FACULTAD
DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico Nº1

Análisis exploratorio

Grupo L.F.F.F

7506 - Organización de Datos

Prof. Argerich , Luis

Luz Fox

Francisco Luna

Francisco Talenti

Federico Krell

1 ÍNDICE DE CONTENIDOS

2 INTRODUCCIÓN

El 25 de abril del 2015 ocurrió un terremoto de 7.8 Mw en el territorio de Gorkha, Nepal. Tras estos hechos el gobierno decidió realizar una encuesta masiva de hogares a través de dispositivos móviles para evaluar los daños ocasionados en el distrito. Esta encuesta también reúne datos socioeconómicos que son de utilidad para realizar una investigación.

El objetivo del presente trabajo es realizar una análisis exploratorio de los datos para encontrar *insights* y extraer información relevante, para luego concluir con un modelado de los mismos a través de tecnologías de machine learning -el mismo será abordado en un trabajo futuro-.

El trabajo se estructura en tres principales procesos: limpieza, exploración y conclusión.

3 PROCESAMIENTO

3.1 DATOS UTILIZADOS

Para el presente estudio los datos se obtuvieron del siguiente enlace:

https://www.drivendata.org/competitions/57/nepal-earthquake/data/

El dataset consiste principalmente de datos correspondientes a estructuras edilicias y del dueño legal. Cada fila representa un edificio específico en la región que fue afectado por el terremoto.

En total hay 39 columnas en el dataset.

Los datos provienen de dos fuentes:

- train_values.csv: consta de 260601 filas y 38 columnas.
- train_labels.csv: consta de 260601 filas y 1 columna.

3.1.1 train_values.csv

Consta de una única columna “damage_grade” con un único índice “building_id”.

- *damage_grade:* representa el grado de daño a la edificación. Hay solamente tres posibles valores:
  - *o* 1: daño leve
  - *o* 2: daño moderado
  - *o* 3: casi destrucción completa
- *building_id*: es un identificador único y random.

3.1.2 train_labels.csv

Consta de una 38 columnas y un único índice “building_id”. Para el caso del índice “building_id” corresponde a las mismas características descriptas previamente. Para el caso de las columnas:

- geo_level_1_id, geo_level_2_id, geo_level_3_id: región geografica en donde se encontraba la edificación. Desde el nivel 1 hasta el nivel 3 (sub región específica).
  - o geo_level_1_id: 0-30
  - o geo_level_2_id: 0-1427
  - o geo_level_3_id: 0-12567
- count_floors_pre_eq: número de pisos en la edificaciñon antes de del terremoto.
- Age: cantidad de años que posee la edificación.
- area_percentage: área normalizada de la superficie del edificio.
- height_percentage: área normalizada de la altura del edificio.
- land_surface_condition: condición de la superficie de la tierra donde se encuentra la edificación.
- foundation_type: tipo de utilización que se le daba a la edificación.

- roof_type: tipo de techo que poseía la edificación.
- ground_floor_type: tipo de suelo de la edificación.
- other_floor_type: tipo de construcciones usadas en edificaciones que posean mas de un piso (exceptuando el techo).
- Position: posición de edificio.
- plan_configuration: configuración de plan de la edificación.
- has_superstructure_adobe_mud: bandera para identificar si la estructura fue hecha con adobe o barro.
- has_superstructure_mud_mortar_stone: bandera para identificar si la estructura fue hecha con un tipo de barro o piedra.
- has_superstructure_stone_flag: bandera para identificar si la estructura fue hecha con piedra.
- has_superstructure_cement_mortar_stone: bandera para identificar si la estructura fue hecha cemento.
- has_superstructure_mud_mortar_brick: bandera para identificar si la estructura fue hecha con ladrillos de barro.
- has_superstructure_cement_mortar_brick: bandera para identificar si la estructura fue hecha con ladrillos de cemento.
- has_superstructure_timber: bandera para identificar si la estructura fue hecha con madera.
- has_superstructure_bamboo: bandera para identificar si la estructura fue hecha con bambú.
- has_superstructure_rc_non_engineered: bandera para identificar si la estructura fue hecha sin hormigón.
- has_superstructure_rc_engineered: bandera para identificar si la estructura fue hecha con hormigón.
- has_superstructure_other: bandera para identificar si la estructura fue hecha con cualquier otro material.
- legal_ownership_status: status legal del dueño de la tierra donde la edificación se encuentra.
- count_families: cantidad de familias que vivían en la edificación.
- has_secondary_use: bandera para identificar si la edificación era usada para otros propósitos.
- has_secondary_use_agriculture: bandera para indicar si fue usada para la agricultura.
- has_secondary_use_hotel: bandera para indicar si poseía uso para la hotelería la edificación.
- has_secondary_use_rental: bandera para indicar si poseía uso para alquilarla.
- has_secondary_use_institution: bandera para indicar si poseía uso como lugar para alguna institución.
- has_secondary_use_school: bandera para indicar si poseía uso como escuela.

- has_secondary_use_industry: bandera para indicar si poseía uso para desarrollar la hotelería.
- has_secondary_use_health_post: bandera para indicar si la edificación era usada para cuidados médicos.
- has_secondary_use_gov_office: bandera para indicar si la edificación era usada como oficina gubernamental.
- has_secondary_use_use_police: bandera para indicar si la edificación era usada como centro policial.
- has_secondary_use_other: bandera para indicar si la edificación era usada para otros usos.

3.2 Tecnologías utilizadas

Para el presente trabajo se utilizaron:

- Python 3
- Jupyter notebook
- GitHub
- Librerías:
    - Numpy.
    - Pandas.
    - Matplotlib.
    - Seaborn.
    - Scipy
    - Time
    - Warnings

3.3 Repositorio

El desarrollo del presente trabajo se encuentra en:

https://github.com/paco3209/tpFiuba

4 Limpieza

4.1 Resultados de la limpieza

Para la limpieza de los datos se realizó una investigación que consistió en:

- Analizar datos faltantes en el dataset.
- Analizar los tipos de datos dentro del dataset.
- Identificar la distribución de los datos.
- Identificar y tratar outliers.

4.1.1 Valores faltantes

No se encontraron valores faltantes en las columnas de ambos dataset.

```
train_values.isnull().sum() * 100 / len(train_values)
```

```
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
dtype: float64
```

```
train_labels.isnull().sum() * 100 / len(train_values)
```

```
damage_grade    0.0
dtype: float64
```

5 ANÁLISIS EXPLORATORIO

5.1 Descripción

El análisis exploratorio consistió en investigar las diferentes variables recibidas y por un lado su relación de daño total intravariable, como puede ser la cantidad de edificios dañados por antigüedad, y por otro lado dado una variable los niveles de daño en la misma, por ejemplo si la casa es de barro cuantos casos de daño alto, medio y bajo tuvo. De esta forma se podrá estimar el nivel de daño que va a recibir una edificación dados sus registros geográficos y constructivos.

