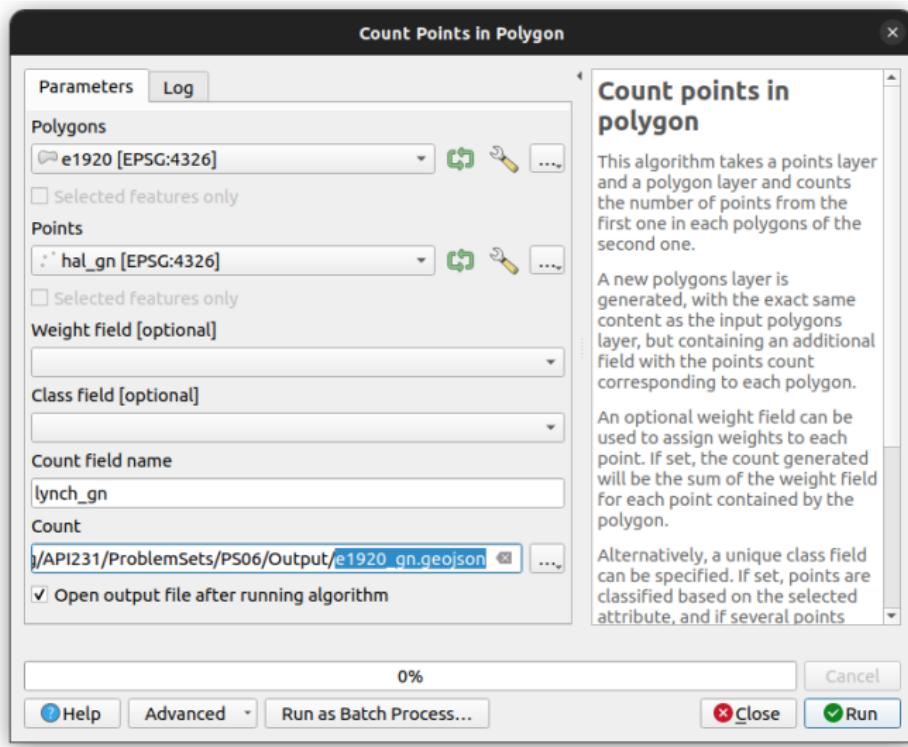
Navigate to Vector menu → Analysis Tools → Count Points in Polygon.
Select Polygons = e1920, Points = hal_osm. Name the count field lynch_osm, and save the output to e1920_osm.geojson. Click Run



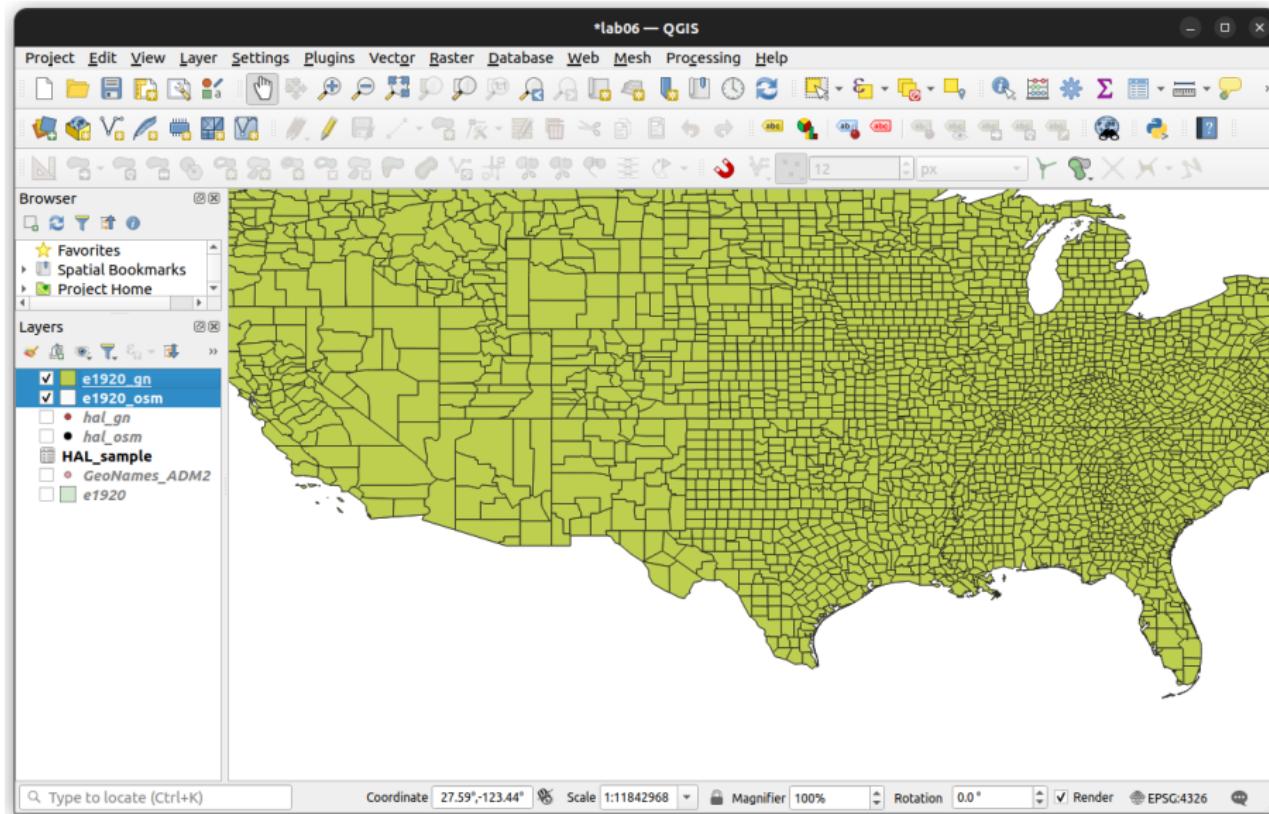
You may see a warning about “No spatial index exists...”. You can ignore it here



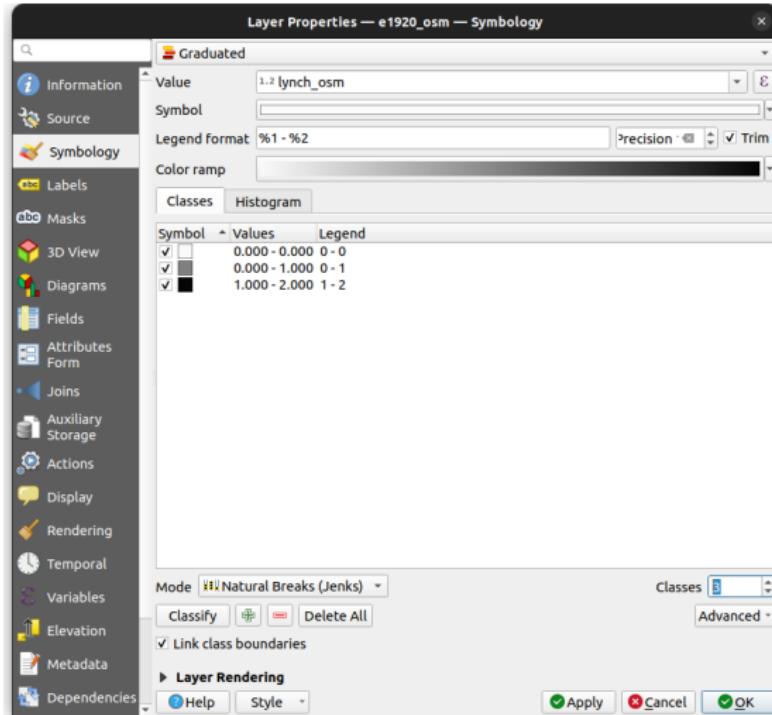
Repeat this process with hal_gn as the points layer. Name the count field lynch_gn, and save the output as e1920_gn.geojson



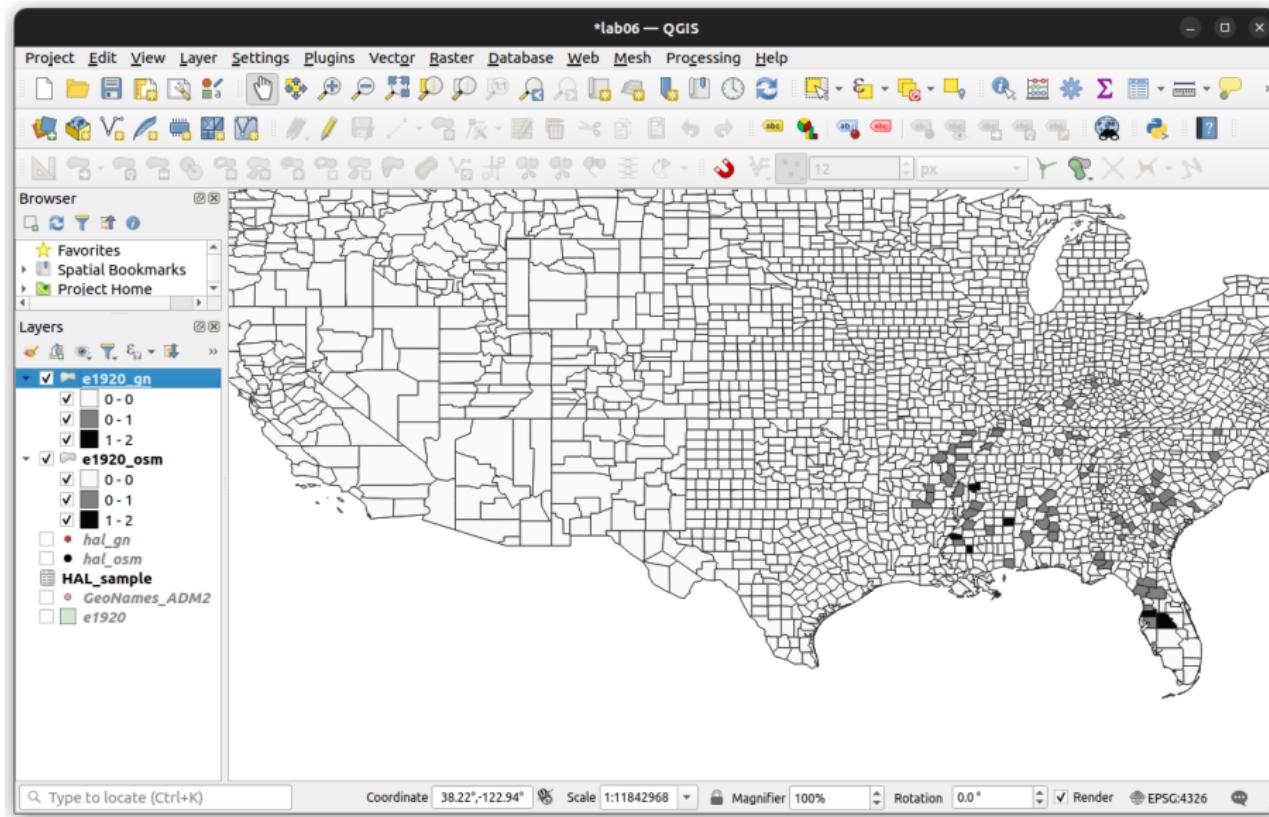
The two new layers e1920_osm and e1920_gn should appear in the project window



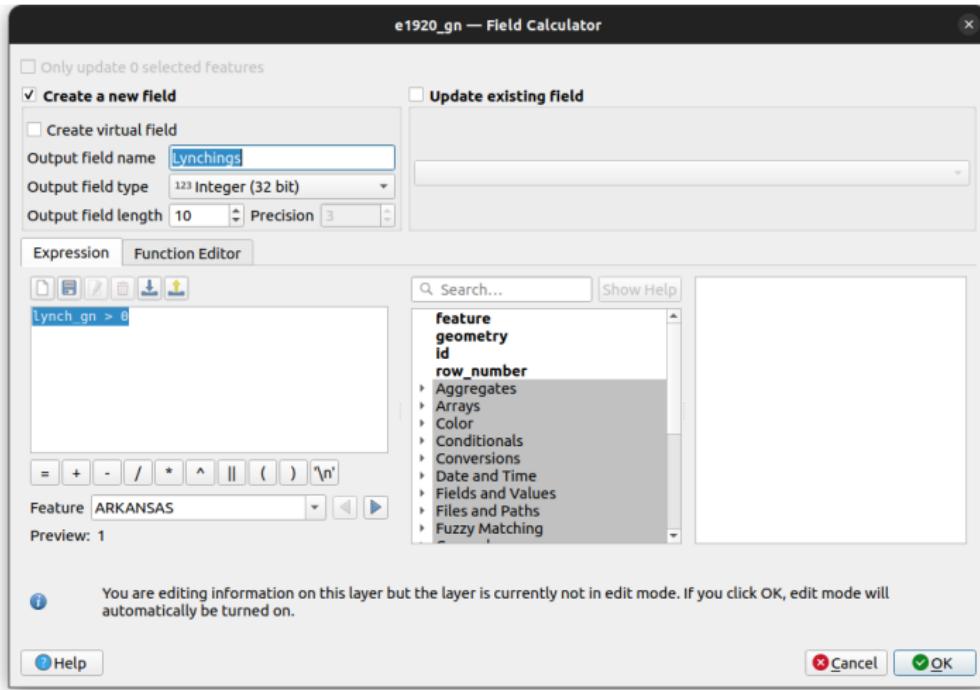
We can try plotting these new count variables through Layer Properties → Symbology. You will notice right away that there are not many unique values. I used Natural Breaks with 3 classes, but you can try other options.



