Date TBD

Course Code & Course Name

MPAD 2002A Introductory Data Storytelling

Student's First Name & Last Name

Gabriel Castillo, Avery O'Donnell, Tina-lise Kostopoulos

Presented to Jean-Sébastien Marier

Exploratory Data Analysis (EDA) & Pitch

Use one hashtag symbol (`#`) to create a level 1 heading like this one.

Foreword

For this assignment, you must extract data from a dataset provided by the instructor. You must then clean and analyze the data, create exploratory charts/visualizations, and find a potential story idea. Your assignment must clearly detail your process. You are expected to write about 1500-2000 words, and to include several screen captures showing the different steps you went through. Your assignment must be written with the Markdown format and submitted on GitHub Classroom.

I have been assigning different versions of this project to my digital journalism and data storytelling students for a few years now. Its structure was inspired by the main sections/chapters of *The Data Journalism Handbook*. This version was further inspired by the Key Capabilities in Data Science program offered by the University of British Columbia (UBC).

Here are some useful resources for this assignment:

- GitHub's *Basic writing and formatting syntax* page
- The template repository for this assignment in case you delete something by mistake

Did you notice how to create a hyperlink? In Markdown, we put the clickable text between square brackets and the actual URL between parentheses.

And to create an unordered list, we simply put a star (`*`) before each item.

1. Introduction

For this assignment, we will be analyzing a City of Ottawa dataset of information in the different wards of Ottawa. The dataset was gathered by survey and includes information about the number of people per household, as well as their income and employment status, divided by their ward/living location. It also states the ages of the people. In the first section, we will explain how we got the dataset and the information that we are using. In the following sections, we will be assessing the accuracy of the dataset as well as exploring the data and analyzing it for a potential story.

Teachers URL: https://raw.githubusercontent.com/jsmarier/files-for-course-assignments/refs/heads/main/2021_Long_Form_Census_-_Ward_Data.csv

Original Dataset: https://open.ottawa.ca/datasets/ottawa::2021-long-form-census-ward-data/explore

2. Getting Data

Use two hashtag symbols (`##`) to create a level 2 heading like this one.

To include a screen capture, use the sample code below. Your images should be saved in the same folder as your `.md` file.

