To: Professor Stephen Coggeshall
From: Alok Abhishek
Date: 01/30/2018
Subject: DSO 562: Fraud Analytics

Data Quality Report: New York Real Estate data set…

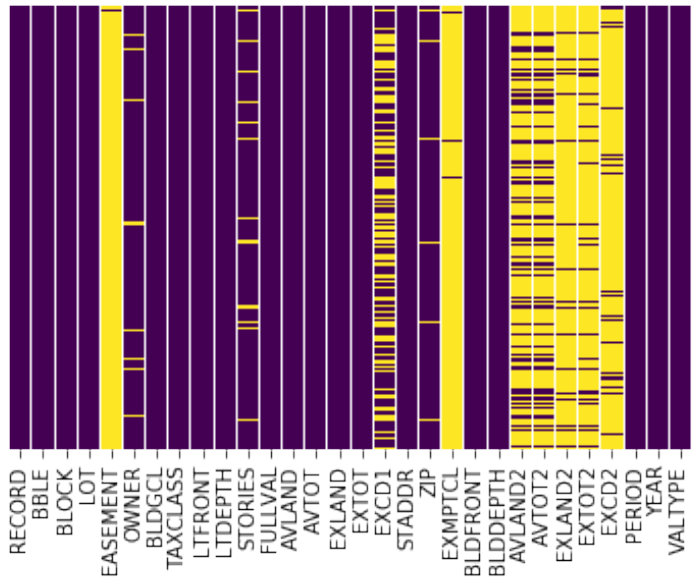
Data Description: The New York City department of Finance values properties in NYC every year to calculate the property tax. This report provides property tax data such as market and assessed values, exemptions, and abatements from the assessment year 2010/11. The information is listed by categories, such as borough, tax class, and type of building.

There are 1048575 records and 29 columns with total data size of ~159 MB.

Data Summary:

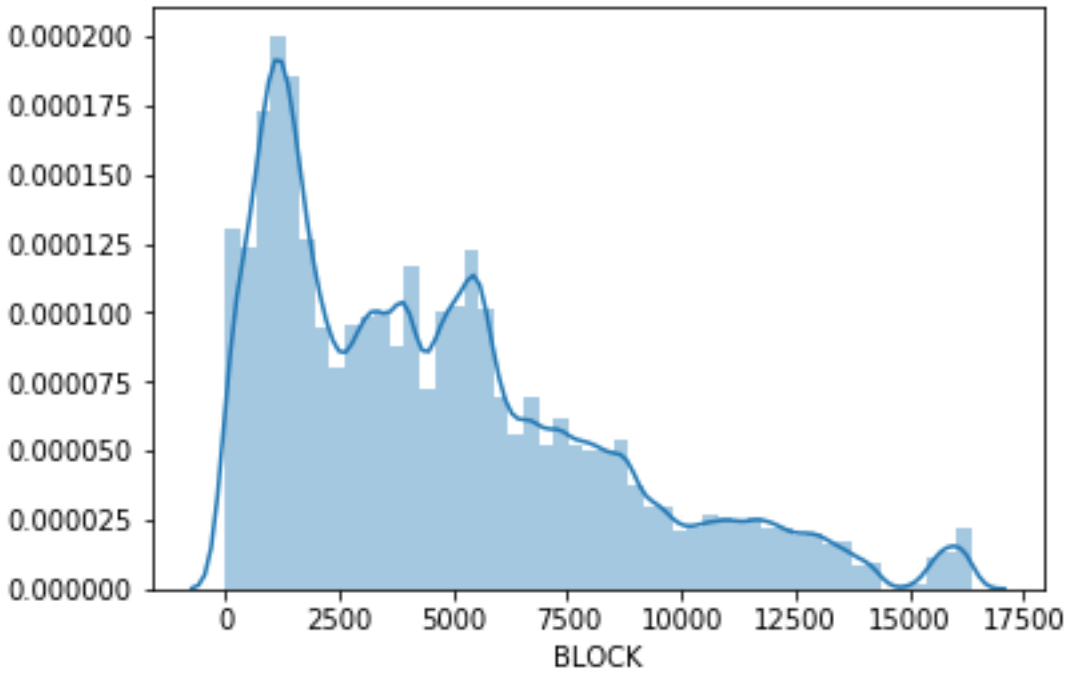
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
RECORD	int64	1,048,575								100.00%	
BLOCK	int64	1,048,575	4,708.87	3,699.55	1.00	1,534.00	3,944.00	6,797.00	16,350	100.00%	13,949
LOT	int64	1,048,575	370.09	860.54	1.00	23.00	49.00	146.00	9,978	100.00%	6,366
LTFRONT	int64	1,048,575	36.17	73.73	0.00	19.00	25.00	40.00	9,999	100.00%	1,277
LTDEPTH	int64	1,048,575	88.28	75.48	0.00	80.00	100.00	100.00	9,999	100.00%	1,336
STORIES	float64	996,433	5.06	8.43	1.00	2.00	2.00	3.00	119	95.03%	111
FULLVAL	int64	1,048,575	880,487.66	11,702,930.00	0.00	303,000.00	446,000.00	619,000.00	6,150,000,000	100.00%	108,277
AVLAND	int64	1,048,575	85,995.03	4,100,755.00	0.00	9,160.00	13,646.00	19,706.00	2,668,500,000	100.00%	70,529
AVTOT	int64	1,048,575	230,758.18	6,951,206.00	0.00	18,385.00	25,339.00	46,095.00	4,668,309,000	100.00%	112,294
EXLAND	int64	1,048,575	36,811.79	4,024,330.00	0.00	0.00	1,620.00	1,620.00	2,668,500,000	100.00%	33,186
EXTOT	int64	1,048,575	92,543.81	6,578,281.00	0.00	0.00	1,620.00	2,090.00	4,668,309,000	100.00%	63,805
EXCD1	float64	622,642	1,604.50	1,388.13	1,010.00	1,017.00	1,017.00	1,017.00	7,170	59.38%	129
ZIP	float64	1,022,219	10,935.32	526.58	10,001.00	10,453.00	11,215.00	11,364.00	33,803	97.49%	196
BLDFRONT	int64	1,048,575	23.02	35.79	0.00	15.00	20.00	24.00	7,575	100.00%	610
BLDDEPTH	int64	1,048,575	40.07	43.04	0.00	26.00	39.00	51.00	9,393	100.00%	620
AVLAND2	float64	280,966	246,365.48	6,199,390.00	3.00	5,705.00	20,059.00	62,338.75	2,371,005,000	26.80%	58,169
AVTOT2	float64	280,972	716,078.71	11,690,170.00	3.00	34,013.50	80,010.00	240,792.00	4,501,180,000	26.80%	110,890
EXLAND2	float64	86,675	351,802.21	10,852,480.00	1.00	2,090.00	3,053.00	31,419.00	2,371,005,000	8.27%	21,996
EXTOT2	float64	129,933	658,114.78	16,129,810.00	7.00	2,889.00	37,116.00	106,629.00	4,501,180,000	12.39%	48,106
EXCD2	float64	90,941	1,371.66	1,105.49	1,011.00	1,017.00	1,017.00	1,017.00	7,160	8.67%	60
EASEMENT	object	4,043								0.39%	12
OWNER	object	1,017,492								97.04%	847,053
BLDGCL	object	1,048,575								100.00%	200
TAXCLASS	object	1,048,575								100.00%	11
STADDR	object	1,047,934								99.94%	820,637
EXMPTCL	object	14,992								1.43%	14
PERIOD	object	1,048,575								100.00%	1
YEAR	object	1,048,575								100.00%	1
VALTYPE	object	1,048,575								100.00%	1