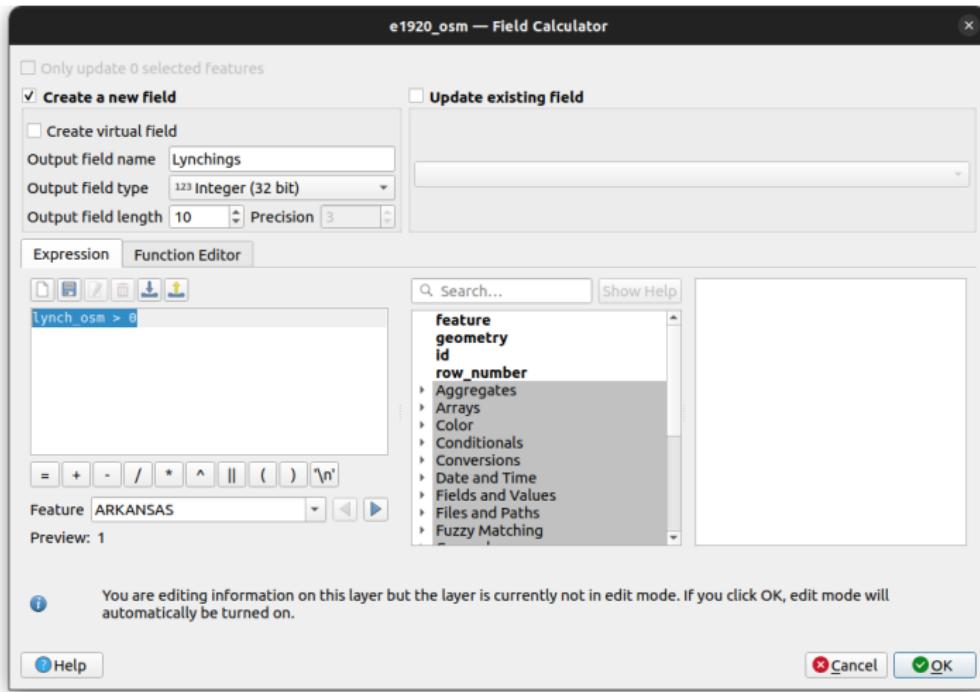
The resulting distribution should look something like this



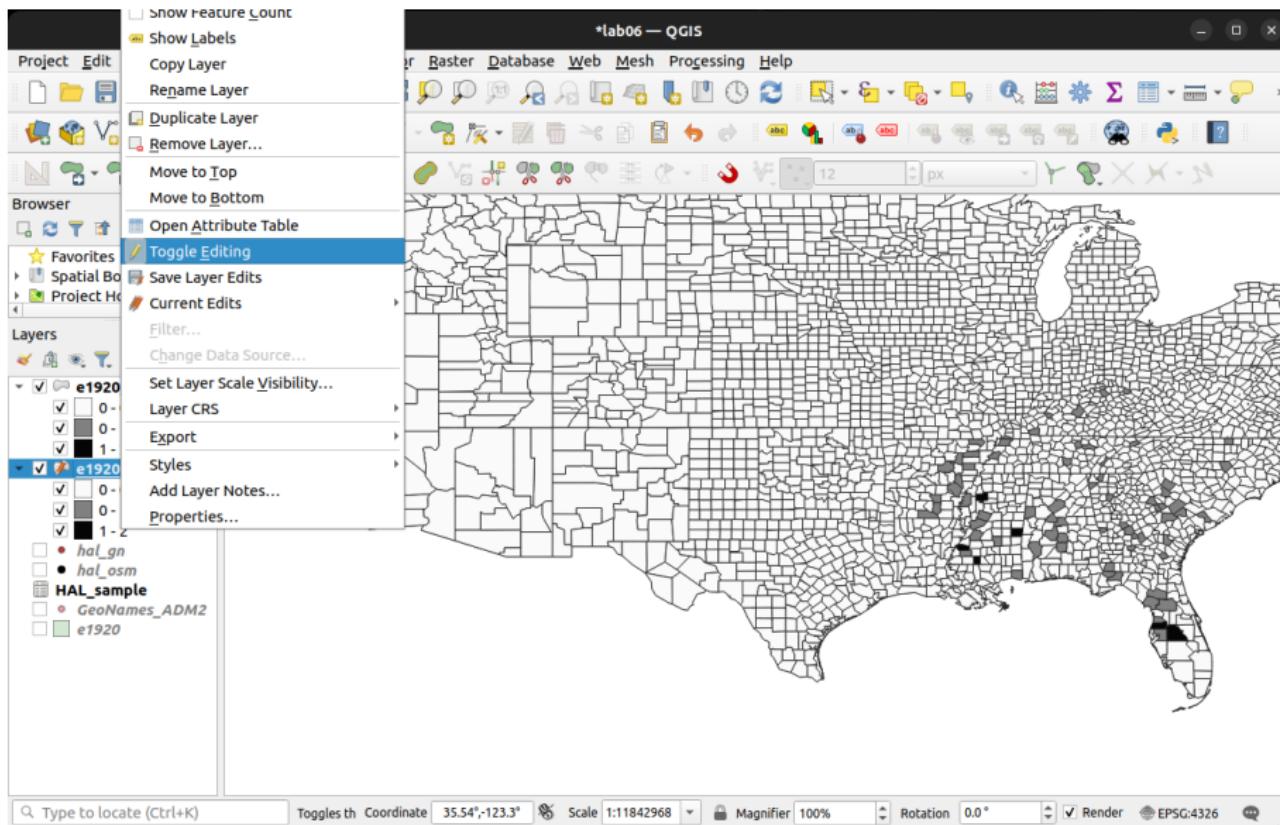
Let's create "dummy" variables indicating whether *at least one* lynching occurred in each county. Open Field Calculator for e1920_gn, create new field, Lynchings of type Integer, with Expression set to lynch_gn > 0. Click OK



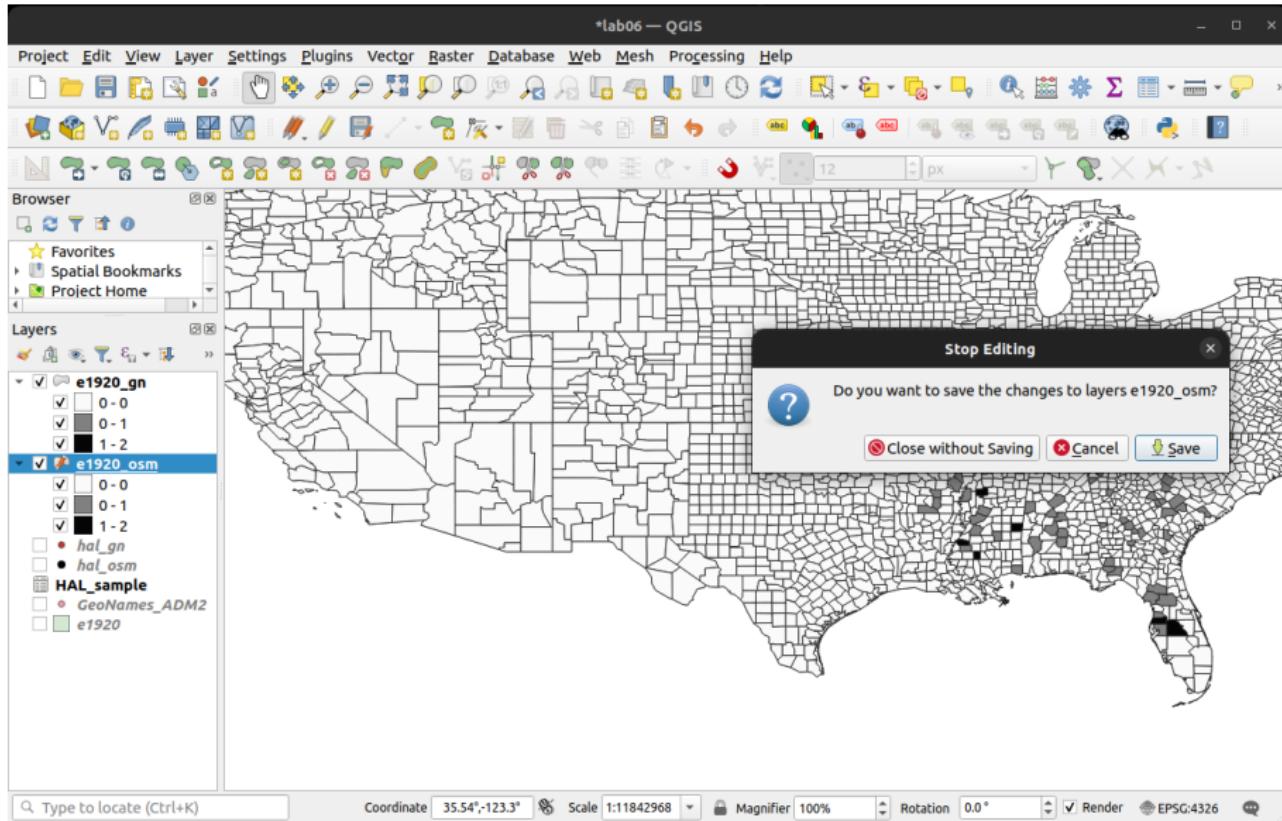
Do the same for e1920_osm, with Expression set to lynch_osm > 0



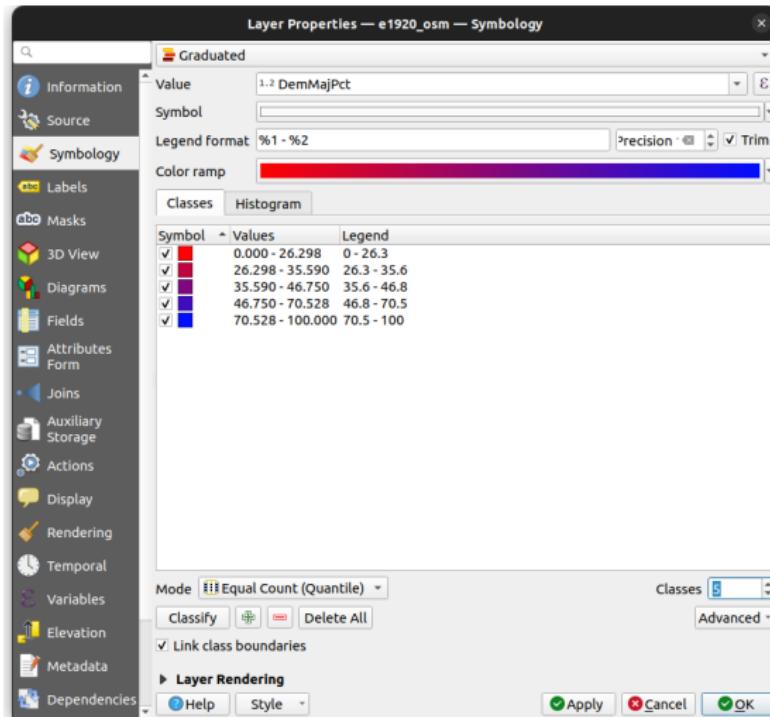
When finished, right-click on e1920_osm and e1920_gn, uncheck Toggle Editing



