Affordable Housing Snapshot (Housing Affordability Index) - Team 2

Deliverable 3 (v1 Draft Complete Report)

1. Data Collection and Preprocessing

Following are the central datasets used for this project, along with a brief description:

1. **Rentsmart Dataset:** As the name suggests, Rentsmart is a dataset consisting of property violation details along with details of the owner, the year the house was built in, property type, and location details. It is a dataset that allows a new tenant to smartly find a new house to rent based on all the relevant details for a particular house. The Rentsmart dataset is central to this problem statement since it enables us to examine almost every house in Boston based on various parameters, including the different violations in the houses.

	date	violation_type	description	address	neighborhood	zip_code	parcel	owner	year built	year remodeled	property_type	latitude	longitude
0	2022-11-20T00:00:00	Enforcement Violations	Improper storage trash: res	325-327 Dorchester St, 02127	South Boston	2127	700214040	THREE-25 -327 DORCHESTER ST CONDO TR	1928.0	2010.0	Condominium Main*	42.33124	-71.05411
1	2022-11-20T00:00:00	Enforcement Violations	Improper storage trash: res	7 Aberdeen St, 02215	Boston	2215	2100139000	SEVEN 21 ABERDEEN STREET	1999.0	1999.0	Condominium Main*	42.34642	-71.10403
2	2022-11-20T00:00:00	Enforcement Violations	Improper storage trash: res	62 H St, 02127	South Boston	2127	603197000	SIXTY-2 H STREET CONDO TR	1890.0	2009.0	Condominium Main*	42.33589	-71.04138
3	2022-11-19T00:00:00	Enforcement Violations	Occupying City prop wo permit	9 Anderson St, 02114	Boston	2114	502218000	EMERALD REALTY CAPITAL LLC MASS LLC	1899.0	2014.0	Residential 7 or more units	42.36076	-71.06799
4	2022-11-19T00:00:00	Enforcement Violations	Improper storage trash: res	9 Anderson St, 02114	Boston	2114	502218000	EMERALD REALTY CAPITAL LLC MASS LLC	1899.0	2014.0	Residential 7 or more units	42.36076	-71.06799

Fig 1. First 5 Rows of the Rentsmart Dataset

```
Index(['date', 'violation_type', 'description', 'address', 'neighborhood',
       'zip_code', 'parcel', 'owner', 'year built', 'year remodeled',
       'property_type', 'latitude', 'longitude'],
      dtype='object')
```

Fig 2. Column Headings of the Rentsmart Dataset

2. **311 Service Requests Dataset:** The 311 Service Requests dataset consists of all Open and Closed service requests in different properties of Boston city. This dataset is essential because it also gives the area information (e.g. Roxbury, Allston, etc.), which is a key demographic we are trying to exploit in our extension project. The 311 Service Requests dataset gives vital

information such as a clear description of the complaint/issue, the department addressing the issue, if the case is overdue, etc.

	case_enquiry_id	open_dt	target_dt	closed_dt	ontime	case_status	closure_reason	case_title	subject	reason	...	police_district	neighbor
0	101004143507	2022-01-22 08:14:00	2022-01-25 08:30:00	NaN	OVERDUE	Open		Pothole	Public Works Department	Highway Maintenance	...	B2	Rc
1	101004232225	2022-03-24 12:36:00	2022-04-14 12:36:52	2022-03-24 12:37:41	ONTIME	Closed	Case Closed Case Invalid	Request for Recycling Cart	Public Works Department	Recycling	...	B3	Dorcl
2	101004194256	2022-02-18 16:48:00	2023-02-18 16:48:41	NaN	ONTIME	Open		Boston Bikes: Bike Racks; Request	Transportation - Traffic Division	Boston Bikes	...	D14	E
3	101004202046	2022-02-26 14:30:00	2022-03-01 08:30:00	NaN	OVERDUE	Open		Request for Pothole Repair	Public Works Department	Highway Maintenance	...	D14	Al Bri
4	101004151776	2022-01-30 11:17:00	NaN	2022-01-30 11:31:43	ONTIME	Closed	Case Closed Case Noted	PublicWorks: Compliment	Mayor's 24 Hour Hotline	Employee & General Comments	...	E5	West Rc

5 rows × 29 columns

Fig 3. First 5 Rows of the 311 Service Requests Dataset

```
Index(['case_enquiry_id', 'open_dt', 'target_dt', 'closed_dt', 'ontime',
       'case_status', 'closure_reason', 'case_title', 'subject', 'reason',
       'type', 'queue', 'department', 'submittedphoto', 'closedphoto',
       'location', 'fire_district', 'pwd_district', 'city_council_district',
       'police_district', 'neighborhood', 'neighborhood_services_district',
       'ward', 'precinct', 'location_street_name', 'location_zipcode',
       'latitude', 'longitude', 'source'],
      dtype='object')
```

Fig 4. Column Headings of the 311 Service Requests Dataset

3. **<u>Building and Property Violations Dataset:</u>** Similar to the 311 Service Requests dataset, the Building and Property Violations dataset consists of building code violations across different regions of Boston city. This dataset gives us a more Building Violations codes view as opposed to the Service Complaints view in the Rentsmart and 311 Service Requests datasets.

	case_no	status_dttm	status	code	value	description	violation_stno	violation_sthigh	violation_street	violation_suffix	...	ward	contact_addr1	co
0	HVIOL-644882	2022-11-18 09:50:19	Open	CMR410.550 B	NaN	Extermination of Insects, Rodents and Skunks -...	19	NaN	Adams	ST	...	15	77 Pond Ave #401	
1	HVIOL-644882	2022-11-18 09:50:19	Open	CMR410.482 A	NaN	Smoke Detectors & Carbon Monoxide Alarms - Own...	19	NaN	Adams	ST	...	15	77 Pond Ave #401	
2	HVIOL-644882	2022-11-18 09:50:19	Open	CMR410.482 A	NaN	Smoke Detectors & Carbon Monoxide Alarms - Own...	19	NaN	Adams	ST	...	15	77 Pond Ave #401	
3	HVIOL-644882	2022-11-18 09:50:19	Open	CRM410.550 D	NaN	Extermination of Insects, Rodents and Skunks -...	19	NaN	Adams	ST	...	15	77 Pond Ave #401	
4	HVIOL-644882	2022-11-18 09:50:19	Open	CMR410.550 B	NaN	Extermination of Insects, Rodents and Skunks -...	19	NaN	Adams	ST	...	15	77 Pond Ave #401	

Fig 5. First 5 Rows of the Building and Property Violations Dataset

```
Index(['case_no', 'status_dttm', 'status', 'code', 'value', 'description',
       'violation_stno', 'violation_sthigh', 'violation_street',
       'violation_suffix', 'violation_city', 'violation_state',
       'violation_zip', 'ward', 'contact_addr1', 'contact_addr2',
       'contact_city', 'contact_state', 'contact_zip', 'sam_id', 'latitude',
       'longitude', 'location'],
      dtype='object')
```

Fig 6. Column Headings of the Building and Property Violations Dataset

The following datasets play a central role in the extension project:

4. **<u>Boston Neighborhood Dataset:</u>** The Boston Neighborhood dataset is Boston city's Census data containing the total number of people residing in each Boston neighborhood. It also classifies the population in each neighborhood on key demographics such as race, age, profession, etc. This dataset helps us better understand any potential exploitation of social vulnerability by bad landlords in each Boston neighborhood.

	field concept	Total:	White alone	Black or African American alone	Hispanic or Latino	Asian, Native Hawaiian and Pacific Islander alone, all ages	Other Races or Multiple Races, all ages	Total:	White alone	Black or African American alone	...	Nursing facilities/Skilled-nursing facilities	Other institutional facilities	Noninstitutionalized population:	College/U student
1	Allston	28621	14634	1451	3657	7173	1706	26668	14022	1294	...	26	0	3364	
2	Back Bay	19588	14056	718	1326	2604	884	18374	13296	669	...	269	0	1701	
3	Beacon Hill	9336	7521	252	537	630	396	8603	6980	231	...	0	0	33	
4	Brighton	48330	30596	2289	4978	7801	2666	44129	28706	1966	...	240	56	3713	
5	Charlestown	19120	13626	990	2075	1650	779	15661	11689	662	...	55	0	55	

Fig 7. First 5 Rows of the Boston Neighborhood Dataset

```
Index(['field concept', 'Total:', 'White alone',
       'Black or African American alone', 'Hispanic or Latino',
       'Asian, Native Hawaiian and Pacific Islander alone, all ages',
       'Other Races or Multiple Races,  all ages', 'Total:', 'White alone',
       'Black or African American alone', 'Hispanic or Latino',
       'Asian, Native Hawaiian and Pacific Islander alone, aged 18+',
       'Other Races or Multiple Races, aged 18+', 'Total:, aged 0-17',
       'White alone, aged 0-17', 'Black or African American alone, aged 0-17',
       'Hispanic or Latino, aged 0-17',
       'Asian, Native Hawaiian and Pacific Islander alone, aged 0-17',
       'Other Races or Multiple Races, aged 0-17', 'household population',
       'Total:', 'Institutionalized population:',
       'Correctional facilities for adults', 'Juvenile facilities',
       'Nursing facilities/Skilled-nursing facilities',
       'Other institutional facilities', 'Noninstitutionalized population:',
       'College/University student housing', 'Military quarters',
       'Other noninstitutional facilities', 'Total:', 'Occupied', 'Vacant',
       'household size'],
      dtype='object', name=0)
```

Fig 8. Column Headings of the Boston Neighborhood Dataset

5. **<u>Climate Ready Boston Social Vulnerability Dataset:</u>** The Climate Ready Boston Social Vulnerability dataset showcases the total number of vulnerable people (based on their age, gender, race, English proficiency, income class, disability, and medical illness). This dataset, combined with the Boston Neighborhood dataset, gives us a chance to analyze the percentage of population vulnerable in each Boston neighborhood and then find the correlation between a high presence of socially vulnerable groups in Boston and a high percentage of bad landlords in a neighborhood.

	FID	GEOID10	AREA_SQFT	AREA_ACRES	POP100_RE	HU100_RE	TotDis	TotChild	OlderAdult	Low_to_No	LEP	POC2	MedIllnes	Name	Shape_
0	1	25025010405	3914567.54	89.8661	5522	994	470	60	331	1191	1522	1755	2131.22	Mission Hill	666100.0
1	2	25025010404	1472713.92	33.8089	5817	1862	299	77	56	2387	2443	1749	2201.14	Fenway	250612.3
2	3	25025010801	1376667.12	31.6039	2783	1899	84	281	390	72	462	447	1214.76	Back Bay	234357.9
3	4	25025010702	3228780.12	74.1226	2400	1643	45	86	285	187	472	320	1014.20	Back Bay	549614.0
4	5	25025010204	2741497.18	62.9361	3173	1283	131	13	36	895	931	1039	1181.78	Fenway	466585.2

Fig 9. First 5 Rows of the Climate Ready Boston Social Vulnerability Dataset

```
Index(['FID', 'GEOID10', 'AREA_SQFT', 'AREA_ACRES', 'POP100_RE', 'HU100_RE',
       'TotDis', 'TotChild', 'OlderAdult', 'Low_to_No', 'LEP', 'POC2',
       'MedIllnes', 'Name', 'Shape__Area', 'Shape__Length'],
      dtype='object')
```

Fig 10. Column Headings of the Climate Ready Boston Social Vulnerability Dataset

All these datasets were either provided to us in the Project Description or collected. Based on the type of data, all the datasets were cleaned and preprocessed. The following steps were taken to preprocess the data:

- **Dealing with NaN (Missing) Values:** Columns which had over 50-60% missing values were dropped from the dataset. If a column was significantly important, then the missing rows were filled with the mean value (for continuously-valued columns) and with the highest occurring value (for discrete-valued columns). If the number of rows with NaN values were low (3-5%), then these rows were simply dropped.
- **Data Normalization:** All the continuously real-valued columns were normalized using z-standardization (subtracting each value by the mean of the column and then dividing the result by the standard deviation of the column).
- **Data (Attribute) Selection:** For each dataset, a few columns (such as ID) were not relevant to the information the data was representing. Hence, these columns were simply dropped from the database.
- **Outlier Analysis:** A thorough outlier analysis was performed on all the available data. However, there were not any outliers in any of the datasets that hindered the data analysis.

