earthquake

July 2, 2022

[45]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[18]:
```
values = pd.read_csv('train_values.csv')
labels = pd.read_csv('train_labels.csv')
test_values = pd.read_csv('test_values.csv')
df =values.merge(labels,on='building_id',how='left')
df.head(5)
```

[18]:

	building_id	geo_level_1_id	geo_level_2_id	geo_level_3_id	count_floors_pre_eq	age	area_percentage	height_percentage	land_surface_condition	foundation_type	…	has_secondary_use_hotel	has_secondary_use_rental	has_secondary_use_institution
0	802906	6	487	12198	2	30	6	5	t	r	…	0	0	0
1	28830	8	900	2812	2	10	8	7	o	r	…	0	0	0
2	94947	21	363	8973	2	10	5	5	t	r	…	0	0	0
3	590882	22	418	10694	2	10	6	5	t	r	…	0	0	0
4	201944	11	131	1488	3	30	8	9	t	r	…	0	0	0

```
4                         0                              0

  has_secondary_use_school has_secondary_use_industry  \
0                        0                           0
1                        0                           0
2                        0                           0
3                        0                           0
4                        0                           0

   has_secondary_use_health_post  has_secondary_use_gov_office  \
0                              0                             0
1                              0                             0
2                              0                             0
3                              0                             0
4                              0                             0

   has_secondary_use_use_police  has_secondary_use_other  damage_grade
0                             0                        0             3
1                             0                        0             2
2                             0                        0             3
3                             0                        0             2
4                             0                        0             3

[5 rows x 40 columns]
```

[19]:
```
df.describe()
```

[19]:
```
         building_id  geo_level_1_id  geo_level_2_id  geo_level_3_id  \
count   2.606010e+05   260601.000000   260601.000000   260601.000000
mean    5.256755e+05       13.900353      701.074685     6257.876148
std     3.045450e+05        8.033617      412.710734     3646.369645
min     4.000000e+00        0.000000        0.000000        0.000000
25%     2.611900e+05        7.000000      350.000000     3073.000000
50%     5.257570e+05       12.000000      702.000000     6270.000000
75%     7.897620e+05       21.000000     1050.000000     9412.000000
max     1.052934e+06       30.000000     1427.000000    12567.000000

        count_floors_pre_eq            age  area_percentage  height_percentage  \
count        260601.000000  260601.000000    260601.000000      260601.000000
mean              2.129723      26.535029         8.018051           5.434365
std               0.727665      73.565937         4.392231           1.918418
min               1.000000       0.000000         1.000000           2.000000
25%               2.000000      10.000000         5.000000           4.000000
50%               2.000000      15.000000         7.000000           5.000000
75%               2.000000      30.000000         9.000000           6.000000
max               9.000000     995.000000       100.000000          32.000000
```

```
       has_superstructure_adobe_mud  has_superstructure_mud_mortar_stone  …  \
count                 260601.000000                       260601.000000  …
mean                       0.088645                            0.761935  …
std                        0.284231                            0.425900  …
min                        0.000000                            0.000000  …
25%                        0.000000                            1.000000  …
50%                        0.000000                            1.000000  …
75%                        0.000000                            1.000000  …
max                        1.000000                            1.000000  …

       has_secondary_use_hotel  has_secondary_use_rental  \
count             260601.000000             260601.000000
mean                   0.033626                  0.008101
std                    0.180265                  0.089638
min                    0.000000                  0.000000
25%                    0.000000                  0.000000
50%                    0.000000                  0.000000
75%                    0.000000                  0.000000
max                    1.000000                  1.000000

       has_secondary_use_institution  has_secondary_use_school  \
count                 260601.000000             260601.000000
mean                       0.000940                  0.000361
std                        0.030647                  0.018989
min                        0.000000                  0.000000
25%                        0.000000                  0.000000
50%                        0.000000                  0.000000
75%                        0.000000                  0.000000
max                        1.000000                  1.000000

       has_secondary_use_industry  has_secondary_use_health_post  \
count                260601.000000                  260601.000000
mean                      0.001071                       0.000188
std                       0.032703                       0.013711
min                       0.000000                       0.000000
25%                       0.000000                       0.000000
50%                       0.000000                       0.000000
75%                       0.000000                       0.000000
max                       1.000000                       1.000000

       has_secondary_use_gov_office  has_secondary_use_use_police  \
count                 260601.000000                 260601.000000
mean                       0.000146                      0.000088
std                        0.012075                      0.009394
min                        0.000000                      0.000000
25%                        0.000000                      0.000000
50%                        0.000000                      0.000000
```

```
75%                          0.000000                      0.000000
max                          1.000000                      1.000000

       has_secondary_use_other   damage_grade
count             260601.000000  260601.000000
mean                   0.005119       2.238272
std                    0.071364       0.611814
min                    0.000000       1.000000
25%                    0.000000       2.000000
50%                    0.000000       2.000000
75%                    0.000000       3.000000
max                    1.000000       3.000000

[8 rows x 32 columns]
```

