Geocoding Applications with Rstudio

- Presenter: Aaron Majek, MS (aaron.Majek@sulross.edu

- TAIR Conference 2023

Presentation Objectives

- Understand the applications of Rstudio for geocoding in IR
- Demonstrate utilization of Rstudio geocoding packages / processes
- Iterate on several use cases for geocoding with Rstudio

SUL ROSS
The FRONTIER University of Texas

Need for Geocoding

- Need to inform stakeholders about their targeted audience such that they may adopt strategies better suited to their service population.

- Need for data driven and data informed strategies at institutions of higher education.

- Census geocoded can substitute for instances wherein institutional data is missing or lacking.

What is Rstudio?

- Rstudio is an Integrated development environment (IDE) providing users with an environment for writing and executing code.

- It is available with a GNU General public license meaning it is open source and available to users at no cost.

- Rstudio contains a library of ‘packages’ – sets of code and documentation which may be accessed from a centralized depository.

- As of Feb. 2023, there are at least 19,254 of such packages available for public use.

SUL ROSS
The FRONTIER University of Texas

Rstudio Geocoding Packages

- TidyVerse: Contains sets of R packages that are standard to most Rstudio functions.

- Tigris: Load Census TIGER/Line Shapefiles. Connects with US census API facilitating upload of shp files to your Rstudio IDE.

- Tidycensus: Connects with US census API to allow upload of census data files to your Rstudio IDE (IE: American Community Survey).

- Sf: supports a standardized way of encoding spatial vector data. Allows for conversion and manipulation of shp files.

- Tidygeocoder: allows for a the stepwise execution of queries made to numerous geocoding API services at once.

SR SUL ROSS
The FRONTIER University of Texas

TidyGeocode API

- The TidyGeocode Package is a tool used in Rstudio to geocode student addresses and extract lat and long coordinates.
- To the right are a set of geocoding APIs that can be queried using the Tidygeocode package.
- Tidygeocode has a ‘cascade’ feature that allows users to assign prioritization to certain API calls over others in an iterative sequence
- For this analysis, we are only using the two free API calls: the census batch query api and the Nominatim API

Details

The API documentation for each service is linked to below:

- Nominatim
- US Census
- ArcGIS
- Geocodio
- Location IQ
- Google
- OpenCage
- Mapbox
- HERE
- TomTom
- MapQuest
- Bing

SF Package

- The SF package supports the standardization of spatial vector data by representing geographic information as a dataframe.

- It interfaces with the GEOS C/C++ library used in Geographic information systems (GIS) software allowing transformations on projected geographic points.

- Additionally, it interfaces with PROJ, a coordinate transformation library that allows for performing conversions between cartographic projections.

- The combination of these two interfacing capacities (with GEOS and PROJ) facilitates the transformation of otherwise uni-dimensional data frames into multi-dimensional datafiles across spatial interfaces.

- Together with Tidygeocode and other packages – you can merge census data into your student records.

