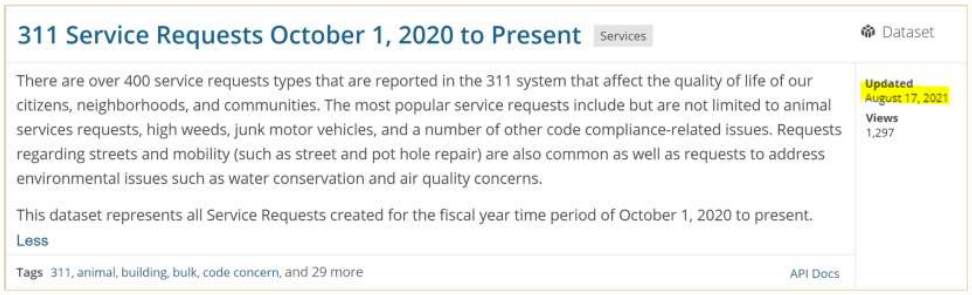
Lab Package 1.1: Doing Data Science in ArcGIS Pro

Monday, August 16, 2021 7:51 PM

In this package, we will practice some ArcGIS Pro tools for a simple data science project.
The research project's goal is to analyze the 311 calls in the City of Dallas and identify where, when and what services were requests during the data period (October 1, 2020 to present). Where are the census tracts with high and low calls? What kinds of calls? Any temporal pattern? Open data are now the common policy for many cities, including Dallas. The vast readily available data drive advances in data science.

Go to the Dallas open data site at https://www.dallasopendata.com/browse and find the 311 data (below). (yes, I could have just include the direct link to 311 data, but I think that searching for the data will give you an opportunity to learn what data are available at this site.



The data are updated daily, so the data change every day. For this exercise, create a folder in OneDrive first (with a folder name starting with A, such as ASDSPkg1) and please download the data here to the OneDrive folder

1. **When you are on the data download page, check out the meta data.**
   How many columns are in the data? _____.
   Which column records what has been done to the request? Note the column name here _____.
   What is the data type for Lat_Long Location? _____

Download the data in CSV file.

Data Engineering

a) Pay attention to the detail in every step
b) The lat_long location is in the format of (latitude, longitude). The comma will cause problems when loading the data into ArcGIS Pro because ArcGIS Pro (and many software tools) because we need to have latitude and longitude in separate columns as x and y to plot locations. The easiest way to do this is to open the file with Notepad, and use "find and replace" to find "(and)" and replace them with nothing (leave the replace blank). Then you also need to find Lat_Long Location in the end of the first line, and change it to Latitude, Longitude.
c) Add the data to ArcGIS Pro and display the locations with latitudes and longitudes, save the data to a feature class for easy access and analysis in ArcGIS Pro, and remove the standalone table csv file to save RAM space.
d) Inspect the data, identify issues, and correct the data: remove data with no lat/long and correct data with coding errors.

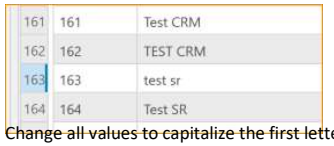
2. **How many service request types are there? _____**
   Which service type has the highest frequency of service calls? _____
   Is Bike share a significant problem in Dallas? Yes, if more than 500 service calls, or otherwise no _____
   (note that the answers will change as the data updated daily.)

d) Add a new field column, Service_Request_Type_Simple, with simplified service type by removing "- xxx"
Inspect the result statistics table, find potential duplicates, and correct the values (hint: street lighting).

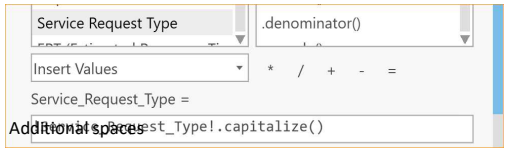
3. **How many simple service request types are there? _____**
   How many simple service request types are related to water? _____
   Among the top 10 most calls, how many types are related to sanitation? _____

e) The questions ask to map the calls in census tracts. However, more people in a census tract may lead to more service calls. Therefore, it is important to consider population when making the comparison of service calls across census tracts. Search census demonstration in the Living Atlas, and add Census Demonstration Products v5 April 2021 > Tracts v5 to the map, and clip out the tracts in the City of Dallas. Name the output feature class Tracts5DallasPopulation

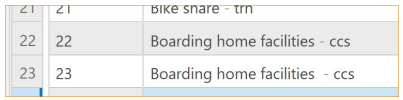
Need to identify inconsistent encoding in the data:

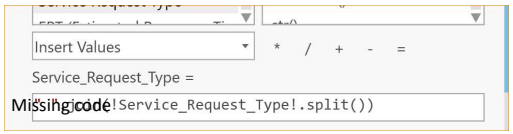


Change all values to capitalize the first letter of the first word

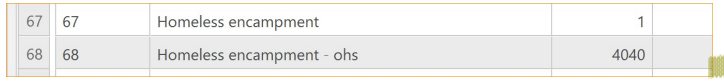


Additional spaces





Missing code



Calculate Service_Request_Type_Simple

e) The questions ask to map the calls in census tracts. However, more people in a census tract may lead to more service calls. Therefore, it is important to consider population when making the comparison of service calls across census tracts. Search census demonstration in the Living Atlas, and add Census Demonstration Products v5 April 2021 > Tracts v5 to the map, and clip out the tracts in the City of Dallas. Name the output feature class Tractv5DallasPopulation

f) Cognitively, we need to limit types to 10 or less. More categories can be difficult to perceive any patterns. Look at the top 10 most frequent service request types. Add a new field column, Top_Service_Request_Type, in the statistics table and re-encode all the service types to the following. Use "Top Service Request Types.xlsx" here reclassify simple service request types to top service request types.
   a. Code concern or violation
   b. Sanitation issues
   c. Water or wastewater issues
   d. Traffic issues
   e. Parking issues
   f. Animal issues
   g. Noise
   h. Others

g) After you added the Top Service Request Types in the statistics table, the table is now a "look-up" table. The table can be joined with the Service Calls, so that every record on the Service Calls has the Top Service Request Type. After joining the tables, we need to make the result permanent. Export it to a feature class (e.g., TopCategoryServices), and use the new feature class for the rest of the analysis.

h) There are too many attributes in the feature class. Remove attributes that we are not going to use in the project can speed up analysis. Use delete field to keep only the following attributes: Created Date, Top_Service_Request_Type

i) Recall our research questions:
Where are the census tracts with high and low calls? What kinds of calls? Any temporal patterns?

First, we will convert the field column "Created Date" from Text data type to a Time field, using Data Engineering > Format > Convert Time Field.

A little visualization here to look for temporal patterns.

Use the time clock and calendar heat chart to show temporal cycles

4. Which month in the data showed the highest number of service calls? _____
Which day in June had the lowest number of service calls? _____
Which month had the least numbers of calls throughout the entire month? _____

Back to Data Engineering

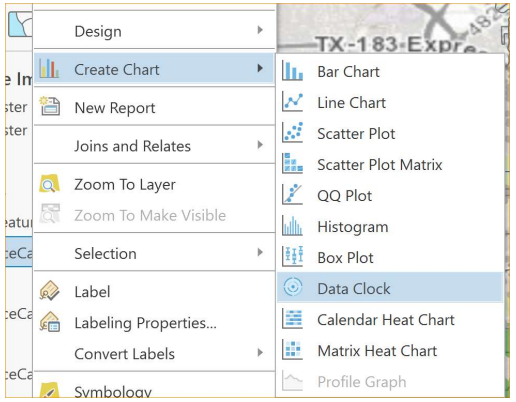
Now we prepare the data for analysis based on census tracts. Add the census tracts with census 2020 population. Download the Dallas City boundary shapefile from Dallas Open Data at https://gis.dallascityhall.com/shapefileDownload.aspx to clip the census tracts in the city (e.g. DallasTracts). Use SF1 Total Population for the analysis, so only keep geographic identification code and SF1 population in DallasTracts to speed up the analysis. (Note that there are suites of demographic data that can help answer many other questions beyond the project).

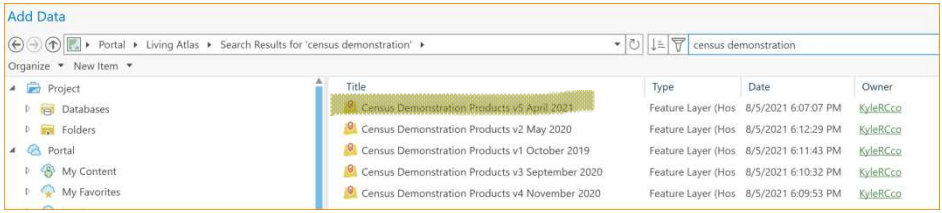
For spatial distribution of service calls, think about how to arrange data in a table for the analysis.

