

11.205 – Intro to Spatial Analysis – Fall 2019

QGIS Exercise

Week 5 – Geocoding: Objectives

Part 1. The Geocoding Process

1. Obtain Data
2. Format Data
3. Geocoding with Online Geocoder
4. Visualize Using GIS
5. Use a Spatial Join to Summarize our Data
6. Format Data for QGIS Geocoder
7. Use QGIS Geocoder
8. Use Open Street Maps Geocoder Through QGIS.
9. Merge geocoded file to have one complete file.

Part 2. PARSING ADDRESSES AND GEOCODING WITH DESKTOP GIS

- VI. Parsing addresses for geocoding.
- VII. Geocoding using a desktop GIS

In lab this week we are going to focus on geocoding. The lab will be split into two parts; the first will introduce the geocoding process using a free, web-based geocoder. The second part will explore geocoding a bit deeper, building on the first part to format untidy data, explore alternatives, and identify errors.

The Cambridge Public Health Department is seeking a list and map of all grocery and convenience stores in the city in order to inform policy around the alleviation of food deserts. They have asked you to find a list of grocery stores and convenience stores in the city of Cambridge, create a map of their locations, and see how many grocery stores are in each neighborhood.

In lab this week we are going to focus on geocoding. The lab will be split into two parts; the first will introduce the geocoding process using OpenCage. The second part will explore geocoding using desktop GIS software.

The Cambridge Public Health Department is seeking a list and map of all grocery and convenience stores in the city in order to inform policy around the alleviation of food deserts. They have asked you to find a list of grocery stores and convenience stores in the city of Cambridge, create a map of their locations, and see how many grocery stores are in each neighborhood.

PART 1: THE GEOCODING PROCESS

Locating and plotting addresses and point locations is a fundamental function of GIS software. The process of spatially locating tabular address data is called **geocoding**. In lab today, we are going to

geocode a tabular dataset containing information on grocery stores in Cambridge. We will download the data, detail a web-based geocoding workflow, then look in-depth at how to clean, parse, and ultimately geocode addresses using a Desktop GIS.

OBJECTIVE 1: OBTAIN DATA

Let's get started! The first step is to acquire a tabular address dataset. This dataset can be any tabular data that contains addresses for each feature. In this exercise, we are going to download a table of grocery and convenience store addresses from a business database called Reference USA. Reference USA contains information on business locations and characteristics, and is a subscription service. MIT has a subscription, and so do many major public libraries across the country. It is a valuable source of business related information.

1. Navigate to Reference USA (www.referenceusa.com).

The screenshot shows the ReferenceUSA website. At the top, there is a navigation menu with links for 'ABOUT US', 'HOW DO I ACCESS?', 'DATABASES', 'LEARNING CENTER', and 'FAQ'. Below the menu is a blue banner with the text 'The premier source of information for reference and research' and 'Offers Accurate Data on 44M Businesses and 273M Consumers'. The main content area is titled 'Available Databases' and contains a grid of database options. The 'U.S. Businesses' option is highlighted with a red dashed box. To the right of the grid is a detailed description of the 'U.S. Businesses' database, including a 'SEARCH' button.

Database Name	Records
U.S. Businesses	44 Million Businesses
U.S. Historical Businesses	183 Million Historical Records
Canadian Businesses	1.9 Million Businesses
U.S. Jobs / Internships	2.5 Million Job Postings
U.S. Healthcare	1.3 Million Physicians & Dentists
U.S. Consumers / Lifestyles	273 Million Individuals
Canadian White Pages	12 Million Individuals

ReferenceUSA screenshot © Data Axle. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

2. Click on U.S. Businesses. Then select Advanced Search. The dialog below will open.

U.S. Businesses Database

Quick Search Advanced Search

Expand All

Select All

- Company Name
- Executives
- Business Type
- Geography
- Phone
- Business Size
- Ownership
- Financial Data
- Special Selects
- Exclusions

To start your search...

Click on the headings to the left to start your search. If you are uncertain what to search for, look for Search Tips within each section to help you along the way.

To further customize your search, select the Record Type you would like to search below to ensure you get the records you are looking for.

Record Type Search Tips | Collapse ⁺

- Verified Businesses** (Phone verified and quality checked)
- Include Unverified Businesses** (Not yet fully verified, may not be accurate)
- Include Closed / Out of Business Records** (Suspected to be out of business)

ReferenceUSA screenshot © Data Axle. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>.

In the Custom Search dialog, we want to specify the data we are looking for.

U.S. Businesses Database

Quick Search Advanced Search

Expand All
Select All

- Company Name
- Executives
- Business Type ✓
- Keyword/SIC/NAICS**
- Major Industry Group
- Geography ✓
- Map Based Search
- City / State**
- Metro Area
- ZIP Codes
- Radius
- County
- Street Address
- Neighborhood
- Phone
- Business Size
- Ownership
- Financial Data
- Special Selects
- Exclusions

Record Type Search Tips Collapse +

- Verified Businesses (Phone verified and quality checked)**
- Include Unverified Businesses (Not yet fully verified, may not be accurate)
- Include Closed / Out of Business Records (Suspected to be out of business)

Keyword/SIC/NAICS Search Tips Collapse + Remove ✕

- Search All SICs
- Search Primary SIC Only
- Search All NAICS**
- Search Primary NAICS Only

 SHOW 2 - 4 DIGIT CODES

ex Restaurants

Results:

Selected:

Enter 2-, 4-, or 6-digit [NAICS Codes](#) in the boxes below.