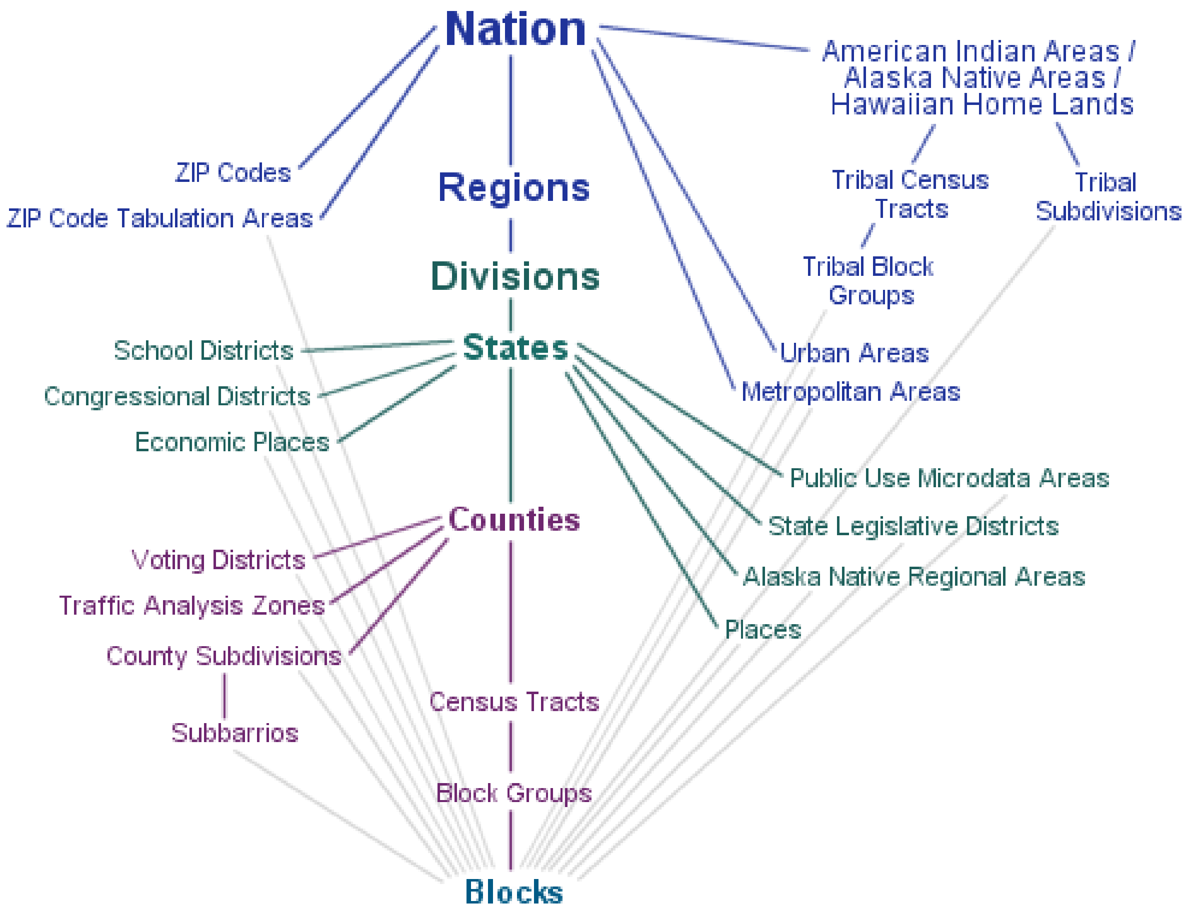
Census Geographies

- Census Geographies are the unit of analysis for the US census.

- Some census data instruments are only available at certain geographies

    - (IE S1701 poverty data is available at census tract level but not the block group level)

- The most granular census geography for which data may be matched and merged to student datafiles is the census block group level



Putting Rstudio to work: I

- Step 1 – install your packages.

```
#step1b - adding stage A packages
library("tidyverse")
library('readxl')
library("tidygeocoder")

#step1c - adding stage B packages
library(tidycensus)
library(tigris)
library(sf)
```

Putting Rstudio to work: II

- Step 2 – Initialize the geocoding function. Here we created the function 'geocode_chunked'. It breaks large address data frames into smaller ones and iterates a stepwise geocoding procedure on them. This was done for two reasons:

```
#this code executes a stepwise process for geocoding addresses by breaki
geocode_chunked <- function(df, chunk_size){
  chunk_count <- ceiling(nrow(df) / chunk_size)
  results_list <- list()

  for (i in 1:chunk_count) {
    start_index <- (i - 1) * chunk_size + 1
    end_index <- min(i * chunk_size, nrow(df))
    df_chunk <- df[start_index:end_index, ]

    df_chunk <- df_chunk %>%
      geocode_combine(
        queries = list(list(method= 'census'), list(method= 'osm')),
        global_params = list(address='ADDRESS'), cascade = TRUE
      )

    results_list[[i]] <- df_chunk
  }

  newdf <- bind_rows(results_list)
```

- Reason one: The census geocoding API only works with files less than 10,000 rows long. Breaking the dataframe into chunks allows users to geocode in one batch.