2. Data Visualization and Analysis

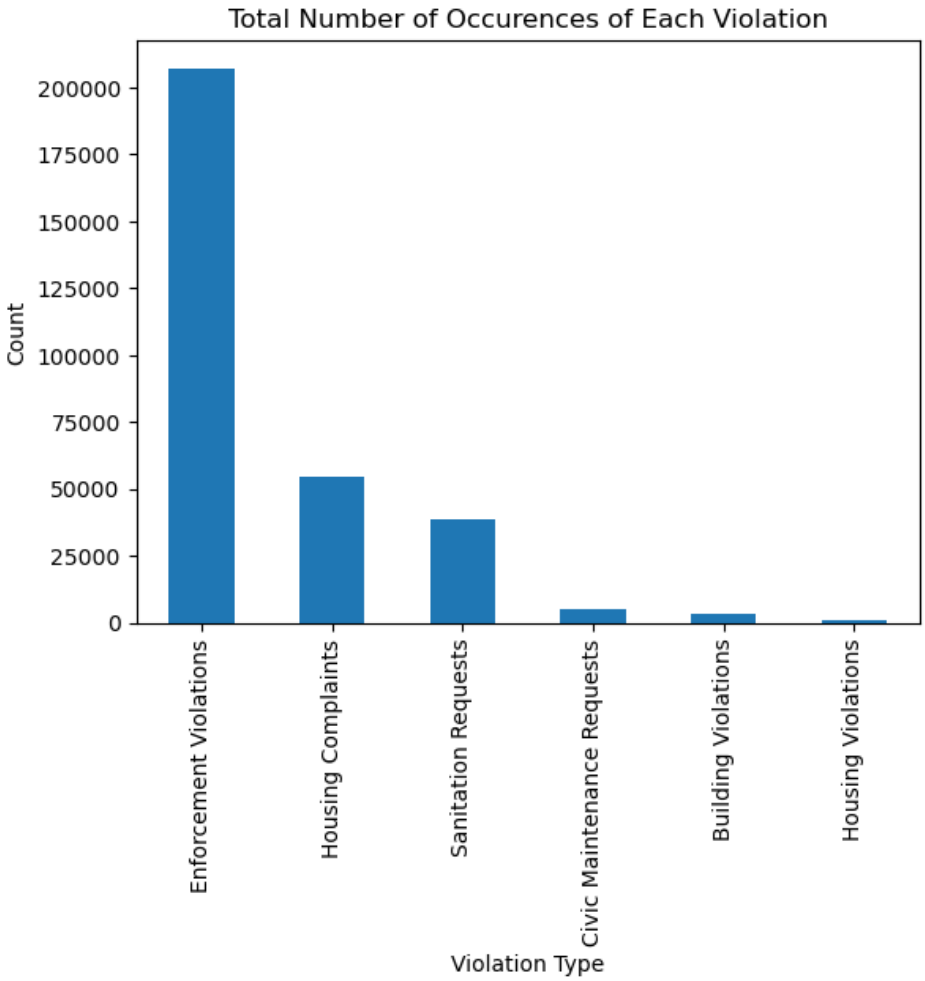

Fig 11. Number of Occurrences of Each Type of Violation

Fig 11 showcases the number of each type of violation that has occurred in Boston city (based on the Rentsmart dataset). Following are some of the Enforcement violations:

- Improper Storage Trash
- Unregistered Motor Vehicles
- Failure to Obtain Inspection
- Failed Multiple Rental Inspections
- Overgrown Weeds on the Property
- No Number on the Building
- Overfilling of a Dumpster

As can be seen from the above graph and list, even though the number of Enforcement Violations is very high in Boston city, these violations do not carry a lot of weight since they are not life-threatening violations.

However, if we look a little deeper into each of these violations, the three main violations that need to be addressed by the landlord right away (i.e. violations that are either life-threatening or violations that make a property inhabitable) are Sanitation Requests (Rodent Activity, Abandoned Building, Rat Bite, Mosquitoes, Pigeon Infestation), Building Violations (Unsafe Structures, Emergency Escape, and Rescue, Minimum Number of Exits), and Housing Complaints (Mice Infestation, Illegal Occupancy, Pest Infestation, Illegal Rooming House, Bed Bugs, Sewage/Septic Backup, Overcrowding). Landlords who ignore these complaints when compared to the other types of violations must be penalized more and deemed “Bad Landlords”.

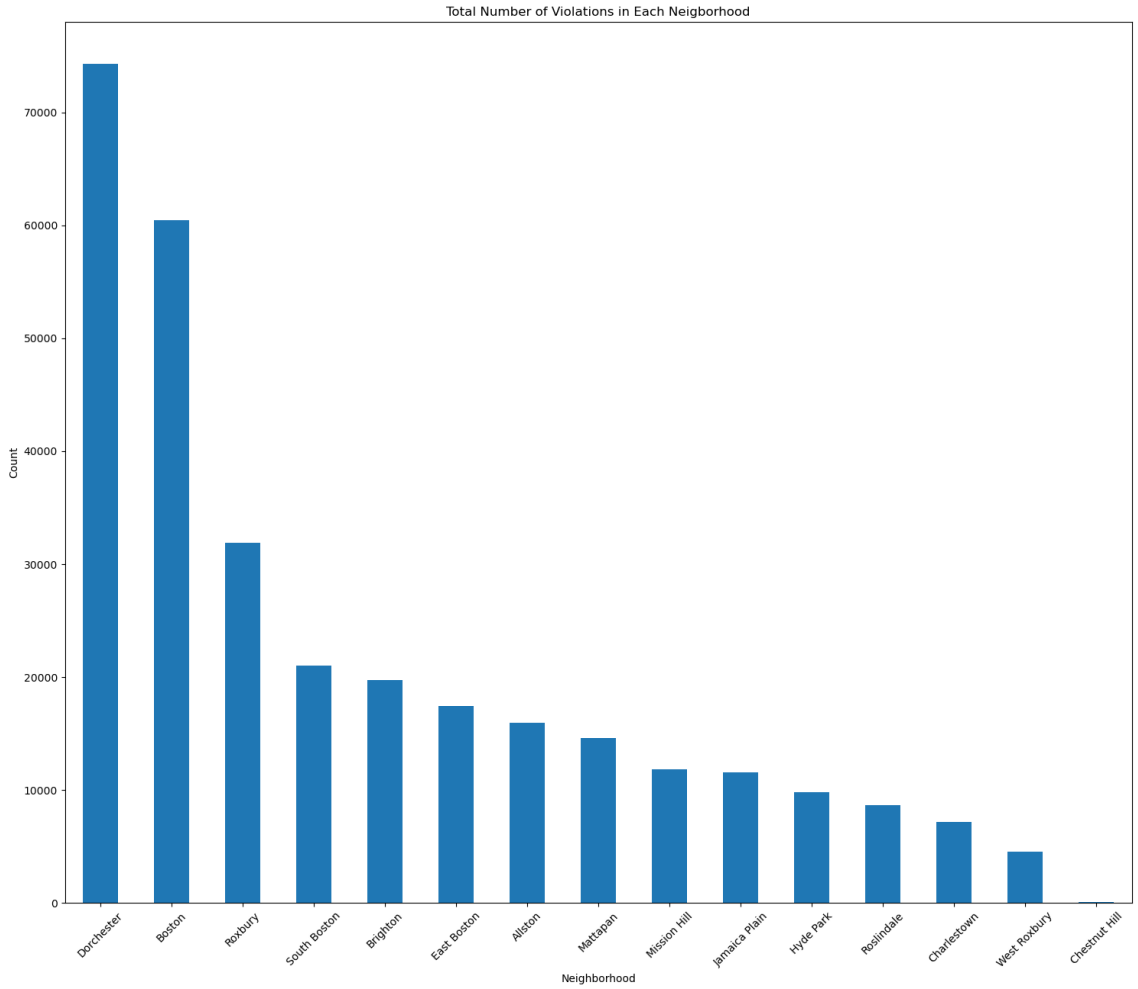

Fig 12. Number of Violations per Boston Neighborhood

The above graph shows the number of violations in each major Boston neighborhood. Dorchester appears to be the words neighborhood of all, whereas there are negligible number of violations in Chestnut Hill. However, if we only look at the three major types of violations (mentioned above), there are a few neighborhoods that improve and a few that worsen.

