USC Marshall
School of Business

Data Quality Report

New York Property Valuation and Assessment Data

DSO 562 Fraud Analytics
Chutong Yan
January 23, 2020

Table of Contents

Section 1: High-level Data Description

The dataset represents NYC (New York City) properties assessments to calculate Property Tax, grant eligible properties exemptions and/or abatements. It contains 1,070,994 records and 32 fields (including field RECORD, a unique key to identify each record).

The dataset is online from NYC Open Data. It was collected and entered into the system by various City employee, like Property Assessors, Property Exemption specialists, ACRIS reporting, Department of Building reporting, etc.

The data covers the current values for Fiscal year 2010/2011 assessments, and the Tentative and Final values for the 2011/2012 Fiscal Year.

Section 2: Summary Tables of Fields

Table 1 Summary Table of Numerical Fields

Name	Type	# Not null	% Populated	# Unique	# Zero	Mean	St dev	Min	Max
LTFRONT	Numerical	1,070,994	100.00%	1,297	169,108	36.64	74.03	0	9,999
LTDEPTH	Numerical	1,070,994	100.00%	1,370	170,128	88.86	76.40	0	9,999
STORIES	Numerical	1,014,730	94.75%	112	0	5.01	8.37	1	119
FULLVAL	Numerical	1,070,994	100.00%	109,324	13,007	874,264.51	11,582,430.99	0	6,150
AVLAND	Numerical	1,070,994	100.00%	70,921	13,009	85,067.92	4,057,260.06	0	266,850
AVTOT	Numerical	1,070,994	100.00%	112,914	13,007	227,238.17	6,877,529.31	0	4,668,308,947
EXLAND	Numerical	1,070,994	100.00%	33,419	491,699	36,423.89	3,981,575.79	0	2,668,500,000
EXTOT	Numerical	1,070,994	100.00%	64,255	432,572	91,186.98	6,508,402.82	0	4,668,308,947
BLDFRONT	Numerical	1,070,994	100.00%	612	228,815	23.04	35.58	0	7,575
BLDDEPTH	Numerical	1,070,994	100.00%	621	228,853	39.92	42.71	0	9,393
AVLAND2	Numerical	282,726	26.40%	58,592	0	246,235.72	6,178,962.56	3	2,371,005,000
AVTOT2	Numerical	282,732	26.40%	111,361	0	713,911.44	11,652,528.95	3	4,501,180,002
EXLAND2	Numerical	87,449	8.17%	22,196	0	351,235.70	10,802,210.00	1	2,371,005,000
EXTOT2	Numerical	130,828	12.22%	48,349	0	656,768.30	16,072,510.00	7	4,501,180,000

* Field RECORD is not specified in the table above because summary statistics of unique record key is not meaningful.

Table 2 Summary Table of Categorical Fields

Name	Type	# Not null	% Populated	# Unique	# zero	Most Common Field Value
BBLE	Categorical	1,070,994	100.00%	1,070,994	0	N/A
B	Categorical	1,070,994	100.00%	5	0	4
BLOCK	Categorical	1,070,994	100.00%	13,984	0	3944
LOT	Categorical	1,070,994	100.00%	6,366	0	1
EASEMENT	Categorical	4,636	0.43%	13	0	E
OWNER	Categorical	1,039,249	97.04%	863,347	0	PARKCHESTER PRESERVAT
BLDGCL	Categorical	1,070,994	100.00%	200	0	R4
TAXCLASS	Categorical	1,070,994	100.00%	11	0	1
EXT	Categorical	354,305	33.08%	4	0	G
EXCD1	Categorical	638,488	59.62%	130	0	1017
STADDR	Categorical	1,070,318	99.94%	839,281	0	501 SURF AVENUE
ZIP	Categorical	1,041,104	97.21%	197	0	10314
EXMPTCL	Categorical	15,579	1.45%	15	0	X1
EXCD2	Categorical	92,948	8.68%	61	0	1017
PERIOD	Categorical	1,070,994	100.00%	1	0	FINAL
YEAR	Date/time	1,070,994	100.00%	1	0	2010/11
VALTYPE	Categorical	1,070,994	100.00%	1	0	AC-TR

Section 3: Explorations of Fields

3.1 RECORD

RECORD field is the unique integer label for each record, ranging from 1 to 1,070,994.

3.2 BBLE

BBLE field is the concatenation of values in field B, BLOCK, LOT and EASEMENT. All BBLE in the dataset are unique, so mode is not meaningful. The table below listed sample values of the BBLE field (the first 5 records).

Table 3 Sample Records in BBLE Field

Records	BBLE
1	1000010201
2	1000020001
3	1000020023
4	1000030001
5	1000030002

3.3 B

B field represents the borough codes, where 1 = MANHATTAN, 2 = BRONX, 3 = BROOKLYN,

4 = QUEENS, 5 = STATEN ISLAND. The table below shows the frequency of values in B field.

Table 4 Frequency Table of Values in B Field

Value - B	Frequency
4	358,046
3	323,243
1	146,220
5	136,200
2	107,285

3.4 BLOCK

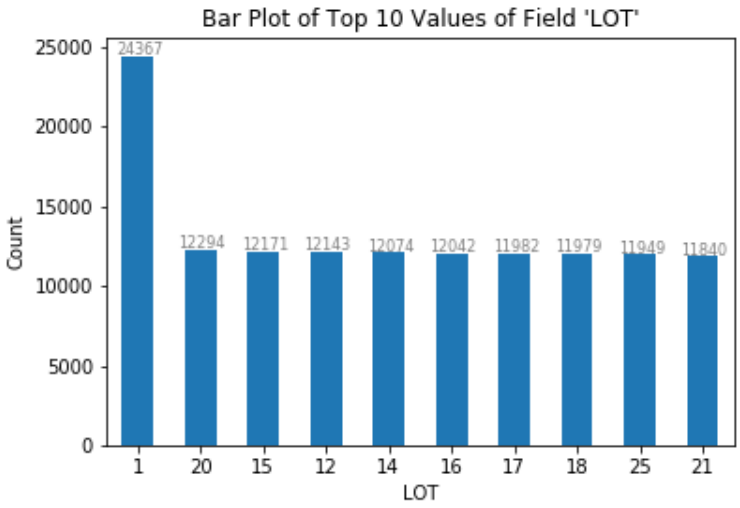
BLOCK field represents the valid block ranges by boroughs, where MANHATTAN = 1 TO 2,255, BRONX = 2,260 TO 5,958, BROOKLYN = 1 TO 8,955, QUEENS = 1 TO 16,350 and STATEN ISLAND = 1 TO 8,050. The table below shows the top 10 most common values of BLOCK field.