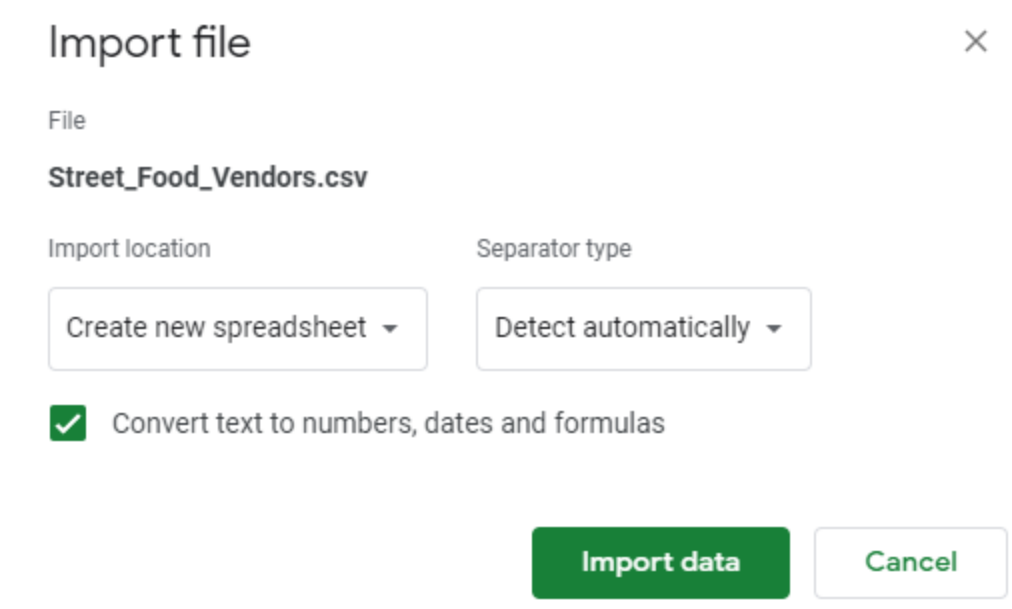
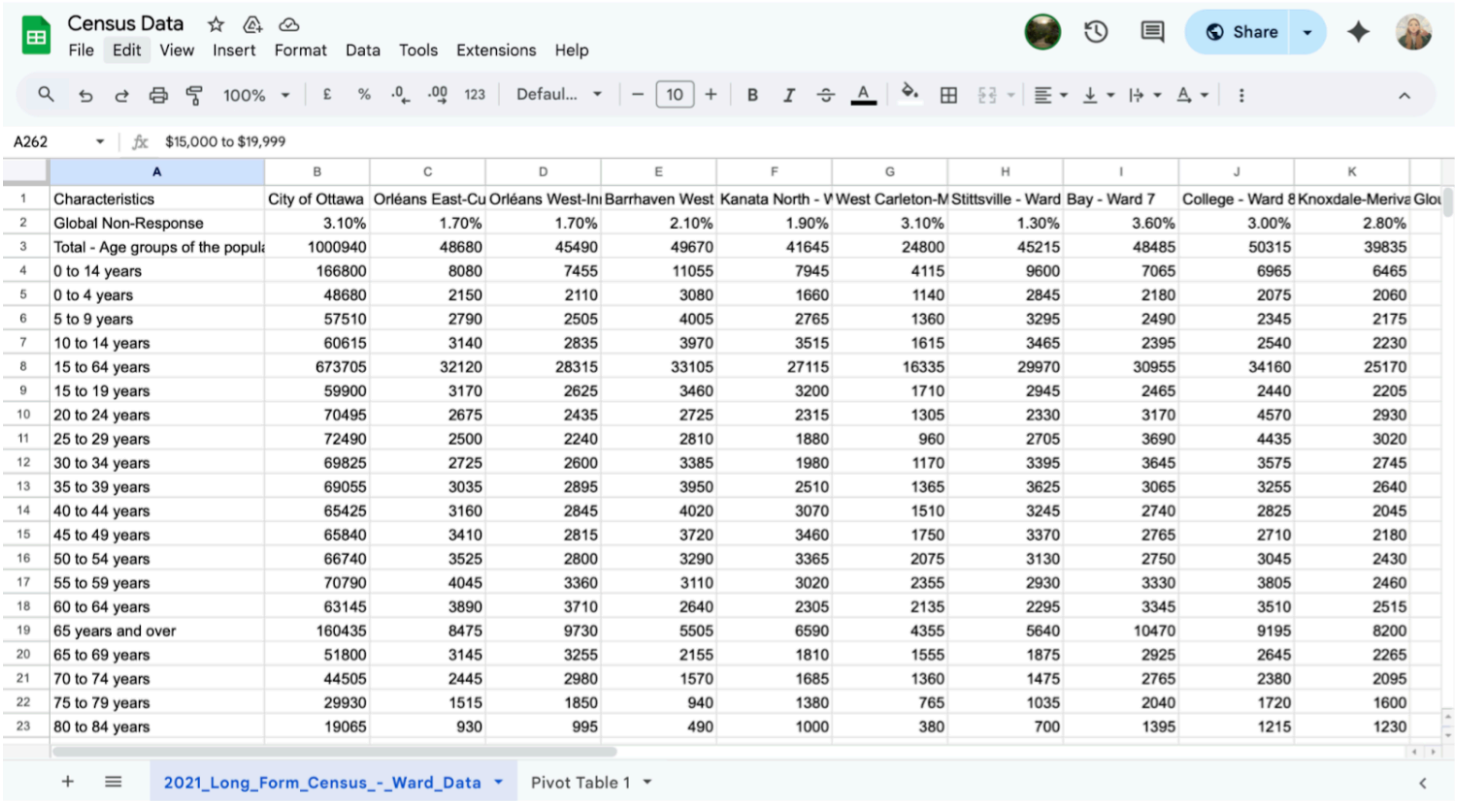


Figure 1: The "Import file" prompt on Google Sheets.