[20]:
```
df[df.duplicated()] # no duplicated data
```

[20]:
```
Empty DataFrame
Columns: [building_id, geo_level_1_id, geo_level_2_id, geo_level_3_id,
count_floors_pre_eq, age, area_percentage, height_percentage,
land_surface_condition, foundation_type, roof_type, ground_floor_type,
other_floor_type, position, plan_configuration, has_superstructure_adobe_mud,
has_superstructure_mud_mortar_stone, has_superstructure_stone_flag,
has_superstructure_cement_mortar_stone, has_superstructure_mud_mortar_brick,
has_superstructure_cement_mortar_brick, has_superstructure_timber,
has_superstructure_bamboo, has_superstructure_rc_non_engineered,
has_superstructure_rc_engineered, has_superstructure_other,
legal_ownership_status, count_families, has_secondary_use,
has_secondary_use_agriculture, has_secondary_use_hotel,
has_secondary_use_rental, has_secondary_use_institution,
has_secondary_use_school, has_secondary_use_industry,
has_secondary_use_health_post, has_secondary_use_gov_office,
has_secondary_use_use_police, has_secondary_use_other, damage_grade]
Index: []

[0 rows x 40 columns]
```

[21]:
```
df.isna().sum()/df.shape[0] #percentage of null values
```

[21]:
```
building_id                             0.0
geo_level_1_id                          0.0
geo_level_2_id                          0.0
geo_level_3_id                          0.0
count_floors_pre_eq                     0.0
age                                     0.0
area_percentage                         0.0
height_percentage                       0.0
```

```
land_surface_condition                    0.0
foundation_type                           0.0
roof_type                                 0.0
ground_floor_type                         0.0
other_floor_type                          0.0
position                                  0.0
plan_configuration                        0.0
has_superstructure_adobe_mud              0.0
has_superstructure_mud_mortar_stone       0.0
has_superstructure_stone_flag             0.0
has_superstructure_cement_mortar_stone    0.0
has_superstructure_mud_mortar_brick       0.0
has_superstructure_cement_mortar_brick    0.0
has_superstructure_timber                 0.0
has_superstructure_bamboo                 0.0
has_superstructure_rc_non_engineered      0.0
has_superstructure_rc_engineered          0.0
has_superstructure_other                  0.0
legal_ownership_status                    0.0
count_families                            0.0
has_secondary_use                         0.0
has_secondary_use_agriculture             0.0
has_secondary_use_hotel                   0.0
has_secondary_use_rental                  0.0
has_secondary_use_institution             0.0
has_secondary_use_school                  0.0
has_secondary_use_industry                0.0
has_secondary_use_health_post             0.0
has_secondary_use_gov_office              0.0
has_secondary_use_use_police              0.0
has_secondary_use_other                   0.0
damage_grade                              0.0
dtype: float64
```

