Team 1 Deliverable 1:

Landlords analysis:
There are 5498 entries of evictions in MA with 5009 lines with the owner's name, including 3117 different owners/companies. 1292 of them are corporations, 1825 of them are not corporations. The most number of evictions with an owner is 76, the least is 1.
Top 5 number of evictions of an individual owner is shown below:
WORCESTER HOUSING AUTHORITY: 76
JEFFERSON AT EDGEWATER HILLS: 44
LINCOLN STREET REALTY COMPANY: 41
GPT-RG FALL RIVER LLC: 28
WINDSOR GARDENS PROPCO LLC: 27

For all owners, 3 owners with more than 30 evictions; 7 owners with 20-30 evictions; 26 owners with 10-20 evictions; 624 owners with 2-10 evictions; 2456 owners with 1 eviction.

For the owner who has more than 20 evictions, we figure out whether it is a corporation:
('WORCESTER HOUSING AUTHORITY', True)
('JEFFERSON AT EDGEWATER HILLS', False)
('LINCOLN STREET REALTY COMPANY', True)
('GPT-RG FALL RIVER LLC', True)
('WINDSOR GARDENS PROPCO LLC', True)
('GS STONEGATE PROJECT OWNER LLC', True)
('FRAMINGHAM HOUSING AUTHORITY', True)
('600 MAIN STREET,LLC', True)
('ARCHSTONE NORTH POINT I LLC,', True)
('CLARENDON HILL SOMERVILLE LP', True)

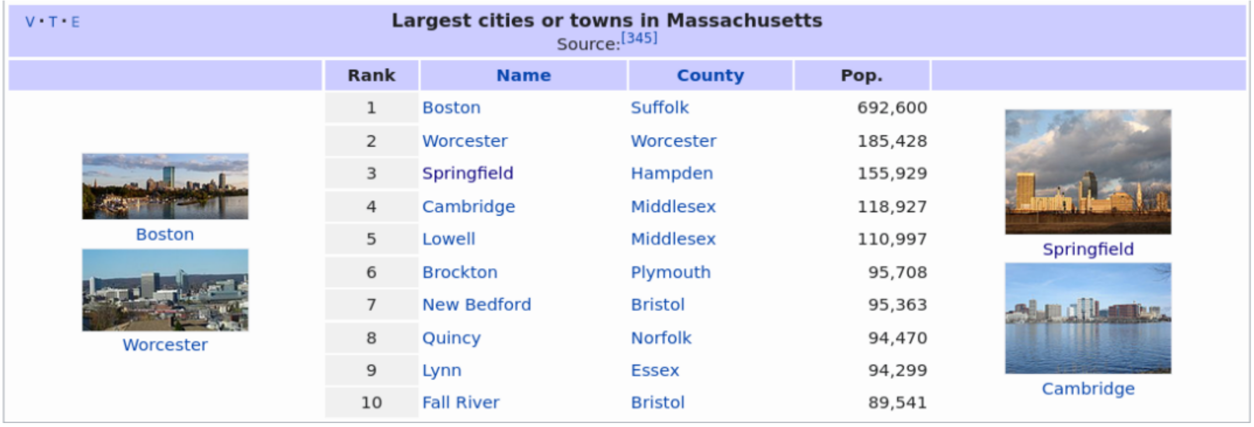
We can see except JEFFERSON AT EDGEWATER HILLS, all other owners are corporations.

Municipality Evictions Analysis:
There were 243 distinct municipalities with at least 1 eviction in the given dataset. Exactly 10 municipalities have more than 100 evictions and they are as follows :

Municipality	**Count**
Worcester	576
Boston	483
Fall River	354
New Bedford	293
Springfield	171
Framingham	169
Fitchburg	121
Brockton	119
Lynn	115

Lowell	115

Upon looking at this list, we immediately compared it with the most populated cities of Massachusetts.



