Modeling Earthquake Damage

Machine Learning Project
Milestone 1

Presented by:

Omowonuola Molayosi Akintola
Rabina Twayana

Photo Source CircleOfBlue

Université Bretagne Sud
COPERNICUS MASTER IN DIGITAL EARTH
Co-funded by the European Union

Agenda

1. Background
2. Data Overview
3. Data Analysis
4. Data Preparation
5. Data Preprocessing
6. Protocol

1. Background

- Survey conducted after the 2015 Gorkha earthquake in Nepal. Trained enumerators visited the affected areas and collected information on the level of damage, structural characteristics (foundation, roof, building materials), household information, age, and use for over 250,000 buildings.

- The primary goal of the survey was to identify beneficiaries eligible for government assistance for housing reconstruction but the data now serves other purposes including ML tasks.

- The main task is to predict the damage level for buildings affected by the earthquake.

2. Data Overview

Samples

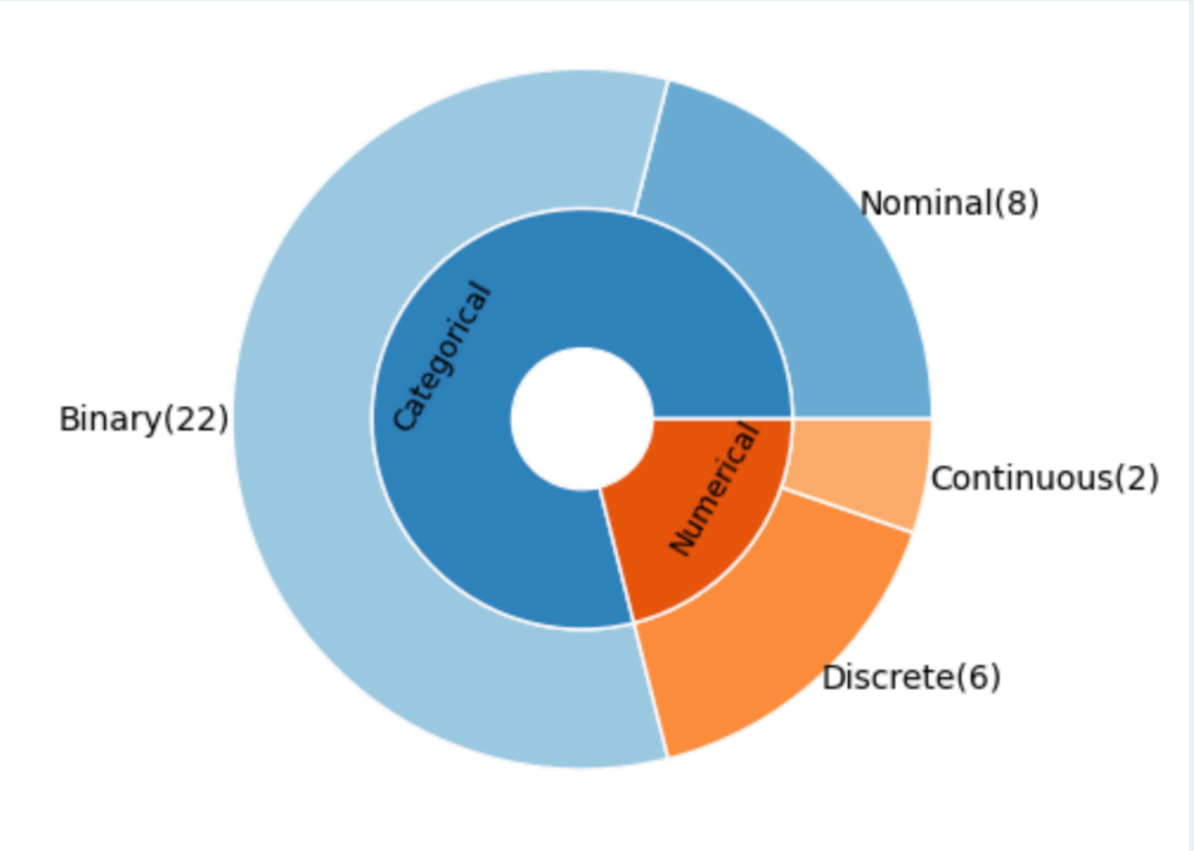
260,600

Features

39

Target

1



The data contains both numerical **(continuous and discrete)** and categorical (**binary and nominal)** variables**.**

Target Variable

The target variable represents the level of damage to the buildings. There are 3 grades