[22]:
```
train = pd.merge(values, labels, on='building_id')
```

0.1 Exploratory Data Analysis

[23]:
```
print('Shape of DF:',df.shape)
print(df.dtypes,'\n') #there is no null values
df.info()
```

```
Shape of DF: (260601, 40)
building_id                               int64
geo_level_1_id                            int64
geo_level_2_id                            int64
geo_level_3_id                            int64
```

```
count_floors_pre_eq                       int64
age                                       int64
area_percentage                           int64
height_percentage                         int64
land_surface_condition                   object
foundation_type                          object
roof_type                                object
ground_floor_type                        object
other_floor_type                         object
position                                 object
plan_configuration                       object
has_superstructure_adobe_mud              int64
has_superstructure_mud_mortar_stone       int64
has_superstructure_stone_flag             int64
has_superstructure_cement_mortar_stone    int64
has_superstructure_mud_mortar_brick       int64
has_superstructure_cement_mortar_brick    int64
has_superstructure_timber                 int64
has_superstructure_bamboo                 int64
has_superstructure_rc_non_engineered      int64
has_superstructure_rc_engineered          int64
has_superstructure_other                  int64
legal_ownership_status                   object
count_families                            int64
has_secondary_use                         int64
has_secondary_use_agriculture             int64
has_secondary_use_hotel                   int64
has_secondary_use_rental                  int64
has_secondary_use_institution             int64
has_secondary_use_school                  int64
has_secondary_use_industry                int64
has_secondary_use_health_post             int64
has_secondary_use_gov_office              int64
has_secondary_use_use_police              int64
has_secondary_use_other                   int64
damage_grade                              int64
dtype: object

<class 'pandas.core.frame.DataFrame'>
Int64Index: 260601 entries, 0 to 260600
Data columns (total 40 columns):
 #   Column                                  Non-Null Count   Dtype
---  ------                                  --------------   -----
 0   building_id                             260601 non-null  int64
 1   geo_level_1_id                          260601 non-null  int64
 2   geo_level_2_id                          260601 non-null  int64
 3   geo_level_3_id                          260601 non-null  int64
 4   count_floors_pre_eq                     260601 non-null  int64
```

```
 5   age                                     260601 non-null  int64
 6   area_percentage                         260601 non-null  int64
 7   height_percentage                       260601 non-null  int64
 8   land_surface_condition                  260601 non-null  object
 9   foundation_type                         260601 non-null  object
 10  roof_type                               260601 non-null  object
 11  ground_floor_type                       260601 non-null  object
 12  other_floor_type                        260601 non-null  object
 13  position                                260601 non-null  object
 14  plan_configuration                      260601 non-null  object
 15  has_superstructure_adobe_mud            260601 non-null  int64
 16  has_superstructure_mud_mortar_stone     260601 non-null  int64
 17  has_superstructure_stone_flag           260601 non-null  int64
 18  has_superstructure_cement_mortar_stone  260601 non-null  int64
 19  has_superstructure_mud_mortar_brick     260601 non-null  int64
 20  has_superstructure_cement_mortar_brick  260601 non-null  int64
 21  has_superstructure_timber               260601 non-null  int64
 22  has_superstructure_bamboo               260601 non-null  int64
 23  has_superstructure_rc_non_engineered    260601 non-null  int64
 24  has_superstructure_rc_engineered        260601 non-null  int64
 25  has_superstructure_other                260601 non-null  int64
 26  legal_ownership_status                  260601 non-null  object
 27  count_families                          260601 non-null  int64
 28  has_secondary_use                       260601 non-null  int64
 29  has_secondary_use_agriculture           260601 non-null  int64
 30  has_secondary_use_hotel                 260601 non-null  int64
 31  has_secondary_use_rental                260601 non-null  int64
 32  has_secondary_use_institution           260601 non-null  int64
 33  has_secondary_use_school                260601 non-null  int64
 34  has_secondary_use_industry              260601 non-null  int64
 35  has_secondary_use_health_post           260601 non-null  int64
 36  has_secondary_use_gov_office            260601 non-null  int64
 37  has_secondary_use_use_police            260601 non-null  int64
 38  has_secondary_use_other                 260601 non-null  int64
 39  damage_grade                            260601 non-null  int64
dtypes: int64(32), object(8)
memory usage: 81.5+ MB
```

[24]:
```
df.describe()
```

[24]:

	building_id	geo_level_1_id	geo_level_2_id	geo_level_3_id	\
count	2.606010e+05	260601.000000	260601.000000	260601.000000	
mean	5.256755e+05	13.900353	701.074685	6257.876148	
std	3.045450e+05	8.033617	412.710734	3646.369645	
min	4.000000e+00	0.000000	0.000000	0.000000	
25%	2.611900e+05	7.000000	350.000000	3073.000000	
50%	5.257570e+05	12.000000	702.000000	6270.000000	

```
75%    7.897620e+05       21.000000     1050.000000     9412.000000
max    1.052934e+06       30.000000     1427.000000    12567.000000

       count_floors_pre_eq            age  area_percentage  height_percentage  \
count        260601.000000  260601.000000    260601.000000      260601.000000
mean              2.129723      26.535029         8.018051           5.434365
std               0.727665      73.565937         4.392231           1.918418
min               1.000000       0.000000         1.000000           2.000000
25%               2.000000      10.000000         5.000000           4.000000
50%               2.000000      15.000000         7.000000           5.000000
75%               2.000000      30.000000         9.000000           6.000000
max               9.000000     995.000000       100.000000          32.000000

       has_superstructure_adobe_mud  has_superstructure_mud_mortar_stone  …  \
count                 260601.000000                       260601.000000  …
mean                       0.088645                            0.761935  …
std                        0.284231                            0.425900  …
min                        0.000000                            0.000000  …
25%                        0.000000                            1.000000  …
50%                        0.000000                            1.000000  …
75%                        0.000000                            1.000000  …
max                        1.000000                            1.000000  …

       has_secondary_use_hotel  has_secondary_use_rental  \
count            260601.000000             260601.000000
mean                  0.033626                  0.008101
std                   0.180265                  0.089638
min                   0.000000                  0.000000
25%                   0.000000                  0.000000
50%                   0.000000                  0.000000
75%                   0.000000                  0.000000
max                   1.000000                  1.000000

       has_secondary_use_institution  has_secondary_use_school  \
count                 260601.000000             260601.000000
mean                       0.000940                  0.000361
std                        0.030647                  0.018989
min                        0.000000                  0.000000
25%                        0.000000                  0.000000
50%                        0.000000                  0.000000
75%                        0.000000                  0.000000
max                        1.000000                  1.000000

       has_secondary_use_industry  has_secondary_use_health_post  \
count               260601.000000                  260601.000000
mean                    0.001071                       0.000188
std                     0.032703                       0.013711
```

```
min                       0.000000                       0.000000
25%                       0.000000                       0.000000
50%                       0.000000                       0.000000
75%                       0.000000                       0.000000
max                       1.000000                       1.000000

       has_secondary_use_gov_office  has_secondary_use_use_police  \
count                 260601.000000                 260601.000000
mean                       0.000146                      0.000088
std                        0.012075                      0.009394
min                        0.000000                      0.000000
25%                        0.000000                      0.000000
50%                        0.000000                      0.000000
75%                        0.000000                      0.000000
max                        1.000000                      1.000000

       has_secondary_use_other   damage_grade
count            260601.000000  260601.000000
mean                  0.005119       2.238272
std                   0.071364       0.611814
min                   0.000000       1.000000
25%                   0.000000       2.000000
50%                   0.000000       2.000000
75%                   0.000000       3.000000
max                   1.000000       3.000000

[8 rows x 32 columns]
```

[25]:
```
df[df.duplicated()] # no duplicated data
```

[25]:
```
Empty DataFrame
Columns: [building_id, geo_level_1_id, geo_level_2_id, geo_level_3_id,
count_floors_pre_eq, age, area_percentage, height_percentage,
land_surface_condition, foundation_type, roof_type, ground_floor_type,
other_floor_type, position, plan_configuration, has_superstructure_adobe_mud,
has_superstructure_mud_mortar_stone, has_superstructure_stone_flag,
has_superstructure_cement_mortar_stone, has_superstructure_mud_mortar_brick,
has_superstructure_cement_mortar_brick, has_superstructure_timber,
has_superstructure_bamboo, has_superstructure_rc_non_engineered,
has_superstructure_rc_engineered, has_superstructure_other,
legal_ownership_status, count_families, has_secondary_use,
has_secondary_use_agriculture, has_secondary_use_hotel,
has_secondary_use_rental, has_secondary_use_institution,
has_secondary_use_school, has_secondary_use_industry,
has_secondary_use_health_post, has_secondary_use_gov_office,
has_secondary_use_use_police, has_secondary_use_other, damage_grade]
Index: []
```

```
[0 rows x 40 columns]
```