<input type="text" value="445110"/>	<input type="text" value="445120"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
-------------------------------------	-------------------------------------	----------------------	----------------------	----------------------

[Paste NAICS codes](#) [Clear Field\(s\)](#)

City / State Search Tips Collapse + Remove ✕

Search

 GO

Results

Results	Selected
Burlington, MA <i>City</i>	<input checked="" type="checkbox"/> Cambridge, MA <i>City</i>
Buzzards Bay, MA <i>City</i>	
Byfield, MA <i>City</i>	
Cambridge, MA <i>City</i>	
Canton, MA <i>City</i>	

ReferenceUSA screenshot © Data Axle. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <https://ocw.mit.edu/help/faq-fair-use/>

Fill out the form with the following:

- **Record Type:** Include only Verified Businesses
- **Business Type:** Keyword/SIC/NAICS - Search All NAICS -> 445110 (Grocery Stores) and 445120 (Convenience Stores) See the entire listing of NAICS here: <http://www.naics.com/search>

- **City/State:** Cambridge, MA

When done click View Results.

Side Note: What is NAICS?

From the Census [website](#): “The North American Industry Classification System (NAICS, pronounced Nakes) was developed under the direction and guidance of the Office of Management and Budget (OMB) as the standard for use by Federal statistical agencies in classifying business establishments for the collection, tabulation, presentation, and analysis of statistical data describing the U.S. economy. [...] NAICS is based on a production-oriented concept, meaning that it groups establishments into industries according to similarity in the processes used to produce goods or services.”

3. You will now see your resulting table from our query; we saw 70 results, but as this database is regularly updated, you may get slightly more or fewer results. It will contain listings for all the verified grocery stores and convenience stores in Cambridge. Click Download to download a tabular table of the data and save it to your working directory.

After clicking download, select:

1. Excel
2. Summary
3. All Records
4. Download Records and save as cambridge_grocery.xls.

OBJECTIVE 2: FORMAT DATA

The first step towards working with tabular data is cleaning and preparing the dataset. To do this, open the file in Microsoft Excel. You may see a warning message reading ‘The file format and extension of ‘Summary[timestamp].xls don’t match. The file could be corrupted or unsafe. Unless you trust its source, don’t open it. Do you want to open it anyway?’ This is fine – go ahead and open it.

When you open the data, notice that there is an Address field that contains the street number and name. Also note there are City, State, and Zip Code fields. These fields are necessary for our geocoding, as they contain the information we need to create an Address Locator and locate points. Other fields might be useful as well, though we won’t necessarily use them this time around.

1. Delete the Fax Number, IUSA Number, Primary SIC Code, Primary SIC Description, SIC Code 1, and Record Type field.
2. Remove spaces from field headers and reduce their length. These should **make sense to you**. Most GIS system can not handle field with character greater than 8. I changed to use these names:

Company Name	name
Executive First Name	execFrst
Executive Last Name	execLast
Address	address
City	city
State	state
ZIP Code	Zip

Credit Score Alpha	credit
Executive Gender	execGdr
Executive Title	execTtl
Location Employee Size Range	empSize
Location Sales Volume Range	salesVol
Phone Number Combined	phone
SIC Code 1 Description	type

3. Finally, we also need to create a **unique identifying column**. Create a new column called 'ID' at the left of your sheet, place a number 1 in the first row. In the second row, use the formula **=A2+1** to increment by 1. Then fill this to the bottom of the table. Click the column, copy it, right click the same column, and select Paste Special > Values.
4. We need the address to be in one field so we need to format it.
5. Create a new field called **Address2**.
6. Click on the cell and copy the following formula so that address is in one line.

=CONCATENATE(F2, ", ", G2, ", ", H2, ", ", I2)

Notice that formula has added spaces and comma so the program know the differences between the town, state, and zip.

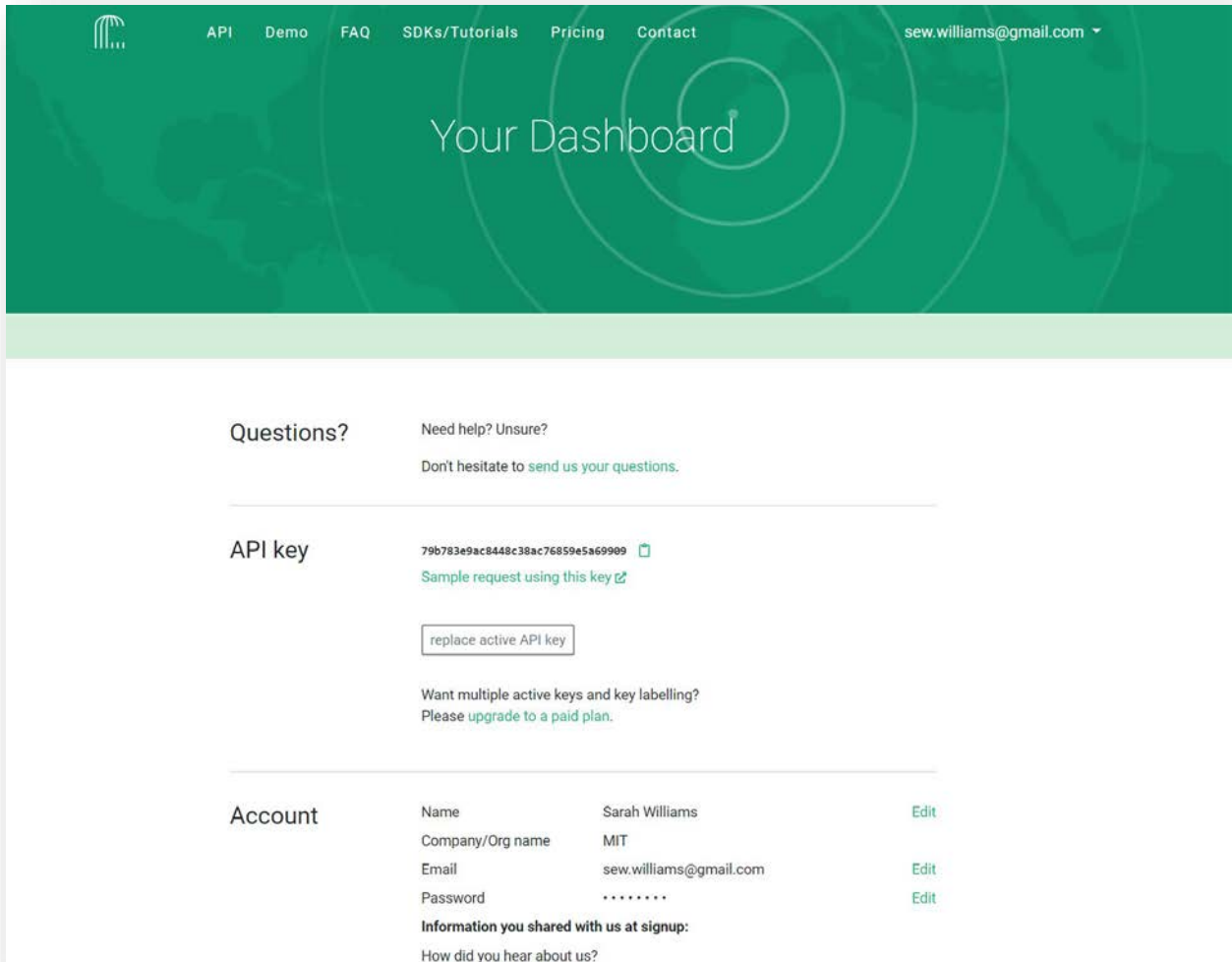
Double-click on the right-hand side of the cell to make the formula go all the way down the addresses.