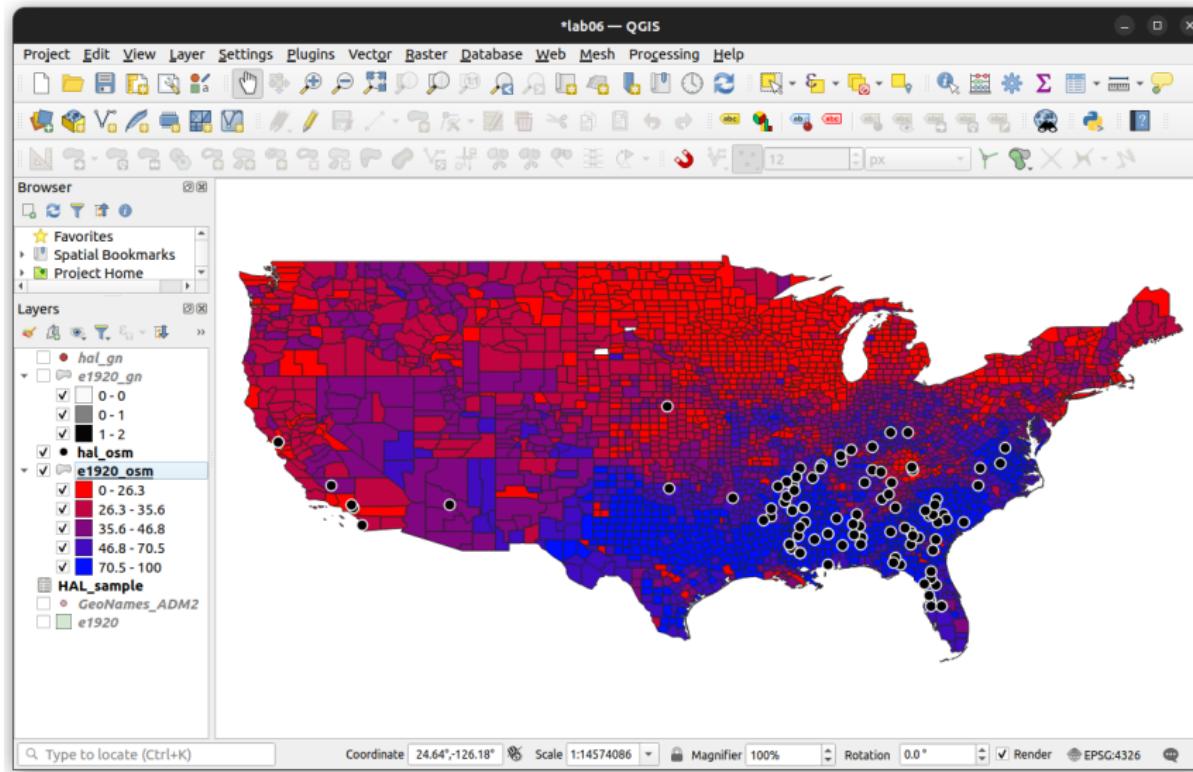
Save your changes when prompted. Do this for both layers



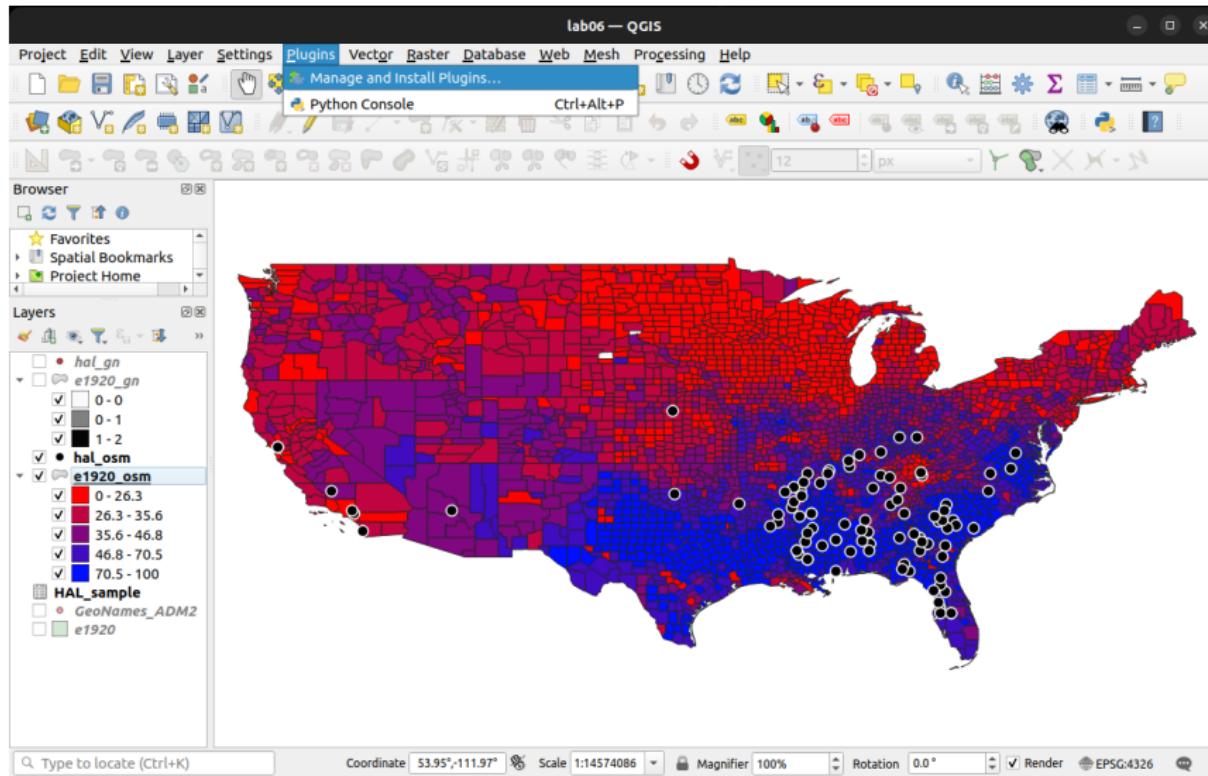
Let's examine the relationship between lynching locations and local electoral preferences. Plot the variable DemMajPct in e1920_osm, with a red-to-blue color ramp (Equal Count with 5 classes)



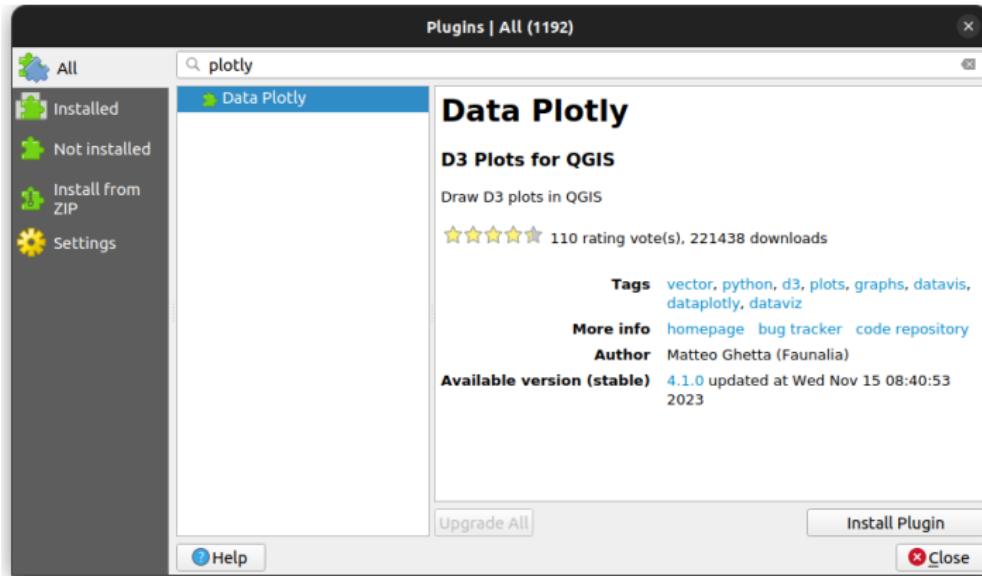
The lynching locations appear to be mostly in southern Democratic Party strongholds, but let's look at this more systematically



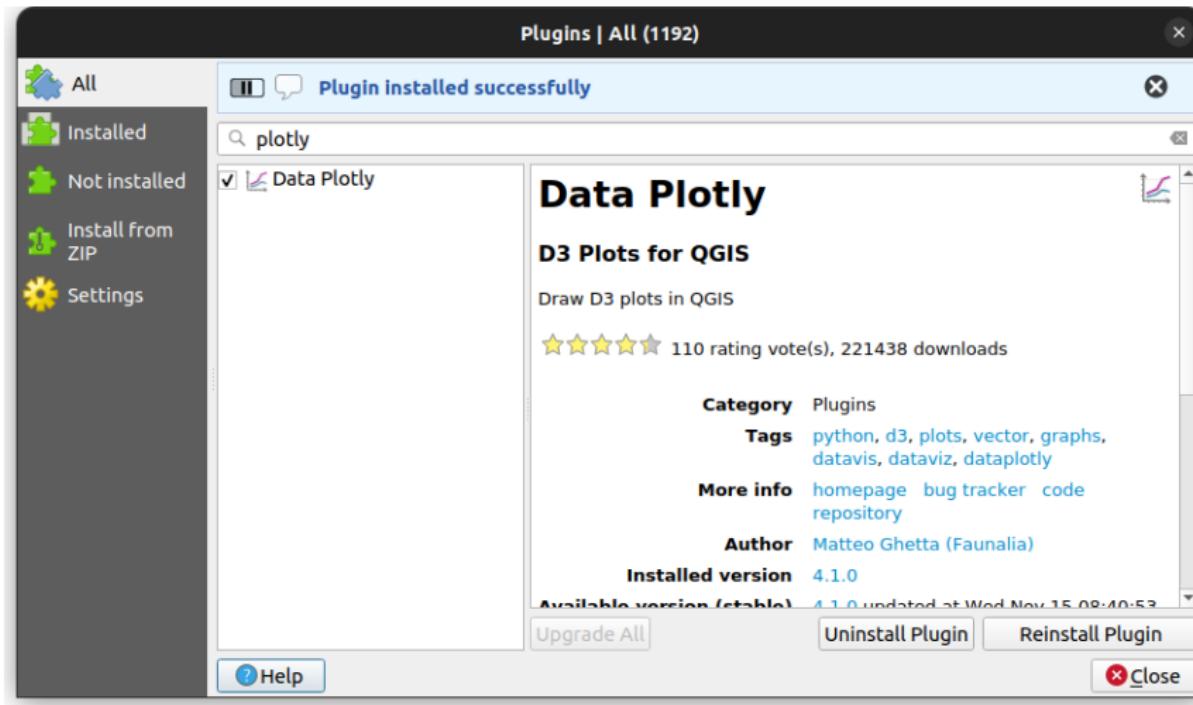
To enable a variety of statistical plotting functions in QGIS, we need to enable the *Data Plotly* plugin. Go to Plugins → Manage and Install Plugins



