Affordable Housing Snapshot (Housing Affordability Index) - Team 2

Deliverable 5 (Final Project Deliverable)

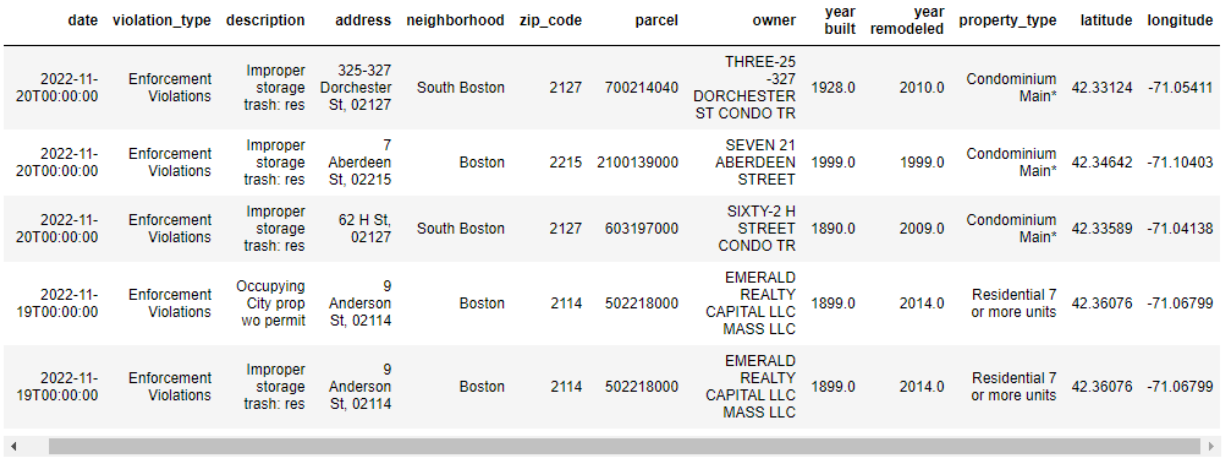
Aakash Bhatnagar
Sai Krishna Sashank Madipally
Zhiqi Jin
Tessa Sharma

INDEX

1. Data Collection and Preprocessing

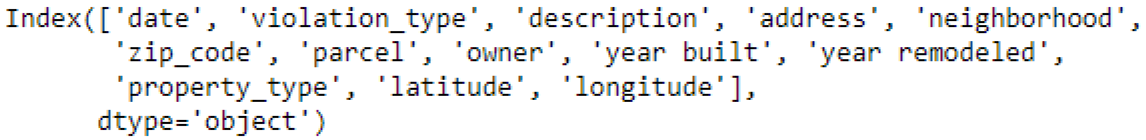
Following are the central datasets used for this project, along with a brief description:

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
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
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
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
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
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

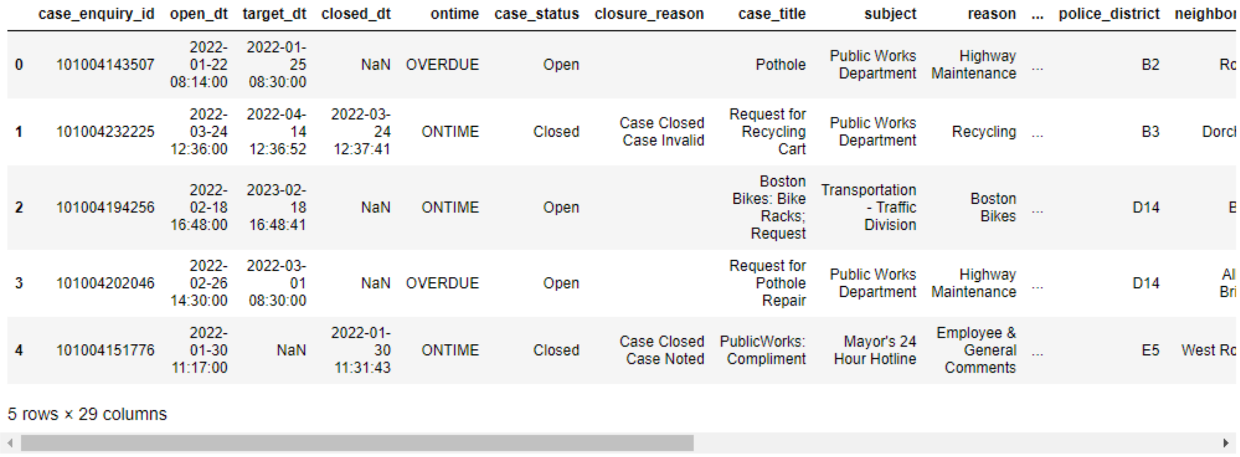


```
Index(['date', 'violation_type', 'description', 'address', 'neighborhood',
       'zip_code', 'parcel', 'owner', 'year built', 'year remodeled',
       'property_type', 'latitude', 'longitude'],
      dtype='object')
```

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

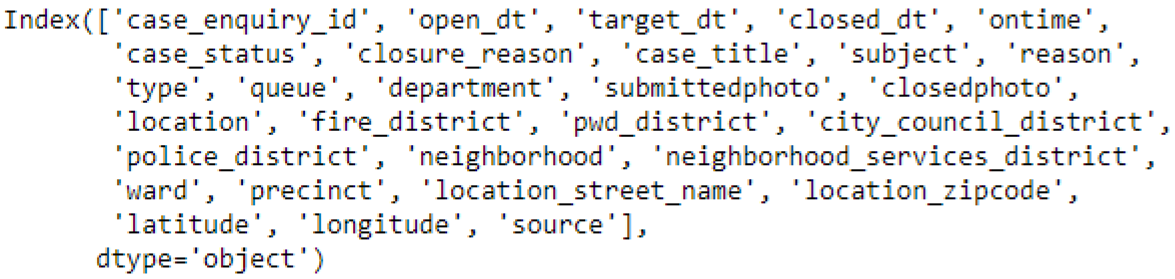
information such as a clear description of the complaint/issue, the department addressing the issue, if the case is overdue, etc.



	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	Ro
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	B
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 Ro

5 rows × 29 columns

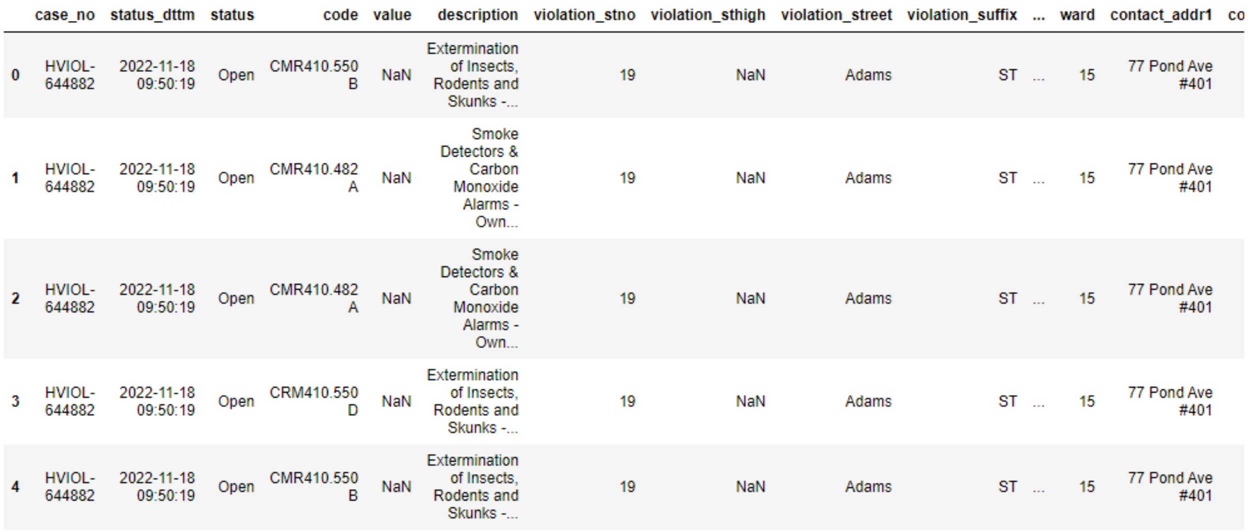
Fig 3. First 5 Rows of the 311 Service Requests Dataset



```
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')
```

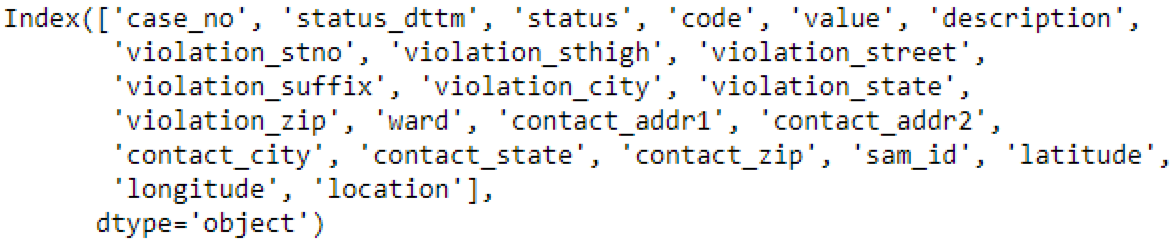
Fig 4. Column Headings of the 311 Service Requests Dataset

3. **Building and Property Violations Dataset:** 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

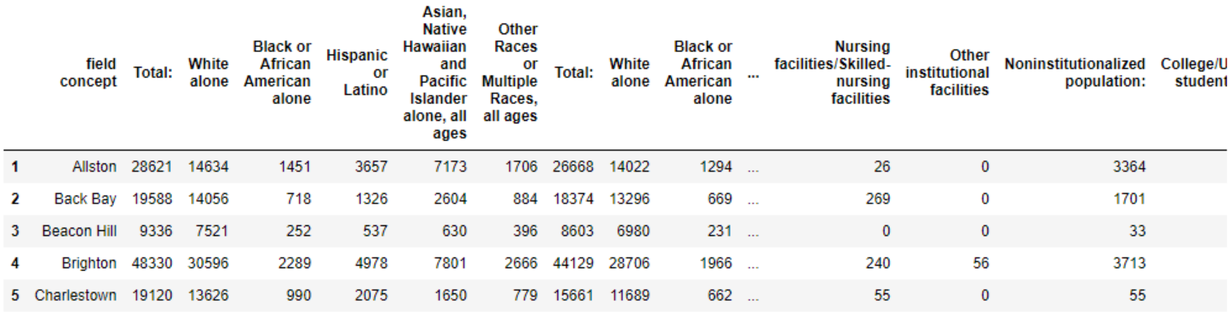


```
Index(['case_no', 'status_dttm', 'status', 'code', 'value', 'description',
       'violation_stno', 'violation_sthigh', 'violation_street',
       'violation_suffix', 'violation_city', 'violation_state',
       'violation_zip', 'ward', 'contact_addr1', 'contact_addr2',
       'contact_city', 'contact_state',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

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

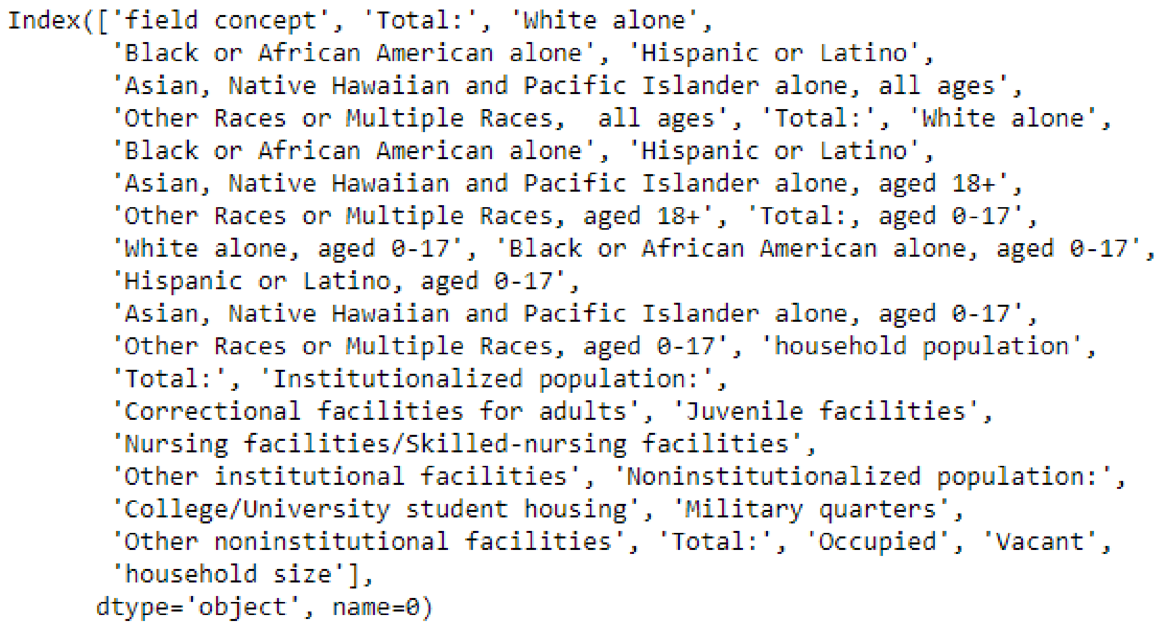
The following datasets play a central role in the extension project:

4. **Boston Neighborhood Dataset:** 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	

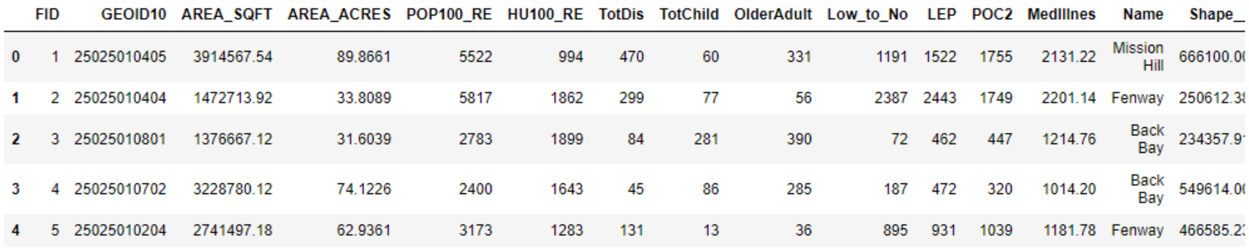
Fig 7. First 5 Rows of the Boston Neighborhood Dataset



