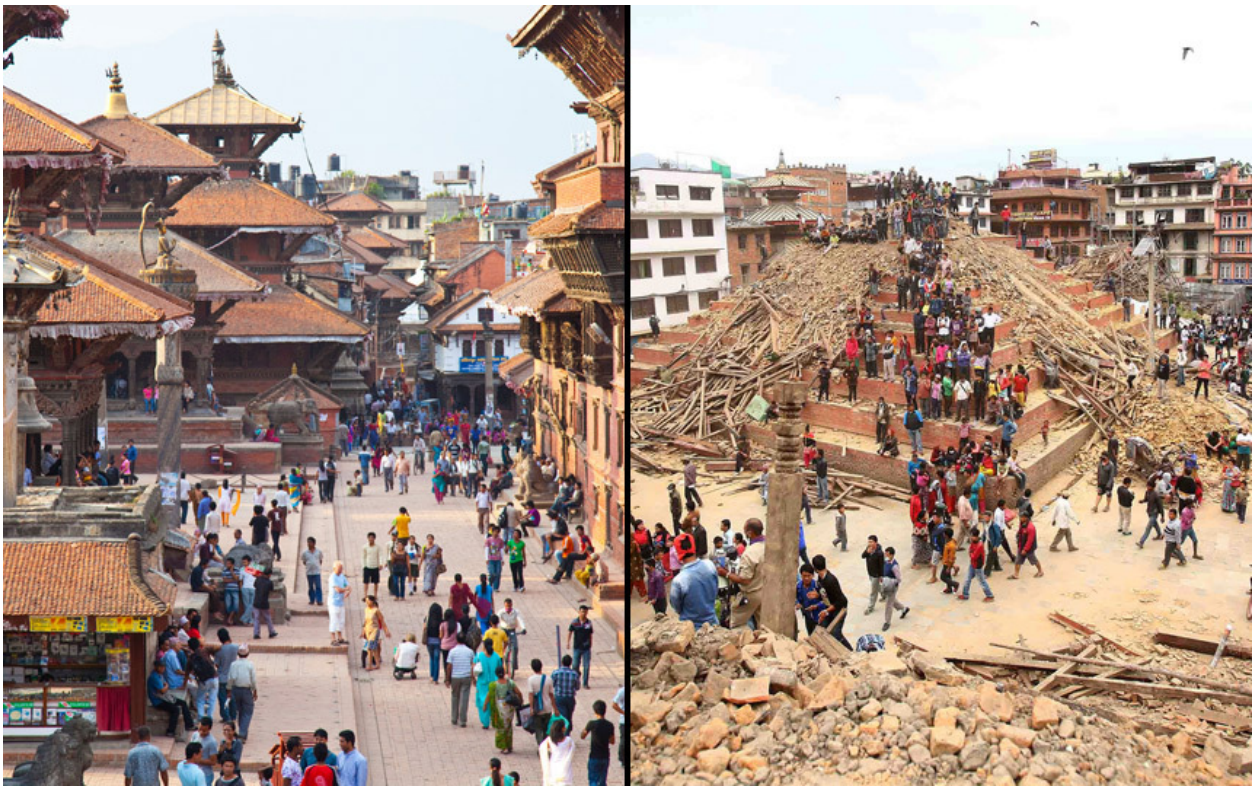
Report Harvard Data Science Capstone

Predicting Earthquake Damage in Nepal

Author: philk1337

2019-02-01



Contents

1 Executive summary

The data used for this analysis represents aspects of building location and construction after an earthqauake in nepal.

It was collected through surveys by the Central Bureau of Statistics that work under the National Planning Commission Secretariat of Nepal. This survey is one of the largest post-disaster datasets ever collected, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics.

In the analysis we start by getting a descriptive overview of the data in the dataset. Then we will try to find correlations in the data and have a look at the distribution of data types. After having an overview of the data we dive in and analyze specific features to find out how specific features like height are correlated with others.

Next part is a description of the methods used to predict the damage category.

At the end of this report you will find a section with more sophisticated findings and recommendations.

