2024/10/30

Mpad 2003 Data Storytelling

Maxine Zeng

Presented to Jean-Sébastien Marier

Midterm Project: Exploratory Data Analysis (EDA)

Foreword

This data exploratory analysis focus on learning objectives of data storytelling:

1. Import, cleaning data from scratch.
2. Learning basics of github operations.
3. Learning how to write markdown files.

1. Introduction

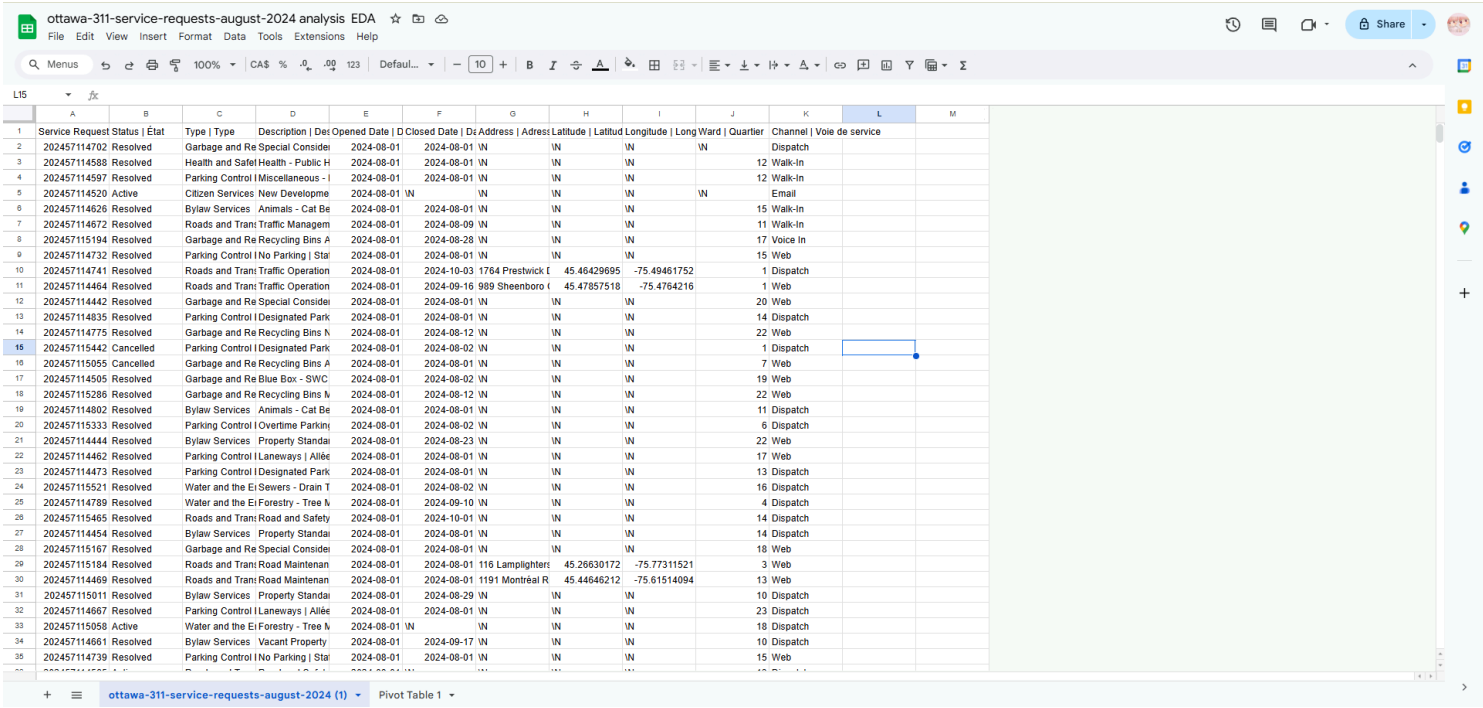
This exploratory data analysis will be analyzing the City of Ottawa dataset of citizen requests.
The City of Ottawa receives requests from four sources: 311 contact centre, client service centre, 311 email and web-based self- service portal. Which includes variety of content such as from Bylaw Services, Citizen Services, City Facilities and Garbage and Recycling, etc. It includes attributes of Service Request ID, Status, Description, Type, Opened Date, Closed Date, Address, Latitude, Longitude and Channel. They collected these requests and documented each one of them down into a dataset. The dataset is meant to update daily.

This analysis will including sections of getting data, understanding data, potential story, conclusion and references list.

- Link to the original dataset from City of Ottawa (updating daily):
- Link to the csv version I used.

