

Assignment #1 – DQR of NY Property Data

- Dhvani Kapadia

High-Level Description:

This report consists of a preliminary quantitative analysis of the Property Valuation and Assessment Data derived from NYC OpenData. It primarily covers data from November 2010.

- There are a total 1,070,994 rows and 32 columns.
- There is a mix of numerical and categorical variables.

Following table summarizes characteristics of different variables, 31 columns, present in the NY Property Data ('Record' variable has been discarded as it is just a unique identifier).

Table of Summary Characteristics of Variables

- **Numerical Variables**

Field Name	Field Type	# Records that have a value	% Populated	# Unique values	# Records with value 0	Min	Max	Mean	Standard deviation
LTFRONT	Numerical	1070994	100	1297	169108	0	9999	36.6	74
LTDEPTH	Numerical	1070994	100	1370	170128	0	9999	8.88	76.3
STORIES	Numerical	1014730	94.74	112	0	1	119	5	8.36
FULLVAL	Numerical	1070994	100	109324	13007	0	6.15e+09	8.7e+05	1.15e+07
AVLAND	Numerical	1070994	100	70921	13009	0	2.6e+09	8.5e+04	4.0e+06
AVTOT	Numerical	1070994	100	112914	13007	0	4.66e+09	2.27e+05	6.87e+06
EXLAND	Numerical	1070994	100	33419	491699	0	2.6e+09	3.6e+04	3.9e+04
EXTOT	Numerical	1070994	100	64255	432572	0	4.6e+09	9.1e+04	6.5e+06
BLDFRONT	Numerical	1070994	100	612	228815	0	7575	23.04	35.57
BLDDEPTH	Numerical	1070994	100	621	228853	0	9393	39.92	42.70
AVLAND2	Numerical	282726	26.4	58592	0	3	2.37e+09	2.46e+05	6.17e+06
AVTOT2	Numerical	282732	26.4	111361	0	3	4.5e+09	7.13e+05	1.16e+07
EXLAND2	Numerical	87449	8.17	22196	0	1	2.37e+09	3.51e+05	1.08e+07
EXTOT2	Numerical	130828	12.2	48349	0	7	4.5e+09	6.56e+05	1.60e+07

- **Categorical Variables**

Field Name	Field Type	# records that have value	% Populated	# unique values	#records with zero value	Most common value (MCV)	Freq. of MCV
BBLE	Categorical	1070994	100	1070994	0	NA	NA
B	Categorical	1070994	100	5	0	4	358046
BLOCK	Categorical	1070994	100	13984	0	3944	3888
LOT	Categorical	1070994	100	6366	0	1	24367
EASEMENT	Categorical	4636	0.43	13	0	E	4148
OWNER	Text	1039249	97	863346	0	PARKCHESTER PRESERVAT	6020
BLDGCL	Categorical	1070994	100	200	0	R4	139879
TAXCLASS	Categorical	1070994	100	11	0	1	660721
EXT	Categorical	354305	33.08	4	0	G	266970
EXCD1	Categorical	638488	59.62	130		1017	425348
STADDR	Categorical	1070318	99.93	839280	0	501 Surf Avenue	902
ZIP	Categorical	1041104	97.21	197	0	10314	24606
EXMPTCL	Categorical	15579	1.45	15	0	X1	6912
EXCD2	Categorical	92948	8.68	61	0	1017	65777
PERIOD	Categorical	1070994	100	1	0	FINAL	1070994
YEAR	DateTime	1070994	100	1	0	2010/11	1070994
VALTYPE	Categorical	1070994	100	1	0	AC-TR	1070994

Now, we will explore in detail, the characteristics of each variable by plotting various graphs based on the category of the variable.

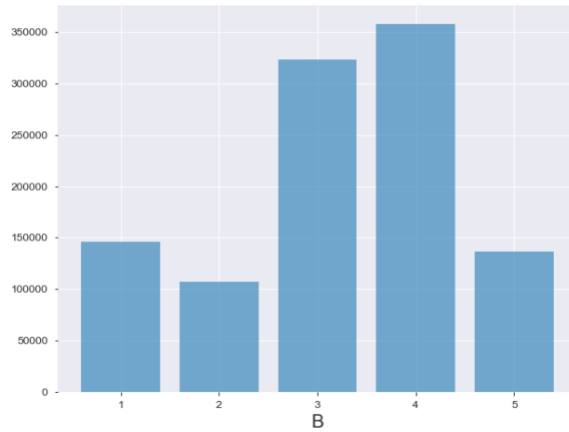
BBLE:

- Each of the column values under the 'BBLE' column has a unique value (Similar to 'RECORD').
- It is concatenation of AV_BORO, AV_BLOCK, AV_LOT and AV_EASEMENT. It is used to uniquely identify the area of each property

B:

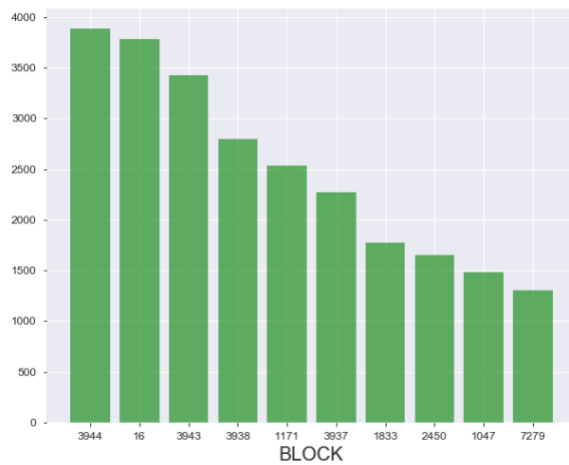
- B here means BORO Codes of the property.
- The following list denotes the meaning of each value under this variable:
 - 1: Manhattan

- 2: Bronx
- 3: Brooklyn
- 4: Queen's
- 5: Staten Island



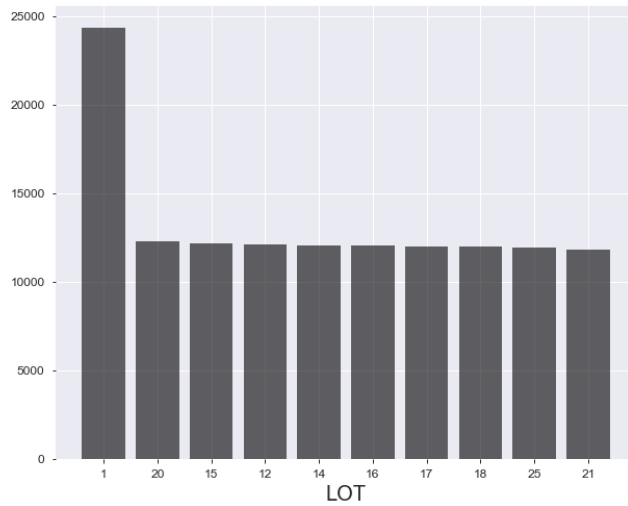
BLOCK:

- BLOCK denotes the block number (different for each BORO) of the property.
- The following is a valid range of BLOCK variables:
 - MANHATTAN - 1 TO 2,255
 - BRONX - 2,260 TO 5,958
 - BROOKLYN - 1 TO 8,955
 - QUEENS - 1 TO 16,350
 - STATEN ISLAND - 1 TO 8,050
- The following graph shows the top 10 BLOCKs:

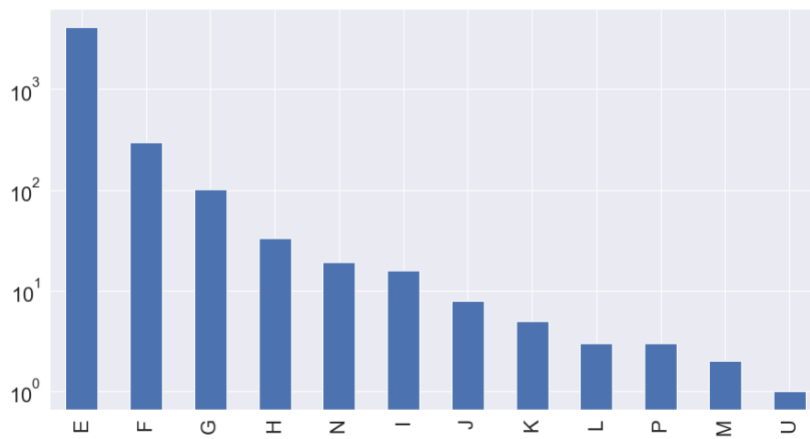


LOT:

- It is a unique area number within a BLOCK/BORO
- The following graph shows the top 10 LOTs:

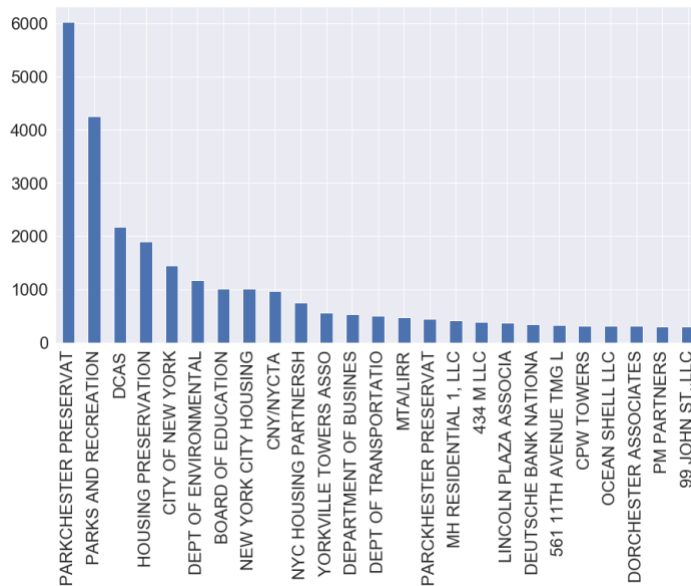


EASEMENT: It's a categorical variable. Following bar graph shows the spread of NY properties across different classes of 'EASEMENT' variable

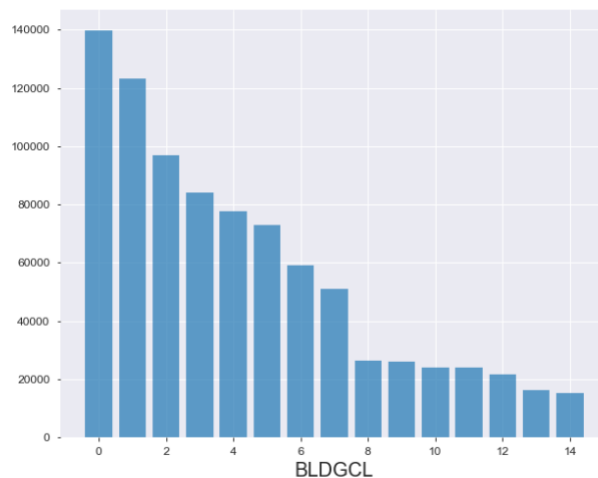


OWNER:

- This field has the name of the owner of the property.

**BLDGCL:**

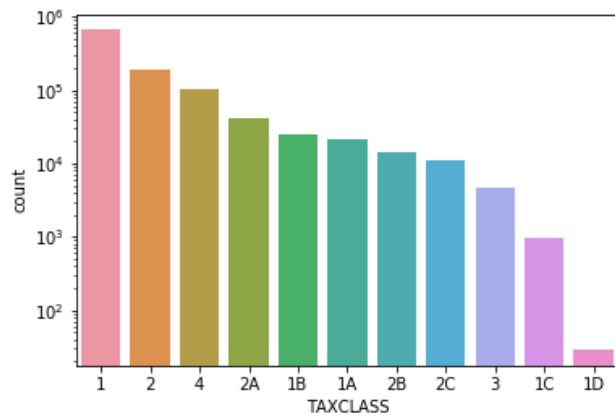
- It's a categorical variable that shows the class of the building
- The following are the top 15 classes in the New York data:

**TAXCLASS:**

- It corresponds to the Current Property Tax Class Code (NYS Classification). The following are valid values for this column:
 - 1 = 1-3 unit residences
 - 1a = 1-3 story condominiums originally a condo
 - 1b = residential vacant land
 - 1c = 1-3 unit condominiums originally tax class 1
 - 1d = select bungalow colonies
 - 2 = apartments

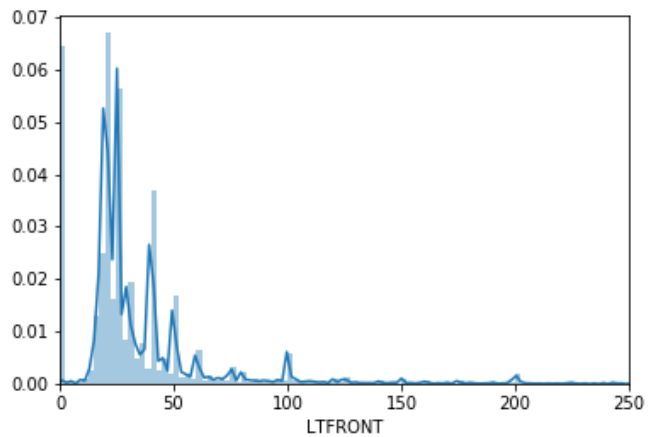
- 2a = apartments with 4-6 units
- 2b = apartments with 7-10 units
- 2c = coops/condos with 2-10 units
- 3 = utilities (except ceiling RR)
- 4a = utilities - ceiling railroads
- 4 = all others