To begin, we imported the dataset into Google Sheets to make it easier to view, clean and analyze. We opened a new Google Sheet, clicked File, Import, Upload, and selected the CSV file from my computer. When prompted,we chose “Insert new sheet(s)” so the data appeared in a new tab while keeping the headers intact. Once the file loaded, the column headings automatically appeared in the first row, followed by 2,602 rows of data.
Below is a screenshot of the dataset immediately after importation:



	A	B	C	D	E	F	G	H	I	J	K
1	Characteristics	City of Ottawa	Orléans East-Cu	Orléans West-In	Barrhaven West	Kanata North - W	West Carleton-M	Stittsville - Ward	Bay - Ward 7	College - Ward 8	Knoxdale-Meriva
2	Global Non-Response	3.10%	1.70%	1.70%	2.10%	1.90%	3.10%	1.30%	3.60%	3.00%	2.80%
3	Total - Age groups of the popul	1000940	48680	45490	49670	41645	24800	45215	48485	50315	39835
4	0 to 14 years	166800	8080	7455	11055	7945	4115	9600	7065	6965	6465
5	0 to 4 years	48680	2150	2110	3080	1660	1140	2845	2180	2075	2060
6	5 to 9 years	57510	2790	2505	4005	2765	1360	3295	2490	2345	2175
7	10 to 14 years	60615	3140	2835	3970	3515	1615	3465	2395	2540	2230
8	15 to 64 years	673705	32120	28315	33105	27115	16335	29970	30955	34160	25170
9	15 to 19 years	59900	3170	2625	3460	3200	1710	2945	2465	2440	2205
10	20 to 24 years	70495	2675	2435	2725	2315	1305	2330	3170	4570	2930
11	25 to 29 years	72490	2500	2240	2810	1880	960	2705	3690	4435	3020
12	30 to 34 years	69825	2725	2600	3385	1980	1170	3395	3645	3575	2745
13	35 to 39 years	69055	3035	2895	3950	2510	1365	3625	3065	3255	2640
14	40 to 44 years	65425	3160	2845	4020	3070	1510	3245	2740	2825	2045
15	45 to 49 years	65840	3410	2815	3720	3460	1750	3370	2765	2710	2180
16	50 to 54 years	66740	3525	2800	3290	3365	2075	3130	2750	3045	2430
17	55 to 59 years	70790	4045	3360	3110	3020	2355	2930	3330	3805	2460
18	60 to 64 years	63145	3890	3710	2640	2305	2135	2295	3345	3510	2515
19	65 years and over	160435	8475	9730	5505	6590	4355	5640	10470	9195	8200
20	65 to 69 years	51800	3145	3255	2155	1810	1555	1875	2925	2645	2265
21	70 to 74 years	44505	2445	2980	1570	1685	1360	1475	2765	2380	2095
22	75 to 79 years	29930	1515	1850	940	1380	765	1035	2040	1720	1600
23	80 to 84 years	19065	930	995	490	1000	380	700	1395	1215	1230

The dataset contains 26 columns and 2,602 rows, each representing demographic information from the City of Ottawa’s census data. Overall, the spreadsheet appears well structured, but there are a few inconsistencies and missing entries. Some ward names differ slightly from those listed on the City of Ottawa’s official website, which may indicate older ward boundaries or formatting errors. There are also a few blank cells where data may have been suppressed for privacy or unavailable at the time of collection.
Analyzing 3 columns:

Column A (“Characteristics”) contains nominal variables describing each demographic measure (for example, total population, age groups, or median age).

Column B (“City of Ottawa”) holds numerical data summarizing citywide totals or percentages.
Column C through Z represent individual wards, each containing quantitative values for the same characteristics.

One notable observation is that some wards with similar total populations have very different median ages and household structures. This raises the question: how do demographic characteristics, such as household sizes or age distribution, influence income levels across Ottawa's wards?

Here are examples of functions and lines of code put in grey boxes:

1. If you name a function, put it between "angled" quotation marks like this: `IMPORTHTML` .
2. If you want to include the entire line of code, do the same thing, albeit with your entire code: `=IMPORTHTML("https://en.wikipedia.org/wiki/China"; "table", 5)` .
3. Alternatively, you can put your code in an independent box using the template below:

```
=IMPORTHTML("https://en.wikipedia.org/wiki/China"; "table", 5)
```

This also shows how to create an ordered list. Simply put `1.` before each item.