Table 5 Frequency Table of Values in BLOCK Field

Value - BLOCK	Frequency
3944	3,888
16	3,786
3943	3,424
3938	2,794
1171	2,535
3937	2,275
1833	1,774
2450	1,651
1047	1,480
7279	1,302

3.5 LOT

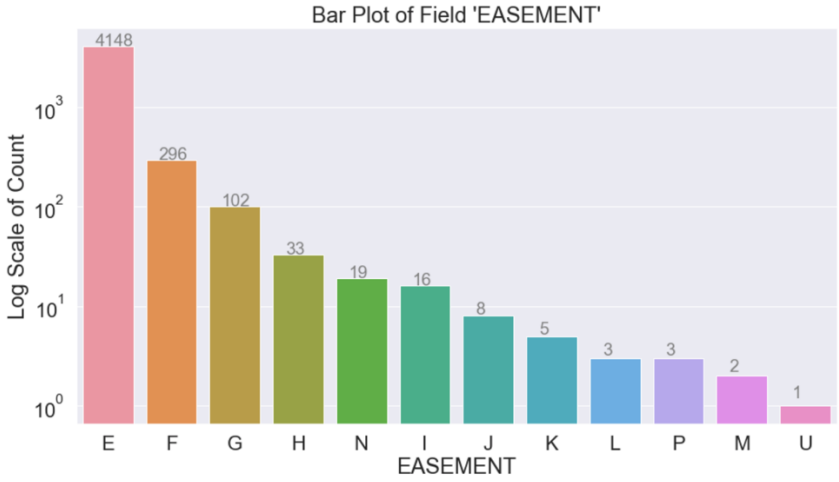
LOT field represents the unique number of lots within borough/block. The plot below shows the top 10 most common values of LOT field.



(figure 1)

3.6 EASEMENT

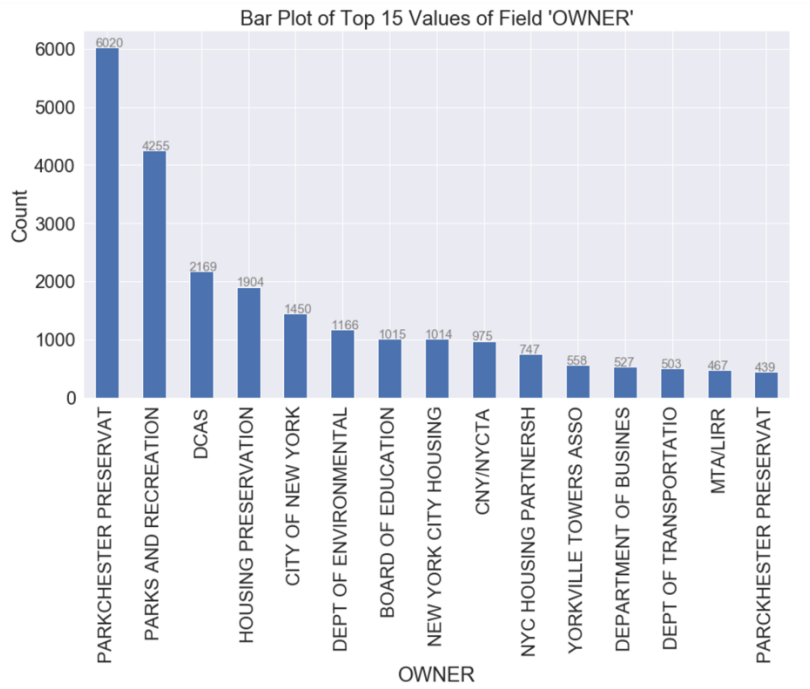
EASEMENT field specifies the easement details of the lot, where SPACE indicates the lot has no Easement, A indicates the portion of the Lot that has an Air Easement, B indicates Non-Air Rights, E THRU M indicates the portion of the lot that has a Land Easement, N indicates Non-Transit Easement, P indicates Piers, R indicates Railroads, S indicates Street and U indicates U.S. Government. The Bar plot below shows the (log scale of) count of values in EASEMENT field.



(figure 2)

3.7 OWNER

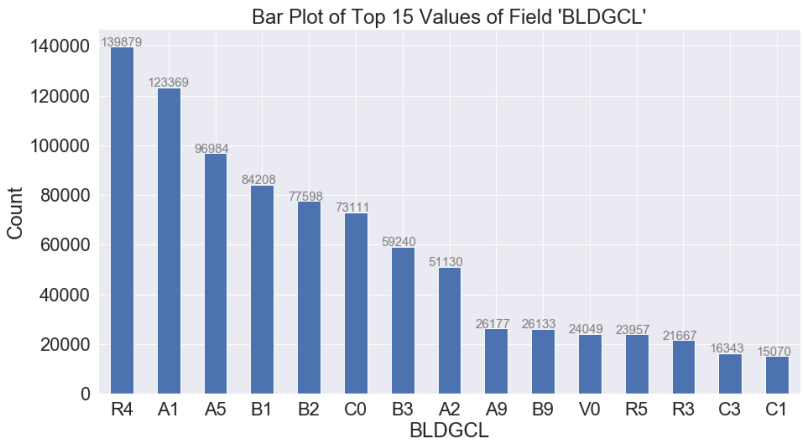
OWNER field represents the owner's names of the properties. The Bar plot below shows the top 15 most common values of OWNER field.