2. Getting Data

First, I download the data from the github link assigned with a shortcut Ctrl + S, place it under my github folder. Then open google sheet, select tab "file", then "import", upload the file from saved location.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Service Request	Status | État	Type | Type	Description | Des	Opened Date | D	Closed Date | Da	Address | Adress	Latitude | Latitud	Longitude | Long	Ward | Quartier	Channel | Voie de service		
2	202457114702	Resolved	Garbage and Re	Special Conside	2024-08-01	2024-08-01	\N	\N	\N	\N	Dispatch		
3	202457114588	Resolved	Health and Safet	Health - Public H	2024-08-01	2024-08-01	\N	\N	\N	12	Walk-In		
4	202457114597	Resolved	Parking Control I	Miscellaneous - I	2024-08-01	2024-08-01	\N	\N	\N	12	Walk-In		
5	202457114520	Active	Citizen Services	New Developme	2024-08-01	\N	\N	\N	\N	\N	Email		
6	202457114626	Resolved	Bylaw Services	Animals - Cat Be	2024-08-01	2024-08-01	\N	\N	\N	15	Walk-In		
7	202457114672	Resolved	Roads and Tran	Traffic Managem	2024-08-01	2024-08-09	\N	\N	\N	11	Walk-In		
8	202457115194	Resolved	Garbage and Re	Recycling Bins A	2024-08-01	2024-08-28	\N	\N	\N	17	Voice In		
9	202457114732	Resolved	Parking Control I	No Parking | Sta	2024-08-01	2024-08-01	\N	\N	\N	15	Web		
10	202457114741	Resolved	Roads and Tran	Traffic Operation	2024-08-01	2024-10-03	1764 Prestwick D	45.46429695	-75.49461752	1	Dispatch		
11	202457114464	Resolved	Roads and Tran	Traffic Operation	2024-08-01	2024-09-16	989 Sheenboro (	45.47857518	-75.4764216	1	Web		
12	202457114442	Resolved	Garbage and Re	Special Conside	2024-08-01	2024-08-01	\N	\N	\N	20	Web		
13	202457114835	Resolved	Parking Control I	Designated Park	2024-08-01	2024-08-01	\N	\N	\N	14	Dispatch		
14	202457114775	Resolved	Garbage and Re	Recycling Bins N	2024-08-01	2024-08-12	\N	\N	\N	22	Web		
15	202457115442	Cancelled	Parking Control I	Designated Park	2024-08-01	2024-08-02	\N	\N	\N	1	Dispatch		
16	202457115055	Cancelled	Garbage and Re	Recycling Bins A	2024-08-01	2024-08-01	\N	\N	\N	7	Web		
17	202457114505	Resolved	Garbage and Re	Blue Box - SWC	2024-08-01	2024-08-02	\N	\N	\N	19	Web		
18	202457115286	Resolved	Garbage and Re	Recycling Bins N	2024-08-01	2024-08-12	\N	\N	\N	22	Web		
19	202457114802	Resolved	Bylaw Services	Animals - Cat Be	2024-08-01	2024-08-01	\N	\N	\N	11	Dispatch		
20	202457115333	Resolved	Parking Control I	Overtime Parkin	2024-08-01	2024-08-02	\N	\N	\N	6	Dispatch		
21	202457114444	Resolved	Bylaw Services	Property Standa	2024-08-01	2024-08-23	\N	\N	\N	22	Web		
22	202457114462	Resolved	Parking Control I	Laneways | Allée	2024-08-01	2024-08-01	\N	\N	\N	17	Web		
23	202457114473	Resolved	Parking Control I	Designated Park	2024-08-01	2024-08-01	\N	\N	\N	13	Dispatch		
24	202457115521	Resolved	Water and the Ei	Sewers - Drain T	2024-08-01	2024-08-02	\N	\N	\N	16	Dispatch		
25	202457114789	Resolved	Water and the Ei	Forestry - Tree M	2024-08-01	2024-09-10	\N	\N	\N	4	Dispatch		
26	202457115465	Resolved	Roads and Tran	Road and Safety	2024-08-01	2024-10-01	\N	\N	\N	14	Dispatch		
27	202457114454	Resolved	Bylaw Services	Property Standa	2024-08-01	2024-08-01	\N	\N	\N	14	Dispatch		
28	202457115167	Resolved	Garbage and Re	Special Conside	2024-08-01	2024-08-01	\N	\N	\N	18	Web		
29	202457115184	Resolved	Roads and Tran	Road Maintenan	2024-08-01	2024-08-01	116 Lamplighters	45.26630172	-75.77311521	3	Web		
30	202457114469	Resolved	Roads and Tran	Road Maintenan	2024-08-01	2024-08-01	1191 Montréal R	45.44646212	-75.61514094	13	Web		
31	202457115011	Resolved	Bylaw Services	Property Standa	2024-08-01	2024-08-29	\N	\N	\N	10	Dispatch		
32	202457114667	Resolved	Parking Control I	Laneways | Allée	2024-08-01	2024-08-01	\N	\N	\N	23	Dispatch		
33	202457115058	Active	Water and the Ei	Forestry - Tree M	2024-08-01	\N	\N	\N	\N	18	Dispatch		
34	202457114661	Resolved	Bylaw Services	Vacant Property	2024-08-01	2024-09-17	\N	\N	\N	10	Dispatch		
35	202457114739	Resolved	Parking Control I	No Parking | Sta	2024-08-01	2024-08-01	\N	\N	\N	15	Web		

Figure 1: The imported dataset in Google Sheets.

