Assignment #1 – DQR of NY Property Data

**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 | 1.15 e+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 |
| **BLDFRON T** | Numerical | 1070994 | 100 | 612 | 228815 | 0 | 7575 | 23.04 | 35.57 |
| **BLDDEPT**  **H** | 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 | 6.56+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 | 35804  6 |
| **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 | 13987  9 |
| **TAXCLASS** | Categorical | 1070994 | 100 | 11 | 0 | 1 | 66072  1 |
| **EXT** | Categorical | 354305 | 33.08 | 4 | 0 | G | 26697  0 |
| **EXCD1** | Categorical | 638488 | 59.62 | 130 |  | 1017 | 42534  8 |
| **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 | 10709  94 |
| **YEAR** | DateTime | 1070994 | 100 | 1 | 0 | 2010/11 | 10709  94 |
| **VALTYPE** | Categorical | 1070994 | 100 | 1 | 0 | AC-TR | 10709  94 |

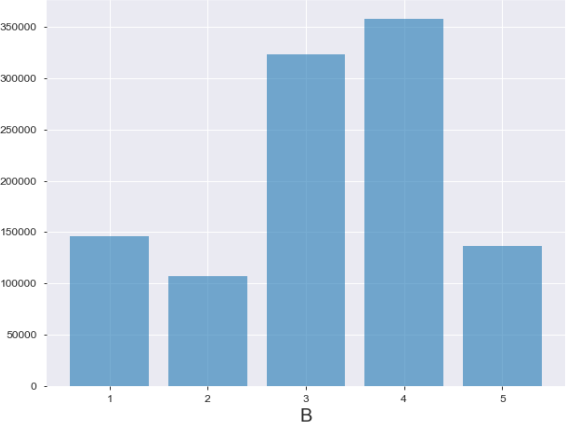
Now, we will explore in detail, the characteristics of each variable by plotting various graphs based on the category of the variable.

# 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



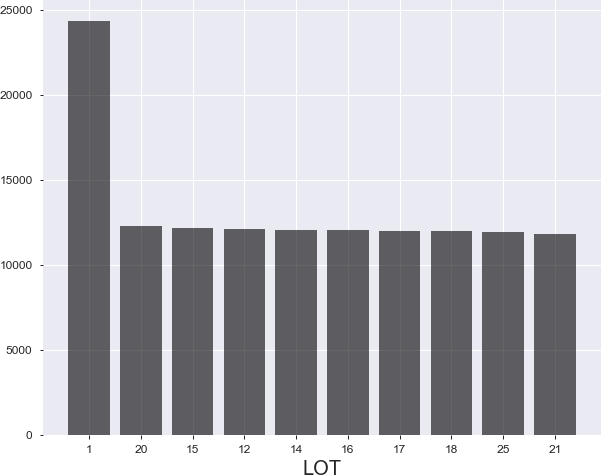
# BLOCK:

* + 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:

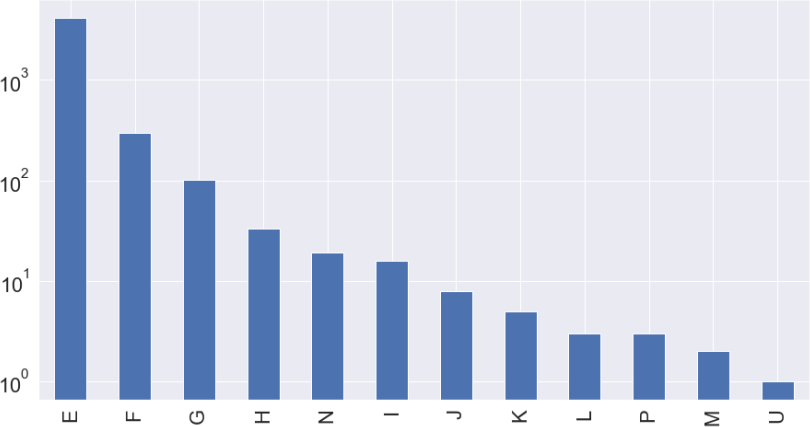


# LOT:

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

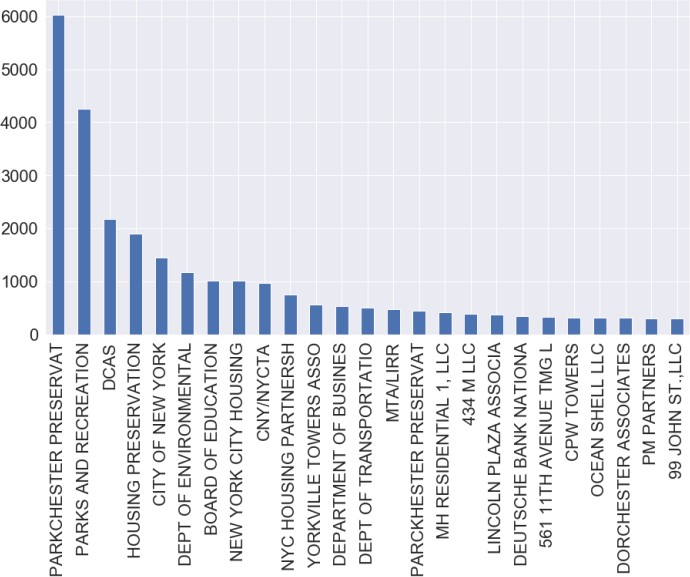


**EASEMENT:** It’s a categorical variable. Following bar graph shows the spread of NY properties across different classes of ‘EASEMENT’ variable



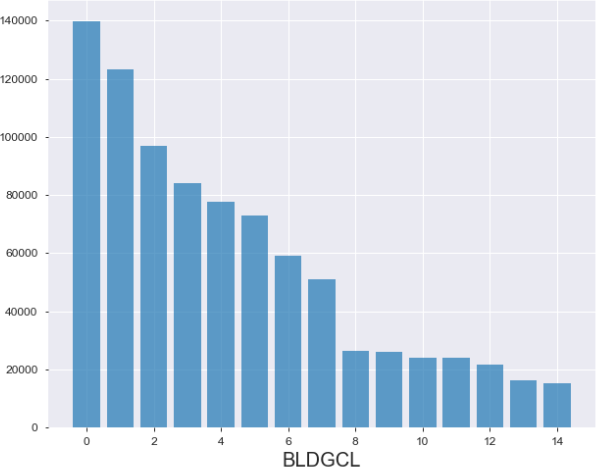
# OWNER:

* + This field has the name of the owner of the property.



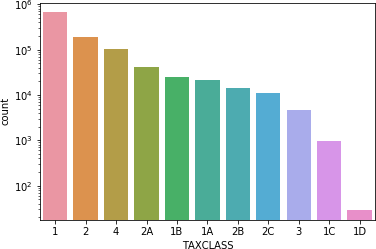
# BLDGCL:

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



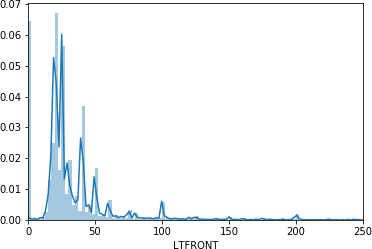
# TAXCLASS:

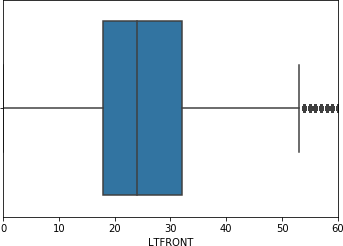
* + It corresponds to the Current Property Tax Class Code (NYS Classification). The following are valid values for this column:
    - 1 = 1-3 unit residences
    - 1a = 1-3 story condominiums originally a condo
    - 1b = residential vacant land
    - 1c = 1-3 unit condominiums originally tax class 1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

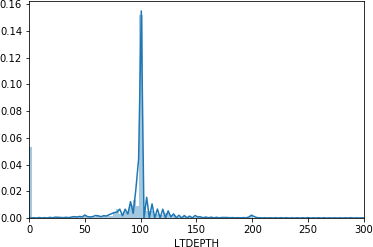
* + It stands for Lot Frontage in feet.
  + The Density plot and Boxplot below show the spread of values of ‘LTFRONT’ variable:

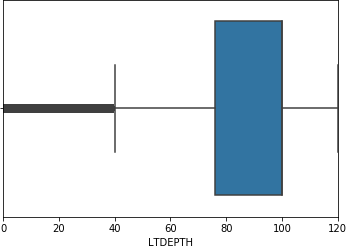




# LTDEPTH:

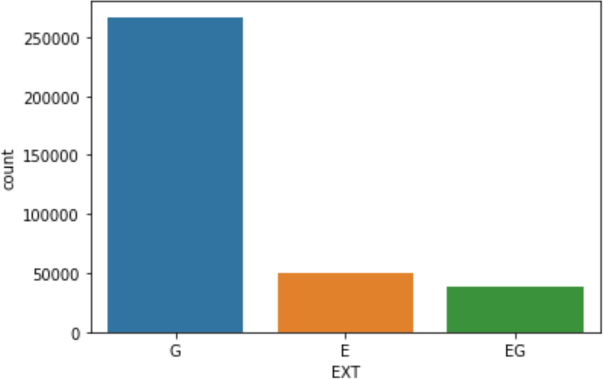
* + It stands for Lot Depth in feet.
  + The Density plot and Boxplot below show the spread of values of ‘LTDEPTH’ variable:





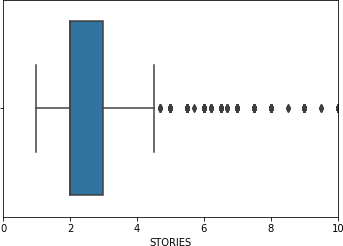
# EXT:

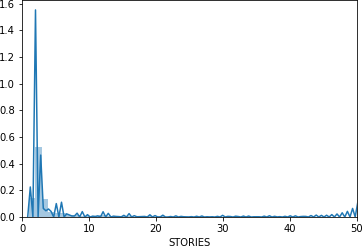
* + It’s a categorical variable, meaning extension.
  + Following bar graph shows the spread of properties across different classes of ‘EXT’ variable.



# STORIES:

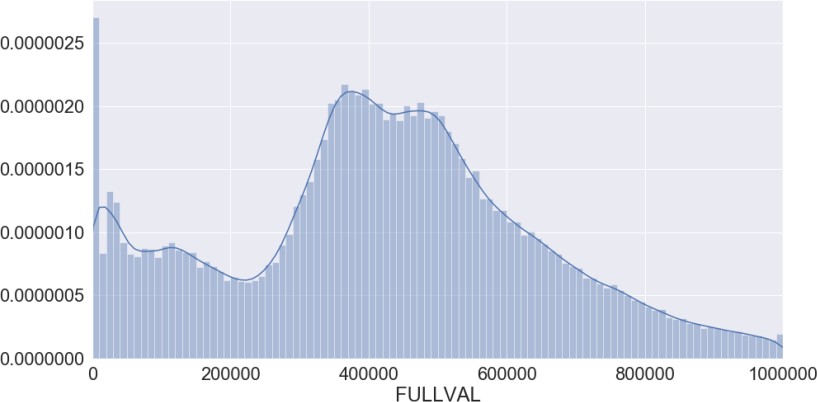
* + It’s a numerical variable showing the number of stories.
  + Following Density plot and Boxplot show the spread of values of ‘STORIES’ variable





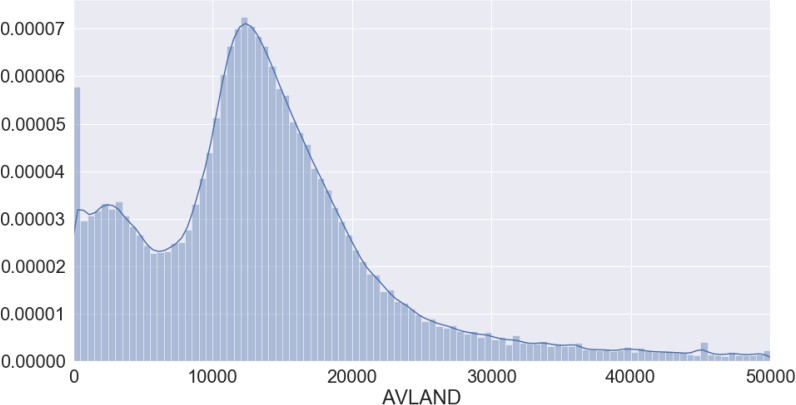
# FULLVAL:

* + It’s a numerical variable showing the total value of the property.
  + Following Density plot shows the distribution of values of ‘FULLVAL’ variable



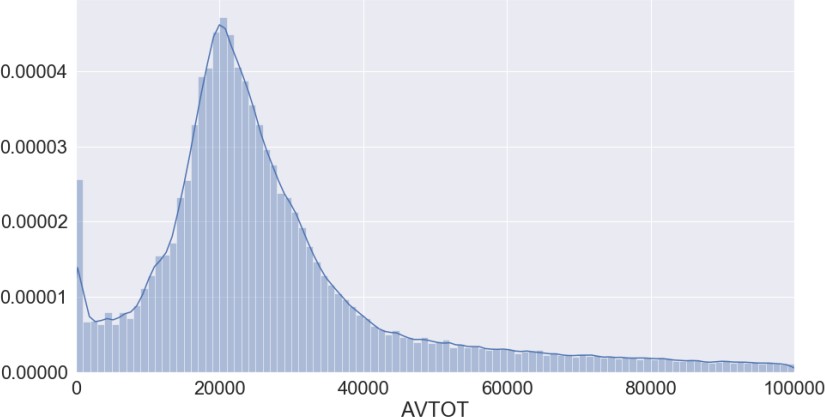
# AVLAND:

* + It’s a numerical variable.
  + Following Density plot shows the spread of values of ‘AVLAND’ variable



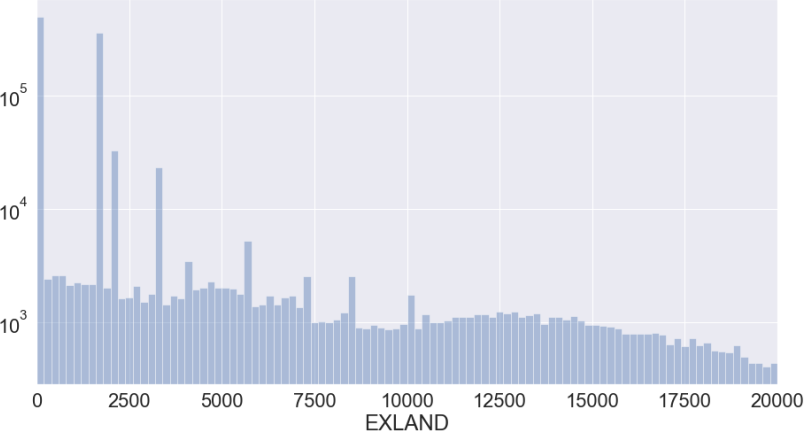
# AVTOT:

* + It’s a numerical variable.
  + Following density plot shows the spread of values of the ‘AVTOT’ variable



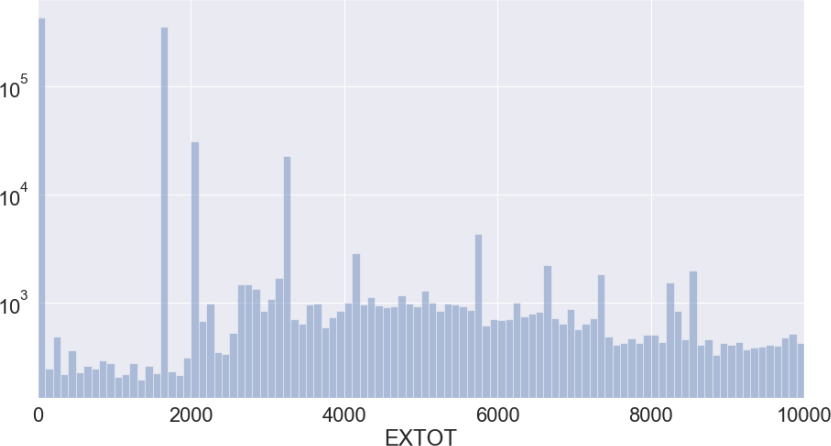
# EXLAND:

* + It’s a numerical variable. Following density plot shows the spread of values of the ‘EXLAND’ variable