V • T • E
Boston
Worcester
Springfield
Cambridge

Largest cities or towns in Massachusetts
Source:[345]

Rank	Name	County	Pop.
1	Boston	Suffolk	692,600
2	Worcester	Worcester	185,428
3	Springfield	Hampden	155,929
4	Cambridge	Middlesex	118,927
5	Lowell	Middlesex	110,997
6	Brockton	Plymouth	95,708
7	New Bedford	Bristol	95,363
8	Quincy	Norfolk	94,470
9	Lynn	Essex	94,299
10	Fall River	Bristol	89,541

Except for the absence of Cambridge and Quincy, we notice that the municipalities with the most evictions are in sync with the top 10 most populated cities of Massachusetts. This suggests that there is a correlation between population and eviction numbers.

44 municipalities have exactly one eviction and 168 municipalities have less than 10 evictions. 65 municipalities have between 10 and 100 evictions.

Preliminary analysis:

Helpful columns:

'muni': Usage: Eviction rate analysis

'owner_name': Name of property owner. Usage: evictee profile analysis

'geometry': Altitude and latitude extraction. Usage: geopandas

'corp': bool value for whether the owner is corp or not. Usage: evictee profile analysis

'ARC_Street': Street of property

'poly_typ': FEE, TAX, etc. Usage: evicted profile analysis

'ARC_ZIP': area zip code. Usage: Eviction rate analysis

'case_numbe': Usage: extract more information online based on case number

'court_divi': Usage: eviction rate analysis

'file_date': date eviction was filed. Usage: different moratorium milestones (e.g. CDC expiration, state expiration, city expiration) and it’s impact

'yr_built': year the property was built. Usage: maybe show gentrification trends. High eviction + new date

'ls_price': last sold price. Usage: evicted profile analysis

'ls_date': last sell date. Usage: evicted profile analysis ‘propert_3’: Size of property. Usage: evicted profile analysis. Family size high + property area low = more help needed

‘Initiation’: Reasons for evicting. Usage: evicted profile analysis

‘realesttyp’: One-digit MAPC-assigned use category for all properties based on land use codes and FAR (density values) (fig. 2). Usage: Evicted profile analysis. Business or home.

'STATEFP': State FIPS Code

'COUNTYFP': County Code

'TRACTCE': Census Tract Code

'AFFGEOID': Unknown

'GEOID': Metropolitan statistical area/micropolitan statistical area code. GEOIDs are numeric codes that uniquely identify all administrative/legal and statistical geographic areas for which the Census Bureau tabulates data.

'NAME': Same as TRACTCE

'LSAD': The legal/statistical area description (LSAD) codes describe the particular typology for each geographic entity. CT means census tract.