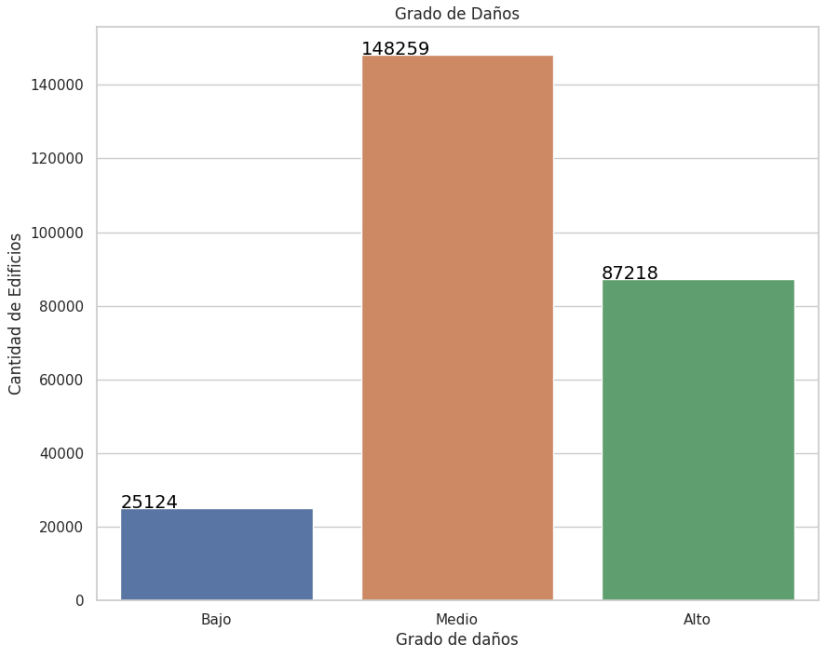
5.2 Primer abordaje del dataset

5.2.1 Tipos de datos

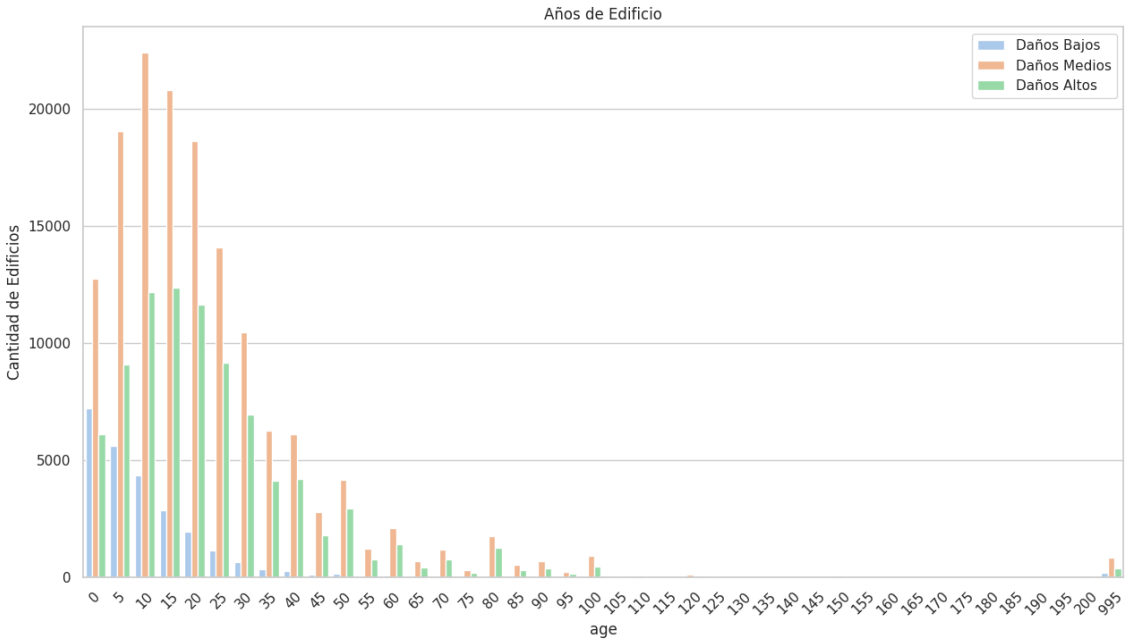
Los datos se muestran consistentes

```
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
```