- Link to my dataset imported in Google Sheets

From A to K, there are 11 columns and 28539 row of data for each individual cases. Apparently, the original dataset is quite overwhelming for analyzing. The data looks quite clean and well structured, but it can still benefit from some modification for a clear readability.

The variables in different columns can be identify as following:

1. Categorical variables: refers to a value that can not be quantifiable, either nominal or ordinal.
2. Nominal variables: One that describes a name, label or category without natural order.
3. Ordinal variables: The value can be defined with an order relation in between.

(Statistic Canada, 2020)

Identification for specific columns:

1. Column A "Service Request ID | Numéro de demande" contains numerical ordinal ID numbers, which are all unique to its own case.

2. Column B "Status | État" contains nominal categorical data, which only has "Resolved", "Cancelled" and "Active" as input, they have no order relation in between.

3. Column C "Type | Type" is nominal categorical data defining the request under a big category. It has 11 categories in total.

4. Column D "Description | Description" Contains 554 sub category data under each big category, which is also nominal categorical data.

5. Column E "Opened Date | Date d'ouverture" and Column F "Closed Date | Date de fermeture" are ordinal date data, there will always be a specific opened data on column e, but Column F can be mark as "/N" because it is not resolved.

6. Column G "Address | Adresse" includes text information of address that are nominal. A lot of them were marked as "\N".

7. Column H "Latitude | Latitude" and I "Longitude | Longitude" are discrete data indicating the coordinate of the address. Can be marked as "\N" with address

8. Column J "Ward | Quartier" is 24 ordinal categorical data of the ID of wards in Ottawa, but some of them were marked as "\N" as the 25th category. The ID of wards makes the data cleaner and easier to use compare to data record with ward names, which may have different versions.

9. Column K "Channel | Voie de service" is nominal categorical data of how the case being requested, there are 6 category in total.

After understanding the data, I wonder if there is a correlation between wards and type categories of issues.

- Statistics: Power from Data! 4.2 Types of variables

3. Understanding Data

3.1. VIMO Analysis

This VIMO Analysis is focusing on Column C Type, Column D Description and Column J Ward Columns:

1. Valid:

Valid data meets 3 conditions: It is not blank or missing, it is within a valid range, and it is correct. (Statistics Canada, 2020)

Most data are valid under each category except for the missing ones.
They are identify as Micro data, which are summarized into a big category.

2. Invalid:

Invalid data refers to certain unreasonable or impossible value under a specific category.(Statistics Canada, 2020)
I check through Google sheet's inside column stat and filtering, there are no data seems to be Invalid except for the missing ones.

3. Missing:

Missing data means the cell was left with blank.(Statistics Canada, 2020)
In our dataset missing values are marked as "\N".

There are 1549 rows of data is being marked as "\N" under column of wards.
There are 628 rows are being marked as "\N" under column Description
There are 2 rows of data being marked as "\N" under column of types.

- Data Accuracy and Validation: Methods to ensure the quality of data

4. Outlier:

Outlier in a dataset refers to a value that is either extremely large or small. (Statistics Canada, 2020)

I noticed that for the type Water and the Environment, there are 834 row of ward ID being marked as missing, which is noticeably higher than other types. This problem could be due to location ambiguity, which some issues location were not clearly defined within ward boundaries or have cross ward boundaries, leading to reporting confusions.

3.2. Cleaning Data

To clean the data, I duplicated the original dataset into a new sheet called "Type and ward", then I hided most of the columns except the three columns (Type, Description and Ward) that I needed. Then I removed all the French labels after the divider for each column on first row. Next, I selected the first row and freeze them as my header, they will stay on top of when I scroll down.

For the Type column, I found all the "and" using the find functions with Ctrl F, the replace all of them to a divider " | ", resulting a shorter value for each row with more clear readability. After that I removed all of "the" for the same purpose. Then I sorted the data from A to Z in Type column.

For the description tab, I placed spilt functions `=SPLIT(D2, "|")` to split the French description to two extra columns. Then I copy the value in the English description column with Ctrl + Shift + V to copy plain text only to get rid of the split formula. Then deleted the French description column and the original one. This action refers to the data cleaning tutorial in week 5, which have made the table looks cleaner and simpler. (Marier, 2024)

I further split the description tab with `=SPLIT(D2, "-")` , which splits it into 3 columns. Then I copy and pasted the three column as plain values, then I rejoin the last two columns together with `=CONCATENATE(F2," ", G2)` . The concatenate column is named as Sub description, which is useful for looking at data within a smaller range then the original Description categories. Then I used the trim white space tool under the data cleaning of Google onto the new Description and Sub Description column.

