Analysis of Earthquake Damage in Nepal

Joe Song, April 2018

Executive Summary

This document presents an analysis of data concerning the 2015 Nepal earthquake and its damage. The analysis is based on 10,000 observations of data collected through surveys by the Central Bureau of Statistics that work under the National Planning Commission Secretariat of Nepal, containing valuable information on earthquake impacts, household conditions, and socio-economic-demographic statistics.

We're trying to predict the ordinal variable 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

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the data, several potential relationships between building characteristics and damage were identified. After exploring the data, a predictive model to classify damage into three categories was created.

After performing the analysis, the author presents the following conclusions: While many factors can help indicate the damage of a building, significant features found in this analysis were:

- `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: numerical): number of floors in the building before the earthquake.
- `age` (type: numerical): age of the building in years.
- `area` (type: numerical): plinth area of the building in m^2.
- `height` (type: numerical): height of the building in m.
- `land_surface_condition` (type: categorical): surface condition of the land where the building was built.
- `foundation_type` (type: categorical): type of foundation used while building.

- `roof_type` (type: categorical): type of roof used while building.
- `ground_floor_type` (type: categorical): type of the ground floor.
- `other_floor_type` (type: categorical): type of constructions used in higher than the ground floors (except of roof).
- `position` (type: categorical): position of the building.
- `count_families` (type: numerical): 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.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics

Summary statistics for mean, standard deviation, minimum, first quartile, median, maximum, third quartile, and maximum were calculated for numeric columns, and the results taken from 10,000 observations are shown here:

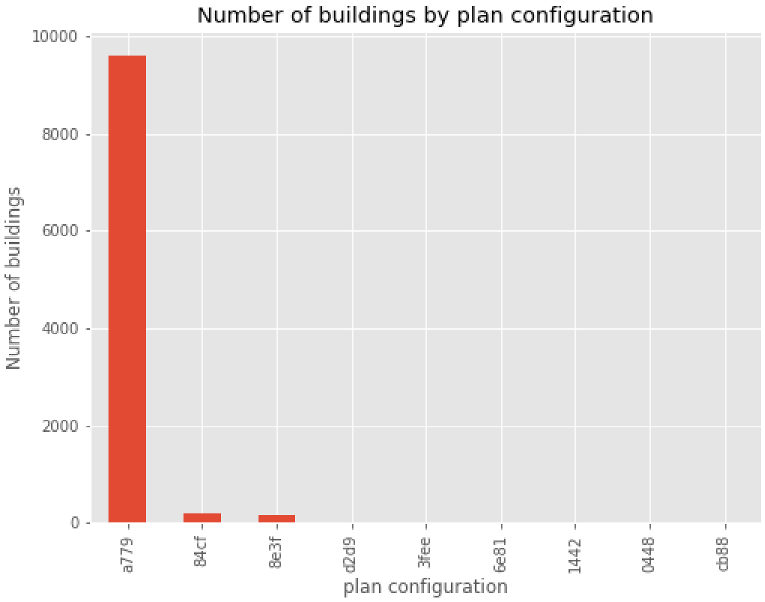
Column	mean	std	min	25%	50%	75%	max
building_id	9987.16	5800.800829	1	4998.75	9963.5	15044.75	19999
geo_level_1_id	7.1356	6.225567	0	2	6	10	30
geo_level_2_id	296.9303	279.390651	0	60	219	466	1411
geo_level_3_id	2678.6179	2520.663769	0	606.75	1937.5	4158	12151
count_floors_pre_eq	2.1467	0.736365	1	2	2	3	9
age	25.3935	64.482893	0	10	15	30	995
area	38.4381	21.265883	6	26	34	44	425
height	4.6531	1.792842	1	4	5	5	30
count_families	0.9846	0.423297	0	1	1	1	7