2 Analysis

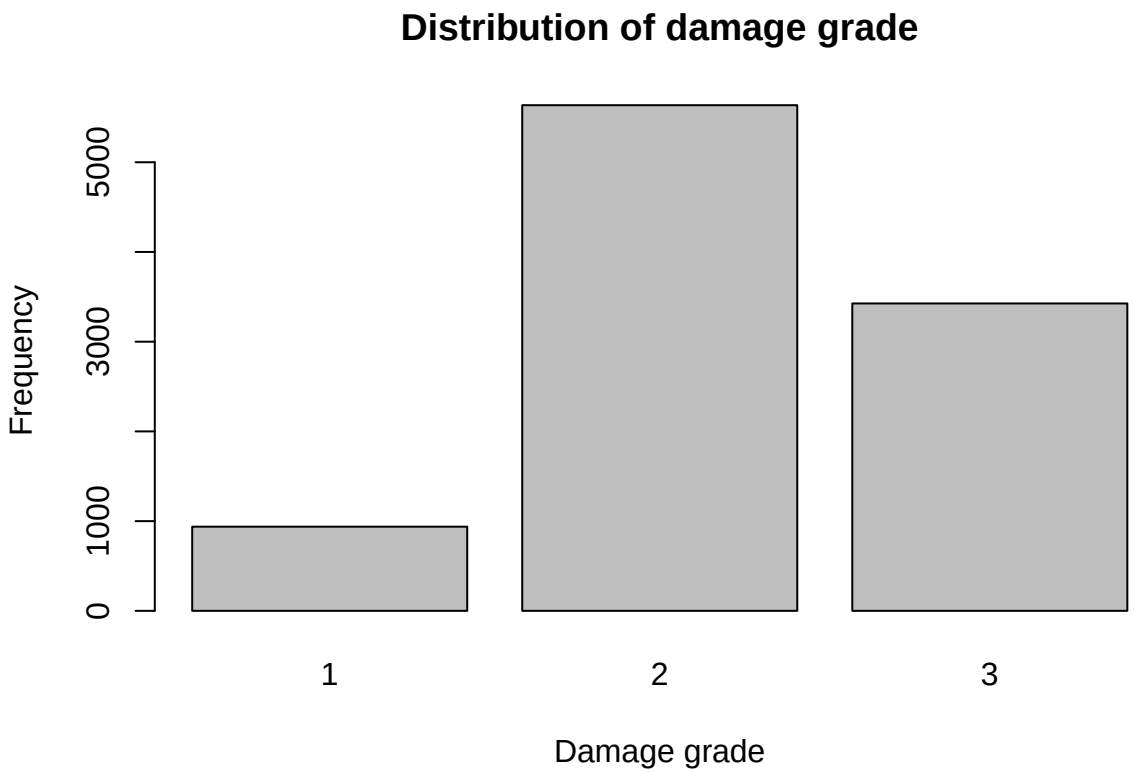
In a first step we will get an overwiew of the dataset and the variables in the dataset. This will be done in the chapter “descriptive statistics”. After that we will try to bring some of them together and find correlations between them in th chapter “correlations”