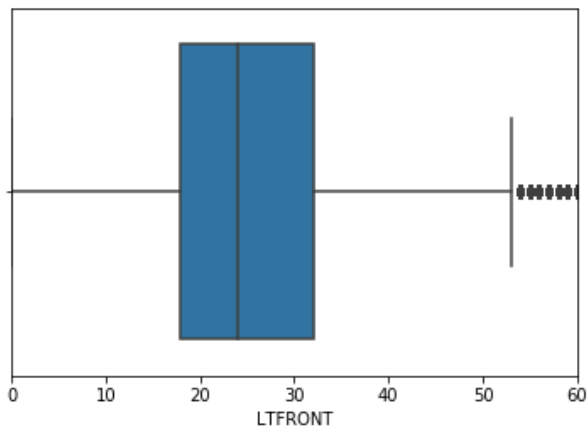
- The bar graph below shows the spread of NY properties across different classes of 'TAXCLASS' variable:



LTFRONT:

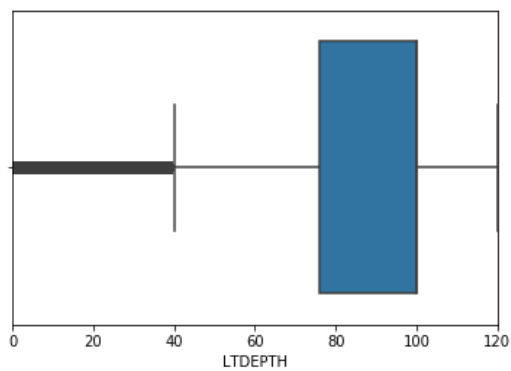
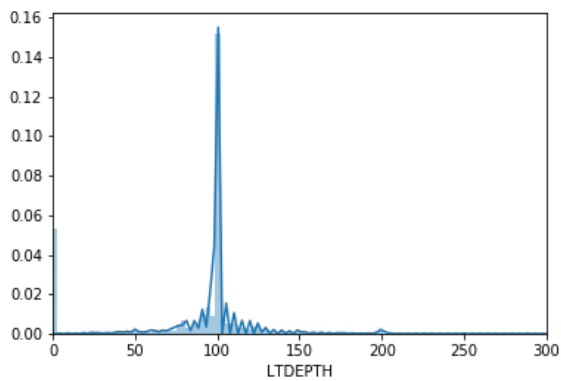
- It stands for Lot Frontage in feet.
- The Density plot and Boxplot below show the spread of values of 'LTFRONT' variable:





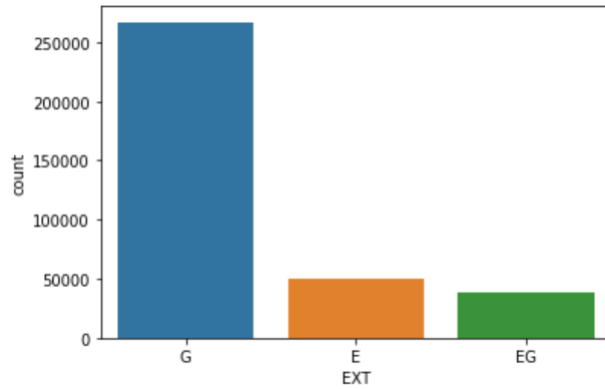
LTDEPTH:

- It stands for Lot Depth in feet.
- The Density plot and Boxplot below show the spread of values of 'LTDEPTH' variable:



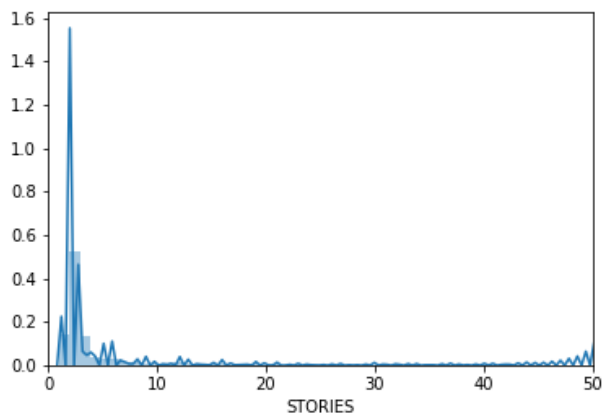
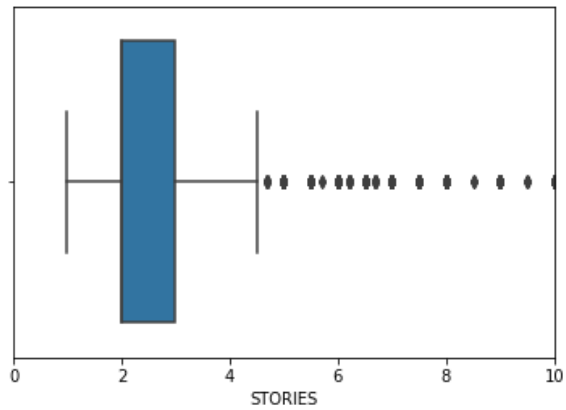
EXT:

- It's a categorical variable, meaning extension.
- Following bar graph shows the spread of properties across different classes of 'EXT' variable.