5.3 Frecuencia de la variable de interés: nivel de daño (damage - grade)

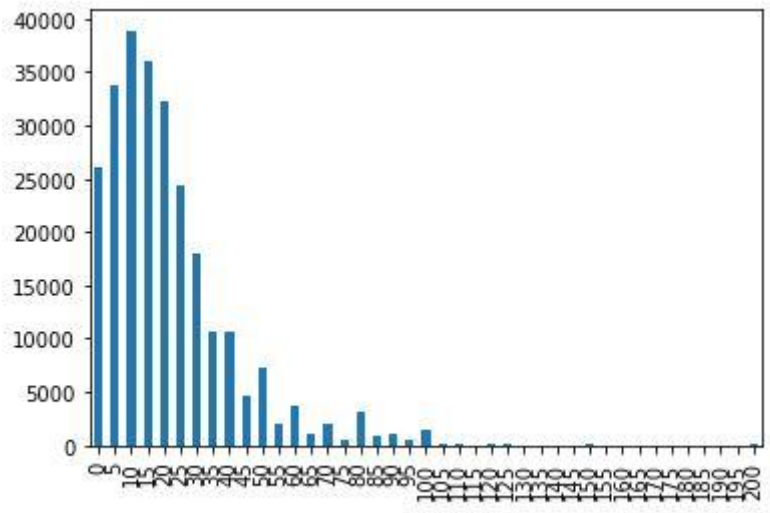


5.4 Daños por antigüedad de Edificio

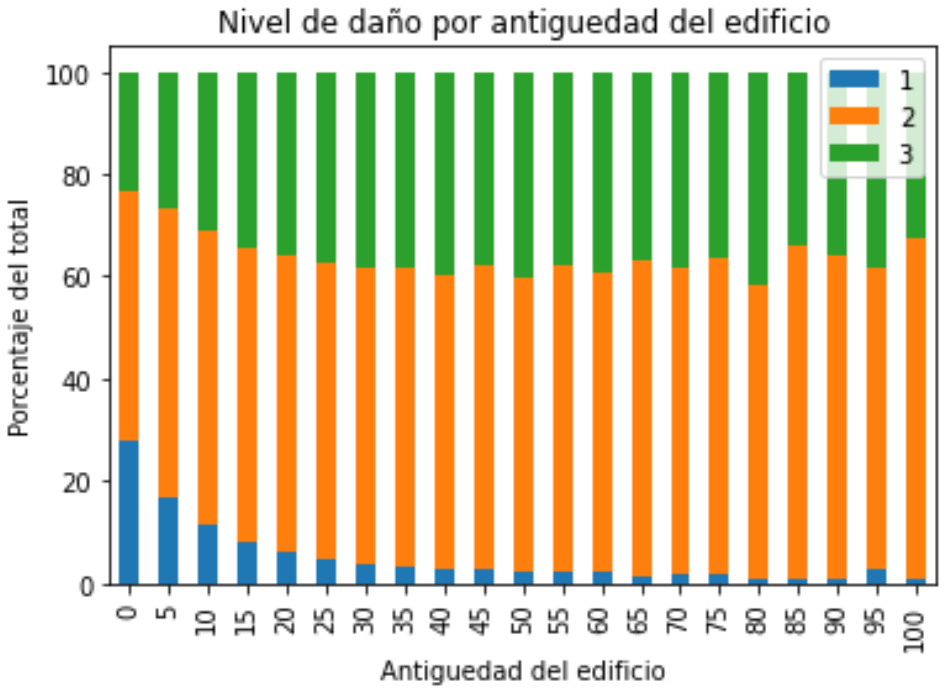


En la distribución inicial puede verse que hay algunos datos de edad, los edificios con 995, que corresponden a un etiquetado diferenciado, hasta los 100 años pareciera haber información utilizable y los de 995 son simplemente más viejos. De todas formas son una estricta minoría y en análisis posteriores podrían ser simplemente limpiados para poder cubrir con algún modelo la mayoría de los casos. Se puede ver una distribución estadística normal si se dejan esos casos afuera.

Dicha distribución puede verse si se deja solo el gráfico de edades y edificios dañados:

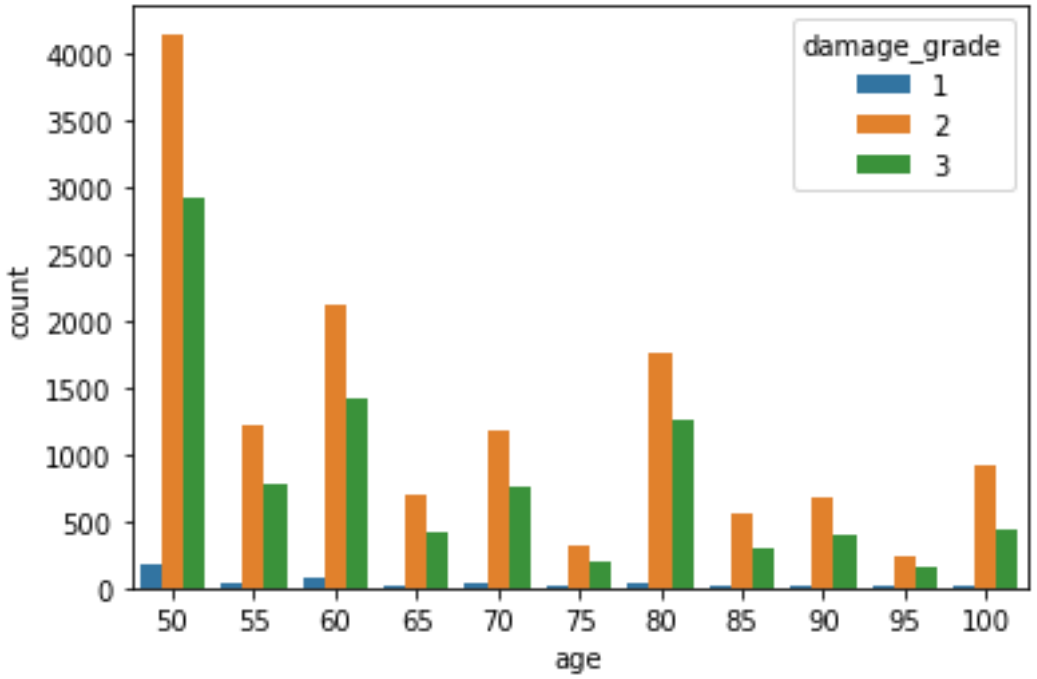


Por otro lado, ya con los datos filtrados se analiza separando por año cuál es el porcentaje de nivel de daño normalizado por casos en intervalos de 5 años de antigüedad:



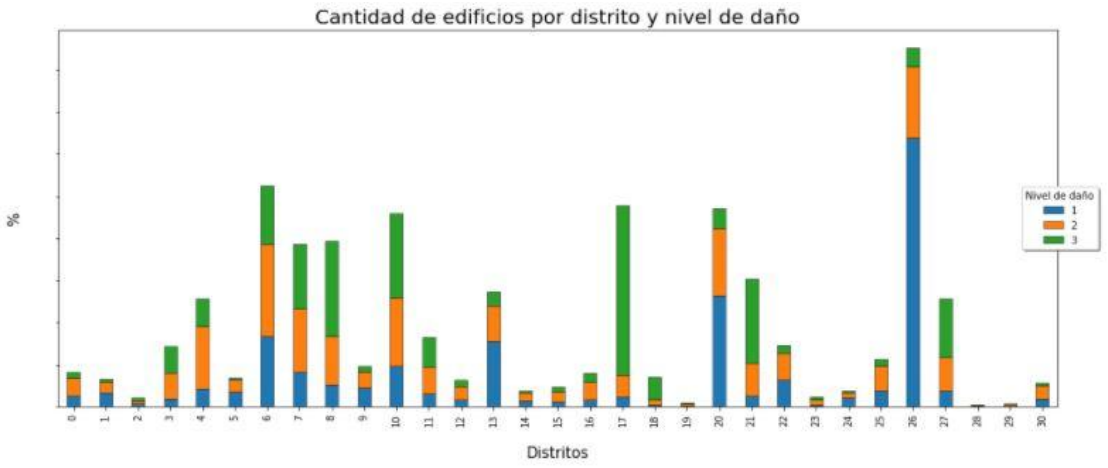
Estos daños pueden ser utilizados para caracterizar por edad el posible daño que se espera.

En los casos de los edificios de más de 50 años se puede esperar que la distribución sea característica por los materiales utilizados y el tiempo de desgaste.



En efecto, la cantidad de casos con daño bajo son pocos.