Heat map of missing values in the data set…



BLOCK – Block #

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLOCK	int64	1,048,575	4,708.87	3,699.55	1.00	1,534.00	3,944.00	6,797.00	16,350	100.00%	13,949



Bock # to area mapping:

Manhattan - 1 to 2,255

Bronx - 2,260 to 5,958

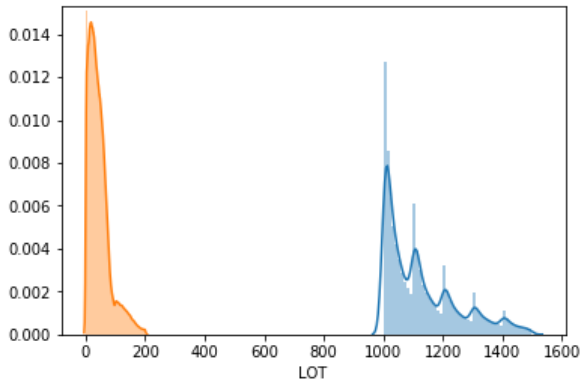
Brooklyn - 1 to 8,955

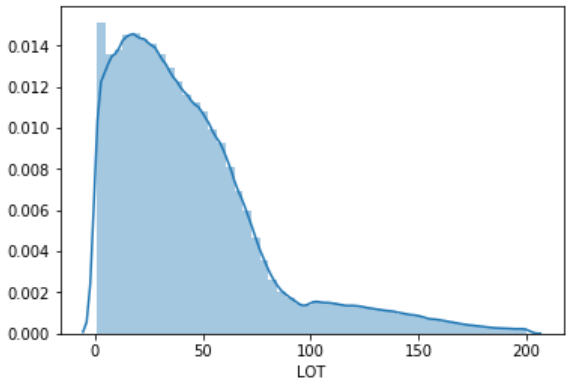
Queens - 1 to 16,350

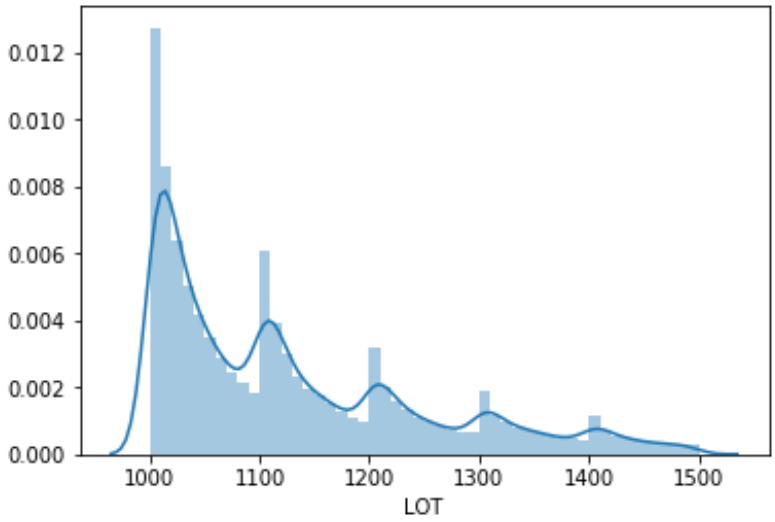
Staten Island - 1 to 8,050

LOT – Lot # within Block

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LOT	int64	1,048,575	370.09	860.54	1.00	23.00	49.00	146.00	9,978	100.00%	6,366

