2018

Visualizing Crime inter-Alberta and within Calgary

CPSC 599 – DATA VISUALIZATION

DEBBIE TEE MACROHON

Table of Contents

Introduction

I chose 3 different data sets, the first one outlining the different crimes in Alberta, and between cities in Alberta such as Calgary, Edmonton and Medicine Hat, from 1998 to around 2014. The second data set I chose is detailing the number of categorized crimes by community in Calgary from 2012 to 2016. The third data set is a shape file outlining the boundaries of each community, along with when structure details of it. I initially chose the second data set because I wanted to know if Forest Lawn, the community that I live in, was safe. However, this data set was not rich enough because it had a limited number of categories and spanned across only a few years. I decided to integrate the former data set because there are more datasets available for provinces, and not the varies communities in the cities that reside in that province. I hope to understand the trend of crime over the years for Alberta, Calgary, and various communities in Calgary and see if Calgary is becoming a more dangerous place.

Data Sets

Here are the following data sets that were selected.

Alberta Crime Data Set

The first data set was chosen from this website:

https://open.canada.ca/data/en/dataset/5d203dae-26d2-4f21-85de-c69e33d9f254

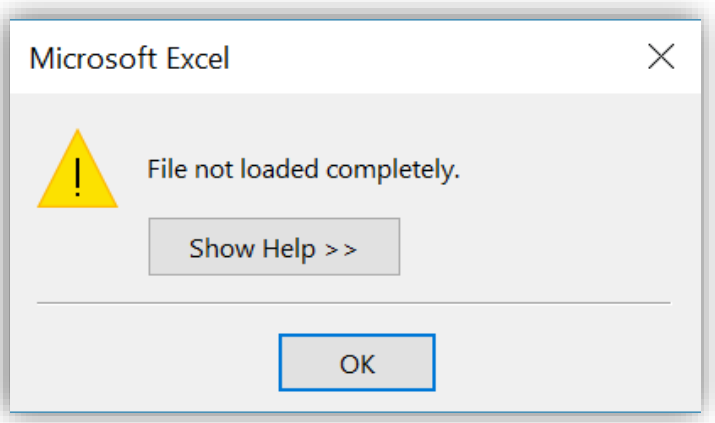
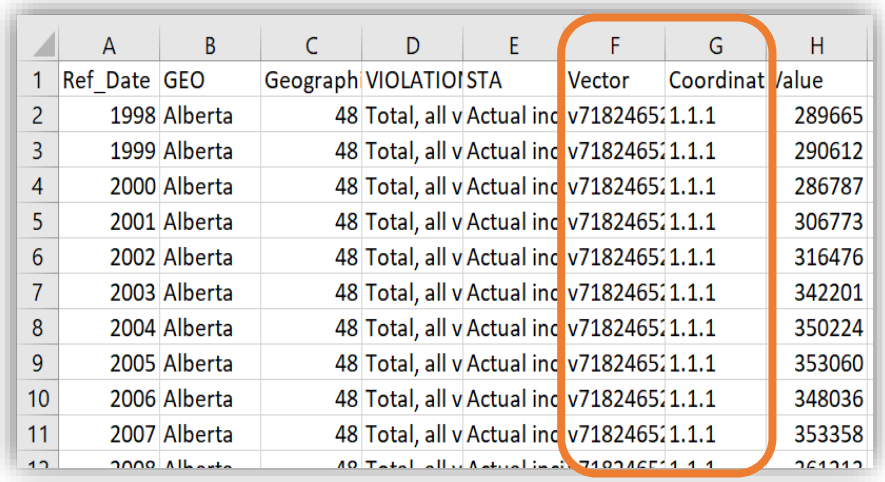


Figure 1 - File was too big to load into Excel



	A	B	C	D	E	F	G	H
1	Ref_Date	GEO	Geograph	VIOLATIOI	STA	Vector	Coordinat	Value
2	1998	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	289665
3	1999	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	290612
4	2000	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	286787
5	2001	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	306773
6	2002	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	316476
7	2003	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	342201
8	2004	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	350224
9	2005	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	353060
10	2006	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	348036
11	2007	Alberta	48	Total, all v	Actual inc	v7182465	1.1.1	353358

Figure 1 - removing unnecessary columns

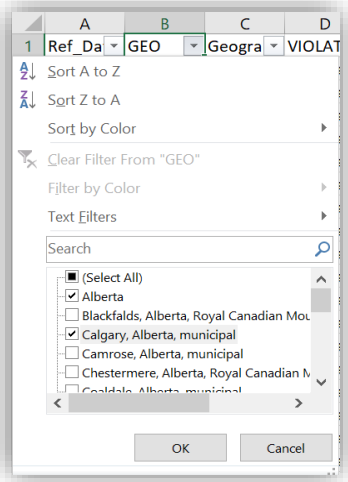
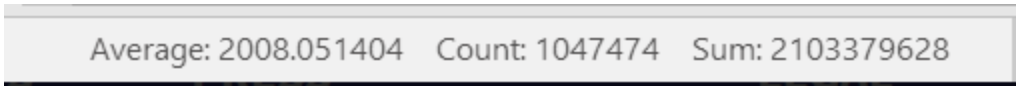