- MPAD 2003: 5.1 Cleaning Data in Google Sheets

3.3. Exploratory Data Analysis (EDA)

The cleaned dataset now only includes four column, Type, Description, Sub Description and Ward. Which I can used to see patterns of the issues between wards.

A pivot table `Type and Ward Stats` was created to analysis the relationship between Wards and Type. Which only contains nominal variables. The table is also an ordinal level of measurement with these following requirements met(TouhidulIslam, 2021):

1. The data classifications of the columns are mutually exclusive and exhaustive.
2. The data can be meaningfully ranked by its number, from low to high.

- [Levels of Measurement (Nominal, Ordinal, Interval, Ratio) in Statistics] (https://www.datasciencecentral.com/levels-of-measurement-nominal-ordinal-interval-ratio-in/)

In the pivot table, I also calculated the mean of all Types using `=DIVIDE(AA2, 24)` , and the median with `=MEDIAN((B2:Z2))` . The average of Requests between 24 wards is approximately 1190 per ward. The highest one is ward 12 Rideau-Vanier with 1628 request and Ward 20 Osgoode with 648 requests, the median is 1168 requests from ward 18 Cumberland.

Type\Ward IDs	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	Missing (\N)	Grand Total	Average	Median	Outlier
Missing (\N)	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0.08333333333	0	
Bylaw Services	173	169	243	145	80	236	228	287	159	274	126	611	246	485	292	278	279	225	255	100	156	208	154	162	72	5643	235.125	225	
Citizen Services	7	12	11	13	8	30	15	8	4	7	2	16	3	15	8	10	7	10	16	11	7	59	8	3	148	438	18.25	10	
City Facilities	3	1	0	2	1	2	17	11		5	3	12	1	4	0	4	3	6	5	2	5	5	3	6	0	101	4.208333333	3	
Garbage | Recycling	582	450	694	356	170	520	346	474	311	432	248	331	332	331	472	361	388	340	639	300	366	521	449	411	433	10257	427.375	388	
Health | Safety	2	2	1	1	4	3	3	10	1	5	1	8	8	28	23	5	14	14	2	0	1	2	3	2	3	146	6.083333333	3	
Licenses | Permits	0	0	0	0	5	3	0	7	2	4	3	5	4	3	9	1	6	5	2	6	3	0	1	1	0	70	2.916666667	3	
Parking Control Enforcement	58	88	176	82	14	142	143	169	122	134	73	264	176	300	224	128	245	143	138	16	43	156	115	111	15	3275	136.4583333	134	
Recreation | Culture	38	45	36	68	19	44	36	33	15	64	19	50	30	25	33	38	35	27	39	11	10	54	41	30	4	844	35.16666667	35	
Roads | Transportation	241	170	95	131	314	162	181	187	158	202	122	204	160	254	171	220	188	199	167	158	138	100	148	108	36	4214	175.5833333	167	
Social Community Service	1	1	0	3	0	4	18	23	9	1	4	36	2	35	36	17	9	2	1	0	0	2	24	3	4	235	9.791666667	3	
Water | Environment	99	103	106	106	45	121	141	145	68	110	54	91	90	90	141	99	151	197	102	44	55	91	137	93	834	3313	138.0416667	102	
Grand Total	1204	1041	1362	907	660	1267	1128	1354	850	1238	655	1628	1052	1571	1409	1161	1325	1168	1366	648	784	1198	1083	930	1549	28538	1189.083333	1168	1628

Figure 2: This pivot table shows the relationship between Ward and Type

Based on the pivot table, I created a few simple visualizations in Google sheets. In the graph visualizing type of requests per Ward on figure 3, the Garbage and Cleaning category in orange is the most common value. It can also be marked as significant data since it went over 30% of the data.

- Interpretation of VIMO table

In the second visualization in figure 4, you can also spot the outlier that was found earlier, which is from Water and Environment that are on the column for requests's ward being marked as missing.