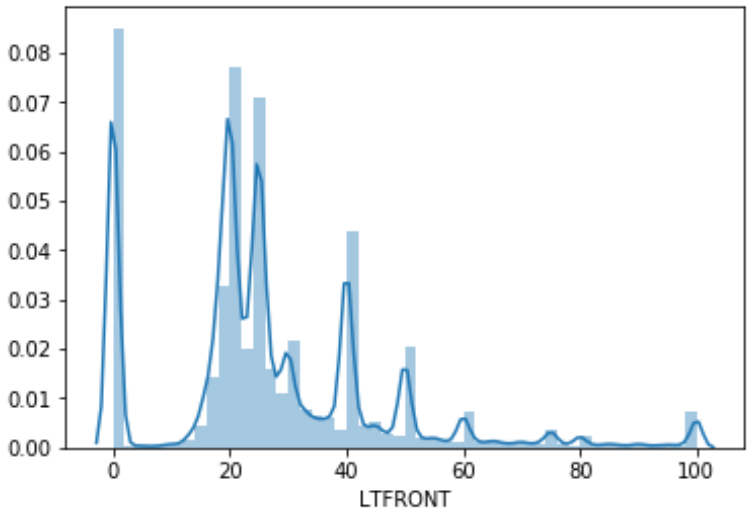






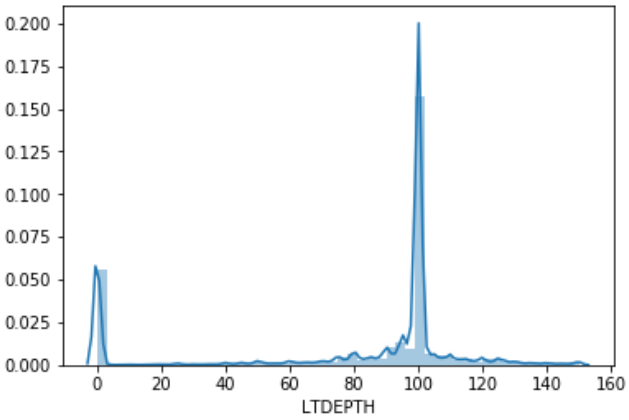
LTFRONT – Lot Width

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LTFRONT	int64	1,048,575	36.17	73.73	0.00	19.00	25.00	40.00	9,999	100.00%	1,277



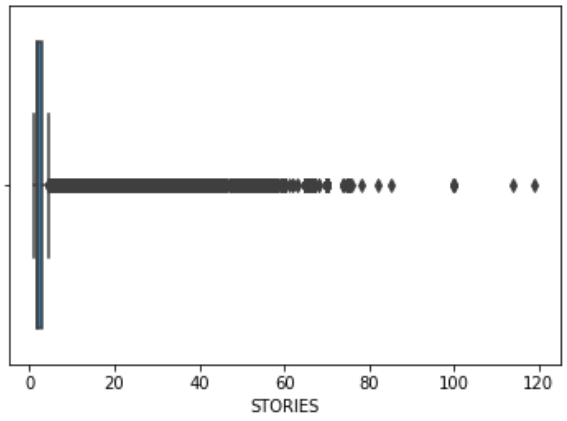
LTDEPTH – Lot Depth

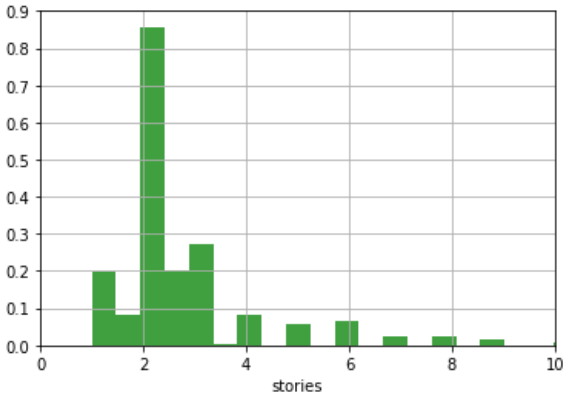
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
LTDEPTH	int64	1,048,575	88.28	75.48	0.00	80.00	100.00	100.00	9,999	100.00%	1,336

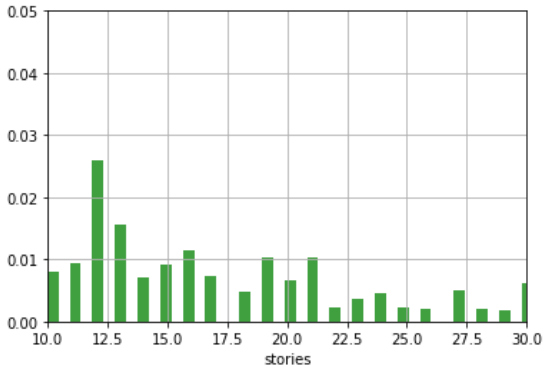


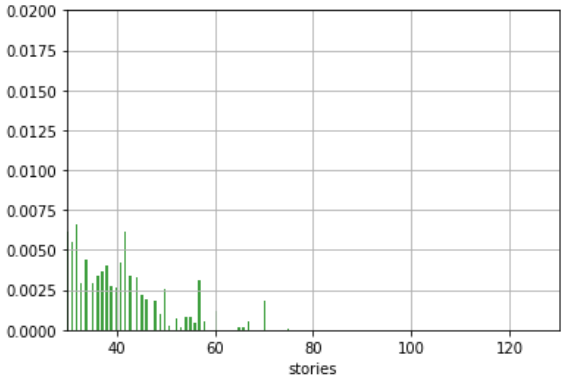
STORIES – Number of stories in the building

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
STORIES	float64	996,433	5.06	8.43	1.00	2.00	2.00	3.00	119	95.03%	111