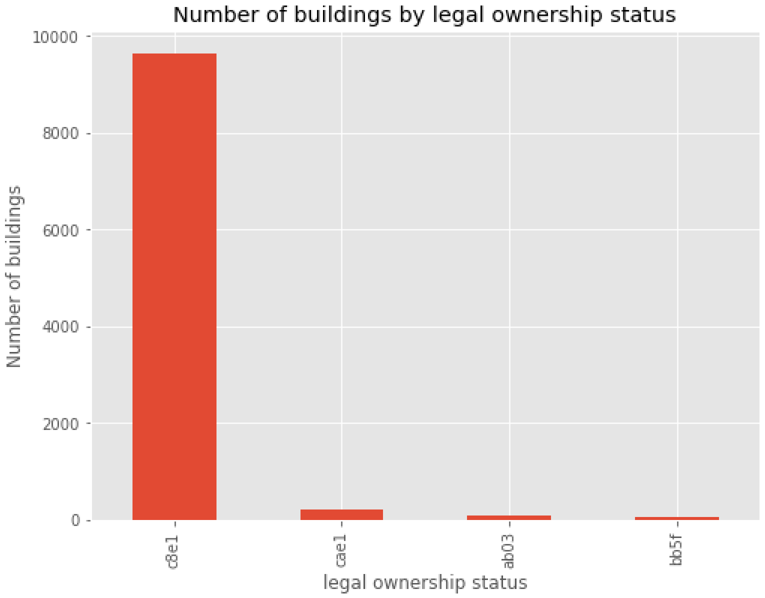
In addition to the numeric values, there are observations include categorical features, including **land_surface_condition**, **foundation_type, roof_type**, **ground_floor_type**, **position**, **plan_configuration and legal_ownership_status**.

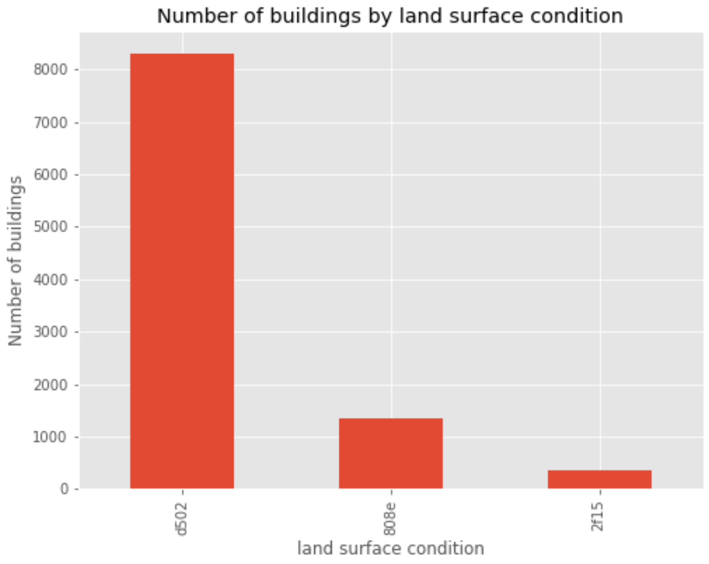
Bar charts were created to show frequency of these features, and indicate the following:

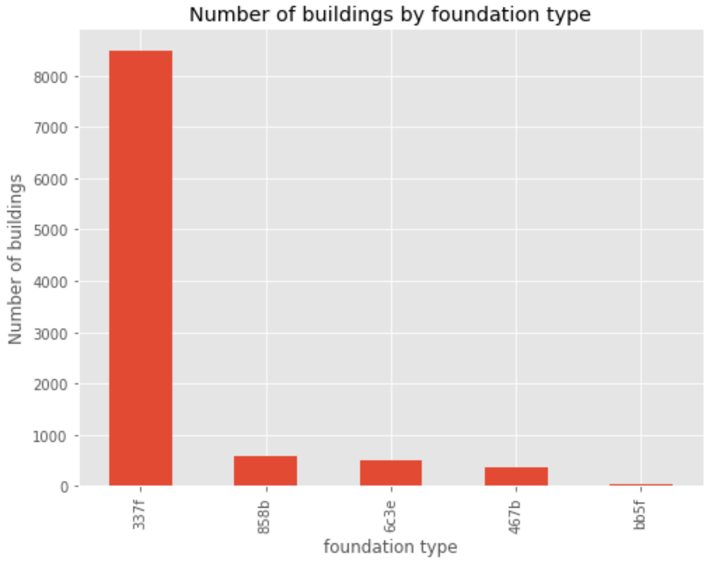
- For every category feature, there is a dominant type that accounts for the most part of each feature.