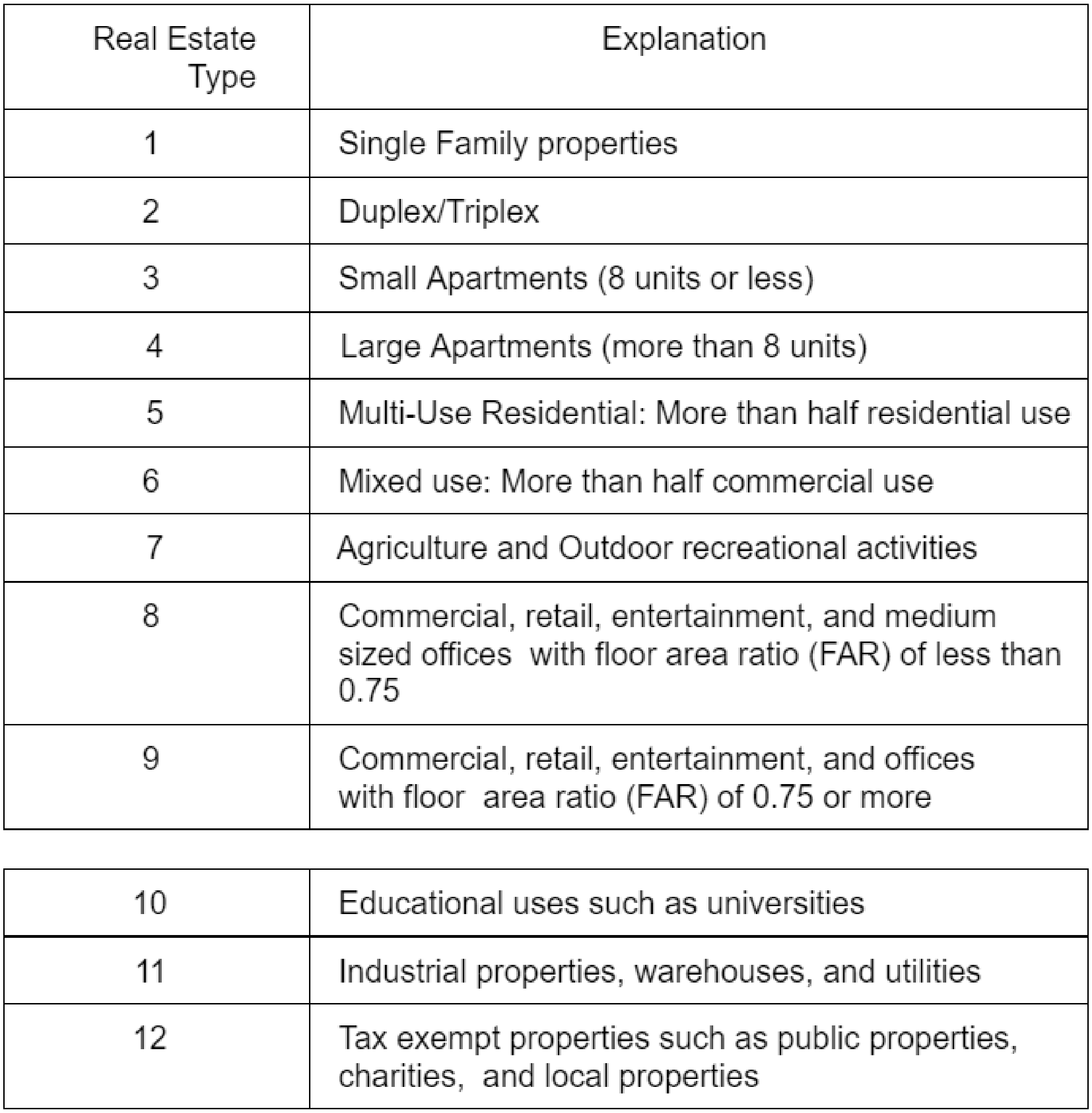
'ALAND': Land Area (square meters)

'AWATER': Water Area (square meters)

‘Index_right’: Unknown

Columns unknown:

1. Unsure of the differences between these columns and their uses. 'Mapc_id', 'parloc_id', 'map_num', 'mappar_id', 'loc_id_cnt', 'luc_1', 'luc_2', 'luc_adj_1', 'luc_adj_2'
2. Unsure if this has to do with the rate of price change overtime. Maybe helpful for figuring out increased prices. 'pct_imperv', 'pct_bldg', 'bldgv_psf', 'bldlnd_rat', 'sqm_imperv', 'sqm_pave’
3. Not in MA_Parcel database: 'temp', 'st_area_sh', 'st_length_'



Real Estate Type	Explanation
1	Single Family properties
2	Duplex/Triplex
3	Small Apartments (8 units or less)
4	Large Apartments (more than 8 units)
5	Multi-Use Residential: More than half residential use
6	Mixed use: More than half commercial use
7	Agriculture and Outdoor recreational activities
8	Commercial, retail, entertainment, and medium sized offices with floor area ratio (FAR) of less than 0.75
9	Commercial, retail, entertainment, and offices with floor area ratio (FAR) of 0.75 or more
10	Educational uses such as universities
11	Industrial properties, warehouses, and utilities
12	Tax exempt properties such as public properties, charities, and local properties

Fig. 2: realestate type

The purpose of analysing the database is to better understand how to accommodate those evicted and whether certain areas require more assistance and why.

