Analysis of quake damage

Seema G, April 2018

Executive Summary

In 2015, a high magnitude earthquake in Nepal caused significant destruction. Some of these casualties happened in buildings that collapsed in the earthquake, and may have been preventable if they had withstood the initial ground motion or resulting aftershocks. This analysis attempts to analyze data on buildings in the affected area and how they were impacted by the earthquake, and model risk of damage.
https://datasciencecapstone.org/competitions/4/earthquake-damage/ has detailed information.

Labeled data contains 10K rows and 38 features and includes information on earthquake impacts, household conditions, and socio-economic-demographic statistics. Target variable for prediction is the ordinal value damage_grade, which represents a level of damage to the building that was hit by the earthquake. There are 3 grades of the damage:

- 1 represents low damage
- 2 represents a medium amount of damage
- 3 represents almost complete destruction

This is a multi-class classification problem. Performance is measured using F-micro score that balances precision and recall.

Initial step involved exploring data, gathering summary statistics and creating visualizations of data to identify potential relationships. After exploring the data, a multi-class classification model was created to predict the damage level.

After performing the analysis, in conclusion, while many factors can help predict damage level, significant features found in this analysis were:

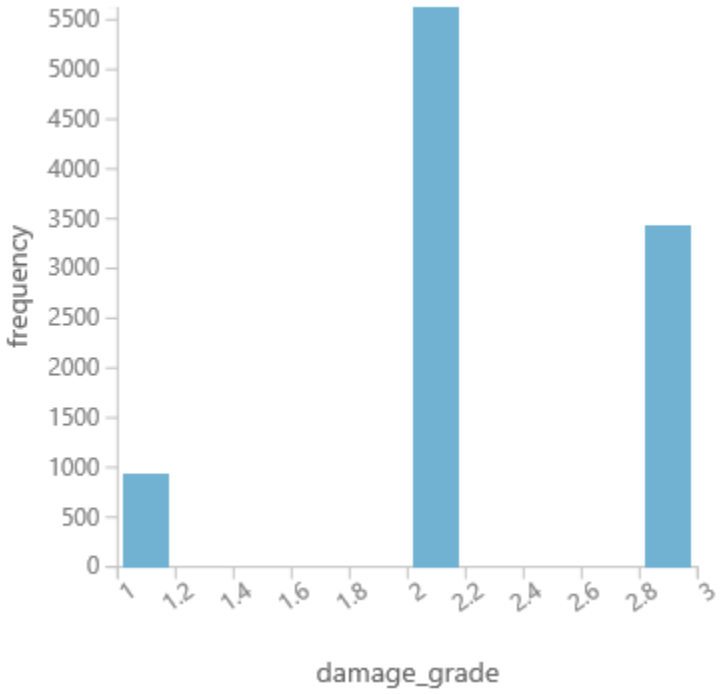
Geo_level: geographic region in which building exists, from largest to most specific sub-region.

Superstructure: flag variables that indicate if the superstructure was made of cement mortar – stone, or timber or mud mortar – stone, or mud mortar – brick or other type.

Age, Area, height, Position: Age, area, height and position of the building. Most of the buildings with low damage were below average height. Most of the buildings with an above average area had medium damage.

Initial Data Exploration

Dependent variable damage_grade has the following distribution. This shows most of the buildings suffered medium damage.



Summary statistics for the building age, height and area are as follows.

	age	area	height	damage_grade
count	10000.000000	10000.000000	10000.000000	10000.000000
mean	25.393500	38.438100	4.653100	2.248800
std	64.482893	21.265883	1.792842	0.611993
min	0.000000	6.000000	1.000000	1.000000
25%	10.000000	26.000000	4.000000	2.000000
50%	15.000000	34.000000	5.000000	2.000000
75%	30.000000	44.000000	5.000000	3.000000
max	995.000000	425.000000	30.000000	3.000000

Following observations were derived from the data.

The mean age of buildings with a damage_grade of 2 is higher than for buildings with a damage_grade of 1.