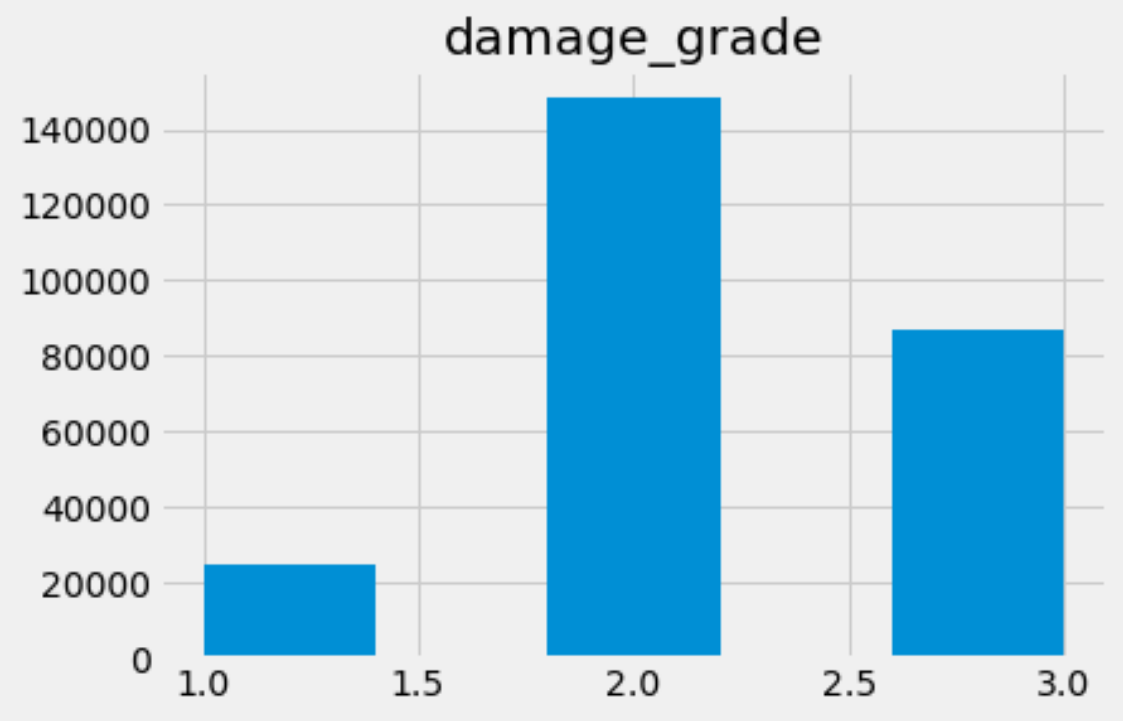
[26]:
```
df.isna().sum()/df.shape[0] #percentage of null values
```

[26]:
```
building_id                             0.0
geo_level_1_id                          0.0
geo_level_2_id                          0.0
geo_level_3_id                          0.0
count_floors_pre_eq                     0.0
age                                     0.0
area_percentage                         0.0
height_percentage                       0.0
land_surface_condition                  0.0
foundation_type                         0.0
roof_type                               0.0
ground_floor_type                       0.0
other_floor_type                        0.0
position                                0.0
plan_configuration                      0.0
has_superstructure_adobe_mud            0.0
has_superstructure_mud_mortar_stone     0.0
has_superstructure_stone_flag           0.0
has_superstructure_cement_mortar_stone  0.0
has_superstructure_mud_mortar_brick     0.0
has_superstructure_cement_mortar_brick  0.0
has_superstructure_timber               0.0
has_superstructure_bamboo               0.0
has_superstructure_rc_non_engineered    0.0
has_superstructure_rc_engineered        0.0
has_superstructure_other                0.0
legal_ownership_status                  0.0
count_families                          0.0
has_secondary_use                       0.0
has_secondary_use_agriculture           0.0
has_secondary_use_hotel                 0.0
has_secondary_use_rental                0.0
has_secondary_use_institution           0.0
has_secondary_use_school                0.0
has_secondary_use_industry              0.0
has_secondary_use_health_post           0.0
has_secondary_use_gov_office            0.0
has_secondary_use_use_police            0.0
has_secondary_use_other                 0.0
damage_grade                            0.0
dtype: float64
```