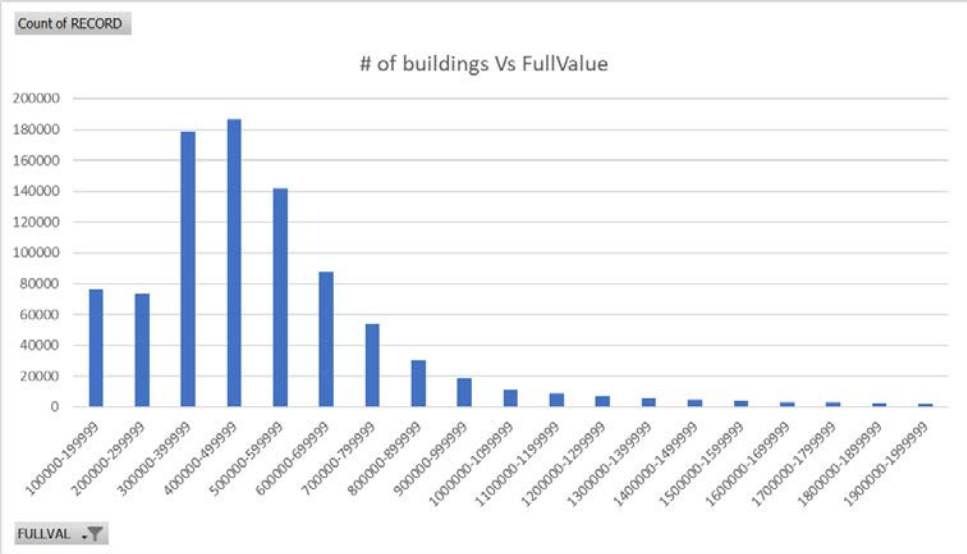


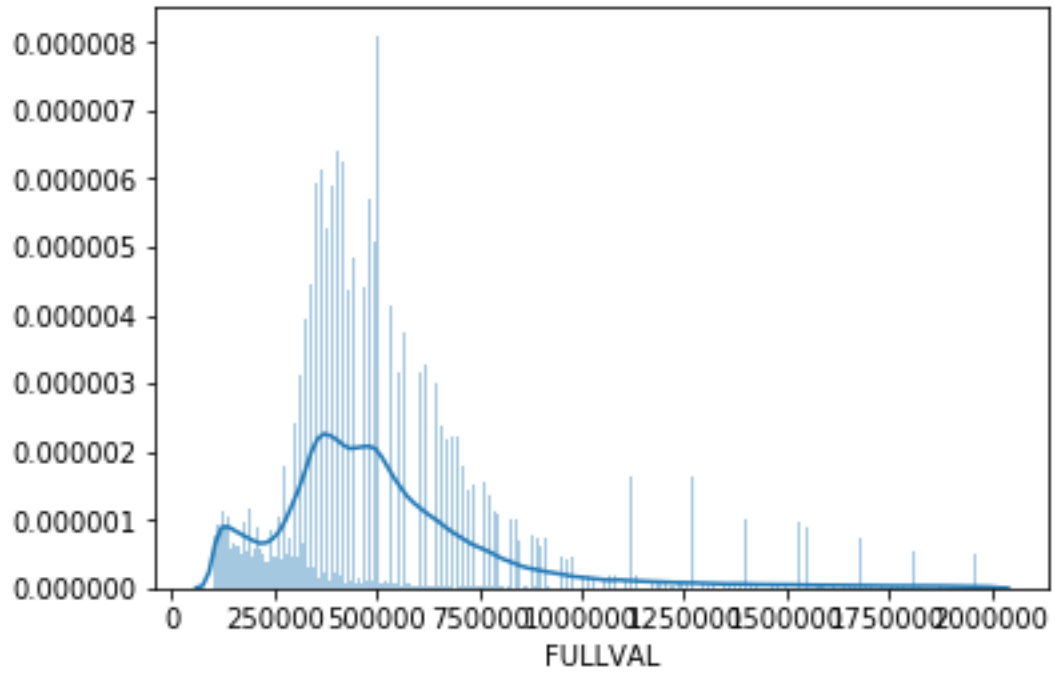




FULLVAL – Market Value

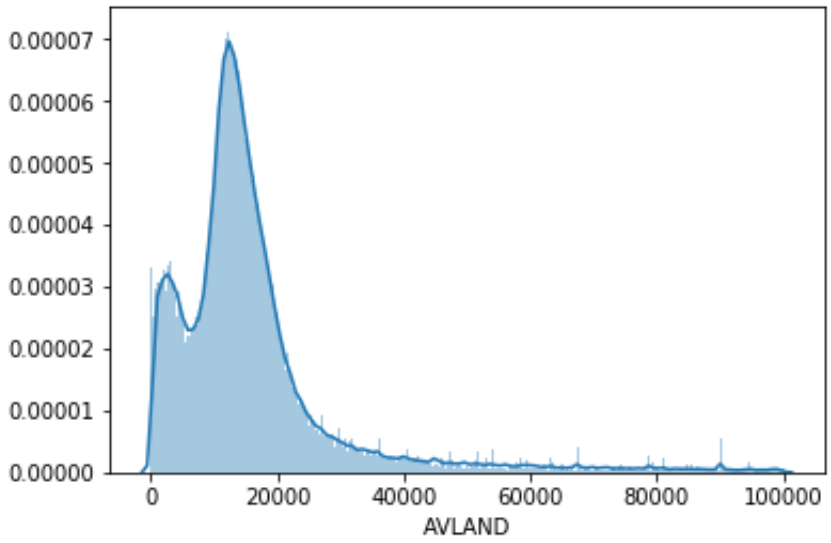
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
FULLVAL	int64	1,048,575	880,487.66	11,702,930.00	0.00	303,000.00	446,000.00	619,000.00	6,150,000,000	100.00%	108,277