2.1 Descriptive statistics

Table 1: Descriptive Statistics

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
building_id	1	10000	9987.1600	5800.8008293	9963.5	9986.137500	7432.2738	1	19999	19998	0.0029010	-1.2059950	58.0080083
geo_level_1_id	2	10000	7.1356	6.2255673	6.0	6.262000	5.9304	0	30	30	1.1749932	0.9525824	0.0622557
geo_level_2_id	3	10000	296.9303	279.3906512	219.0	258.789500	268.3506	0	1411	1411	1.0625580	0.5759075	2.7939065
geo_level_3_id	4	10000	2678.6179	2520.6637691	1937.5	2324.264750	2283.9453	0	12151	12151	1.1089550	0.6635190	25.2066377
count_floors_pre_eq	5	10000	2.1467	0.7363649	2.0	2.129375	0.0000	1	9	8	0.8558617	2.6972923	0.0073636
age	6	10000	25.3935	64.4828928	15.0	18.685000	14.8260	0	995	995	13.6373213	201.5544702	0.6448289
area	7	10000	38.4381	21.2658833	34.0	35.509000	13.3434	6	425	419	3.8652558	37.9451244	0.2126588
height	8	10000	4.6531	1.7928418	5.0	4.542375	1.4826	1	30	29	1.4191292	8.2509723	0.0179284
land_surface_condition*	9	10000	2.7964	0.4812174	3.0	2.913875	0.0000	1	3	2	-2.3505165	4.7720630	0.0048122
foundation_type*	10	10000	1.3333	0.8560867	1.0	1.081250	0.0000	1	5	4	2.4907634	4.8993146	0.0085609
roof_type*	11	10000	2.1835	0.5154166	2.0	2.176750	0.0000	1	3	2	0.2270849	0.1411536	0.0051542
ground_floor_type*	12	10000	2.0210	0.4668841	2.0	2.000875	0.0000	1	5	4	0.9320789	6.8210824	0.0046688
other_floor_type*	13	10000	3.3005	1.0912923	4.0	3.500625	0.0000	1	4	3	-1.3063160	0.1472132	0.0109129
position*	14	10000	2.3615	0.7596548	2.0	2.215375	0.0000	1	4	3	1.5107564	0.6878868	0.0075965
plan_configuration*	15	10000	6.9381	0.3694348	7.0	7.000000	0.0000	1	9	8	-6.0633343	53.1889752	0.0036943
has_superstructure_adobe_mud	16	10000	0.0897	0.2857658	0.0	0.000000	0.0000	0	1	1	2.8712954	6.2449617	0.0028577
has_superstructure_mud_mortar_stone	17	10000	0.7626	0.4255107	1.0	0.828250	0.0000	0	1	1	-1.2341581	-0.4769013	0.0042551
has_superstructure_stone_flag	18	10000	0.0299	0.1703200	0.0	0.000000	0.0000	0	1	1	5.5196464	28.4693428	0.0017032
has_superstructure_cement_mortar_stone	19	10000	0.0190	0.1365315	0.0	0.000000	0.0000	0	1	1	7.0452859	47.6408173	0.0013653
has_superstructure_mud_mortar_brick	20	10000	0.0688	0.2531264	0.0	0.000000	0.0000	0	1	1	3.4066530	9.6062453	0.0025313
has_superstructure_cement_mortar_brick	21	10000	0.0725	0.2593270	0.0	0.000000	0.0000	0	1	1	3.2966664	8.8688964	0.0025933
has_superstructure_timber	22	10000	0.2561	0.4364995	0.0	0.195125	0.0000	0	1	1	1.1174152	-0.7514584	0.0043650
has_superstructure_bamboo	23	10000	0.0877	0.2828723	0.0	0.000000	0.0000	0	1	1	2.9148053	6.4967396	0.0028287
has_superstructure_rc_non_engineered	24	10000	0.0400	0.1959690	0.0	0.000000	0.0000	0	1	1	4.6941511	20.0370586	0.0019597
has_superstructure_rc_engineered	25	10000	0.0138	0.1166659	0.0	0.000000	0.0000	0	1	1	8.3340818	67.4636664	0.0011667
has_superstructure_other	26	10000	0.0141	0.1179092	0.0	0.000000	0.0000	0	1	1	8.2411109	65.9225009	0.0011791
legal_ownership_status*	27	10000	2.9957	0.2582405	3.0	3.000000	0.0000	1	4	3	-3.6140521	38.1983754	0.0025824
count_families	28	10000	0.9846	0.4232975	1.0	1.000000	0.0000	0	7	7	1.7967935	18.9505864	0.0042330
has_secondary_use	29	10000	0.1086	0.3111522	0.0	0.010750	0.0000	0	1	1	2.5155586	4.3284680	0.0031115
has_secondary_use_agriculture	30	10000	0.0673	0.2505534	0.0	0.000000	0.0000	0	1	1	3.4536094	9.9284110	0.0025055
has_secondary_use_hotel	31	10000	0.0294	0.1689334	0.0	0.000000	0.0000	0	1	1	5.5708692	29.0374875	0.0016893
has_secondary_use_rental	32	10000	0.0064	0.0797476	0.0	0.000000	0.0000	0	1	1	12.3778217	151.2255915	0.0007975
has_secondary_use_institution	33	10000	0.0007	0.0264496	0.0	0.000000	0.0000	0	1	1	37.7510861	1423.2868289	0.0002645
has_secondary_use_school	34	10000	0.0007	0.0264496	0.0	0.000000	0.0000	0	1	1	37.7510861	1423.2868289	0.0002645
has_secondary_use_industry	35	10000	0.0008	0.0282744	0.0	0.000000	0.0000	0	1	1	35.3076017	1244.7512130	0.0002827
has_secondary_use_health_post	36	10000	0.0002	0.0141414	0.0	0.000000	0.0000	0	1	1	70.6788600	4994.0006500	0.0001414
has_secondary_use_gov_office	37	10000	0.0002	0.0141414	0.0	0.000000	0.0000	0	1	1	70.6788600	4994.0006500	0.0001414
has_secondary_use_use_police	38	10000	0.0001	0.0100000	0.0	0.000000	0.0000	0	1	1	99.9700020	9993.0006000	0.0001000
has_secondary_use_other	39	10000	0.0053	0.0726116	0.0	0.000000	0.0000	0	1	1	13.6245688	183.6472385	0.0007261
damage_grade	40	10000	2.2488	0.6119935	2.0	2.303250	0.0000	1	3	2	-0.2012394	-0.5806365	0.0061199