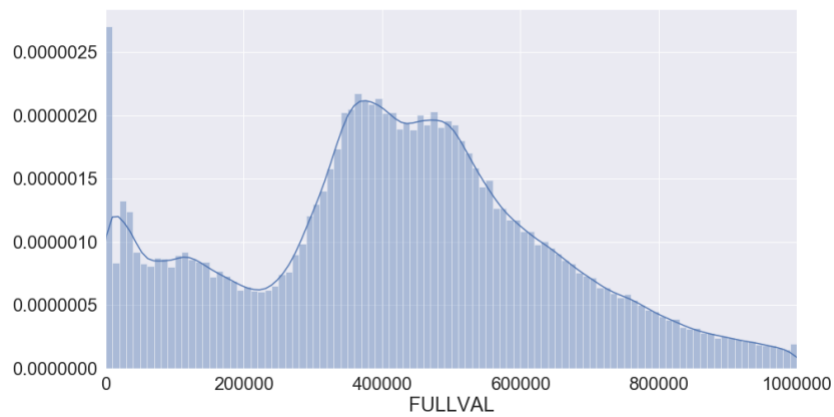
STORIES:

- It's a numerical variable showing the number of stories.
- Following Density plot and Boxplot show the spread of values of 'STORIES' variable



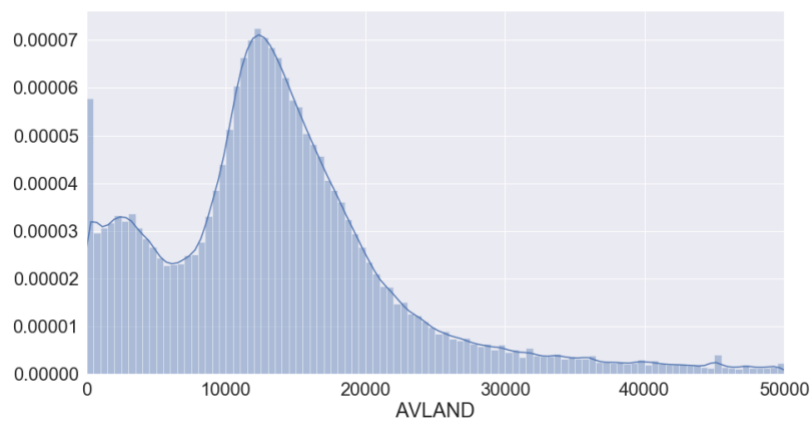
FULLVAL:

- It's a numerical variable showing the total value of the property.
- Following Density plot shows the distribution of values of 'FULLVAL' variable



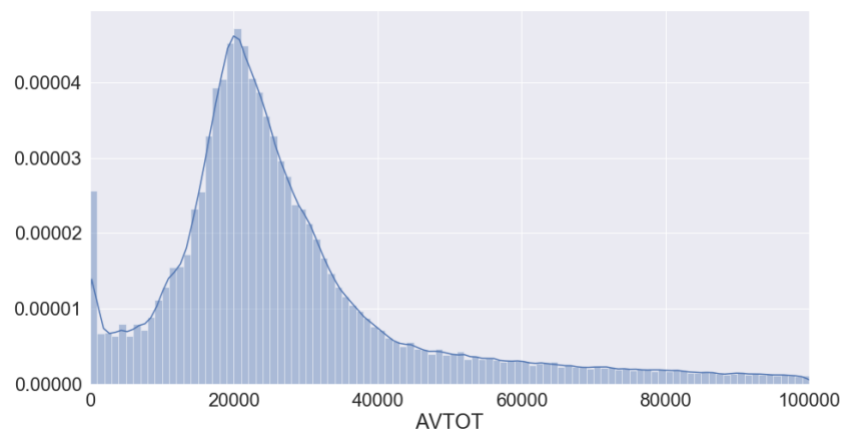
AVLAND:

- It's a numerical variable.
- Following Density plot shows the spread of values of 'AVLAND' variable