[27]:
```
df['damage_grade'].value_counts() #not goodly balanced
```

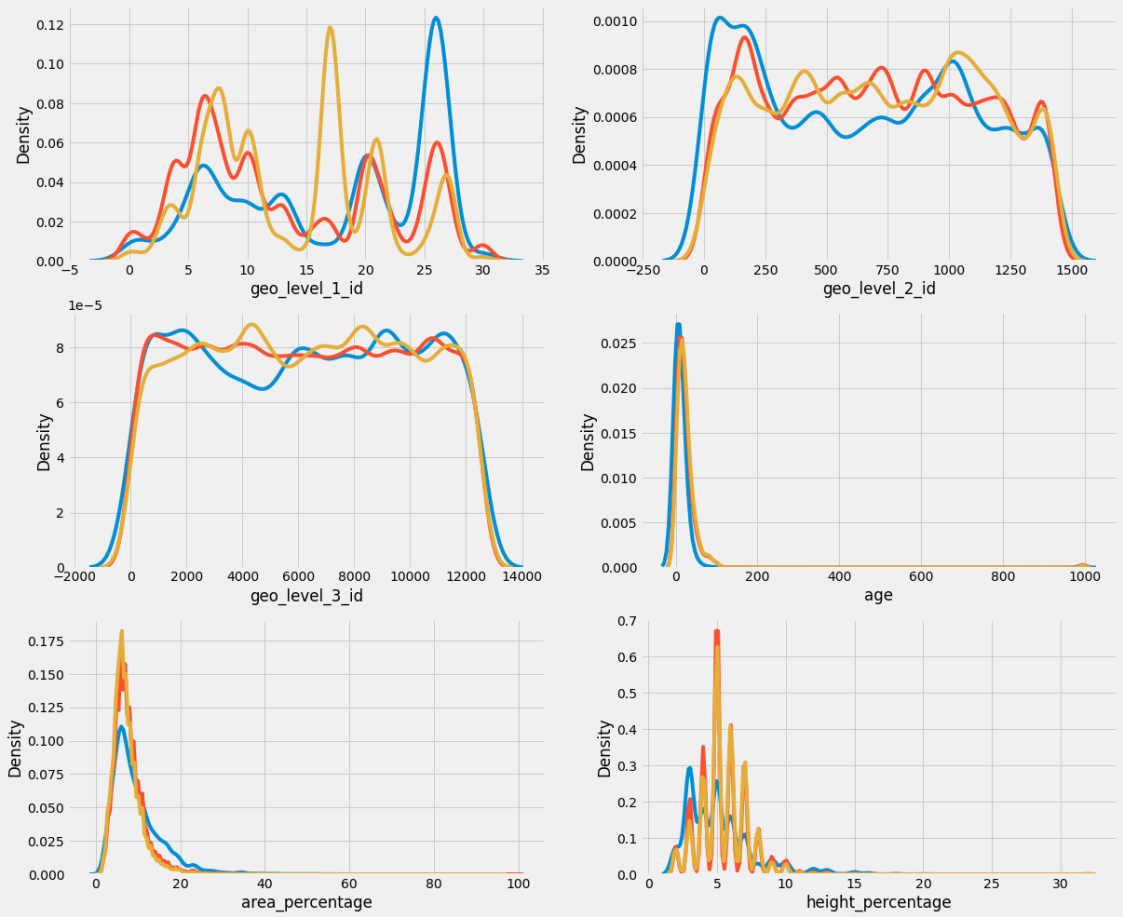
[27]:
```
2    148259
3     87218
1     25124
Name: damage_grade, dtype: int64
```