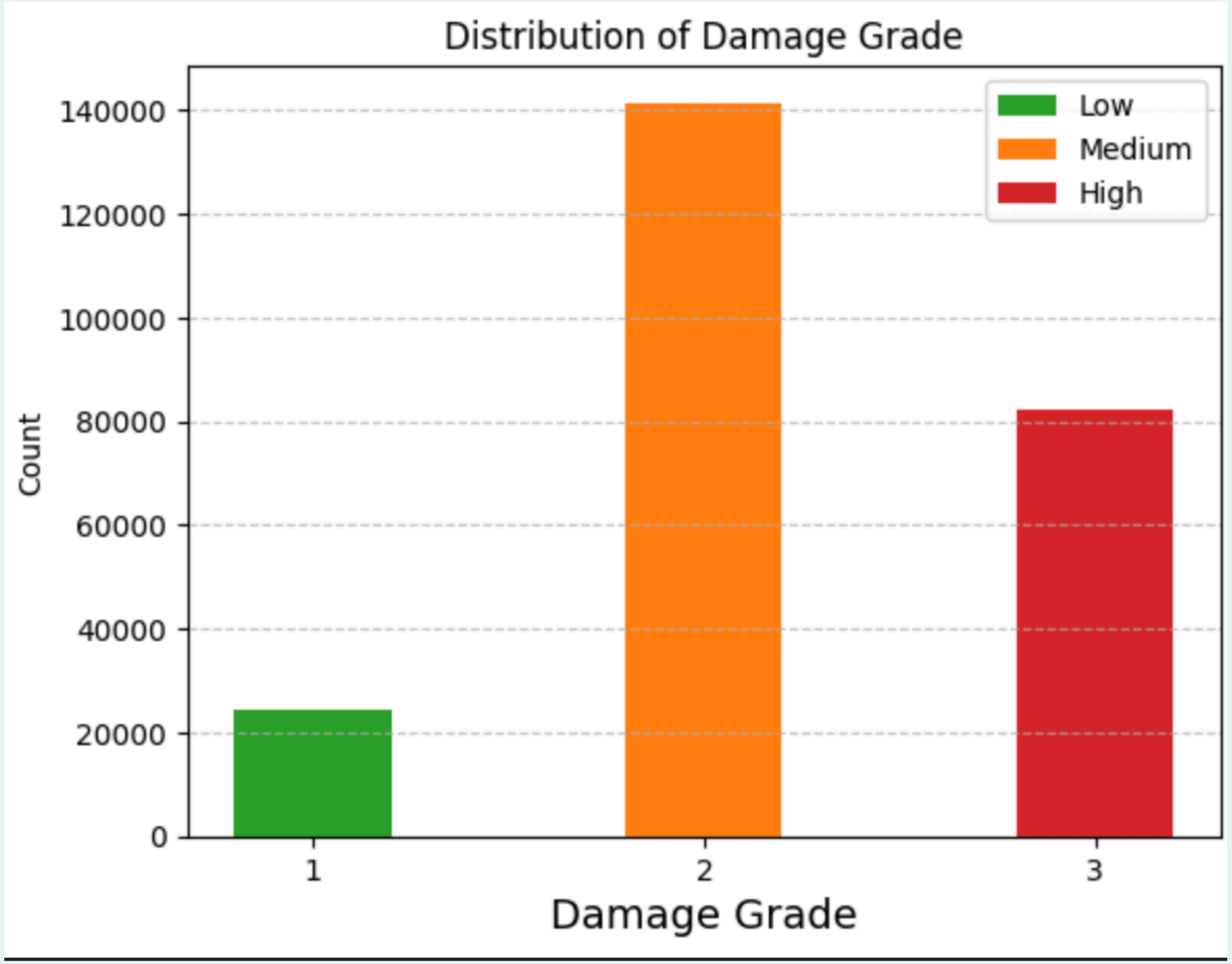
01 Low Damage

02 Medium Damage

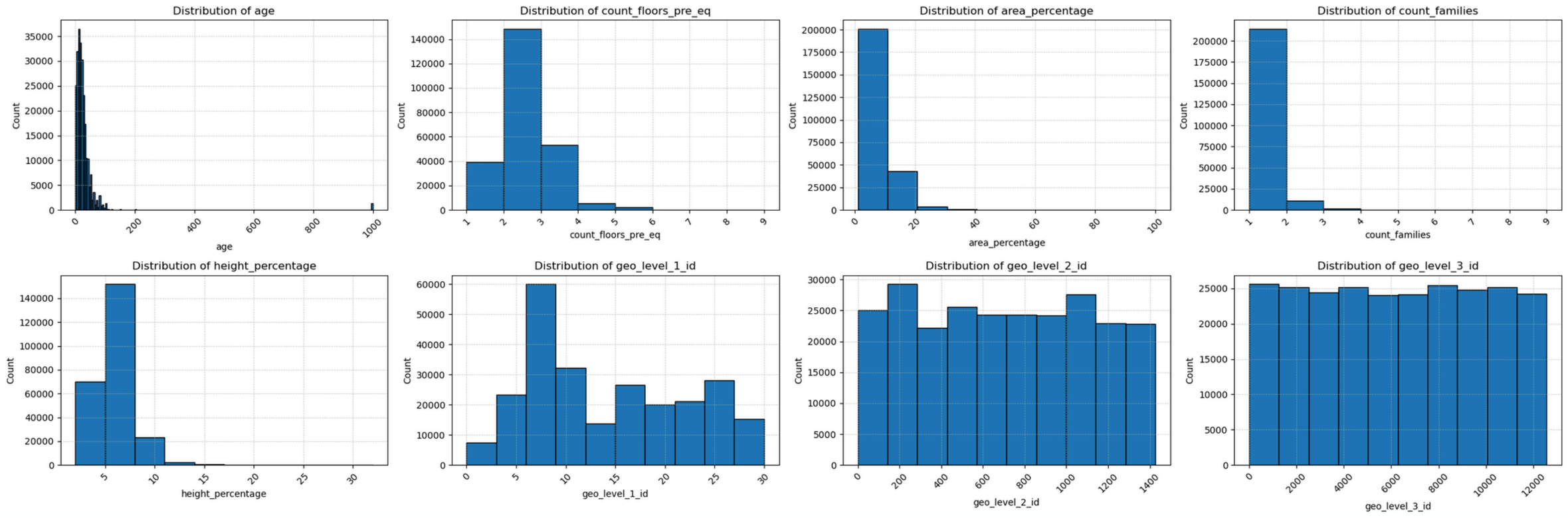
03 Almost Complete Destruction

The ML task for this project is considered a multiclassification task, where one of the three damage grades are assigned to the buildings.

Highly Imbalanced Data !!!