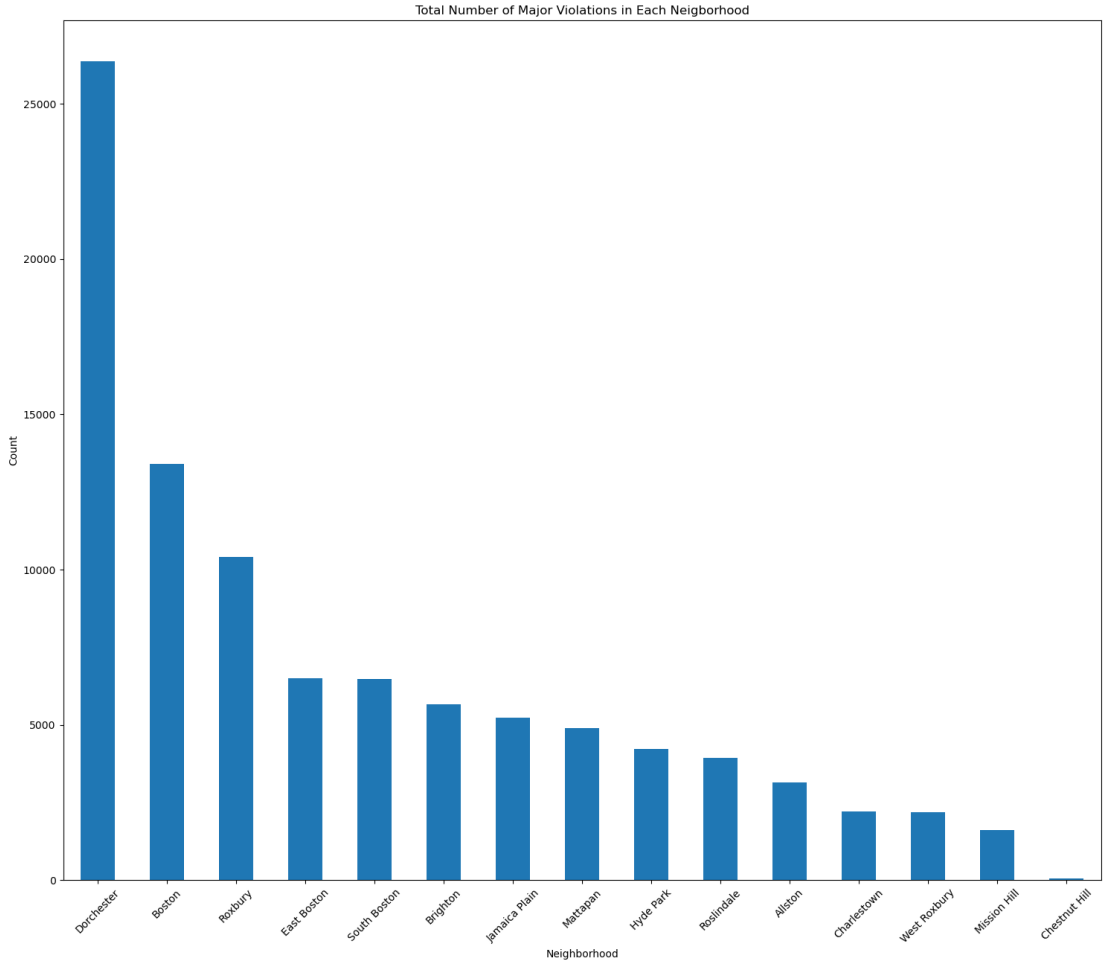

Fig 13. Number of Major Violations per Boston Neighborhood

Dorchester is still the worst neighborhood in Boston city. However, East Boston worsens as a neighborhood, whereas Allston and Mission Hill significantly improve. This clearly implies that we also need to look into these major violations while deeming a particular landlord as bad instead of just looking at the number of violations they have.

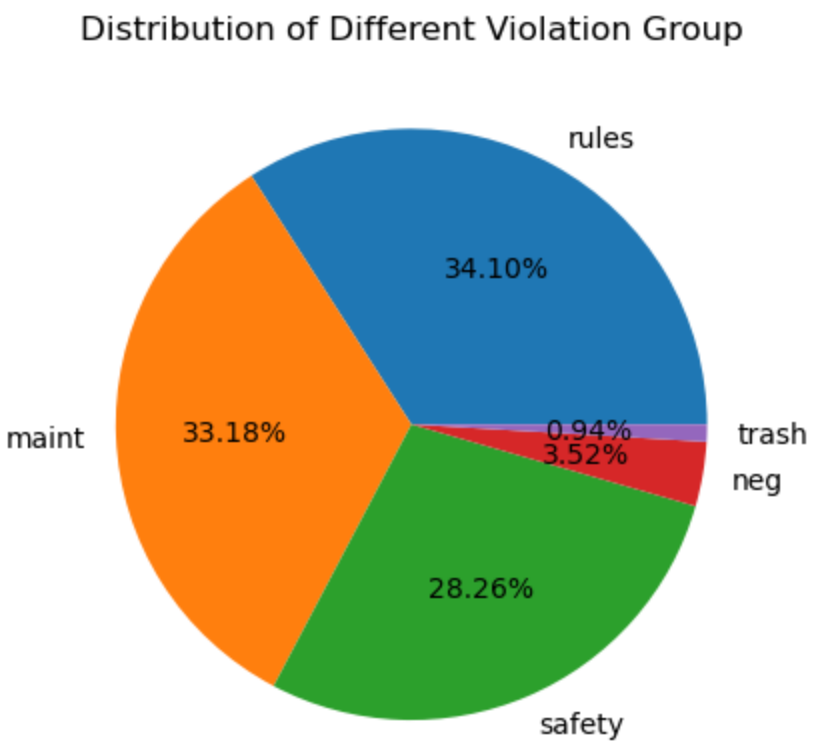

Fig 14. Percentage of Different Types of Violations in Boston City

The above pie chart depicts the percentage of each type of violation (with a one-word description of what each type of violation is associated with). Rules and Maintenance are the two most common violations in Boston city. Safety (such as unsafe building structure) is a violation that shall be addressed right away. However, it is also a widely occurring violation in the city. Trash-related violations are the least occurring in the city.