[28]:
```
hist = df.hist('damage_grade',bins=5)
plt.show()
```

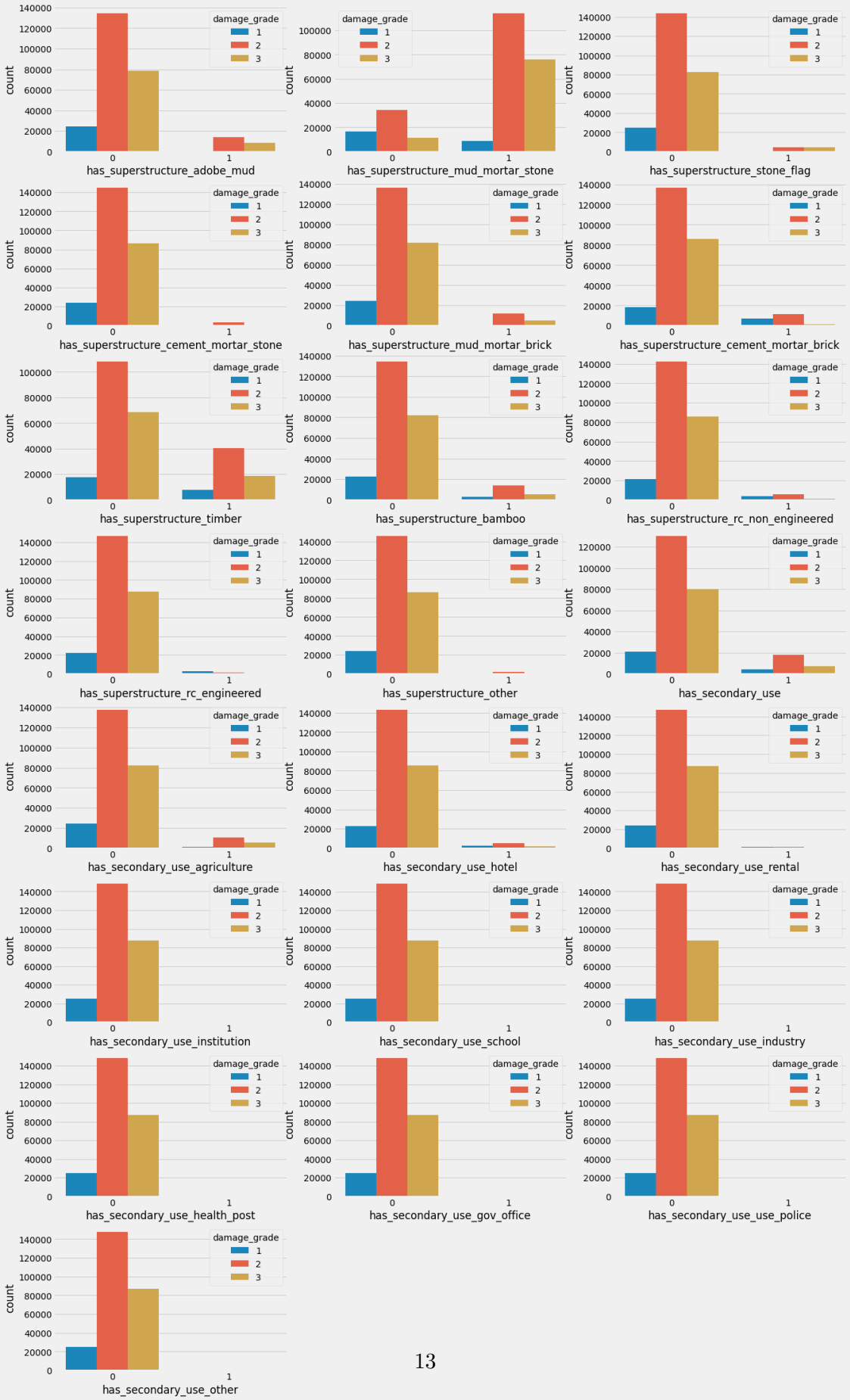


[29]:
```
continous_values =␣
 ↪['geo_level_1_id','geo_level_2_id','geo_level_3_id','age','area_percentage','height_percent

def densityPlot(continous_values):
  fig = plt.figure(figsize=(18,16))
  plt.style.use('fivethirtyeight')
  for i,txt in enumerate(continous_values):
    ax = fig.add_subplot(3,2,i+1)
    sns.kdeplot(train.loc[train['damage_grade'] == 1, txt], ax=ax,␣
 ↪label='damage_grade==1')
    sns.kdeplot(train.loc[train['damage_grade'] == 2, txt], ax=ax,␣
 ↪label='damage_grade==2')
    sns.kdeplot(train.loc[train['damage_grade'] == 3, txt], ax=ax,␣
 ↪label='damage_grade==3')
```

```
  plt.show()
densityPlot(continous_values)
```