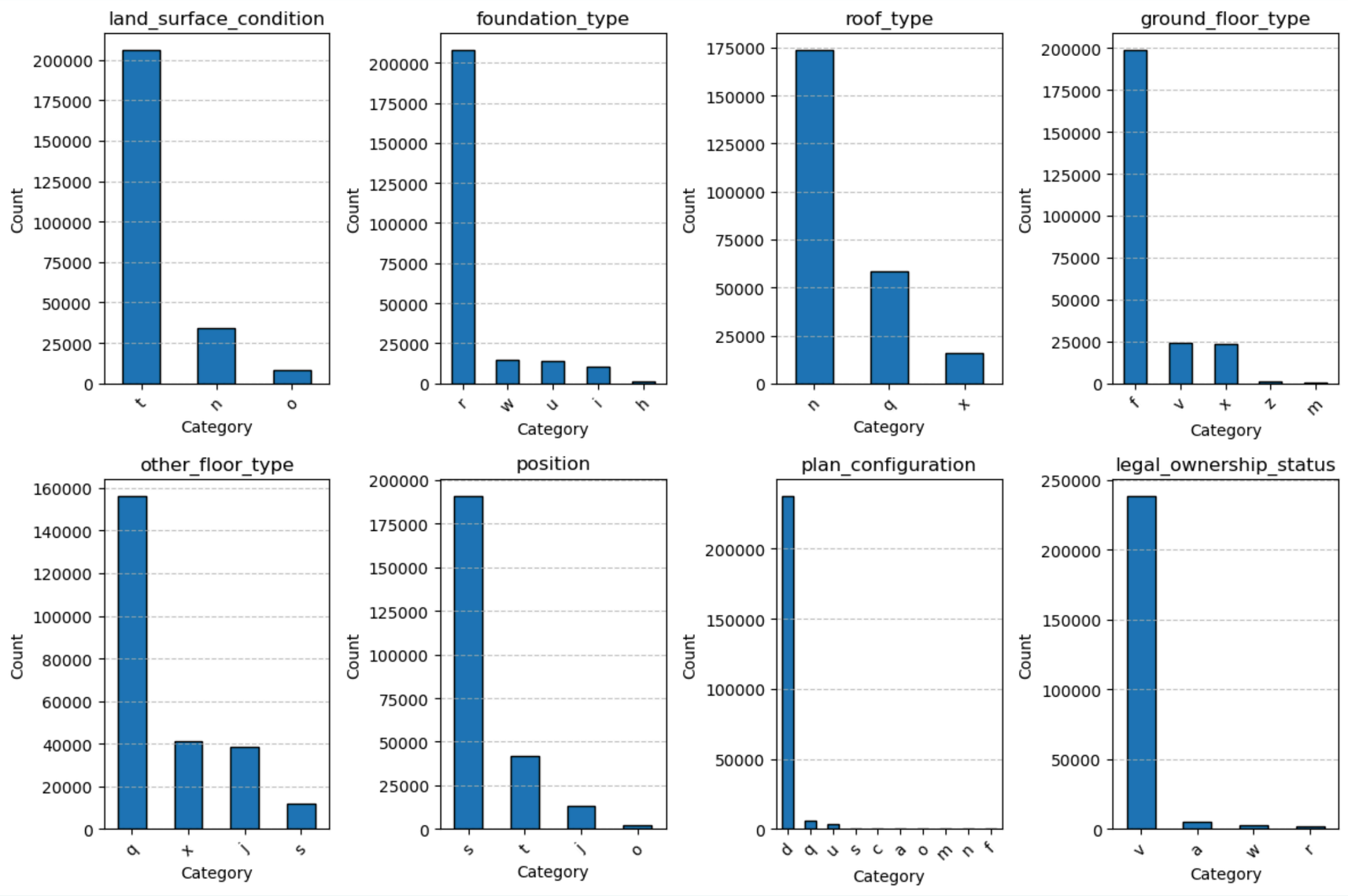


3. Data Analysis: Numerical Features Distribution



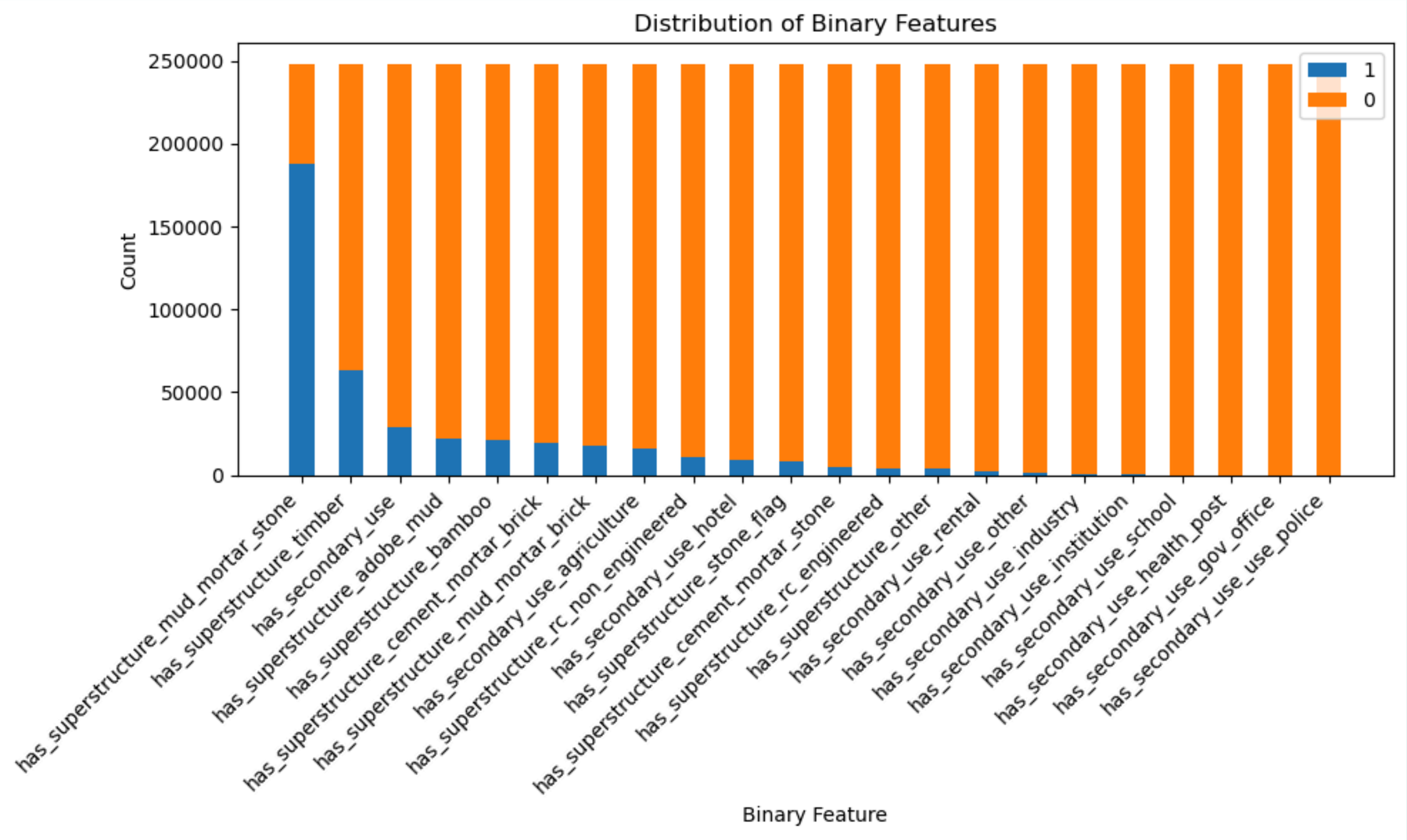
- Most features are right-skewed
- geo_level_1_id have distinct peaks
- geo_level_2_id and geo_level_3_id is more uniform.
- As it is real field data collection, extreme values is most likely to be genuine rather than outliers

3. Data Analysis: Categorical Features Distribution



The categorical variable has been ascii coded so no meaningful interpretation can be given to the values

3. Data Analysis: Binary Features Distribution



- Strong class imbalance
- 0 as dominating class (except 1)
- 8 features have a count of 1s (<1%)

4. Data Preparation

01 **Removal of Data Duplicates**

Found 12319 duplicate rows.