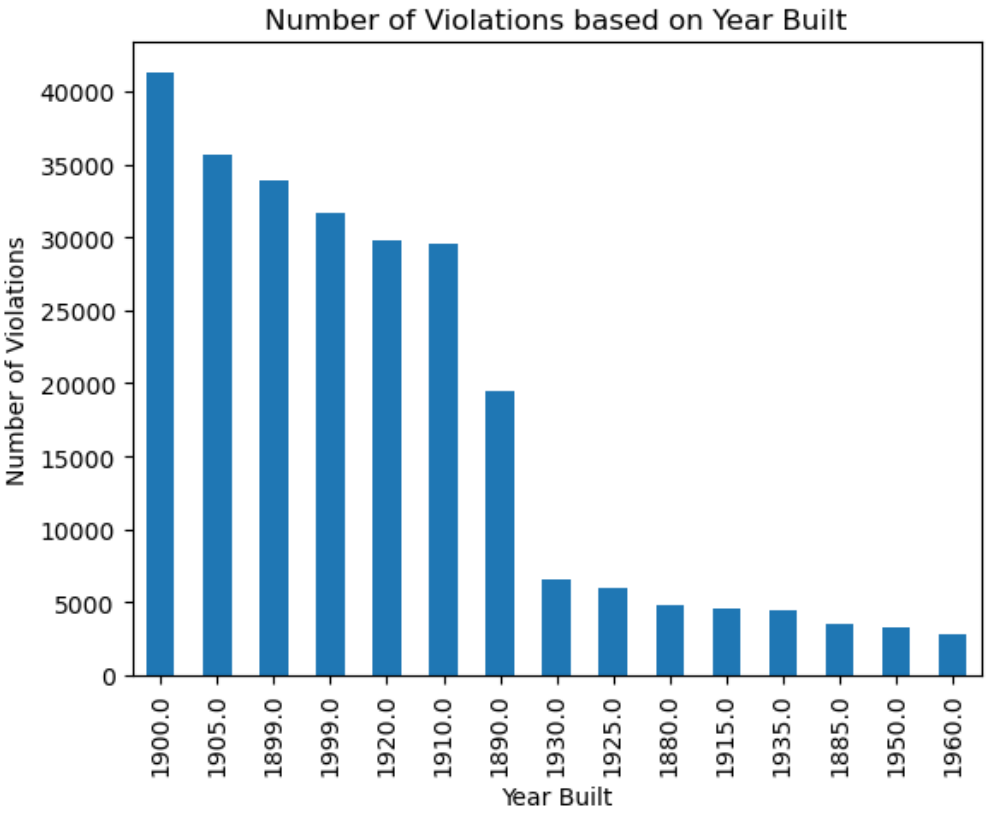

Fig 15. Highest Number of Violations based on the Year a House was Built in

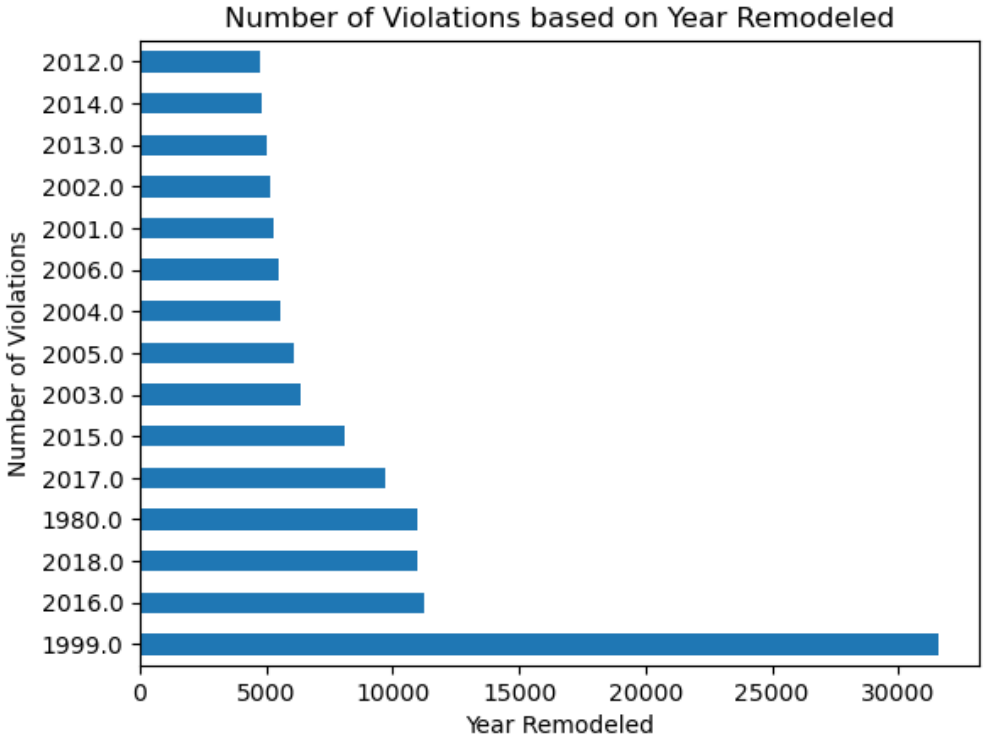

Fig 16. Highest Number of Violations based on the Year a House was Remodeled in

The above two bar graphs show the top 15 years with highest number of violations. Here, years means year built (in Fig 15.) and year remodeled (in Fig 16.). It is noteworthy that the number of violations in houses remodeled in the year 1999 are more than double of the year, with the second highest number of violations (2016). This might also be due to the fact that many houses were remodeled in the year 1999 in Boston. The maximum number of violations based on year built are also in the year range of 1899 - 1950, which signifies the fact that the majority of houses built in Boston city were during this time period.

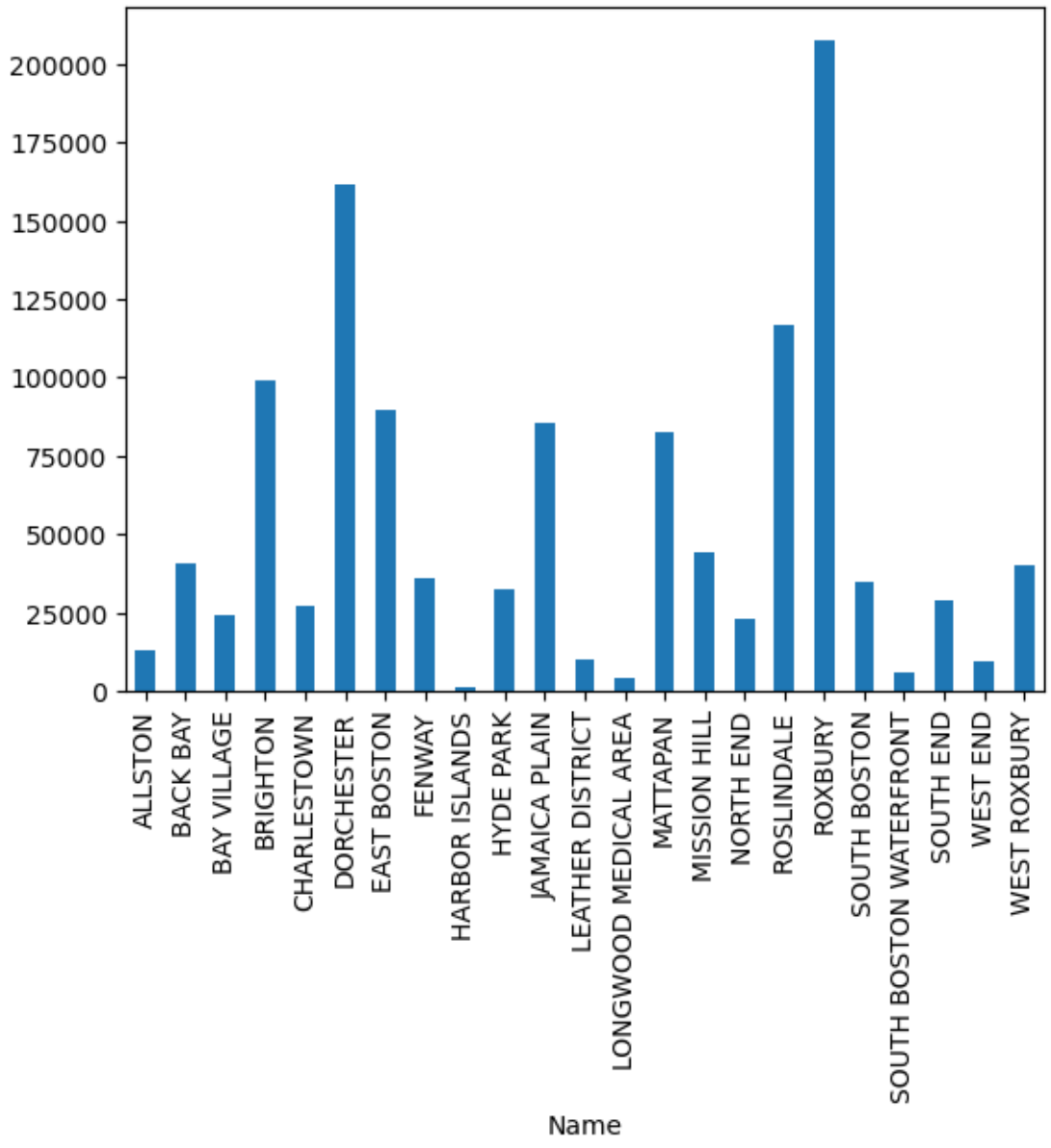

Fig 17. Total Vulnerable Population per Neighborhood in Boston City

Fig 17 depicts the total vulnerable population in different neighborhoods of Boston city. Roxbury is home to the largest vulnerable population in the city, followed by Dorchester and Roslindale. Places like Longwood Medical Area, West End, Harbor Islands, and South Boston Waterfront host the least number of vulnerable people. Two major areas in the city (Brighton and Allston) host about 100000 and 12500 vulnerable people, respectively.