(figure 3)

3.8 BLDGCL

BLDGCL field represents the building class of the properties. There is a direct correlation between the Building Class and the Tax Class. The Bar plot below shows the Top 15 Values of BLDGCL field.



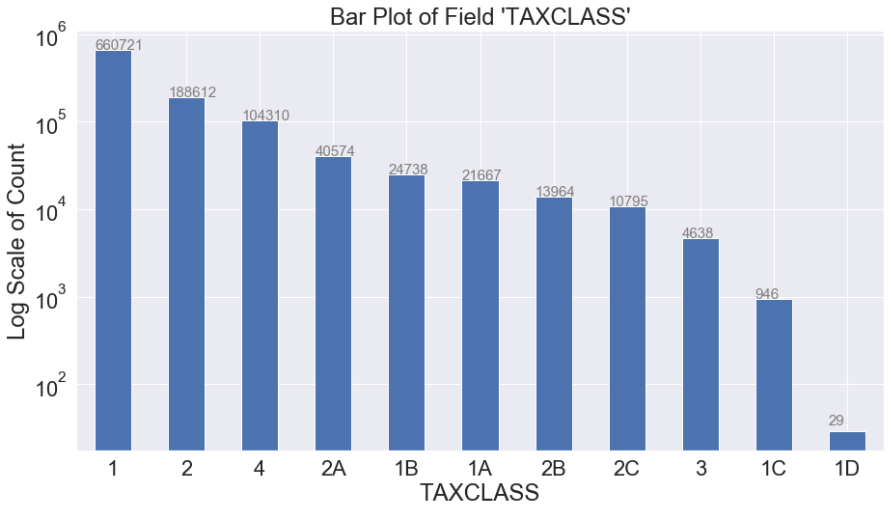
(figure 4)

3.9 TAXCLASS

TAXCLASS field represents the tax class of the properties. Explanations of tax classes are listed below:

TAX CLASS 1 = 1-3 UNIT RESIDENCES,

TAX CLASS 1B = RESIDENTIAL VACANT LAND,
TAX CLASS 1C = 1-3 UNIT CONDOMINUMS,
TAX CLASS 1D = SELECT BUNGALOW COLONIES,
TAX CLASS 2 = APARTMENTS,
TAX CLASS 2A = APARTMENTS WITH 4-6 UNITS,
TAX CLASS 2B = APARTMENTS WITH 7-10 UNITS,
TAX CLASS 2C = COOPS/CONDOS WITH 2-10 UNITS,
TAX CLASS 3 = UTILITIES (EXCEPT CEILING RR),
TAX CLASS 4A = UTILITIES - CEILING RAILROADS,
TAX CLASS 4 = ALL OTHERS.
The Bar plot below shows the frequency of values in the TAXCLASS field.

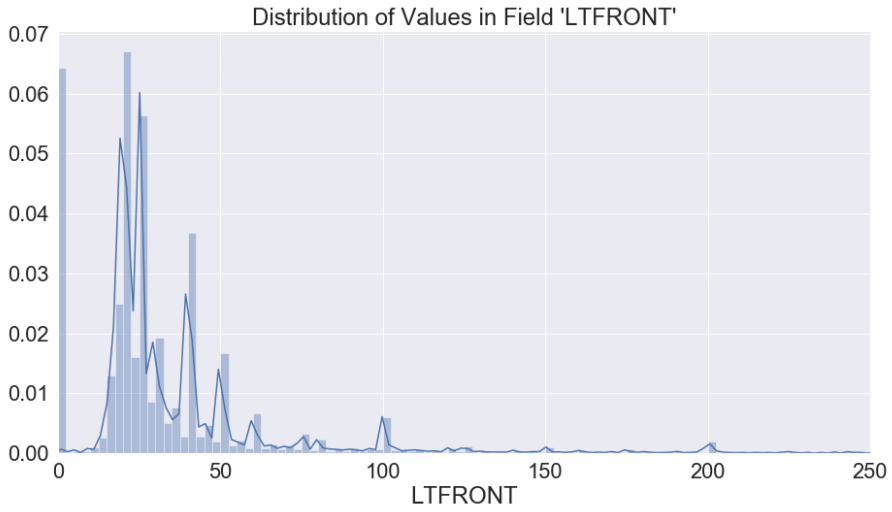


(figure 5)

3.10 LTFRONT

LTFRONT field represents the lot frontage in feet. The plot below shows the distribution of values in LTFRONT field (omitted values greater than 250, that is 11,854 records and 1.11% of total valid records in this field).

It is also worth noticing that there are 170,128 records equal to zero, which make up 15.89% of total valid records in the LTFRONT field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.

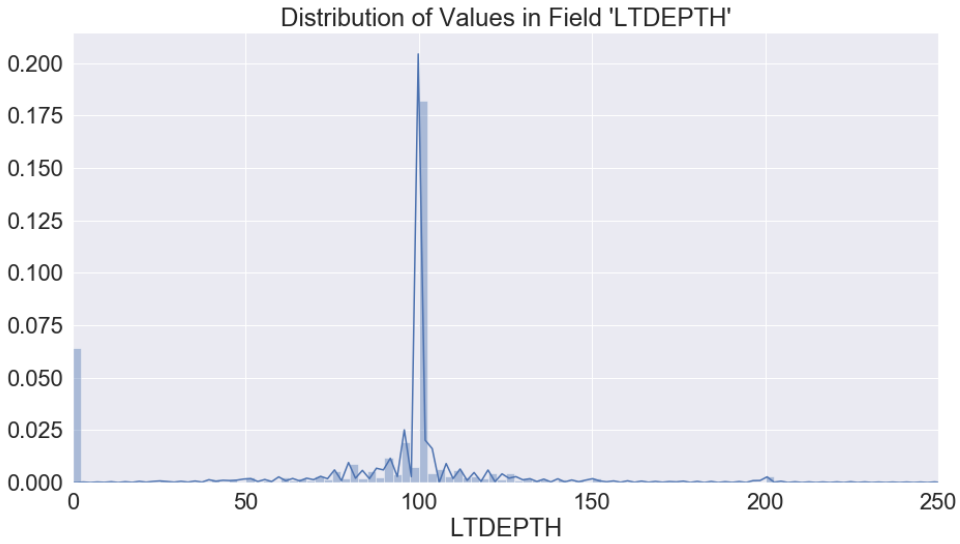


(figure 6)

3.11 LTDEPTH

LTDEPTH field represents the lot depth in feet. The plot below shows the distribution of values in LTDEPTH field (omitted values greater than 250, that is 10,033 records and 0.94% of total valid records in this field).