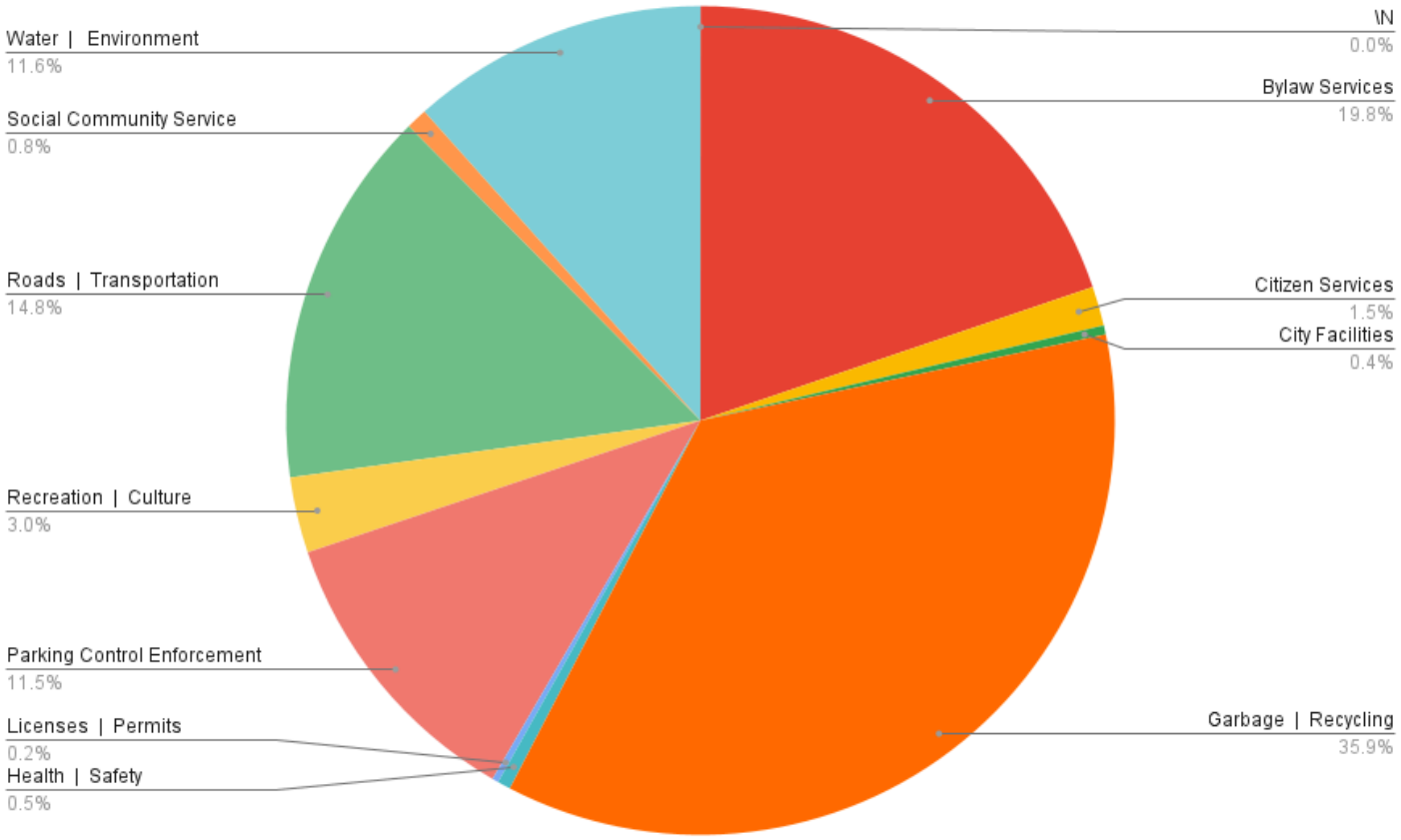


Figure 3: Percentage of Types

I further dived into the missing value in percentage, Citizen Service has 25.3% ward information missing and Water and Environment has 20.1% percentage missing. These two are the only ones that is significantly higher than others.