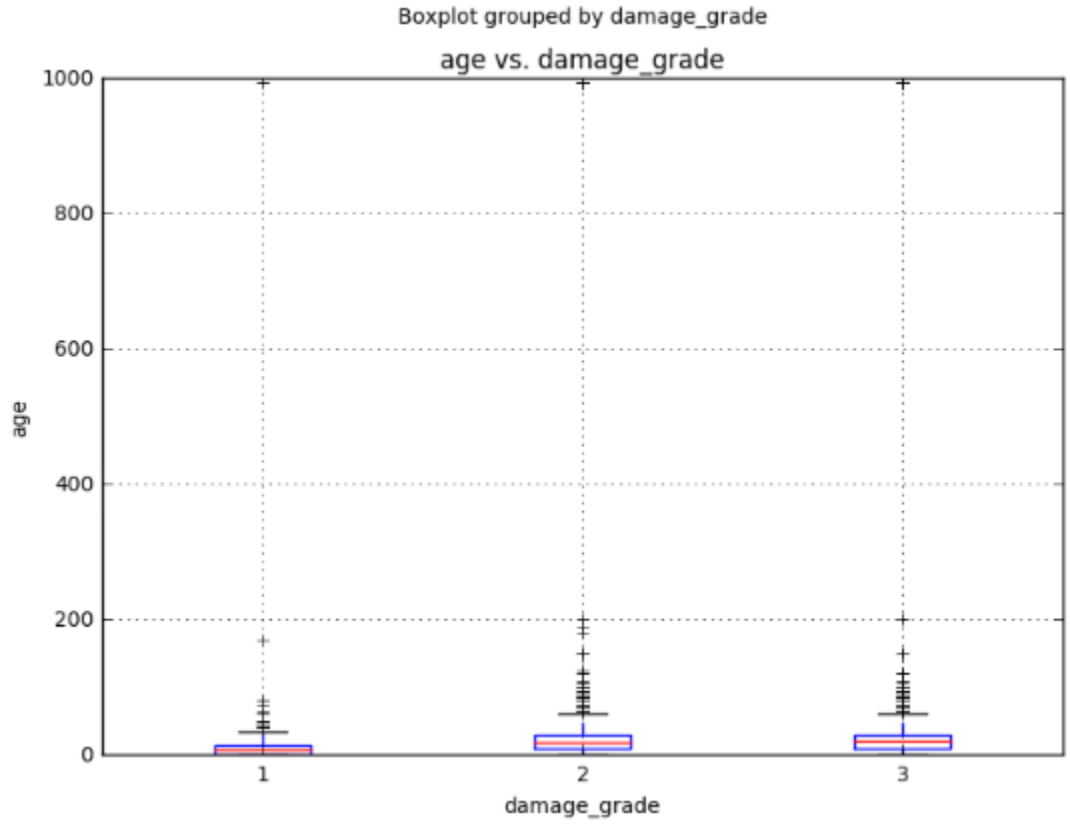
The mean age of buildings with a damage_grade of 2 is higher than for buildings with a damage_grade of 3.

Most of the buildings with a damage_grade of 1 are below average height.

Most of the buildings with an above average area have a damage_grade of 2.

Outliers

Outliers were observed in the data when damage_grade was plotted against age.