It is also worth noticing that there are 170,128 records equal to zero, which make up 15.79% of total valid records in the LTDEPTH field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.



(figure 7)

3.12 EXT

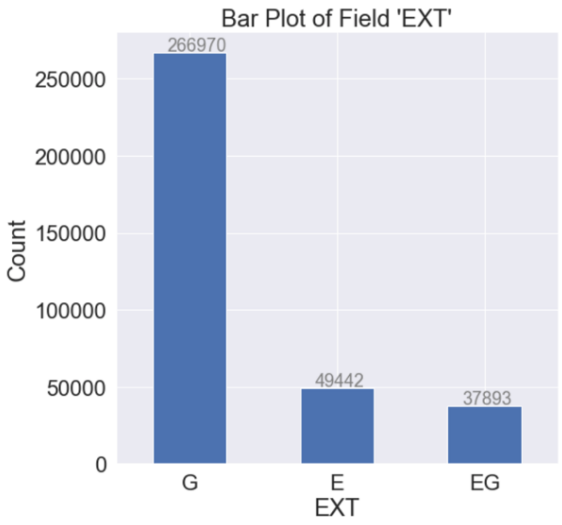
EXT field represents the extension of the properties. Below are the explanations of the abbreviations:

'E' = EXTENSION

'G' = GARAGE

'EG' = EXTENSION AND GARAGE

The Bar plot below shows the frequency of values in the EXT field.



(figure 8)

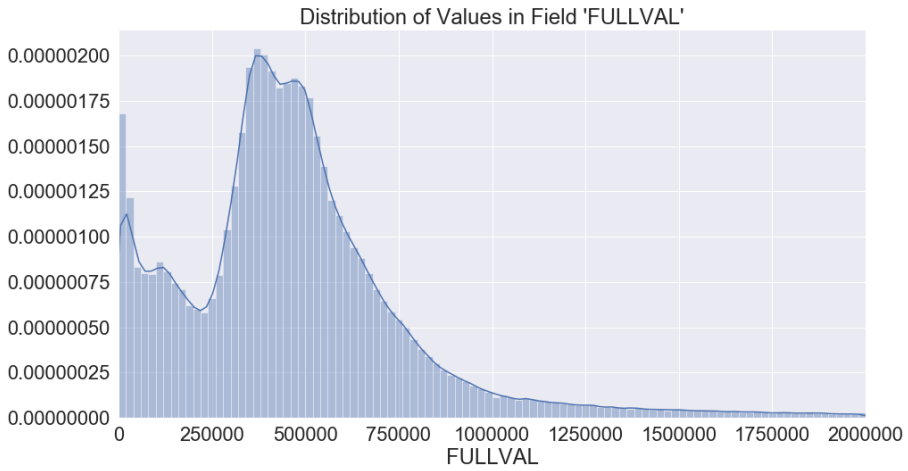
3.13 STORIES

STORIES field represents the number of stories for the building (# of Floors). The plot below shows the distribution of values in STORIES field (omitted values greater than 15, that is 898,673 records and 7.26% of total valid records in this field).

(figure 9)

3.14 FULLVAL

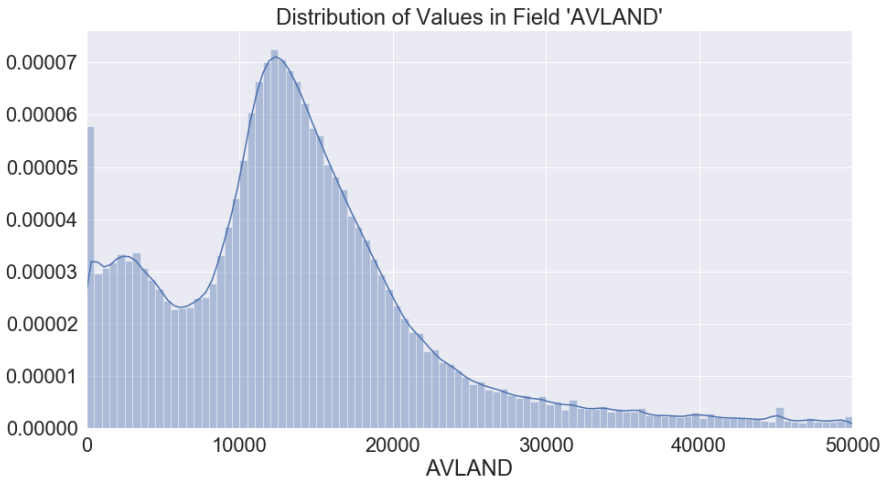
FULLVAL field represents the total market value of the property (if the value is not zero). The plot below shows the distribution of values in FULLVAL field (omitted values greater than 2,000,000, that is 39,496 records and 3.69% of total valid records in this field).



(figure 10)

3.15 AVLAND

AVLAND field represents the actual land value. The plot below shows the distribution of values in AVLAND field (omitted values greater than 50,000, that is 101,453 records and 9.47% of total valid records in this field).