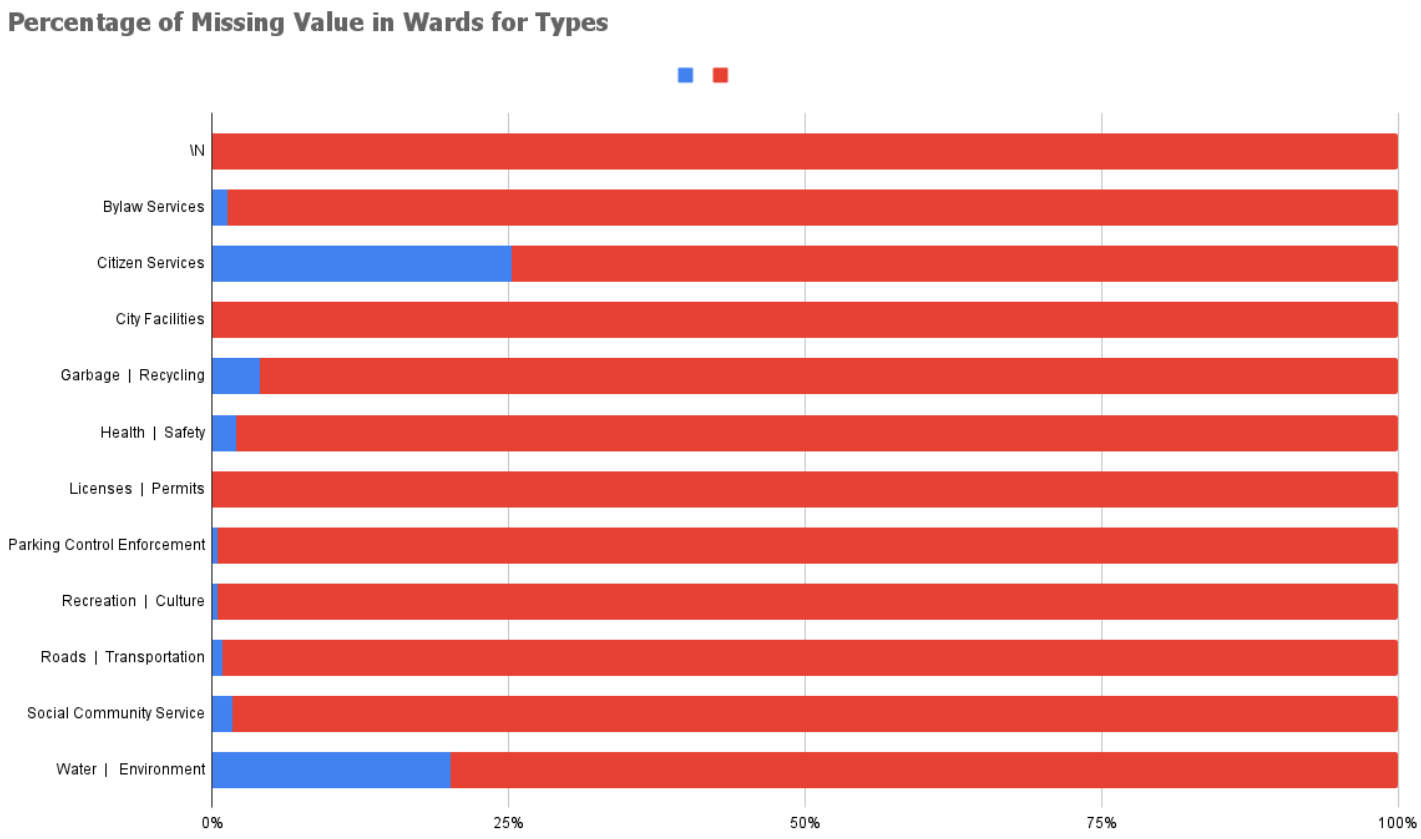


Figure 4: Percentage of Types

The Licenses | Permit rows are most rare requests among all the categorical, which is not surprising, but Health | Safety row being the second least seems to be odd.

Covariations happened between the number of the types and the total number of requests per ward. For example, ward 12 Rideau-Vanier has the highest total number (1628) and the most Bylaw Services (611) request at the same time. Ward 20 Osgoode has the fewest number of total number and fewest number of Water Requests. (Wickham 2017)