02 **Check for NaN values**

No missing values

5. Data Preprocessing

01 Removing highly correlated Numerical Variables

02 Train_Test Split

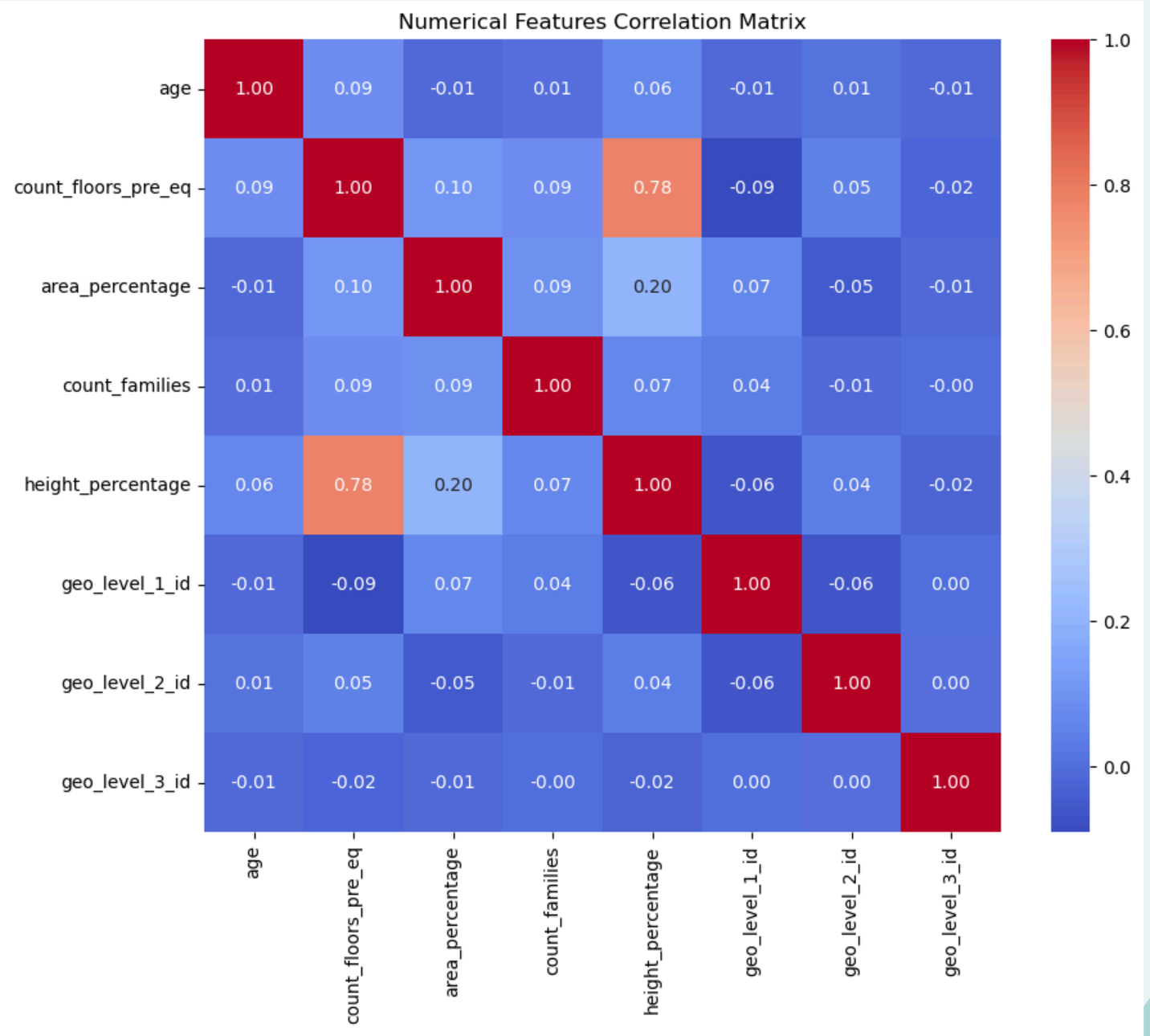
03 Encoding Categorical Variables

04 Dimensionality Reduction with MCA

05 Correlation between Features and Target Variable

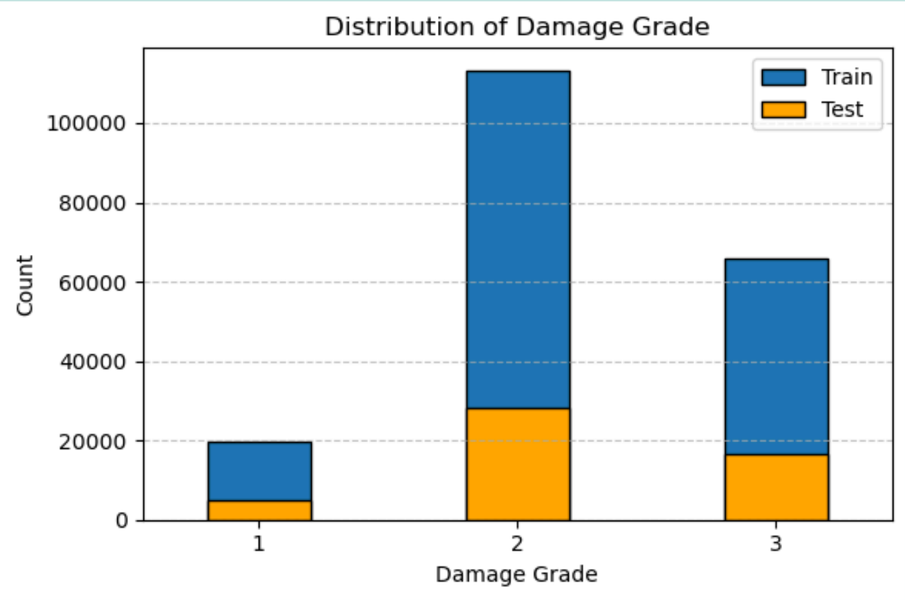
01 Removing Highly Correlateed Numerical Variables

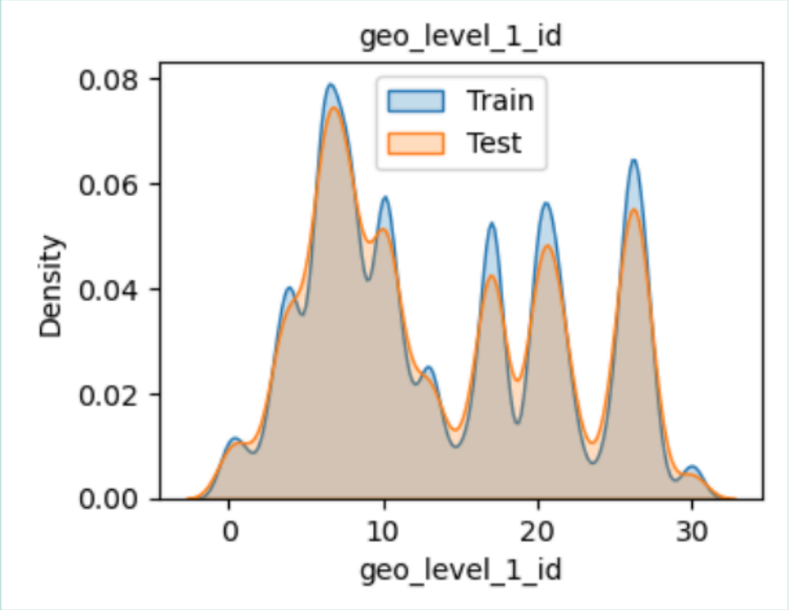
- High correlation between count_floors_pre_eq and height_percentage.
- Consider as redundant and only keep one in the data