[TODO] Add more visualizations and analysis to this section

3. Answering Key Questions

Main Questions

1. How to determine if a landlord is a bad landlord?
   Answer:
   As mentioned earlier, there are main types of violations occurring across Boston city. Out of these, Housing Complaints, Sanitation Requests, and Building Violations are the worst violations of all.

   Hence, while trying to classify landlords as bad and good, we should not look at the number of complaints or violations that a landlord makes but at the number of violations in each category and then give the above violations more weight.

   Then, if a landlord has many violations but very few violations in the above-given classes, whereas another landlord has fewer violations with majority of violations in the above classes, the second landlord must be deemed the worse landlord.

   [TODO] Improve upon this description of bad landlord with a few insights and more parameters

2. Who is causing the problems?
   Answer:
   [TODO]

3. What factors are correlated with the violations?
   Answer:
   [TODO]

Basic Questions

1. What are the types of violations? How many violations for each type?

Answer:

There are mainly five types of violations. The following table shows the type of violation and the total number of occurrences of that violation in Boston city:

Table 1. Types of Violations and their respective Number of Occurences

Sr No.	Violation	Number of Occurences
1	Enforcement Violations	207180
2	Housing Complaints	54230
3	Sanitation Requests	38806
4	Civic Maintenance Requests	4947
5	Building Violations	3279
6	Housing Violations	666

2. Who has the most violations? Any punishments?

Answer:

[TODO]

3. Where are the violations in Boston?

Answer:

As we can see from Figure 12, Dorchester is the Boston neighborhood with the maximum number of violations. Roxbury is another prominent area with a large chunk of violations in the city. On the other hand, the neighborhoods of West Roxbury and Chestnut Hill have the least violations in the whole city.

a. (Compare offcampus and on campus housing violation)

Answer:

[TODO]

b. (Compare student housing and non-student housing)

Answer:

[TODO]

c. Distribution of % of violations across Boston

Answer:

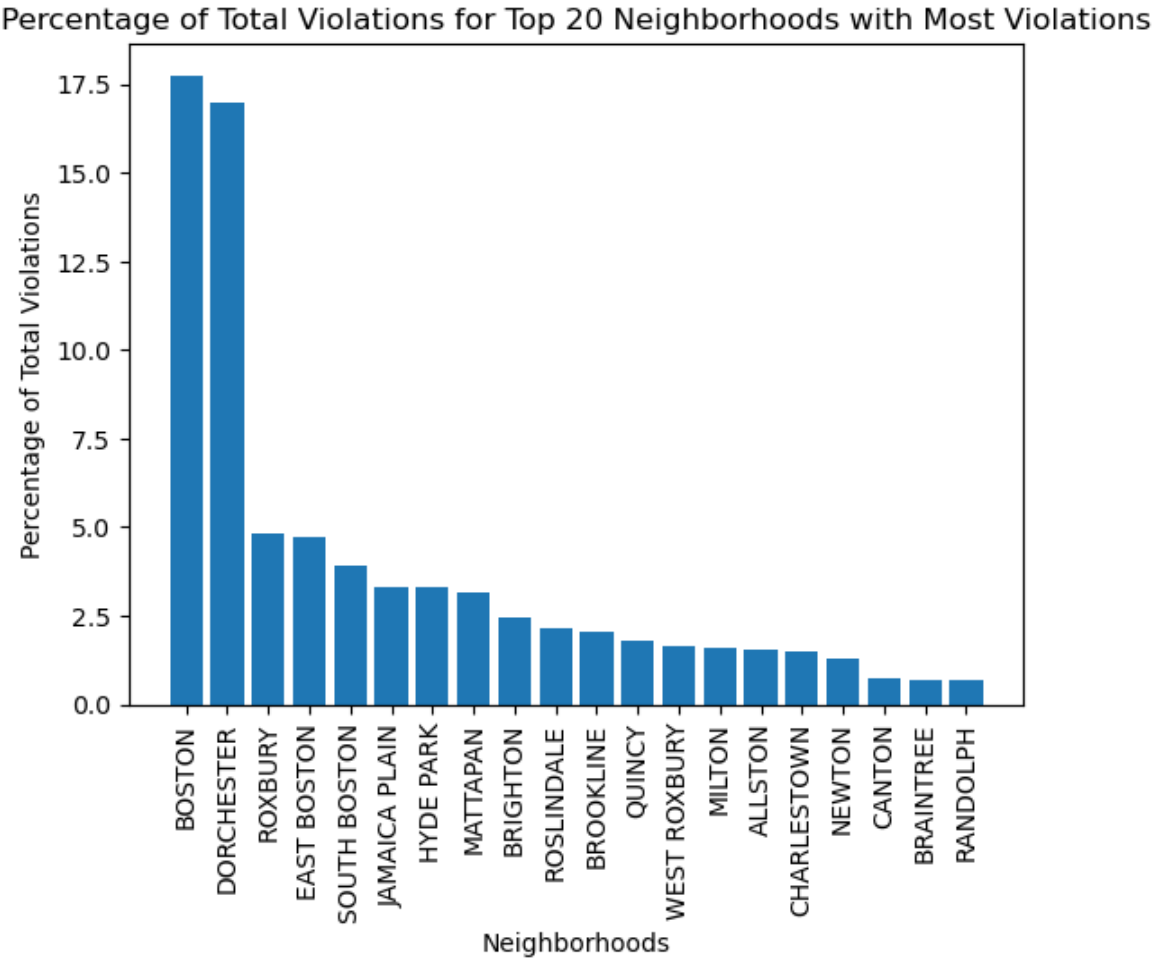

Fig 18. Percentage of Total Violation for Top 20 Neighborhoods

4. Who are the renters and landlords in the block groups? What are their demographics?

Answer:

[TODO]

5. Who are the landlords in the block groups? What are their demographics?

Answer:

[TODO]

6. How old are the buildings? How are the renovation status? How many property violations per year?

Answer:

The oldest house in Boston city was built in the year 1700, and the newest building in Boston (Rentsmart dataset) was built in 2019. Majority of the buildings were built before the year 1950.

Similarly, the first renovation in Boston occurred in the year 1900, and the latest renovation in Boston city occurred in the year 2019. Majority of the renovations in the city occurred after the year 1950.