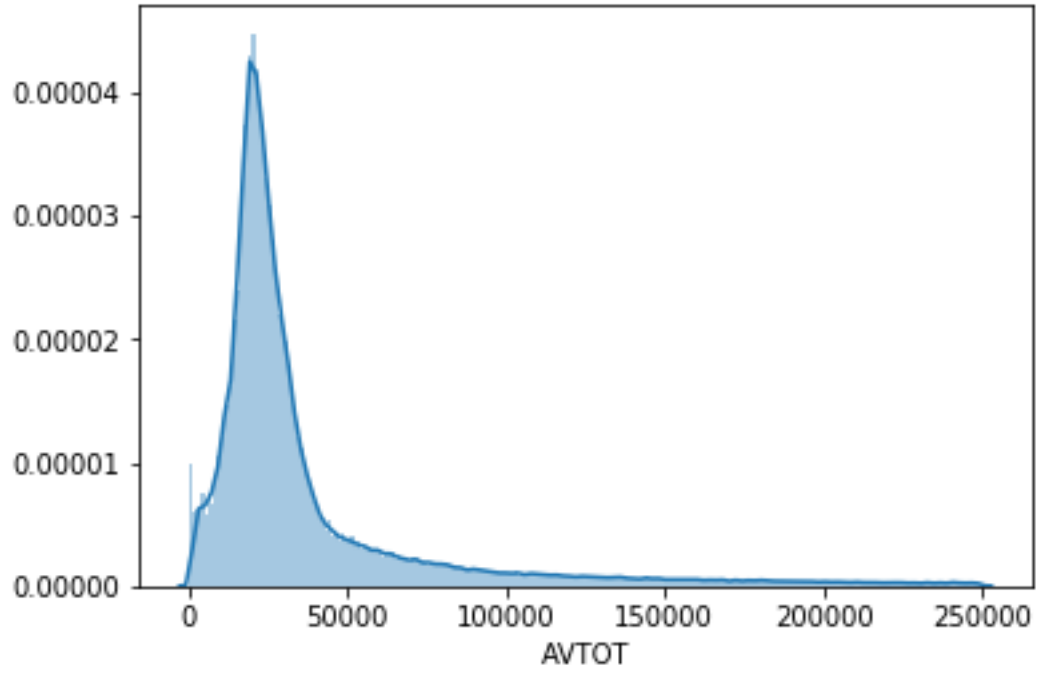
AVLAND – Actual Land Value

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVLAND	int64	1,048,575	85,995.03	4,100,755.00	0.00	9,160.00	13,646.00	19,706.00	2,668,500,000	100.00%	70,529



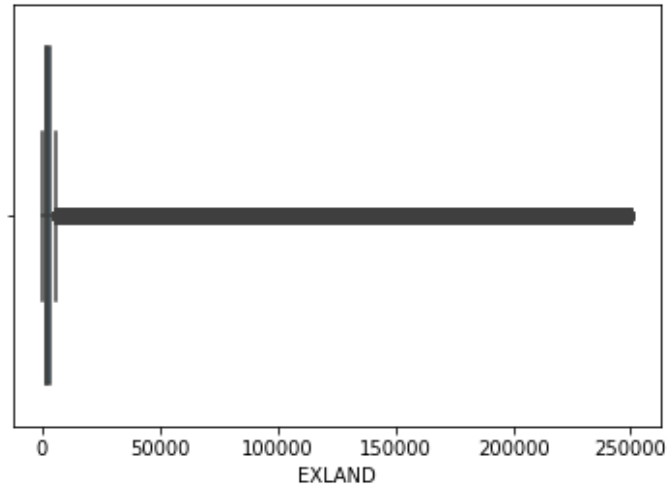
AVTOT – Actual Total Value

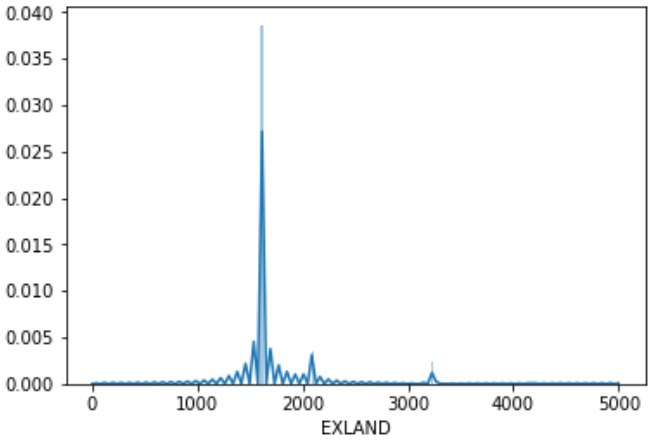
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVTOT	int64	1,048,575	230,758.18	6,951,206.00	0.00	18,385.00	25,339.00	46,095.00	4,668,309,000	100.00%	112,294



EXLAND – Actual Exempt Land Value

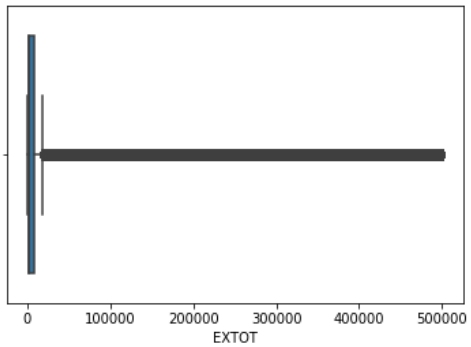
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXLAND	int64	1,048,575	36,811.79	4,024,330.00	0.00	0.00	1,620.00	1,620.00	2,668,500,000	100.00%	33,186