- R for Data Science

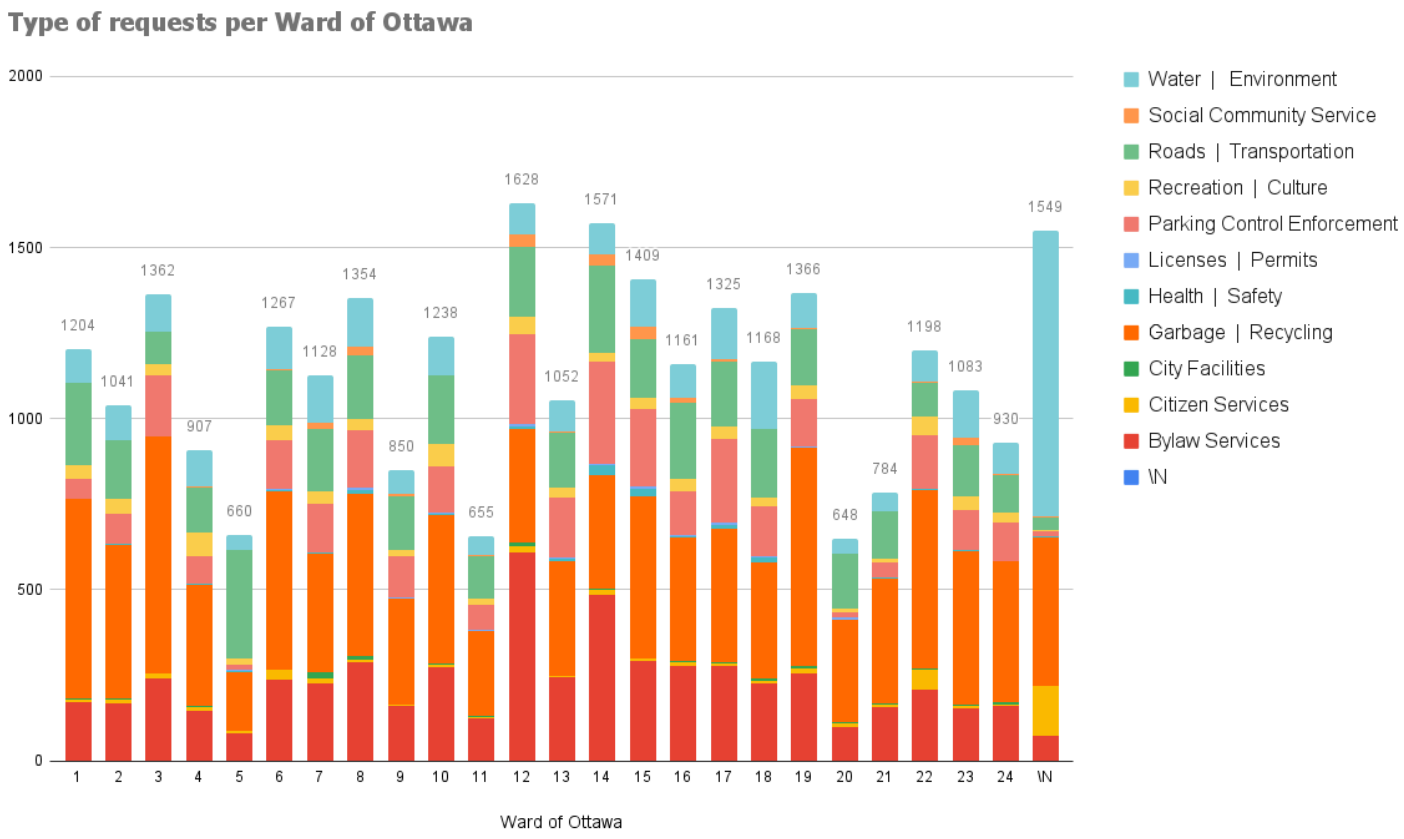


Figure 5: This graph visualized the type of requests per district

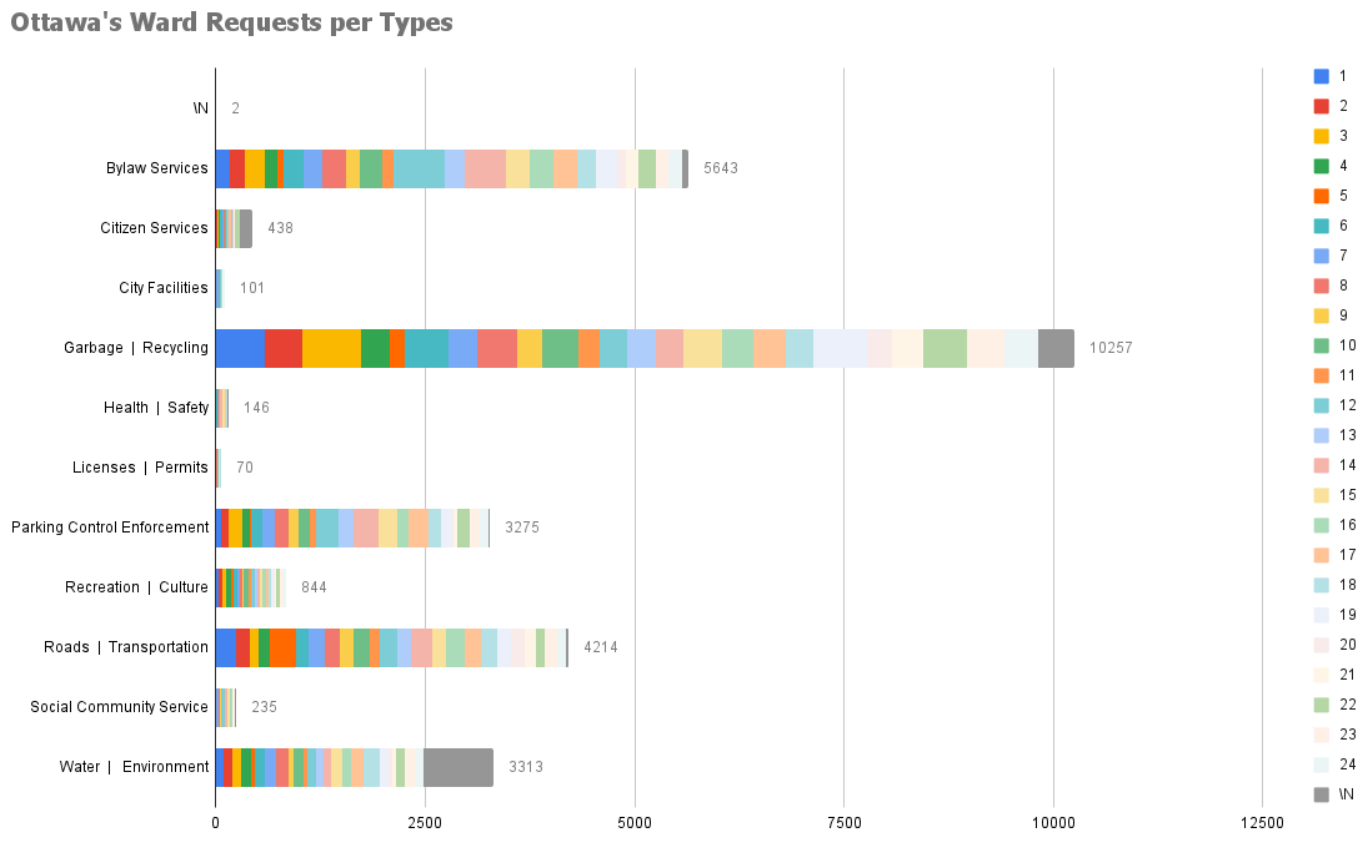


Figure 6: This graph visualized each district's request number in different types

4. Potential Story

A potential story in my analysis can be sharing which ward in Ottawa is the most habitable and safe, by analyzing which ward has less issues. For example we can create a heat map using the ward map of Ottawa, using the color red to indicate which wards are tend to be problematic.

This story can be further narrowed down targeting a certain group, for example the analysis can focus on elders. Then we can focus on types like Roads and Transportation, Health and safety, and Bylaw services that closer binds with their daily life. If a district has high amount of noise complaints under bylaw services, poor road management, and active health concerns, it will be identify as less suitable for elders to live in there.

An example of this type of data storytelling, CTV Ottawa reported noise issue in 2015 ward by ward. The former bylaw Chief Roger Chapman claimed that the most popular noise complaints they received are

loud musics, construction noises and machine noises, anything above 50 db is consider violation. The top 1 complaint ward is 12 Rideau Vanier. (CTV Ottawa 2015)

- What's all the noise? A ward-by-ward breakdown of noise complaints in Ottawa

For further information onto my topic, it would be a great help to interview the current city Councillors of Ottawa or bylaw chief, and the local residents who are currently experiencing the issues.

5. Conclusion

The most challenging part of this assignment is managing to clean the data itself. For a large dataset containing all most 30k rows of data, it is a heavy loaded work for my laptop to process, it glitched and crushed the browser for a few times. Reflecting on this experiencing, I recognized that I needed to narrow down my study to a smaller range with a more focused topic. Thus it would provided me with a lighter data handling.