Figure 3 - Filtering out unnecessary locations

I opened the file, however it was not loaded completely because Excel can only handle around 2 million rows. The columns include the year, geography, type of violation, STA, and value, and includes data of Alberta as well as its municipalities. I also removed the Vector and Coordinate columns to save on space. I filtered the GEO column by Alberta, Medicine Hat, Calgary and Edmonton. We end up with around a million rows, and 5 columns:



Cleansing, Augmenting, Profiling the Data

I transformed the data into two different data sets, one pivoted by violation and the other pivoted by the column STA. I do this to calculate the correlations between violations or STA, which will allow for multiple views of the same data.

By Violation

Using Trifacta Wrangler, I pivoted the Violations Column so I could later do analysis on the different correlations between each type of crime.

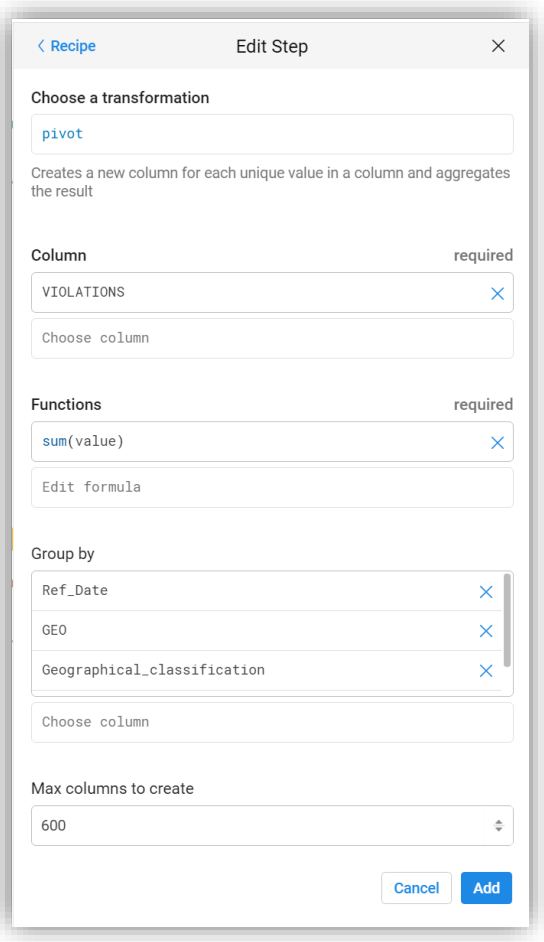


Figure 4 - Pivoting the Violations column. Summing the value column is fine because it is grouped by the year, location and geographical classification.