5.5 Nivel de daño por distrito.

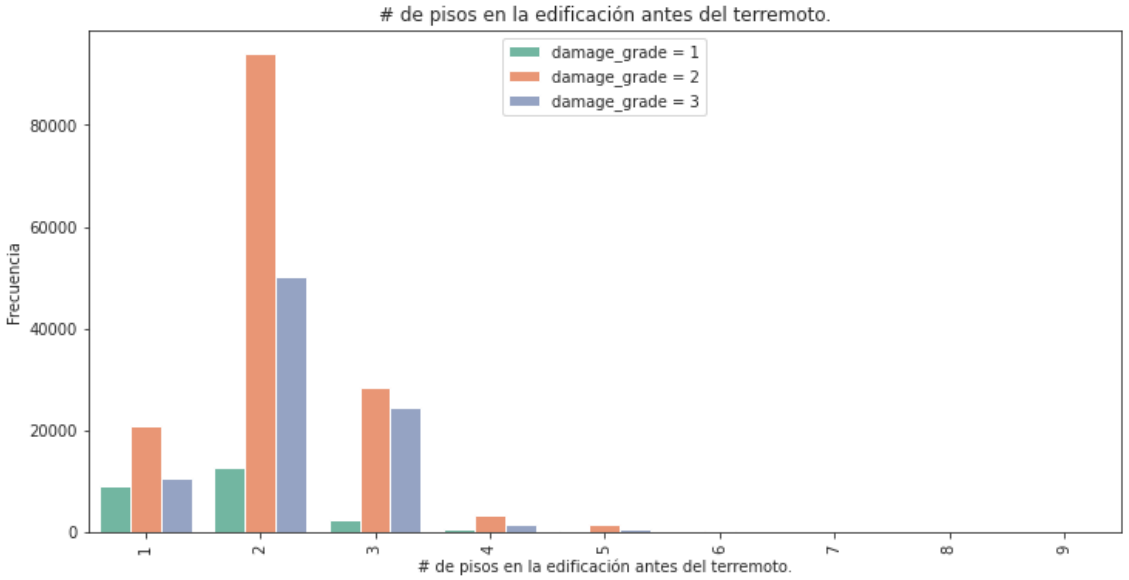


Se obtuvo una caracterización por distrito y nivel de daño. Aquí puede verse que hay distritos en específico que tienen una mayor cantidad de edificios dañados y también puede verse que hay distritos en los que el daño en cantidad fue menor pero fue más

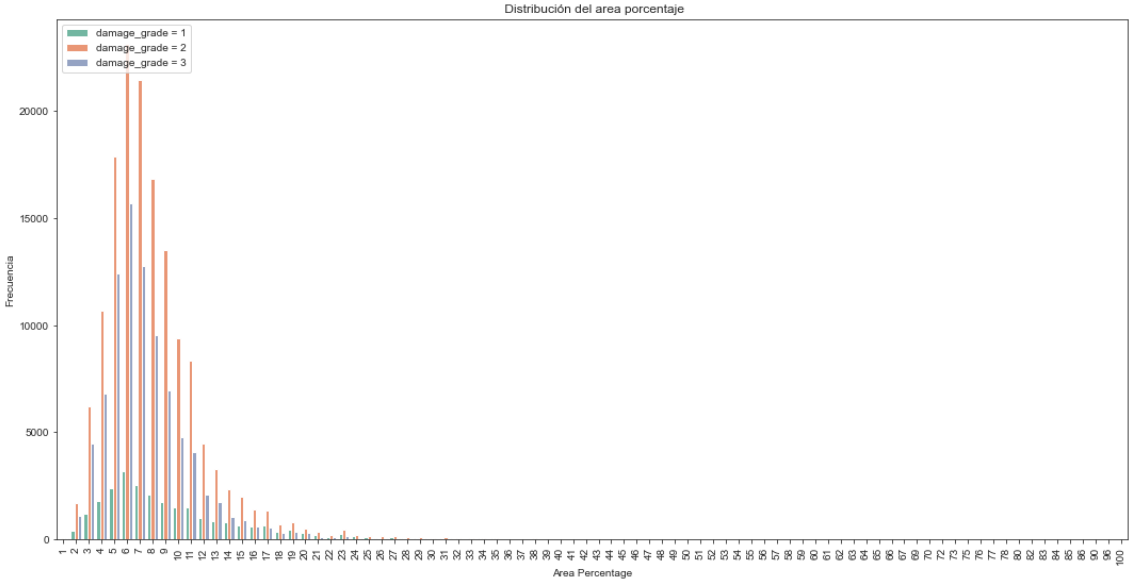
grave. El análisis se hizo con el GEO_ID 1 pero puede extenderse el mismo análisis en los otros niveles y se obtendrá con mayor precisión cuales son los sectores más vulnerables a ser dañados y cuales con mayor nivel de daño.

5.6 Distribución del número de pisos en la edificación antes del terremoto.

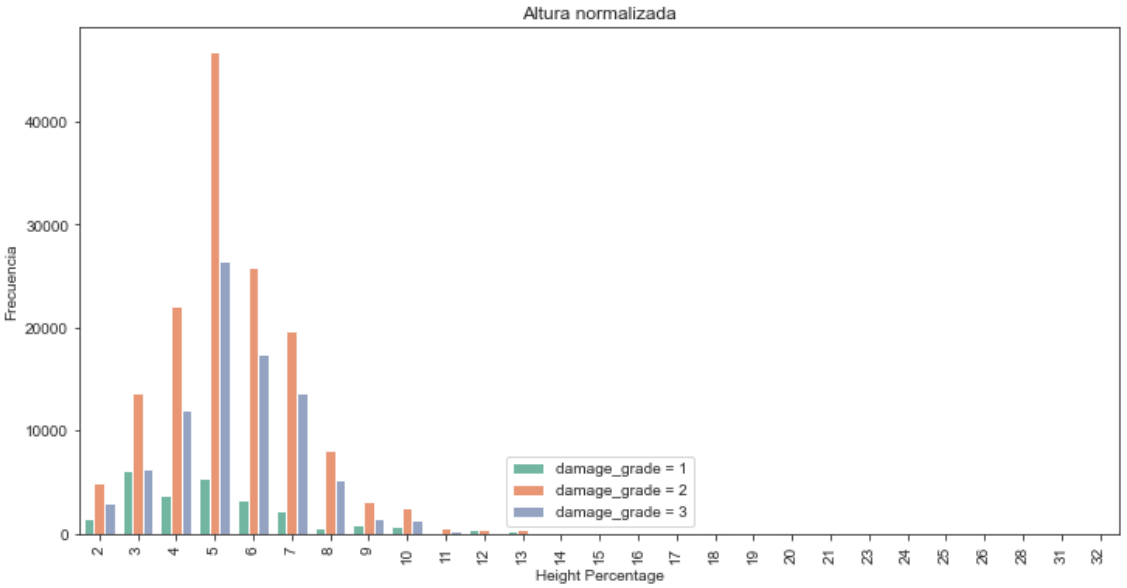
Puede observarse que la mayoría de los edificios dañados eran de dos pisos, sin embargo entre 1 y 3 pisos tienen suficiente cantidad de edificios como para hacer un análisis estadístico que relaciona los niveles de daños con su cantidad de pisos.



5.7 Superficie ocupada y altura normalizadas.



Se observa una distribución en la que las curvas de todos los niveles de daños hacen pico y su declive por los mismos porcentajes. Probablemente sea difícil relacionar esta variable geométrica con el nivel de daño.