3. Understanding Data

3.1. VIMO Analysis

Our Census dataset from the City of Ottawa includes a wide range of information. The information types are separated by wards, such as Orleans East/West, Barrhaven, etc. These sections allow for all the data to be presented to the user in an organized, accessible way. The data can range from finance (incomes, rents, etc.) to the number of people per household, who is employed, etc. We decided to focus on the number of people in each household and how it correlates to general household income. The data to figure out the number of people per ward living in houses was found in row 47, and follows along all columns (wards) in the Census. There was also the Average total income in 2019 among recipients ($), which was in row 142. We conducted a detailed VIMO analysis to deeper analyze the data.

The VIMO analysis covers four main sections: Valid, Invalid, Missing, and Outlier data, in order to identify potential data quality issues and ensure data accuracy before being used in our research and analysis. When it came to performing this analysis, we were able to confirm with our initial sweep that none of the data was missing (every section was filled out), and none of the data was invalid (every section was filled out properly, not with invalid or misplaced data). In terms of data Validity, it was an officially conducted City of Ottawa Census, so we count the data as valid. In terms of outliers, there weren't any that really stood out in either category. The following images display a more detailed analysis of possible outliers, as well as quartiles, etc.

Image 1: Individuals Living in Households

First quartile:	$Q1 = \frac{x_7 + x_6}{2} = \frac{37035 + 37020}{2} = 37027.5$
Third quartile:	$Q3 = \frac{x_{18} + x_{19}}{2} = \frac{48680 + 48700}{2} = 48690$
Interquartile range:	IQR = Q3 - Q1 = 48690 - 37027.5 = 11662.5
Outliers(numbers less than Q1 – 1.5×IQR or greater than Q3 + 1.5×IQR):	**None** (Q1 - 1.5×IQR = 19533.75 | Q3 + 1.5×IQR = 66183.75)

Image 2: Average Incomes in 2019:

First quartile:	$Q1 = \frac{x_7 + x_6}{2} = \frac{54650 + 54350}{2} = 54500$
Third quartile:	$Q3 = \frac{x_{18} + x_{19}}{2} = \frac{67000 + 68700}{2} = 67850$
Interquartile range:	IQR = Q3 - Q1 = 67850 - 54500 = 13350
Outliers(numbers less than Q1 – 1.5×IQR or greater than Q3 + 1.5×IQR):	**None** (Q1 - 1.5×IQR = 34475 | Q3 + 1.5×IQR = 87875)

3.2. Cleaning Data

After the initial analyses of the data, as well as our VIMO analysis, we then moved on to cleaning our data. We used 3 of the methods we had learned in our Media Production and Design: Introductory Data Storytelling. The methods were Google Sheets data-cleaning tools, which we used to get rid of

unnecessary/extra spacing, duplicates, data validity tests, etc. We did have some work to do when it came to removing spaces, trimming white spaces, as well as fixing hyphens in the ward names. But, the process was made much easier in Google Sheet's data cleanup tool set. Aside from the initial cleanup, we then experimented with freezing the rows and columns we needed in order to keep the data accessible to the user. We did this to a smaller dataset of just our isolated variables, and transposed the data. We then froze the top row, which contains the titles of our two categories. (As shown in the following screenshot). Instead, we went into OpenRefine, using Clusters. In OpenRefine, we used facets (text facets, numeric facets), clustering, and used it as well to find extreme highs/lows in the income and household columns, but as mentioned in the previous VIMO analysis, there were no obvious outliers or imposing errors in the data.