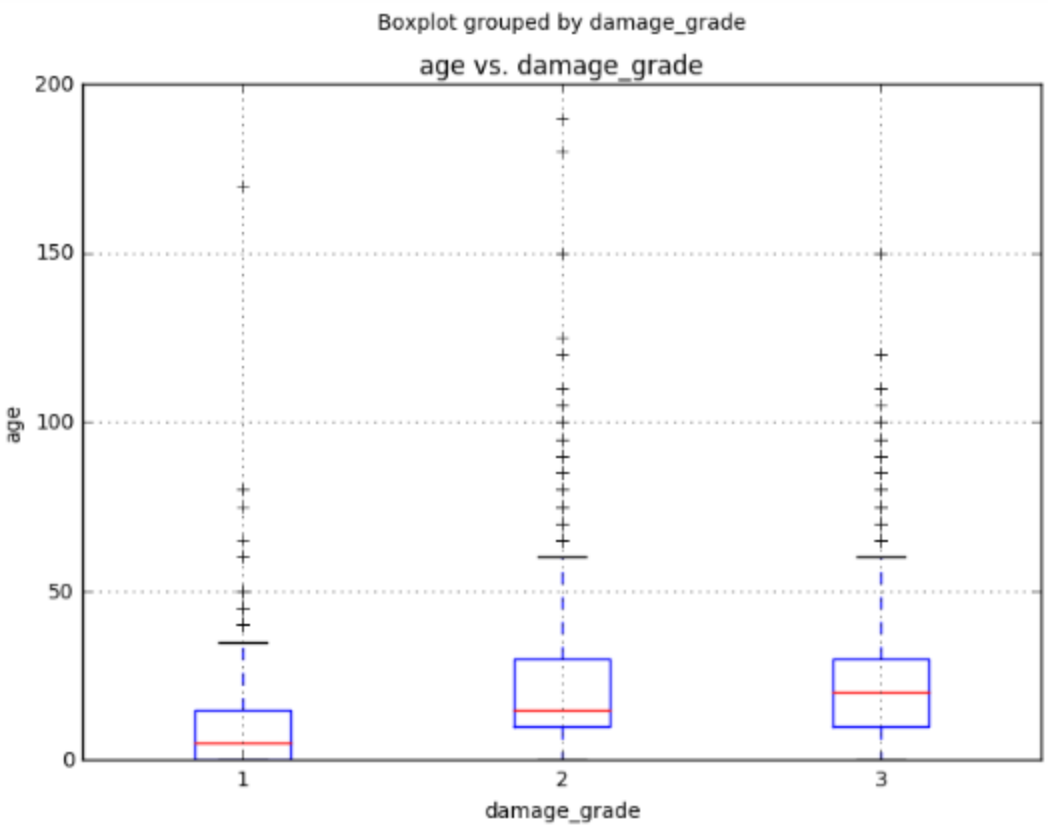


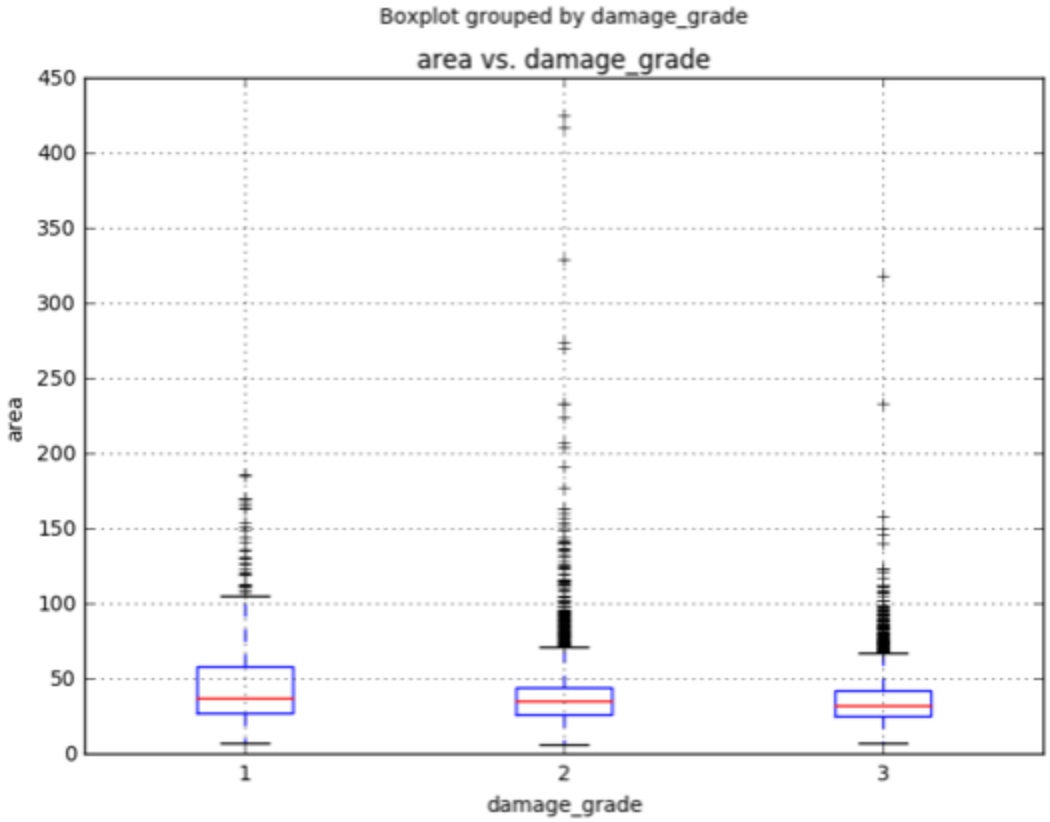
This helped identify the following rows of outliers where age of the building was specified to be 995 years.