7. Now we need to make this concatenated address permanent. Click the column, copy it, right click the same column, and select Paste Special > Values. You will know you did the correctly if you don't see a formula in the cell.
8. Save the prepared file as '**cambridge_grocery_cleaned.xls**'. This will be our starting point for additional cleaning work!

OBJECTIVE 3: GEOCODING WITH THE OPENCAGE

There are a large number of geocoding services available and they change all the time – one free option if you are geocoding something that is under 2,500 records a day is [OpenCage](#). What's great about OpenCage is they use a number of different opensource geocoders in the background. This means that your data is still owned by you. If you use Google's Geocoding API technically they own your data. OpenCage prides itself in following the guidelines of GDPR (the EU's General Data Protection Regulation) which you own our own data.

1. Go to the OpenCage Website and sign-up for an account. Once you have your account set-up you should see on your dashboard an API key. You will need that to geocode your addresses. Copy your API key somewhere. You will need it later.



2. OpenCage makes it easy to geocode using a lot of different format everything from Java to R. We are going to use the option that uses the least amount of coding – Google Sheets. Yes you can geocode directly from Google Sheets.

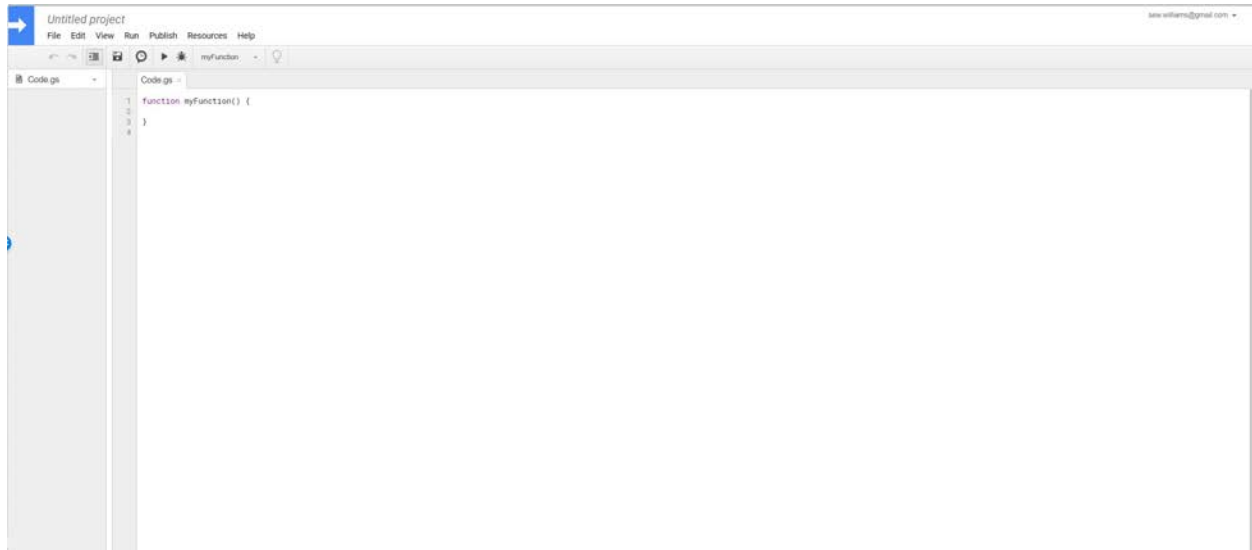
Load the Excel file you just created into Google Drive. If you do not have Google Drive account, we suggest that you sign up for it at <https://www.google.com/drive>.

1. In **Google Sheets** select File> Import > select **cambridge_grocery_cleaned.xls**. Do not simply upload the **.xls** file as we need the file to be natively in Google Sheets. Name the file **Cambridge Grocery**.