Search for `plotly` and install the plugin

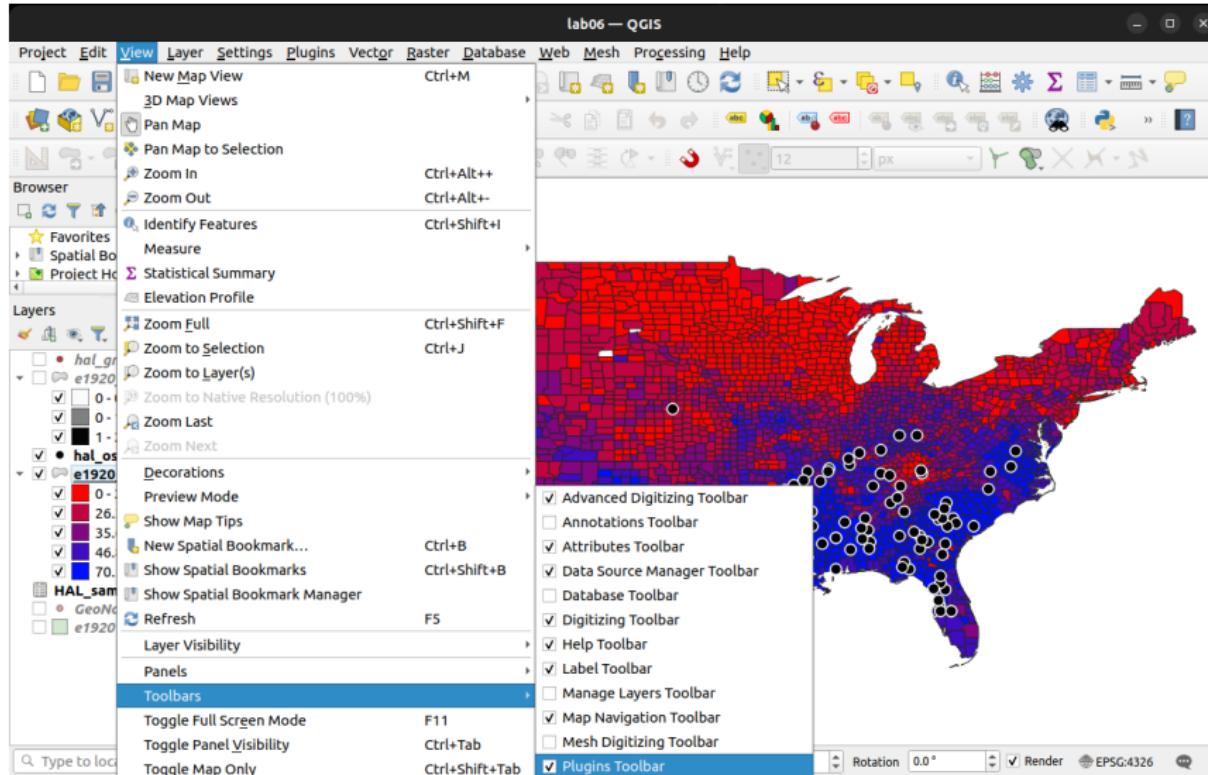


Make sure the the box is checked next to Data Plotly after installation



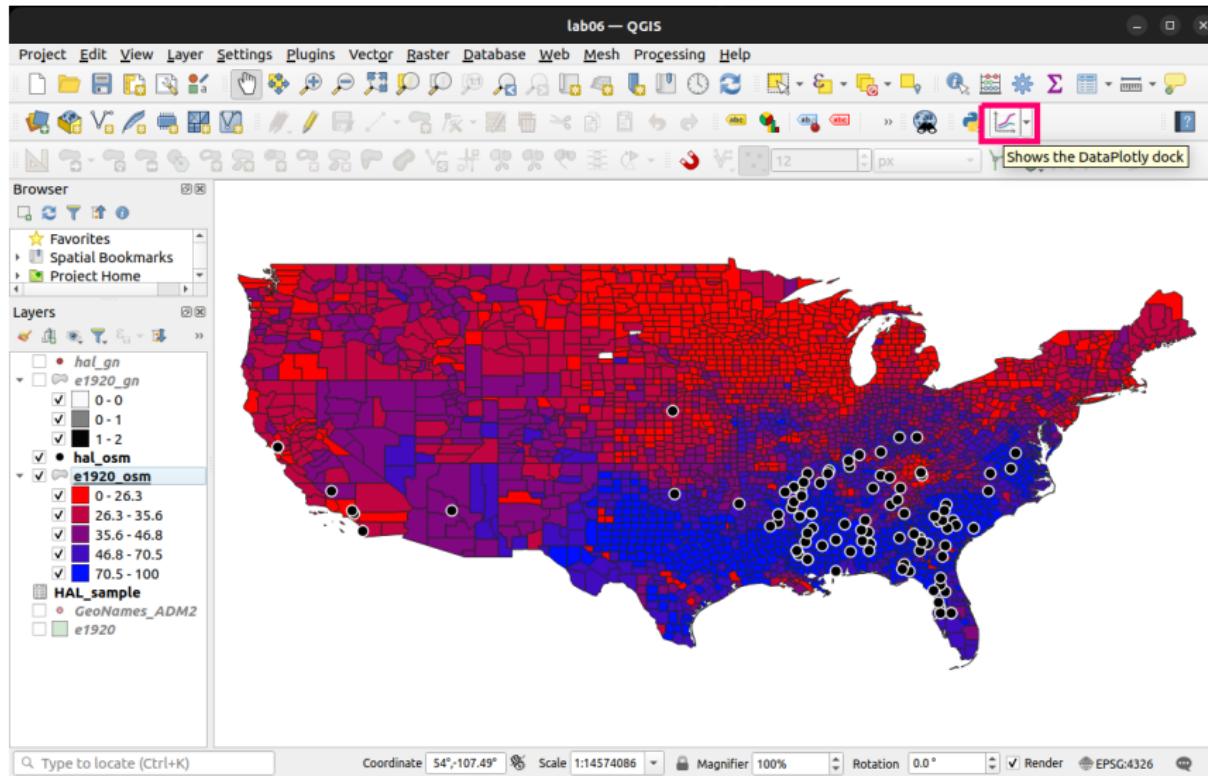
To access Data Plotly, activate the Plugins Toolbar.

Go to View → Toolbars → check box next to Plugins Toolbar

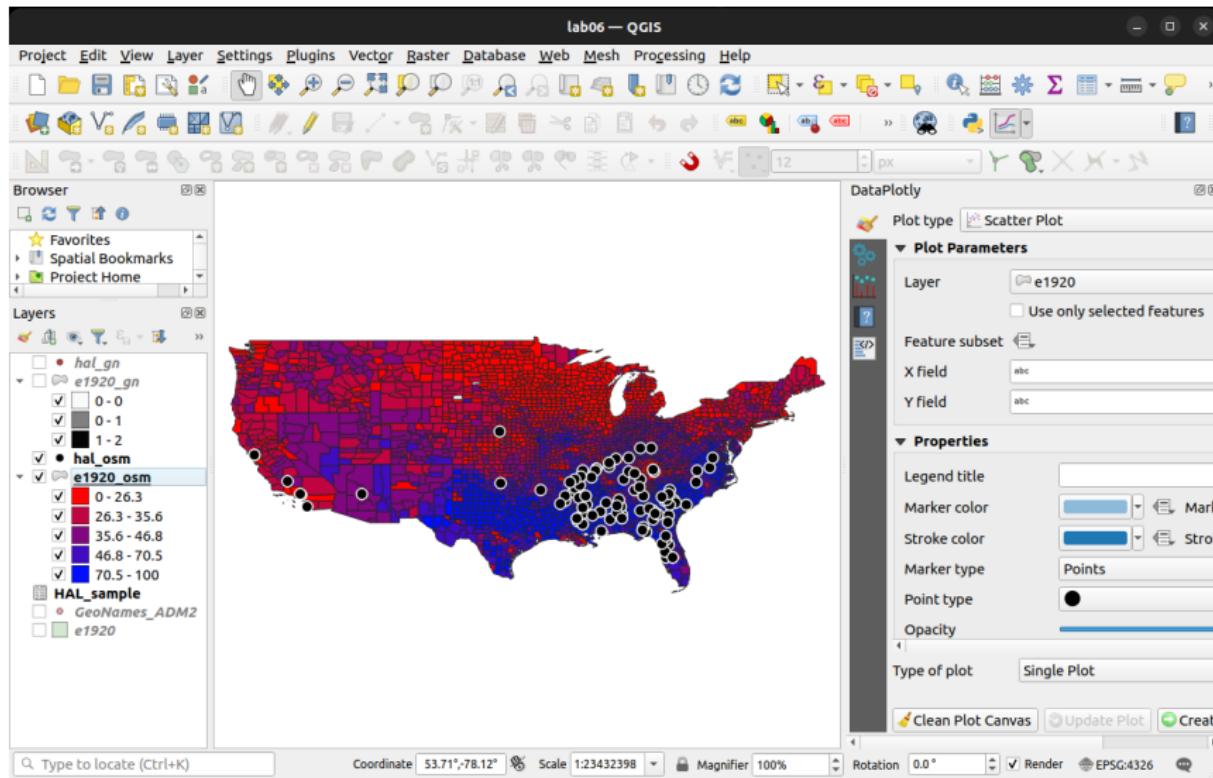


The icon for Data Plotly can be tricky to find.

Look for an icon that looks like a statistical graphic, as shown here ↓. Click on it

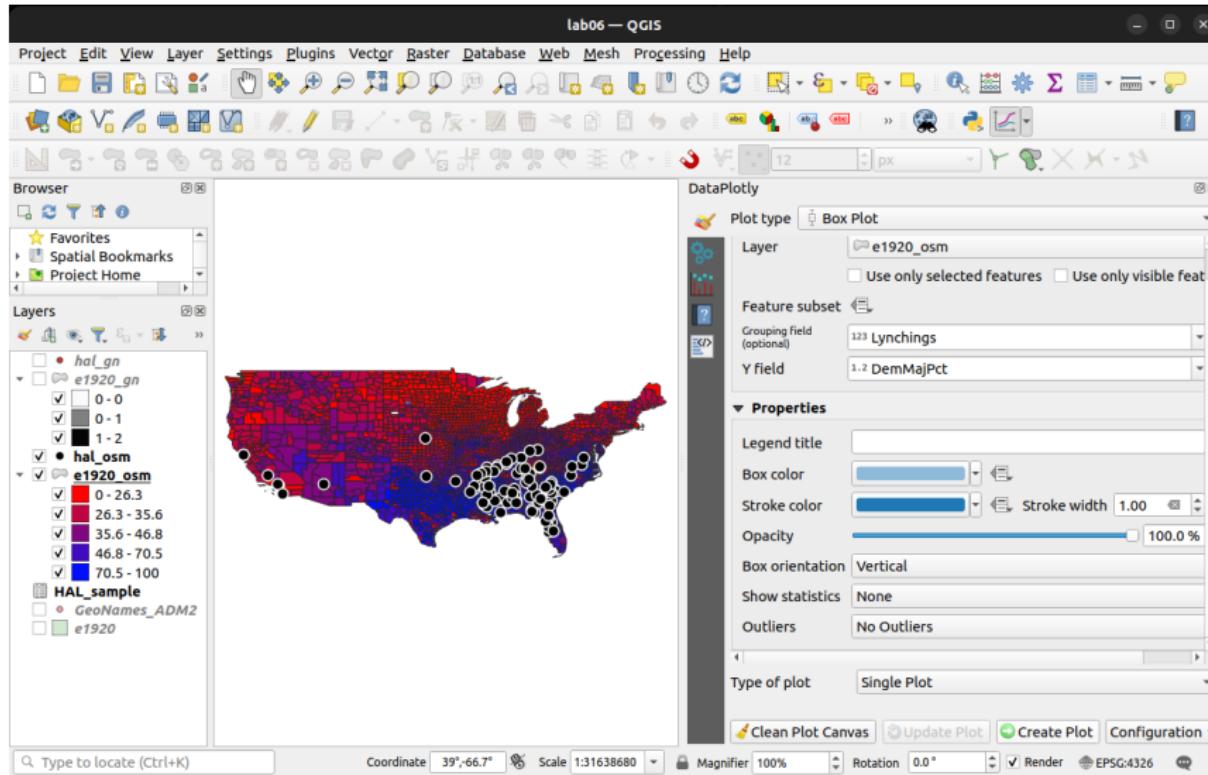


This should open a new DataPlotly panel in the project window.