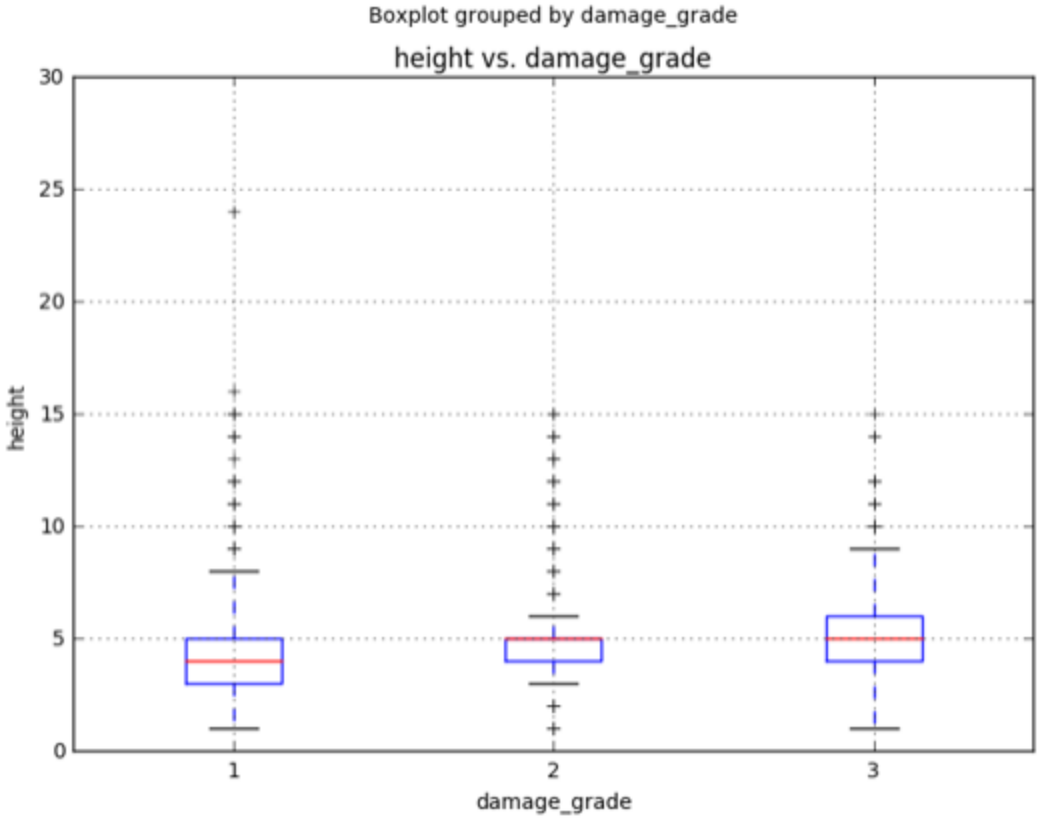
	building_id	age	area	height	damage_grade
182	15472	995	18	6	2
326	9693	995	30	6	3
625	14944	995	36	4	2
800	7744	995	76	8	2

	building_id	age	area	height	damage_grade
1529	7495	995	23	4	3
1640	11940	995	70	4	1
2205	14549	995	18	7	3
2244	19066	995	19	7	2
2246	2343	995	64	6	2
2427	13615	995	42	5	2
2654	14874	995	36	4	3
2708	19450	995	53	5	2
3318	17918	995	24	8	2
3367	11026	995	37	2	2
3488	6599	995	28	5	2
3645	12142	995	35	5	3
3688	19774	995	35	4	2
4486	18979	995	40	4	2
4866	15209	995	30	4	2
5071	14622	995	37	4	2
5076	5677	995	53	5	3
5196	4147	995	24	5	2
5204	4098	995	28	3	2
5296	14768	995	70	6	1
5859	570	995	30	6	2
6267	19983	995	58	3	2
6280	18327	995	35	7	2
7065	5068	995	19	4	3
7180	1151	995	26	1	3
7244	19977	995	92	2	3
7246	17809	995	46	7	3
7327	16018	995	42	2	2
7574	488	995	20	9	2
7580	16087	995	50	2	2
7626	2191	995	30	4	2
8099	2244	995	44	5	2
8174	193	995	33	6	2
8778	1203	995	56	12	2
9148	16148	995	14	5	2
9433	7486	995	32	6	2

After removing outliers, damage grade is distributed as follows by age, area and height.