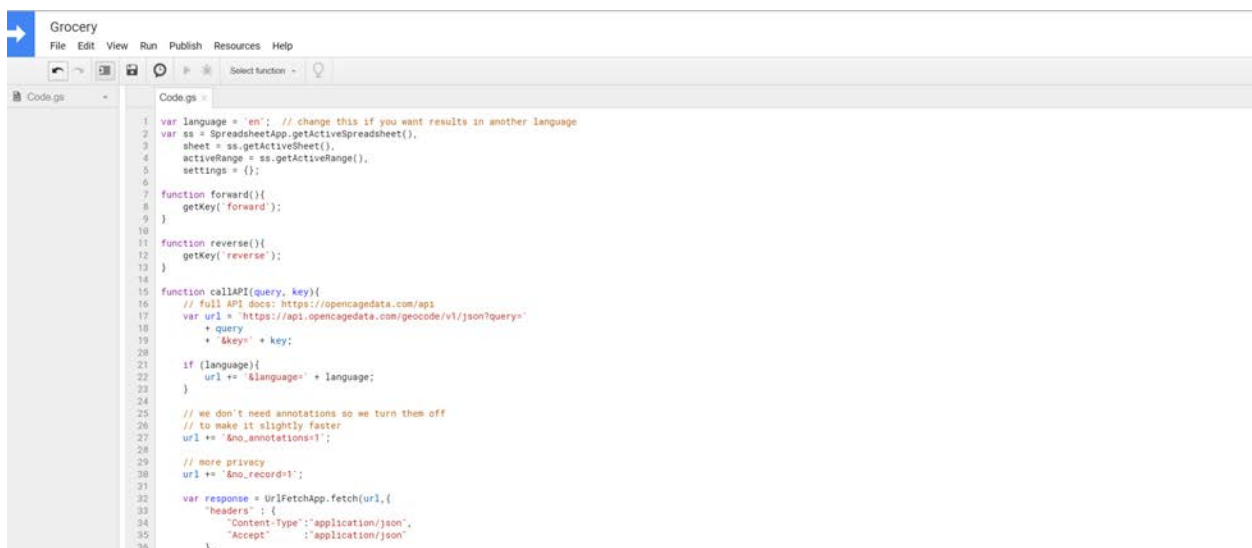
2. Once Opened in Google Sheets, add 2 new fields - Latitude and Longitude.

Add the Open Cage Script to Google Sheets

1. In your Google Sheet click **Tools> Script Editor** The following will appear. We need to add our OpenCage script into the dialog.



2. Go to OpenCage and click on **SDKs/Tutorials** from the main menu (<https://opencagedata.com/code>). Scroll down and pick the Google Docs Tutorial.
3. The tutorial asks you to go to [opencage.js](https://opencagedata.com/code).
4. Now copy all the code in the code space and go back your Google sheets script editor and completely replace with this script. Call the script project **Grocery** and **save the script using the save icon**. It should look like below.



5. Go back to your Cambridge Grocery Google Sheet (not the script editor). You should see a separate tab appear after you reload the page. After about 30 seconds you should see a new tab in your sheet called **Geocode**.
6. Select these three columns: "Address2, Latitude, Longitude "
7. Then go to the Geocode tab and select Geocode > Address > to Latitude, Longitude
8. You will be prompted to give the script permission to run. You will need to select "Continue".
9. A popup will appear asking you which Google account you want to authorize. Select one.
10. The popup will display a warning that our app isn't verified. Click "Advanced"
11. Scroll down and then click on the link to "Go to your-project-name (unsafe)"
12. Now you will need to give the project access to your spreadsheets. Click "Allow"
13. A popup will appear in your spreadsheet asking your for your OpenCage API key. Enter it and click "Geocode". If you forgot where you got this, OpenCage sent you an email or it's in your Opencage dashboard.
14. Now we wait. According to Opencage's website it take roughly 7 seconds for each address to be found through their API. $7 \times 77 = 539$ seconds when you divide that by 60 you get roughly **9 minutes** (it's usually a little quicker). So take a break to ask a question about the process so far, organize your files or anything else you want to do.

You will notice that it tries to geocode the Header field. Rename the headers to Latitude and Longitude. If the geocoder gives you all errors in the latitude and longitude field the first time, fear not. Just try it again and it should work the second time around.

You will also notice that there are a number of errors. What do you think the reason for this might be? It's because of the qualifying numbers such as 468 Broadway #2. The #2 is confusing the geocoder service. Edit that row in your sheet highlight the Address, Latitude and Longitude fields for only that row that you corrected and then run the geocoder again on just that record. It should fix the situation.

Now go ahead and fix all the Addresses that have errors. Can you think of systemic ways to fix this rather than going through them one by one?

Once you have corrected the addressed and geocoded all 70 address let's download them so we can visualize in GIS. Use **File>Download** and download as a **CSV**. I called my file `cambridge_grocery_results.csv`.

OBJECTIVE 4: VISUALIZE USING GIS

1. Open QGIS.
2. You can add this CSV to QGIS using the 'Add Delimited Text' tool. Go to 'Layer' in the top menu > 'Add Layer' > 'Add Delimited Text Layer.'
3. Your settings should look like the below image. Make sure that your X field is 'LON' and your Y field is 'LAT'!
4. Set your Geometry CRS to EPSG:4326 WGS 84.
5. Click 'Add' when you are done.

File Name ...

Layer Name Encoding

File format

CSV (comma separated values)

Regular expression delimiter

Custom delimiters

Record and fields options

Geometry definition

Point coordinates X field

Well known text (WKT) Y field

No geometry (attribute only table) DMS coordinates

Geometry CRS

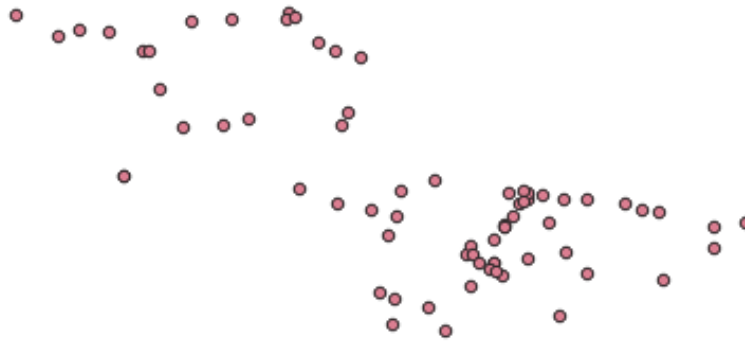
Layer settings

Sample data

	FID	ADDRESS	MATCH	PRECISION	
1	35	2225 Massachusetts Ave, Cambridge, MA, 02140	Match	Exact	2225 MASSACHUSE
2	36	875 Cambridge St, Cambridge, MA, 02141	Match	Exact	875 CAMBRIDGE ST
3	33	48 Jfk St, Cambridge, MA, 02138	Match	Exact	48 JFK ST, CAMBRI
4	34	87 Pearl St, Cambridge, MA, 02139	Match	Exact	87 PEARL ST, CAME