```
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)
```

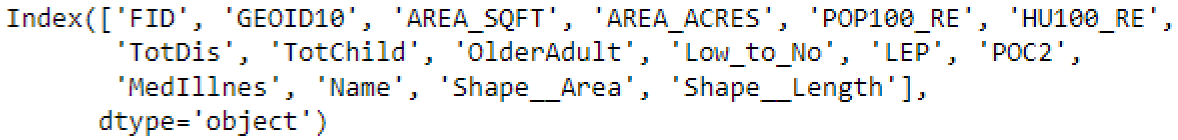
Fig 8. Column Headings of the Boston Neighborhood Dataset

5. **Climate Ready Boston Social Vulnerability Dataset:** 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.
- **Processing Textual Columns:** We also analyzed the text in the data. We performed some basic preprocessing, such as stripping the spaces off the sides of the text, capitalizing letters, and checking to see if there are no duplicates due to a single character's difference (usually, these are typos that occur during data entry).

2. Data Visualization and Analysis

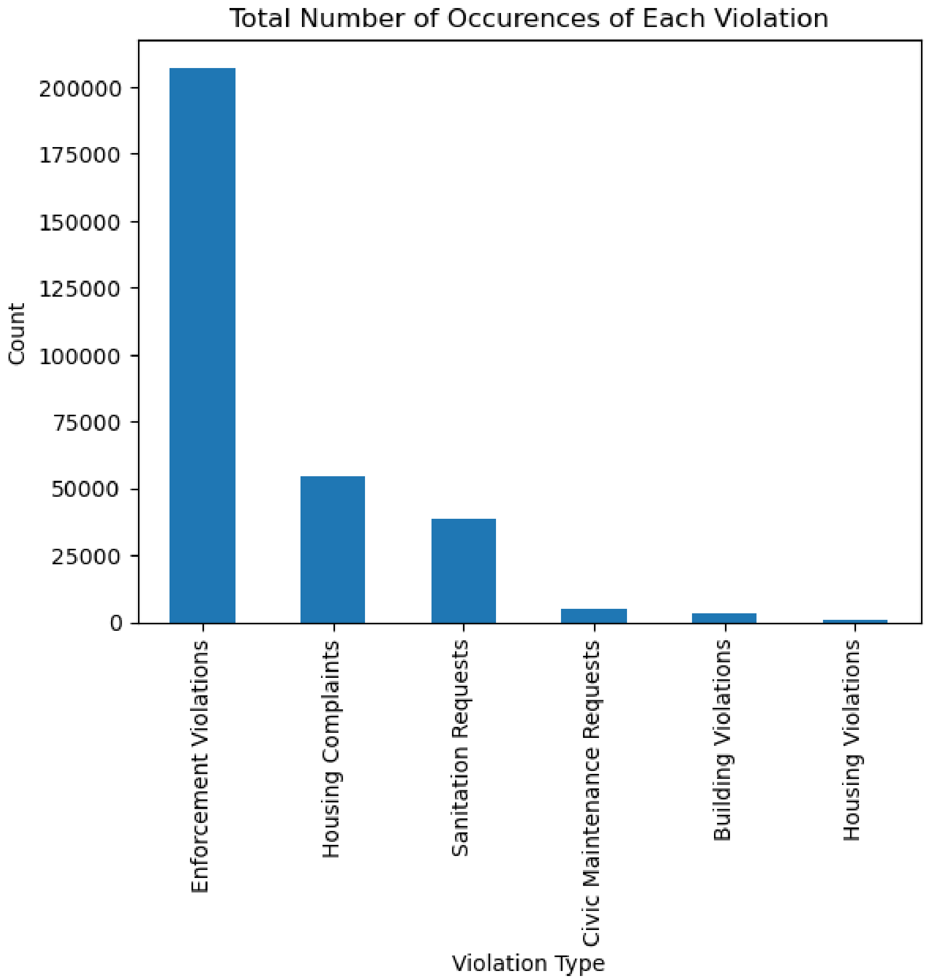


Fig 11. Number of Occurrences of Each Type of Violation

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".

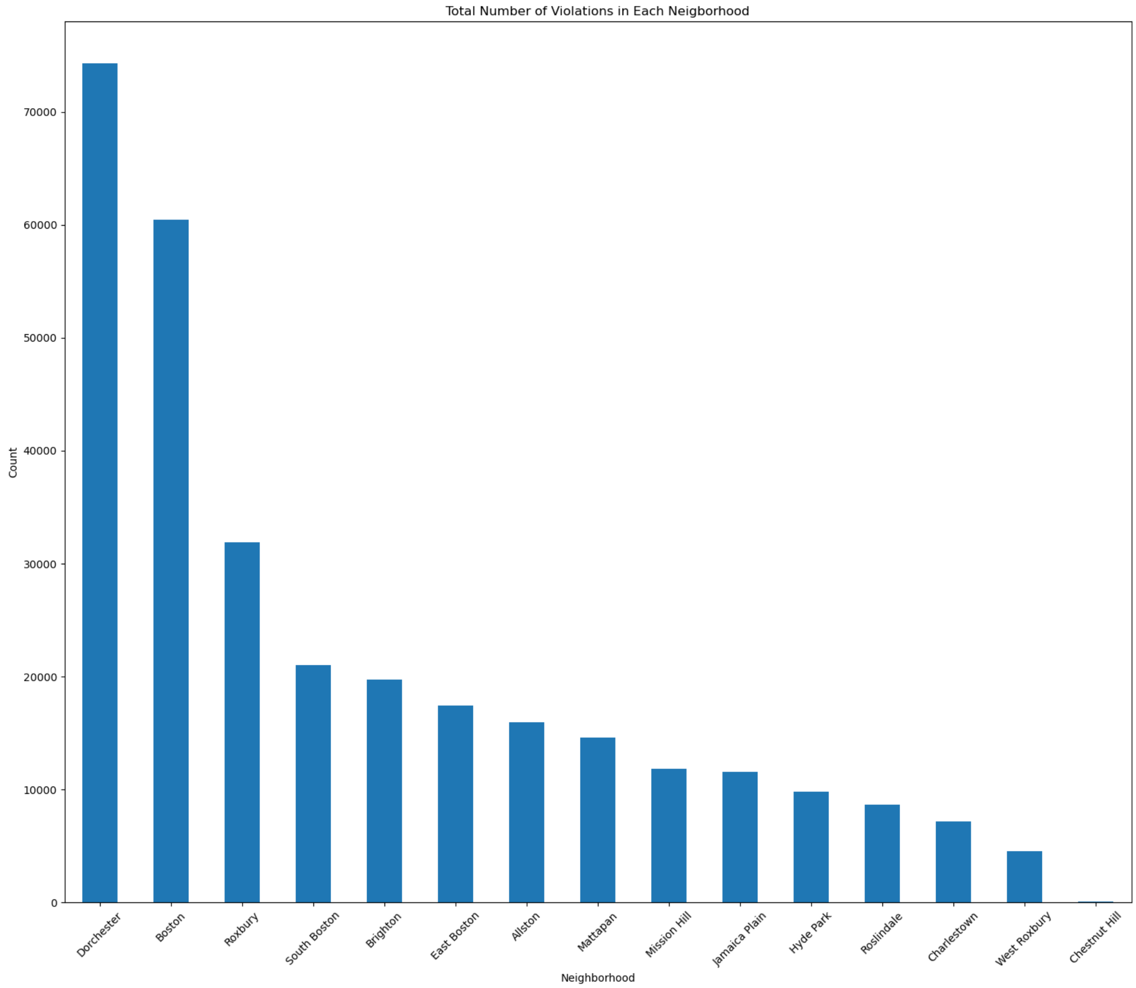


Fig 12. Number of Violations per Boston Neighborhood

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.

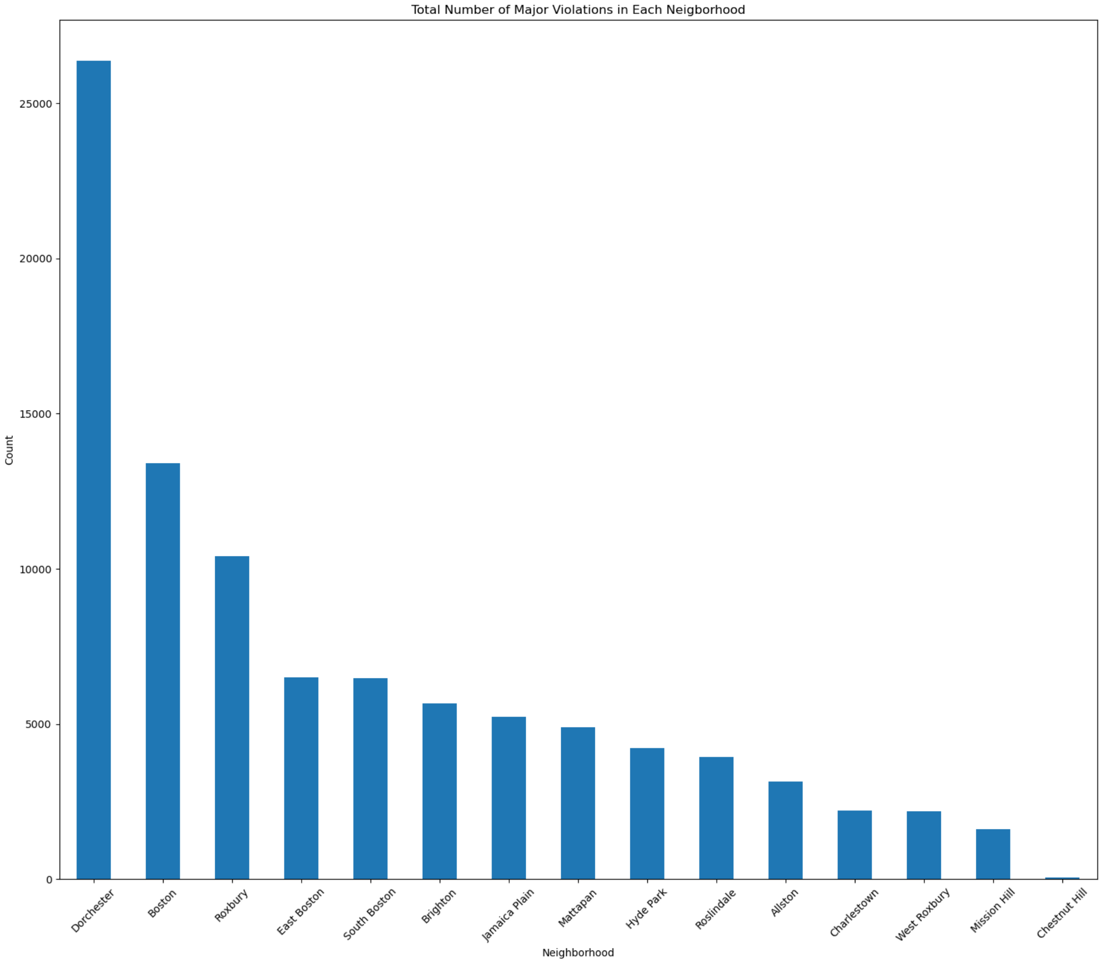


Fig 13. Number of Major Violations per Boston Neighborhood

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.

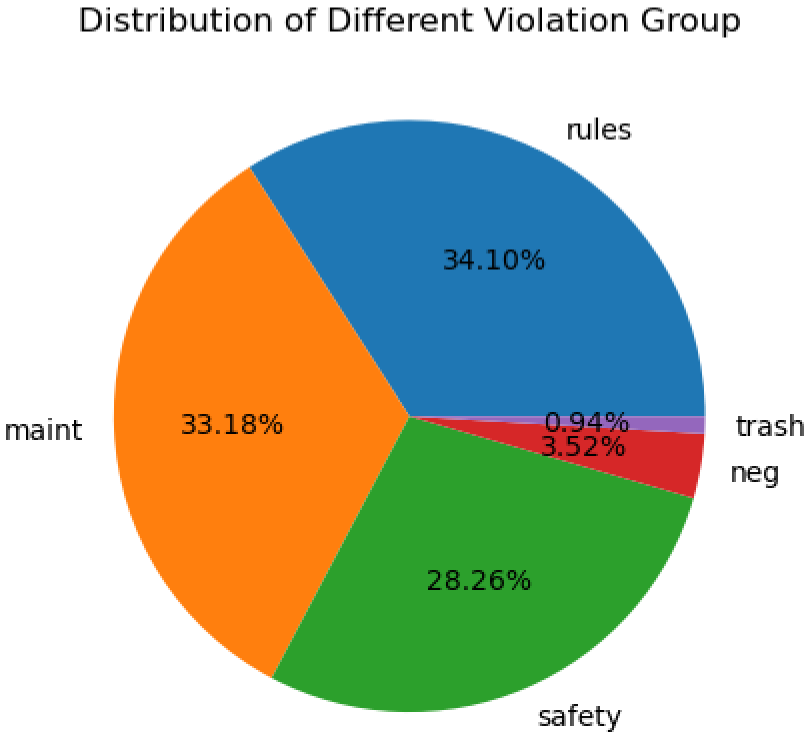


Fig 14. Percentage of Different Types of Violations in Boston City

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 an 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.

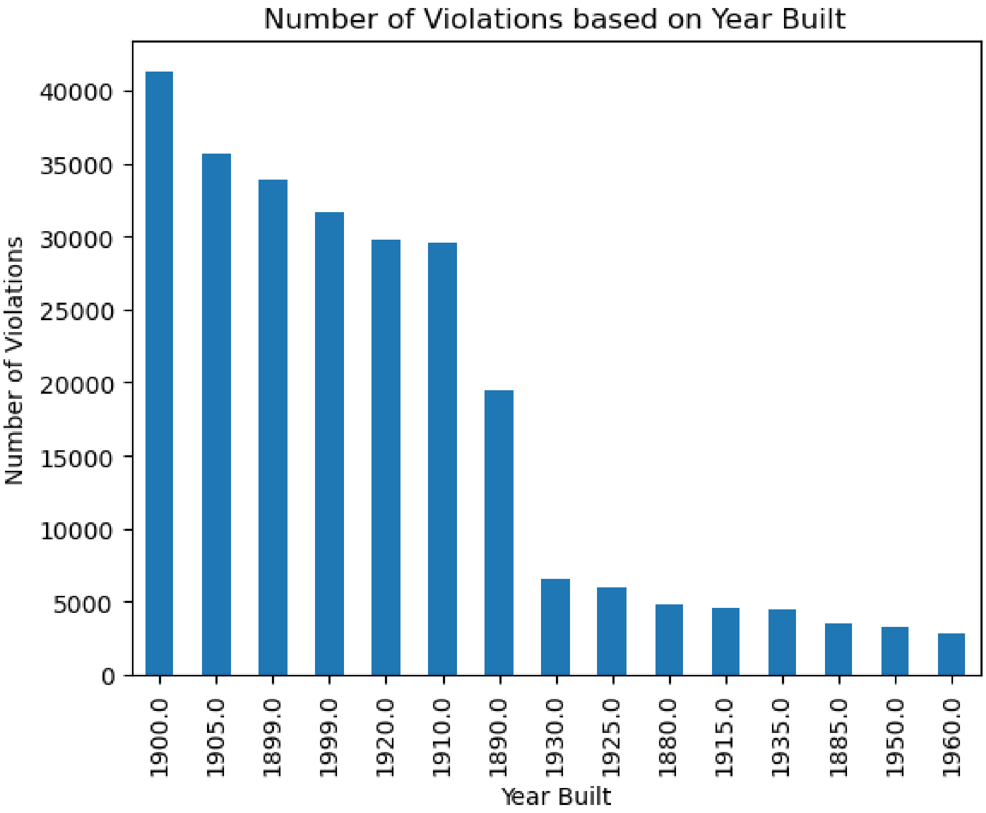


Fig 15. Highest Number of Violations based on the Year a House was Built in

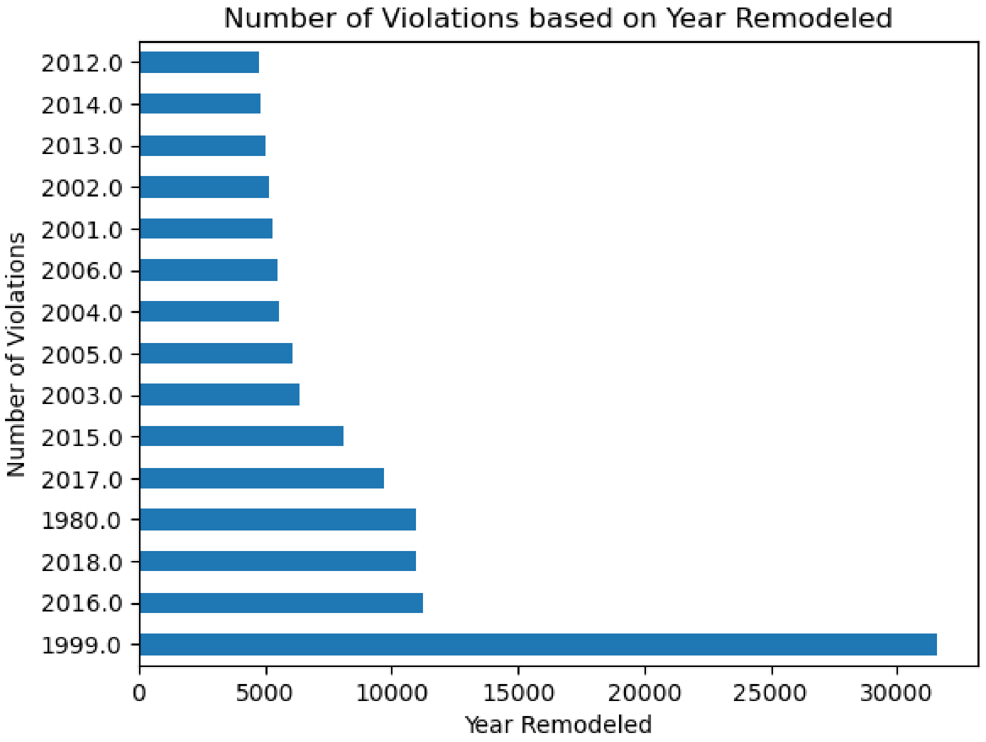


Fig 16. Highest Number of Violations based on the Year a House was Remodeled in

The above two bar graphs show the top 15 years with the 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 is 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 the year built is 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.

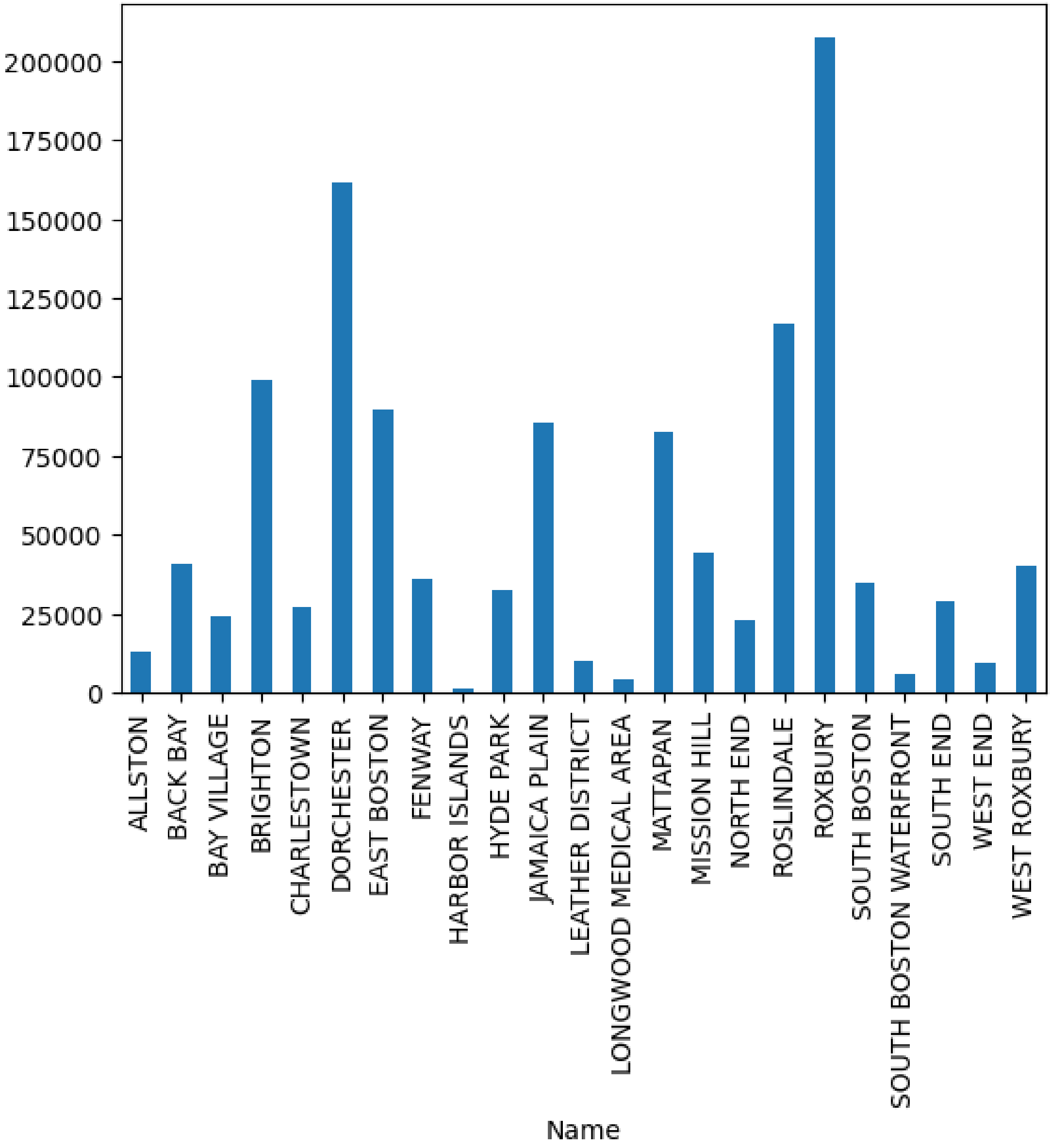


Fig 17. Total Vulnerable Population per Neighborhood in Boston City