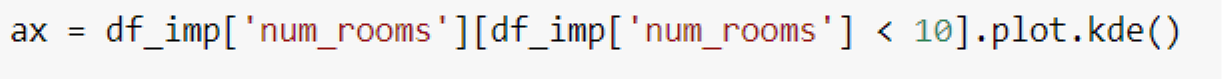
With that being said, I think understanding correlations between race, family composition, property price, property price fluctuations, house size, year_built, year sold, and room numbers is an important aspect of this analysis. I think in order to accommodate them. These details will allow us to understand poverty trends in a certain area. It might be that their family composition is large and their home is small. It means they are in a higher urgent need for help. It would be weighted more than other cases. The room number can be removed. Property_3 columns hold the size of the place evicted.

There are a number of duplicates or columns that mean similar things. For example, owner_name, clean_name, and entityname essentially mean the same thing. However, there are slight differences in naming. Therefore, they are not complete duplicates. Furthermore, there are about 8 columns thare state the address with the same missing value numbers. Therefore, I used all non-repeated address columns.

Some of the columns used here can be used to understand whether the area is poverty-stricken. I am unsure of whether to manually estimate it from this dataset or use census data for poverty rates per area. For example, some columns mention building density. That is not a good example as it could just be a strategic spot like Downtown Boston. Furthermore, if detailed information like the eviction experience is like quitting. It can be estimated using the data that we have and demographic census data like education or job type. For example, area with low education and high eviction rate could mean

Missing values:
There are about 2000 case numbers out of 7000 that do not have corresponding MA_PARCEL data. This includes geometric data, where instead of NaN, missing data is written as POINT(-inf, inf). This could make our analysis not very representative of reality. Team 2 is possibly working on filling such information.

Fig. 1 shows that the overwhelming number of rooms in num_room column are 0 as this distribution shows even after removing all 2000 NaN values. It could be that it is a studio. However, it should say 1 instead. This would make it hard to find this column useful. In Fig. 1, I am showing the values between 0 to 10. Therefore, the meaningless values are 0, which is 1977 rows and -0.9999, 7 rows. Since it is an overwhelming number of rows, it would not be good to replace it with the median. To conclude, this row would be meaningless for our analysis and we could use census data for poverty rate estimations per zip code or county (see this link for details: https://www.census.gov/quickfacts/fact/table/MA#).



```
ax = df_imp['num_rooms'][df_imp['num_rooms'] < 10].plot.kde()
```

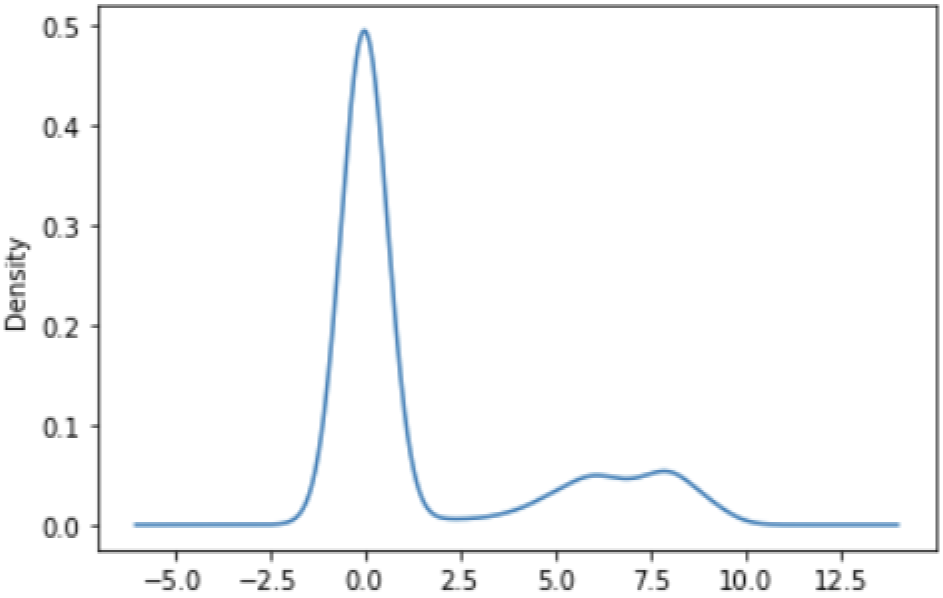
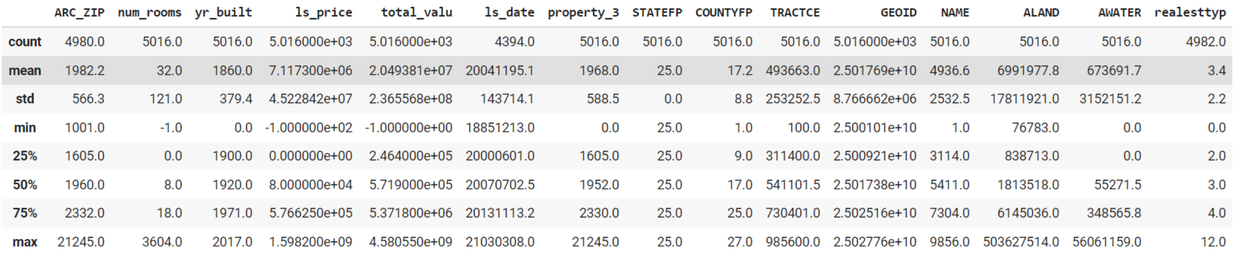