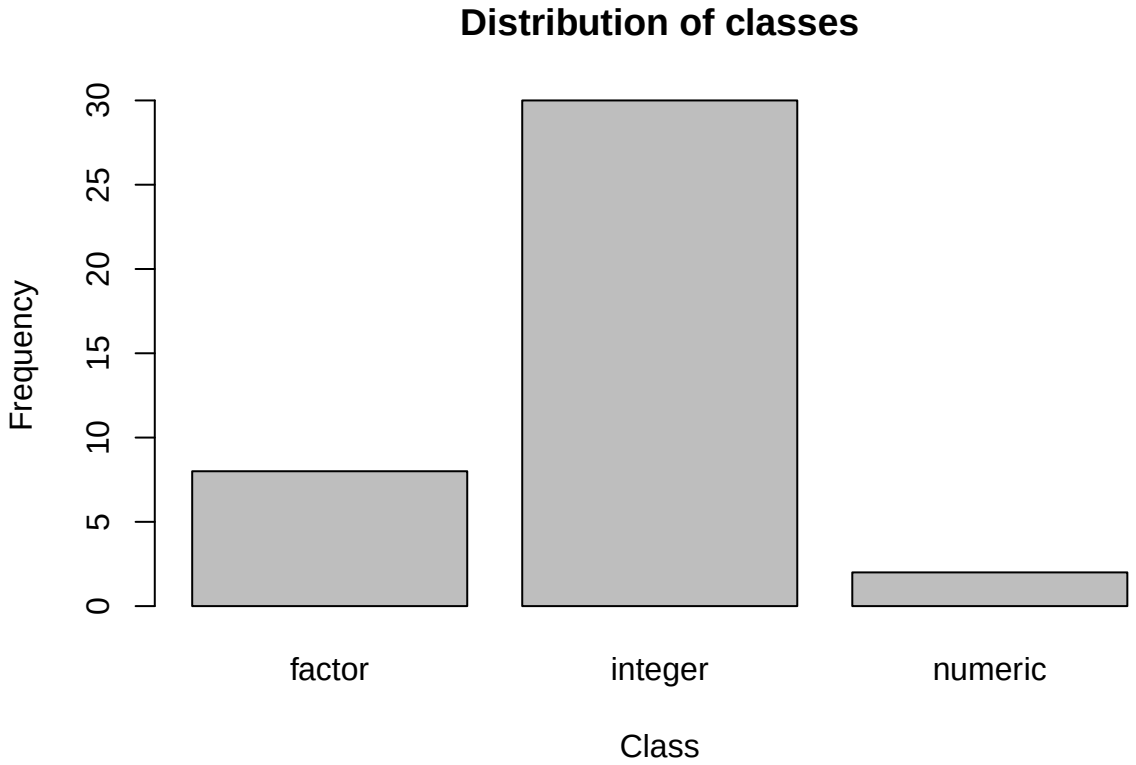
2.2 Damage grade



It is clearly visible that the most frequent damage_level is “2”, followed by damage_level “3” and then “1”.