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.

As mentioned by the clients in recent meetings, we also worked on finding the ratio of the number of violations per neighborhood to the number of houses per neighborhood. Fig 18 depicts this ratio for each Boston Neighborhood.

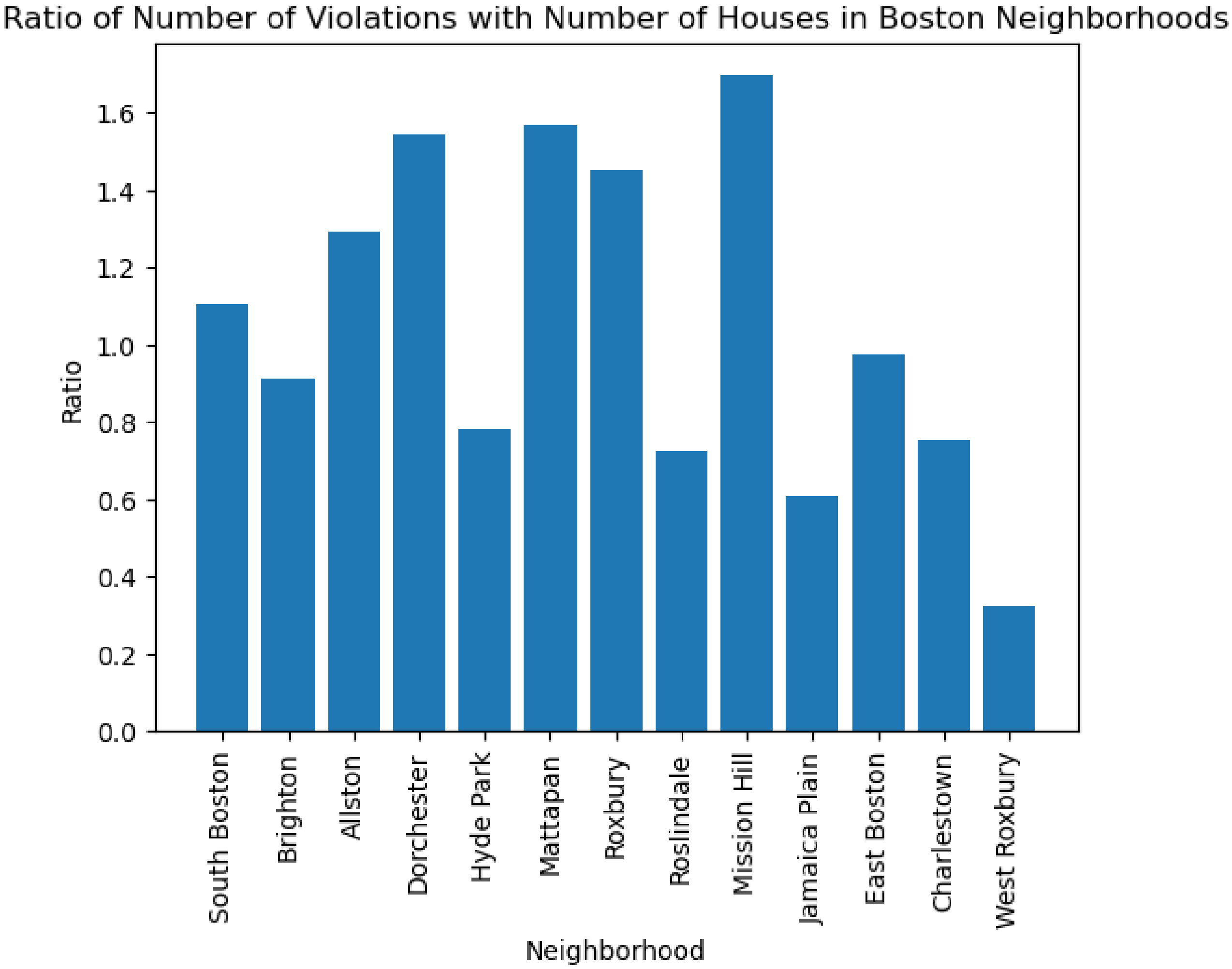


Fig 18. Ratio of the Number of Violations to the Number of Houses in each Boston Neighborhood

As it can be seen from the above bar graph, Mission Hill has the worst (about 1.7) Violations to House ratio. This means each house in Mission Hill has an average of 1.7 violations. According to this metric, West Roxbury is an excellent neighborhood to live in. Its Violation to Houses ratio is the lowest among all neighborhoods (about 0.35), meaning each house in West Roxbury has only 0.35 violations (i.e. many houses might have 0 violations, and some might have 1, 2, and so on). From our previous analysis, we already saw that Dorchester doesn't seem to be the best neighborhood to live in. With the third highest Violation to House ratio (about 1.57), Dorchester is again not an ideal neighborhood to live in. Brighton, one of the major neighborhoods in Boston, has a Violation to House ratio of less than 1, making it a good neighborhood to live in.

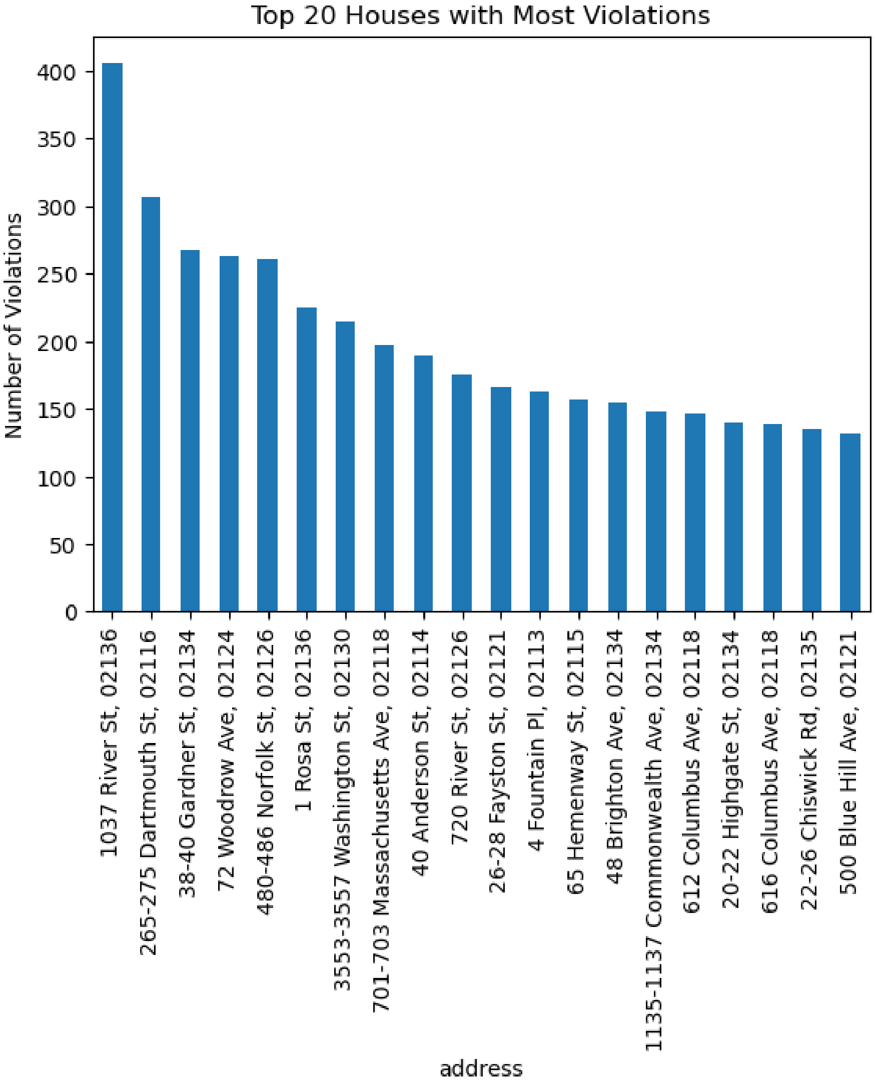
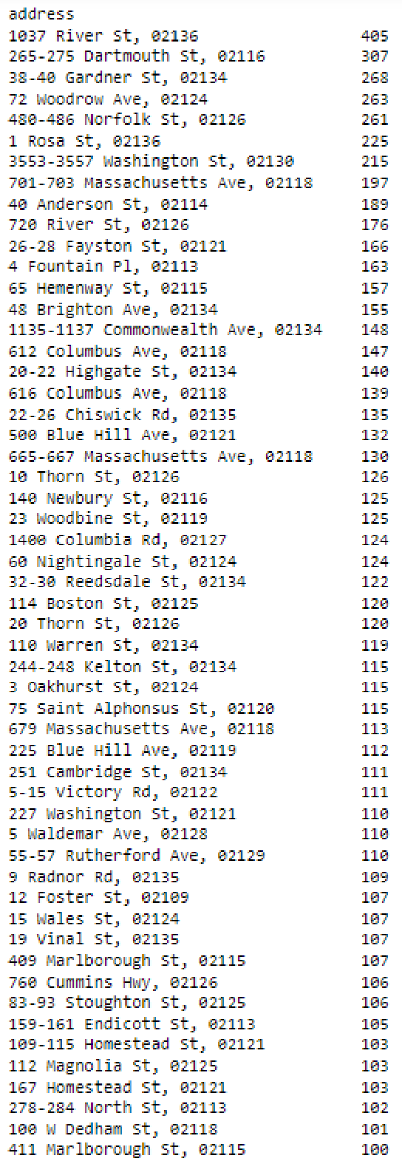


Fig 19. Addresses of the Houses (Top 20) with Most violations

Fig 19 showcases the top 20 houses with the most violations in Boston City. As it can be seen, the house at 1037 River Street (02136) is the worst house in the city, with over 400 violations to its name. There are many houses in Boston city with over 100 violations. Fig 20 lists the addresses of all the houses with 100 or more violations. There are a total of 54 houses with 100 or more violations in Boston city. These are the worst houses in terms of just the number of violations.