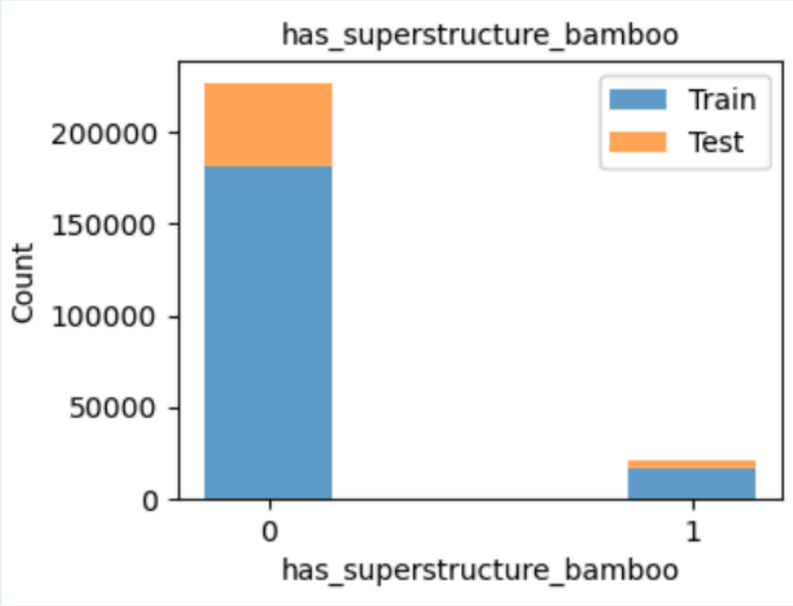


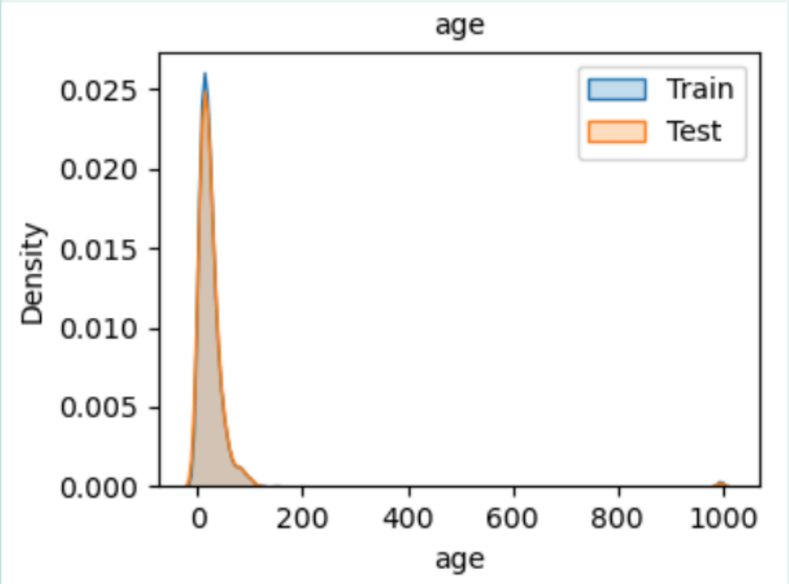
02 Train_Test Split

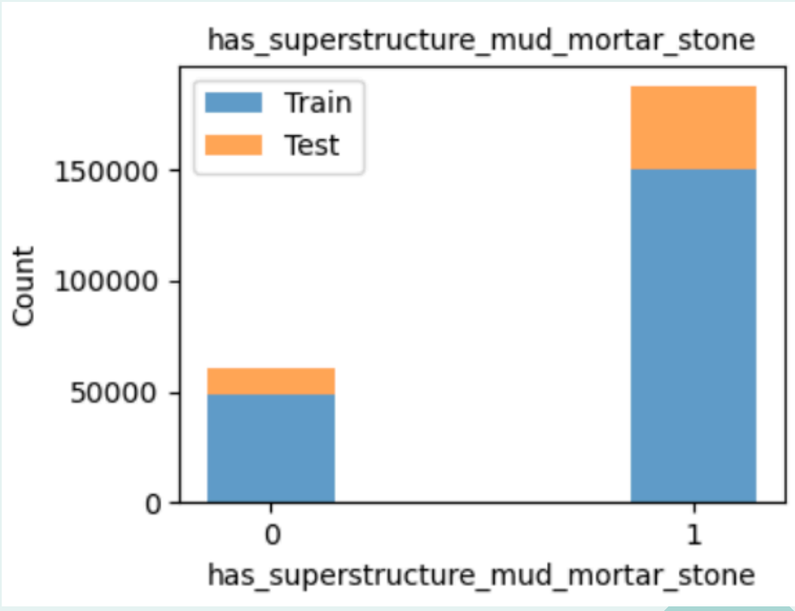
Due to the class imbalance in the dataset, we used stratified sampling to split the data and then checked the distribution of the training and test sets to ensure that all classes were proportionally represented in both splits.