SUL ROSS
The FRONTIER University of Texas

Putting Rstudio to work: III

- Step 2 – Initialize the geocoding function. Here we created the function 'geocode_chunked'. It breaks large address data frames into smaller ones and iterates a stepwise geocoding procedure on them. This was done for two reasons:

```
Passing 20 addresses to the US Census batch geocoder
Query completed in: 0.7 seconds
Passing 6 addresses to the Nominatim single address geocoder
[==============================================================================]


Passing 18 addresses to the US Census batch geocoder
Query completed in: 0.6 seconds
Passing 6 addresses to the Nominatim single address geocoder
[==============================================================================]


Passing 12 addresses to the US Census batch geocoder
Query completed in: 0.4 seconds
Passing 3 addresses to the Nominatim single address geocoder
[==============================================================================]
```

- Reason Two: Large dataframes take a long time to geocode. Breaking the frames into chunks allow you to see that the geocoding process is working

SR | SUL ROSS The FRONTIER University of Texas

Putting Rstudio to work: IV

- Step 3 – Review the results. When the geocoding procedure is completed, it will produce at lat and long coordinate set in two respective columns. It will also generate a column indicating which geocoding API was used to produce a positive result. If no results were located then the lat, long, and query column will be NULL.

	INST_NAME	INST_IPEDS_CODE	DIGIT_ANON	ADDRESS	lat	long	query
34	Sul Ross State University	228501			30.35556	-103.67845	census
35	Sul Ross State University	228501			30.10718	-98.42077	census
36	Sul Ross State University	228501			30.35313	-103.61429	census
37	Sul Ross State University	228501			30.34860	-103.68219	census
38	Sul Ross State University	228501			31.76151	-106.30815	census
39	Sul Ross State University	228501			29.79140	-95.13488	census
40	Sul Ross State University	228501			29.75774	-99.02438	census
41	Sul Ross State University	228501			30.50989	-97.78177	census
42	Sul Ross State University	228501			29.56478	-104.36340	census
43	Sul Ross State University	228501			31.82239	-106.46536	census

SUL ROSS
The FRONTIER University of Texas

Putting Rstudio to work: IV

- Step 4 – Review the results. When the geocoding procedure is completed, it will produce at lat and long coordinate set in two respective columns. It will also generate a column indicating which geocoding API was used to produce a positive result. If no results were located then the lat, long, and query column will be NULL.

	INST_NAME	INST_IPEDS_CODE	DIGIT_ANON	ADDRESS	lat	long	query
34	Sul Ross State University	228501			30.35556	-103.67845	census
35	Sul Ross State University	228501			30.10718	-98.42077	census
36	Sul Ross State University	228501			30.35313	-103.61429	census
37	Sul Ross State University	228501			30.34860	-103.68219	census
38	Sul Ross State University	228501			31.76151	-106.30815	census
39	Sul Ross State University	228501			29.79140	-95.13488	census
40	Sul Ross State University	228501			29.75774	-99.02438	census
41	Sul Ross State University	228501			30.50989	-97.78177	census
42	Sul Ross State University	228501			29.56478	-104.36340	census
43	Sul Ross State University	228501			31.82239	-106.46536	census