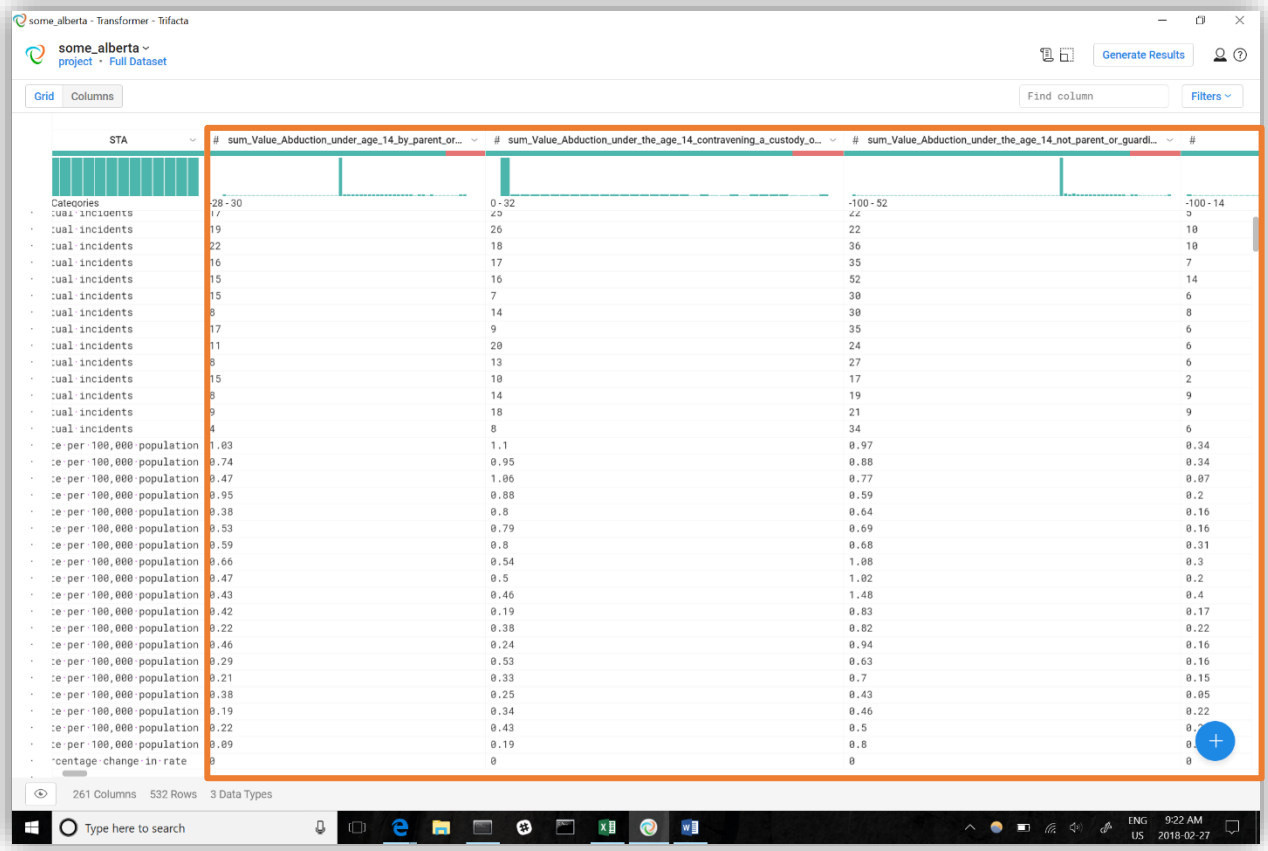
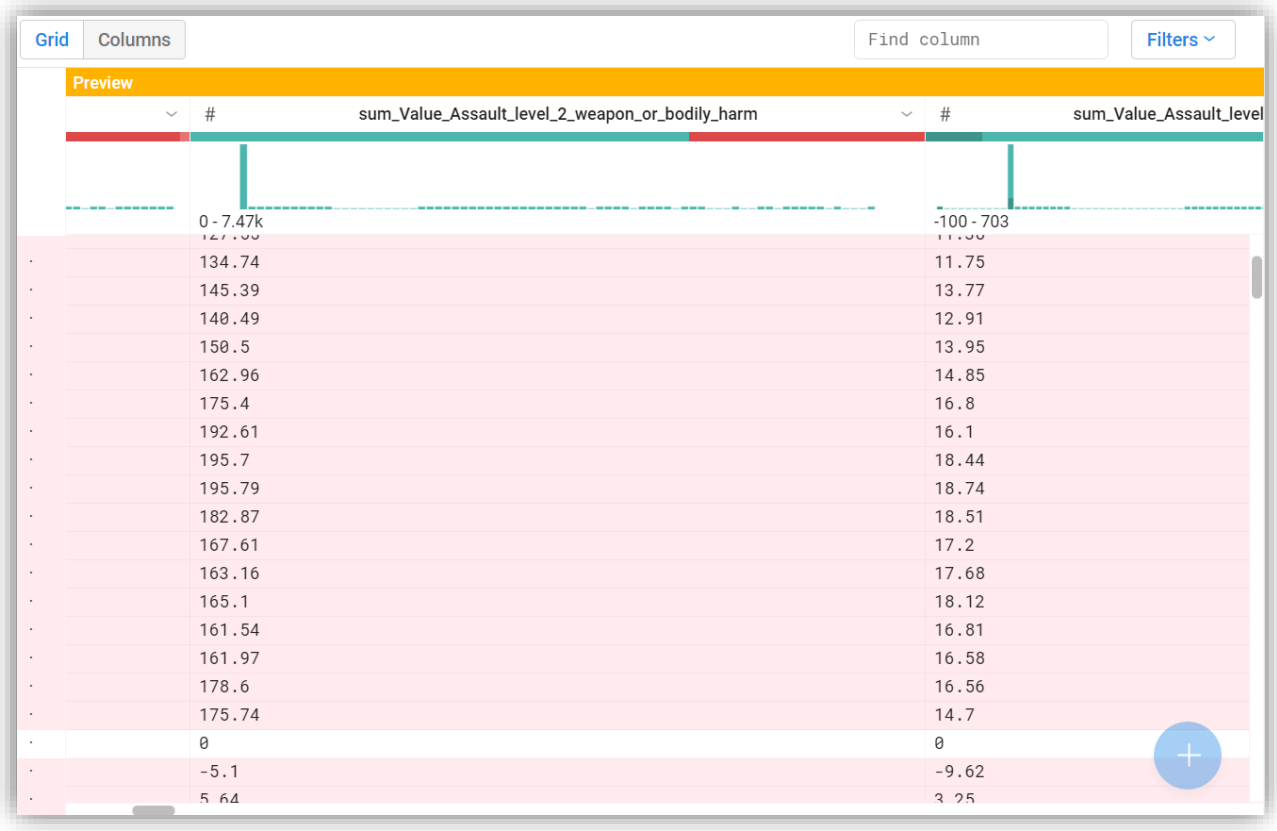
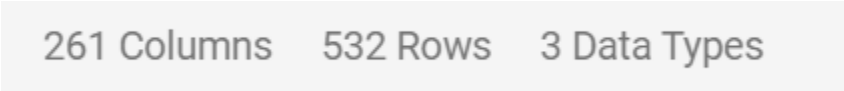