Boxplot grouped by damage_grade
age vs. damage_grade
200
150
100
50
0
age
1
2
3
damage_grade


Boxplot grouped by damage_grade
area vs. damage_grade
area
damage_grade
0
50
100
150
200
250
300
350
400
450
1
2
3


Boxplot grouped by damage_grade
height vs. damage_grade
height
damage_grade
0
5
10
15
20
25
30
1
2
3

Data preparation

AzureML was used for data preparation, modeling and predictions. Data preparation involved the following steps.

- Making categorical features categorical using Edit Meta Data module.
- Z-score normalization for numerical features of age, height and area.
- Removing feature flag for building_id.

Permutation Feature Importance

PFI is an algorithm that computes importance scores for each of the feature variables of a dataset. The importance measures are determined by computing the sensitivity of a model to random permutations of feature values. In other words, an importance score quantifies the contribution of a certain feature to the predictive performance of a model in terms of how much a chosen evaluation metric deviates after permuting the values of that feature. The intuition behind permutation importance is that if a feature is not useful for predicting an outcome, then altering or permuting its values will not result in a significant reduction in a model's performance [1]. PFI output is as follows.

Feature	Score
geo_level_1_id	0.148092
geo_level_2_id	0.037651
has_superstructure_cement_mortar_brick	0.008534
position	0.005522
has_superstructure_timber	0.005522
has_secondary_use	0.00502
age	0.004016
area	0.004016
count_families	0.003012
has_superstructure_mud_mortar_stone	0.00251
has_superstructure_mud_mortar_brick	0.00251
plan_configuration	0.002008
has_superstructure_bamboo	0.002008
geo_level_3_id	0.001506
ground_floor_type	0.001506
has_superstructure_rc_non_engineered	0.001004
has_superstructure_adobe_mud	0.000502
has_superstructure_other	0.000502
foundation_type	0
has_secondary_use_institution	0
has_secondary_use_school	0
has_secondary_use_industry	0
has_secondary_use_health_post	0

has_secondary_use_gov_office	0
has_secondary_use_use_police	0
has_secondary_use_other	0

Based on these observations, legal_owndership_status and has_secondary_use_* features were removed from the dataset.

Multi-class classification

Prepared labeled data was split into 70% training data and 30% test data. Data was trained using multi-class logistic regression to predict the dependent variable damage_grade. Performance metrics and confusion matrix are as follows. F-micro score is 0.702.

F-micro score balances between precision and recall and is used for optimizing the model performance [2].

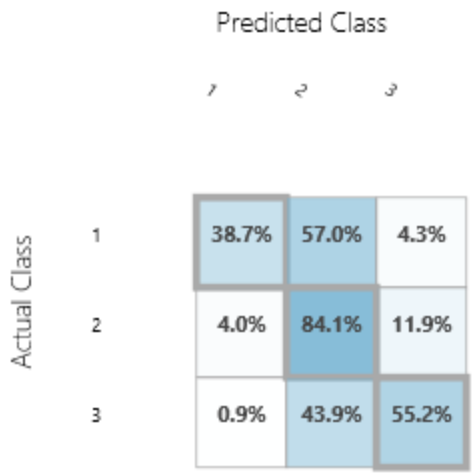
$$F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}}$$

where

$$P_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3}(TP_k + FP_k)}, R_{micro} = \frac{\sum_{k=1}^{3} TP_k}{\sum_{k=1}^{3}(TP_k + FN_k)}$$

and TP is True Positive, FP is False Positive, FN is False Negative, and k represents each class in $1, 2, 3$.

Overall accuracy	0.702811
Average accuracy	0.801874
Micro-averaged precision	0.702811
Macro-averaged precision	0.668668
Micro-averaged recall	0.702811
Macro-averaged recall	0.593266



Labeled data had 10K rows and private test data also had 10K rows. Then the same steps were performed on the private test data to predict damage_grade. After submission F-micro score for private test data was 0.7086.

Conclusion

This analysis has shown that the amount of damage to the buildings during a quake can be predicted (within a reasonable factor) from the geographic region (geo_level_*), superstructure (has_superstructure_*), age, area, height and position information.