	Variable1	Variable2	...	...	...
Unit of analysis	value	value	...	...	...
Unit of analysis	value	value	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

5. Challenge (required for graduate students;10 bonus points for undergraduate students who come up with a working solution): Create a table below for analysis.

Calculate Service_Request_Type_Simple

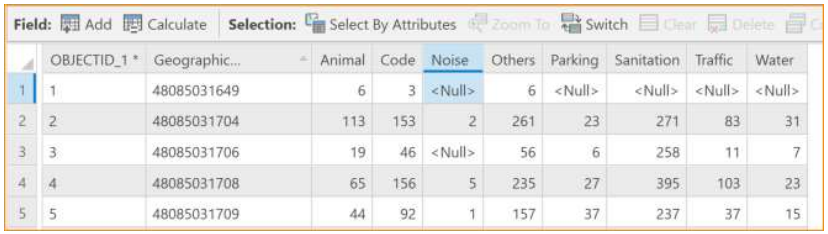




The table presents census tracts with service calls. The Geographic Identification Code is the census tract identifier in the census TIGER file.

Note that if the first Geographic Identification Code is <Null>. Look at the Service Calls table. There are many entries with <Null> in address, latitude, and longitude. These calls are without locations. However, there are some calls with addresses and latitudes/longitudes but without GEOID. Plot these points on the map to see why and find solutions to locate these points to corresponding census tracts.

Raise your hand or Send me a text in the class TEAM channel if you find the solution. Enter the sequence of tools in your solution (e.g., clip to the city of Dallas boundary, sum the total population over tracts, …) You can only get the credits in class.


Field: Add Calculate Selection: Select By Attributes Zoom To Switch Clear Delete

	OBJECTID_1 *	Geographic...	Animal	Code	Noise	Others	Parking	Sanitation	Traffic	Water
1	1	48085031649	6	3	<Null>	6	<Null>	<Null>	<Null>	<Null>
2	2	48085031704	113	153	2	261	23	271	83	31
3	3	48085031706	19	46	<Null>	56	6	258	11	7
4	4	48085031708	65	156	5	235	27	395	103	23
5	5	48085031709	44	92	1	157	37	237	37	15

Next, use find and replace to replace <Null> with 0, delete the total population (already in the DallasTracts), save the table, and Join the table to DallasTracts, save to a new feature class, named TractCallPop

Spatial Analysis and Machine Learning

1. Do an optimized hot spot analysis on the service calls. We must project the data to a Cartesian coordinate system first before the hot spot analysis because the hot spot analysis uses Euclidean distance in the calculation. Use Data Management Tools > Project to reproject to NAD_1983_StatePlane_Texas_North_Central_FIPS_4202_Feet (or use the City Limit layer as the project reference). Also project TractCallPop.

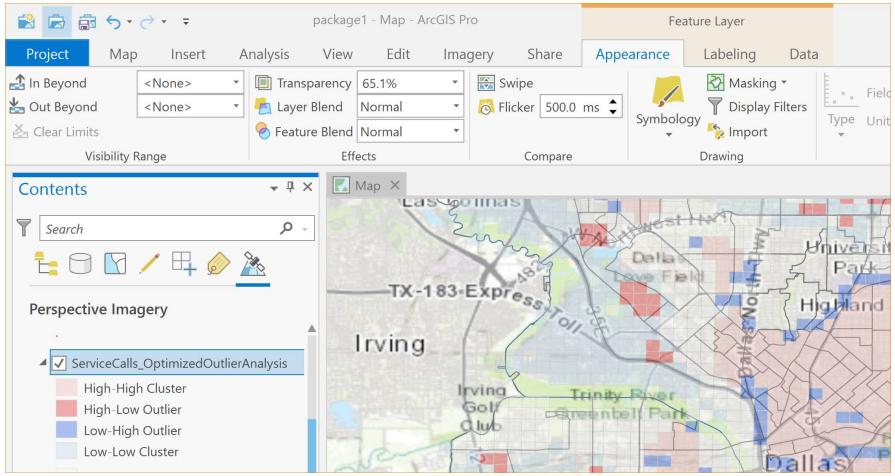
   Perform Optimized Hot Spot Analysis, Optimized Outlier Analysis, and Density-based clustering with DBSCAN and 2000 minimum features per cluster. Set layers transparent one at a time, so that we can geographically contextualize the hot spots, cold spots and outliers.