Figure 5 - new columns that have generated from the pivot

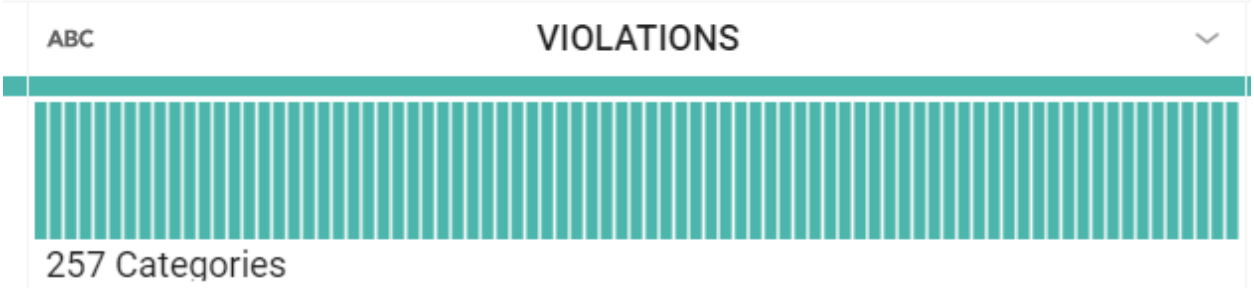
I didn't run into any problems here. It was straightforward and the only challenge was learning how the pivot worked. When profiling the data:



There were many mismatched values but this is because the column is expecting integers, not float values. The total row count came to 532 as shown below:

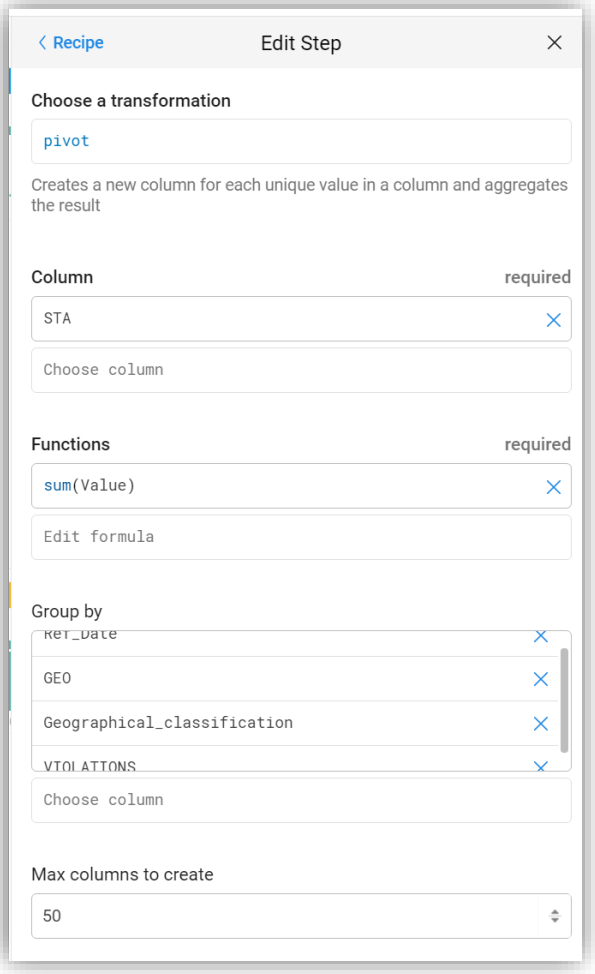


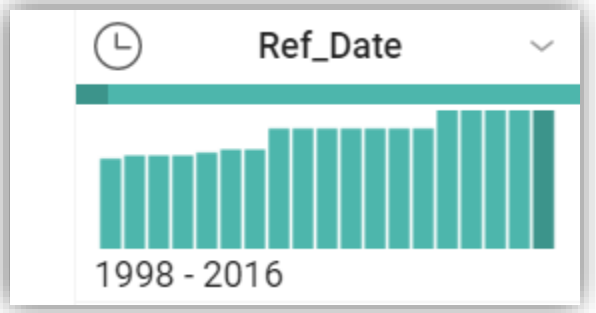
The number of violations were 257:



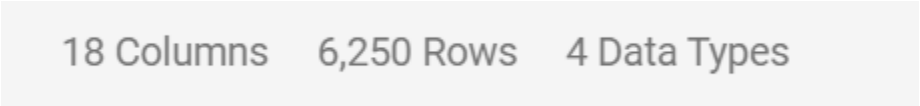
This initially didn't make sense because 87,618 rows /257 categories= 340, not 532. But analyzing the data again, I found that some crimes did not exist in later years. Thus, it's not an equal distribution of crimes for each year, since for example 1999 may have just 150 categories, compared to 2015 where we have 250. This means we can't evenly divide the data, but the statistic makes sense.