# EXTOT:

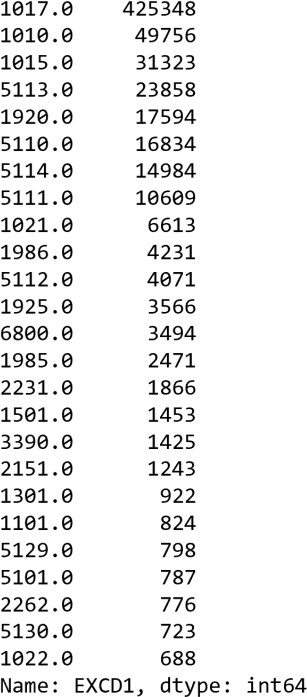
* + It’s a numerical variable.
  + Following histogram shows the spread of values of the ‘EXTOT’ variable



# EXCD1:

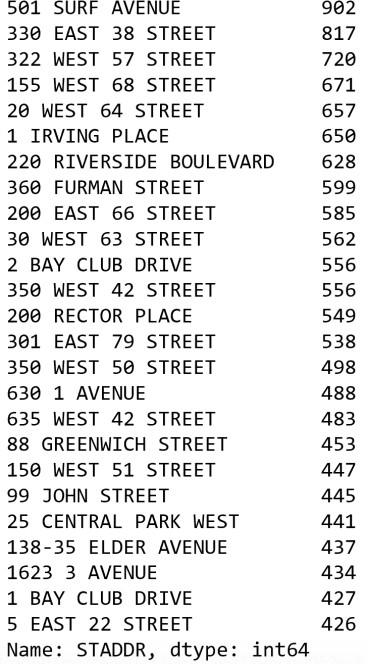
* + t’s a categorical variable. Following table shows the spread of NY properties across different values of ‘EXCD1’ variable

# Value Count



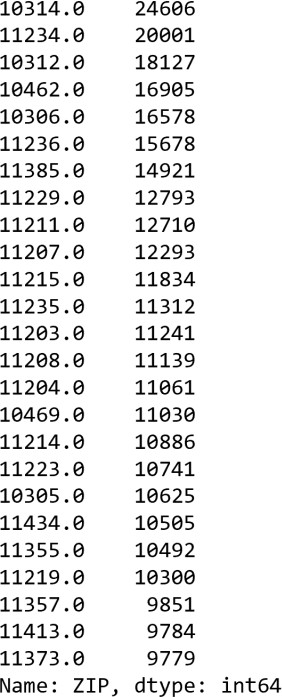
**STADDR**:

* + 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



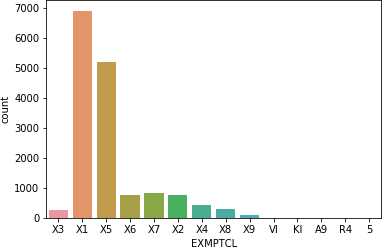
# ZIP:

* 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



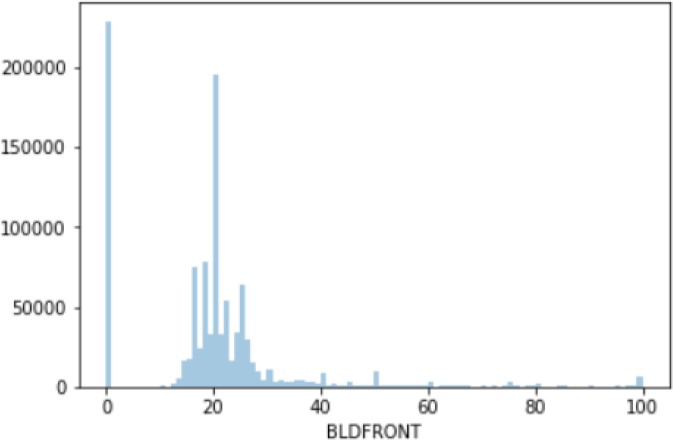
# EXMPTCL:

* + 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:

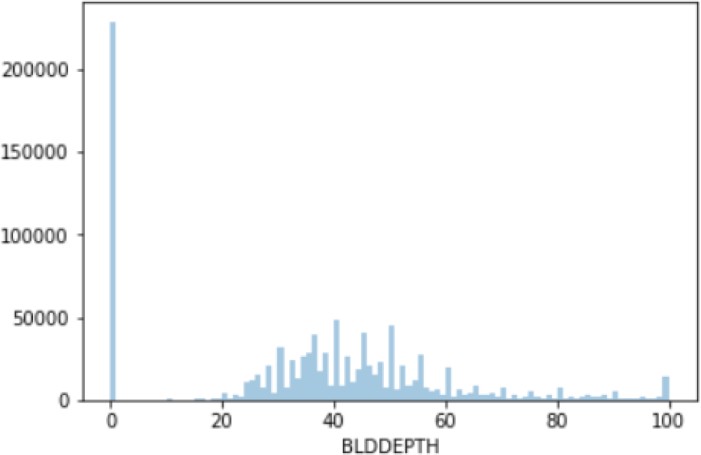
* + It’s a numerical variable.
  + Following histogram shows the spread of values of the ‘BLDFRONT’ variable



# BLDDEPTH:

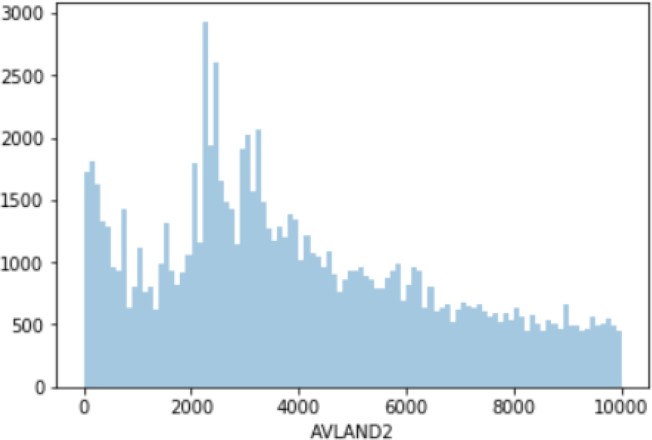
* + It’s a numerical variable, stands for depth of the building in feet.
  + Following histogram shows the spread of values of the ‘BLDDEPTH’ variable

-



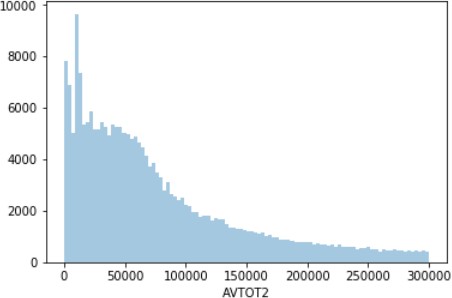
# AVLAND2:

* + It’s a numerical variable.
  + Following histogram shows the spread of values of the ‘AVLAND2’ variable



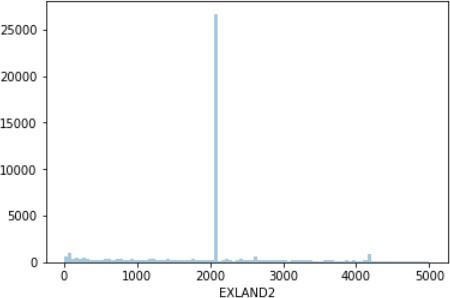
# AVTOT2:

* + It’s a numerical variable.
  + - Following histogram shows the spread of values of the ‘AVTOT2’ variable



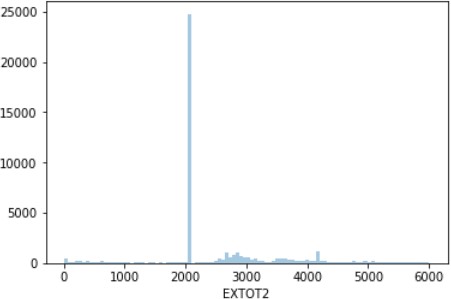
# EXLAND2:

* + It’s a numerical variable.
  + Following histogram shows the spread of values of the ‘EXLAND2’ variable



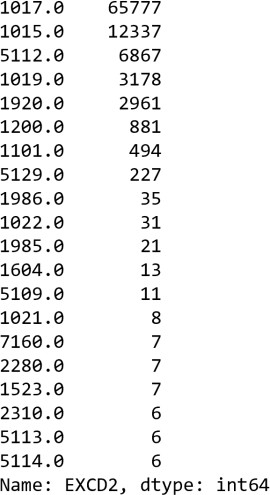
# EXTOT2:

* + 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



# 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’.