Putting Rstudio to work: V

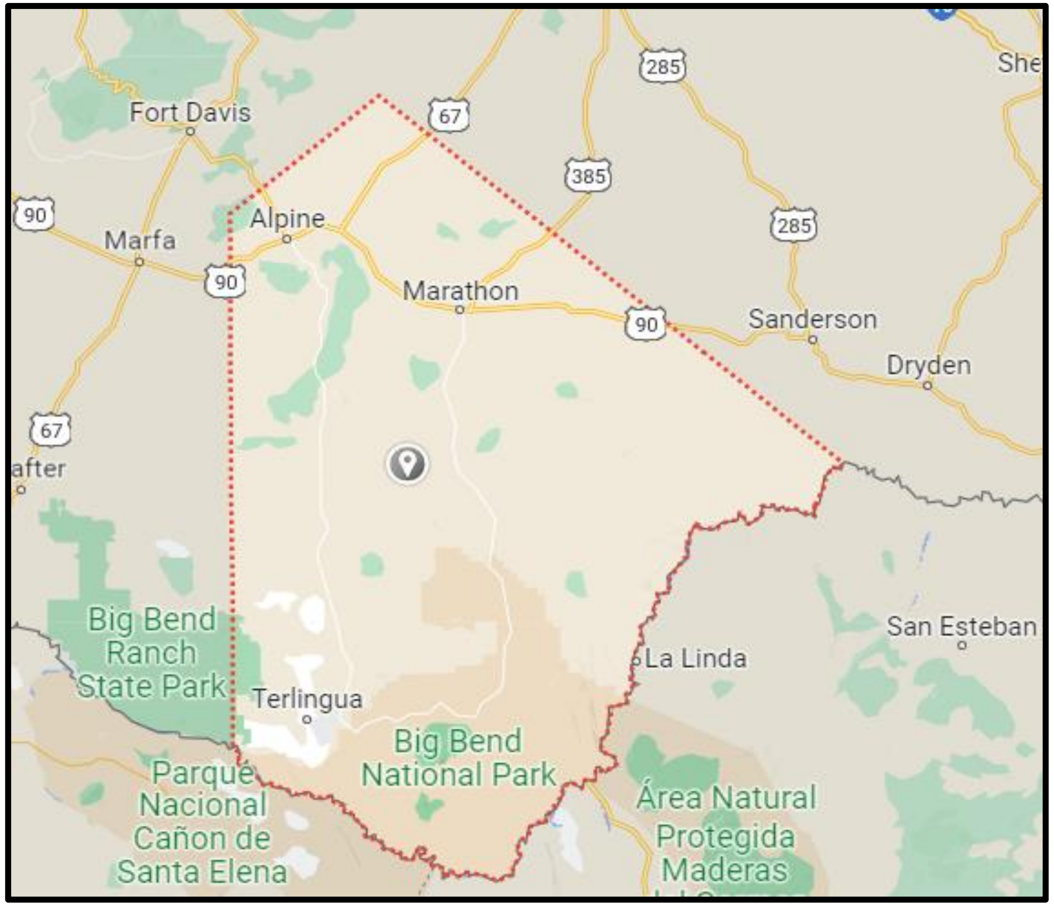
- Step 5 – Review data and make corrections. The code below identifies certain strings of letters in the Address dataframe and flags them for removal or separate analysis.

```
# this function creates a new column that is a flag identifying th
detect_address_types <- function(df, col_name) {
  df$address_type <- ""
  for (i in 1:nrow(df)) {
    row_text <- tolower(df[i, col_name])
    if (grepl("\\bpo box\\b", row_text)) {
      df[i, "address_type"] <- "PO BOX"
    } else if (grepl("\\bste\\b", row_text)) {
      df[i, "address_type"] <- "STE"
    } else if (grepl("\\bapt\\b", row_text)) {
      df[i, "address_type"] <- "APT"
    } else if (grepl("\\bunit\\b", row_text)) {
      df[i, "address_type"] <- "UNIT"
    } else if (grepl("\\bapartment\\b", row_text)) {
      df[i, "address_type"] <- "APARTMENT"
    } else if (grepl("\\bdepartment\\b", row_text)) {
      df[i, "address_type"] <- "DEPARTMENT"
    } else if (grepl("\\bdept\\b", row_text)) {
      df[i, "address_type"] <- "DEPT"
    }
  }
  return(df)
}


IPEDS_ID_2015 <-detect_address_types(IPEDS_ID_2015,"ADDRESS")
```

Putting Rstudio to work: VI

- Use Tidycensus and Tigris to query and pull census data files and census shp files into your Rstudio environment. Merge census data using sf functions:



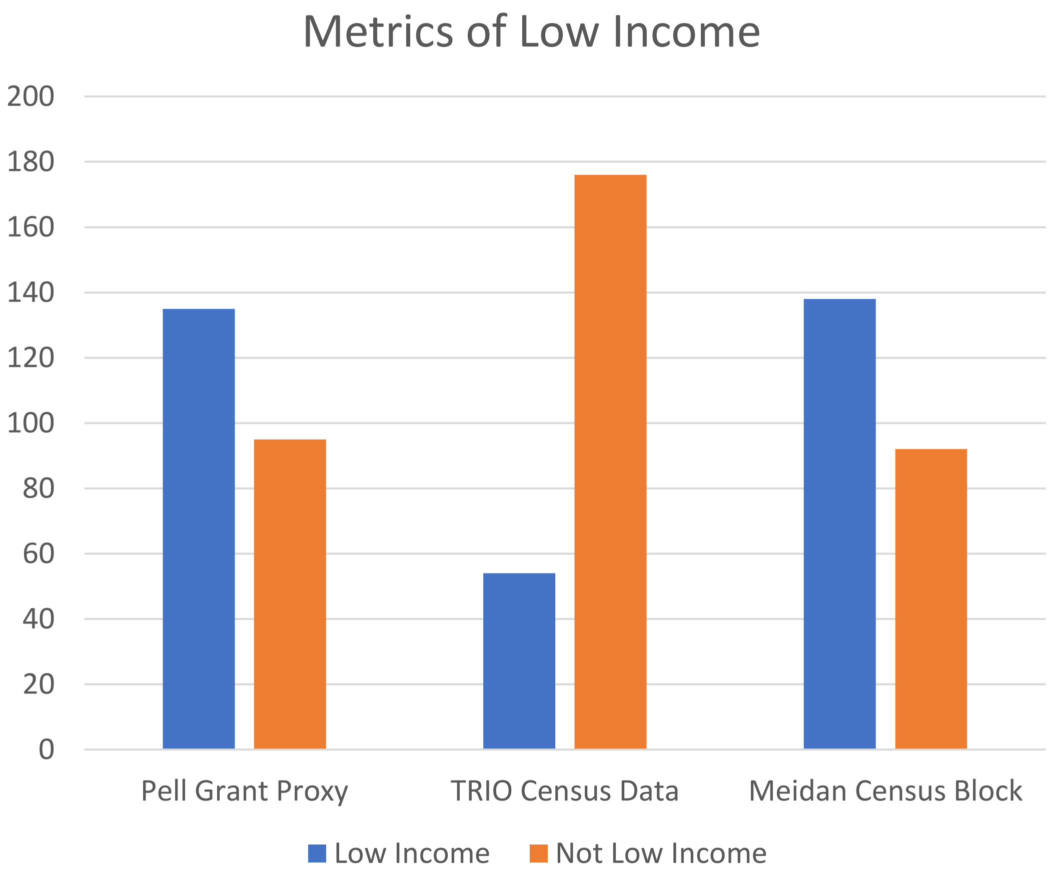
- st_transform – function in sf packaged used to convert a dataframe into a shp file. Also used to set CRS

- st_join – function in sf package used to join two separate shape file

- IE: If point X is contained within shape Y, execute merge

Low Income Proxy

- Situation: Pell grant recipient status is an imperfect proxy for low income status.