Help Add Close

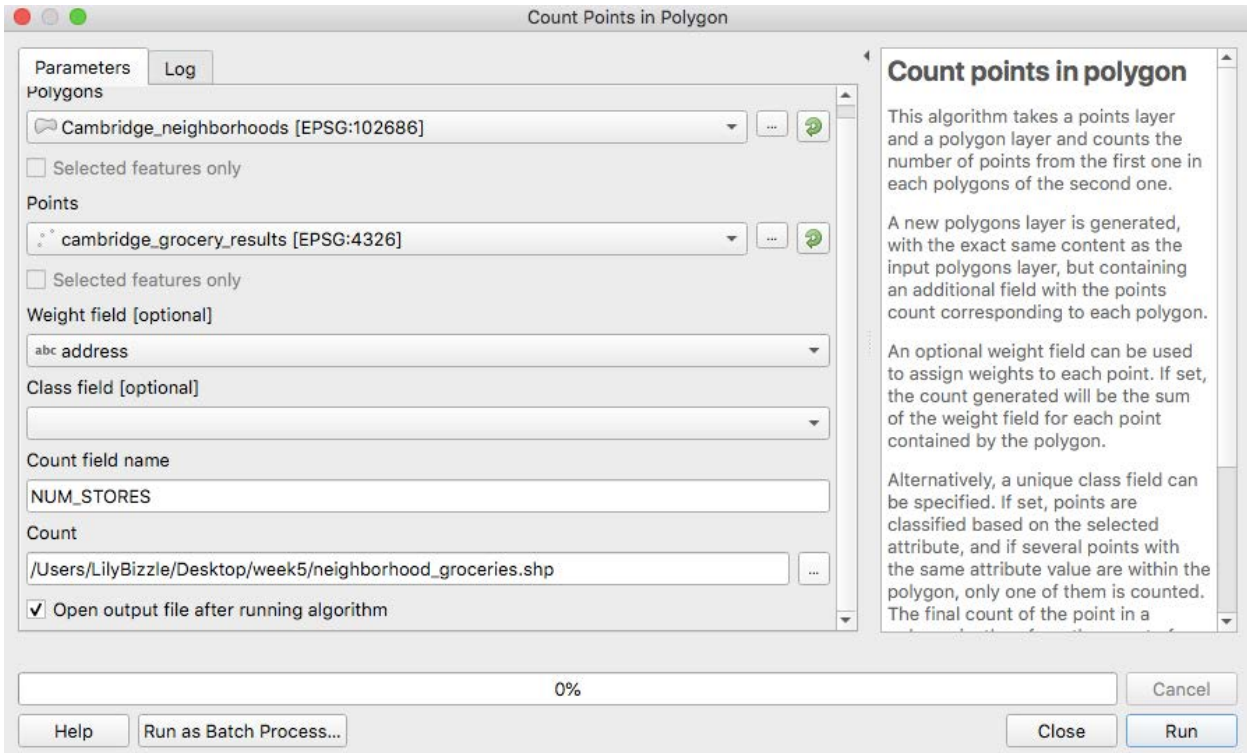
You should see something like this:



OBJECTIVE 5: USE POINTS IN POLYGON (SPATIAL JOIN) TO SUMMARIZE OUR DATA

We now have a file containing points all verified grocery store and convenience store locations in Cambridge. Using a Spatial Join, we are going to find out how many grocery stores and convenience stores are in each Cambridge neighborhood. Create a new QGIS document called Week5.qgis

1. Add **Cambridge_StreetCenterlines.shp** to your project, along with and **Cambridge_neighborhoods.shp**.
2. We are going to count the number of points in each polygon. Click the **vector** drop-down menu and select **Analysis > Count Points in Polygon**. Alternatively, search in your toolbox for **'Count Points in Polygon.'**
3. In the dialog that appears, choose the **neighborhoods polygons as your 'Polygons'** layer, your **geocoded points as your 'Points'** layer, and name the count field name **'num_stores'**. Set the output to save into your workspace, and name it **'neighborhood_groceries.shp'**. Click **'Run.'** Save a temporary layer for now.



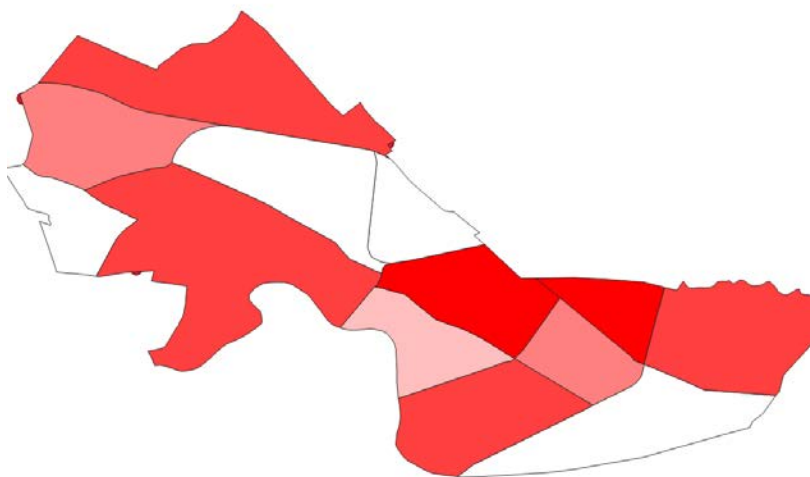
9. Open the attribute table and see the summary information.