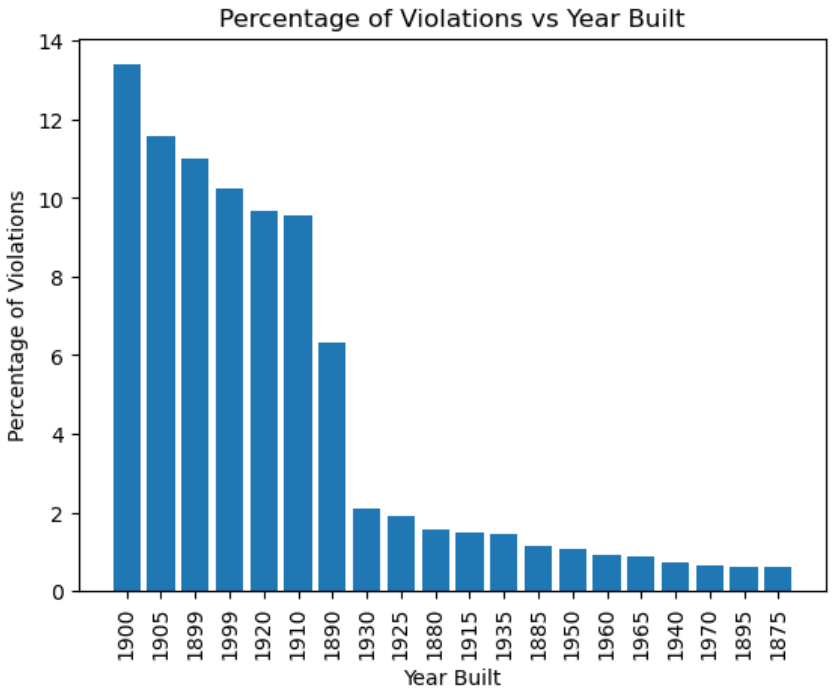

Fig 19. Percentage of Violation for each year

Fig 19 shows the Year Built of buildings against the percentage of violations in buildings built in that particular year. The buildings built in 1900 have the most number of violations.

7. Whether affordability affects the amount of property violations?

Answer:

[TODO]

4. Limitations of the Project

Following are a few limitations of this project:

- There is a lot of ambiguity and lack of cohesion among the datasets.
- One of the biggest challenges faced by us during initial preprocessing of the data was building violation codes. If these codes were better explained to us in some way, it would have been much easier to understand and analyze the data.
- The datasets provided in the Project Proposal cover a lot of bases when it comes to the project questions. However, these datasets cannot be merged together due to the lack of a common column in all datasets (like primary keys in SQL tables).
- The project involves multiple datasets coming from multiple sources. This usually creates a lot of confusion.
- Most of the datasets (including the Rentsmart dataset) were last updated in 2019. The pandemic has changed many key demographics related to housing and realty, which are not covered by the datasets we used in the project.
- Throughout the extension project, we look at certain Socially Vulnerable Groups in Boston city. Machine Learning and Data Science have conventionally shown bias when they are trained on racial data.
- The insights of this project could be used by landlords to exploit socially vulnerable groups (if they are not already) and worsen the situation of bad landlords in Boston city.
- Housing affordability was the main theme of the project before the pivot. We were interested in incorporating an Affordability vs Landlord Behaviour view to the project after the pivot. However, Property affordability data in Boston city is scarce.
- We always risk Data Violation and Data Security while handling huge amounts of data in Data Science.

5. Extension Project

As an extension to this project, we propose finding insights into the socially vulnerable groups in Boston and the correlation between the presence of socially vulnerable groups and bad behavior of landlords.

For this extension, we use the Boston Neighborhood Data (Census data consisting of population statistics in Boston city and classification of the population based on key demographics) and the Climate Ready Social Vulnerability Dataset.