2.3 Class distribution

First of all lets have a look at the distribution of the classes in the dataset:



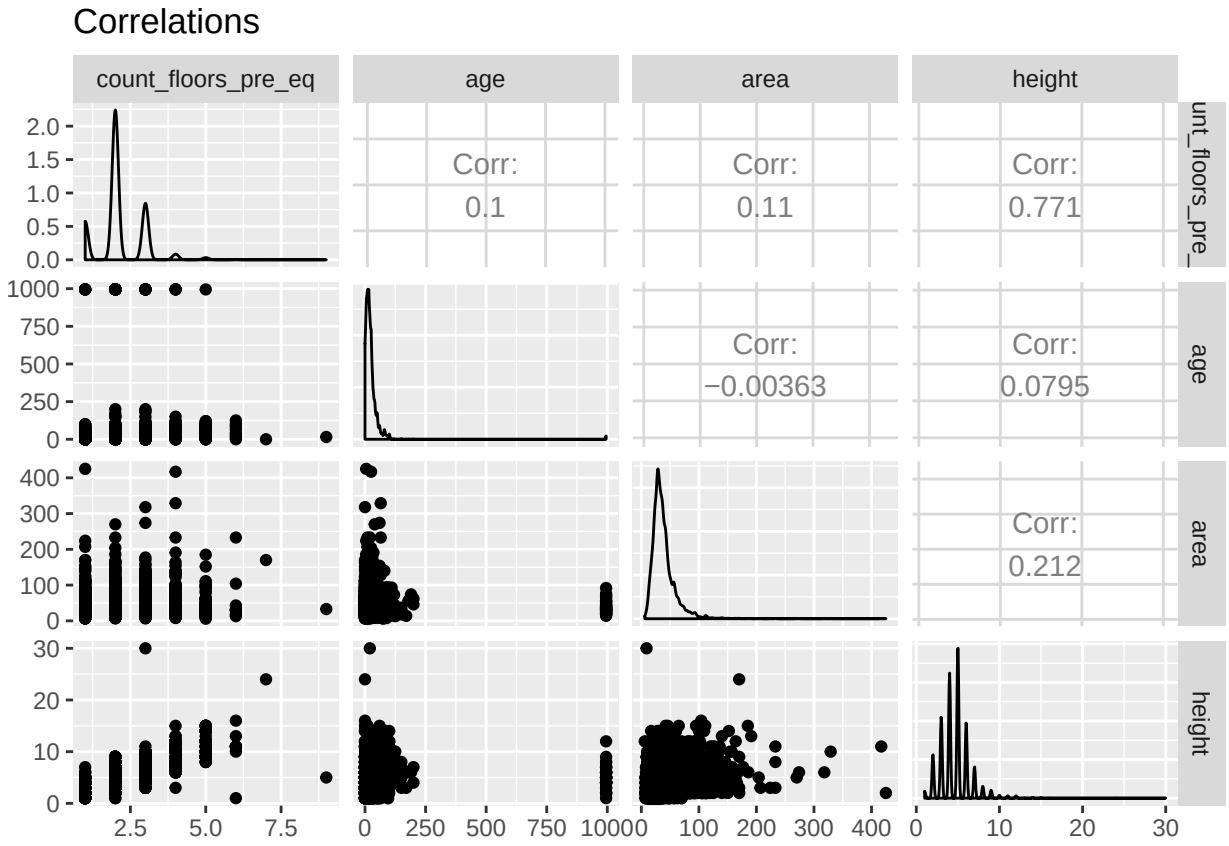
As we can see there are - 8 variables with the class **factor** - 30 variables with the class **integer** - 2 variables with the class **numeric**,

2.4 Correlations

First of all we have to remove the correlation-irrelevant variables like **building_id**. Then in a next step we will check the correlations between the numeric variables in the dataset. In a third step we try to correlate the non-numeric factor variables in the dataset.