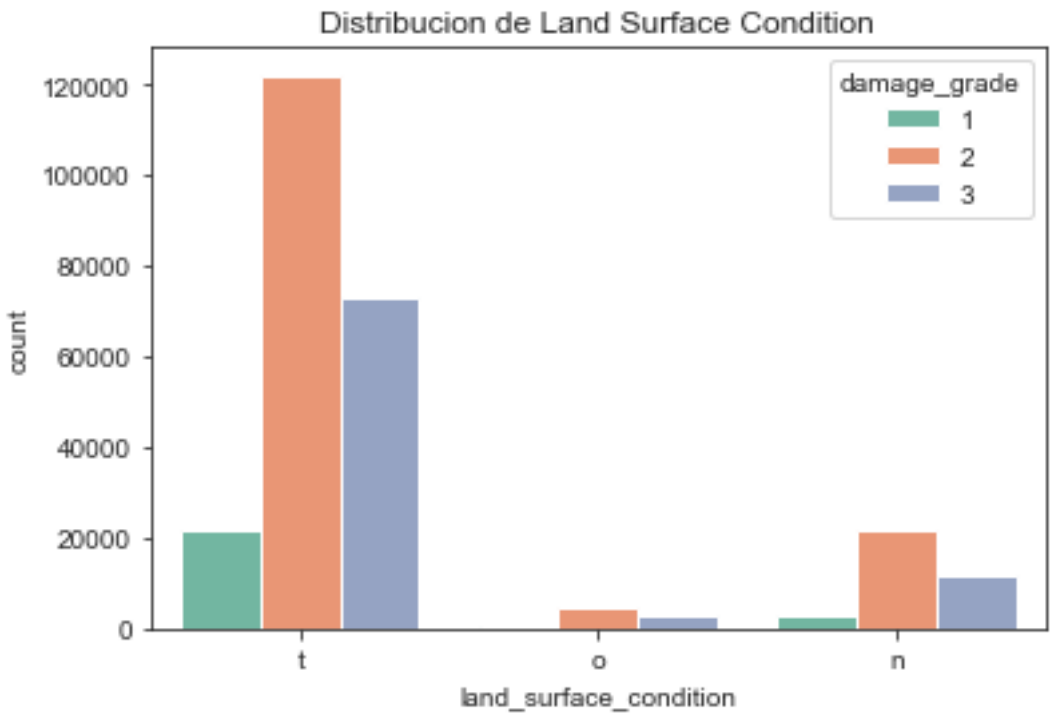
Por otro lado la altura pareciera ser una variable con mayor influencia en el nivel de daño teniendo los daños 1 una curva con menores alturas y los daños 2 y 3 una mayor ocurrencia con alturas un poco mayores.

5.8 Variables categóricas.

Se realizó un análisis de las variables categóricas, las mismas serán utilizadas para obtener una idea de los valores que darían una mayor incidencia de daño y al mismo por tener una mayor incidencia servirán para ver la distribución de los niveles de daño (1, 2 o 3) dada esa categoría.

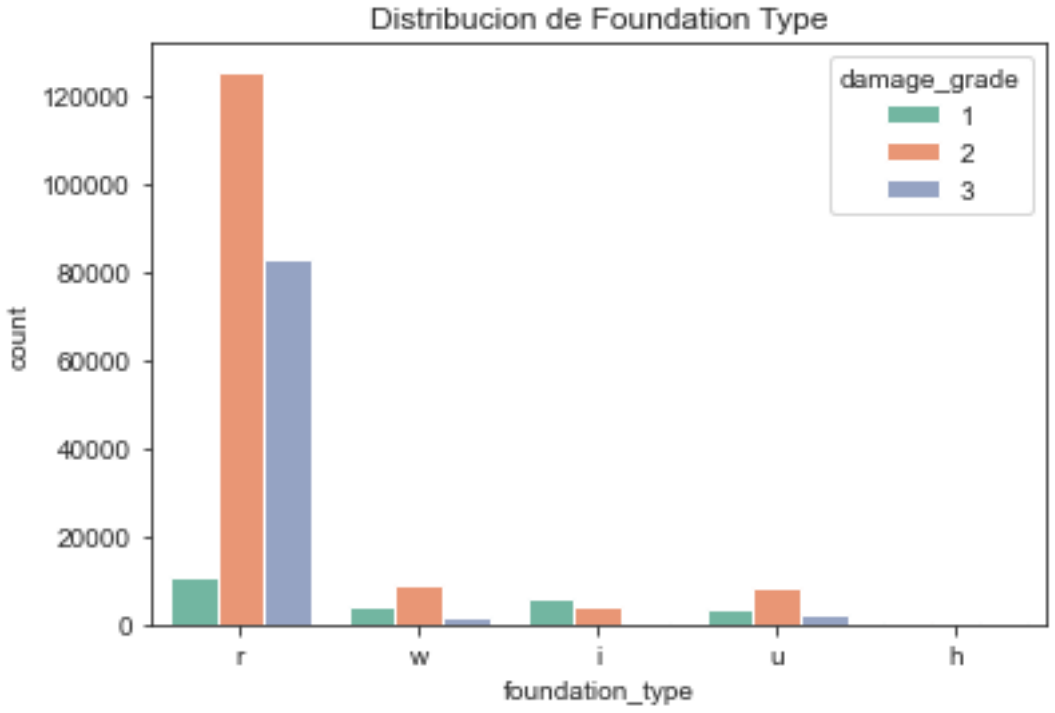
Distribución de la condición de la superficie terrestre donde el edificio fue construido.

Podemos observar la variable categórica que indica la condición de la superficie terrestre donde el edificio fue construido. Los valores posibles son n, o, t. El valor con mayor ocurrencia de daño grado 3 es t. En principio este categórico podría indicar una mayor vulnerabilidad.



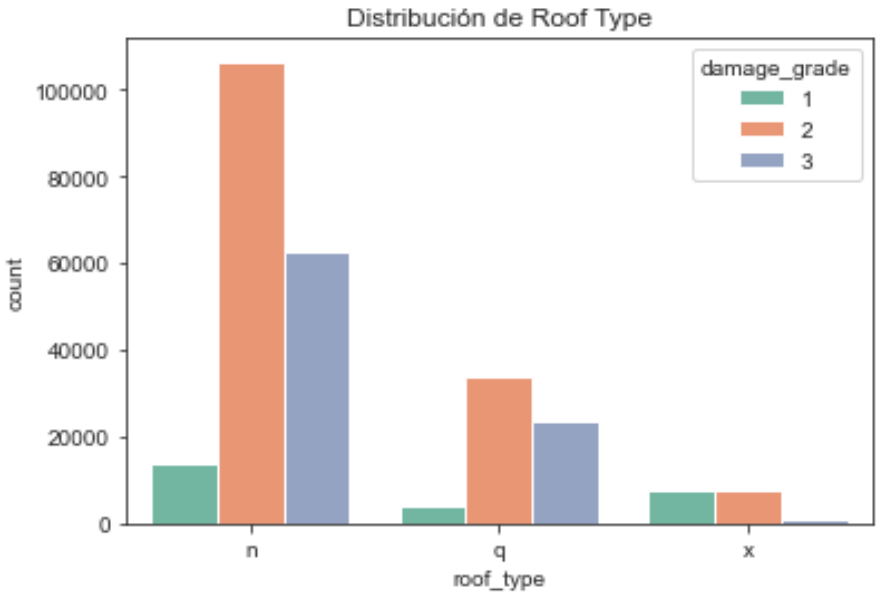
DISTRIBUCIÓN DE TIPO DE CIMIENTOS USADOS CUANDO SE CONSTRUYÓ LA EDIFICACIÓN

Variable de tipo categórico que indica el tipo de cimientos usados cuando se construyó la edificación. Los valores posibles pueden ser h, i, r, u, w. El valor r podría indicar una mayor vulnerabilidad.