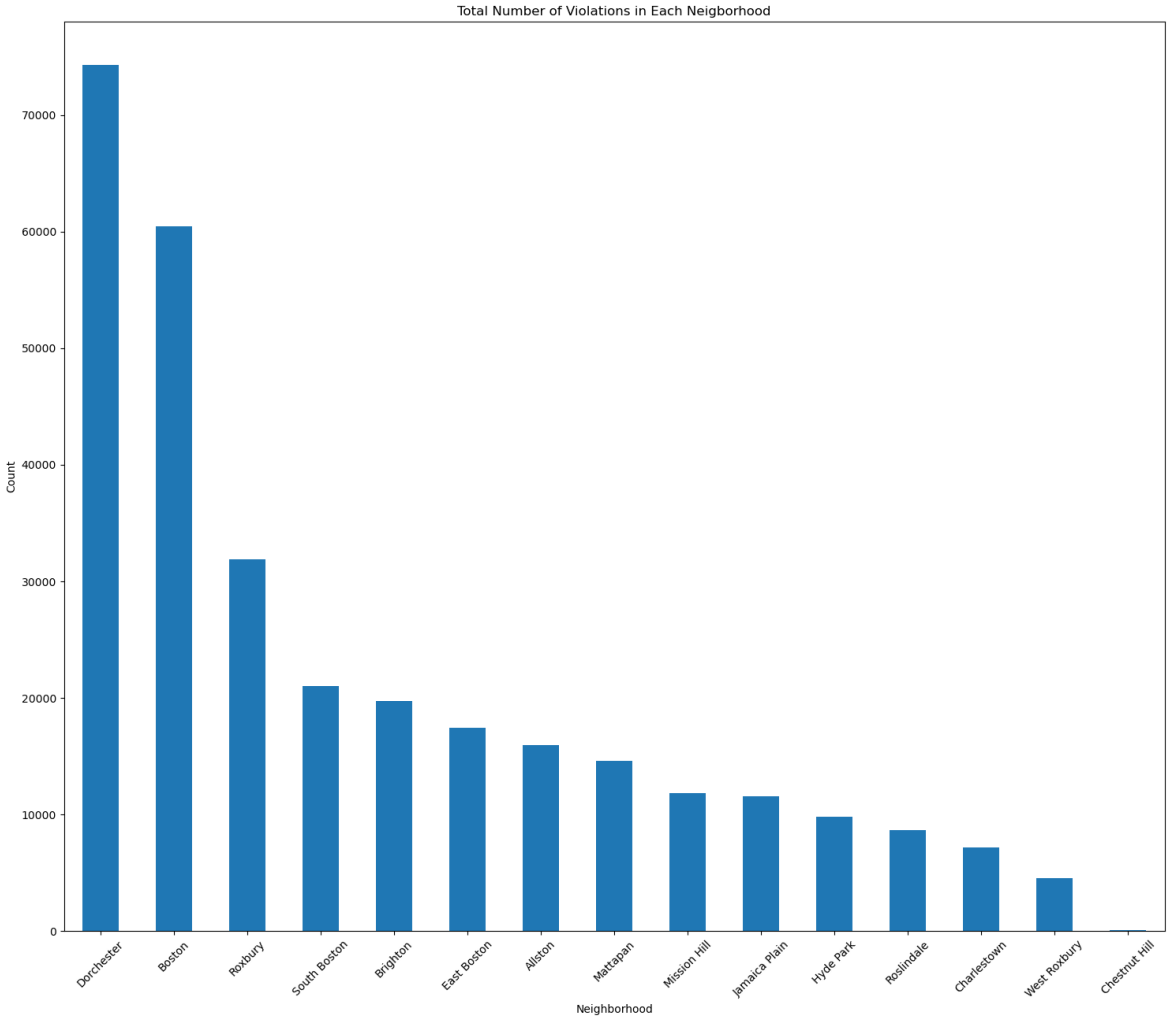

Fig 20. Number of Violations per Boston Neighborhood

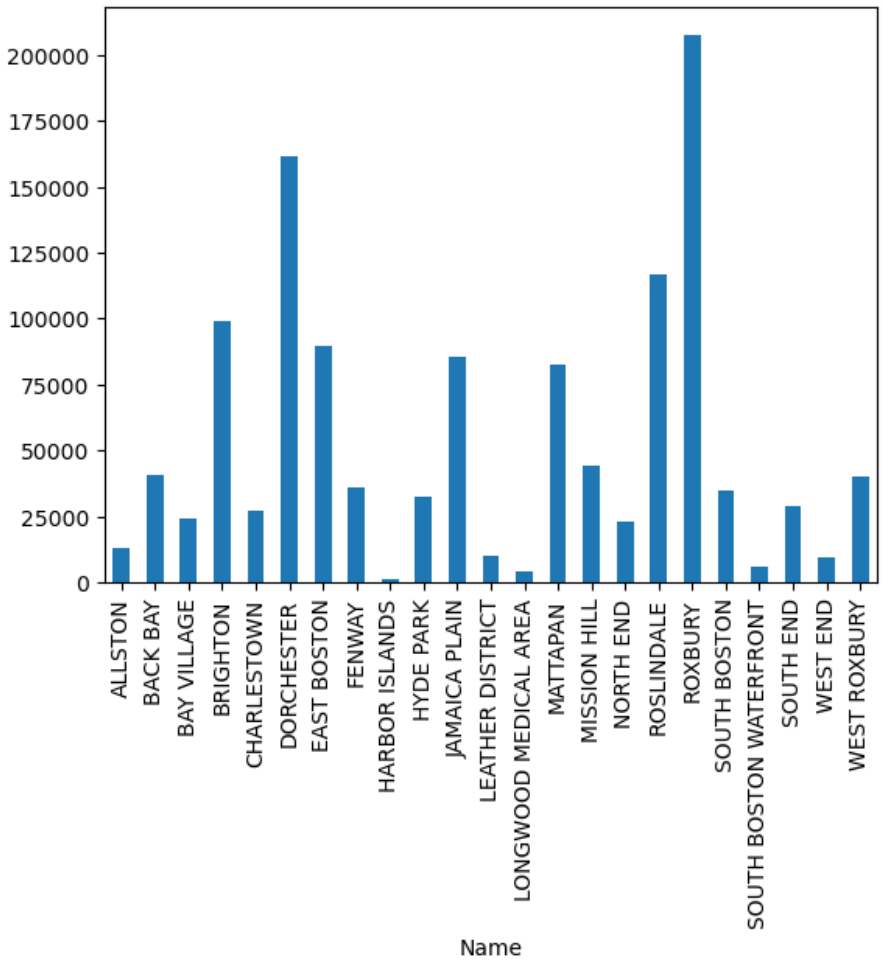

Fig 21. Total Vulnerable Population per Neighborhood in Boston City

The above graphs (also shown in the Data Visualization and Data Analysis part) showcase the number of violations per neighborhood and the total vulnerable population per neighborhood in Boston city.

The following insights can be drawn from these two graphs regarding the correlation between high vulnerable populations and number of violations per neighborhood:

- Roxbury has the highest vulnerable population in the city. Incidentally, Roxbury is the neighborhood with the third-highest number of violations in the city.
- Dorchester has the highest number of violations, and it is host to the second largest vulnerable population.

- West Roxbury has the second-lowest number of violations. West Roxbury also does not have a large vulnerable population when compared to other neighborhoods.
- Brighton is home to the fourth-largest vulnerable population and has the fifth-highest number of violations.

There are a few other trends that support the fact that there might be an underlying correlation between the presence of socially vulnerable populations and the bad behavior of landlords.

Fig 22 (shown below) depicts a bar graph with the ranking of each major Boston neighborhood with respect to the number of violations and the percentage of vulnerable population residing in the neighborhood. Out of the 12 neighborhoods shown in the graph, 9 neighborhoods (75%) have a close ranking in both the cases. This is a clear indication of the correlation we are trying to prove in our Extension Project.

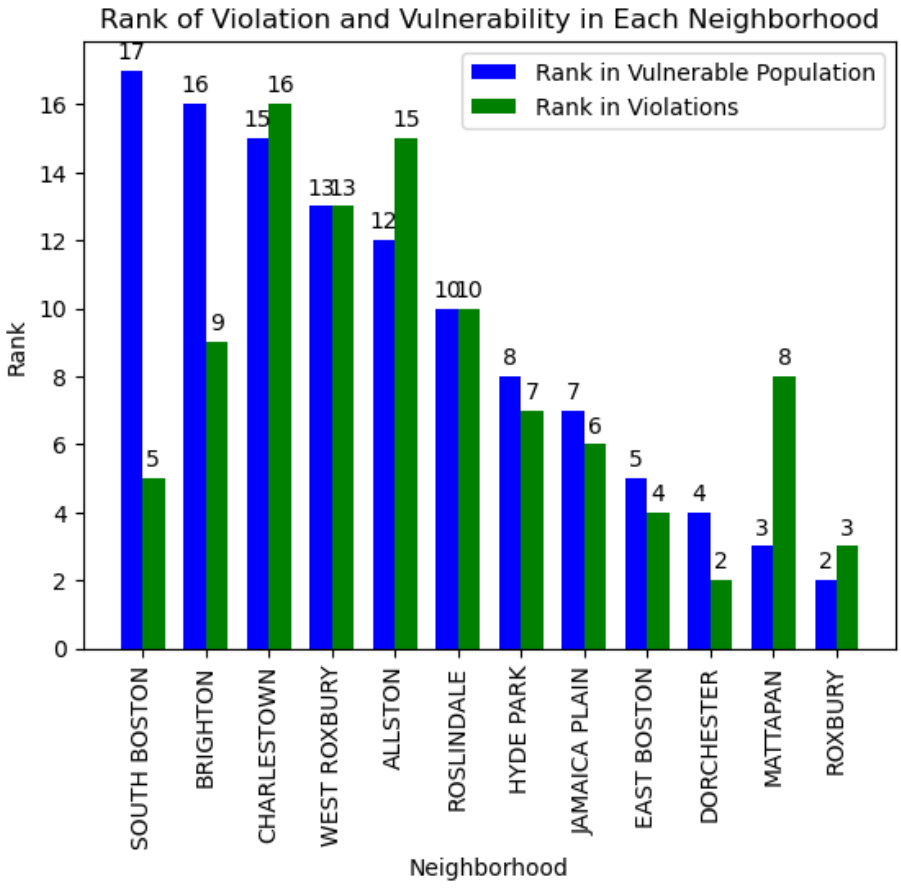

Fig 22. Neighborhood vs Rank for Vulnerable Population and Number of Violations

[TODO] Gain more insights into this relationship and add more visualizations to this section