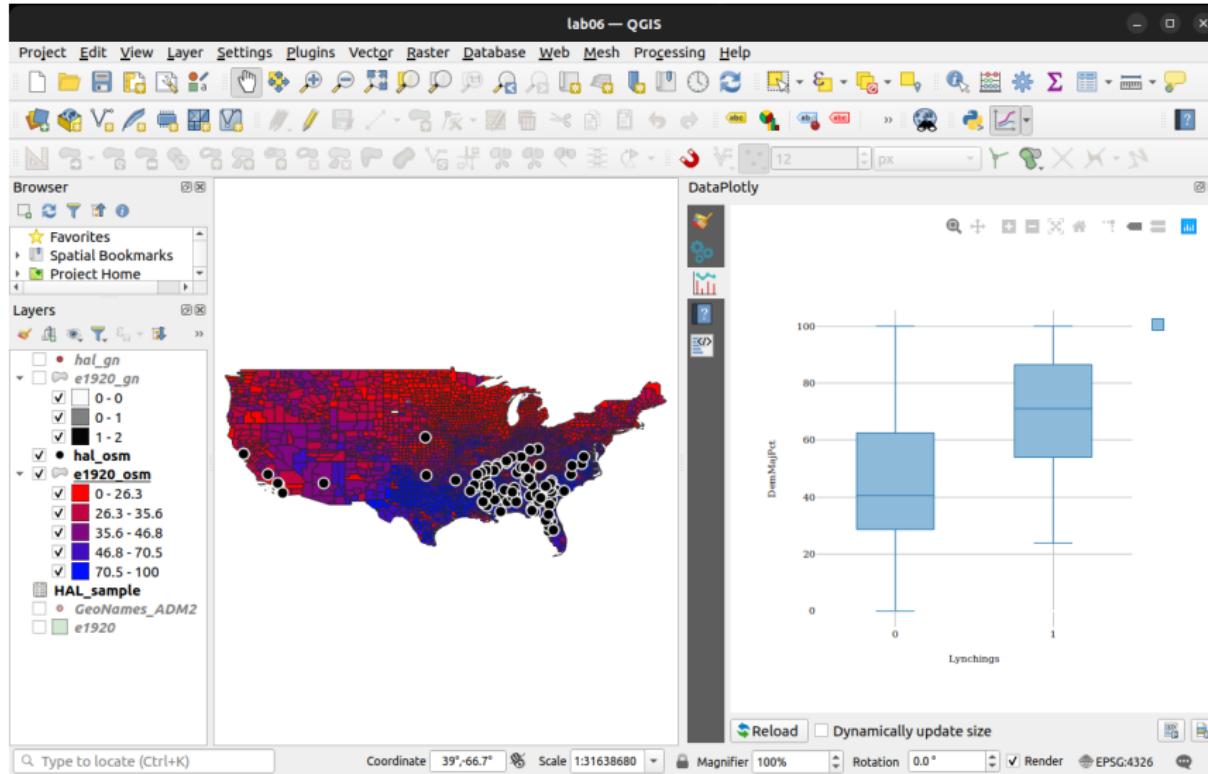


Set Plot type = Box Plot, Layer = e1920_osm.

Set Grouping = Lynchings, Y field = DemMajPct. Click Create Plot

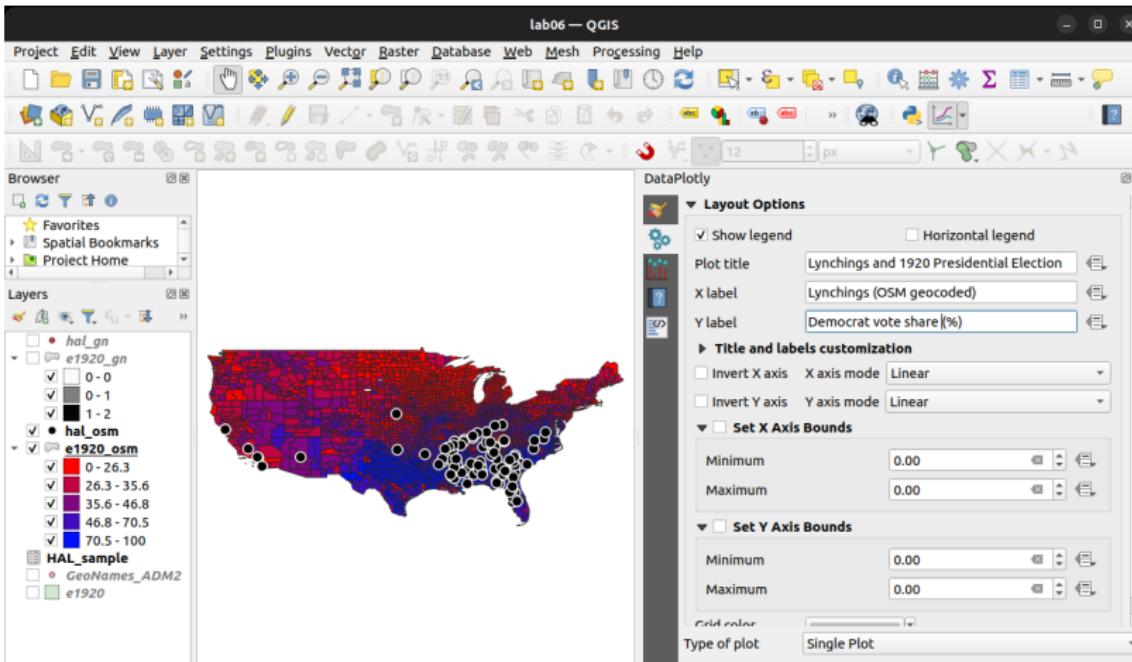


The average Democratic vote share was indeed significantly higher for counties with lynchings than counties without lynchings

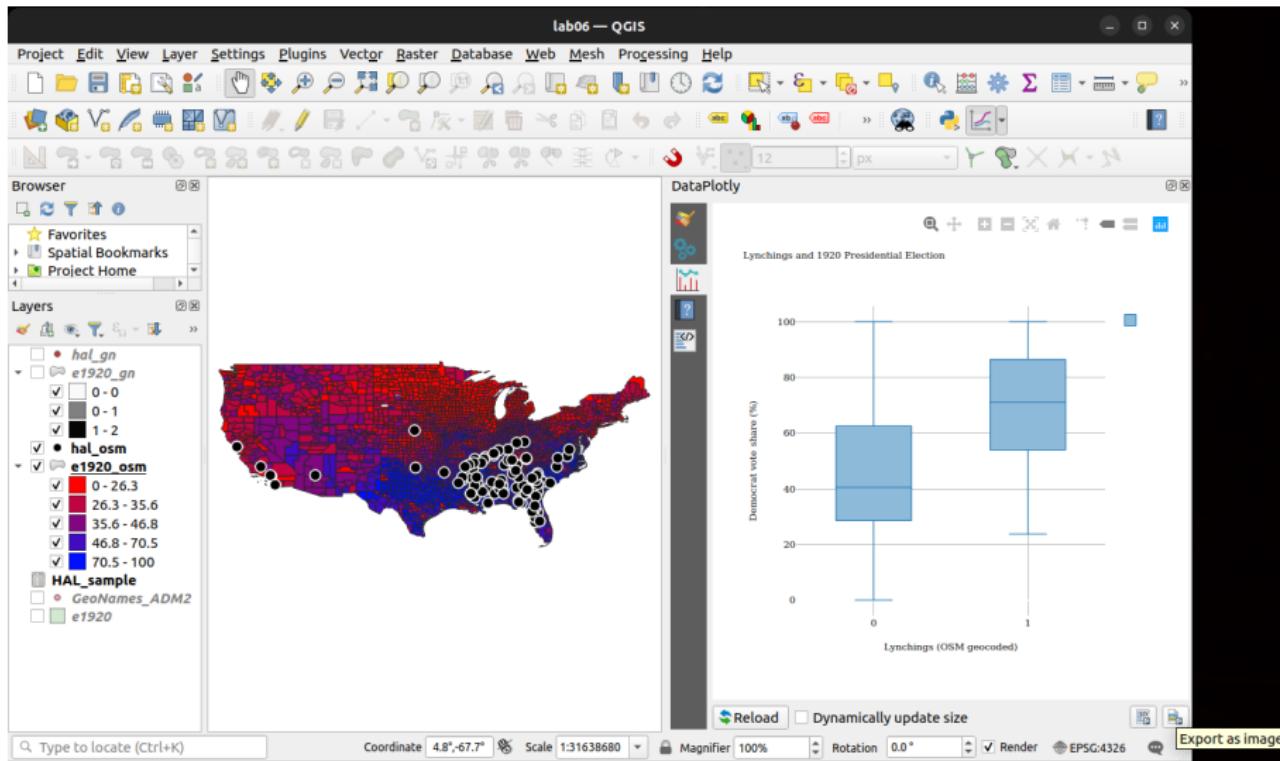


Let's clean up the axis labels. Click on the settings button (gears icon).

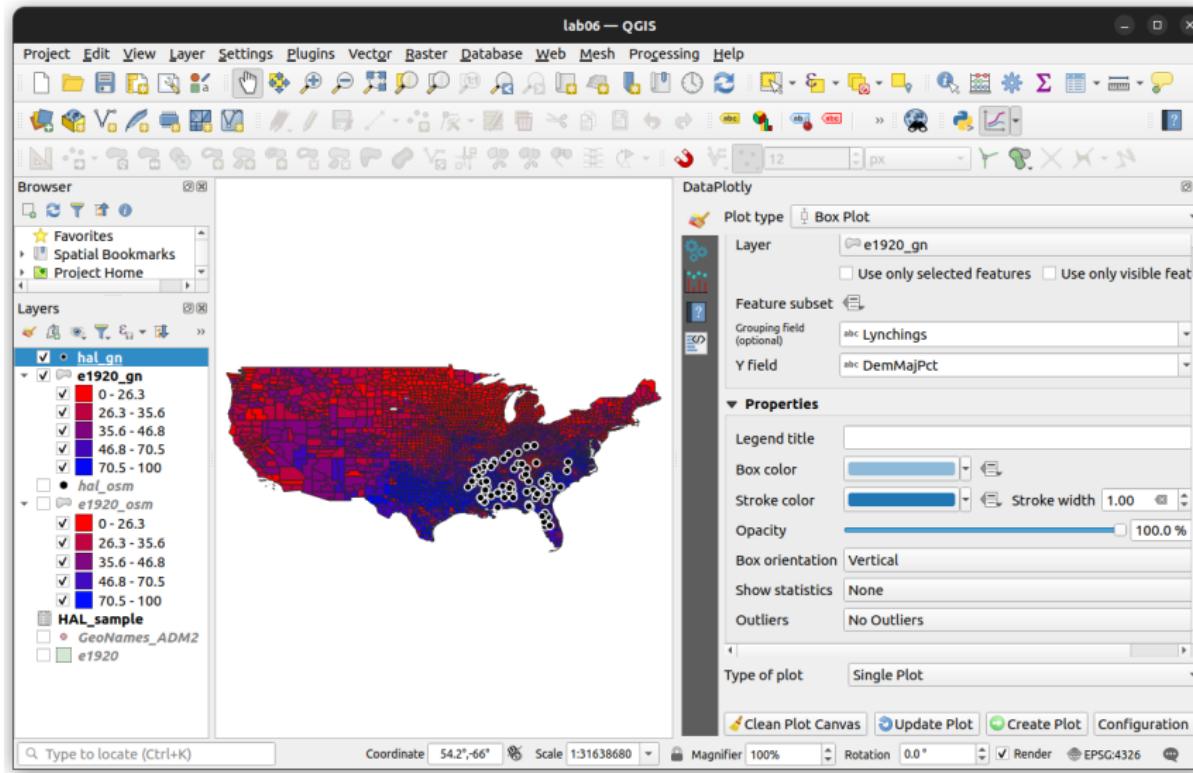
- Set Plot Title = Lynchings and 1920 Presidential election
- Set X Label = Lynchings (OSM geocoded) and Y Label = Democratic vote share in 1920 (\%) Click Update Plot