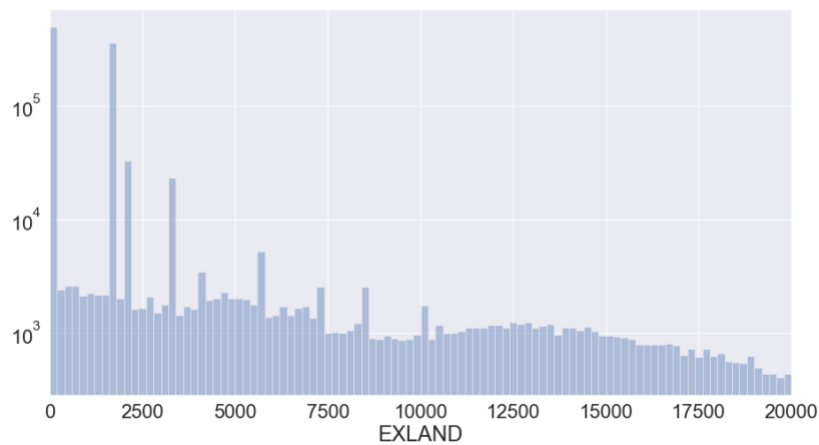
AVTOT:

- It's a numerical variable.
- Following density plot shows the spread of values of the 'AVTOT' variable

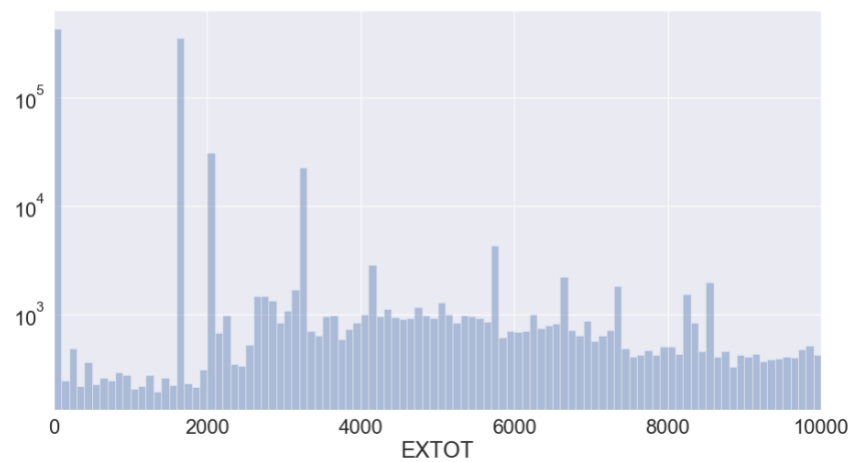


EXLAND:

- It's a numerical variable. Following density plot shows the spread of values of the 'EXLAND' variable

**EXTOT:**

- It's a numerical variable.
- Following histogram shows the spread of values of the 'EXTOT' variable

**EXCD1:**

- It's a categorical variable. Following table shows the spread of NY properties across different values of 'EXCD1' variable

Value	Count
-------	-------

1017.0	425348
1010.0	49756
1015.0	31323
5113.0	23858
1920.0	17594
5110.0	16834
5114.0	14984
5111.0	10609
1021.0	6613
1986.0	4231
5112.0	4071
1925.0	3566
6800.0	3494
1985.0	2471
2231.0	1866
1501.0	1453
3390.0	1425
2151.0	1243
1301.0	922
1101.0	824
5129.0	798
5101.0	787
2262.0	776
5130.0	723
1022.0	688

Name: EXCD1, dtype: int64

STADDR:

- It's a categorical variable.
- The following table shows the spread of NY properties across some of the top occurring addresses present in 'STADDR' variable

501 SURF AVENUE	902
330 EAST 38 STREET	817
322 WEST 57 STREET	720
155 WEST 68 STREET	671
20 WEST 64 STREET	657
1 IRVING PLACE	650
220 RIVERSIDE BOULEVARD	628
360 FURMAN STREET	599
200 EAST 66 STREET	585
30 WEST 63 STREET	562
2 BAY CLUB DRIVE	556
350 WEST 42 STREET	556
200 RECTOR PLACE	549
301 EAST 79 STREET	538
350 WEST 50 STREET	498
630 1 AVENUE	488
635 WEST 42 STREET	483
88 GREENWICH STREET	453
150 WEST 51 STREET	447
99 JOHN STREET	445
25 CENTRAL PARK WEST	441
138-35 ELDER AVENUE	437
1623 3 AVENUE	434
1 BAY CLUB DRIVE	427
5 EAST 22 STREET	426

Name: STADDR, dtype: int64

ZIP:

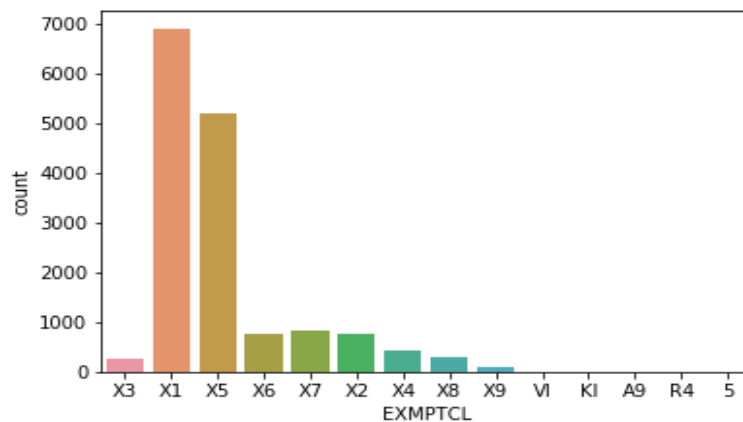
- It's a categorical variable, denoting the zip-code of the property.
- The following table shows the spread of NY properties across top 20 Zip codes

10314.0	24606
11234.0	20001
10312.0	18127
10462.0	16905
10306.0	16578
11236.0	15678
11385.0	14921
11229.0	12793
11211.0	12710
11207.0	12293
11215.0	11834
11235.0	11312
11203.0	11241
11208.0	11139
11204.0	11061
10469.0	11030
11214.0	10886
11223.0	10741
10305.0	10625
11434.0	10505
11355.0	10492
11219.0	10300
11357.0	9851
11413.0	9784
11373.0	9779

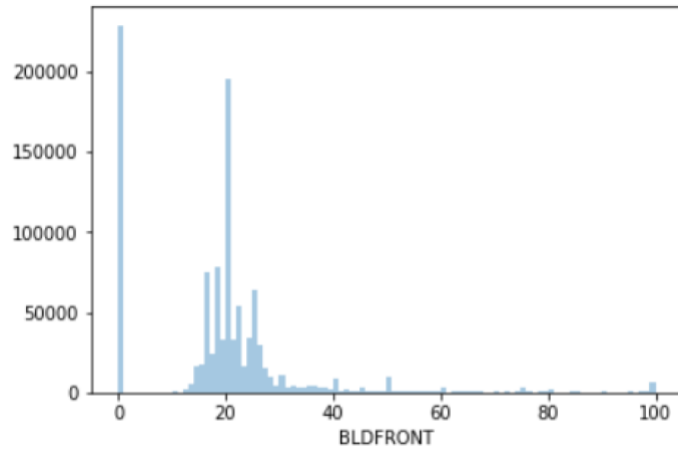
Name: ZIP, dtype: int64

EXMPTCL:

- It's a categorical variable used for fully exempt properties.
- Following bar graph shows the spread of properties across different classes of 'EXMPTCL' variable

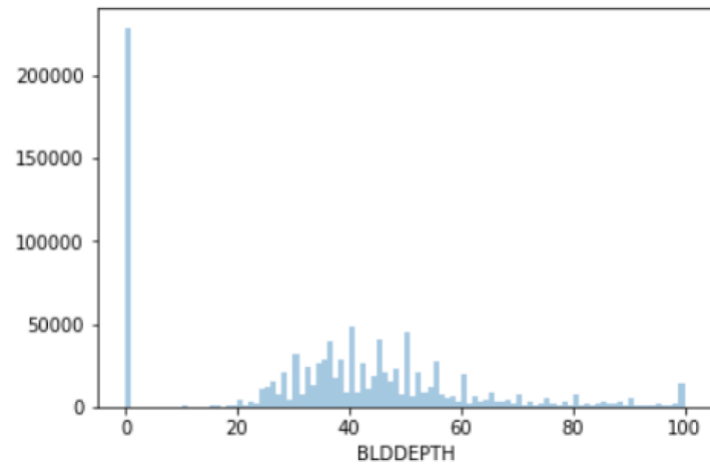
**BLDFRONT:**

- It's a numerical variable.
- Following histogram shows the spread of values of the 'BLDFRONT' variable



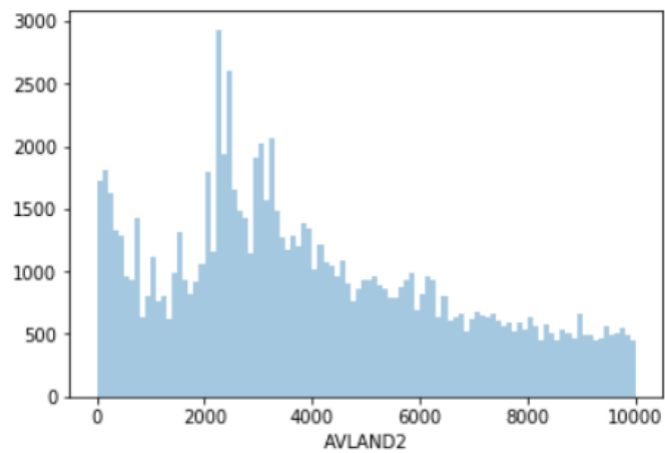
BLDDEPTH:

- It's a numerical variable, stands for depth of the building in feet.
- Following histogram shows the spread of values of the 'BLDDEPTH' variable
-



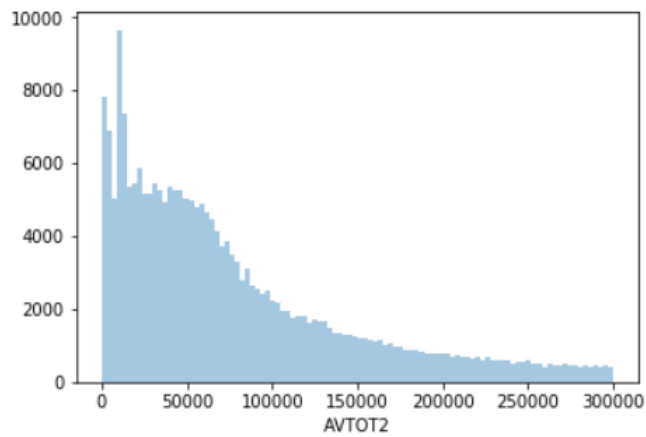
AVLAND2:

- It's a numerical variable.
- Following histogram shows the spread of values of the 'AVLAND2' variable

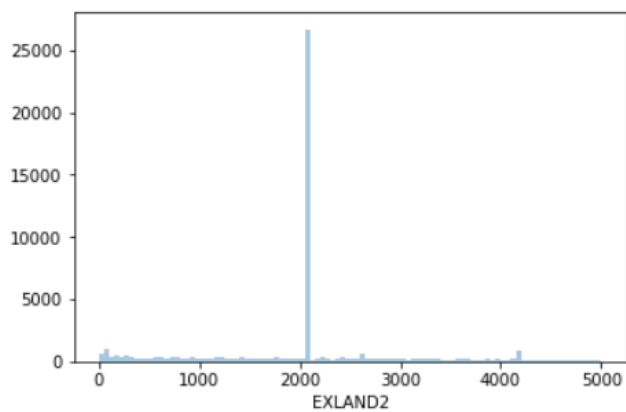


AVTOT2:

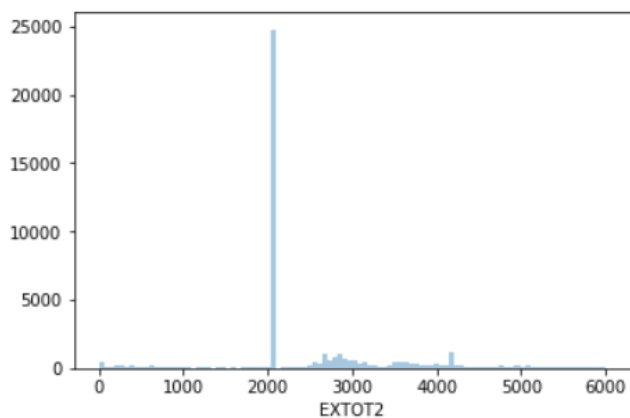
- It's a numerical variable.
- Following histogram shows the spread of values of the 'AVTOT2' variable

**EXLAND2:**

- It's a numerical variable.
- Following histogram shows the spread of values of the 'EXLAND2' variable

**EXTOT2:**

- It's a numerical variable.
- Following histogram shows the spread of values of the 'EXTOT2' variable



EXCD2:

- It's a categorical variable.
- Following table shows the spread of NY properties across different values of 'EXCD2' variable

1017.0	65777
1015.0	12337
5112.0	6867
1019.0	3178
1920.0	2961
1200.0	881
1101.0	494
5129.0	227
1986.0	35
1022.0	31
1985.0	21
1604.0	13
5109.0	11
1021.0	8
7160.0	7
2280.0	7
1523.0	7
2310.0	6
5113.0	6
5114.0	6

Name: EXCD2, dtype: int64

PERIOD:

- PERIOD variable has only 1 value across all the NY properties, which is 'FINAL'.

YEAR:

- YEAR variable has only 1 value across all the NY properties, which is '2010/11'.

VALTYPE:

- VALTYPE variable has only 1 value across all the NY properties, which is 'AC-TR'.