Count :: Features Total: 13, Filtered: 13, Selected: 0

	Webpage	N_HOOD	NAME	GlobalID	NUM_STORES
1	http://www.c...	6	Mid-Cambrid...	{54BA721E-5...	11
2	http://www.c...	3	Wellington-H...	{BCF123EA-5...	9
3	http://www.c...	5	Cambridgeport	{BA795F74-0...	8
4	http://www.c...	10	West Cambri...	{FC9B6813-1...	8
5	http://www.c...	11	North Cambri...	{CB9A6B5D-...	8
6	http://www.c...	1	East Cambrid...	{1533C532-2...	7
7	http://www.c...	12	Cambridge H...	{846719F3-3...	6
8	http://www.c...	4	The Port	{93279187-C...	6
9	http://www.c...	7	Riverside	{BC1F44A9-...	4
10	http://www.c...	2	Area 2/MIT	{7E7AA8AC-...	1
11	http://www.c...	8	Agassiz	{63D00AF5-...	1
12	http://www.c...	9	Neighborhoo...	{8ABEBD64-...	1
13	http://www.c...	13	Strawberry Hill	{E9ABF249-...	0

Show All Features

5. Symbolize the map by the number of grocery stores. Your map document should look similar to this.



Now **continue on to part two**, where we'll be learning to geocode using QGIS!

PART 2: GEOCODING WITH A DESKTOP GIS

In Part 1 of this exercise, we used OpenCage to geocode data downloaded from Reference USA. This is convenient, but it hides much of the inner workings from us; as a service, it's great, but it leaves us with much to learn! It's worth knowing how to use a Desktop GIS to accomplish the same task.

If you still have QGIS open from your last exercise please close it now.

OBJECTIVE 1: PARSING ADDRESS DATA

It is often the case that data we find will not be formatted in a way that is intelligible to a geocoder. Often, addresses are in a single field (e.g., '77 Massachusetts Ave, Cambridge, MA 02139') like OpenCage required. Other times a geocoder will require that an address be split into parts (e.g., its street number - '77' – its street name - 'Massachusetts Ave'). Furthermore, there may be errors or ambiguities in the data that will confuse a given geocoding service. For example, when we had the error of the OpenCage geocoder not being able to geocode addresses where it had the unit number.

Geocoding errors and ambiguities can be found in either the reference dataset (the street file used to locate geocoded points) or the address dataset to be geocoded.

It is impractical and often impossible to fix reference datasets. However, it is always within your power to make sure an address dataset is formatted to support geocoding.

To start, let's work with the file we already cleaned for OpenCage. Go to the Google Sheet file you created for the OpenCage Geocoder.

Making New Fields for the Geocoding Service

In the sample set, you see that our address is entirely contained in one column. You are probably used to entering a full street address, including city, state, and zip code, into Google Maps. Although you do not see this, Google Maps and the major geocoders parse this address into its components before running it through their geocoding engine, which is in essence a very large Address Locator like we created in Part 1. Most geocoding methods require you to parse address data into constituent components before processing.

For the QGIS geocoder that we will be using it want the data to have the following fields:

1. Street & Type
2. Number
3. Zip

We already have the ZIP field isolated but we need the "Street and Type" and "Number" separated as unique fields. Making these new fields are easy in Google Sheets or Excel. Let's do this now.

In Google Sheets add **two fields** one called **Street** and one call **number**.

Now we need to parse the data so that we get these values.

Let's start with the number field. Enter the following formula into your Google Sheets.

```
=LEFT(H2,FIND(" ",H2)-1)
```

What this formula does is find the first space in the field H. In the example below the first space happens at character "5". Once the formula finds the character location (5) the LEFT function extracts that number of character – 1 on the left hand side of the field H.

Double-click on the blue dot on the corner of the file (red circle above) to populate the formula all the way down.

Double-check that the values are all correct. Now we want to make the formula permanent. Select the entire Number Field. Right-click and pick "Copy". Then Right-click again and select Paste>Paste Values Only. To check that you did this correctly, click on a cell and see if a formula pop's-up. If no formula pops up then you did this correctly.

Now we need to make a formula that produces just the street Name and Type. Ideas of how we might create this formula?

This time the formula looks like the following:

```
=RIGHT(H2, (LEN(H2)) - (FIND(" ", H2)))
```

What this formula does is find the Length of field H and subtract by the first character location. In the case below that means 22-5 = 17. Then the formula gets 17 characters starting from the right.

You will want to do the same thing you did last time. Double click on the blue dot to bring the formula all the way down. Then check to make sure the values are correct. Then make the change permanent. Select the entire Number Field. Right-click and pick "Copy". Then Right-click again and select Paste>Paste Values Only. To check that you did this correctly, click on a cell and see if a formula pop's-up. If no formula pops up then you did this correctly.

Standardize the various components of the address

Last time we geocoded this file the inclusion of the "Suite Number" gave the program an error. Can you think of a way to automatically remove these so this problem does not happen again? We can do it with another formula. Make a new field called Street2. Add the formula below.

This time the formula looks like below:

```
=IF(REGEXMATCH(J2, "#"), LEFT(J2, (FIND("#", J2) - 1)), J2)
```

This formula will search if the value "#" exists. If it does it will run the "LEFT(J2, (FIND("#", J2) - 1))" if it doesn't exist it will keep the original values.

You will want to do the same thing you did last time. Double click on the blue dot to bring the formula all the way down. Then check to make sure the values are correct. Then make the change permanent. Select the entire Number Field. Right-click and pick "Copy". Then Right-click again and select Paste>Paste Values Only. To check that you did this correctly, click on a cell and see if a formula pop's-up. If no formula pops up then you did this correctly.

Now we can delete the first street field-- we no longer need it.