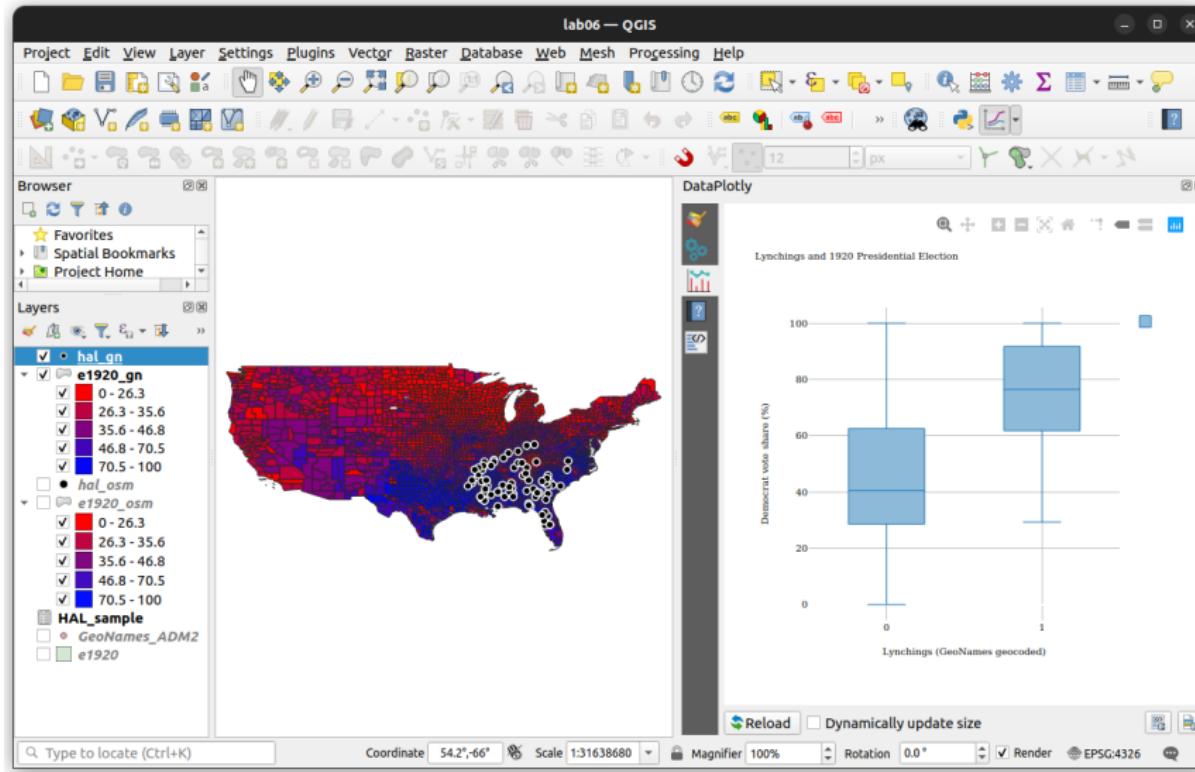
To export the box plot as a .png file, click on the icon in the lower-right corner.
Name it `boxplot_osm.png`



Repeat this process for GeoNames-geocoded e1920_gn, with similar labels



Export the resulting image



Problem Set 6

Your assignment (if using QGIS): create a map to go along with the box plot

- make and export a box plot, as just demonstrated
 - name the file `boxplot_osm.png` or `boxplot_gn.png`
(depending on which geocoder you're using)
- create and export a new map layout, showing:
 - points: lynching locations (either OSM- or GeoNames-geocoded)
 - polygons: counties colored by Democratic vote share (`DemMajPct`)
 - legend (only the layers you're using should be on legend)
 - scale bar (optional)
- name the file `map_osm.png` or `map_gn.png`
(depending on which geocoder you're using)
- upload map and accompanying box plot (**2 files total!**) to Canvas
- AGAIN: you can do this with *either* OpenStreetMap or GeoNames, you are not required to do both

Either this:

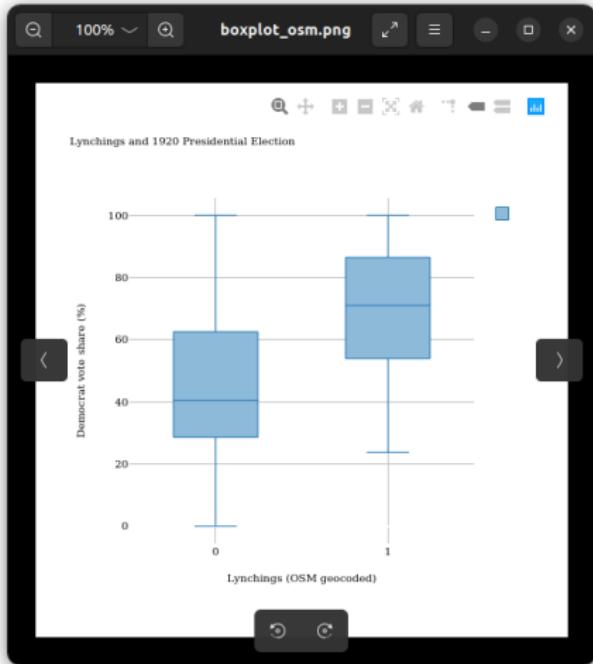


Figure 10: OSM boxplot

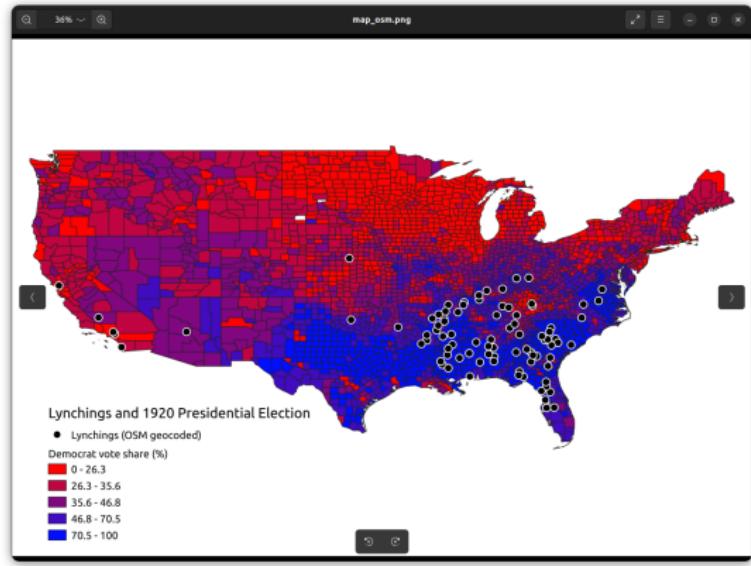


Figure 11: OSM map

... or this:

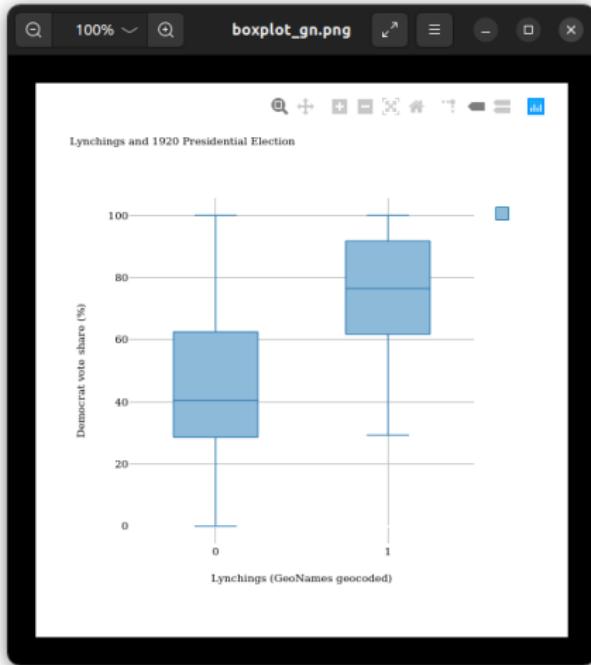


Figure 12: GeoNames boxplot

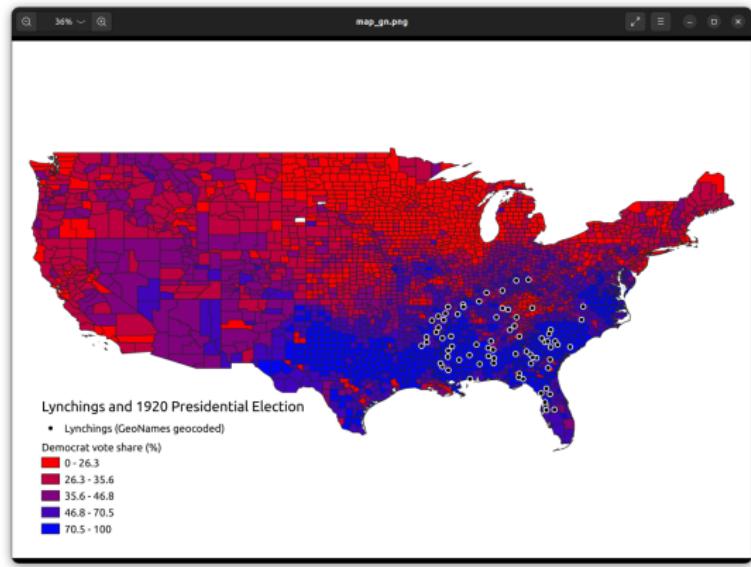


Figure 13: GeoNames map

R

Loading R packages

To implement these steps in R, we will be using the `sf` package, and two others (`RCurl`, `jsonlite`) that help R compose HTTP requests and process the results returned by online servers:

```
library(sf)
library(RCurl)
library(jsonlite)
```

NOTE: The code to produce the maps and boxplots in R is in `ps05_demo.R` on RStudio Cloud, and in `PS05.zip` (posted on Canvas).

Geocoding

As with QGIS, we will geocode in R using two methods:

1. Online, using a web service (OSM/Nominatum API)
2. Offline, using gazetteer data (GeoNames)

Method 1: Geocode the addresses using OSM/Nominatum

Step 1: define a function `url_geo()` that sends queries to OSM/Nominatum, and returns geographic information from server:

```
url_geo = function(query, return.call = "json", sensor = "false") {  
  root = "https://nominatim.openstreetmap.org/search?q="  
  sfxx = "&format=json&polygon=1&addressdetails=1"  
  u = paste(root, query, sfxx, sep = "")  
  return(URLencode(u))  
}
```

Step 2: define a wrapper function geoCode_OSM(), that sends the query through url_geo() and parses the result:

```
geoCode_OSM = function(query,match.num=1){  
  address=NA; longitude=NA; latitude=NA  
  u = url_geo(query)  
  doc = RCurl::getURL(u,httpheader = c('User-Agent' = "contact info"))  
  if(nchar(doc)>2){  
    dat = jsonlite::fromJSON(doc)  
    if(nrow(dat)>0){  
      address = dat$display_name[match.num]  
      longitude = as.numeric(as.character(dat$lon[match.num]))  
      latitude = as.numeric(as.character(dat$lat[match.num]))  
    }  
    return(data.frame(  
      address=address,longitude=longitude,latitude=latitude  
    ))  
  }  
}
```

Let's test the geocoding function!

```
geoCode_OSM("79 John F. Kennedy Street, Cambridge, MA")
```

```
##  
## 1: 79, John F. Kennedy Street, Old Cambridge, Cambridgeport, Cambridge, Middlesex County, Mass  
##   longitude latitude  
## 1: -71.12136 42.37129
```

```
geoCode_OSM("Harvard Kennedy School")
```

```
##  
## 1: Harvard Kennedy School Library, John F. Kennedy Street, Old Cambridge, Cambridgeport, Camb  
##   longitude latitude  
## 1: -71.12165 42.37117
```

```
geoCode_OSM("Harvard")
```

```
##  
## 1: Harvard University, Western Avenue, Barry's Corner, Allston, Boston, Suffolk County, Massa  
##   longitude latitude  
## 1: -71.12221 42.36574
```

```
geoCode_OSM("Harvard", match.num=2)
```

```
##                                     address longitude latitude  
## 1: Harvard, McHenry County, Illinois, United States -88.61371 42.42224
```

Load, pre-process Project HAL data

Load the tabular dataset using `read.csv()`, and preview the first few rows:

```
hal = read.csv("Data/HAL/HAL_sample.csv")
head(hal)

##   State Year Mo Day      Victim    County Race Sex Mob        Offense
## 1   SC 1921  9  8 Mansfield Butler     Aiken  Blk Male Muderous assault
## 2   FL 1897  1 24 Pierson Taylor      Leon  Blk Male Attempted rape
## 3   MS 1898 11 26 Unnamed Negro Lauderdale Blk Male          Assault
## 4   AL 1888  3 29 Theo Calloway    Lowndes  Blk Male          Murder
## 5   FL 1899  6 11 Unnamed Negro     Marion  Blk Male Blk Aided in lynching
## 6   NC 1900  3 20 George Rittle     Moore  Blk Male           Informer
## Note X2nd.Name X3rd.Name    County_full
## 1                               Aiken county
## 2                               Leon county
## 3 Uncertain                   Lauderdale county
## 4                               Lowndes county
## 5                               Marion county
## 6       George Ritter         Moore county
```

