1. Understanding of Datasets

We managed to find data to address a portion of the housing-related questions that the councilor had. For example, we obtained data from the Boston Census to determine the number of total housing units, owner-occupied and renter-occupied, which gave us a clearer understanding of the housing situation in District 7.

1.1. Income-restricted housing inventory

We found data on **income-restricted housing**, which refers to units designated as affordable for individuals or families with lower incomes. While this primarily applies to rental units, there are also instances where restrictions limit who can purchase properties, such as through affordable housing programs.

Questions from the councilor focus on how this number has changed or grown over time, which may highlight the increasing difficulty people face in obtaining affordable housing and the need for more government subsidies. Another question we addressed was related to **Section 8 vouchers**, a form of government assistance that helps low-income families, elderly individuals, and people with disabilities afford housing in the private market. This part of the analysis sought to use the data to determine the number of housing units that accept these vouchers, indicating the availability of subsidized housing options for those receiving this type of assistance.

1.2. RentSmart

We also found the **RentSmart** dataset that we can use for detailed analysis of the inspectional services, which is to protect the health and safety of business and residential communities by enforcing building, housing, health, and environmental regulations effectively. Ranging from 2016 to 2024, this dataset allows us to gain insights into how well the health and safety of the residents living in District 7 using the fields in the dataset, including violation type, addresses, neighborhood, zip code, property types, and years the building was built and remodeled.

With this information, we can answer the questions of how the total violations and properties have changed over time and what the correlation between them might be. From the preliminary analysis, enforcement violations have the highest number among all other violations and have been strikingly fluctuating on the higher end since 2020. It also shows that the number of households with 3 residents is the highest across all areas in this district. Therefore, we can find out which types of violations are under control and which require more effort while taking into account the number of residents living in a household.

1.3 Boston Census

This is an extensive dataset that contains detailed information about housing occupancy, such as the number of units per structure, the number of rooms, whether they're renter or owner occupied, utilities etc., With this information, we can answer questions about housing density and distribution across different areas of District 7, and potentially do a cross-comparison with other districts. We can also look at housing affordability and availability across different neighborhoods, examine occupancy and ownership trends, and analyze utility and infrastructure needs by neighborhood. These could be helpful guides for future urban planning and development for District 7.

2. Data cleaning and preprocessing

2.1. Income-restricted housing dataset

For the data cleaning and processing of the income restricted housing dataset, we tried different cleaning methods due to the problems encountered when associating the data with their respective districts, but we settled with the following:

- We dropped some missing data points from the dataset, as they represented a small percentage—less than 5%.
- The data type for ZIP code after loading to dataframe is double. We corrected the data type for zip codes by first transforming them into integers and then into strings, followed by adding 0 at the beginning to properly format them as zip codes.
- We filtered the data based on the ZIP codes that are part of the four neighborhoods under District 7. We noticed that some ZIP codes are shared with neighborhoods outside the four that fall under Councilor Anderson's jurisdiction. To solve this problem, we used the neighborhood column to further filter the data. First, we cleaned some of the neighborhood names for consistency, lowercased them, and removed leading spaces.

Now that the data is consistent and in string format, we filtered it for the four neighborhoods, giving us a dataset containing the relevant zip codes under the jurisdiction of the councilor.

2.2. RentSmart dataset

To use the RentSmart dataset, we cleaned it using the following methods that align with our requirements for analysis. While keeping in mind the association among the data, we were able to find the optimal way to display the dataset as follows:

- The date column in the original dataset contains a long list of specific hours, minutes, and seconds of each record. To understand the data more efficiently, we removed those elements from the date and only retained the year, month, and day.
- In the description column, we also noticed that a few descriptions detailing the violation types are extremely similar in terms of their meaning, upper and lower cases, and thus, those descriptions were grouped together. For example, "Work without permit" and "Work w/o Permit" were grouped as "Work Without Permit."
- After downloading the dataset and opening it as a CSV file, we found that each zip code was missing 0 in the front, and each year_built and year_remodeled was followed by .0 at the end. Therefore, we reformatted them to add necessary and remove unnecessary information.
- To ensure the accuracy of our analysis for District 7, we filtered the zip code and neighborhood columns. Similar to what we have done for the income-restricted housing dataset, we observed that filtering the data with zip codes still includes other areas not in the District 7 region. Therefore, we further defined the data by specifying the neighborhoods on top of zip codes.

2.3 Boston Census Dataset

In terms of preliminary data cleaning and preprocessing for this dataset, we did the following:

- Selected the relevant zip-codes that we concluded as belonging to District 7, and downloaded the housing data from the Boston Census website for the years 2011 - 2022 (all available years). We expected a single dataset, however, we ended up with 12 different datasets for each year, each with over 500 columns

- Upon closer inspection, each dataset column primarily had 4 sub-columns with information about the estimate of the value for that column, the margin of error, the percentage, and the percentage margin of error. For our preliminary pre-processing, we focused on the estimate and percentage-of-whole columns.
- The column names were quite messy and not easy to understand - unlike the Boston Census website, where information was quite clear. For example, the column containing information about the estimated Total Number of Housing Units in an area, was named 'Estimate!!HOUSING OCCUPANCY!!Total housing units.' We are working to get this into a readable format, and then apply the changes over each separate dataset
- To perform a year-wise comparison across different zip-codes, we need a single dataframe with information for each year and each zip-code. This is now our focus for this dataset - to concatenate the sub-datasets into a single dataset, but where there is a clear difference between values for each year and for each zip code.

3. Difficulties and Challenges

3.1. District 7 Jurisdiction

We are having some difficulty determining from the **income-restricted housing** dataset that we received which projects are in District 7 since the area is given by neighborhood. For projects where the neighborhood is South End in particular, we are in the process of determining the best way to extract only projects in the part of South End that belongs to District 7. We suspect that the ZIP code provided in the ZIP code column for projects in South End may be incorrect because South End should only have one ZIP code, but there are many different ZIP codes provided for projects in South End in the dataset.

We have reached out to the research guide librarian at Mugar Memorial Library, Lucy Flamm, who owns the guide for researching Boston's data and statistics. She recommended we use the Find My Councilor tool to determine which district the ambiguous addresses in South End belong to. For any address where the district cannot be determined by the tool, we will look at the Massachusetts Interactive Property Map and compare it to Boston's precinct map to determine the district. Fortunately, this only needs to be done for projects that are in South End, so it is feasible by hand.

A CSV file containing only restricted-income housing projects in District 7 is now on GitHub, but it isn't used as a basis dataset for cleaning and preprocessing for this milestone due to the time it takes to do the manual verification. We plan to use this new file as the basis for all restricted housing related analysis going forward.

4. Insufficient Data

We are lacking an efficient way to determine whether an address in South End is a part of District 7. For income-restricted housing, the verification can be done by manually searching for a street address and then using Find My Councilor tool, but for bigger datasets like RentSmart, it is impossible to do it by hand. We think that reaching out to the Boston city's data team for assistance might be good, as Lucy Flamm also recommended to do in case they have those data stored privately.