Image of frozen rows and isolated data:

H3 | *fx*

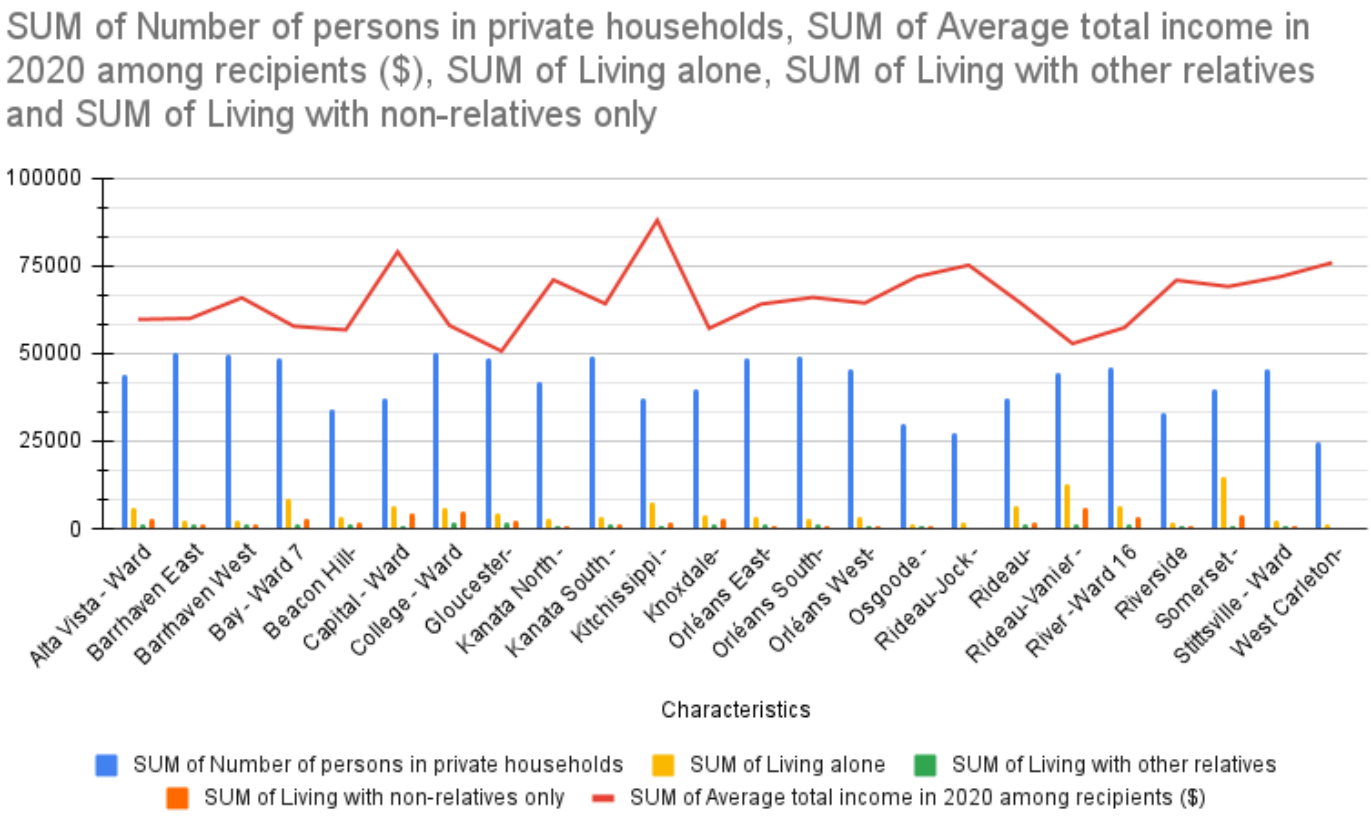
	A	B	C	D	E	F
1	Wards (24)	Number of persons in private households	Average total income in 2019 among recipients ($)			
2	City of Ottawa	1000940	60900			
3	Orléans East-Cumberland - Wa	48680	62000			
4	Orléans West-Innes - Ward 2	45490	61550			
5	Barrhaven West - Ward 3	49670	63500			
6	Kanata North - Ward 4	41645	67000			
7	West Carleton-March - Ward 5	24800	73200			
8	Stittsville - Ward 6	45215	68700			
9	Bay - Ward 7	48485	54650			
10	College - Ward 8	50315	54150			
11	Knoxdale-Merivale - Ward 9	39835	54350			
12	Gloucester-Southgate - Ward 1	48700	47360			
13	Beacon Hill-Cyrville - Ward 11	34050	54200			
14	Rideau-Vanier - Ward 12	44310	48280			
15	Rideau-Rockcliffe - Ward 13	37395	61600			
16	Somerset - Ward 14	39795	58650			
17	Kitchissippi - Ward 15	37035	82100			
18	River - Ward 16	45895	53100			
19	Capital - Ward 17	37020	74400			
20	Alta Vista - Ward 18	44065	55650			
21	Orléans South-Navan - Ward 1	49055	62400			
22	Osgoode - Ward 20	29965	69000			
23	Rideau-Jock - Ward 21	27045	73700			
24	Riverside South-Findlay Creek	32995	65100			
25	Kanata South - Ward 23	49110	61500			
26	Barrhaven East - Ward 24	50360	56750			