By STA

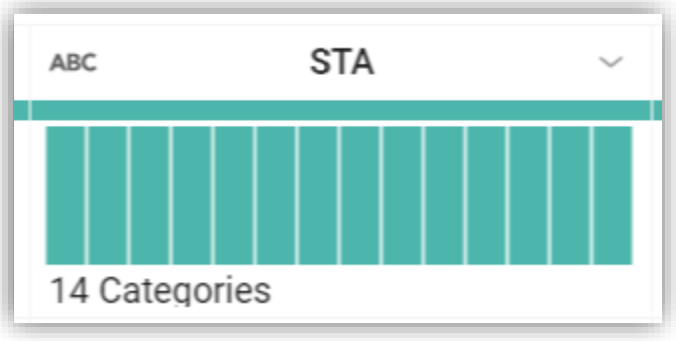




I thought this statistic was strange at first, but it turns out that they add more types of crimes as years go by.



The number of rows make sense, since there are 14 possible values in STA:



87,618 rows/ 14 categories = ~6250 rows

Calgary Communities Crime Data Set

The original file contains the category name, community, months and years, and was taken from this website:

http://www.calgary.ca/cps/Documents/statistical-reports/2017%20Community%20Crime%20Statistics.xls

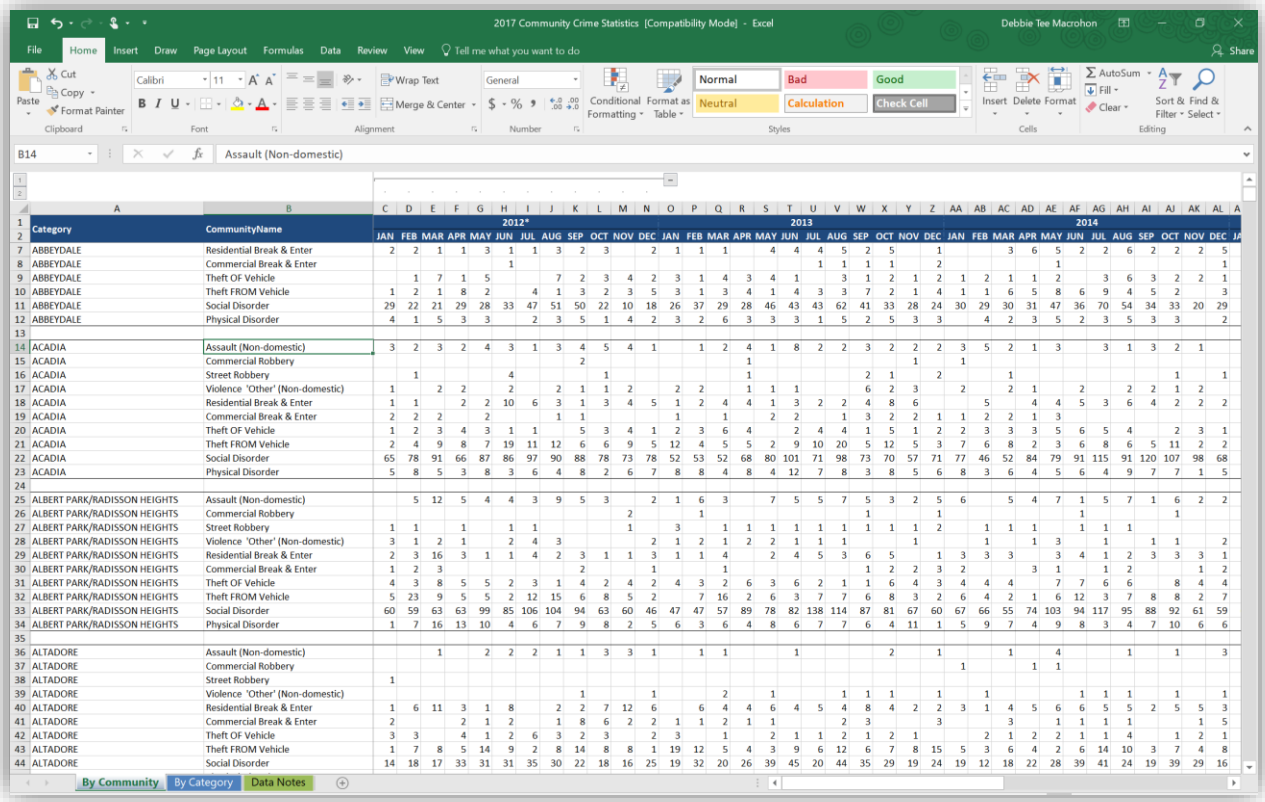


Figure 6 - Calgary community crime data set

Cleansing, Augmenting, Profiling the Data

I initially had trouble figuring out how to wrangle the data, because I had to unpivot 5 large columns(2012 – 2017) and then further unpivot it for each month from January to December.

For Step 2 - I merged all the date columns into one. Each year has 12 months, and since the data spans from 2012 to 2017 we have 5x12 = 72 columns. When unpivoted, the values ranged from column4 to column75.