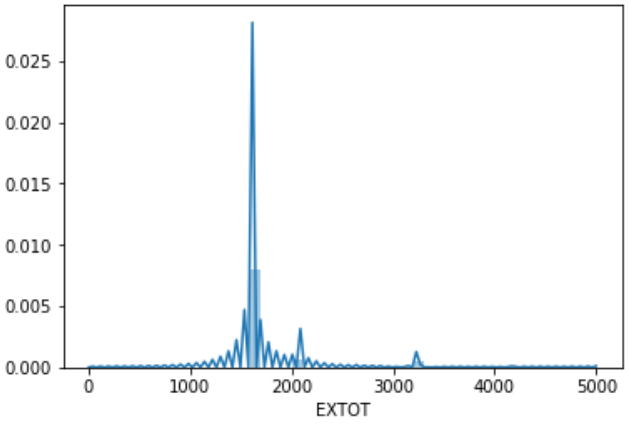




EXTOT – Actual Exempt Land Total

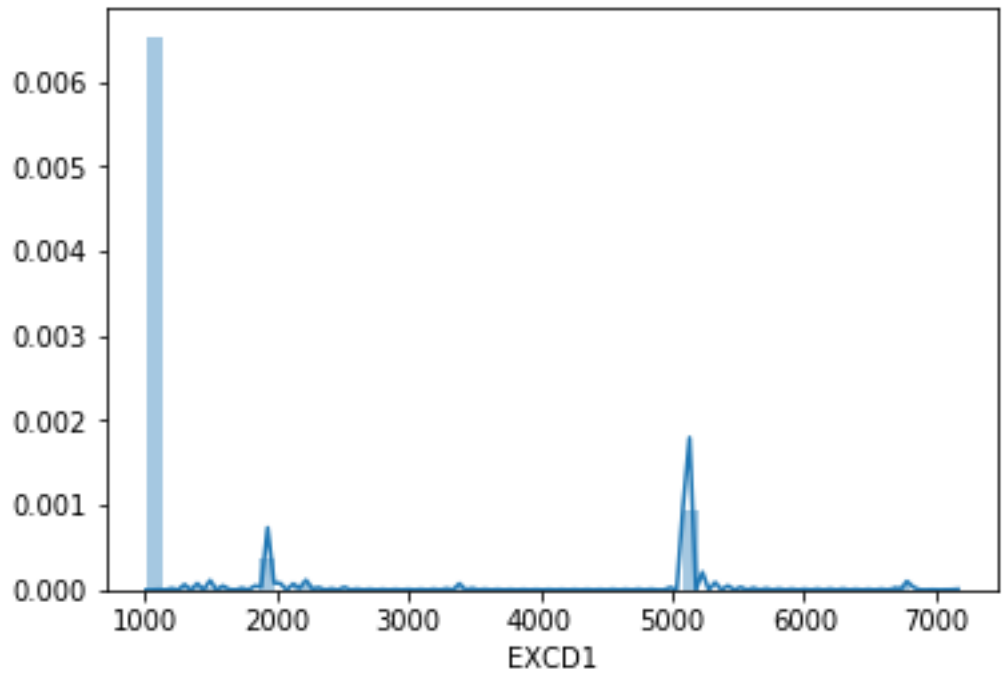
Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXTOT	int64	1,048,575	92,543.81	6,578,281.00	0.00	0.00	1,620.00	2,090.00	4,668,309,000	100.00%	63,805





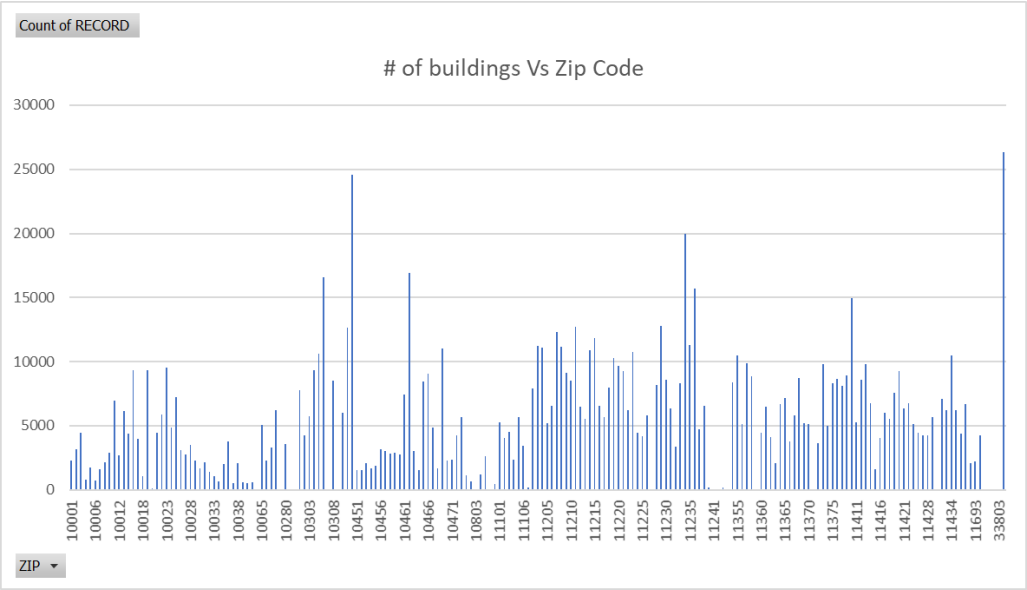
EXCD1 – Exemption Code 1

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXCD1	float64	622,642	1,604.50	1,388.13	1,010.00	1,017.00	1,017.00	1,017.00	7,170	59.38%	129



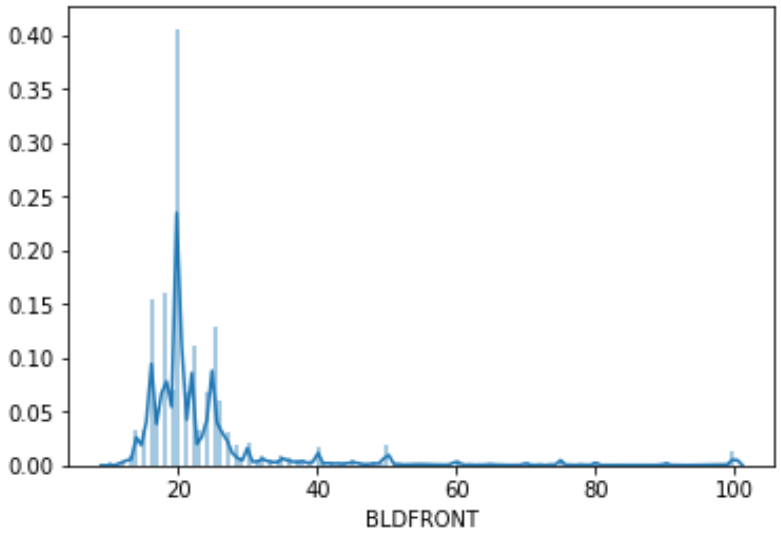
ZIP – Zip Code

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
ZIP	float64	1,022,219	10,935.32	526.58	10,001.00	10,453.00	11,215.00	11,364.00	33,803	97.49%	196