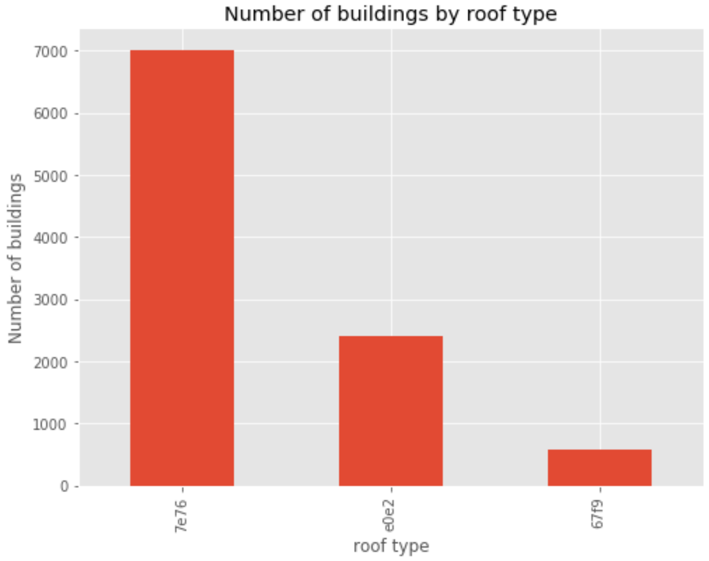
- For **plan_configuration and legal_ownership_status** features, almost every building has the same plan configuration (a779) and legal ownership status (c8e1), so for simplicity, we can get assume all buildings have this kind of features and get rid of these two columns.