```
address
1037 River St, 02136                 405
265-275 Dartmouth St, 02116          307
38-40 Gardner St, 02134              268
72 Woodrow Ave, 02124                263
480-486 Norfolk St, 02126            261
1 Rosa St, 02136                     225
3553-3557 Washington St, 02130       215
701-703 Massachusetts Ave, 02118     197
40 Anderson St, 02114                189
720 River St, 02126                  176
26-28 Fayston St, 02121              166
4 Fountain Pl, 02113                 163
65 Hemenway St, 02115                157
48 Brighton Ave, 02134               155
1135-1137 Commonwealth Ave, 02134    148
612 Columbus Ave, 02118              147
20-22 Highgate St, 02134             140
616 Columbus Ave, 02118              139
22-26 Chiswick Rd, 02135             135
500 Blue Hill Ave, 02121             132
665-667 Massachusetts Ave, 02118     130
10 Thorn St, 02126                   126
140 Newbury St, 02116                125
23 Woodbine St, 02119                125
1400 Columbia Rd, 02127              124
60 Nightingale St, 02124             124
32-30 Reedsdale St, 02134            122
114 Boston St, 02125                 120
20 Thorn St, 02126                   120
110 Warren St, 02134                 119
244-248 Kelton St, 02134             115
3 Oakhurst St, 02124                 115
75 Saint Alphonsus St, 02120         115
679 Massachusetts Ave, 02118         113
225 Blue Hill Ave, 02119             112
251 Cambridge St, 02134              111
5-15 Victory Rd, 02122               111
227 Washington St, 02121             110
5 Waldemar Ave, 02128                110
55-57 Rutherford Ave, 02129          110
9 Radnor Rd, 02135                   109
12 Foster St, 02109                  107
15 Wales St, 02124                   107
19 Vinal St, 02135                   107
409 Marlborough St, 02115            107
760 Cummins Hwy, 02126               106
83-93 Stoughton St, 02125            106
159-161 Endicott St, 02113           105
109-115 Homestead St, 02121          103
112 Magnolia St, 02125               103
167 Homestead St, 02121              103
278-284 North St, 02113              102
100 W Dedham St, 02118               101
411 Marlborough St, 02115            100
```

Fig 20. List of House Addresses with 100 or More Violations in Boston City

As mentioned earlier, there are three violations that should be weighed higher than the remaining (i.e. Sanitation Requests, Building Violations, and Housing Complaints). If only these types of violations are considered, then there is a drastic change in the total number of violations of each house and the worst houses list. Fig 21 showcases the Top 20 houses with the highest number of the above-mentioned three violation types. The house at 225 Blue Hill Avenue (02119) is the only house with over 100 of these types of violations and can be deemed the worst house in Boston City by far.

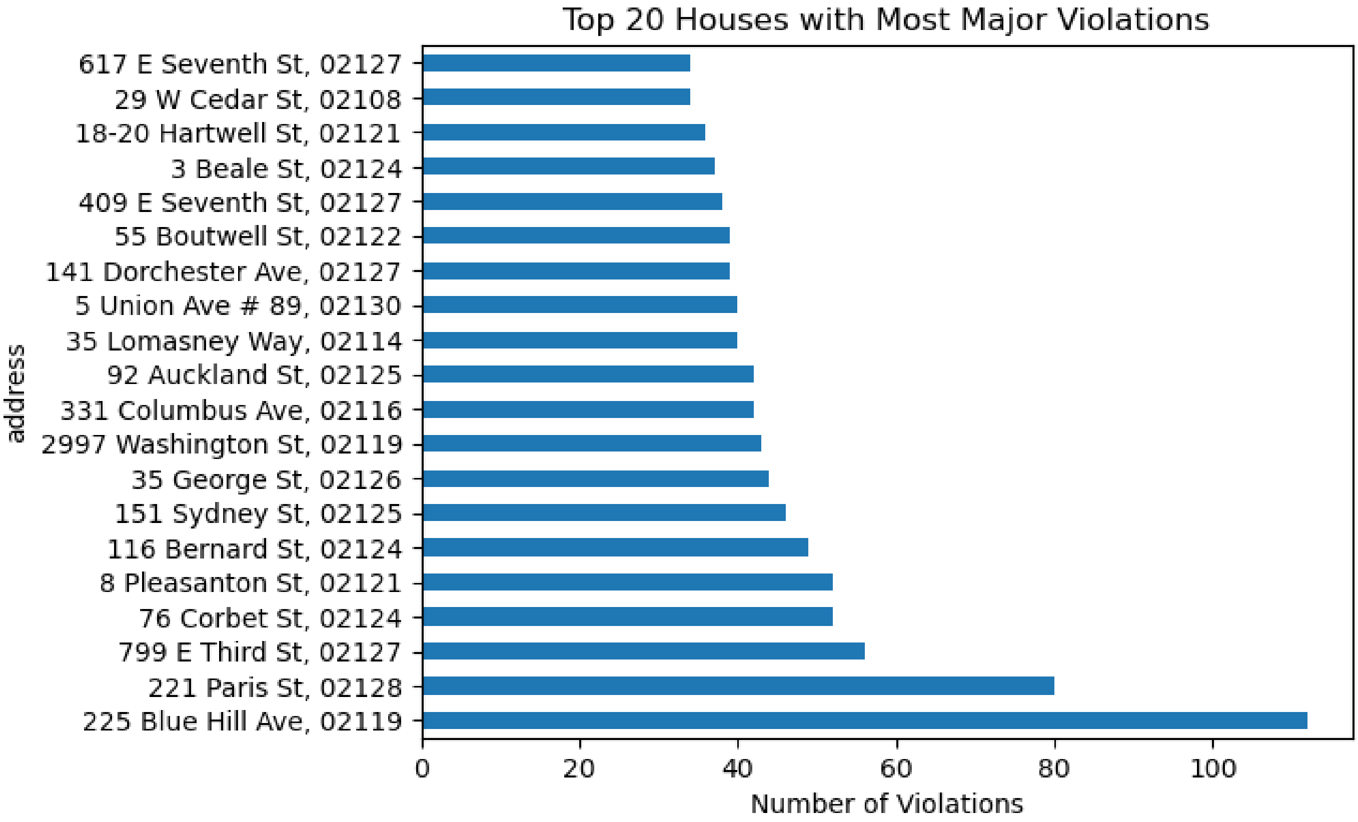


Fig 21. List of House Addresses with 100 or More Violations in Boston City

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.

   To achieve this, we could give different weights to each of these violations (with a higher weight to the above mentioned violations and a lower weight to the others). Finally, in the end, instead of just looking at the number of violations, we will look at the value that comes from multiplying their violations with these weights. The more the weight, the worse the landlord.

   Following are the weights we assigned to each of the violations:

   Housing Complaints: 1.0
   Sanitation Requests: 0.9
   Civic Maintenance Requests: 0.1
   Building Violations: 0.8
   Housing Violations: 0.4
   Enforcement Violations: 0.0 (Enforcement Violations can be ignored as per the latest meeting with the client)

   Fig 21 shows the houses (Top 20) with the highest violation score (the higher this violation score, the worse the house).

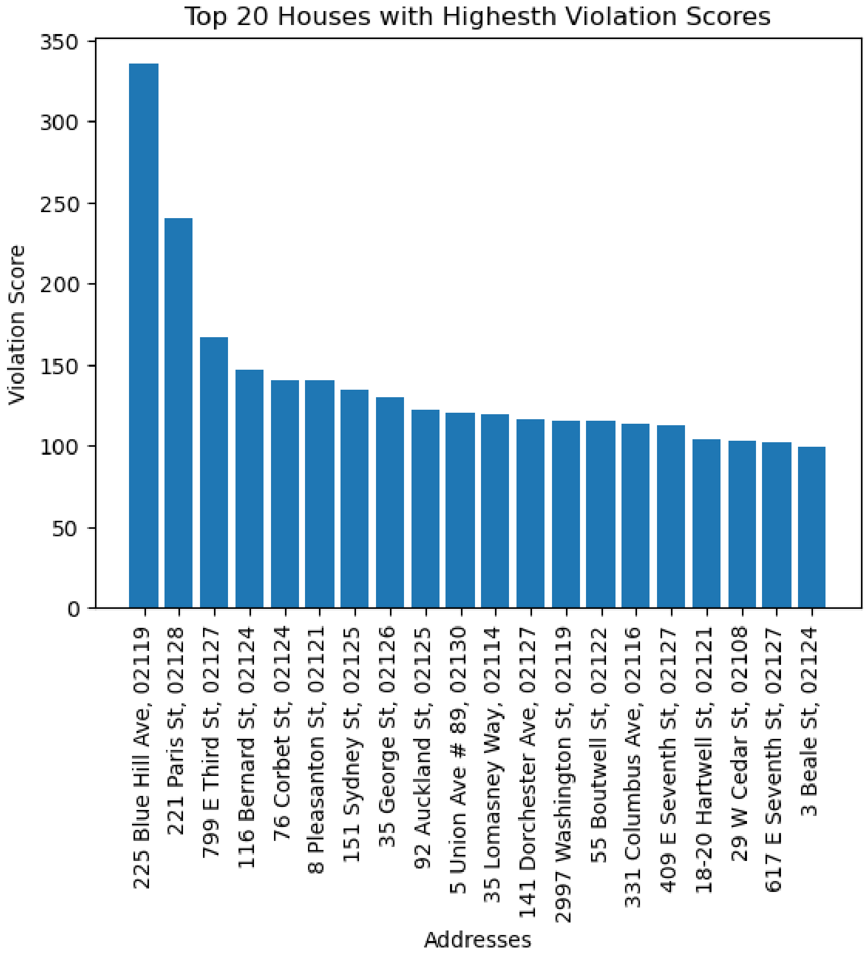


Fig 22. Violation Scores of the Top 20 Worst Houses of Boston According to Our Analysis

Fig 22 correlates pretty well with Fig 20 (List of House Addresses with 100 or More Violations in Boston City). 225 Blue Hill Avenue has the first position in both graphs, and 221 Paris Street is the second house in both of them. There are a few other common houses in both graphs.

A csv file containing the list of all houses and their violation scores can be accessed here.

Adding to this, we also mapped these scores with the owners of each house (which was the end goal of this project). Fig 23 shows the Top 20 Owners with the Highest Violation Scores. GBM Portfolio Owner LLC is the worst landlord (it is a company) in Boston city, with a violation score of almost 40,000. Apart from companies, there are people in this graph who are some of the worst landlords in Boston.

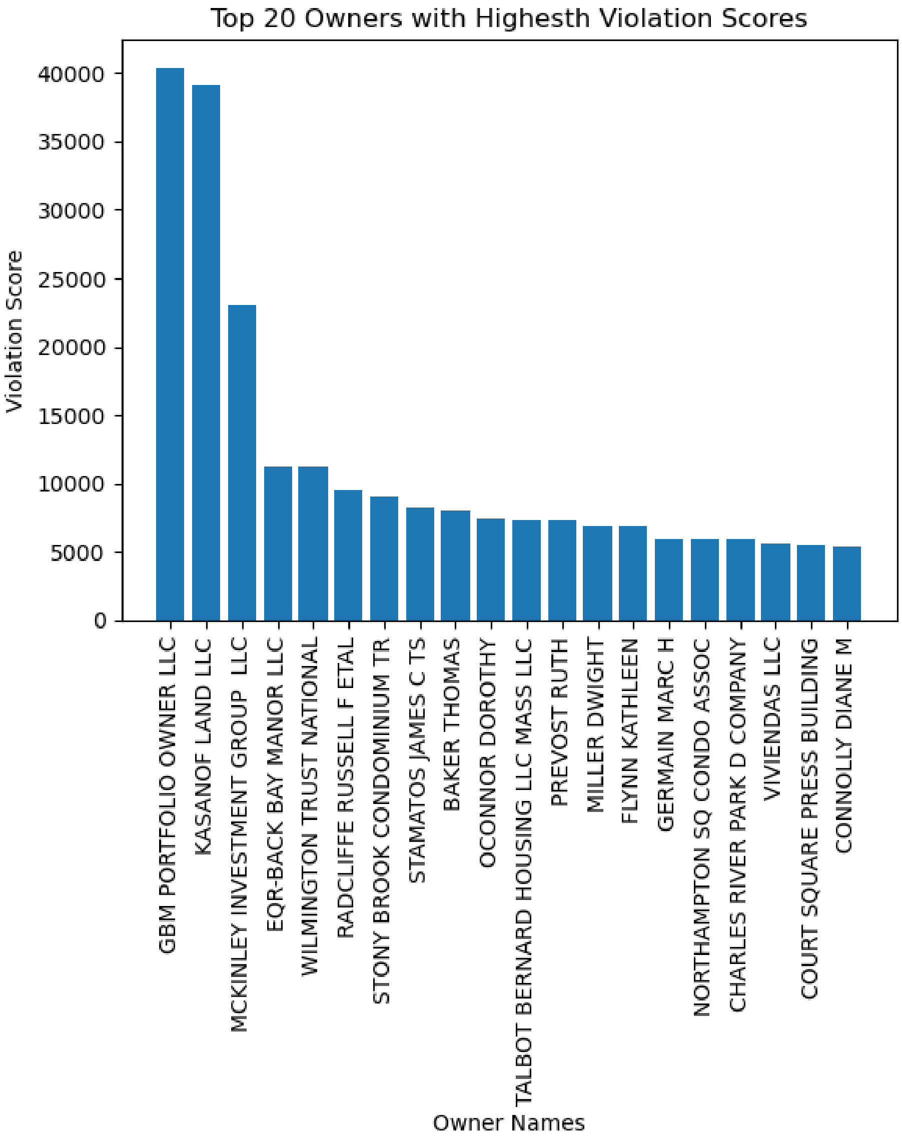


Fig 23. Violation Scores of the Top 20 Worst House Owners (Landlords) of Boston According to Our Analysis

A csv file containing the list of all house owners (or landlords) and their violation scores can be accessed here.

2. Who is causing the problems?

   Answer:

   If we look at the violation scores of properties in Boston city, we see that there are a lot of companies that have poor violation scores as compared to houses or individual landlords.

   Fig 24 shows the violation scores of 20 properties with the worst violation scores in the city.

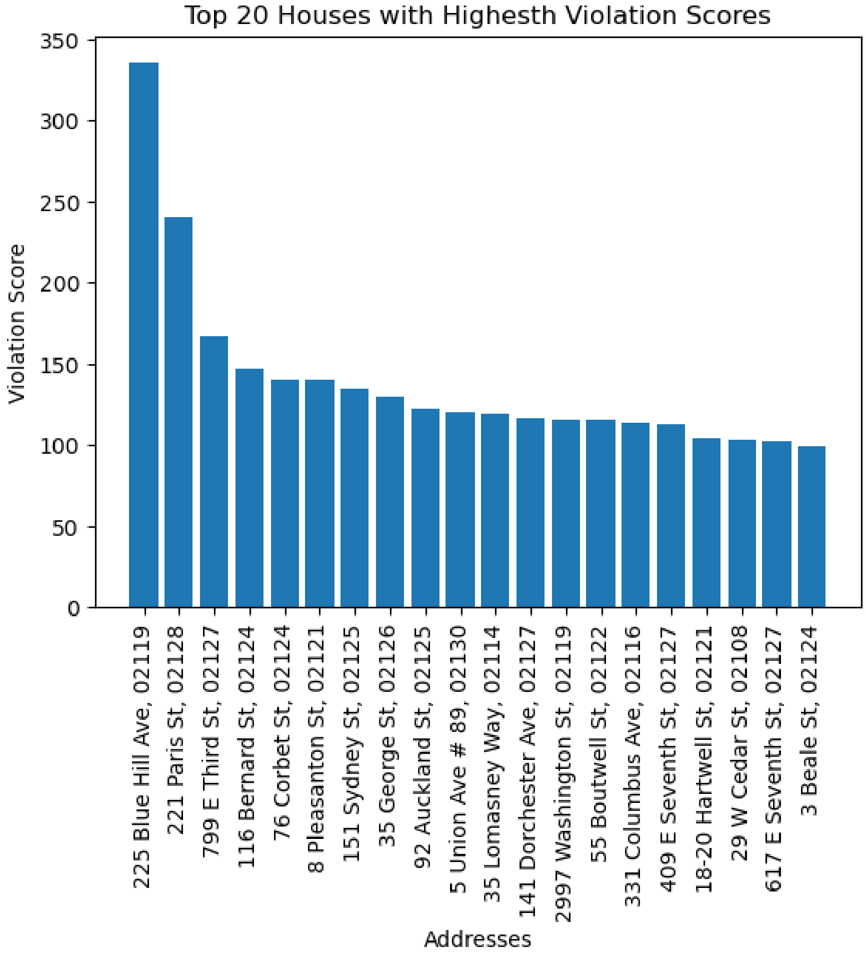


Fig 24. Violation Scores of the Top 20 Worst Houses of Boston

