Boston Property Assessment Data Audit 2014-2017

Tyler Brown

2017-10-13

Summary

The purpose of a data audit is to dig up skeletons in the closet. Let's review the skeletons we found:

- `PID-year` is our unique ID for this dataset. It occurs more than once less than one percent of the time.
- About 27% of the variables are either filled out more than 70% of the time or failed a count of blank values test.
- Location appears to be most easily mapped using `ZIPCODE` instead of Latitude and Longitude because they're missing about 35% of the time.
- `GROSS_TAX` appears to be filled out without any incorrect entries.
- Additional data audit information can be found in the "property-assessment-2014-2017" pdf.

Useful Columns

We can get an initial idea about which columns are useful by seeing how often they're filled out. The following columns are not blank or NA more than 70% of the time.

Table 1: Percentage of columns at least 70% filled out

variables	num_blanks	num_nans	field_types	percent_nan
AV_BLDG	NA	NA	integer	NA
AV_LAND	NA	4	integer	0.00
AV_TOTAL	NA	NA	integer	NA
GROSS_AREA	NA	12669	integer	1.88
GROSS_TAX	NA	NA	integer	NA
LAND_SF	NA	12119	integer	1.80
LIVING_AREA	NA	12662	integer	1.88
LU	NA	NA	character	NA
NUM_FLOORS	NA	63602	numeric	9.46
OWN_OCC	NA	NA	character	NA
OWNER	NA	NA	character	NA
OWNER_MAIL_ZIPCODE	NA	2	character	0.00
PID	NA	NA	character	NA
PTYPE	NA	6	integer	0.00
ST_NAME	NA	NA	character	NA
ST_NAME_SUF	NA	8009	character	1.19
ST_NUM	NA	43726	character	6.50
X1	NA	NA	integer	NA
Year	NA	NA	integer	NA
YR_BUILT	NA	9773	integer	1.45
YR_REMOD	NA	114947	integer	17.10
ZIPCODE	NA	6	character	0.00

Most of our variables are missing from the above table. The next table shows the percentage of variables which are at least 70% populated.

populated_70
27.16

Next we want to double check those columns which came up as `NA` for the 'is blank' tests.

Table 3: NA values for Is Blank Test

variables	num_blanks	num_nans	field_types	percent_nan
AV_BLDG	NA	NA	integer	NA
AV_TOTAL	NA	NA	integer	NA
GROSS_TAX	NA	NA	integer	NA
LU	NA	NA	character	NA
OWN_OCC	NA	NA	character	NA
OWNER	NA	NA	character	NA
PID	NA	NA	character	NA
ST_NAME	NA	NA	character	NA
X1	NA	NA	integer	NA
Year	NA	NA	integer	NA

Come back to this later.

Checking for each Unique ID in Each Row

Concatenated `PID` and `year`. This should be a unique value for each row.

non_unique_ids
0.0066761

It looks like almost all of the time we have unique IDs. However, we will want to filter out the `PID-year` values which are not unique.

Checking for Variables that can be used to figure out location

Since a nontrivial amount of our project involves mapping, we want to make sure we're able to map `PID-year` to a location. In the appendix, `Latitude` and `Longitude` are shown to have `NA` values about 35% of the time. Maybe we can find a way to impute coordinate values if it makes sense.

The `ZIPCODE` variable appears to almost be completely filled out. This doesn't really matter for location purposes if all the zip codes are identical. Let's see how much fidelity we're getting out of `ZIPCODE`.

Table 5: Zip Code by Frequency

ZIPCODE	n
02127_	52171
02130_	45508
02135_	45265
02132_	43685
02124_	43564

ZIPCODE	n
02136_	36976
02116_	36863
02131_	34977
02128_	32983
02118_	32405
02125_	29462
02129_	27382
02119_	23947
02122_	23823
02115_	19549
02126_	19366
02114_	18615
02134_	17631
02121_	17082
02215_	14074
02111_	9822
02113_	8819
02108_	8356
02120_	7580
02109_	7420
02110_	6725
02467_	4103
02210_	3845
02199_	150
02445_	52
02446_	44
_	7
00000_	7
02137_	7
02201_	7
NA	6
0NULL_	5
02090_	4
02112_	4
02133_	4
2118	4
02186_	3
2120	3
2131	2
00003_	1
0000D_	1
00105_	1
00403_	1
00405_	1
00420_	1
2114	1
2116	1

It seems like trying the zip code route for location may be the best approach given currently available data.

Checking variables that can be used to measure value.

Table 6: Gross Tax Frequency, $100 Million Bins

GROSS_TAX_BIN	n
0	671709
1	351
2	80
3	44
4	42
5	22
6	15
7	13
8	2
9	6
10	10
11	3
12	3
13	9
14	2
17	2
19	2

There doesn't seem to be any junk data in `GROSS_TAX`.