Now download this as a CSV file and call it "cambridge_grocery_parsed.csv"

OBJECTIVE 2: USE MMQGIS AND A CAMBRIDGE STREET FILE TO GEOCODE

To begin geocoding the data we just parsed, we can use a QGIS plugin called 'MMQGIS' which includes (among many other exciting features!) a geocoder. First, you'll need to install MMQGIS; to do this, select Plugins > Manage and Install Plugins, and then search for MMQGIS.



You should now have an MMQGIS dropdown menu available.

Add the Cambridge_StreetCenterlines.shp file to your project. We'll be using these streets as a reference dataset against which to locate our addresses!

From the MMQGIS dropdown, select MMQGIS > Geocode > Geocode from Street Layer. You're now presented with a large number of options. This screen should match the configuration on the following page (obviously with file paths that make sense for your directory structure)!

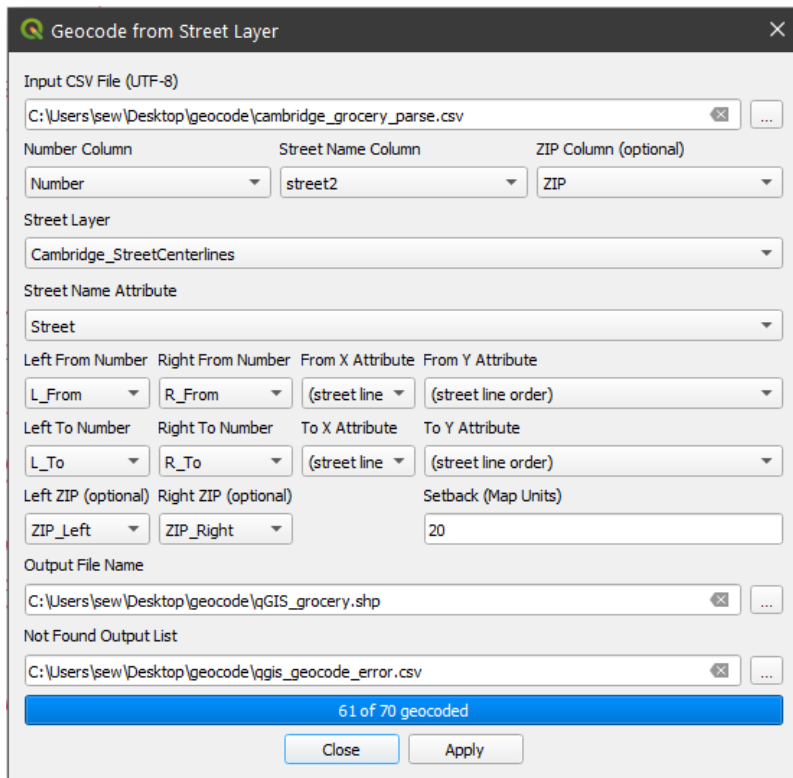
Note: use Street, not Street_Nam for the street layer's Street Name attribute. This is because Street_Nam doesn't include the street type; in contrast, our parsed csv includes the street name and the street type in one field. These need to match!

Enter in the field names as seen in the screen shot below.

The setback of 20 is very important. This make sure that the point is not directly on the street centerline. You know what side of the street it is on. This is very important for spatial joins.

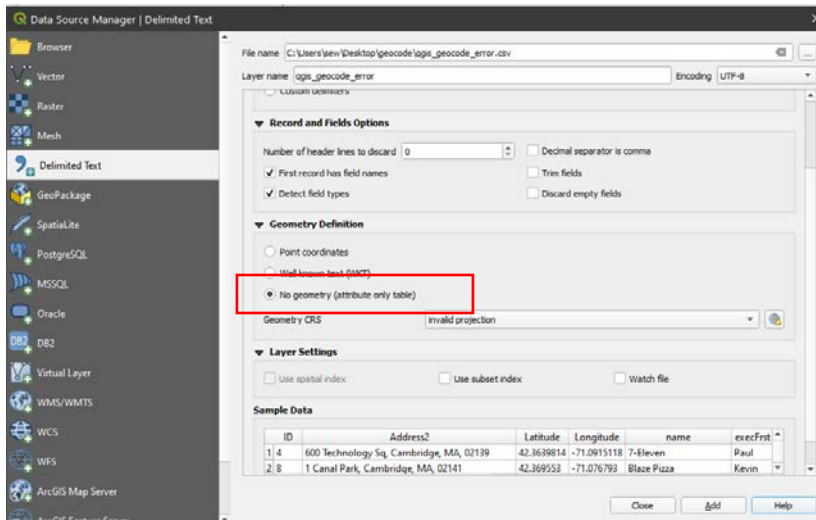
Name you output file **qGIS_grocery.shp** .

Name the errors as **qGIS_geocode_error.csv**. If there is an error in geocoding this will allow you to see what those are.



Click Apply and you get 61 out of 70 geocoded as a match.

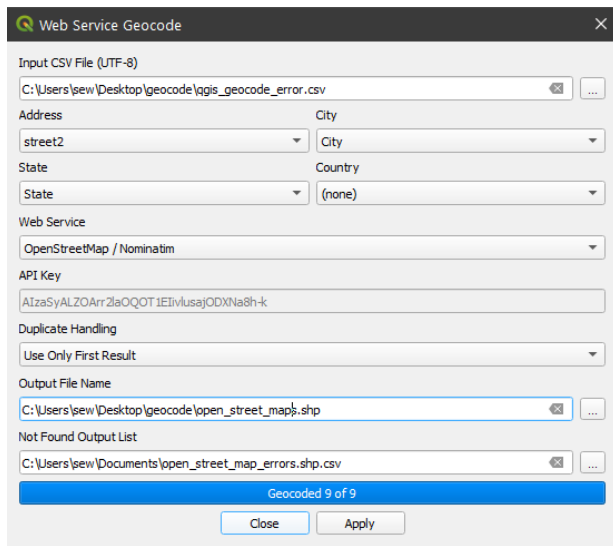
Now let's open the gis_geocode_errro.csv file. Do this by going to the Layer> Add Layer > Add Delimited Text Layer. Choose the qgis_geocode_error.csv you just created. You will need to go to the the Geometry Definition and click "no geometry (attribute only)". You need to do this because you file had the X, Y coordinates from OpenCage still inside. See Below.