These are the properties that are causing the most problems in Boston City. There are many more properties with a violation score of over 100. On the other hand, there are properties in the city with 0 violation score (i.e. houses which do not have any active or past violations in the city).

3. What factors are correlated with the violations?

Answer:

Majority of the factors given in the datasets are categorical. Finding the correlation between categorical and textual features/variables is often not the best way to analyze which factors affect the outcome more and which less.

Even so, we tried to find the correlation between different features in the dataset with the number of violations using the Chi-Square test. We found out that all features were giving a score of 0.0. A Chi-Square score of 0.0 means that all the values of both the features match exactly (which is not possible).

Hence, we did not further examine the correlation between different columns with violations using Chi-Square test.

However, if we compare the violation score of each neighborhood with its population, we find that these two features have a positive correlation of 0.601. Fig 25 shows this correlation in graph form. As it can be seen, both these features have a positive slop showcasing a positive correlation between them and a correlation score of 0.6 depicts a strong correlation.

Hence, we can assume that as the population of a neighborhood increases, so does the number of violations. This conclusion also follows logically since as population increases, so does the number of houses, and as the number of houses increase, so do the number of violations in the neighborhood.

Another factor that is correlated with the number of violations per neighborhood is the social vulnerability of the neighborhood (which is also our extension project). As we see in the Extension, the Social Vulnerability of a neighborhood is pretty strongly correlated with the number of violations of a neighborhood. This also follows logically since if a population is socially vulnerable, then it is easy for landlords and house owners to exploit their vulnerability and behave badly without the fear of punishment.

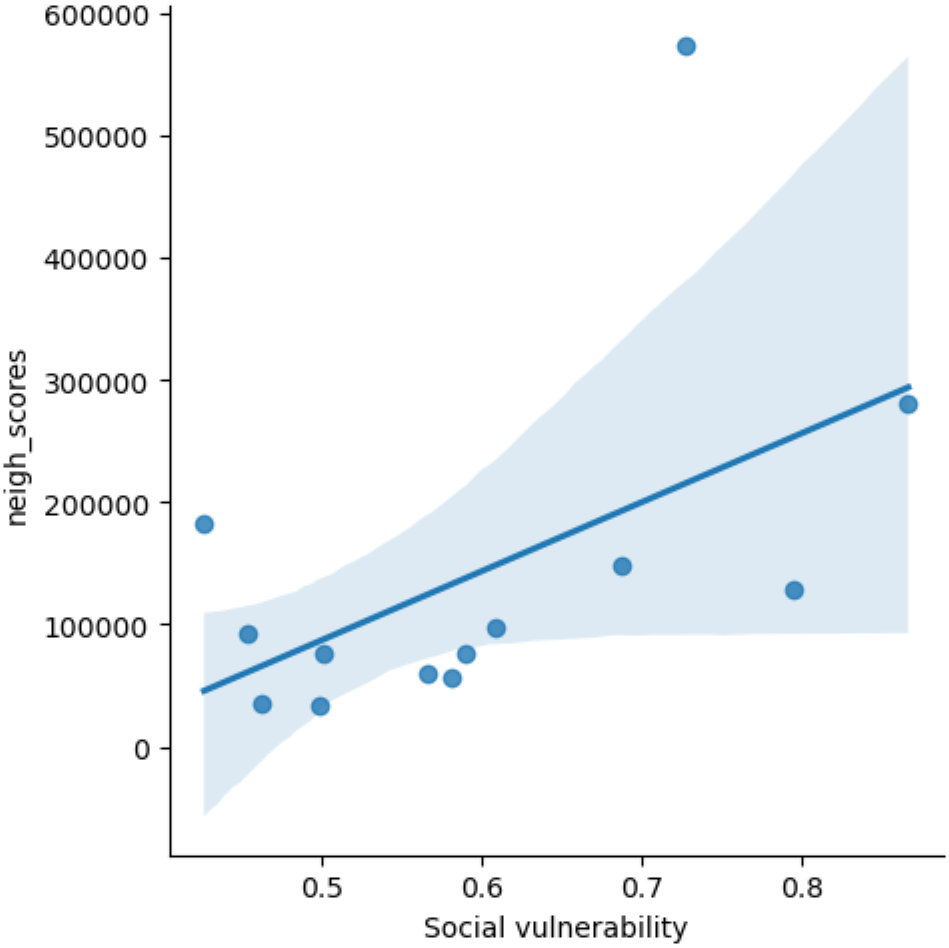


Fig 25. Correlation between Population and Violation Score

Basic Questions

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

Answer:

There are mainly six 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 Occurrences

Sr No.	Violation	Number of Occurrences
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

For most of this analysis, Enforcement violations have been ignored (as mentioned in the client meetings).

2. Who has the most violations?

Answer:

The graph in Fig 19 depicts the addresses of the Top 20 Properties in Boston City with the highest number of violations. 1037, River Street, is the property with over 400 violations making it by far the worst property of Boston City. The second property in this graph (265-275 Dartmouth Street) is the second worst property with around 300 violations.

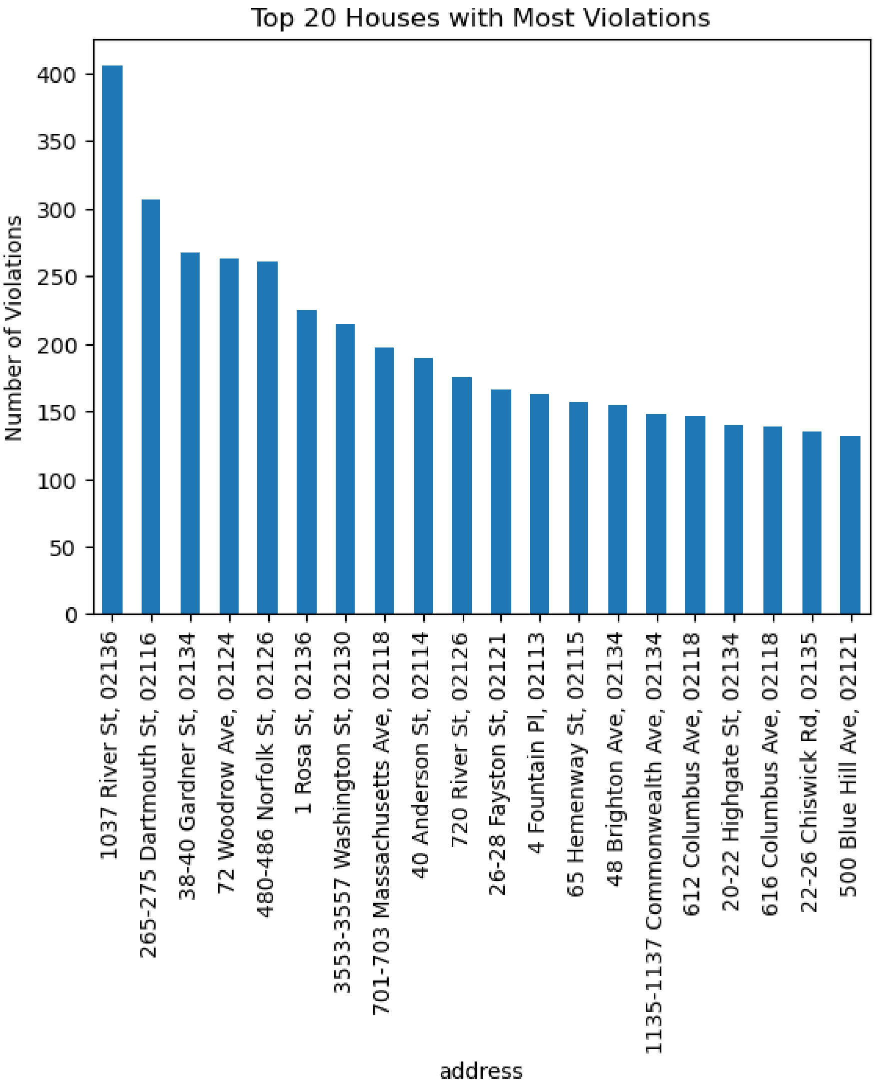


Fig 19. Top 20 houses with Most Violations in Boston City

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. Distribution of % of violations across Boston

<u>**Answer:**</u>

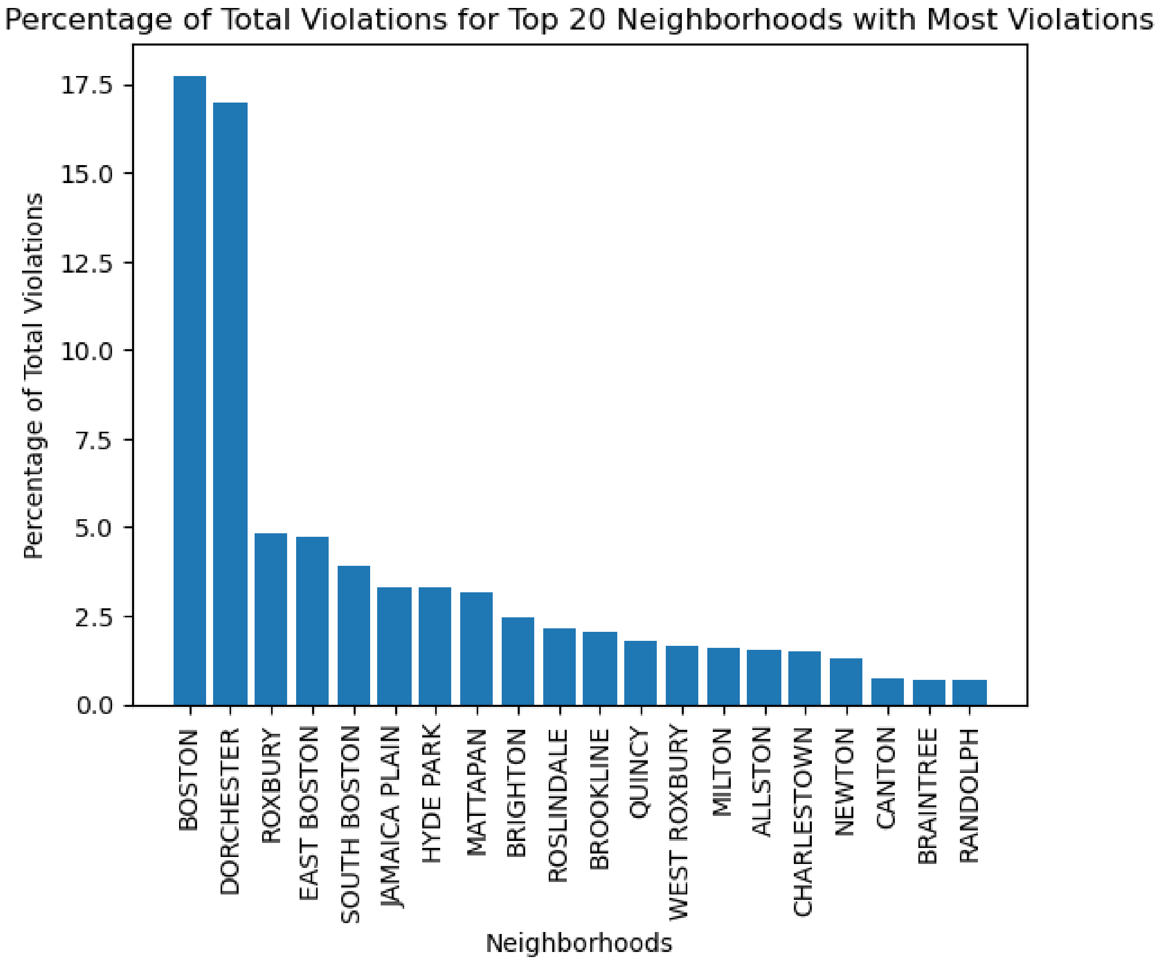


Fig 24. Percentage of Total Violation for Top 20 Neighborhoods

Fig 24 showcases the percentage of violations for the top 20 Neighborhoods with highest number of violations in Boston city. As it can be seen, Dorchester is one major neighborhood with the highest percentage of violations (and it also has the highest number of violations as it was seen in the Data Visualization and Analysis section).

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

Answer:

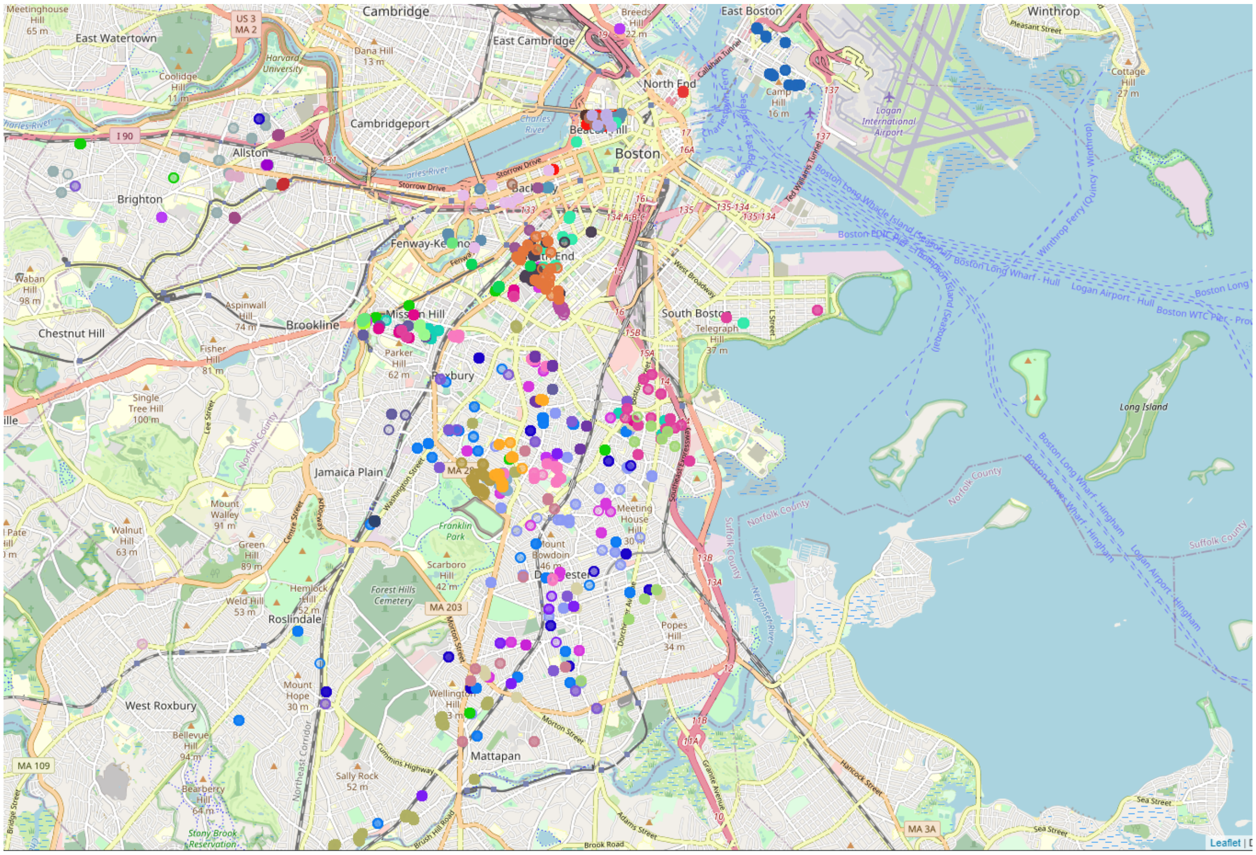
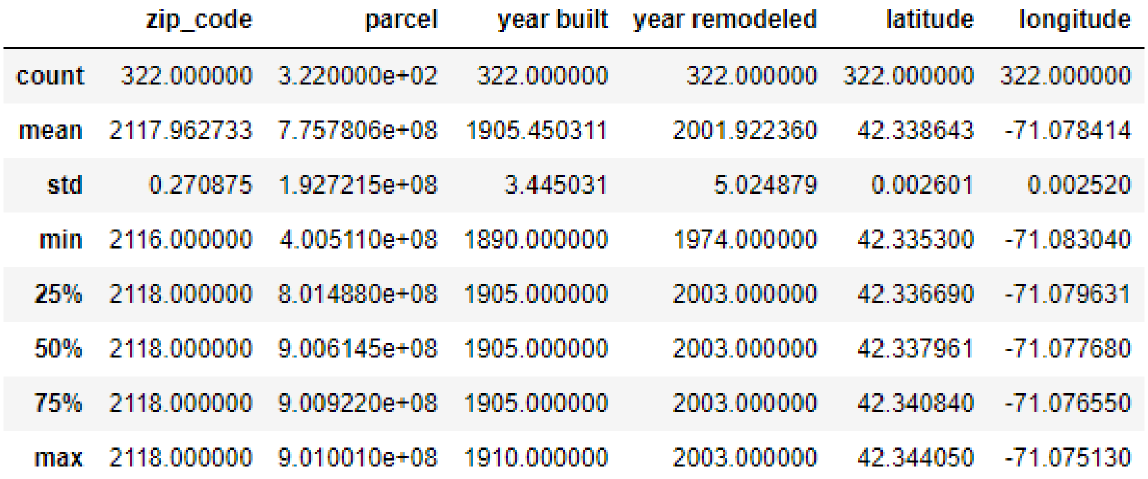


Fig 25. Map Containing the Coordinates of the Top 50 House Owners with Most Violations

From Fig 25, it is certain that there are a few blocks of house owners in Boston city. This map depicts just the Top 50 house owners in terms of the most number of violations. If we go on and analyze more houses, we might see more clusters i.e. more blocks of house owners with violations.

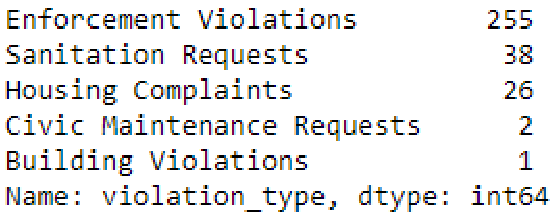
We focus on mainly three of these blocks: 1. The Orange block in South End (in the top middle part of Fig 25), 2. The Woody Brownish block just north of Franklin Zoo Park (just left to the middle part of Fig 25), and 3. The Dark Pink Cluster to the West of UMass/JFK Southeast Express highway (just right to the middle part of Fig 25). We choose these three blocks to showcase that the variance in neighborhood does not affect the formation of these blocks of owners.

The first cluster (Orange) is owned by TENANTS DEVELOPMENT II LP. This company has a bunch of properties with violations in South End. The violations of these companies and their statistics are as follows:



	zip_code	parcel	year built	year remodeled	latitude	longitude
count	322.000000	3.220000e+02	322.000000	322.000000	322.000000	322.000000
mean	2117.962733	7.757806e+08	1905.450311	2001.922360	42.338643	-71.078414
std	0.270875	1.927215e+08	3.445031	5.024879	0.002601	0.002520
min	2116.000000	4.005110e+08	1890.000000	1974.000000	42.335300	-71.083040
25%	2118.000000	8.014880e+08	1905.000000	2003.000000	42.336690	-71.079631
50%	2118.000000	9.006145e+08	1905.000000	2003.000000	42.337961	-71.077680
75%	2118.000000	9.009220e+08	1905.000000	2003.000000	42.340840	-71.076550
max	2118.000000	9.010010e+08	1910.000000	2003.000000	42.344050	-71.075130

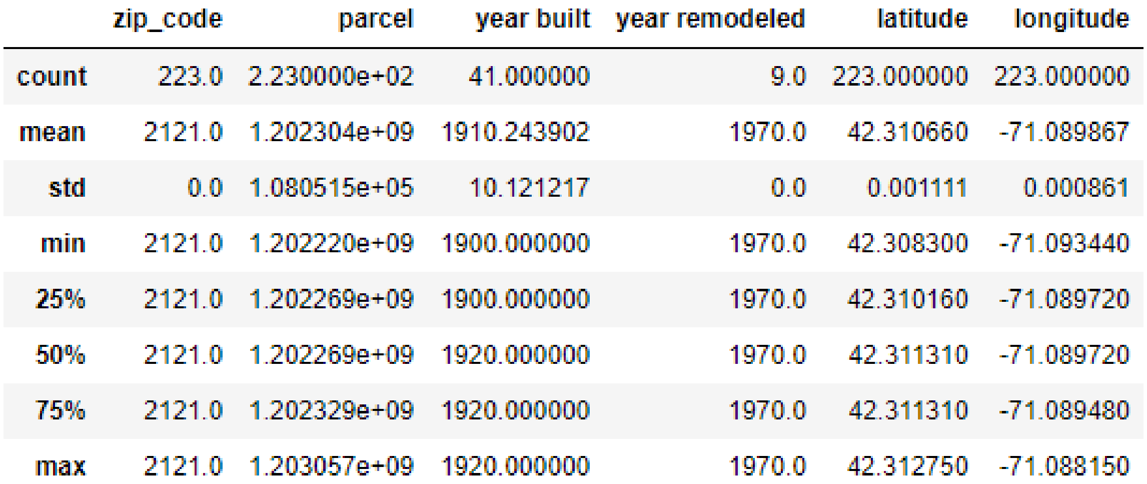
Fig 26. Statistical Description of properties owned by TENANTS DEVELOPMENT II LP