Create new field for “county, state” and placeholders for coordinates

```
hal$address = paste0(hal$County_full, ", ", hal$State)
hal$longitude = NA
hal$latitude = NA
```

To geocode as a batch processing routine, we will write a `for()` loop, which runs the `geoCode_OSM()` function for each address in `hal` and stores the result:

```
for(i in 1:nrow(hal)){
  # Skip past errors
  tryCatch({
    address_geo = geoCode_OSM(hal$address[i])
    # Add coordinates to dataset
    hal$longitude[i] = address_geo$longitude
    hal$latitude[i] = address_geo$latitude
    # Report progress
    print(paste0(i,"/",nrow(hal),"; ",address_geo$address))
  },error=function(e){
    print(paste("Unable to geocode",hal$address[i]))
  })
}
```

Inspect the results of OSM geocoding:

`head(hal)`

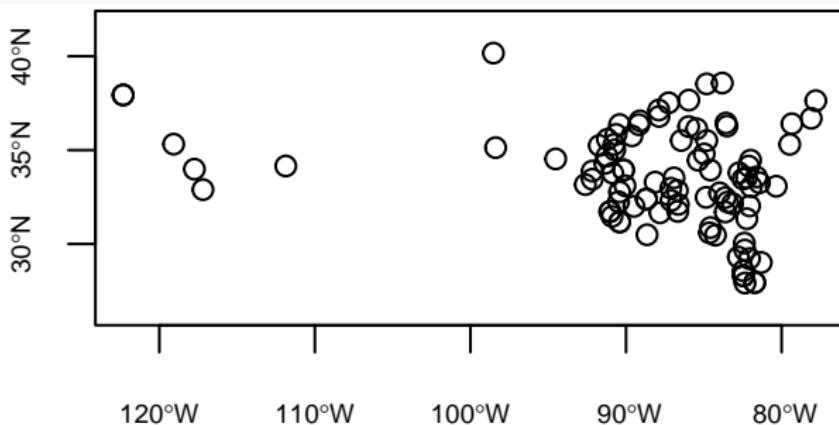
```
##   State Year Mo Day      Victim    County Race Sex Mob      Offense
## 1   SC 1921  9   8 Mansfield Butler     Aiken Blk Male Muderous assault
## 2   FL 1897  1  24 Pierson Taylor      Leon Blk Male Attempted rape
## 3   MS 1898 11  26 Unnamed Negro Lauderdale Blk Male          Assault
## 4   AL 1888  3  29 Theo Calloway     Lowndes Blk Male          Murder
## 5   FL 1899  6  11 Unnamed Negro      Marion Blk Male Blk Aided in lynching
## 6   NC 1900  3  20 George Rittle     Moore Blk Male           Informer
##       Note      X2nd.Name X3rd.Name    County_full            address
## 1                               Aiken county     Aiken county, SC
## 2                               Leon county      Leon county, FL
## 3 Uncertain                  Lauderdale county Lauderdale county, MS
## 4                               Lowndes county    Lowndes county, AL
## 5                               Marion county    Marion county, FL
## 6       George Ritter        Moore county     Moore county, NC
##   longitude latitude
## 1 -81.61821 33.57232
## 2 -84.25491 30.46831
## 3 -88.68964 32.39052
## 4 -86.64025 32.10881
## 5 -82.06269 29.21825
## 6 -79.47612 35.30546
```

Drop observations with missing coordinates:

```
hal_osm = hal[which(!is.na(hal$longitude)),]
```

Convert results to sf object and plot on a map:

```
hal_osm = st_as_sf(hal_osm,  
                    coords=c("longitude","latitude"),crs=4326)  
plot(hal_osm["geometry"],axes=TRUE)
```



Method 2: Geocode the addresses using GeoNames gazetteer

Load gazetteer data:

```
gn = read.csv("Data/GeoNames/GeoNames_ADM2.csv")
```

Create common variables for matching:

```
gn$address_gn = tolower(paste0(gn$asciiname, ", ", gn$admin1_code))
hal$address_gn = tolower(paste0(hal$County_full, ", ", hal$State))
```

Rename OSM coordinates to avoid confusion:

```
hal$longitude_osm <- hal$longitude; hal$latitude_osm <- hal$latitude
hal$longitude <- hal$latitude <- NULL
```

Geocode addresses (i.e. join the datasets):

```
hal_gn = merge(x = hal, y = gn, by = "address_gn")
```

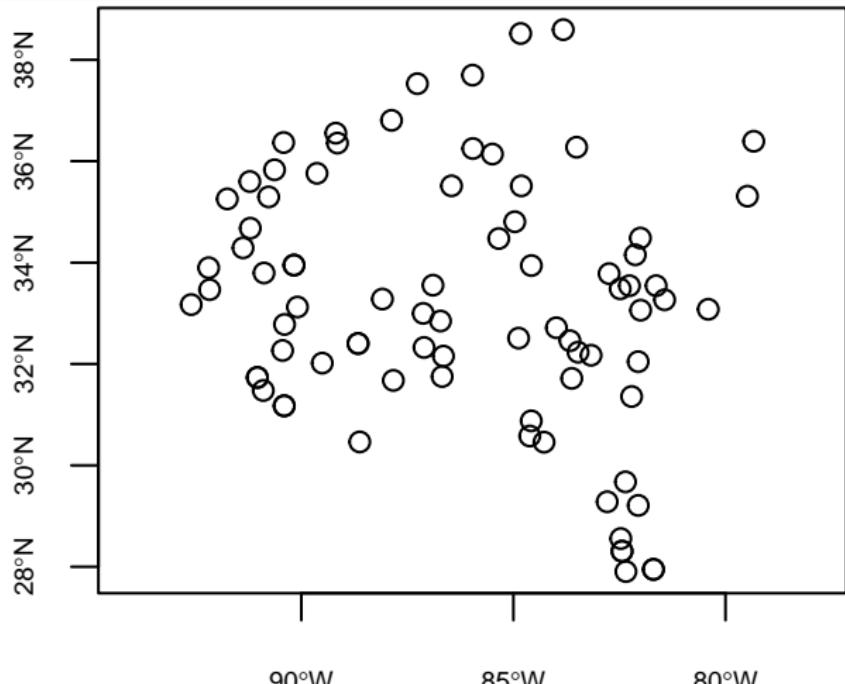
Inspect the results:

```
head(hal_gn)
```

```
##           address_gn State Year Mo Day      Victim County Race Sex Mob
## 1    aiken county, sc   SC 1921  9  8 Mansfield Butler     Aiken Blk Male
## 2  alachua county, fl   FL 1892  9  6      Unnamed Negro  Alachua Blk Male
## 3 arkansas county, ar   AR 1891 12 21       J.A. Smith Arkansas Wht Male
## 4 barnwell county, sc   SC 1890  1  7        Wm. Black Barnwell Blk Male
## 5 bedford county, tn   TN 1912  2 19       Watt Greer Bedford Blk Male
## 6    babb county, al   AL 1904  6 23        Joe Scott Bibb Bibb Blk Male Blk
##           Offense longitude_osm latitude_osm geonameid longitude latitude
## 1 Muderous assault     -81.61821    33.57232  4569073 -81.63474 33.54437
## 2          Arson        -82.36401    29.67557  4145709 -82.35770 29.67476
## 3         Murder       -91.35985    34.29025  4099679 -91.37491 34.29081
## 4       Burglary      -81.41908    33.26410  4570020 -81.43502 33.26606
## 5         Murder      -86.45072    35.50995  4829092 -86.45889 35.51380
## 6         Murder      -87.12271    32.97108  4049189 -87.12644 32.99864
```

Convert results to sf object and plot on a map:

```
hal_gn = sf::st_as_sf(hal_gn,  
                      coords=c("longitude", "latitude"), crs=4326)  
plot(hal_gn["geometry"], axes=TRUE)
```



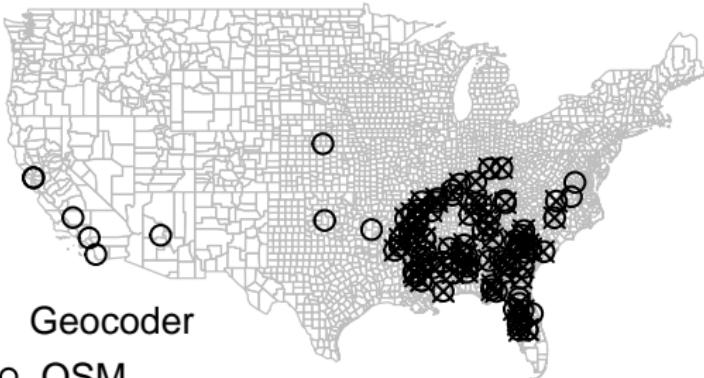
Boxplot

Load 1920 US county boundaries:

```
e1920 = sf::read_sf("Data/Counties/e1920.geojson")
```

Plot overlay with geocoded lynchings

```
plot(e1920["geometry"], border="gray", reset=FALSE)
plot(hal_osm["geometry"], col="black", pch=1, add=TRUE)
plot(hal_gn["geometry"], col="black", pch=4, add=TRUE)
legend("bottomleft", pch=c(1,4), col=c("black", "black"),
       legend=c("OSM", "GeoNames"), title="Geocoder", bty="n")
```



Geocoder

- OSM
- × GeoNames

Point-in-polygon analysis

Overlay points objects (hal_*) and polygons (e1920)

```
o_osm = sf::st_intersects(x = e1920, y = hal_osm)
o_gn = sf::st_intersects(x = e1920, y = hal_gn)
```

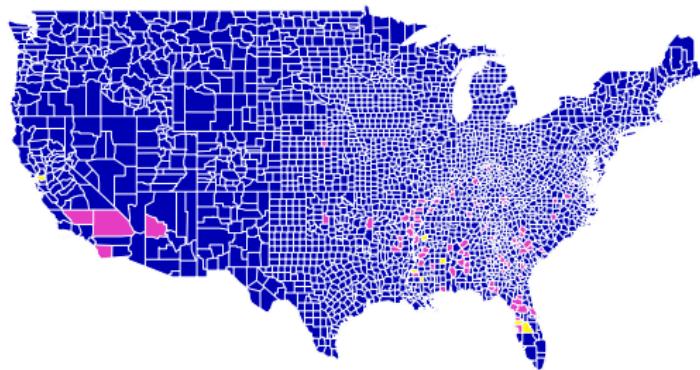
Assign counts to new variables

```
e1920$lynchings_osm = lengths(o_osm)
e1920$lynchings_gn = lengths(o_gn)
```

Plot the results

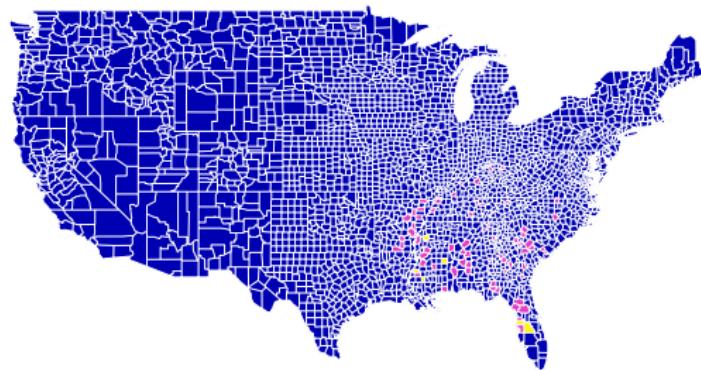
```
plot(e1920["lynchings_osm"], main = "Lynchings (OSM)")  
plot(e1920["lynchings_gn"], main = "Lynchings (GeoNames)")
```

Lynchings (OSM)



0.0 0.5 1.0 1.5 2.0

Lynchings (GeoNames)



0.0 0.5 1.0 1.5 2.0

Calculate new field (at least one lynching per county)

```
e1920$lynchings_osm_1 = 1*(e1920$lynchings_osm>0)
e1920$lynchings_gn_1 = 1*(e1920$lynchings_gn>0)
```

Number of counties with lynching, according to OSM vs. GeoNames

```
sum(e1920$lynchings_osm_1)
```

```
## [1] 89
```

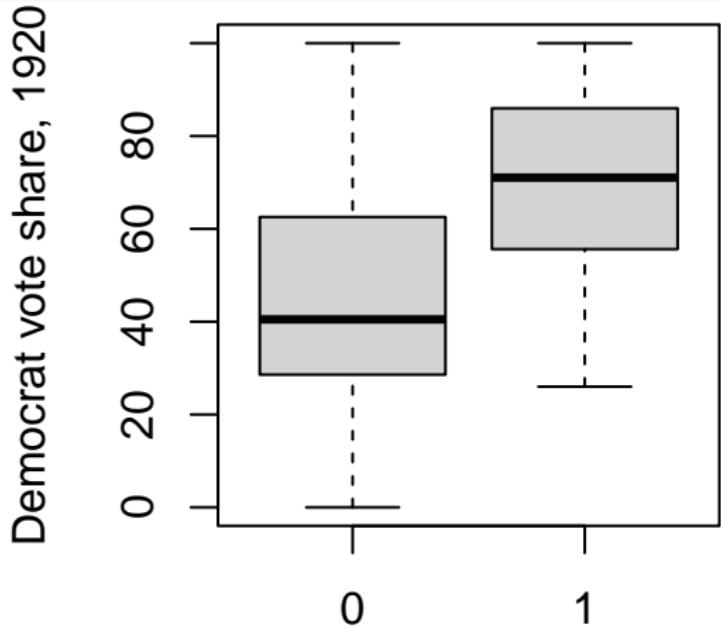
```
sum(e1920$lynchings_gn_1)
```

```
## [1] 74
```

Extra credit for anyone who figures out the source of the discrepancy and (partially) fixes it. Hint: it has to do with pelicans and beignets.