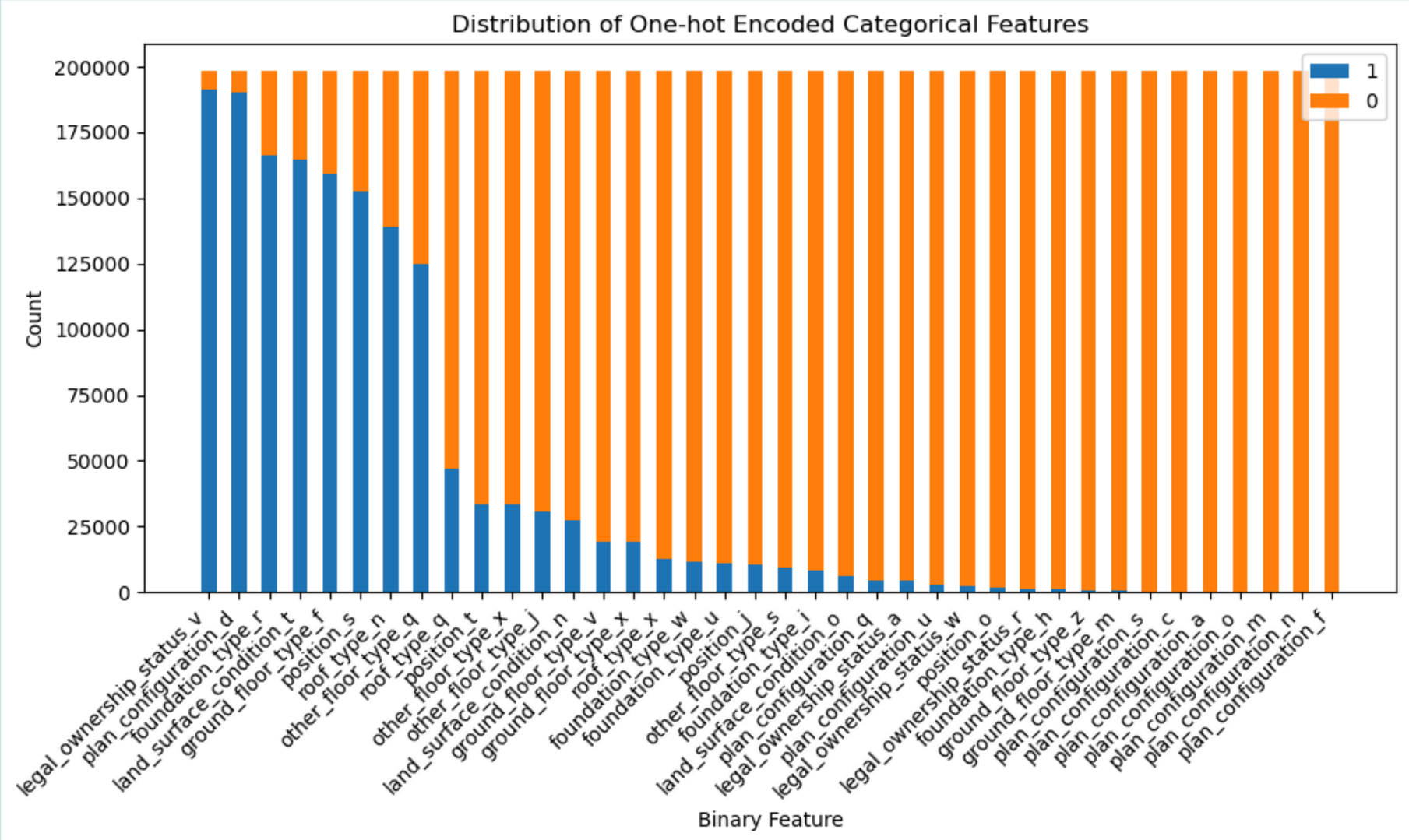




03 Encoding Categorical Variables

Number of categorical Features= 8
After encoding = 38

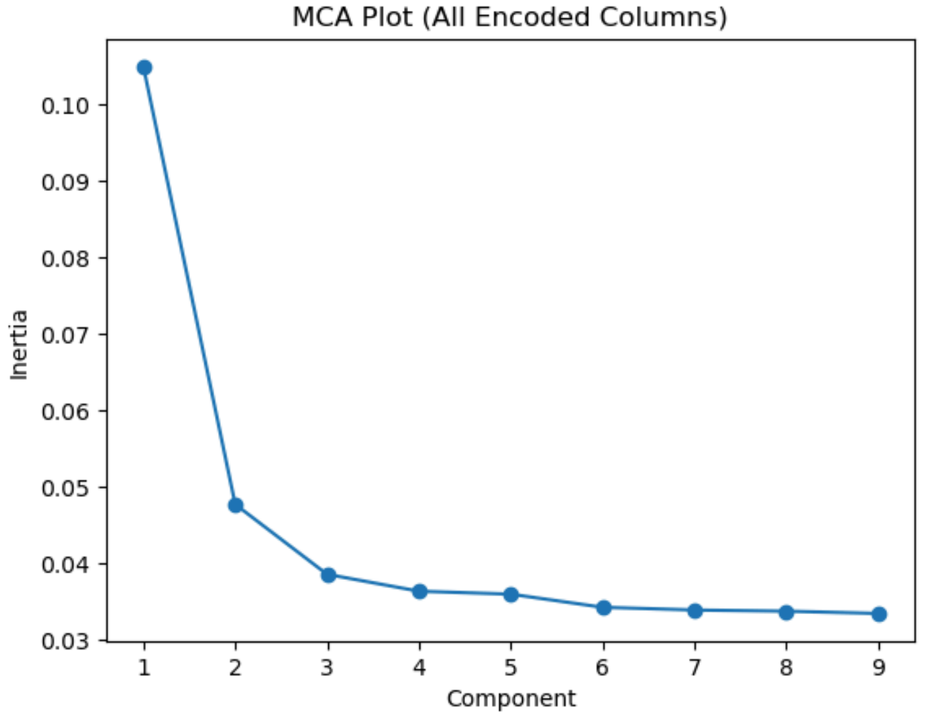
Features less than 1% of 1's sample = 12