Can you guess why some of the addresses were not geocoded? It is because the streets were not represented in the Cambridge Center Lines in this way. For example “Mount Auburn St.” might be “Mt. Auburn St.” Or something like Central Sq is not found as a street.

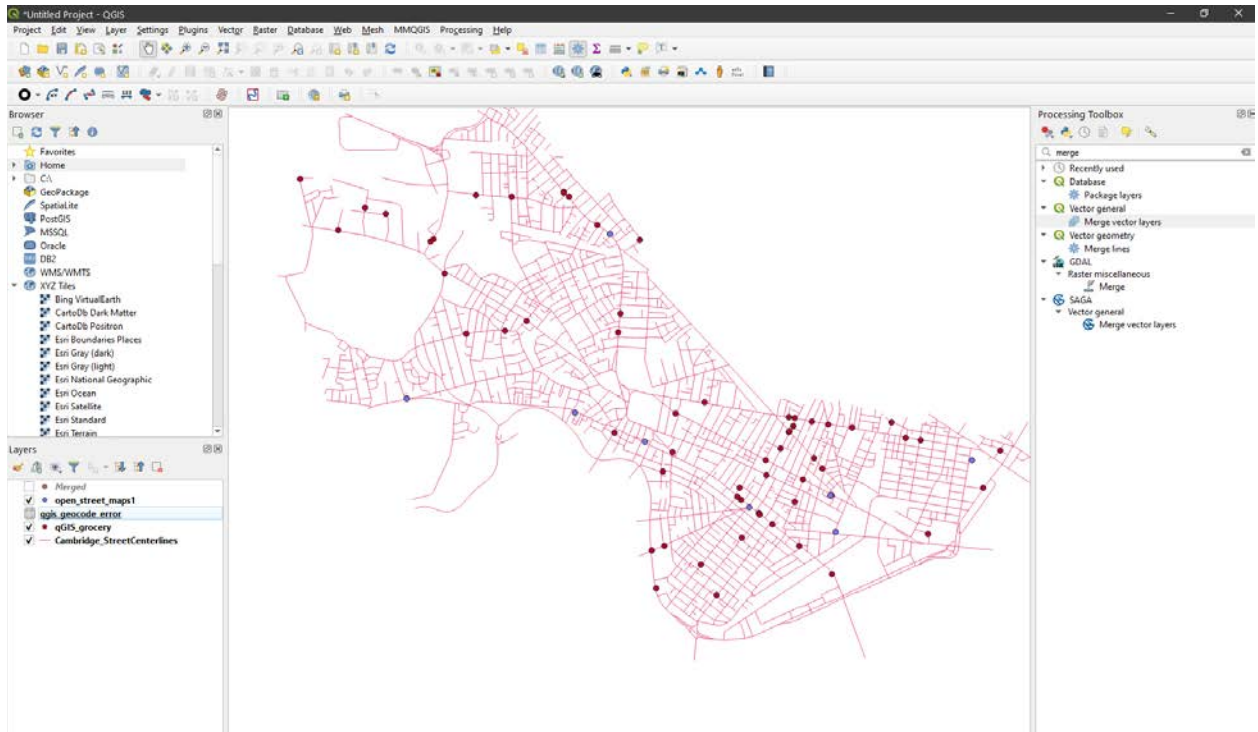
What can we do now? One option is to try to Geocode it with another geocoding service that might have these street names. In fact, that is how OpenCage works, it tries multiple geocoding services until it get all the addresses. This is why it is such a powerful service.

Let’s try the OpenStreetMaps Geocoder for the files we found errors for. Here is how that works. Go to MMQGIS>Geocode>Geocode CSV with Web Service.

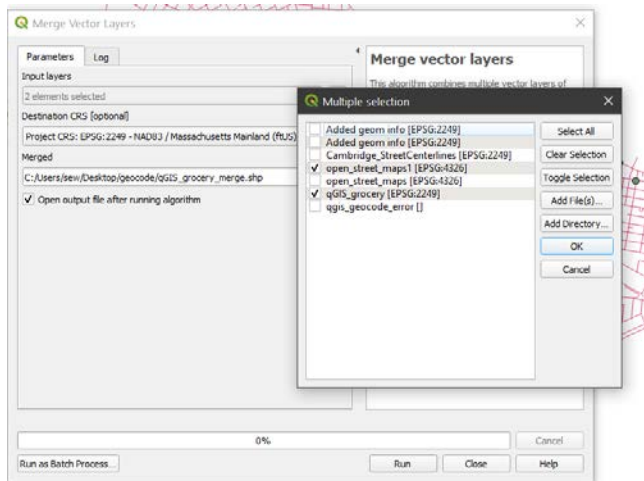


The Open Streets Map Geocoder was able to geocode all 9 of the errors see (Geocoded 9 of 9).

The new file will automatically be added to you map. See below:



Let's merge these files so we have one whole file. Let's go to Processing>Toolbox and then search "merge". To select **Merge Vector Layers**. The dialog box below will appear. Select that you want to merge your open_street_maps file with your qGISfile_grocery file. Make the Destination CRS "Project CRS:EPSG:2249 – NAD83/Massachusetts"



We now have a finished file!! Can you think of a reason we would do the geocoding this way instead of using OpenCage? – Spoiler Alert – Open Cage as a limit of 2500 points a day. You would have to pay if you want more than that. You might have a database of over 2500 points.

MIT OpenCourseWare
<https://ocw.mit.edu/>

11.205 Introduction to Spatial Analysis
Fall 2019

For information about citing these materials or our Terms of Use, visit: <https://ocw.mit.edu/terms>.