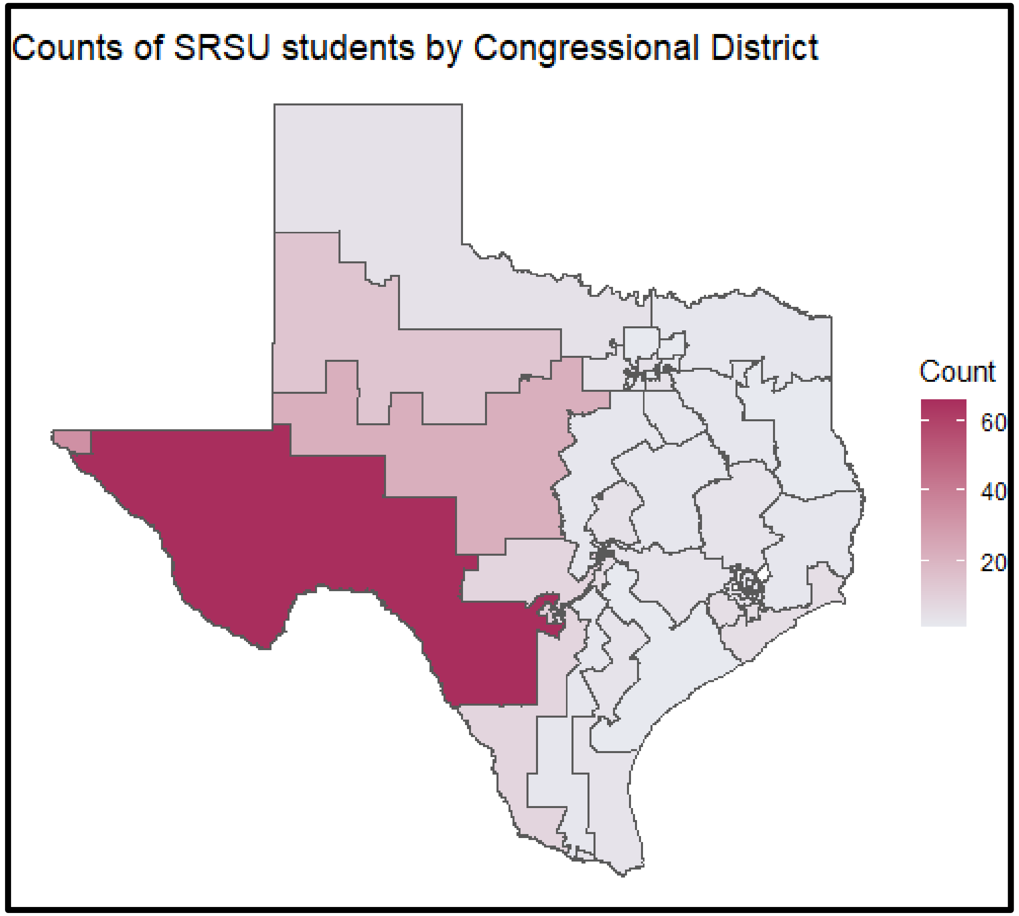


- Solution: You use geocoding procedures to create two separate alternatives to Pell Grant proxy

- The graph here shows a calculated metric of TRIO federal poverty status using B19013_001 (median house hold income) and B25010_001 (average household size)

The Irregular Data request

- Situation: An external data requester asks for a visualization of how many students reside in a given congressional district:



Congressional District	count
Congressional District 23	66
Congressional District 16	33
Congressional District 11	22
Congressional District 19	14
Congressional District 20	10
Congressional District 21	8
Congressional District 28	8
Other-In-State	73
Out-Of-State	6

Recruitment and Marketing

- Situation: Your institution's recruitment dept is collecting limited contact information from prospects who inquire about your institution's programs.

2000 Clusters

Cluster Numbers	Number of Block Groups	Median HH Income	% HH in Poverty	Unemploy-ment Rate	% Hispanic	% White	% Black	% Asi
1	502	$ 32,680.08	14.5%	8.1%	2.9%	94.2%	0.3%	0.7%
2	189	$ 27,998.97	27.0%	13.8%	17.6%	72.8%	1.9%	1.6%
3	474	$ 30,950.98	12.9%	6.5%	2.4%	91.5%	2.1%	2.5%
4	551	$ 44,489.93	7.9%	5.1%	2.3%	94.5%	0.6%	1.4%
5	695	$ 47,723.27	7.5%	5.1%	3.1%	88.1%	3.0%	4.4%
6	391	$ 30,567.18	20.7%	8.9%	4.0%	87.0%	3.6%	3.2%
7	69	$ 25,059.80	33.2%	15.4%	56.3%	38.0%	1.7%	0.7%
8	26	$ 25,762.40	33.7%	18.5%	3.5%	23.4%	0.1%	0.6%
9	73	$ 35,703.35	17.0%	7.9%	3.7%	37.7%	15.9%	40.6
10	66	$ 23,923.16	35.2%	14.3%	4.2%	24.5%	50.1%	18.7
11	651	$ 52,794.20	5.2%	3.9%	2.2%	93.1%	1.1%	2.7%
12	337	$ 78,325.60	2.8%	3.1%	1.7%	92.2%	0.8%	4.8%
13	453	$ 44,343.66	9.1%	4.1%	2.6%	85.2%	4.7%	6.5%
14	118	$ 23,528.77	26.4%	5.8%	4.0%	85.4%	3.5%	4.8%
15	25	$ 18,041.45	43.0%	7.5%	3.0%	85.5%	1.8%	9.1%