```
Enforcement Violations        255
Sanitation Requests            38
Housing Complaints             26
Civic Maintenance Requests      2
Building Violations             1
Name: violation_type, dtype: int64
```

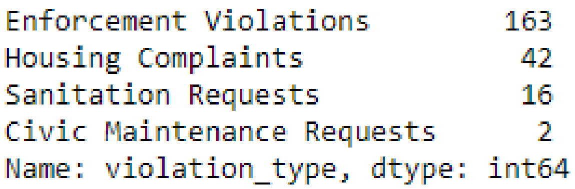
Fig 27. Different Violations (along with the total number of their occurrences) by TENANTS DEVELOPMENT II LP

The second cluster (Woody Brown) is owned by FRANKLIN HIGHLANDS LP. This company has many properties with violations near Franklin Park Zoo (the company might be the owner or curator of the zoo). The violations of these companies and their statistics are as follows:



	zip_code	parcel	year built	year remodeled	latitude	longitude
count	223.0	2.230000e+02	41.000000	9.0	223.000000	223.000000
mean	2121.0	1.202304e+09	1910.243902	1970.0	42.310660	-71.089867
std	0.0	1.080515e+05	10.121217	0.0	0.001111	0.000861
min	2121.0	1.202220e+09	1900.000000	1970.0	42.308300	-71.093440
25%	2121.0	1.202269e+09	1900.000000	1970.0	42.310160	-71.089720
50%	2121.0	1.202269e+09	1920.000000	1970.0	42.311310	-71.089720
75%	2121.0	1.202329e+09	1920.000000	1970.0	42.311310	-71.089480
max	2121.0	1.203057e+09	1920.000000	1970.0	42.312750	-71.088150

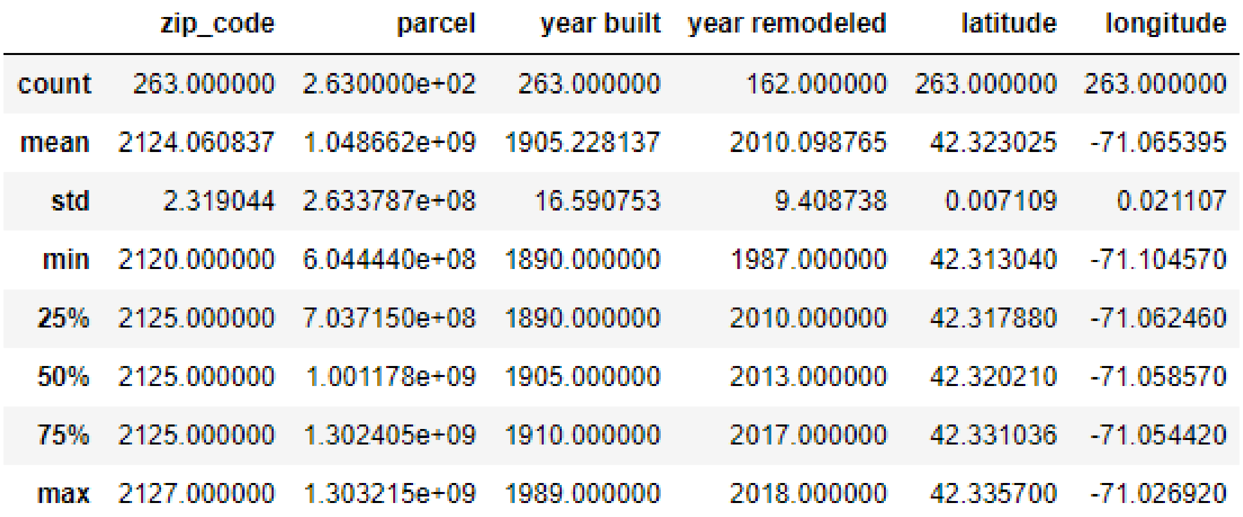
Fig 28. Statistical Description of properties owned by FRANKLIN HIGHLANDS LP