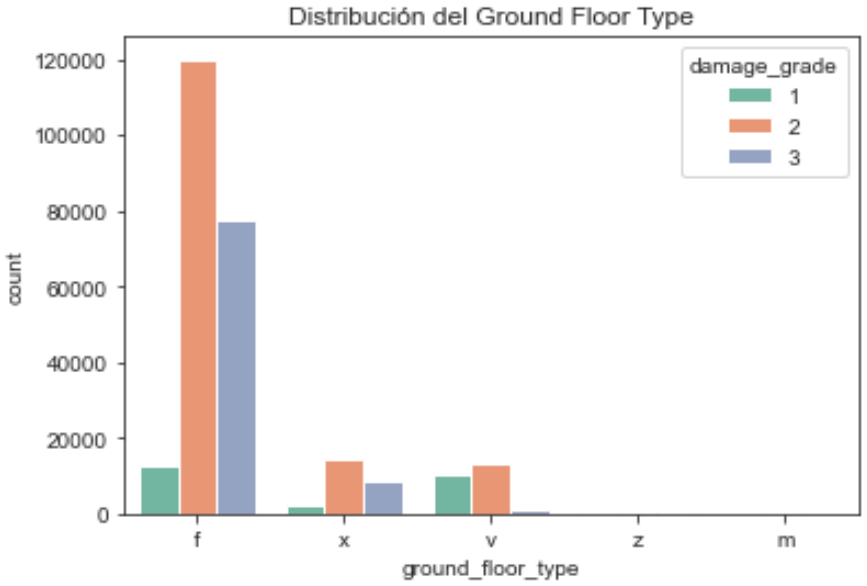
DISTRIBUCIÓN DEL TIPO DE TECHO USADO CUANDO SE CONSTRUYÓ LA EDIFICACIÓN.

Variable de tipo categórico que indica el tipo de techo usado cuando se construyó la edificación. Valores posibles: n, q, x. El valor n podría indicar una mayor vulnerabilidad.



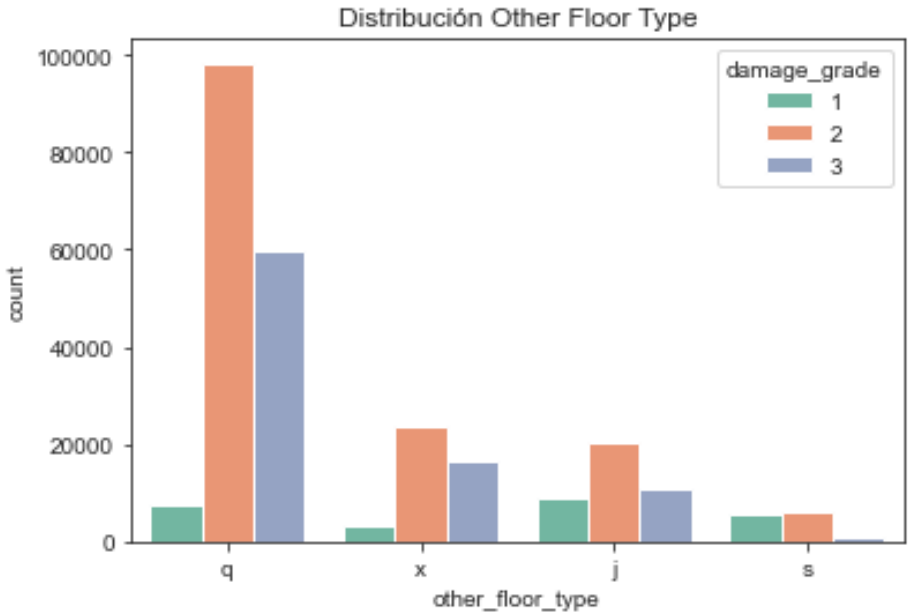
DISTRIBUCIÓN DEL TIPO DE CONSTRUCCIÓN USADO EN LA PLANTA BAJA CUANDO SE CONSTRUYÓ LA EDIFICACIÓN.

Variable de tipo: categórico que indica tipo de construcción usado en la planta baja cuando se construyó la edificación. Valores posibles: f, m, v, x, z. El valor f podría indicar una mayor vulnerabilidad.



Distribución del tipo de construcción usado en otros pisos cuando se construyó la edificación (exceptuando el techo).

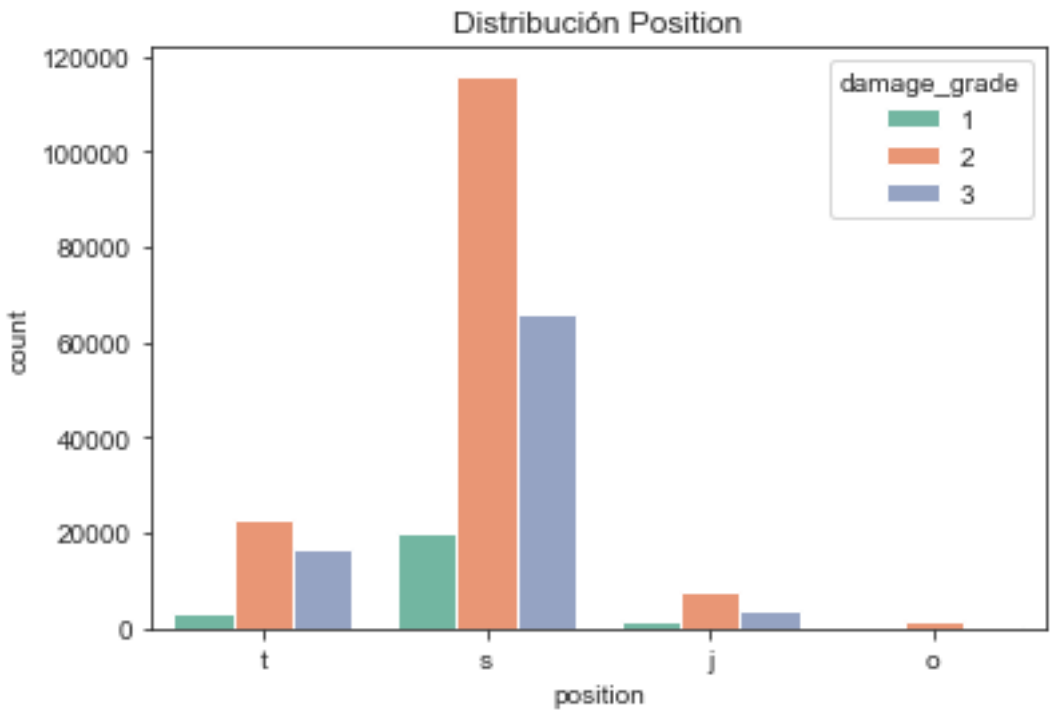
Variable de tipo categórica que tipo de construcción usado en otros pisos cuando se construyó la edificación (exceptuando el techo). Posibles valores: j, q, s, x. El valor q podría indicar una mayor vulnerabilidad.