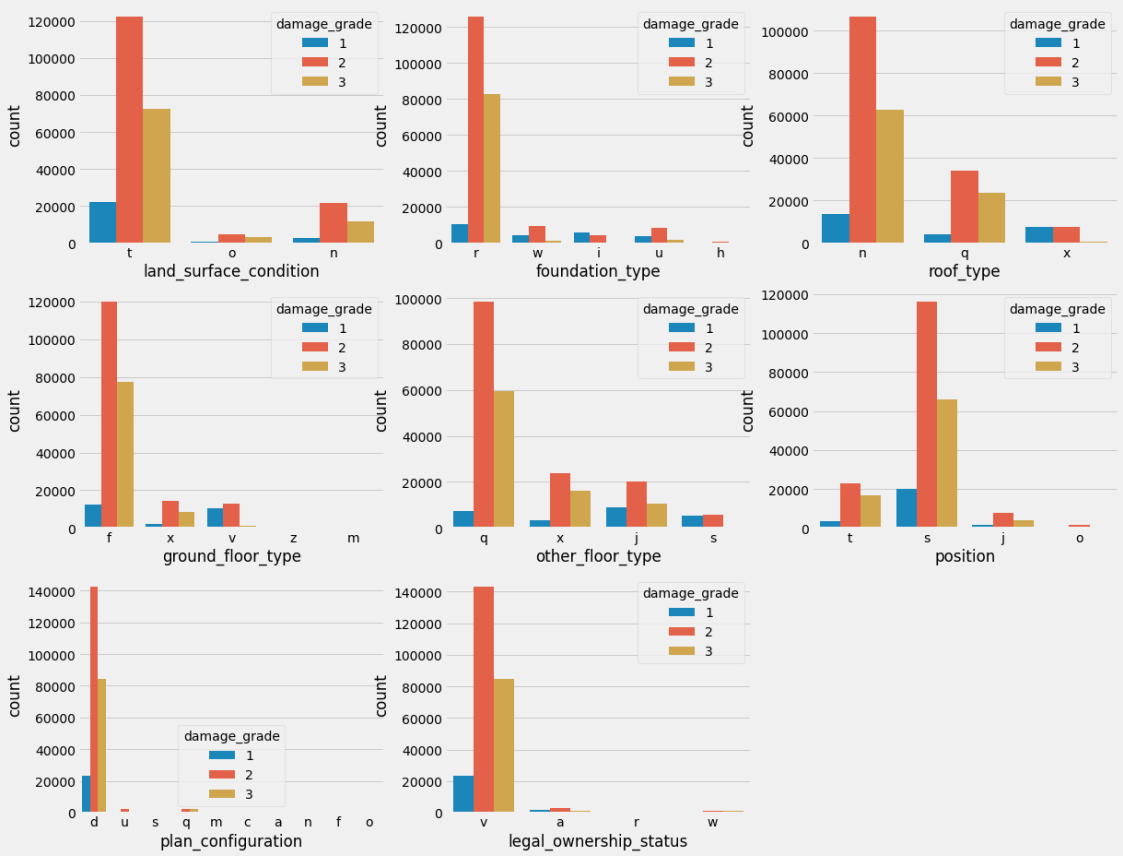


[30]:
```
binary_features = train.columns[train.columns.str.startswith('has')]

def countPlot(binary_features):
  plt.rcParams['font.size'] = 18
  plt.style.use('fivethirtyeight')
  fig = plt.figure(figsize=(20,37))
  for i,txt in enumerate(bin_cols):
    ax = fig.add_subplot(8,3,i+1)
    sns.countplot(x=train[txt], ax=ax, hue=train['damage_grade'])
  plt.show()
countPlot(binary_features)
```


damage_grade
1
2
3
count
has_superstructure_adobe_mud
has_superstructure_mud_mortar_stone
has_superstructure_stone_flag
has_superstructure_cement_mortar_stone
has_superstructure_mud_mortar_brick
has_superstructure_cement_mortar_brick
has_superstructure_timber
has_superstructure_bamboo
has_superstructure_rc_non_engineered
has_superstructure_rc_engineered
has_superstructure_other
has_secondary_use
has_secondary_use_agriculture
has_secondary_use_hotel
has_secondary_use_rental
has_secondary_use_institution
has_secondary_use_school
has_secondary_use_industry
has_secondary_use_health_post
has_secondary_use_gov_office
has_secondary_use_use_police
has_secondary_use_other

Except has_superstructure_cement_mortar_stone other binary features have more zero than 1 and some columns have only zero values

[31]:
```
categorical_features = train.select_dtypes(include=object).columns

def catPlot(categorical_features):
  plt.rcParams['font.size'] = 18
  plt.style.use('fivethirtyeight')
  fig = plt.figure(figsize=(18,15))
  for i,txt in enumerate(categorical_features):
    ax = fig.add_subplot(3,3,i+1)
    sns.countplot(x=train[txt], ax=ax, hue=train['damage_grade'])
  plt.show()
catPlot(categorical_features)
```



0.2 Feature Engineering

[32]:
```
df = pd.concat([train, test_values], axis=0).reset_index(drop=True)
df.shape
```

[32]: (347469, 40)