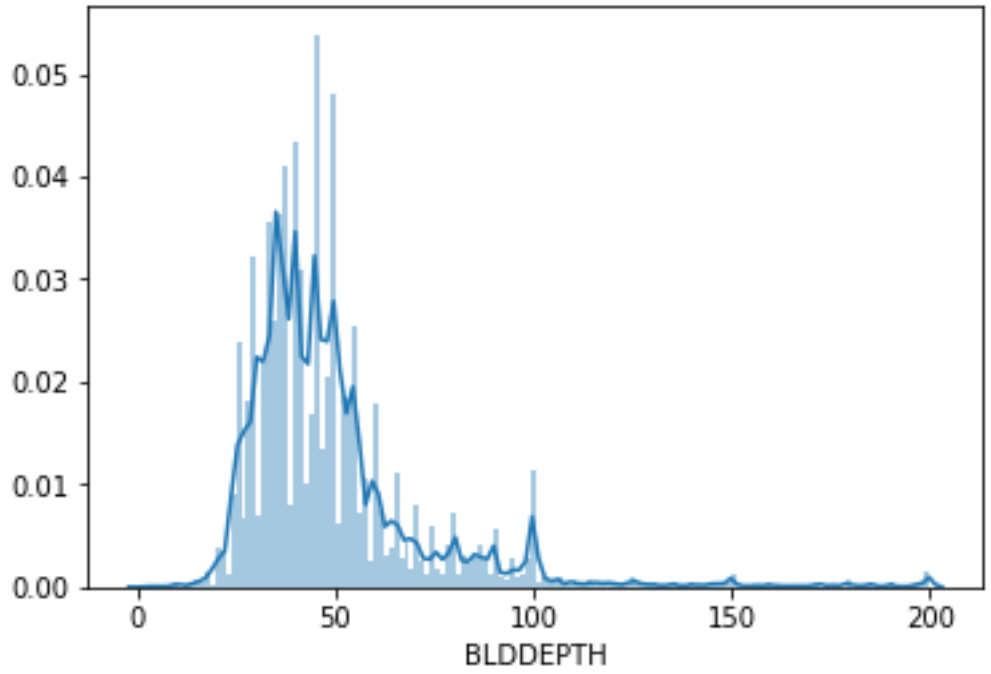
BLDFRONT – Building Width

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLDFRONT	int64	1,048,575	23.02	35.79	0.00	15.00	20.00	24.00	7,575	100.00%	610



BLDDEPTH – Building Depth

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
BLDDEPTH	int64	1,048,575	40.07	43.04	0.00	26.00	39.00	51.00	9,393	100.00%	620



AVLAND2 – Transitional Land Value

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVLAND2	float64	280,966	246,365.48	6,199,390.00	3.00	5,705.00	20,059.00	62,338.75	2,371,005,000	26.80%	58,169

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

AVTOT2 – Transitional Total Value

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
AVTOT2	float64	280,972	716,078.71	11,690,170.00	3.00	34,013.50	80,010.00	240,792.00	4,501,180,000	26.80%	110,890

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

EXLAND2 – Transitional Exempt Land Value

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXLAND2	float64	86,675	351,802.21	10,852,480.00	1.00	2,090.00	3,053.00	31,419.00	2,371,005,000	8.27%	21,996

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

EXTOT2 – Transitional Exempt Land Total

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXTOT2	float64	129,933	658,114.78	16,129,810.00	7.00	2,889.00	37,116.00	106,629.00	4,501,180,000	12.39%	48,106

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

EXCD2 – Exemption Code 2

Predictor	Data Type	count	mean	std	min	25%	50%	75%	max	Percentage Populated	# of unique values
EXCD2	float64	90,941	1,371.66	1,105.49	1,011.00	1,017.00	1,017.00	1,017.00	7,160	8.67%	60

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

EASEMENT – Easement Code

Predictor	Data Type	count	Percentage Populated	# of unique values
EASEMENT	object	4,043	0.39%	12

Very sparsely populated for accurate visualization to develop meaningful intuition of data.

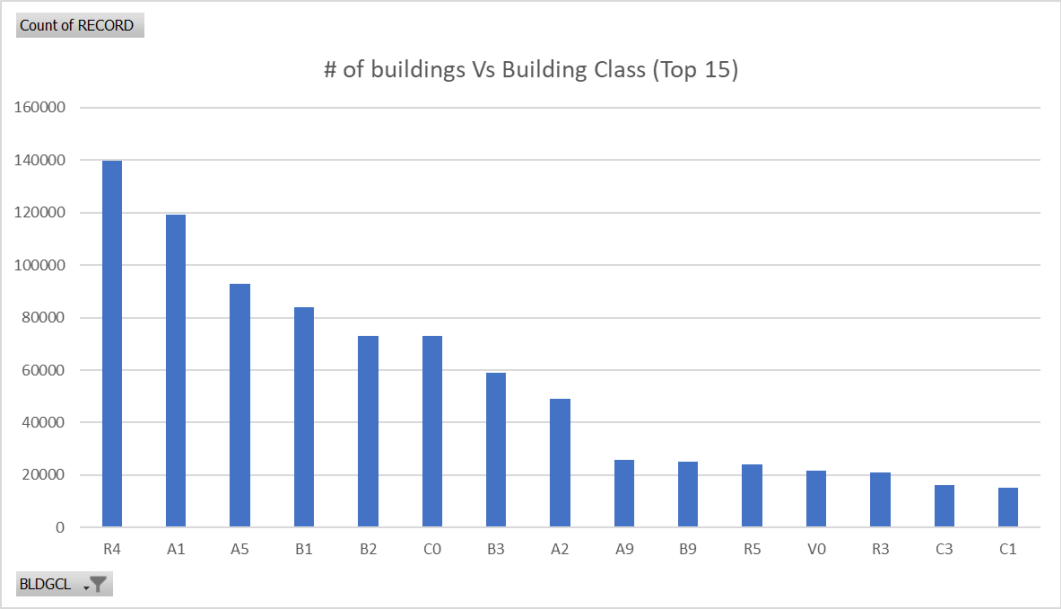
OWNER – Owner Name

Predictor	Data Type	count	Percentage Populated	# of unique values
OWNER	object	1,017,492	97.04%	847,053

Too many unique value therefore not visualizing this categorical variable. A better way to visualize will be to break down owners into public and private and then check the visualization.