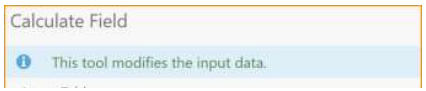
6. **What is the use of fishnet grid in Optimized Hot Spot Analysis and Optimized Outlier Analysis? _____ (a, b, or c) (a: aggregate service calls, b: assign colors for hot or cold spots, c: rasterization for spatial autocorrelation)**
   Is Fair Park in downtown Dallas a hot spot or cold spot of service calls? _____
   Optimized Hot Spots are statistically based, but Density-based clustering is a learning-based method. Based on the analysis, is the following statement true or false? Both methods catch the major hotspots but the Optimized Hot Spot method identified much more confined smaller hot spots than DBSCAN. _____ (T or F)

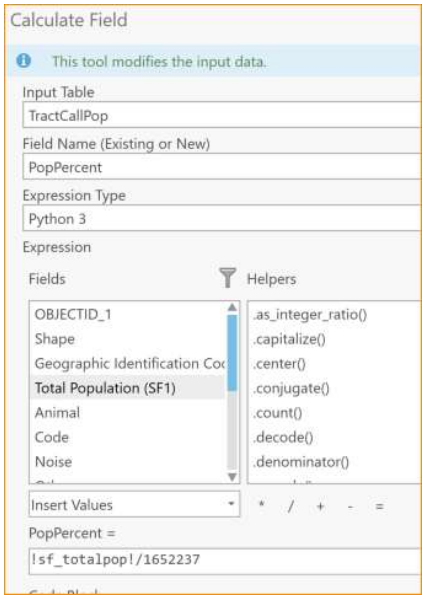
k) To make comparisons among service calls at census tracts, we need to consider population in the tracts, assuming that more people are like to have more calls for services.

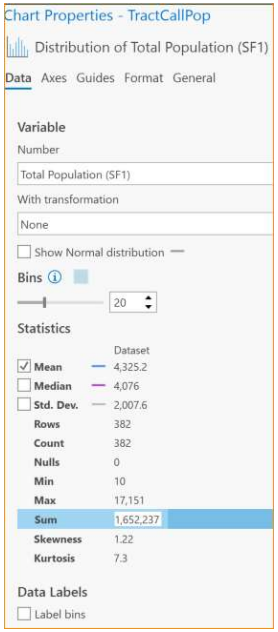
   Before continuing, check the TractCallPop attribute table. Notice that there are many tracts without any service calls. Delete these tracts first. The best way to do this is to sort the top category fields, select, delete, and save.

   Therefore, create a table that includes the following: census tract, geographical identification code, population proportion in the tract against all population in the city, and proportion of each top-service calls in the city. Better to (1) add and set up all new fields make sure that the data type is float, and save and (2) keep both calculate field and chart properties windows open and replace the attribute names for computing through the top categories.

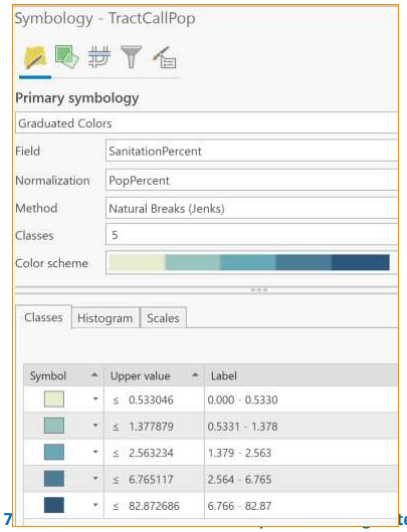


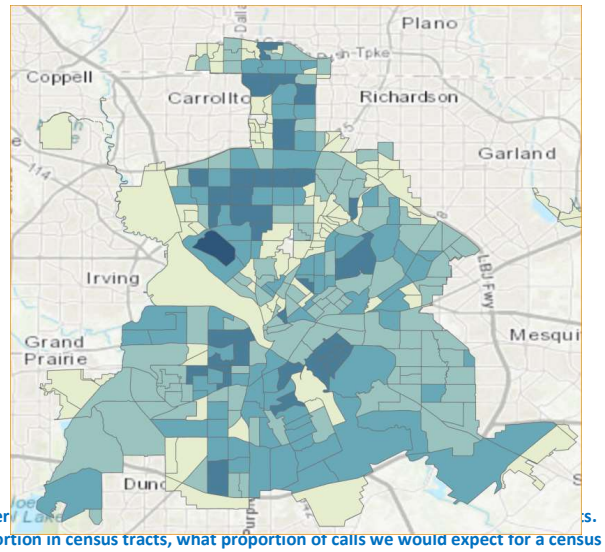






l) So that we can map percentage of a top service-call type normalized by percentage of population in a census tract.