```
Enforcement Violations        163
Housing Complaints             42
Sanitation Requests            16
Civic Maintenance Requests      2
Name: violation_type, dtype: int64
```

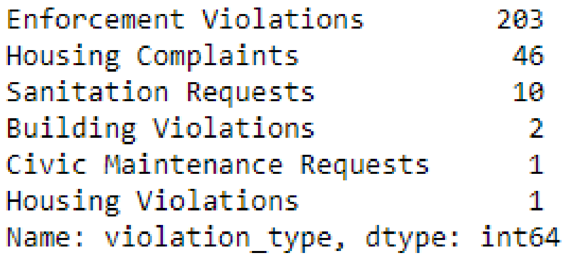
Fig 29. Different Violations (along with the total number of their occurrences) by FRANKLIN HIGHLANDS LP

The third cluster (Dark Pink) is owned by REAL ESTATE BOSTON LLC. This company has many properties with violations near the UMASS-JFK Southeast Expressway. The violations of these companies and their statistics are as follows:



	zip_code	parcel	year built	year remodeled	latitude	longitude
count	263.000000	2.630000e+02	263.000000	162.000000	263.000000	263.000000
mean	2124.060837	1.048662e+09	1905.228137	2010.098765	42.323025	-71.065395
std	2.319044	2.633787e+08	16.590753	9.408738	0.007109	0.021107
min	2120.000000	6.044440e+08	1890.000000	1987.000000	42.313040	-71.104570
25%	2125.000000	7.037150e+08	1890.000000	2010.000000	42.317880	-71.062460
50%	2125.000000	1.001178e+09	1905.000000	2013.000000	42.320210	-71.058570
75%	2125.000000	1.302405e+09	1910.000000	2017.000000	42.331036	-71.054420
max	2127.000000	1.303215e+09	1989.000000	2018.000000	42.335700	-71.026920

Fig 30. Statistical Description of properties owned by REAL ESTATE BOSTON LLC