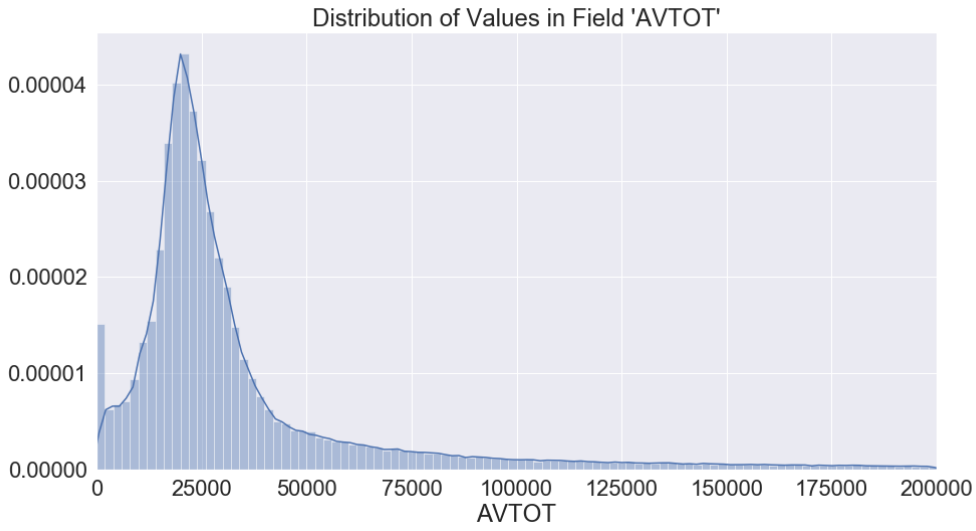


(figure 11)

3.16 AVTOT

AVTOT field represents the actual total value of the building. The plot below shows the distribution of values in AVTOT field (omitted values greater than 200,000 that is 90,264

records and 8.42 % of total valid records in this field).

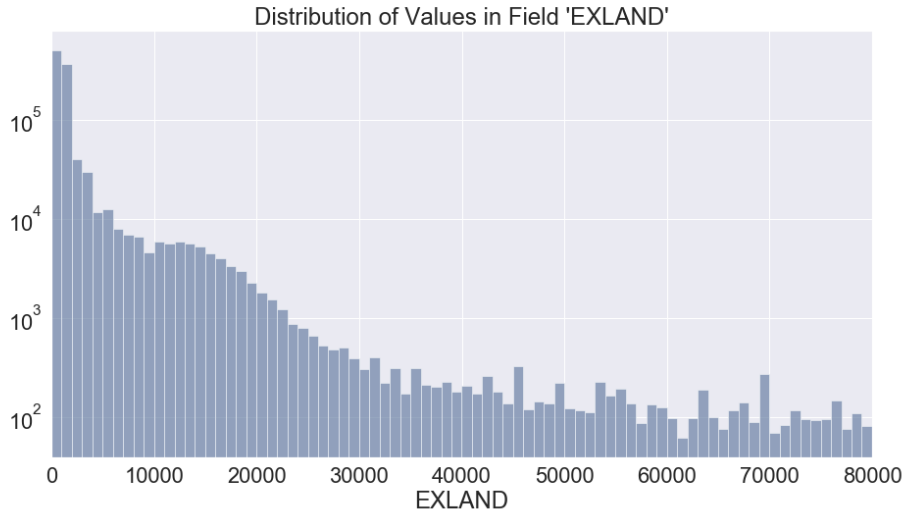


(figure 12)

3.17 EXLAND

EXLAND field represents the actual exempt land value of the building. The plot below shows the distribution of values in EXLAND field (omitted values greater than 80,000, that is 17,257 records and 1.61% of total valid records in this field).

It is also worth noticing that there are 491,699 records equal to zero, which make up 45.91% of total valid records in the EXLAND field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.

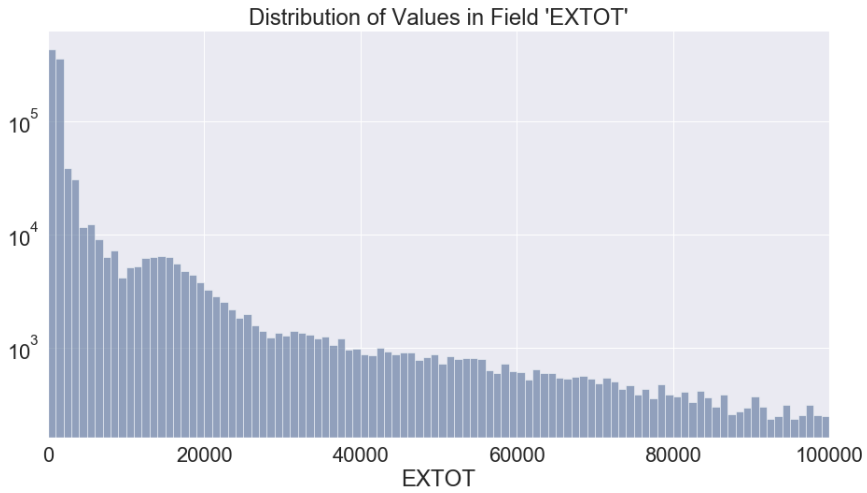


(figure 13)

3.18 EXTOT

EXTOT field represents the actual exempt total value of the building. The plot below shows the distribution of values in EXTOT field (omitted values greater than 100,000, that is 44,209 records and 4.12% of total valid records in this field).

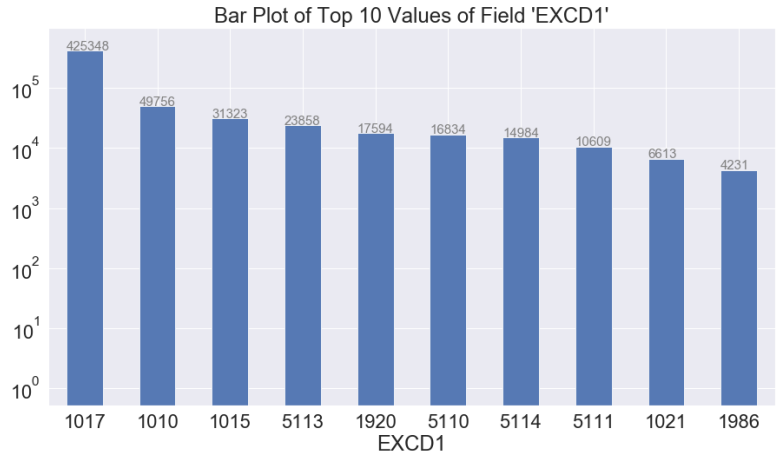
It is also worth noticing that there are 432,572 records equal to zero, which make up 40.39% of total valid records in the EXTOT field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.