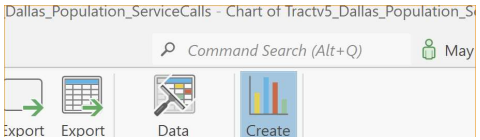


7 ter s.

If the calls are in proportion to population proportion in census tracts, what proportion of calls we would expect for a census tract with 5% of total population in the City of Dallas? _____

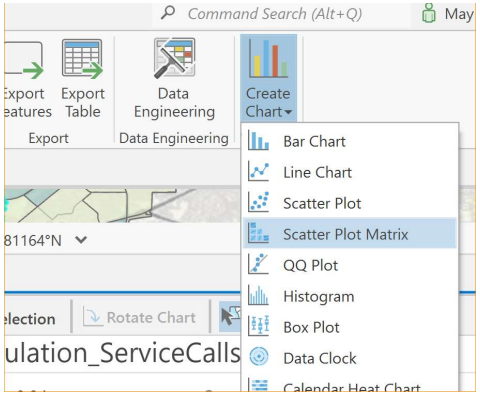
Where were the highest calls for sanitation services in the City of Dallas? Adjust the transparency to see the geographic context. Don't give Geographic Identification Identifier, but name the geographic feature at that location _____

Check out all the service calls, which service call appears relatively confined to central and NW Dallas _____

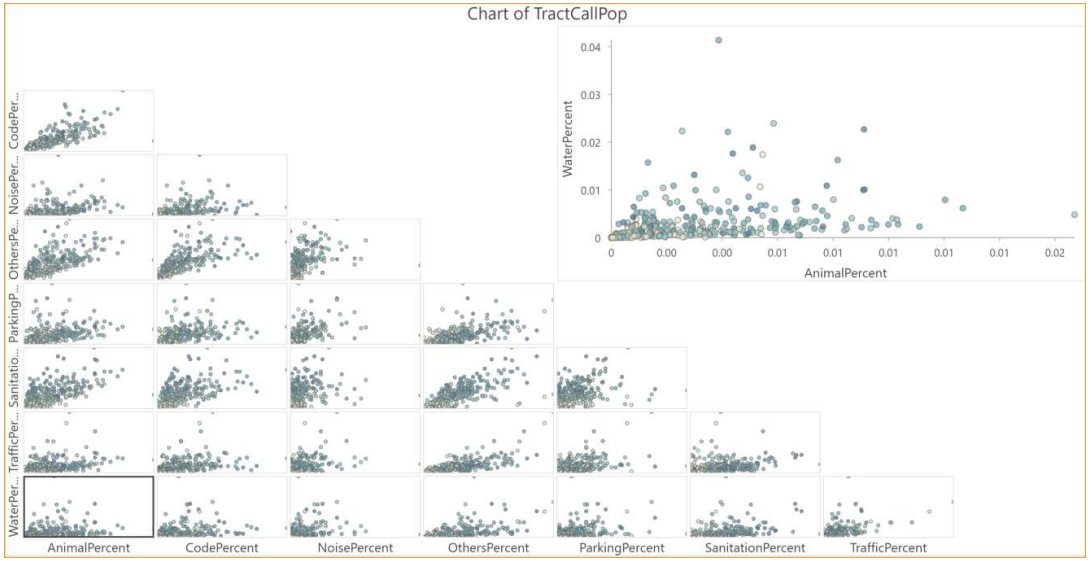


Note that the default classes are based on natural breaks in each service request type, so a darker shade in one may be equivalent to a light

Where were the highest calls for sanitation services in the City of Dallas? Adjust the transparency to see the geographic context. Don't give Geographic Identification Identifier, but name the geographic feature at that location _____
Check out all the service calls, which service call appears relatively confined to central and NW Dallas _____

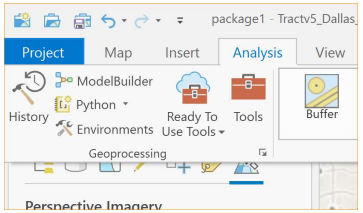


Note that the default classes are based on natural breaks in each service request type, so a darker shade in one may be equivalent to a light shade in another service type. Create Scatter Plot Matrix to see how these service calls relate.