2.5 Numeric scatterplot matrix

Here we visualize the correlation of the variables **age**, **area**, **height** and **count of floors**:



3 Methods

To find the damage grade based of features of the given dataset is seen as a classification problem. Befor starting creating models we had to standadize the numerical features of the dataset to not overweight some of them.

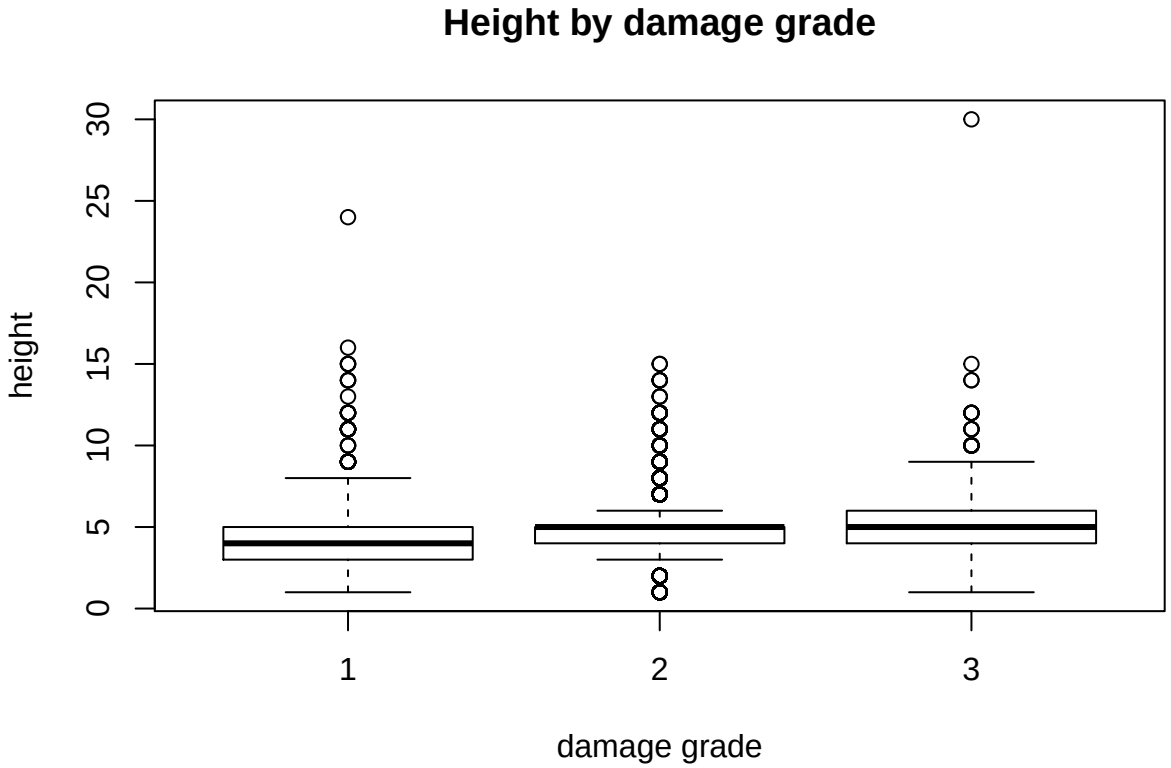
The first approach doing a classification was using a simple **random forest**. Based on the created model we did a prediction with the test-dataset and also plotted the variable importance using VarImpPlot. The result was good but i wanted to do even better. So i decided to use the **CARET package** to finetune the parameters with **crossvalidation**.

Trying another approach i decided to go with a **Support Vector Machine**. Based on some tests it turned out, that using SVM there is a real improvement in RMSE as well as Accuracy of predicted classes.