(figure 14)

3.19 EXCD1

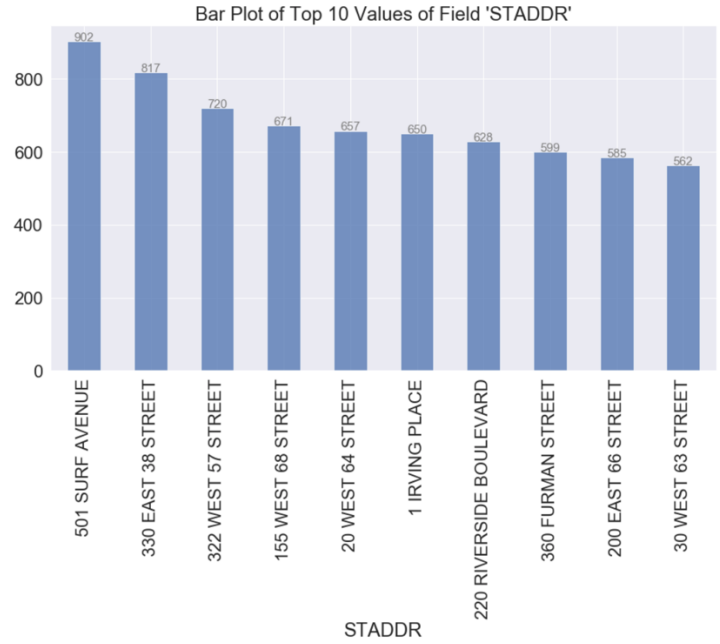
EXCD1 field represents the exemption code 1 of the building. The plot below shows the top 10 most common values of EXCD1 field.



(figure 15)

3.20 STADDR

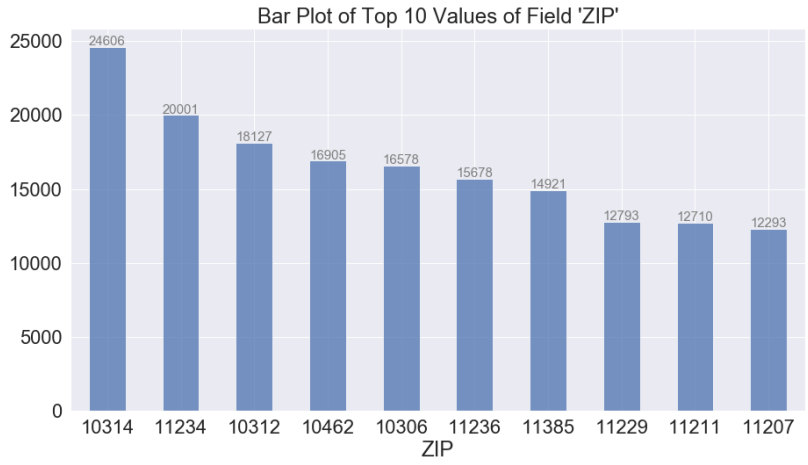
STADDR field represents the street name of the property. The plot below shows the top 10 most common values of STADDR field.



(figure 16)

3.21 ZIP

ZIP field represents the postal zip code of the property. The plot below shows the top 10 most common values of ZIP field.

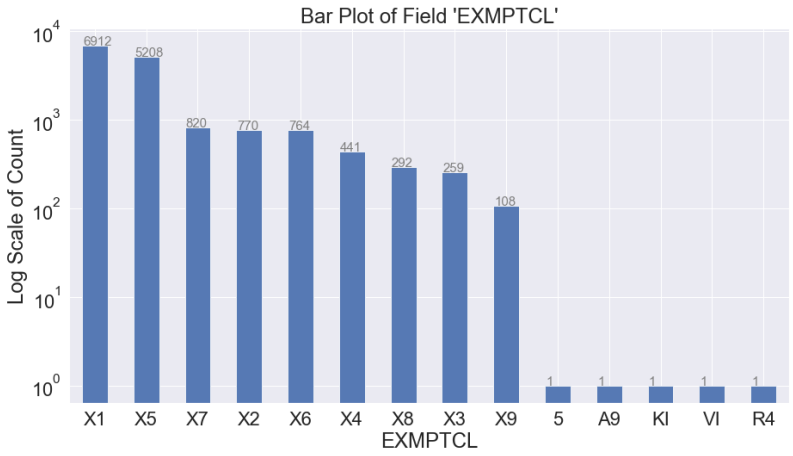


(figure 17)

3.22 EXMPTCL

EXMPTCL field represents the exemption class for the properties. The plot below shows the

distribution of the values in EXMPTCL field.

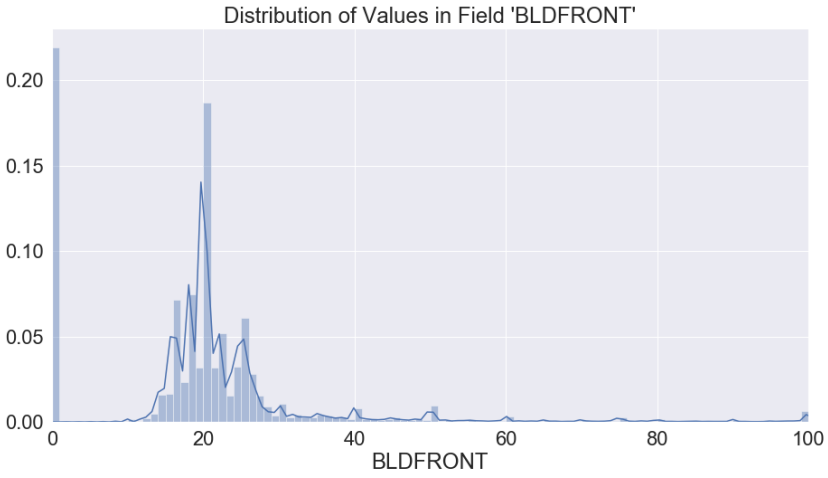


(figure 18)

3.23 BLDFRONT

BLDFRONT field represents building frontage in feet. The plot below shows the distribution of values in BLDFRONT field (omitted values greater than 100, that is 28,142 records and 2.63% of total valid records in this field).