On the other hand, the rewarding aspect came after the data cleaning when I created the visualizations. Seeing the data table and visualization turn out to be clear and readable for data storytelling, which all the efforts on the data before them seen well worth it.

Besides from the the data processing part, I also identified my knowledge gap on analysing the data, I needed to revisit course readings and research for additional resources online to learn and reference how to write more effective exploratory data analysis. Acquiring the new skills for telling a compelling story through data and data visualization made the project feel worthwhile.

In summary, this assignment has been a practice with challenges an rewards, emphasizing the importance of data cleaning and efficiency of data storytelling.

6. References

1. City of Ottawa. (2024, October 3). *2024 service requests. Open Ottawa*. https://open.ottawa.ca/documents/65fe42e2502d442b8a774fd3d954cac5/about

2. CTV Ottawa. (2015, June 19). *What's All the noise? A ward-by-ward breakdown of noise complaints in Ottawa.* https://ottawa.ctvnews.ca/what-s-all-the-noise-a-ward-by-ward-breakdown-of-noise-complaints-in-ottawa-1.2431535

3. Hong, S. P. (2015, March 27). *Interpretation of vimo table. University of Manitoba*. https://umanitoba.ca/manitoba-centre-for-health-policy/sites/manitoba-centre-for-health-policy/files/2021-11/vimo-table-interpretation.pdf

4. Marier, J.-S. (2024, September). *VIDEO | Cleaning Data in Google Sheets*. https://brightspace.carleton.ca/d2l/le/content/290806/viewContent/3855132/View

5. Statistics Canada. (2021, September 2). 4 data exploration 4.2 types of variables. 4.2 Types of variables. https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch8/5214817-eng.htm

6. Statistics Canada. (2022, May 11). *Data accuracy and validation: Methods to ensure the quality of data*. Government of Canada, Statistics Canada. https://www.statcan.gc.ca/en/wtc/data-literacy/catalogue/892000062020008

7. TouhidulIslam, M. (2021, February 25). *Levels of measurement (nominal, ordinal, interval, ratio) in statistics. Data Science Central*. Levels of Measurement (Nominal, Ordinal, Interval, Ratio) in Statistics

8. Wickham , H., & Grolemund, G. (2017). *R for data science. 7 Exploratory Data Analysis*. https://r4ds.had.co.nz/exploratory-data-analysis.html