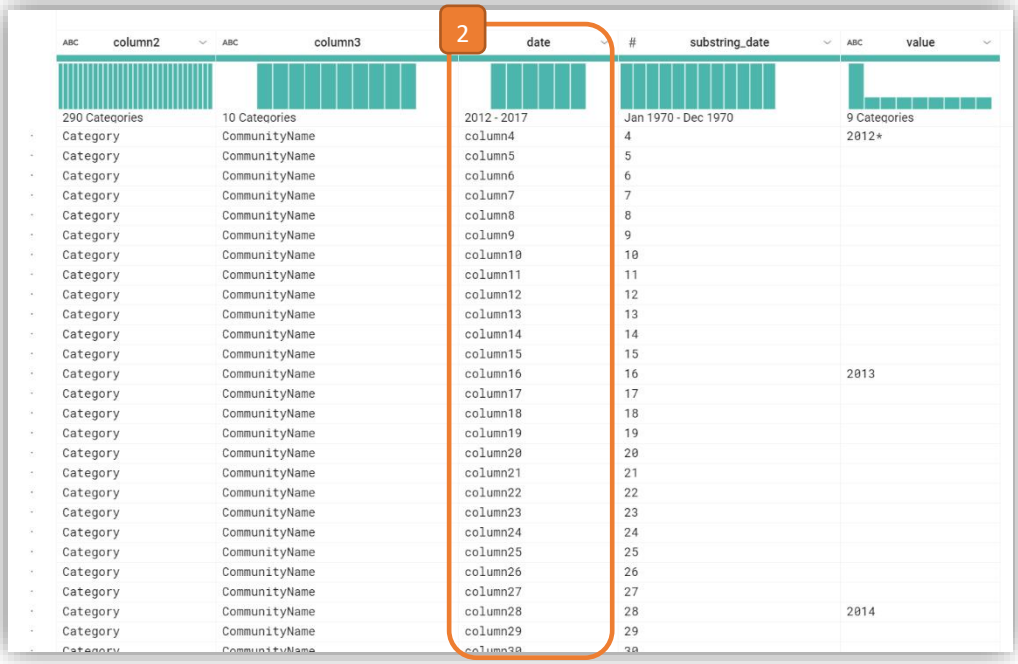


Figure 8- Step 2 of the recipe, date column is the result of unpivoting columns 4 to 75

Step 5 - I decided to manually code it such that it would take the integer value from the newly made column1 (which I eventually renamed to date) created from unpivoting the columns into rows, and then take the mod value such that if an integer mod 12 is equal to a value, it would assign values from 'JAN' to 'DEC'.

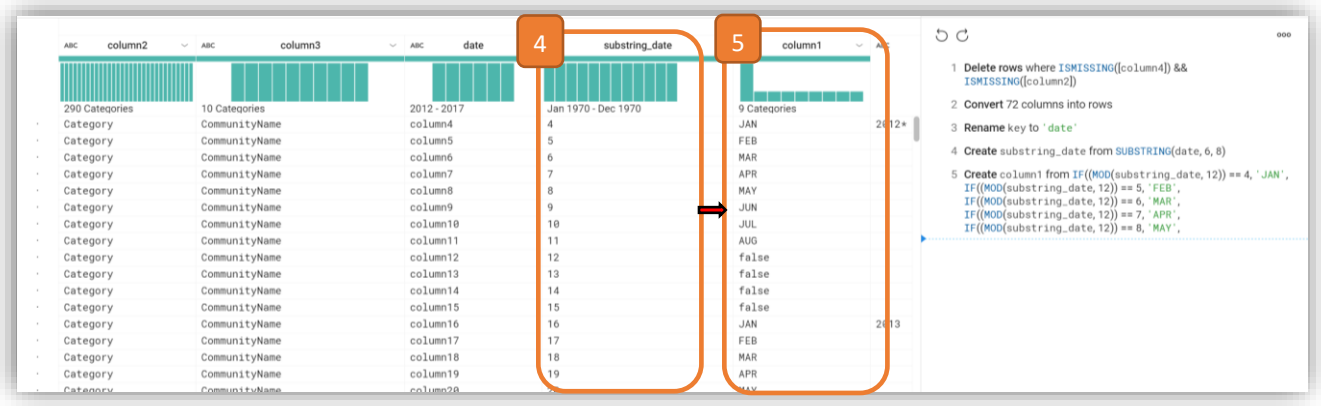


Figure 9 - Steps 4 and 5 from the recipe on the righthand panel

Step 6 - I made another column and set ranges for each year, for example for an integer mod 12, if the value is at a certain range then it will assign values from 2012 to 2017.