Create box plot for OSM-geocoded data

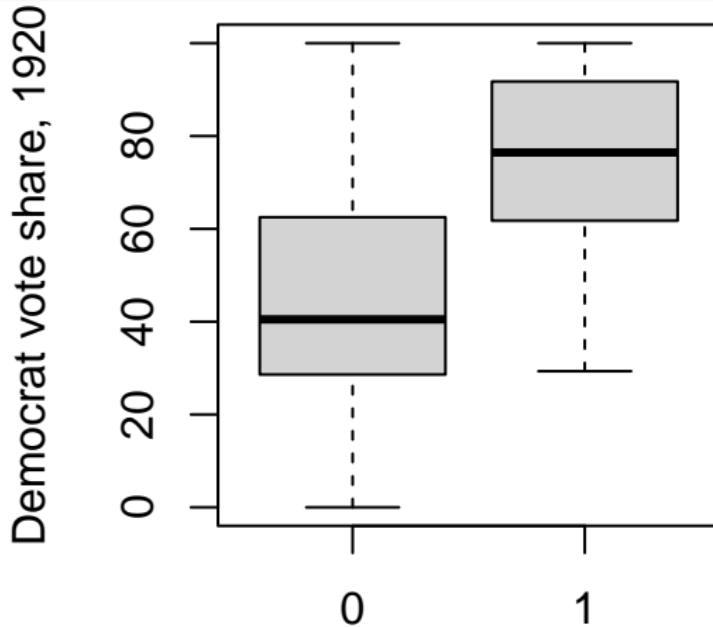
```
boxplot(DemMajPct ~ lynchings_osm_1, data = e1920,  
       xlab = "Lynchings (OSM)", ylab = "Democrat vote share, 1920")
```



Lynchings (OSM)

Create box plot for GeoNames-geocoded data

```
boxplot(DemMajPct ~ lynchings_gn_1, data = e1920,  
       xlab = "Lynchings (GeoNames)", ylab = "Democrat vote share, 1920")
```



Lynchings (GeoNames)

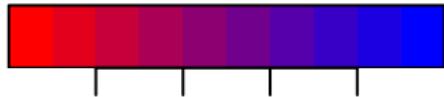
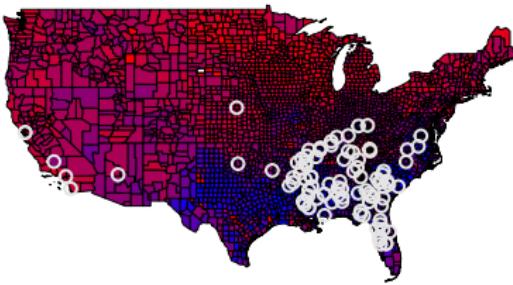
Create red-to-blue color ramp for maps

```
ramp = colorRampPalette(c("red","blue"), space = "rgb")
```

Map the OSM-geocoded locations:

```
plot(e1920["DemMajPct"], pal=ramp(10), reset = FALSE,  
     main = "Lynchings (OSM geocoded) and Democratic vote share in 1920")  
plot(hal_osm["geometry"], add=T, col="grey90", pch=1)
```

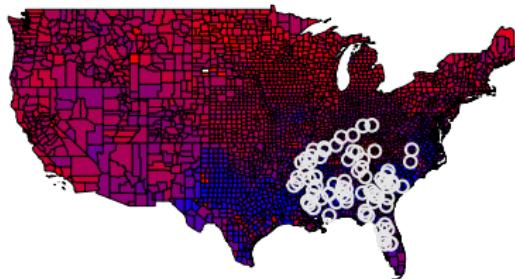
Lynchings (OSM geocoded) and Democratic vote share in 1920



Map the GeoNames-geocoded locations:

```
plot(e1920["DemMajPct"], pal=ramp(10), reset = FALSE,  
     main = "Lynchings (GeoNames) and Democratic vote share in 1920")  
plot(hal_gn["geometry"], add=T, col="grey90", pch=1)
```

Lynchings (GeoNames geocoded) and Democratic vote share in 1920



Bonus exercise: ggplot2

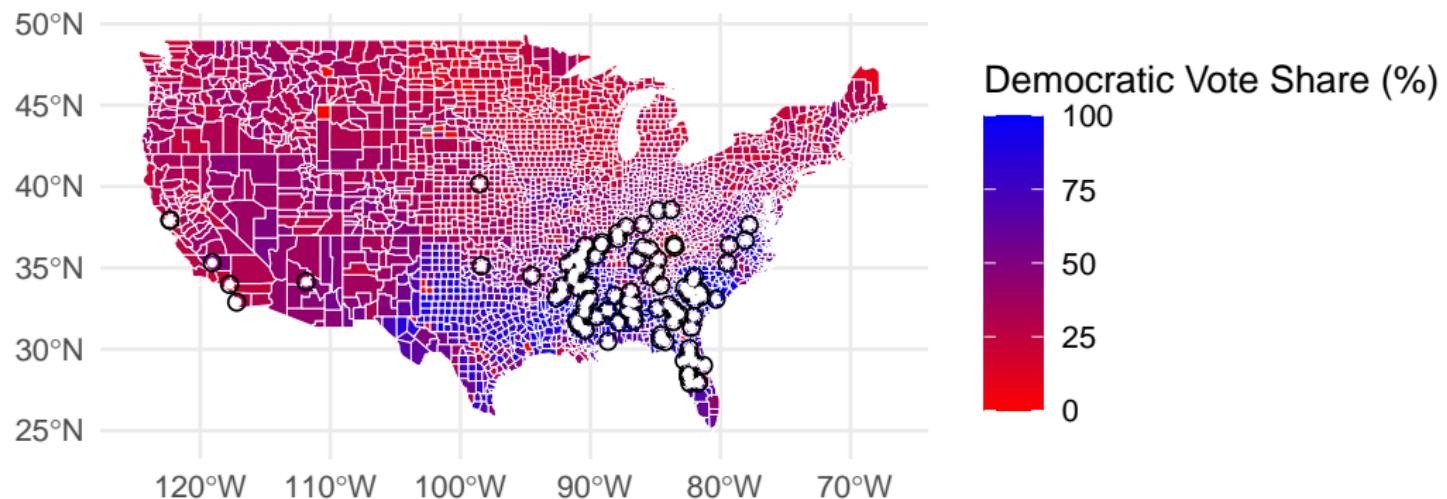
So far, we've been using "base R" plotting functions only. But there is another way to visualize spatial data in R, using the `ggplot2` package.

```
library(ggplot2)
```

The syntax is quite different. OSM example:

```
ggplot2::ggplot() +
  ggplot2::geom_sf(data=e1920, ggplot2::aes(fill=DemMajPct),
    color="white") +
  ggplot2::geom_sf(data=hal_osm, color="black", shape=1, size=2) +
  ggplot2::geom_sf(data=hal_osm, color="white", shape=16, size=1) +
  ggplot2::scale_fill_gradientn(colors=ramp(10)) +
  ggplot2::labs(
    title="Lynchings (OSM geocoded) and Democratic vote share in 1920",
    fill="Democratic Vote Share (%)") +
  ggplot2::theme_minimal()
```

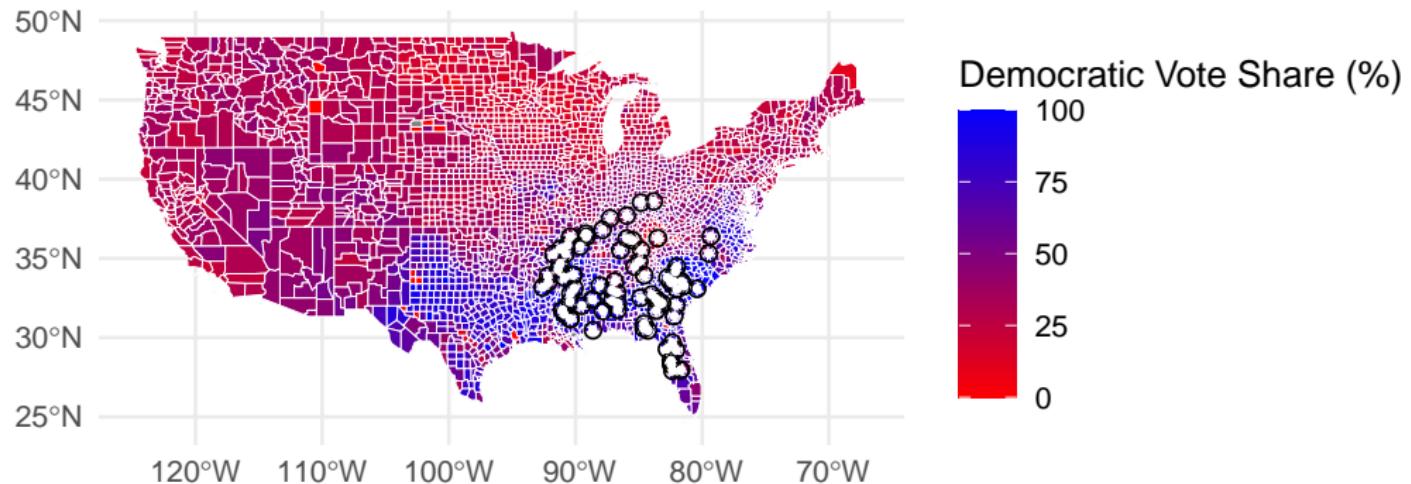
Lynchings (OSM geocoded) and Democratic vote share in 1920



GeoNames example:

```
ggplot2::ggplot() +
  ggplot2::geom_sf(data=e1920, ggplot2::aes(fill=DemMajPct),
    color="white") +
  ggplot2::geom_sf(data=hal_gn, color="black", shape=1, size=2) +
  ggplot2::geom_sf(data=hal_gn, color="white", shape=16, size=1) +
  ggplot2::scale_fill_gradientn(colors=ramp(10)) +
  ggplot2::labs(
    title="Lynchings (GeoNames) and Democratic vote share in 1920",
    fill="Democratic Vote Share (%)"
  ) +
  ggplot2::theme_minimal()
```

Lynchings (GeoNames geocoded) and Democratic vote share in 1920



Problem Set 6

Your assignment (if using R):

- create *one* of these boxplots (for OSM or GeoNames)
- ... and *one* of these maps (for OSM or GeoNames, base or ggplot2)
- save the graphics as `boxplot_osm_R.png` (`boxplot_gn_R.png`) and
`map_osm_R.png` (`map_gn_R.png`)
- upload these files to Canvas (by next Wednesday)