Fig. 1: Density of num_room between 0 and 10

Much like num_rooms, a significant number of rows for Ls_price is 0, 1355 rows, and negative values, 10. It would be difficult to fill the values with the median or other statistical methods. Thus, this column will be discarded as well. We can extract possible prices from the total building price/ estimated number of units. Another way is price per square ft multiplied by property size. The same trend is shown for property_3. Thus, it would be discarded or ignored for now, and would wait for a full dataset.

Next step:
Identify outliers. For example, statistical significance of small population groups or ways to group small populations together.

<u>Numeric column analysis(range, max, mean…):</u>

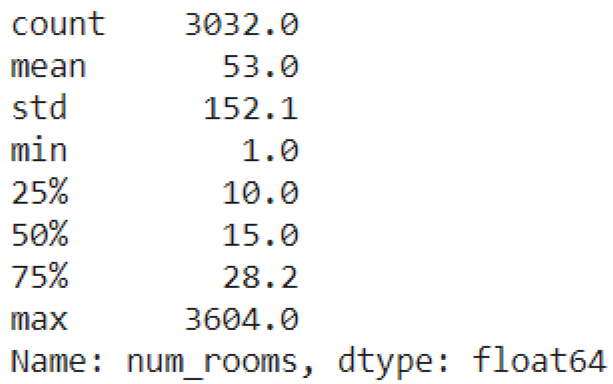


	ARC_ZIP	num_rooms	yr_built	ls_price	total_valu	ls_date	property_3	STATEFP	COUNTYFP	TRACTCE	GEOID	NAME	ALAND	AWATER	realesttyp
count	4980.0	5016.0	5016.0	5.016000e+03	5.016000e+03	4394.0	5016.0	5016.0	5016.0	5016.0	5.016000e+03	5016.0	5016.0	5016.0	4982.0
mean	1982.2	32.0	1860.0	7.117300e+06	2.049381e+07	20041195.1	1968.0	25.0	17.2	493663.0	2.501769e+10	4936.6	6991977.8	673691.7	3.4
std	566.3	121.0	379.4	4.522842e+07	2.365568e+08	143714.1	588.5	0.0	8.8	253252.5	8.766662e+06	2532.5	17811921.0	3152151.2	2.2
min	1001.0	-1.0	0.0	-1.000000e+02	-1.000000e+00	18851213.0	0.0	25.0	1.0	100.0	2.500101e+10	1.0	76783.0	0.0	0.0
25%	1605.0	0.0	1900.0	0.000000e+00	2.464000e+05	20000601.0	1605.0	25.0	9.0	311400.0	2.500921e+10	3114.0	838713.0	0.0	2.0
50%	1960.0	8.0	1920.0	8.000000e+04	5.719000e+05	20070702.5	1952.0	25.0	17.0	541101.5	2.501738e+10	5411.0	1813518.0	55271.5	3.0
75%	2332.0	18.0	1971.0	5.766250e+05	5.371800e+06	20131113.2	2330.0	25.0	25.0	730401.0	2.502516e+10	7304.0	6145036.0	348565.8	4.0
max	21245.0	3604.0	2017.0	1.598200e+09	4.580550e+09	21030308.0	21245.0	25.0	27.0	985600.0	2.502776e+10	9856.0	503627514.0	56061159.0	12.0