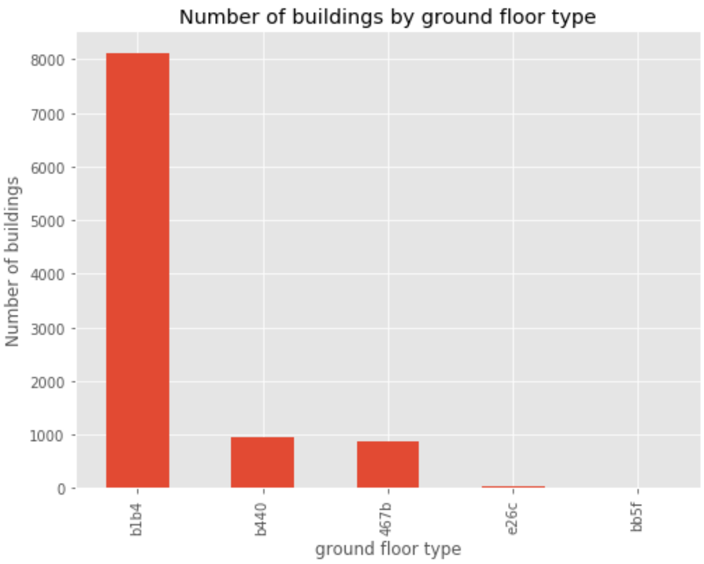


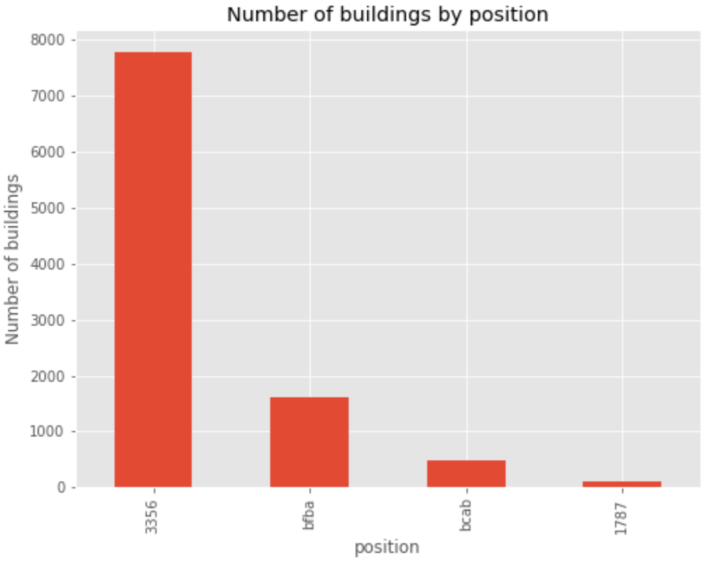



Number of buildings by land surface condition
Number of buildings
0
1000
2000
3000
4000
5000
6000
7000
8000
d502
808e
2f15
land surface condition

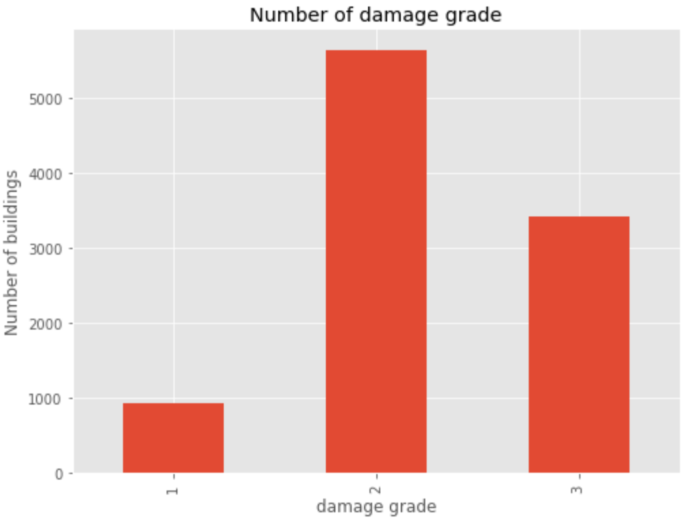

Number of buildings by foundation type
Number of buildings
0
1000
2000
3000
4000
5000
6000
7000
8000
337f
858b
6c3e
467b
bb5f
foundation type


Number of buildings by roof type
Number of buildings
0
1000
2000
3000
4000
5000
6000
7000
7e76
e0e2
67f9
roof type





Since **damage grade** is of interest in this analysis, a bar chart of the **damage grade** column shows that the most buildings are suffered the second-grade damage, as shown here:

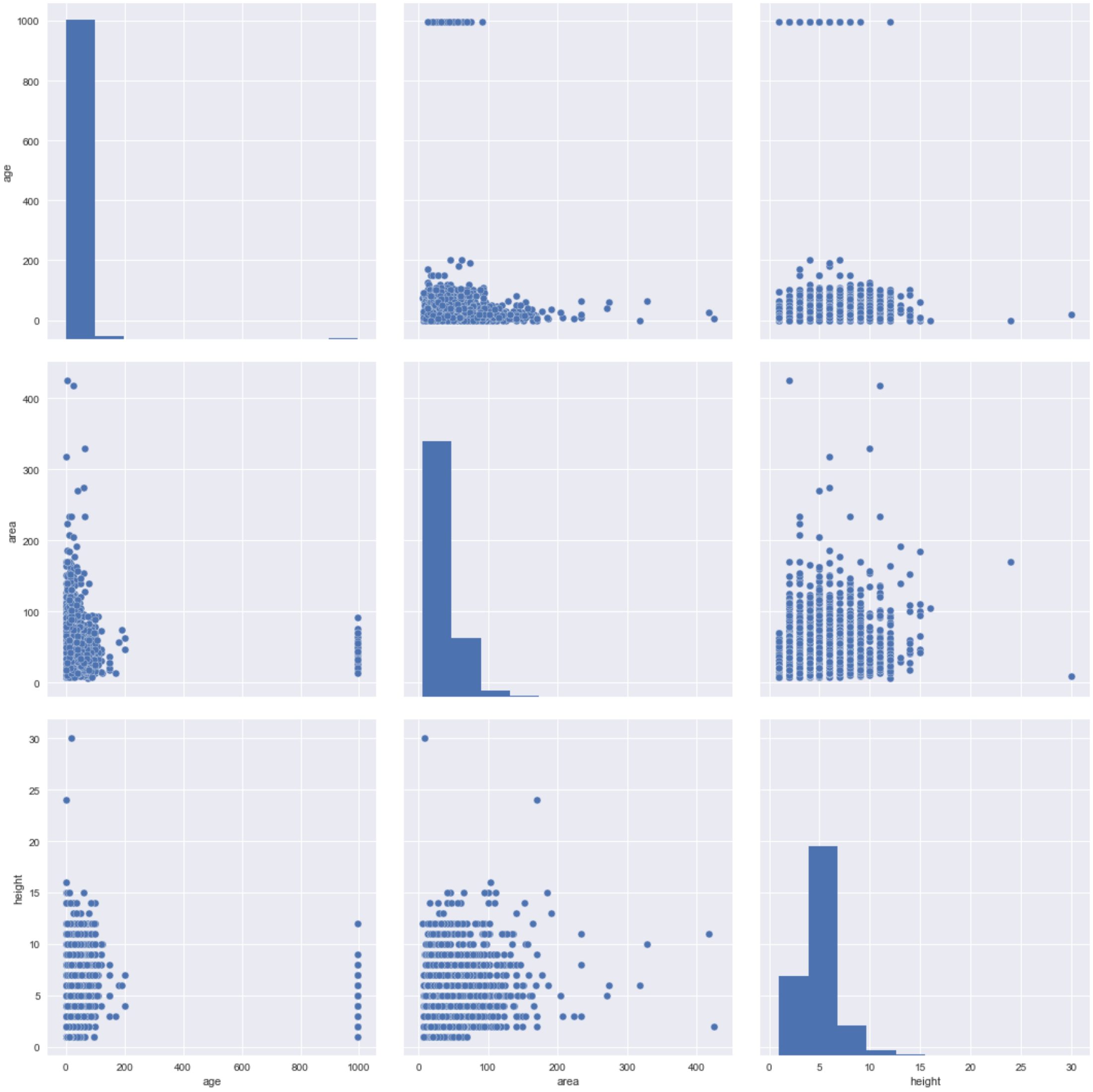


Correlation and Apparent Relationships

After exploring the individual features, an attempt was made to identify relationships between features in the data – in particular, between **damage grade** and the other features.

Scatter Plot

The following scatter-plot matrix was generated initially to compare numeric features (age, area, height) with one another. The key features in this matrix are shown here:

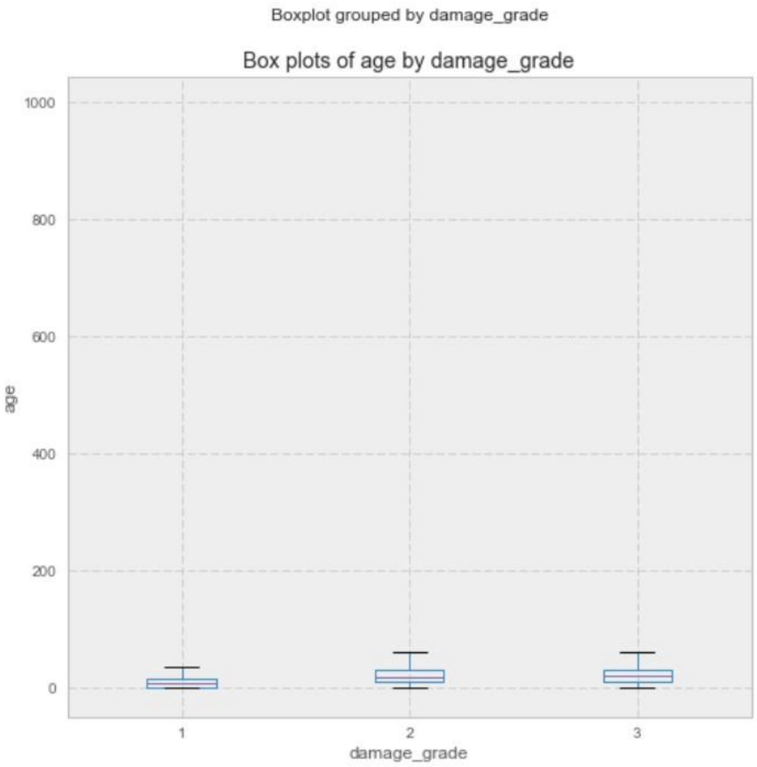


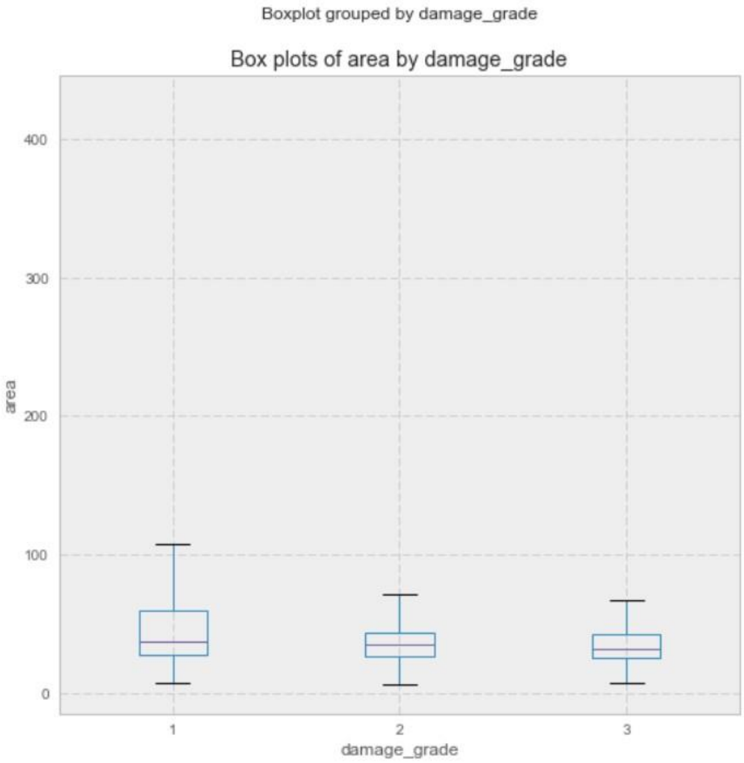
It doesn't show an apparent relationship between each other. The correlation between the numeric columns was then calculated with the following results:

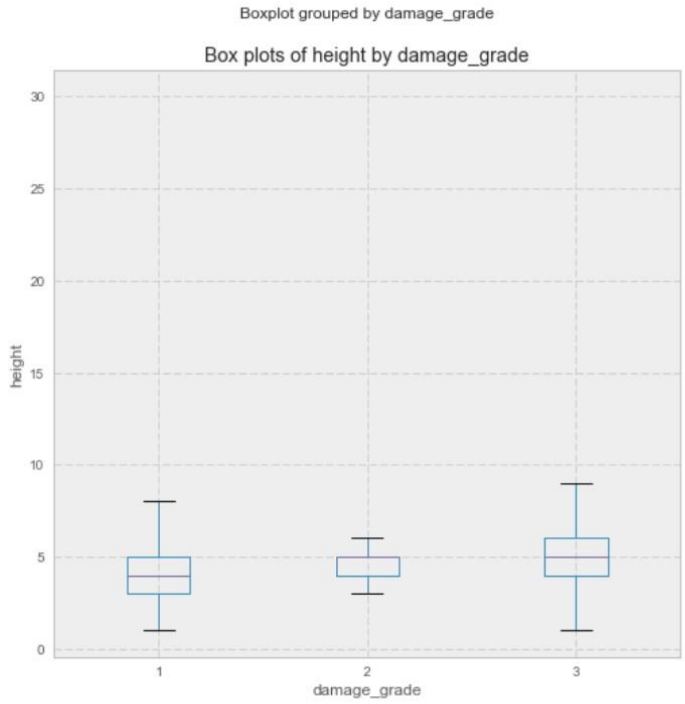
	age	area	height
age	1.000000	-0.003627	0.079549
area	-0.003627	1.000000	0.211601
height	0.079549	0.211601	1.000000

Box plot

Having explored the relationship between numeric features, an attempt was made to discern any apparent relationship between these features and damage grade. The following box-plots show the numeric columns grouped by damage grade:







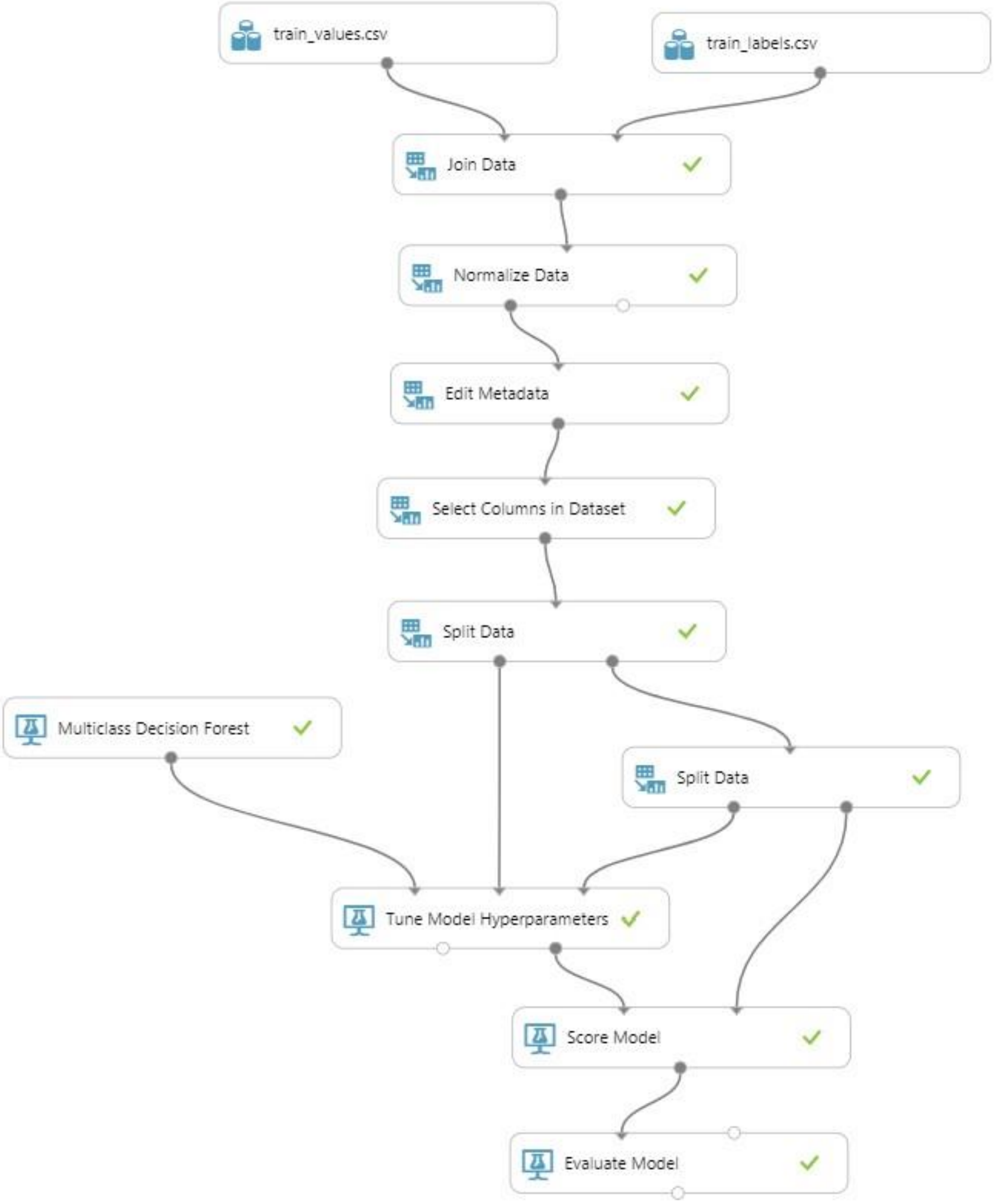
The box plots show some clear differences in terms of the median and range of age, area and height of buildings for different damage grade. For example:

- 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 3 is higher than for buildings with a damage grade of 2.
- 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.

Classification of Damage Grade

Based on the analysis of the data, a predictive model to classify damage grade into three grade categories: 1 represents low damage, 2 represents a medium amount of damage and 3 represents almost complete destruction.

The model was created using the Multiclass Decision Forest algorithm by using Microsoft Azure Machine Learning tool. The process is shown below:

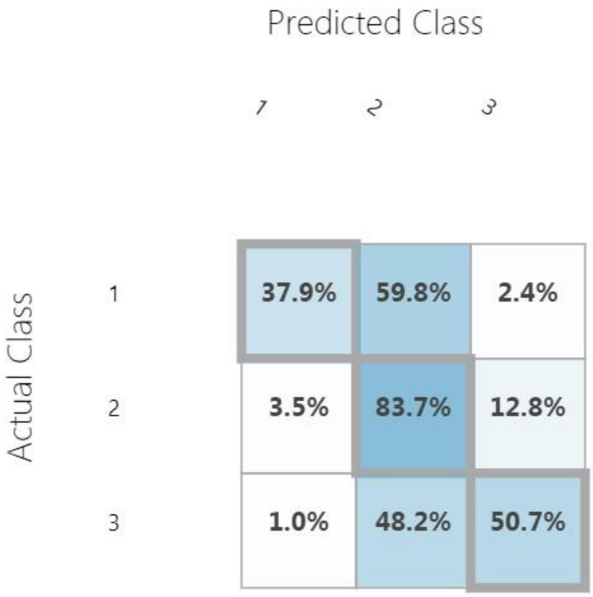


The main processes are here:

- Join Data: Join the features and label together.
- Normalize Data: Use Z-score to transform the numeric columns.
- Edit Metadata: Make categorical columns.
- Select Columns: Exclude non-important features.
- Split Data: Split train data to 60% and 40% for training and testing, respectively.
- Train Model: Use Multiclass Decision Forest algorithm as the train model.

With 60% of the data trained, testing the model with the remaining 40% of the data yielded the following results:

- Overall accuracy: 0.687
- Average accuracy: 0.791333
- Micro-averaged precision: 0.687
- Macro-averaged precision: 0.652714
- Micro-averaged recall: 0.687
- Macro-averaged recall: 0.574254
- Confusion matrix:



Predicted Class

Actual Class

	1	2	3
1	37.9%	59.8%	2.4%
2	3.5%	83.7%	12.8%
3	1.0%	48.2%	50.7%

Conclusion

Using data on buildings in the affected area and how they were impacted by the earthquake, this analysis can model risk of damage. Accurate models of this kind help first responders plan their initial triage after an earthquake and help governments direct scarce resources which may be available to mitigate risk before another earthquake happens.