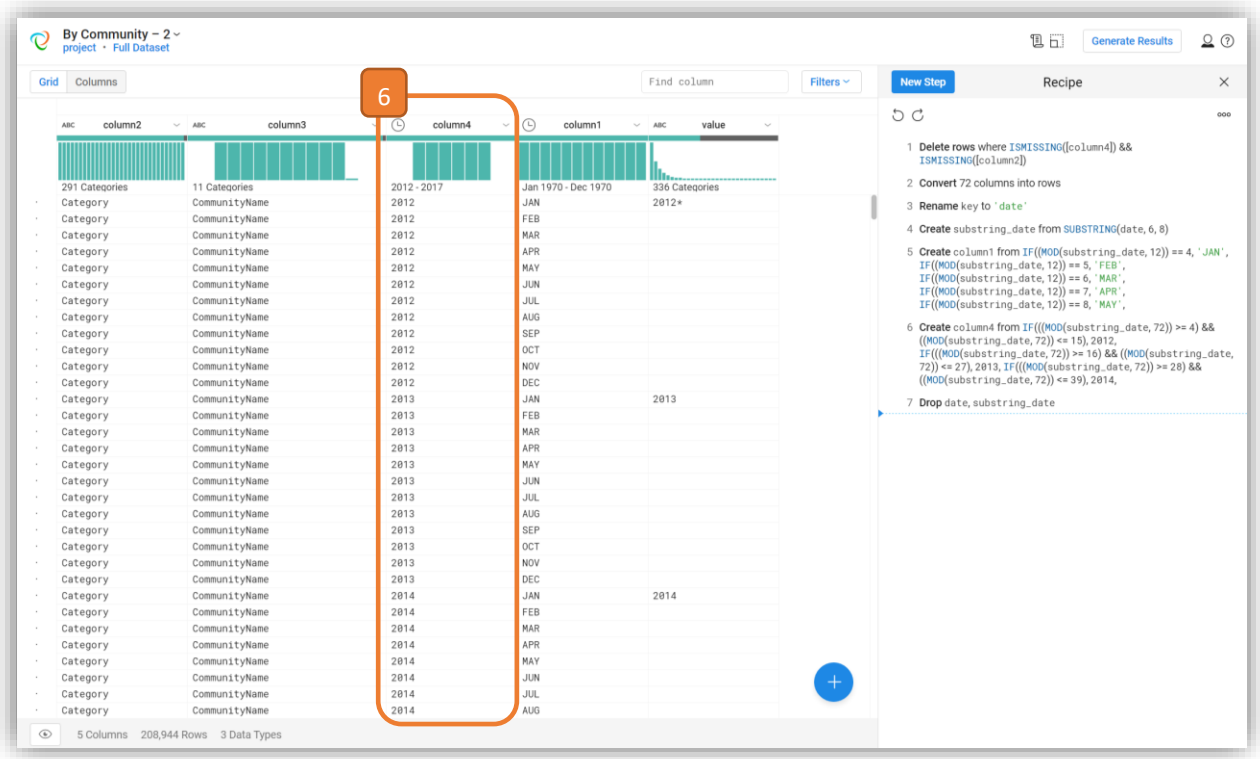


Figure 10 - Step 6 - based on a range, assign a year

Step 8 - I also had to delete some rows in the beginning because it was pivoted against essentially two nested columns being unpivoted.

After all those steps, this is the result:

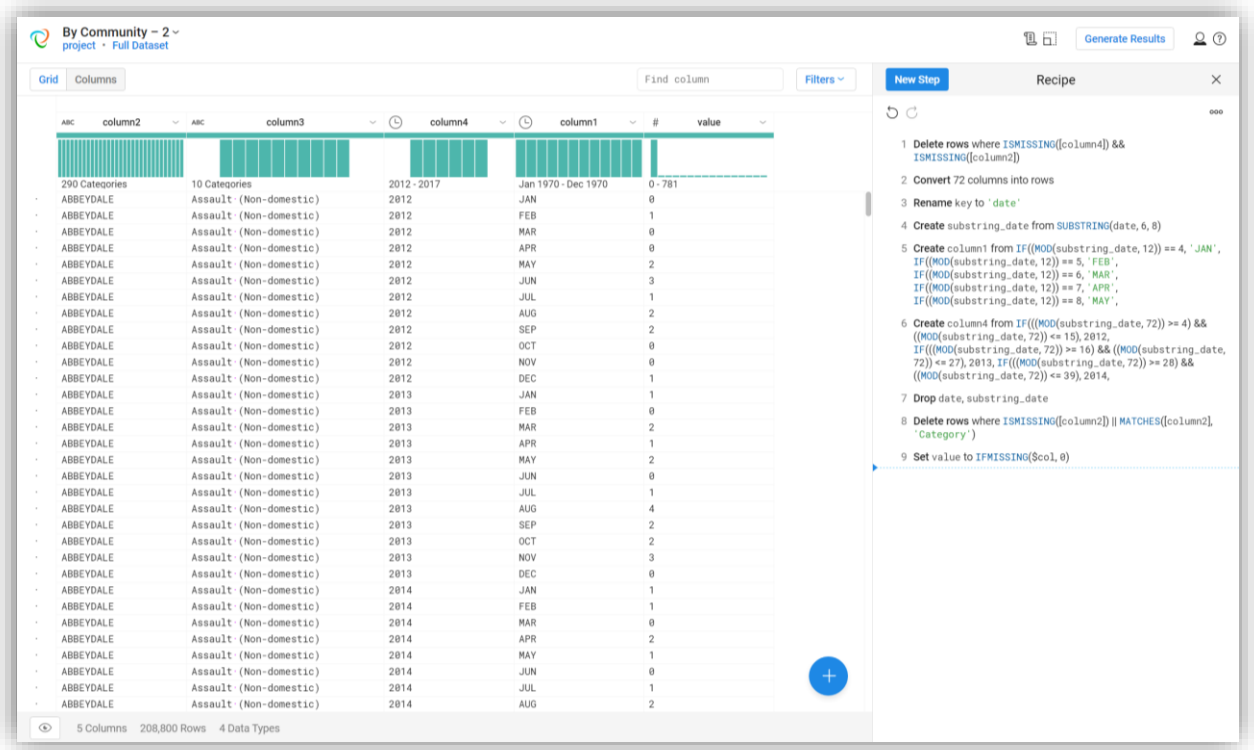


Figure 11 - Final result after all the steps of the recipe

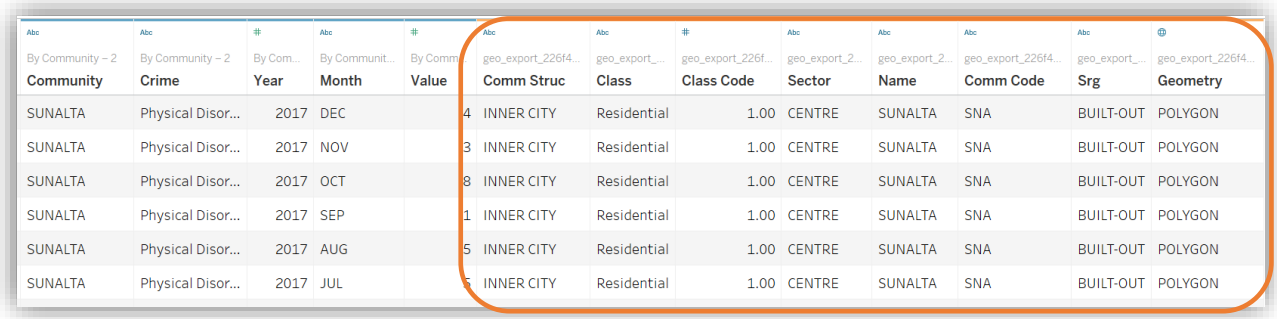
Calgary Shape File

This data set was taken from:

https://data.calgary.ca/Base-Maps/Community-Boundaries/ab7m-fwn6

Using Tableau, I joined together the community file and the shape file, to see what the shape file looks like:

Figure 12 – Inner join with Community (from Community Crime Data Set) and Name column in shape file



By Community – 2 Community	By Community – 2 Crime	By Com... Year	By Communit... Month	By Comm Value	geo_export_226f4... Comm Struc	geo_export_... Class	geo_export_226f... Class Code	geo_export_2... Sector	geo_export_2... Name	geo_export_226f4... Comm Code	geo_export_... Srg	geo_export_226f4... Geometry
SUNALTA	Physical Disor...	2017	DEC	4	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON
SUNALTA	Physical Disor...	2017	NOV	3	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON
SUNALTA	Physical Disor...	2017	OCT	8	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON
SUNALTA	Physical Disor...	2017	SEP	1	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON
SUNALTA	Physical Disor...	2017	AUG	5	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON
SUNALTA	Physical Disor...	2017	JUL	5	INNER CITY	Residential	1.00	CENTRE	SUNALTA	SNA	BUILT-OUT	POLYGON

Figure 13 - What the shape file contains

Using the shape file, I was able to retrieve a lot more attributes like the structure, sector, and class, dimensions which may later be useful for my analysis.

Decisions

A problem that I struggled with was deciding to join these data sets (Alberta Crime Data Set and Calgary Community Crime Data Set) or not. I opted not to join these because of the following reasons:

1. I only have community info for Calgary, since I'm not interested in other municipality communities. Thus it would be a left join with a lot of null values for other municipalities, and would be hard to use for visualization since Calgary values would be bloated.
2. Violations counts are different. We only have 10 different violation categories for the Calgary Community Crime Data Set, whereas the Alberta Crime Data set has 257. There would not be a clean mapping for this.

Research Questions

1. What qualities of communities more prone to crime in Calgary? (older communities, quadrants)
2. What crime is the most prevalent in each city? Are there any significant differences in crime between municipalities?
3. What is the ratio of adult vs youth committing crime, on average? Does it change drastically depending on the crime?

4. What are the top 2 crimes that are strongly correlated with one another? Which crimes can predict a violent crime such as first-degree murder?
5. Is there a month or year that crime rises in Calgary? (e.g. before Christmas, after New Year's where debt is high, summer where the weather is good, recessions)
6. Is Calgary becoming a more dangerous city? Based on the answers to the questions above, and based on the crime statistics alone, would you decide to live in Calgary and if so where?

Backup Questions

7. Is there a relationship between violent crime and non-violent crime? Which non-violent crimes can estimate violent crime?
8. Is there a big difference of actual crime incidents committed versus total charges of these that were cleared?
9. Which municipality is most lenient towards their youth committing crimes?
10. What crimes have experienced significant deviances over the years in Calgary? List at least three.