Distribución de la orientación de la edificación.

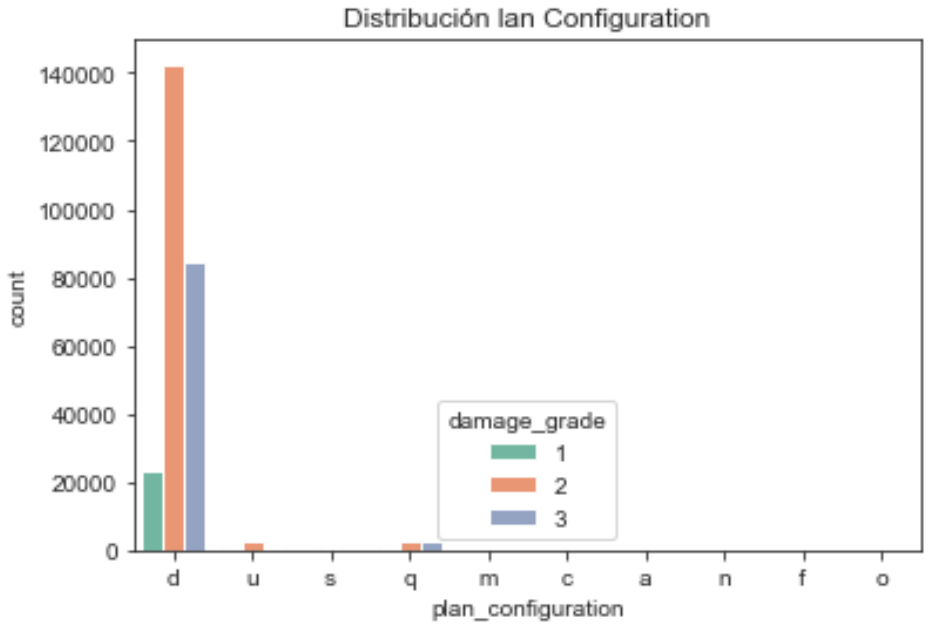
Variable de tipo categórico que indica la orientación de la edificación. Posibles valores: j, o, s, t. La orientación s podría indicar una mayor vulnerabilidad.

.



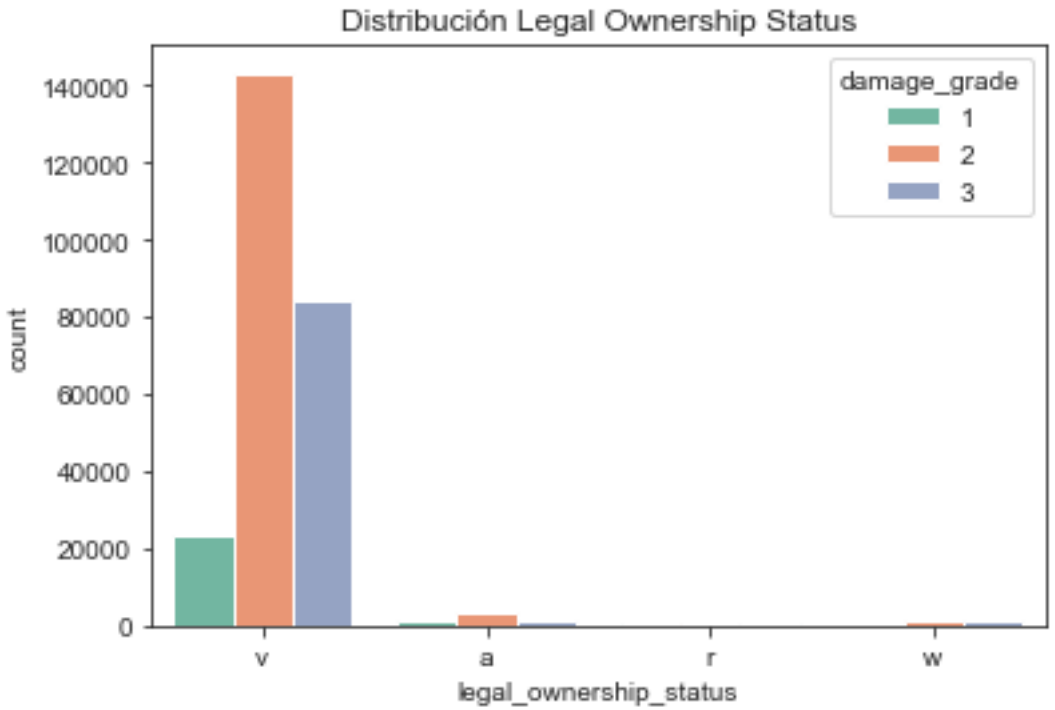
DISTRIBUCIÓN DEL FORMATO DE CONSTRUCCIÓN DE LA EDIFICACIÓN (PARA DISEÑO SÍSMICO).

Variable de tipo categórico) formato de construcción de la edificación (para diseño sísmico). Valores posibles: a, c, d, f, m, n, o, q, s, u. El valor d podría indicar una mayor vulnerabilidad.



Distribución que indica estado legal de la tierra donde la edificación fue construida

Variable de tipo categórico que indica estado legal de la tierra donde la edificación fue construida. Posibles valores: a, r, v, w.

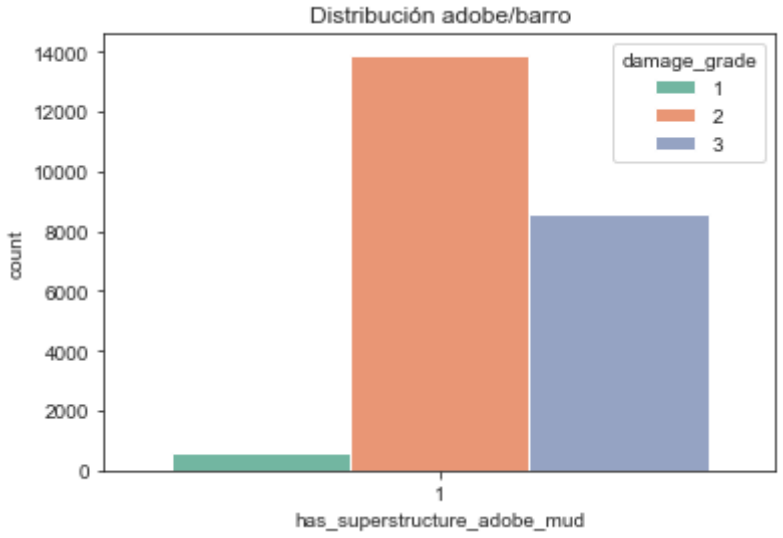


5.9 Variables binarias

Las variables binarias servirán para saber si dada una condición puede hacerse una distribución de daño para esa distribución. Por ejemplo las casa de barro piedra son estadísticamente muchas y tienen una relación de daños específica.