As shown, Yr_built, Is_price, property_3, realesttype has meaningless data.

Here is description for each one after data cleaning (non-representative due to many missing information):

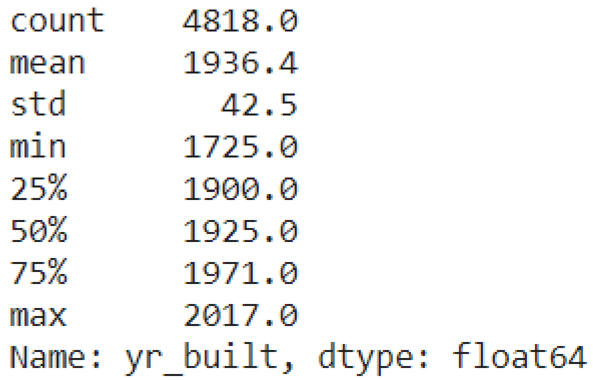
Num_rooms:



```
count    3032.0
mean       53.0
std       152.1
min         1.0
25%        10.0
50%        15.0
75%        28.2
max      3604.0
Name: num_rooms, dtype: float64
```

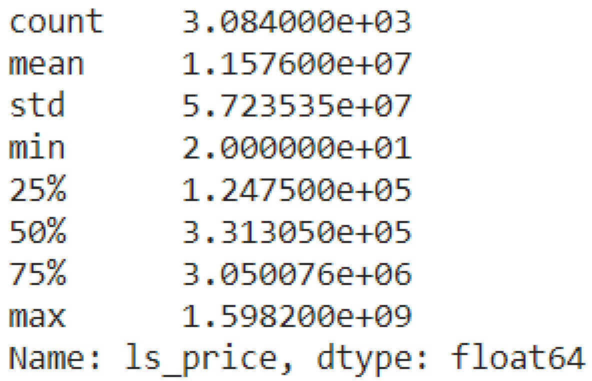
Very high std. There are some very large properties, possibly businesses.

Yr_built:



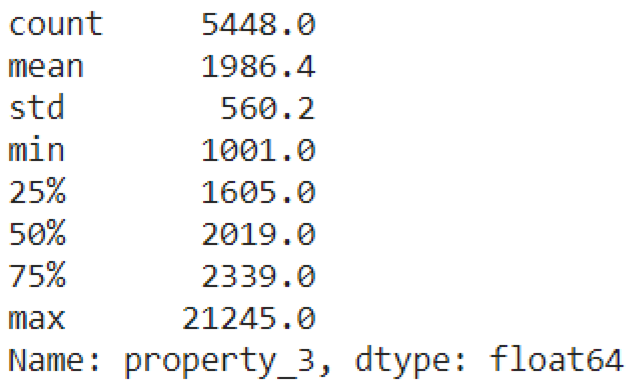
```
count    4818.0
mean     1936.4
std        42.5
min      1725.0
25%      1900.0
50%      1925.0
75%      1971.0
max      2017.0
Name: yr_built, dtype: float64
```

Ls_price:



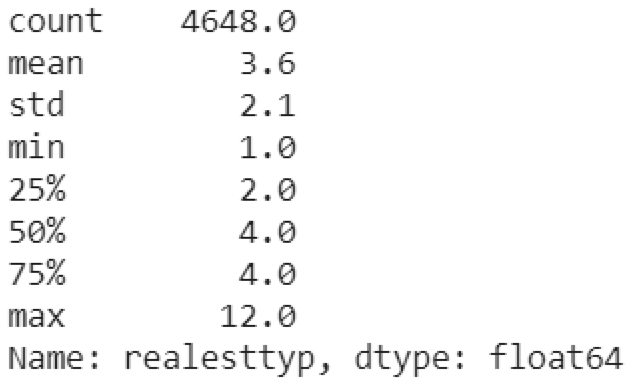
```
count    3.084000e+03
mean     1.157600e+07
std      5.723535e+07
min      2.000000e+01
25%      1.247500e+05
50%      3.313050e+05
75%      3.050076e+06
max      1.598200e+09
Name: ls_price, dtype: float64
```

Property_3:



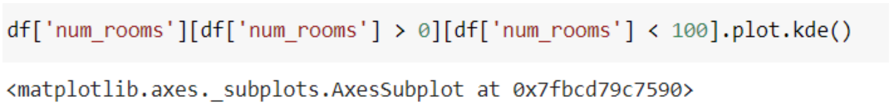
```
count     5448.0
mean      1986.4
std        560.2
min       1001.0
25%       1605.0
50%       2019.0
75%       2339.0
max      21245.0
Name: property_3, dtype: float64
```

Realesttype:
(this may not be representative as there are many unknown values, 4000 rows)



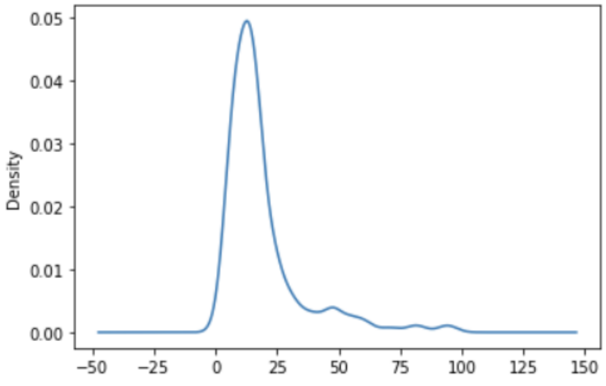
```
count    4648.0
mean        3.6
std         2.1
min         1.0
25%         2.0
50%         4.0
75%         4.0
max        12.0
Name: realesttyp, dtype: float64
```

The median is 4.
This means the median are large apartments.
<u>Basic Distributions:</u>



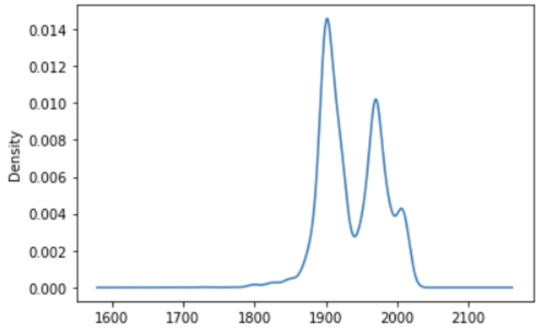
```
df['num_rooms'][df['num_rooms'] > 0][df['num_rooms'] < 100].plot.kde()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd79c7590>
```



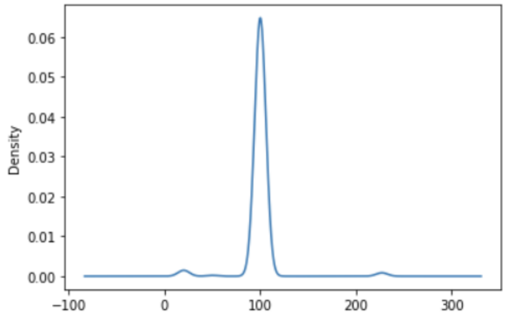
```
df['yr_built'][df['yr_built'] > 0][df['yr_built'] < 2017].plot.kde()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd799df50>
```



```
df['ls_price'][df['ls_price'] > 10][df['ls_price'] < 400].plot.kde()
```