It is also worth noticing that there are 229,815 records equal to zero, which make up 21.36% of total valid records in the BLDFRONT field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.



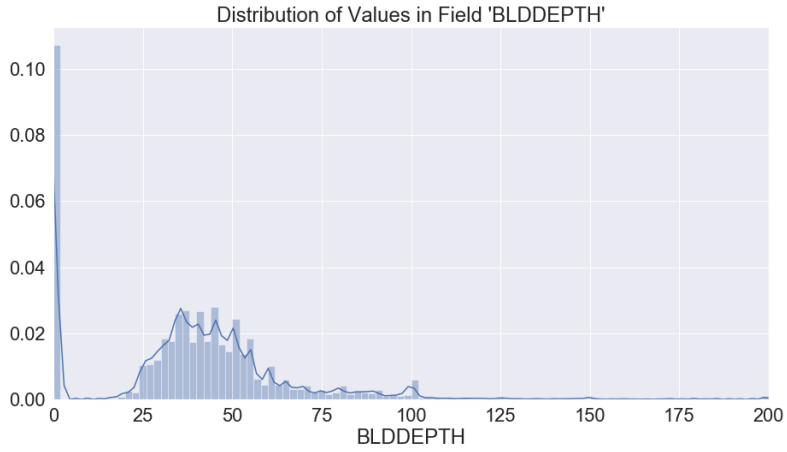
(figure 19)

3.24 BLDDEPTH

BLDDEPTH field represents building depth in feet. The plot below shows the distribution of values in BLDDEPTH field (omitted values greater than 200, that is 4,730 records and 0.44%

of total valid records in this field).

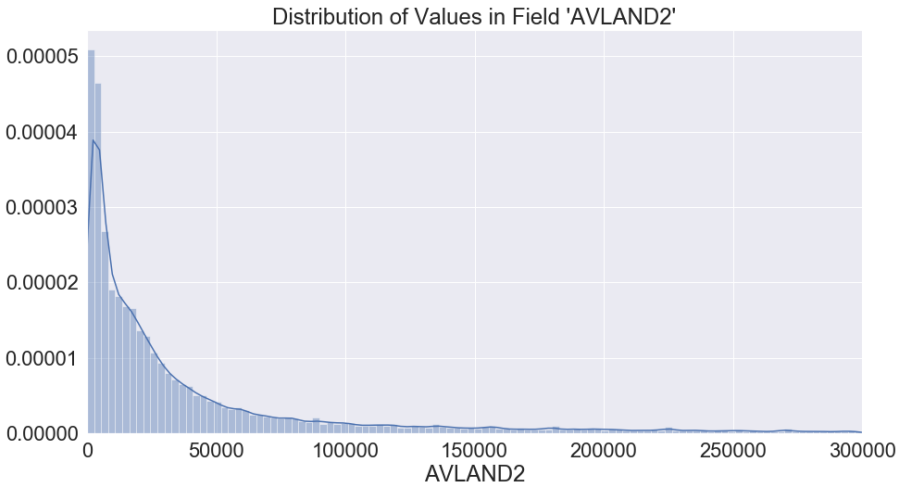
It is also worth noticing that there are 228,853 records equal to zero, which make up 21.37% of total valid records in the BLDDEPTH field. The real-world implications of the zero values and the logic behind the large portion of zeros in this particular field should be further specified.



(figure 20)

3.25 AVLAND2

AVLAND2 field represents transitional land value. The plot below shows the distribution of values in AVLAND2 field (omitted values greater than 300,000, that is 24,077 records and 8.51% of total valid records in this field).

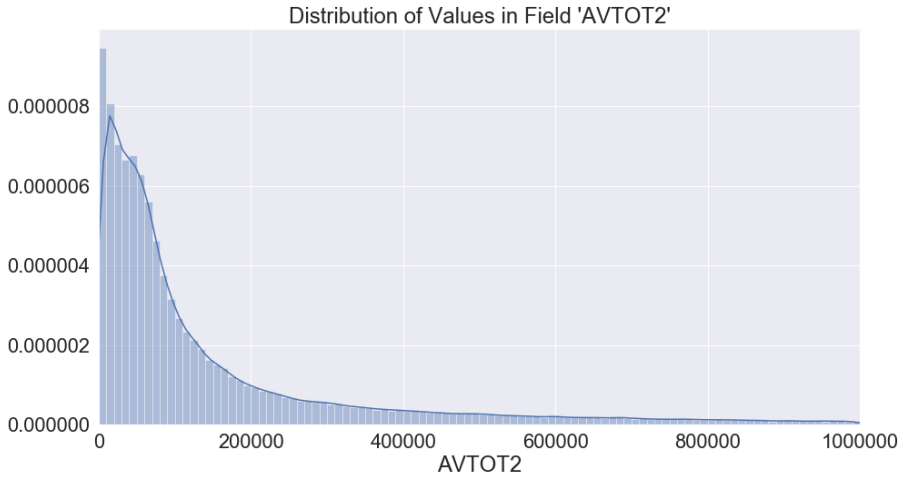


(figure 21)

3.26 AVTOT2

AVTOT2 field represents transitional total value. The plot below shows the distribution of values in AVTOT2 field (omitted values greater than 1,000,000, that is 23,693 records and