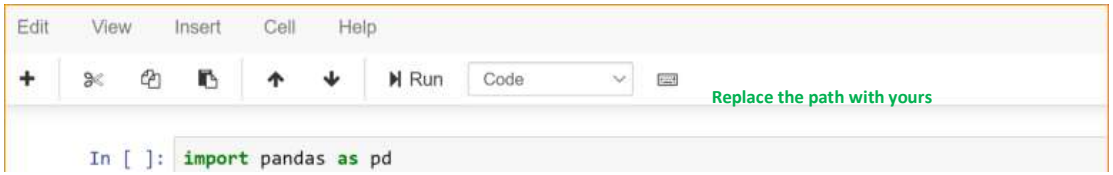
8. **From the scatter plot matrix, which pairs of service request calls are more linearly correlated? a) Water and Animal, b) Code and Traffic, c) Animal and Code? _____(a, b, or c)**
The scatter plot matrix suggests that noise, compared to other types of service calls, are relatively infrequent in most census tracts. _____(T or F)
The scatter plot matrix suggest that the frequency of sanitation calls varies widely across census tracts, but relatively few census tracts have many traffic calls. _____(T or F)

Now we want to find out what type of issue people called for service requests most in each census tract. We need to build a table for the analysis. The table should be something like:



Geographic Identification Code for Tracts	Type of Service calls has the max value
48113014132	Animal
.....	....

m) While we can do this in ArcGIS Pro and Arcpy, but it is too cumbersome and not elegant. Instead, we will do the following
- export the attribute table to an excel file
- Open the excel file and remove all fields but save the following: id, Animal, Code, Noise, Others, Parking, Sanitation, Traffic, and Water
- Open the python notebook in ArcGIS Pro
- Type the following codes in the notebook



Edit View Insert Cell Help

Run Code

Replace the path with yours

```
In [ ]: import pandas as pd

In [ ]: df = pd.read_excel(r'D:\SDSprojects\package1_1\TractCallPop_TableToExcel.xlsx')

In [ ]: df.columns

In [ ]: df.iloc[:, 2:]

In [ ]: df['MaxType'] = df.iloc[:, 2:].idxmax(axis=1)

In [ ]: df['MaxType']

In [ ]: df.to_csv(r'D:\SDSprojects\package1_1\MaxType.csv')
```

Replace the path with yours

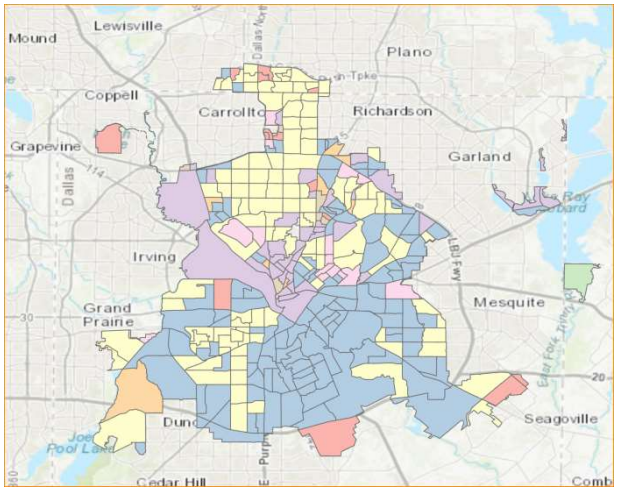
Type equation here.

- Opem the csv file and delete all fields but id and MaxType. Open the output csv file in ArcGIS Pro, We will join it with the Tractv5 _Dallas_Population_ServiceCalls feature class. However, we need to first export the csv file (df.csv) to a geodatabase table (DF). Now join the DF table to the Tract feature class. Click on "Validate Join" to see if any issues.

9. **Why did the join fail? a) Cannot join fields with different names, b) Cannot join fields with different data types, c) Cannot join a feature class with a table. _____ (a, b, or c)**

   How to fix the error? a) Change "id" in DF to "Geographic Identification Code", b) Change "id" to the same data type as "Geographic Identification Code", c) Add a new field "geoid" with the same data type as "Geographic Identification Code" and copy the values from id. _____(a, b, or c)

- After successfully join the tables, map the MaxType. Your map should look something like below:



10. **What was the most frequent types of service requests from most of south Dallas? _____**
    What was the most frequent types of service requests from most of north Dallas? _____

11. **The map shows that census tracts with parking issues as the most frequent types of service requests are relatively much smaller (except for one) than other census tracts. What may be the potential explanation?**

12. **Does the spatial pattern surprise you? Why or why not? What spatial questions does the map prompt you?**