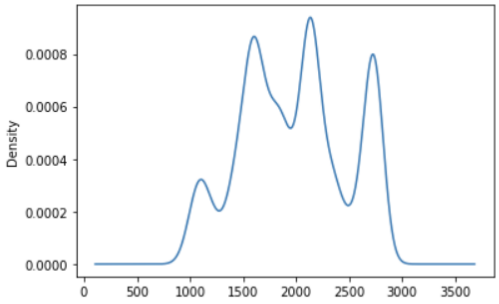
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd793f310>
```

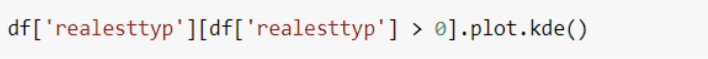


Many outliers in ls_price. Most of the prices are low.

```
df['property_3'][df['property_3'] > 0][df['property_3'] < 20000].plot.kde()
```

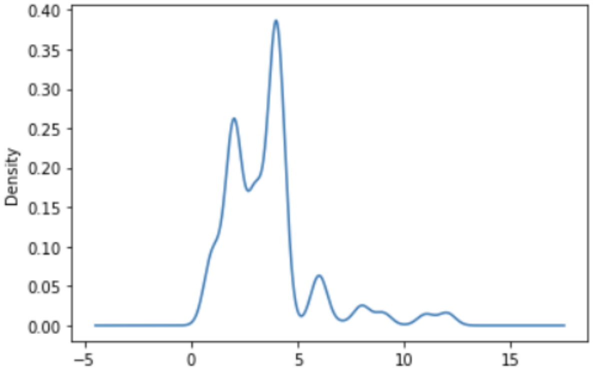
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd7373490>
```

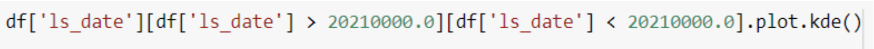




```
df['realesttyp'][df['realesttyp'] > 0].plot.kde()
```

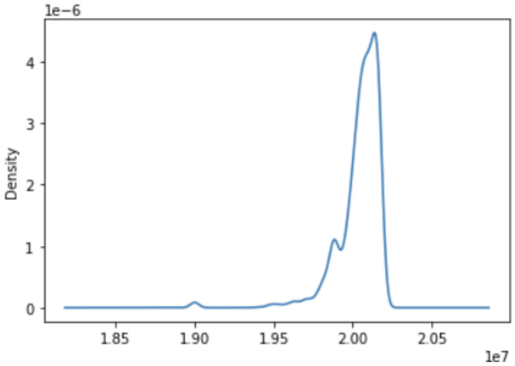
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd72c5550>
```





```
df['ls_date'][df['ls_date'] > 20210000.0][df['ls_date'] < 20210000.0].plot.kde()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcd7116450>
```



Most are very recent. There was some meaningless data like over 2022.

Future analysis:

We would like to analyze evictions based on other factors such as :

1.) House profile(size,locality,etc)
2.) Migration patterns

House profile :

We would like to understand various causes for the filing of an eviction case. Socio-economic factors like locality, house type, house size, proximity of schools etc,can help identify the rationale behind an eviction.

Migration :

Given the pandemic, we would like to see if migrations in and out of the state would've caused eviction conditions directly or indirectly. Intuitively, when there is an exodus, prices of housing go down. Since classes were remote only for many universities across the state and country in early 2020 (hybrid model in Fall), this would mean that out of state students wouldn't be present. This could cause prices to go down and thus, might indirectly affect eviction rates.

Other analysis:
Given the 10 cities listed earlier, we would like to see if the amount of evictions is abnormal given the populations. Initial analysis seems to suggest that the municipalities of Fitchburg and Framingham might be anomalies, especially given that Cambridge and Quincy have fewer evictions than the abovementioned 2 municipalities.

Problems:
A significant problem we have is knowing how to work with the census data. The columns are very confusing, and we're not certain about their meanings. We don't know what the numerical values in the census data signify. It would help if the meaning of the census data would be clarified, as well as knowing how exactly it is supposed to help us. Are we using this census data for deducting evictor profiles, or density of evictions in certain areas? Also, additional information on price changes over time would help.