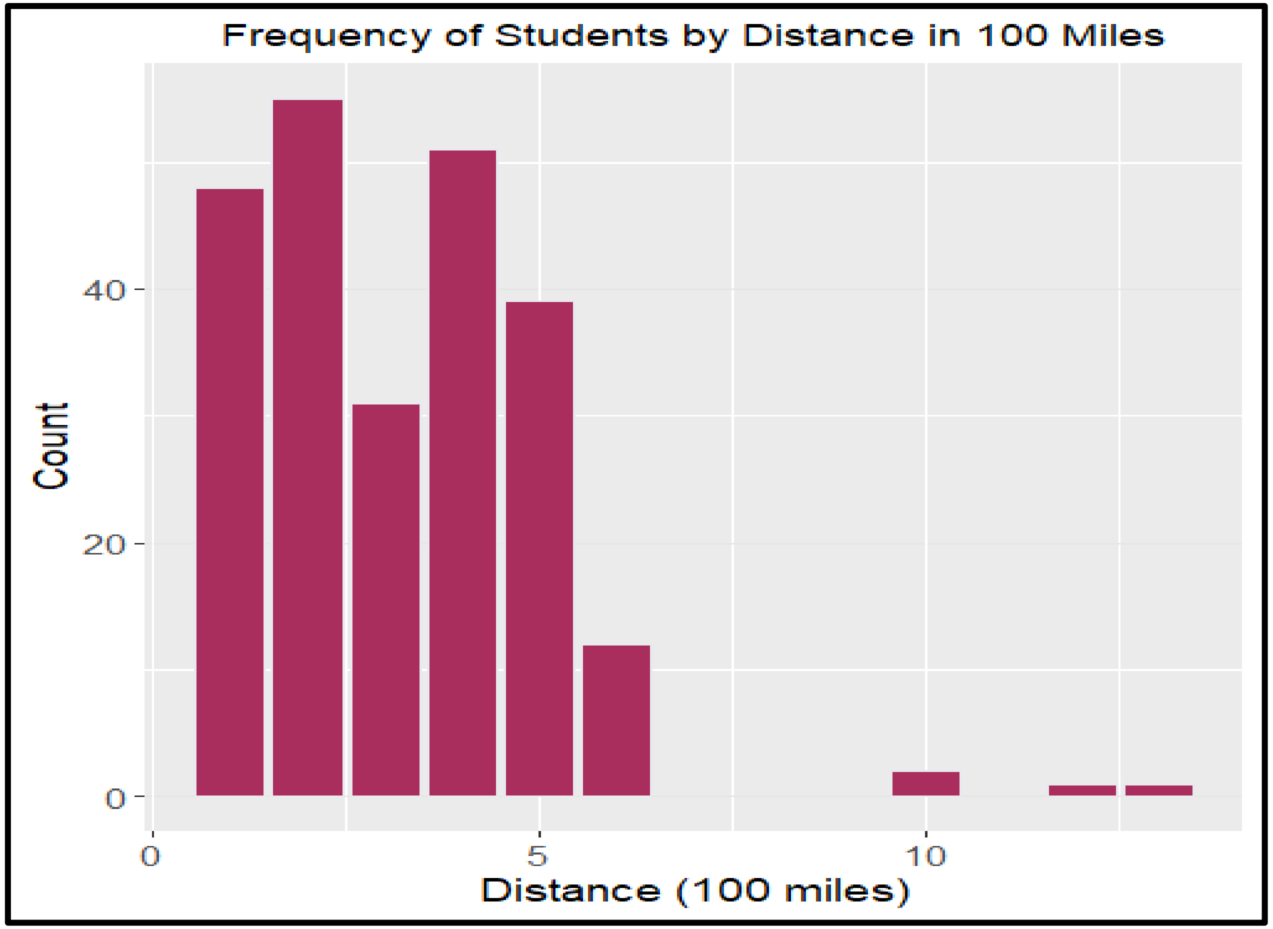
- You can geocode additional contextual information into your analysis and bin students into different groups based on shared census geography traits.

- Using this information, your recruiting or marketing dept may devise alternate marketing methodologies to different groups of students.

- Crosta, P. M., Leinbach, T., & Jenkins, D. (2006). Using Census Data to Classify Community College Students by Socioeconomic Status and Community Characteristics. CCRC Research Tools. Number 1. *Community College Research Center, Columbia University*.

Statistical Analyses

- Situation: Your department wants to conduct an 'At Risk' student analysis but you are missing the variable 'distance from university'.



Distance from Uni in 100 miles	Count
1	48
2	55
3	31
4	51
5	39
6	12
10	2
12	1
13	1

Alumni Affairs

- Situation: Your alumni department wants a metric of upward mobility.

- If your institution has address data from when a student first enrolled vs where they are X years later, you can contrast the economic characteristics of a given census geography against each other.
- Solution: If a student resided in a low income census geography and now resides in a higher income census geography, you may assume that upward mobility took place.