04

Dimensionality Reduction using MCA

- 38 Features retained to 9 MCA component

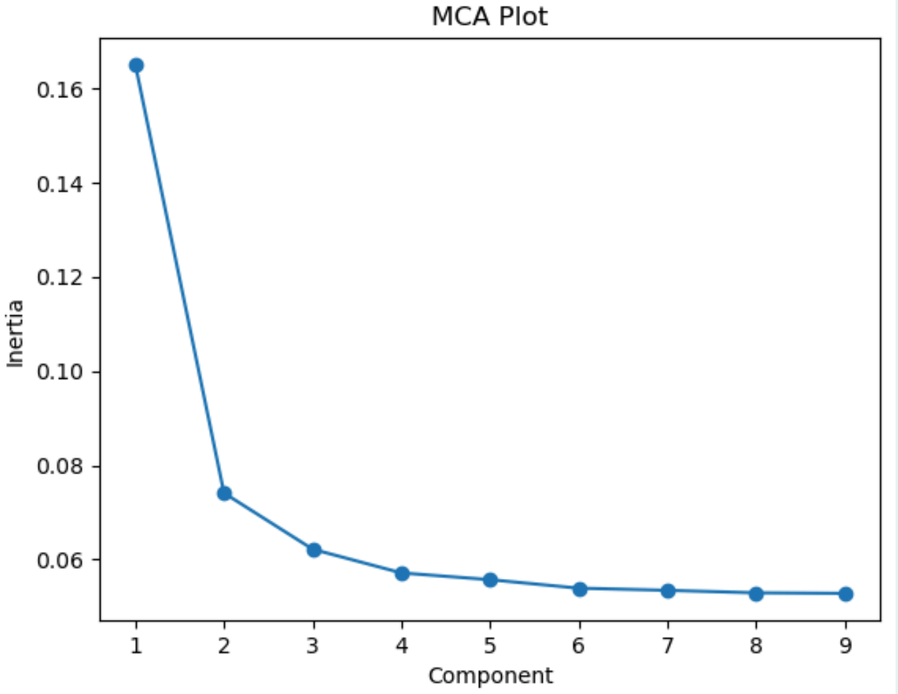


When all 38 encoded features are considered

First 9 MCA components capture 39% of the variance

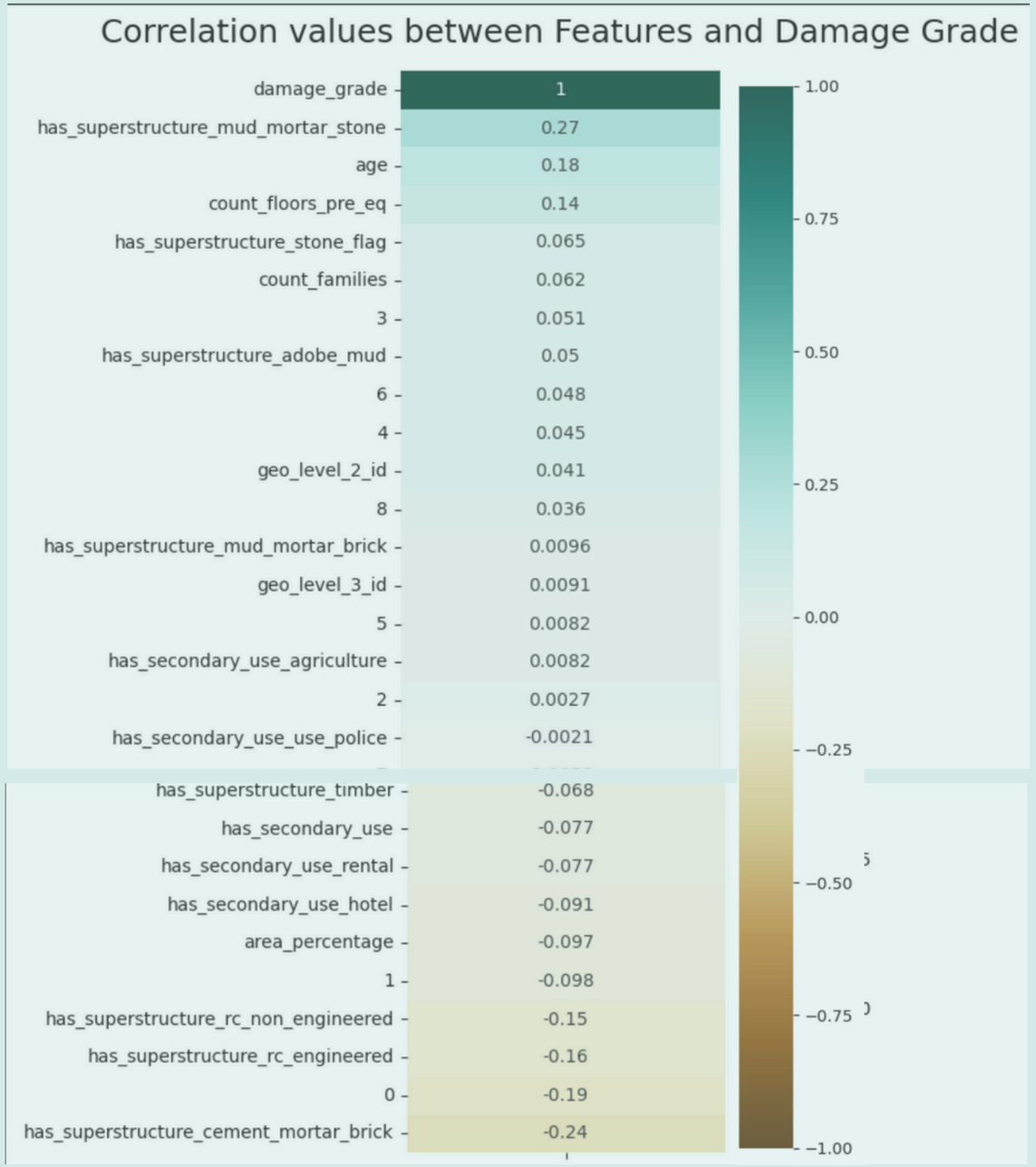
When only 26 features are considered

First 9 MCA components capture 62% of the variance



05 Correlation Between Features and Target

- We computed the correlation mainly to see which features relate most to the target variable.
- This helps decide what decisions to make for model training



Final Features Count

Total= 38
Binary Features =22
MCA components =9
Numeric Features = 7

6. Protocol

01 Data Spliting Strategy and Preventing Data Leakage

- Train and Test ratio = 80:20
- Stratified sampling approach
- Preprocessing steps are performed on the training set after splitting the data to avoid leakage

02 Model Training

- Train two versions of the model: one excluding assumed uncorrelated and irrelevant features from data analysis and another including them, to assess their impact on model performance and evaluation.

03 Hyperparameters Tuning

- Stratified k-fold cross-validation

04 Evaluation metric

- F1-score macro

THANK YOU