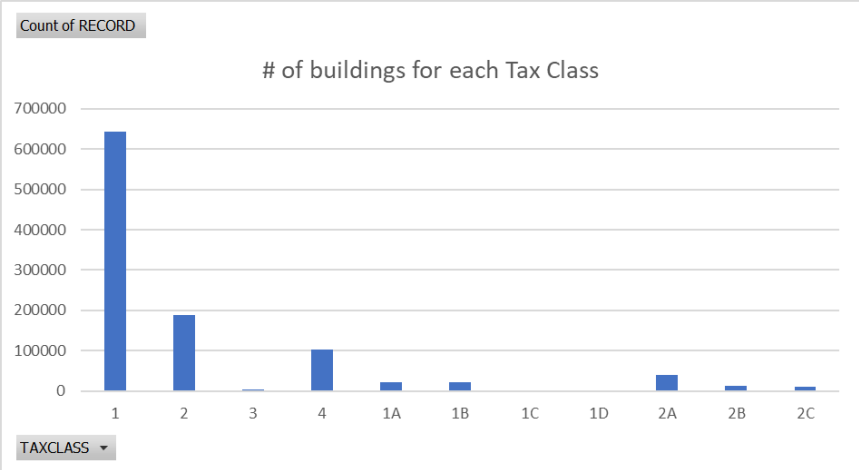
BLDGCL – Building Class

Predictor	Data Type	count	Percentage Populated	# of unique values
BLDGCL	object	1,048,575	100.00%	200



TAXCLASS – Tax Class

Predictor	Data Type	count	Percentage Populated	# of unique values
TAXCLASS	object	1,048,575	100.00%	11



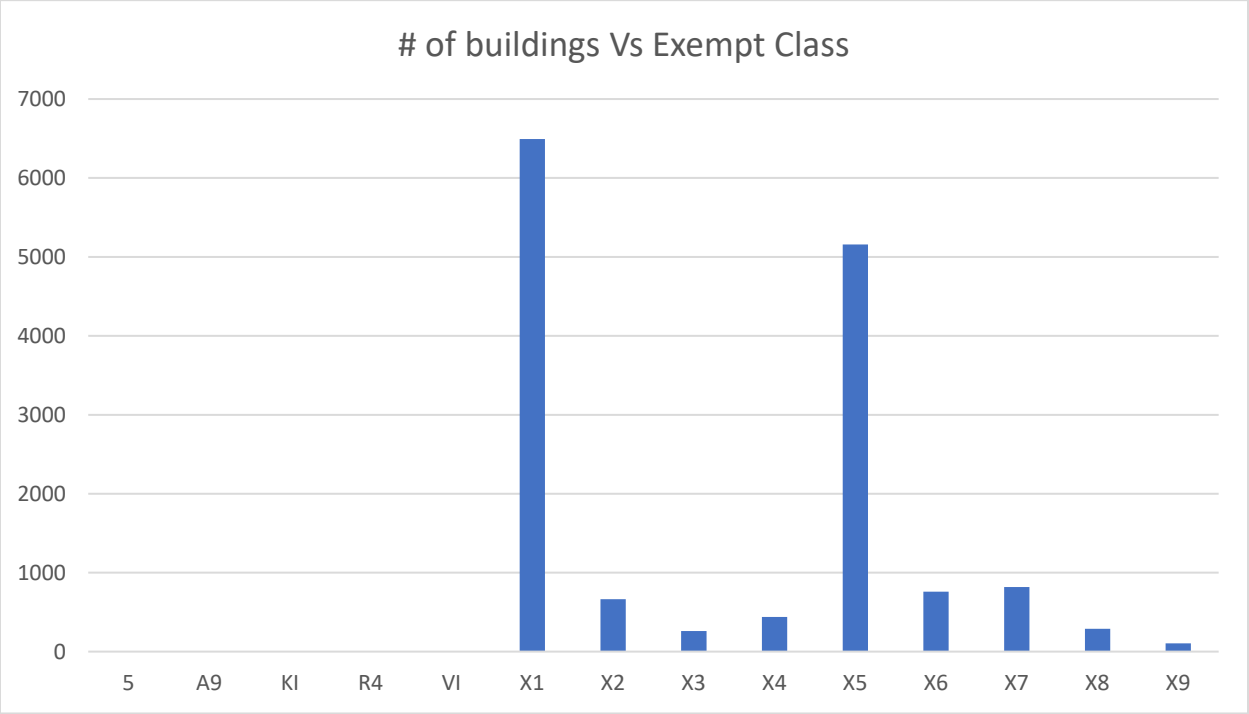
STADDR – Street Address

Predictor	Data Type	count	Percentage Populated	# of unique values
STADDR	object	1,047,934	99.94%	820,637

Too many unique value therefore not visualizing this categorical variable. A better way to visualize will be to put it on google map and create heat map visualization to see which areas are more populus.

EXMPTCL – Exempt Class

Predictor	Data Type	count	Percentage Populated	# of unique values
EXMPTCL	object	14,992	1.43%	14



PERIOD – Assessment Period

Predictor	Data Type	count	Percentage Populated	# of unique values
PERIOD	object	1,048,575	100.00%	1

Single variable – Final.

YEAR – Assessment Year

Predictor	Data Type	count	Percentage Populated	# of unique values
YEAR	object	1,048,575	100.00%	1

Single Variable – 2010/11

VALTYPE –

Predictor	Data Type	count	Percentage Populated	# of unique values
VALTYPE	object	1,048,575	100.00%	1

Single Variable – AC-TR