Anything You want

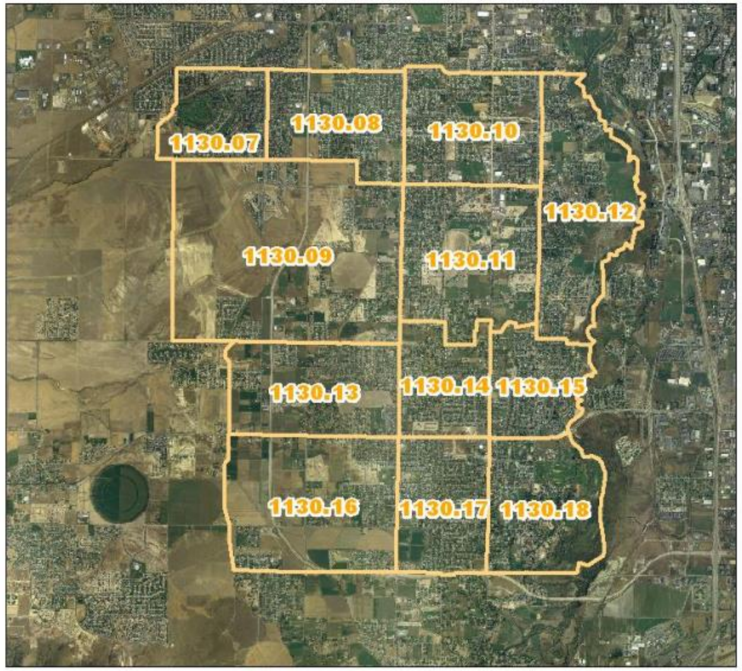
- “The limits of my [shape files] mark the limits of my world.” – ***Not*** Ludwig Wittgenstein, Tractatus logigo-philosphicus, 1922.

- FEMA flood maps: https://coast.noaa.gov/digitalcoast/data/flood.html
- EPA Super Fund sites: https://www.epa.gov/superfund/search-superfund-sites-where-you-live
- Criminal record proximity: https://www.hcdistrictclerk.com/Common/e-services/PublicDatasets.aspx *(data accessibility varies by county in TX)*
- Underserved community identification: https://www.ffiec.gov/cra/distressed.htm
- State Park boundaries: https://tpwd.texas.gov/gis/
- Others: https://catalog.data.gov/dataset/?tags=texas

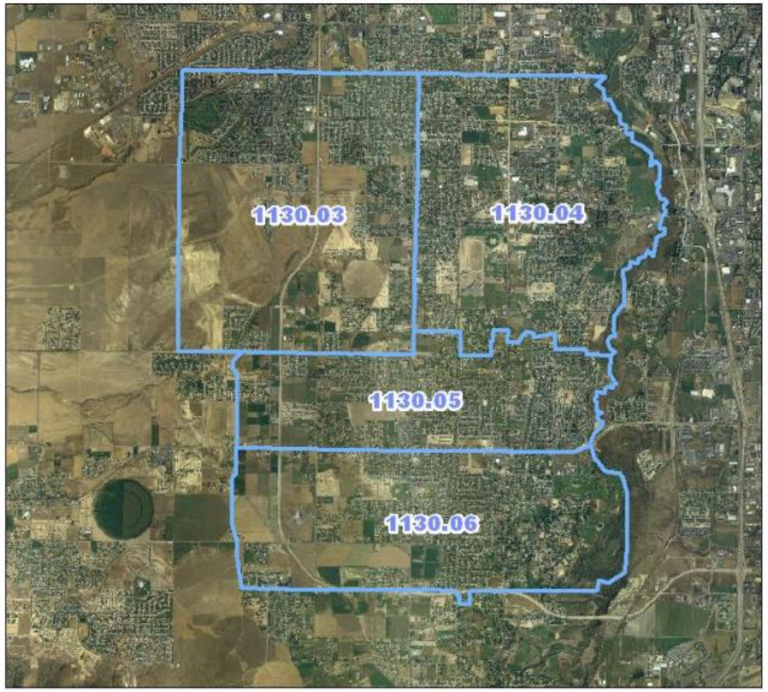
Additional Considerations

- Census Geographies may change with each instance of a census such that a geography bearing ID ‘XYZ’ from 2010 will not have the same shape in 2020.
- Solution: the census maintains a crosswalk of census files across census periods: https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html

2000 Census Tracts



1990 Census Tracts



Additional Considerations

- **Centroids:** centroids are the centre point of a shp file object.

- If you have a PO box address, you cannot use the census block group unit of analysis; however, you can calculate the center point of a city and merge census data at the municipal unit of analysis

- This allows you to include students in your analysis that may otherwise be excluded.