4 Results

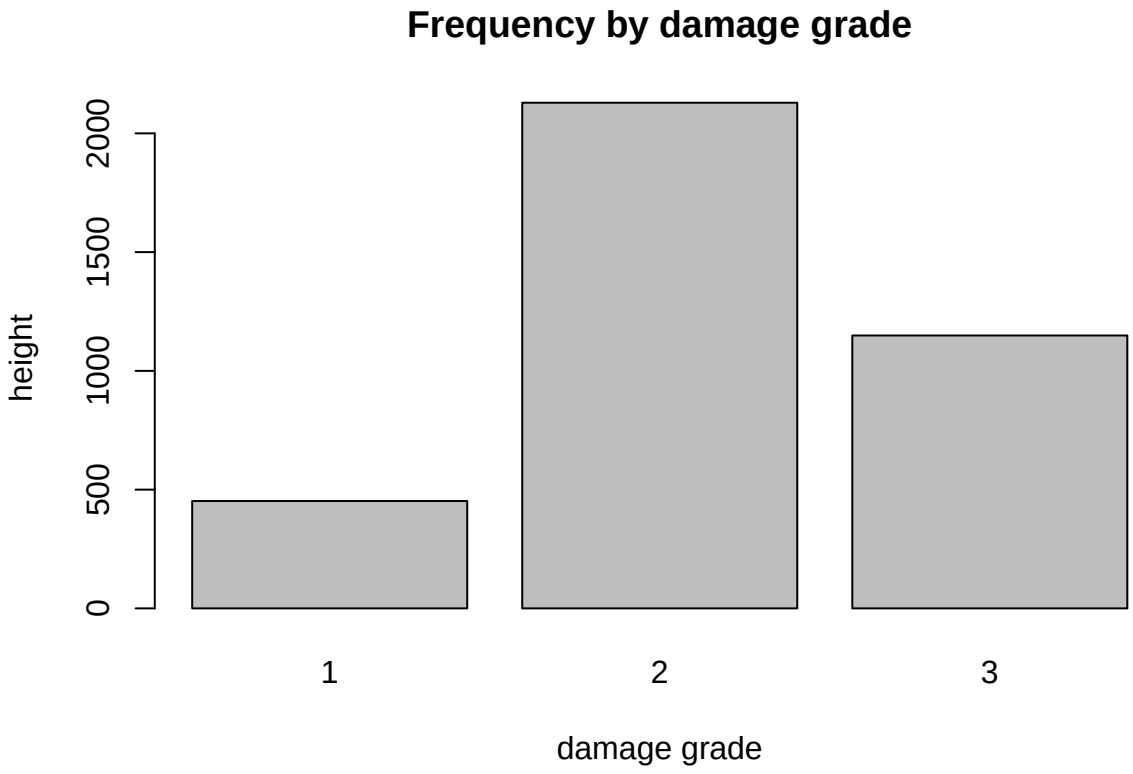
4.1 height by damage category

As visible in the following boxplot only the height of buildings classified with **damage_grade 1** are under the **mean height** of 4.6531:



##

height by area As visible in the following boxplot only the area of buildings classified with **damage_grade 2** are under the **mean height** of 38.4381:



4.2 Prediction using Random Forest

Using Random forest we get an **accuracy** of around **0.37** which is okay cause of the multiclass-classification problem of predictin 3 classes. Simply guessing 3 damage grades would give an accuracy of 0.125.

Also the calculated Root Mean Squared Error (**RMSE**) is **0.94** is under 1 and therefore okay but could be optimized for sure.

4.3 Prediction using SVM

Using Support Vector Machines improved the results as follows:

RMSE: 0.66 Accuracy: 0.56

5 Conclusion

The analysis also clearly shows that there are reliable patterns in the correlation of height, area and damage grade of buildings after an earthquake. Using this data we are able to predict the grade of damage a building will have using the area and the height of a building.

There are good results in predicting the **damage-grade** using Random Forest and even better using SVM. For sure there could be reached even better results by finetuning the used algorithms or using others.