Appendix

Here's the counts of what's blank or not.

Table 7: Count of blanks and NAs

variables	num_blanks	num_nans	field_types	percent_nan
AV_BLDG	NA	NA	integer	NA
AV_LAND	NA	4	integer	0.00
AV_TOTAL	NA	NA	integer	NA
CM_ID	NA	376374	character	55.98
GROSS_AREA	NA	12669	integer	1.88
GROSS_TAX	NA	NA	integer	NA
LAND_SF	NA	12119	integer	1.80
Latitude	NA	238267	numeric	35.44
LIVING_AREA	NA	12662	integer	1.88
Longitude	NA	238267	numeric	35.44
LU	NA	NA	character	NA
MAIL_ADDRESS	NA	332207	character	49.41
MAIL_ADDRESSEE	NA	481657	character	71.64
MAIL_CS	NA	332206	character	49.41
NUM_FLOORS	NA	63602	numeric	9.46
OWN_OCC	NA	NA	character	NA
OWNER	NA	NA	character	NA
OWNER_MAIL_ADDRESS	NA	340109	character	50.59
OWNER_MAIL_CS	NA	340109	character	50.59

variables	num_blanks	num_nans	field_types	percent_nan
OWNER_MAIL_ZIPCODE	NA	2	character	0.00
PID	NA	NA	character	NA
PTYPE	NA	6	integer	0.00
R_AC	NA	420725	character	62.58
R_BDRMS	NA	319514	integer	47.52
R_BLDG_STYL	NA	420720	character	62.58
R_BTH_STYLE	NA	546632	character	81.31
R_BTH_STYLE2	NA	563727	character	83.85
R_BTH_STYLE3	NA	619901	character	92.20
R_EXT_CND	NA	546632	character	81.31
R_EXT_FIN	NA	420727	character	62.58
R_FPLACE	NA	319522	integer	47.53
R_FULL_BTH	NA	319512	integer	47.52
R_HALF_BTH	NA	319512	integer	47.52
R_HEAT_TYP	NA	420717	character	62.58
R_INT_CND	NA	546632	character	81.31
R_INT_FIN	NA	546632	character	81.31
R_KITCH	NA	319512	integer	47.52
R_KITCH_STYLE	NA	546632	character	81.31
R_KITCH_STYLE2	NA	606581	character	90.22
R_KITCH_STYLE3	NA	643226	character	95.67
R_OVRALL_CND	NA	546632	character	81.31
R_ROOF_TYP	NA	420727	character	62.58
R_TOTAL_RMS	NA	319522	integer	47.53
R_VIEW	NA	546632	character	81.31
S_BLDG_STYL	NA	636258	character	94.64
S_EXT_CND	NA	653720	character	97.23
S_EXT_FIN	NA	636258	character	94.64
S_NUM_BLDG	NA	497445	integer	73.99
S_UNIT_COM	NA	480654	integer	71.49
S_UNIT_RC	NA	480652	integer	71.49
S_UNIT_RES	NA	480660	integer	71.49
ST_NAME	NA	NA	character	NA
ST_NAME_SUF	NA	8009	character	1.19
ST_NUM	NA	43726	character	6.50
STRUCTURE_CLASS	NA	336522	character	50.05
U_AC	NA	431052	character	64.11
U_BASE_FLOOR	NA	431059	integer	64.12
U_BDRMS	NA	431054	integer	64.11
U_BTH_STYLE	NA	550513	character	81.88
U_BTH_STYLE2	NA	626465	character	93.18
U_BTH_STYLE3	NA	660481	character	98.24
U_CORNER	NA	431061	character	64.12
U_FPLACE	NA	431061	integer	64.12
U_FULL_BTH	NA	431052	integer	64.11
U_HALF_BTH	NA	431060	integer	64.12
U_HEAT_TYP	NA	431056	character	64.12
U_INT_CND	NA	550513	character	81.88
U_INT_FIN	NA	550513	character	81.88
U_KIT_TYPE	NA	552854	character	82.23
U_KITCH_STYLE	NA	550513	character	81.88
U_KITCH_TYPE	NA	550513	character	81.88

variables	num_blanks	num_nans	field_types	percent_nan
U_NUM_PARK	NA	490465	character	72.95
U_ORIENT	NA	431052	character	64.11
U_TOT_RMS	NA	431067	integer	64.12
U_VIEW	NA	550513	character	81.88
UNIT_NUM	NA	424631	character	63.16
X1	NA	NA	integer	NA
Year	NA	NA	integer	NA
YR_BUILT	NA	9773	integer	1.45
YR_REMOD	NA	114947	integer	17.10
ZIPCODE	NA	6	character	0.00