3.3. Exploratory Data Analysis (EDA)

Insert text here.

Here is the Pivot Table made during the EDA as well as the chart.

	A	B	C	D	E	F
1	*Characteristics*	SUM of Number	SUM of Average	SUM of Living al	SUM of Living w	SUM of Living wi
2	Alta Vista - Ward	44065	59650	6005	1400	2695
3	Barrhaven East -	50360	59900	2535	1450	1385
4	Barrhaven West	49670	65800	2595	1115	1100
5	Bay - Ward 7	48485	57700	8505	1215	2925
6	Beacon Hill-Cyrv	34050	56650	3575	1140	1975
7	Capital - Ward 1	37020	78900	6355	980	4330
8	College - Ward 8	50315	57900	5805	1605	5065
9	Gloucester-Sout	48700	50600	4515	1645	2225
10	Kanata North - W	41645	70900	3055	815	835
11	Kanata South - W	49110	64100	3205	1075	1065
12	Kitchissippi - Wa	37035	87900	7385	625	1780
13	Knoxdale-Meriva	39835	57100	4085	1145	2800
14	Orléans East-Cu	48680	64000	3350	1140	1000
15	Orléans South-N	49055	65900	2635	1100	1005
16	Orléans West-In	45490	64300	3545	1005	1015
17	Osgoode - Ward	29965	71800	1555	575	610
18	Rideau-Jock - W	27045	75100	1675	450	470
19	Rideau-Rockcliff	37395	64300	6605	1065	1970
20	Rideau-Vanier -	44310	52750	12500	1460	5820
21	River -Ward 16	45895	57300	6330	1395	3280
22	Riverside South-	32995	70800	1625	835	585
23	Somerset - Ward	39795	69000	14760	885	4135
24	Stittsville - Ward	45215	71800	2545	885	925
25	West Carleton-M	24800	75800	1470	370	475
26	**Grand Total**	**1000930**	**1569950**	**116215**	**25375**	**49470**
27						



This section should include a screen capture of your pivot table, like so:

	A	B
1	*Position*	SUM of Goals
2	C	10
3	D	4
4	G	0
5	LW	8
6	RW	7
7	**Grand Total**	**29**

Figure 2: This pivot table shows...

This section should also include a screen capture of your exploratory chart, like so:

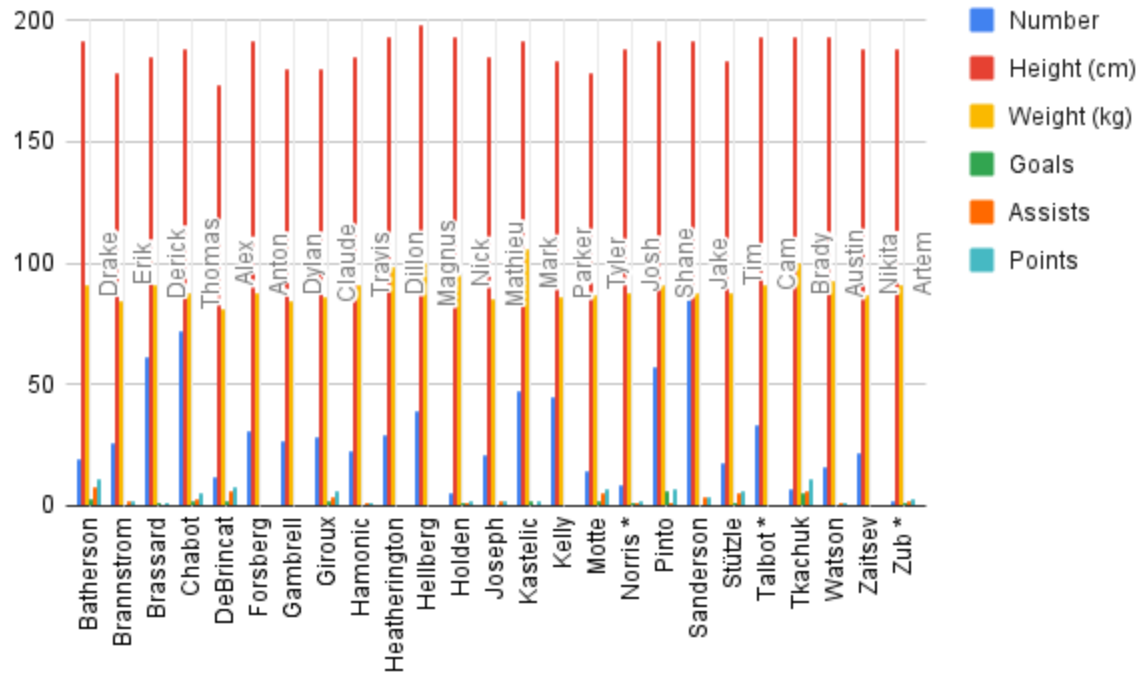


Figure 3: This exploratory chart shows...

4. Potential Story

A potential story could explore how the number of household members across Ottawa's wards relates to total household income, revealing patterns of inequality, affordability, and access to resources across the city. By comparing households ranging from single-person units to those with five or more members, the story could uncover how family size affects financial stability, housing options, and quality of life. To tell this story effectively, ward-level data from the City of Ottawa's Open Data portal and the 2021 Statistics Canada Census would be essential to establish clear links between household composition and income levels. Interviews with residents from different household sizes in both higher and lower-income wards would add a personal dimension, illustrating how income disparities shape lived experiences. City councillors could offer insight into how these differences influence local policy decisions, while economists or sociologists from Carleton University could explain the broader social and economic patterns behind the numbers. Organizations such as United Way East Ontario and Ottawa Community Housing could provide frontline perspectives on how income and family size impact access to housing and community services. By combining quantitative data with qualitative interviews, this story would paint a comprehensive picture of how household size continues to shape economic opportunity and inequality across Ottawa's diverse communities, offering readers both statistical context and human connection.

5. Conclusion

As a group, completing this exploratory data analysis and pitch project was a real learning experience that combined technical work, organization, and collaboration. We started by importing the 2021 Long Form Census Ward Data into Google Sheets, making sure it displayed properly and that we understood the structure of the dataset. Figuring out how to manage all the columns, formatting, and variables took time and teamwork, especially when Sheets didn't interpret the data the way we expected. The most challenging part of this assignment was definitely working with Google Sheets. Cleaning the data, experimenting with functions, and fixing formatting issues required a lot of trial and error. At the same time, this process was one of the most rewarding parts of the project because it helped us see the dataset transform from something confusing into something organized and meaningful. Once we created pivot tables and exploratory charts, we began identifying patterns that could form the basis of a strong story for our term project. This assignment also showed us where we still have room to grow, especially with using Markdown formatting, GitHub for version control, and more advanced data visualization tools. If we could do it again, we'd plan more time for experimenting with chart options and documenting our steps. Overall, this project strengthened our teamwork, problem-solving, and confidence in working with real-world data.

6. References

Include a list of your references here. Please follow APA guidelines for references. Hanging paragraphs aren't required though.

Here's an example:

Bounegru, L., & Gray, J. (Eds.). (2021). *The Data Journalism Handbook 2: Towards A Critical Data Practice*. Amsterdam University Press. https://ocul-crl.primo.exlibrisgroup.com/permalink/01OCUL_CRL/hgdufh/alma9910228900873051 53