```
Enforcement Violations        203
Housing Complaints             46
Sanitation Requests            10
Building Violations             2
Civic Maintenance Requests      1
Housing Violations              1
Name: violation_type, dtype: int64
```

Fig 31. Different Violations (along with the total number of their occurrences) by REAL ESTATE BOSTON LLC

5. 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.

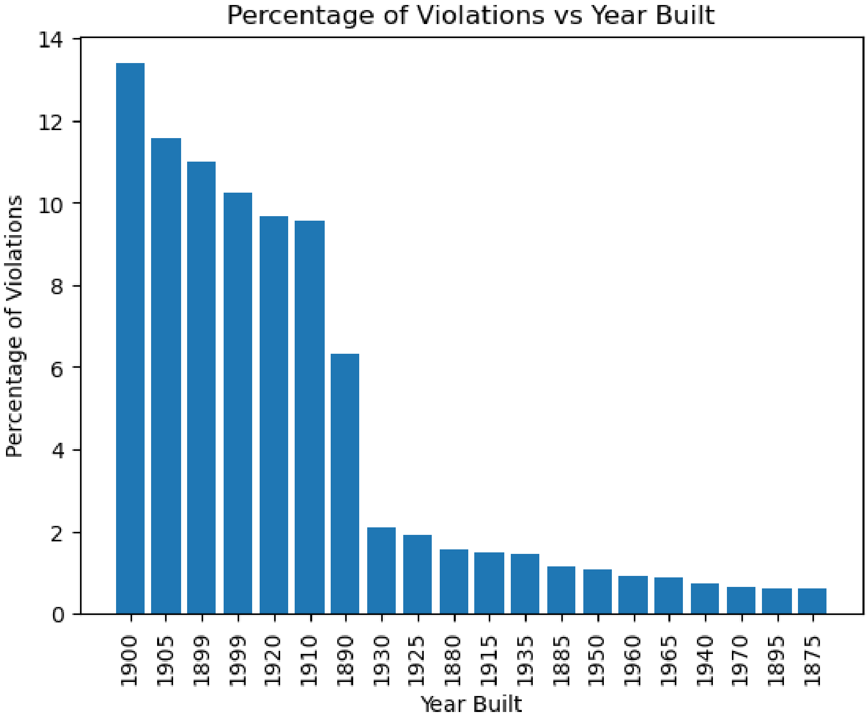


Fig 32. Percentage of Violations for each year

Fig 32 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.

6. Whether affordability affects the amount of property violations?

Answer:

To examine the correlation between affordability and number of property violations, we take a look at the average income and the number of violations of each neighborhood.

When we find the correlation between these two metrics for all neighborhoods, we get a negative correlation score of -0.383. This means that these two metrics are infact inversely correlated to each other. Fig 33 below shows this correlation in graph form.

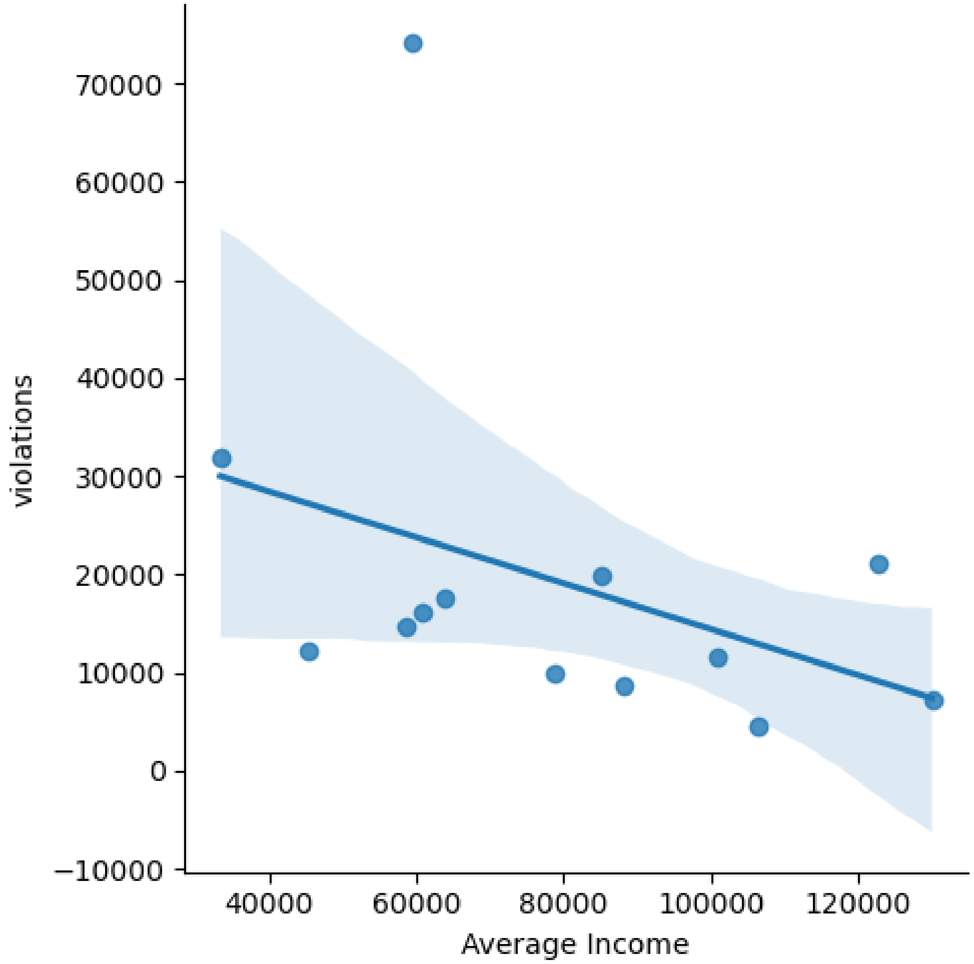


Fig 33. Correlation between Average Income and No. of Violations for Each Neighborhood

It is logical to assume that the affordability of a neighborhood would be directly proportional to the average income of that particular neighborhood. So, with that assumption, we can say that since the average income of each neighborhood is inversely related to the no. of violations of each neighborhood, the affordability of a neighborhood is also inversely proportional to the no. of violations of each neighborhood. This means the less the affordability of a neighborhood, the more the number of violations.

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 during the initial preprocessing of the data was building violation codes. If these codes could be better explained to us in some way in the future, it would be 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.

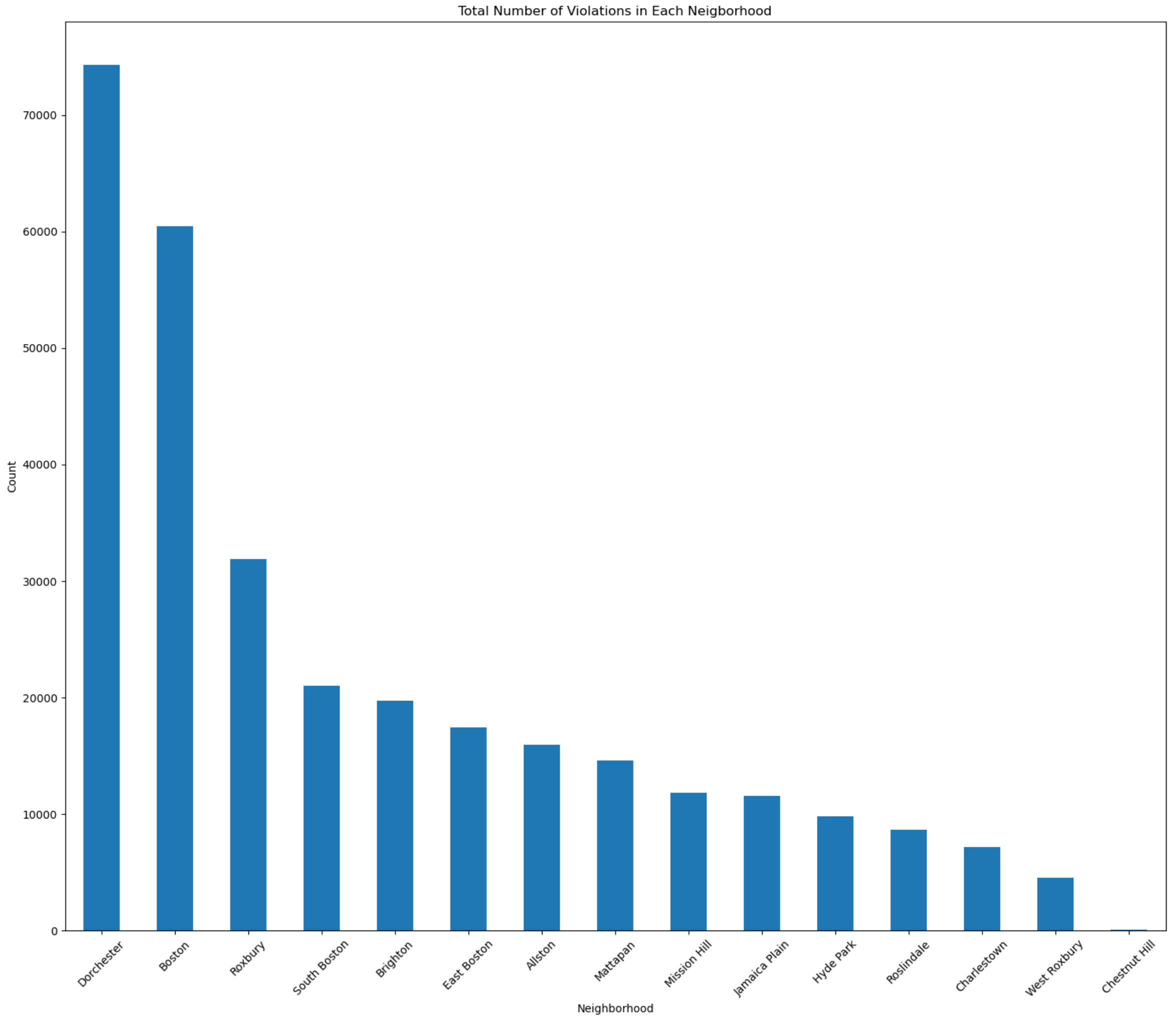


Fig 33. Number of Violations per Boston Neighborhood

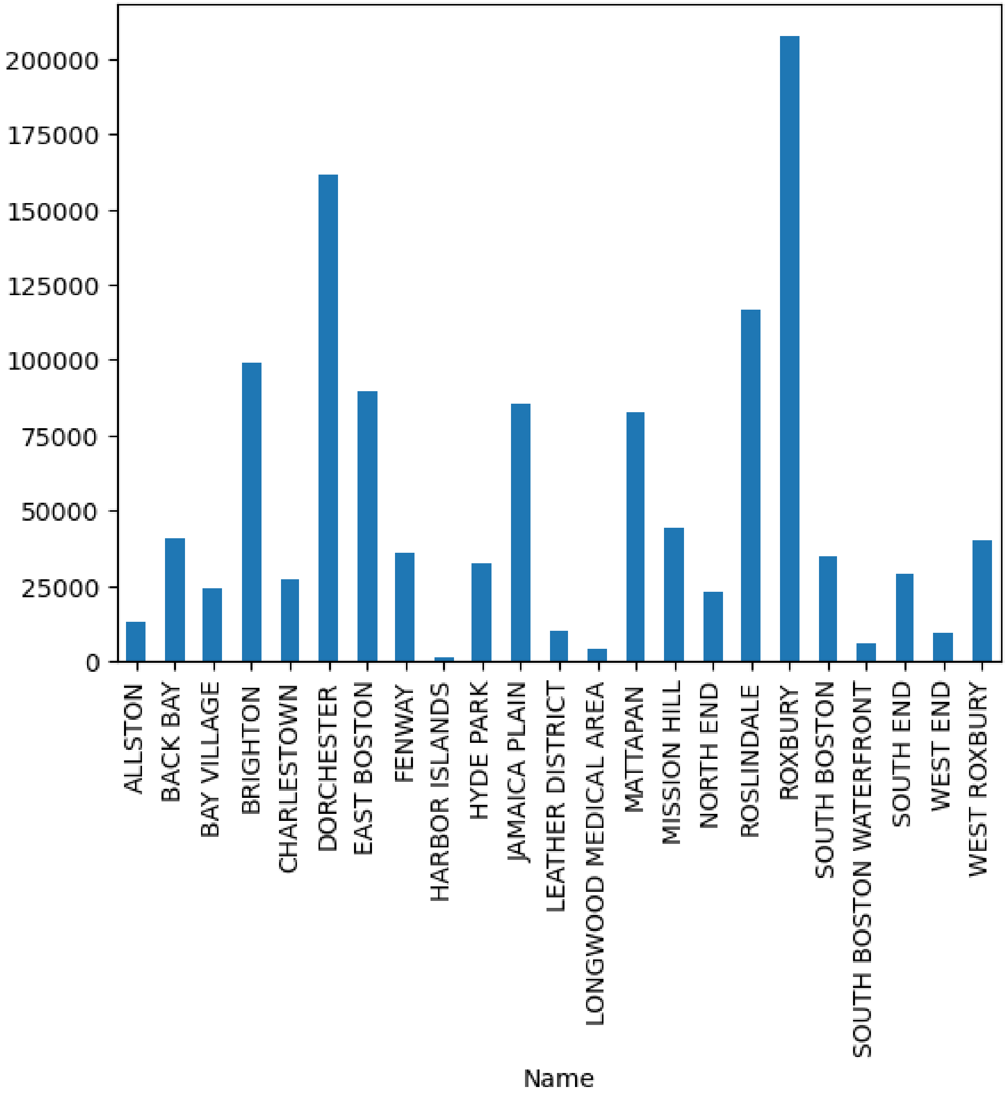


Fig 34. Total Vulnerable Population per Neighborhood in Boston City

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 35 (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.

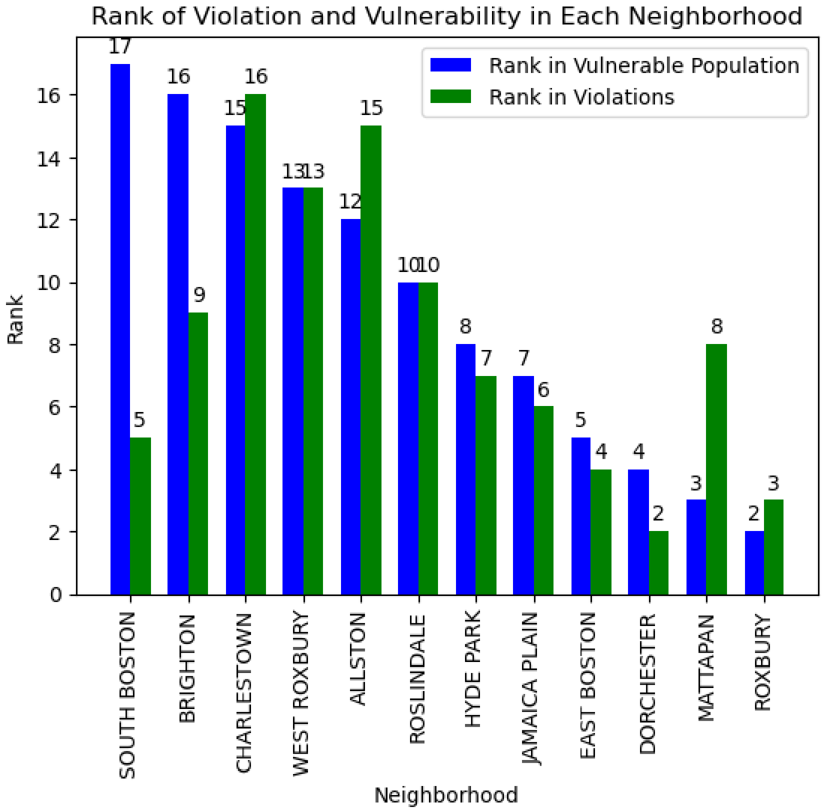


Fig 35. Neighborhood vs Rank for Vulnerable Population and Number of Violations

According to the correlation analysis we ran between Social Vulnerability and the Violation Scores of each neighborhood, there is a moderate to strong correlation between these two parameters. The correlation score between these two parameters was 0.527. Fig 36 showcases this correlation in graph form.

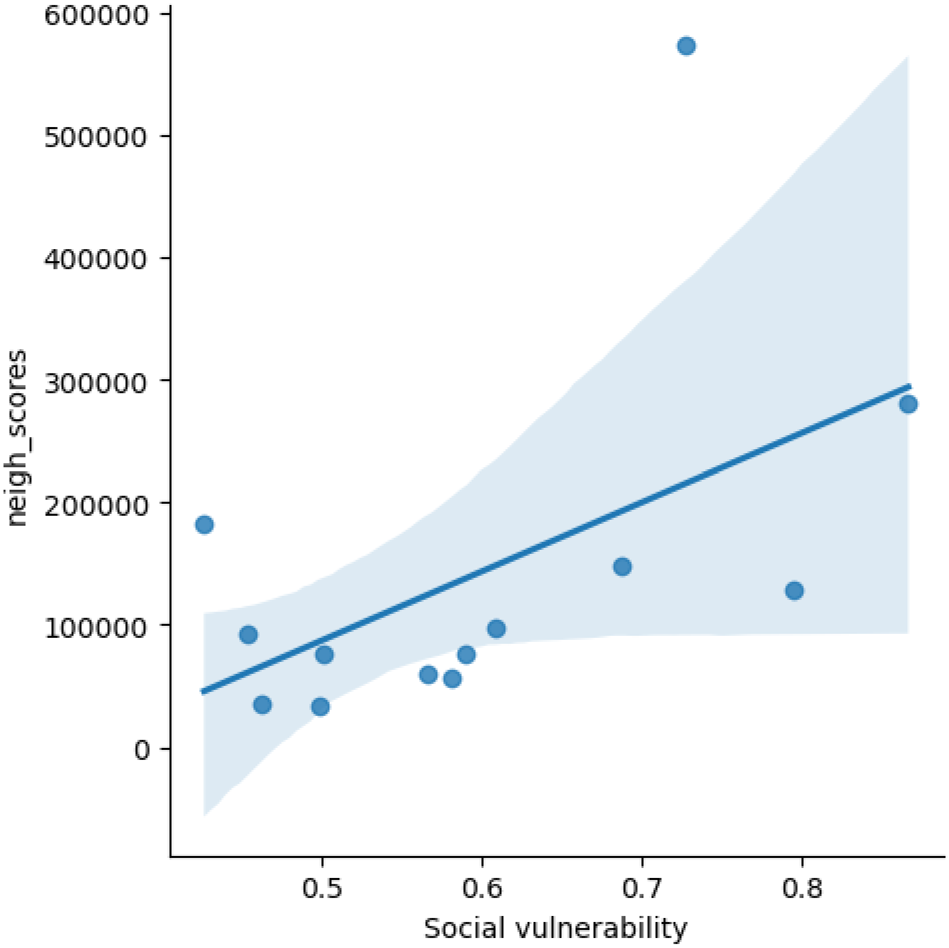


Fig 36. Correlation between Social Vulnerability Score and Violation Score

Similarly, we also ran a correlation analysis for the parameters of Population and Violation Scores of each neighborhood. These two parameters had a stronger correlation with a correlation score of 0.601. This stronger correlation might also be due to the fact that more population requires more houses, and more houses implies more chances of violations. Fig 37 showcases this correlation in graph form.

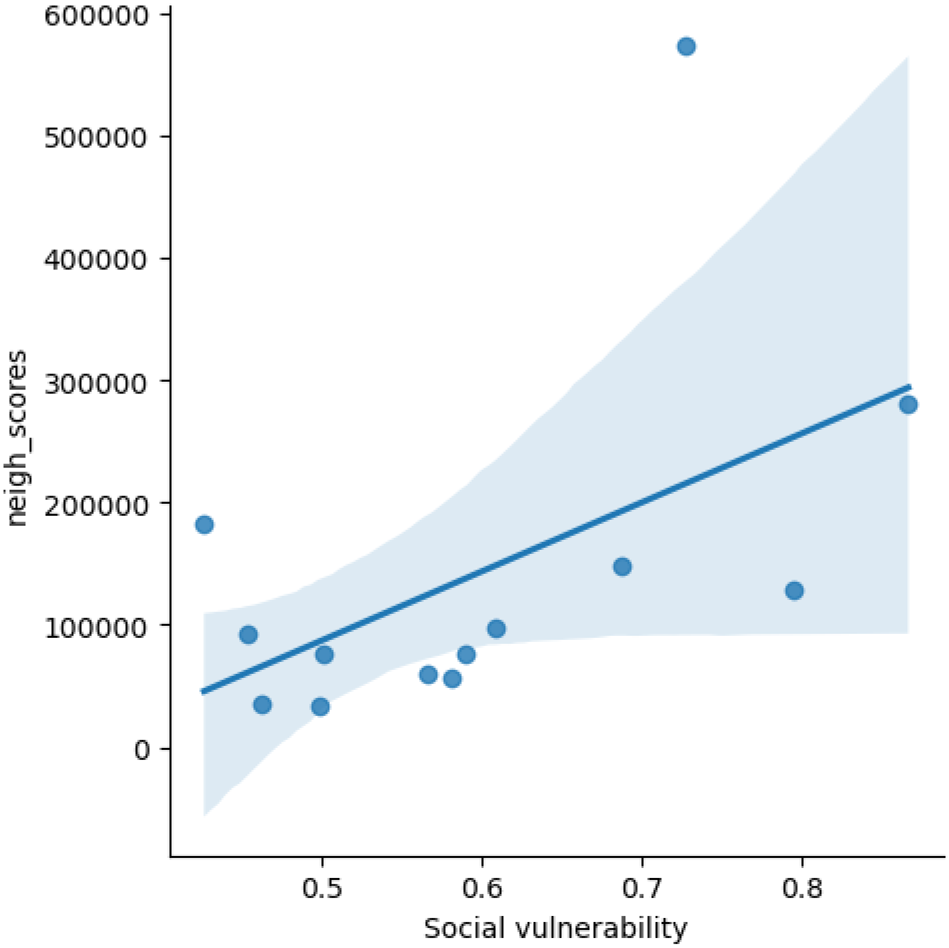


Fig 37. Correlation between Population and Violation Score

With the exception of Mission Hill, if we run a correlation analysis between the Violation Rank and the Vulnerability Rank of all neighborhoods, we see a strong correlation of 0.597. This correlation is reassuring of the fact that there is a core underlying relationship between the presence of vulnerable populations in a neighborhood and the bad behavior of landlords in that very neighborhood. The larger the percentage of vulnerable population in a neighborhood, the worse the landlord's behavior. Fig 38 depicts this relationship.

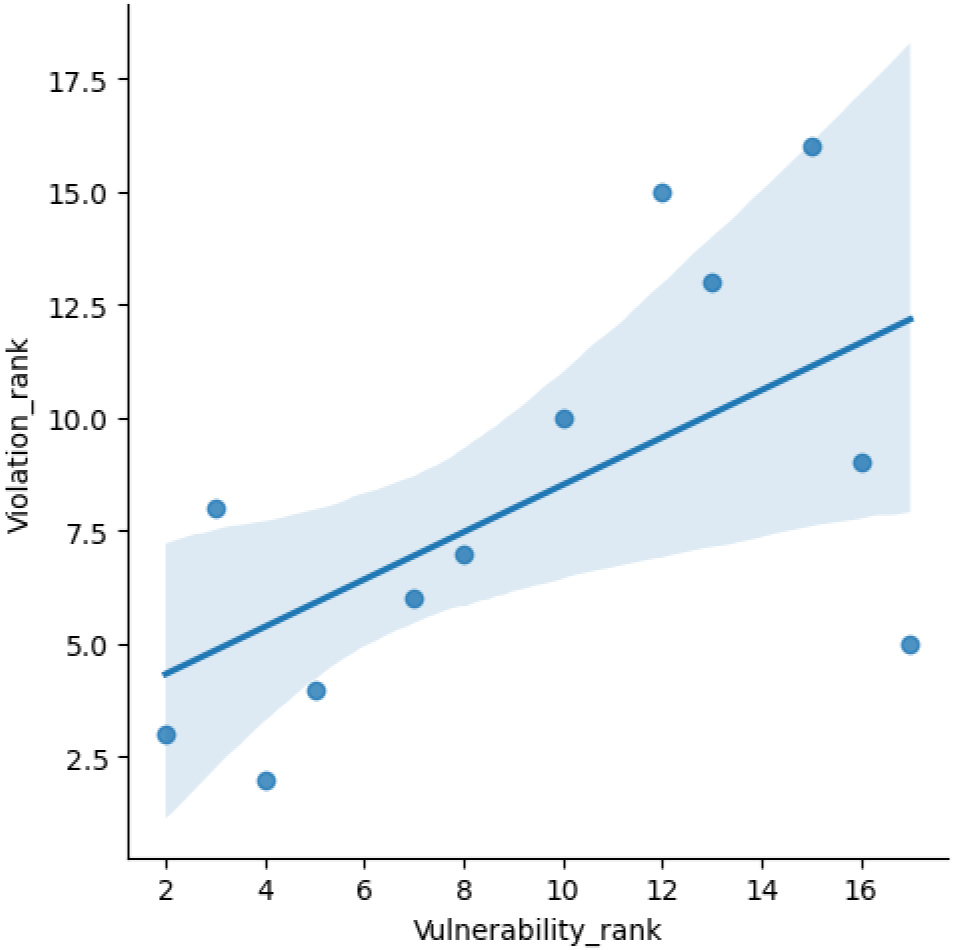


Fig 38. Correlation between Vulnerability Rank and Violation Rank of Neighborhoods

From the above graphs and correlation values, it can be clearly seen that there is an innate positive relationship between the vulnerability of a neighborhood and the number of violations (i.e. bad behavior of landlords) of that neighborhood. This relationship can further be studied in the future.

References

[1]
https://data.boston.gov/dataset/rentsmart/resource/dc615ff7-2ff3-416a-922b-f0f334f085d0 (Rentsmart Dataset)

[2]
https://data.boston.gov/dataset/311-service-requests (311 Service Requests Dataset)

[3]
https://data.boston.gov/dataset/building-and-property-violations1 (Building and Property Violations Dataset)

[4]
https://data.boston.gov/dataset/climate-ready-boston-social-vulnerability (Climate Ready Social Vulnerability Dataset)

[5]
https://data.boston.gov/dataset/2020-census-for-boston (2020 Boston Census Data)

[6]
https://data.boston.gov/dataset/2020-census-for-boston/resource/5800a0a2-6acd-41a3-9fe0-1bf7b038750d (Boston Neighborhood Dataset)

[7]
https://www.bostonplans.org/getattachment/86f801a8-f8a6-4d0c-83ed-b9a63684d6b5 (BPDA-Housing Tenure by Neighborhood)