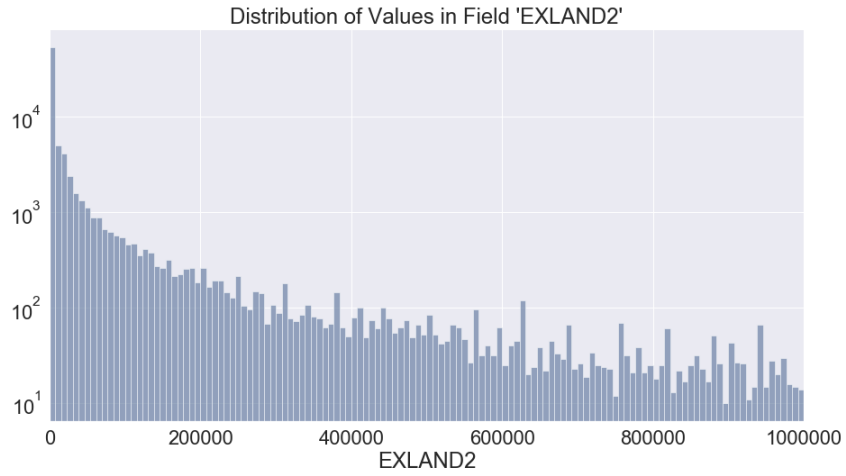
8.38% of total valid records in this field).



(figure 22)

3.27 EXLAND2

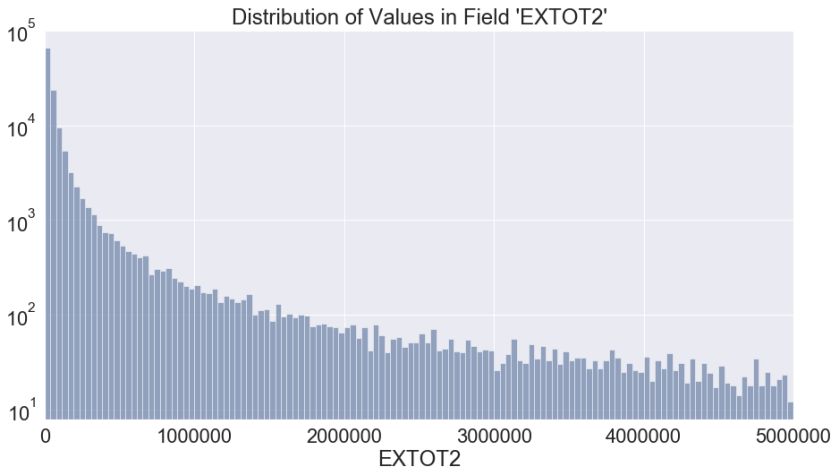
EXLAND2 field represents transitional exempt land value. The plot below shows the distribution of values in EXLAND2 field (omitted values greater than 1,000,000, that is 3,637 records and 4.19% of total valid records in this field).



(figure 23)

3.28 EXTOT2

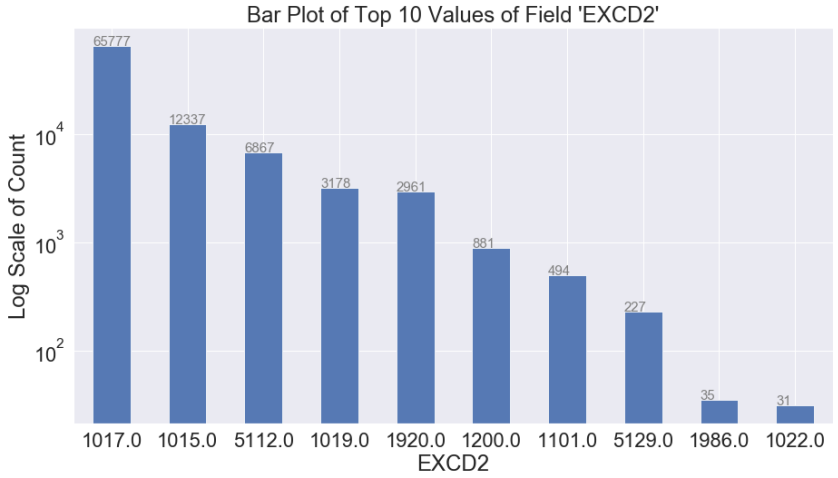
EXTOT2 field represents transitional exempt total value. The plot below shows the distribution of values in EXTOT2 field (omitted values greater than 2,000,000, that is 5,667 records and 4.33% of total valid records in this field).



(figure 24)

3.29 EXCD2

EXCD2 field represents the exemption code 2 of the building. The plot below shows the top 10 most common values of EXCD2 field.



(figure 25)

3.30 PERIOD

PERIOD field represents the assessment period. The table below shows frequencies of values in Period field.

Table 6 Frequency Table of Values in PERIOD Field

Value - PERIOD	Frequency
Final	1,070,994

3.31 YEAR

YEAR field represents the assessment year. The table below shows frequencies of values in Period field.

Table 7 Frequency Table of Values in YEAR Field

Value - YEAR	Frequency
2010/11	1,070,994

3.32 VALTYPE

VALTYPE field represents the value type. The table below shows frequencies of values in VALTYPE field.

Table 8 Frequency Table of Values in VALTYPE Field

Value - VALTYPE	Frequency
AC-TR	1,070,994

Section 4: Potential Problems for Further Clarification

4.1 Fields with Missing Data

Fields including AVLAND2, AVTOT2, EXLAND2, EXTOT2, EASEMENT, EXMPTCL, EXCD2 have more than 70% of null records. Further investigation should be conducted to understand what causes the high null-value rate and if the reasons given are legitimate in this particular business case.

4.2 Fields with High Proportion of Zeros

Fields including LTFRONT, LTDEPTH, EXLAND, EXTOT, BLDFRONT, BLDDEPTH have up to 45.9% of record values as zero. Further investigation should be conducted to understand what is the real-world implications of zero in these particular fields (For example, are they missing values? Does zero make sense for these fields?) and what causes the high proportion of zero values.

Reference

[1] NYC OpenData. (2008). *Property valuation and assessment data* [Data file]. Retrieved from https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8