Reference

[1] https://blogs.technet.microsoft.com/machinelearning/2015/04/14/permutation-feature-importance/

[2] https://datasciencecapstone.org/competitions/4/earthquake-damage/

Appendix A.

https://datasciencecapstone.org/competitions/4/earthquake-damage/ has detailed description of the problem.

Data description:

- `geo_level_1_id`, `geo_level_2_id`, `geo_level_3_id` (type: categorical): geographic region in which building exists, from largest (level 1) to most specific sub-region (level 3).

- `count_floors_pre_eq` (type: int): number of floors in the building before the earthquake.
- `age` (type: int): age of the building in years.
- `area` (type: int): plinth area of the building in m^2m2.
- `height` (type: int): height of the building in mm.
- `land_surface_condition` (type: categorical): surface condition of the land where the building was built. Possible values: d502, 808e, 2f15.
- `foundation_type` (type: categorical): type of foundation used while building. Possible values: 337f, 858b, 6c3e, 467b, bb5f.
- `roof_type` (type: categorical): type of roof used while building. Possible values: 7e76, e0e2, 67f9.
- `ground_floor_type` (type: categorical): type of the ground floor. Possible values: b1b4, b440, 467b, e26c, bb5f.
- `other_floor_type` (type: categorical): type of constructions used in higher than the ground floors (except of roof). Possible values: f962, 9eb0, 441a, 67f9.
- `position` (type: categorical): position of the building. Possible values: 3356, bfba, bcab, 1787.
- `plan_configuration` (type: categorical): building plan configuration. Possible values: a779, 84cf, 8e3f, d2d9, 3fee, 6e81, 0448, 1442, cb88.
- `has_superstructure_adobe_mud` (type: binary): flag variable that indicates if the superstructure was made of Adobe/Mud.
- `has_superstructure_mud_mortar_stone` (type: binary): flag variable that indicates if the superstructure was made of Mud Mortar - Stone.
- `has_superstructure_stone_flag` (type: binary): flag variable that indicates if the superstructure was made of Stone.
- `has_superstructure_cement_mortar_stone` (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Stone.
- `has_superstructure_mud_mortar_brick` (type: binary): flag variable that indicates if the superstructure was made of Mud Mortar - Brick.
- `has_superstructure_cement_mortar_brick` (type: binary): flag variable that indicates if the superstructure was made of Cement Mortar - Brick.
- `has_superstructure_timber` (type: binary): flag variable that indicates if the superstructure was made of Timber.
- `has_superstructure_bamboo` (type: binary): flag variable that indicates if the superstructure was made of Bamboo.
- `has_superstructure_rc_non_engineered` (type: binary): flag variable that indicates if the superstructure was made of non-engineered reinforced concrete.
- `has_superstructure_rc_engineered` (type: binary): flag variable that indicates if the superstructure was made of engineered reinforced concrete.
- `has_superstructure_other` (type: binary): flag variable that indicates if the superstructure was made of any other material.
- `legal_ownership_status` (type: categorical): legal ownership status of the land where building was built. Possible values: c8e1, cae1, ab03, bb5f.
- `count_families` (type: float): number of families that live in the building.
- `has_secondary_use` (type: binary): flag variable that indicates if the building was used for any secondary purpose.
- `has_secondary_use_agriculture` (type: binary): flag variable that indicates if the building was used for agricultural purposes.

- `has_secondary_use_hotel` (type: binary): flag variable that indicates if the building was used as a hotel.
- `has_secondary_use_rental` (type: binary): flag variable that indicates if the building was used for rental purposes.
- `has_secondary_use_institution` (type: binary): flag variable that indicates if the building was used as a location of any institution.
- `has_secondary_use_school` (type: binary): flag variable that indicates if the building was used as a school.
- `has_secondary_use_industry` (type: binary): flag variable that indicates if the building was used for industrial purposes.
- `has_secondary_use_health_post` (type: binary): flag variable that indicates if the building was used as a health post.
- `has_secondary_use_gov_office` (type: binary): flag variable that indicates if the building was used fas a government office.
- `has_secondary_use_use_police` (type: binary): flag variable that indicates if the building was used as a police station.
- `has_secondary_use_other` (type: binary): flag variable that indicates if the building was secondarily used for other purposes.