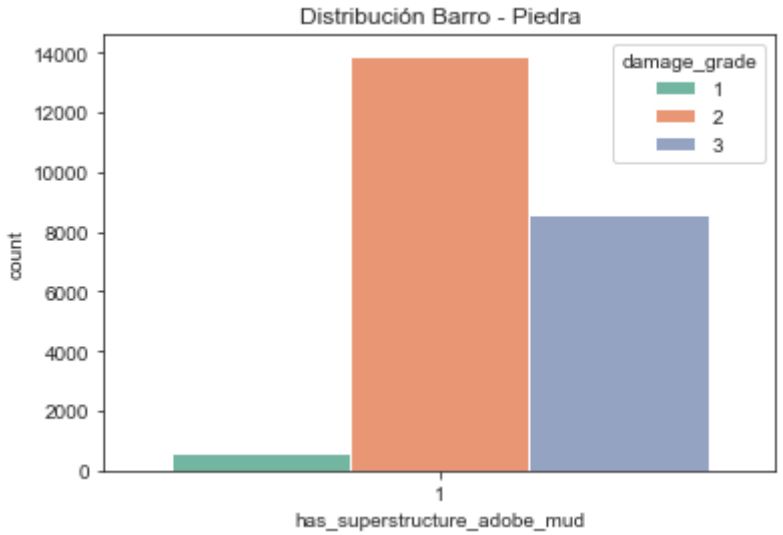
Distribución que indica si la edificación fue construida con adobe/barro.

Variable (tipo: binario) que indica si la edificación fue construida con adobe/barro.



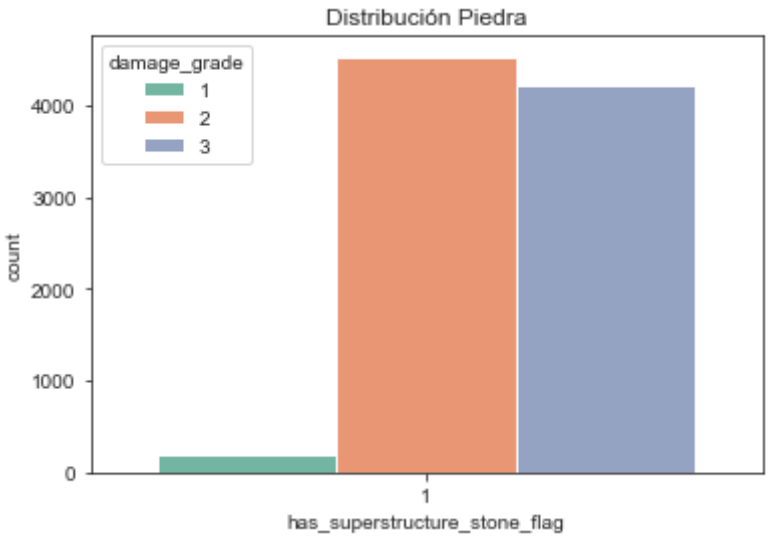
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON BARRO - PIEDRA.

Variable (tipo: binario) que indica si la edificación fue construida con Barro - Piedra.



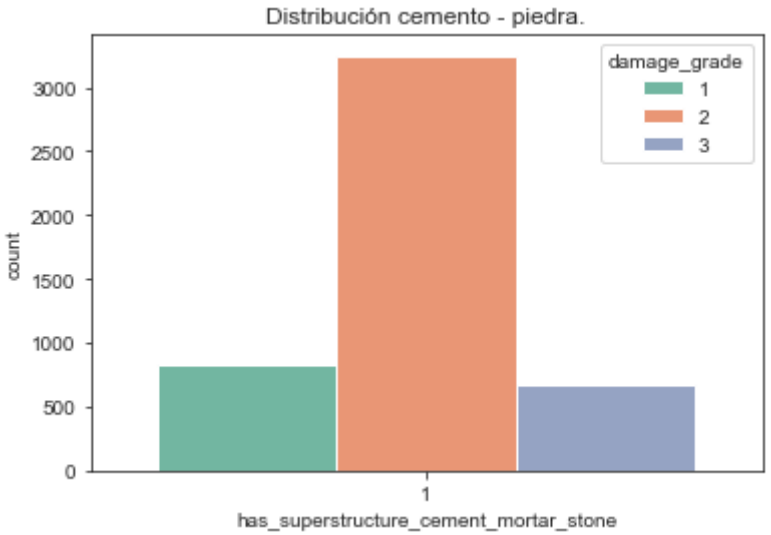
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON PIEDRA

Variable (tipo: binario) que indica si la edificación fue construida con Piedra.



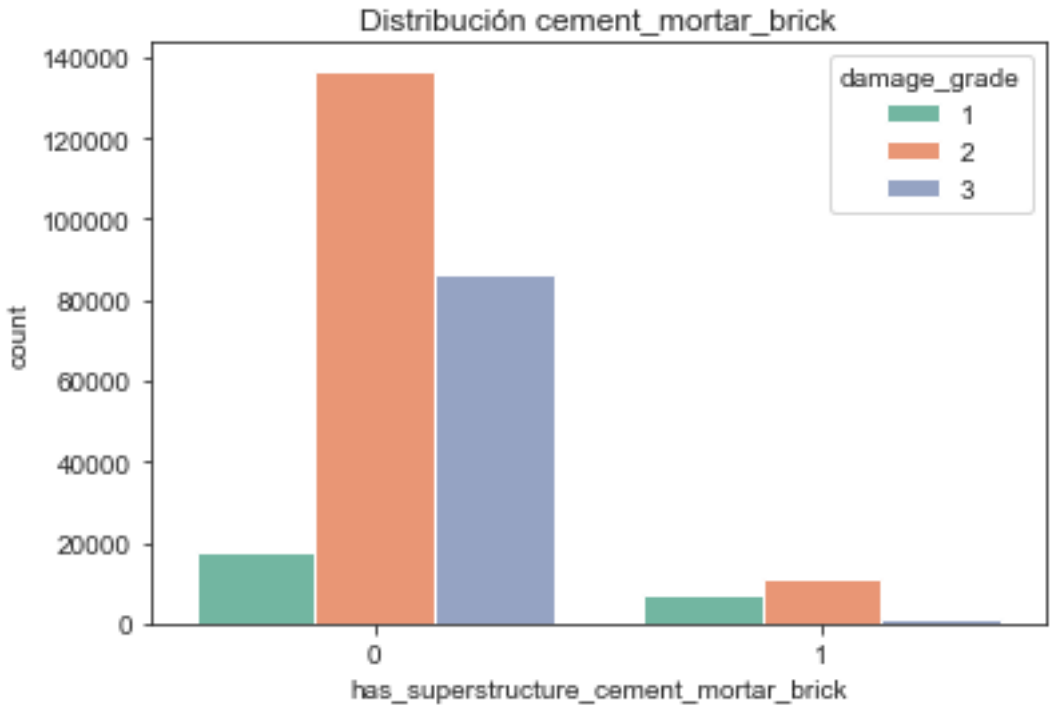
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON CEMENTO - PIEDRA.

Variable (tipo: binario) que indica si la edificación fue construida con cemento - piedra.



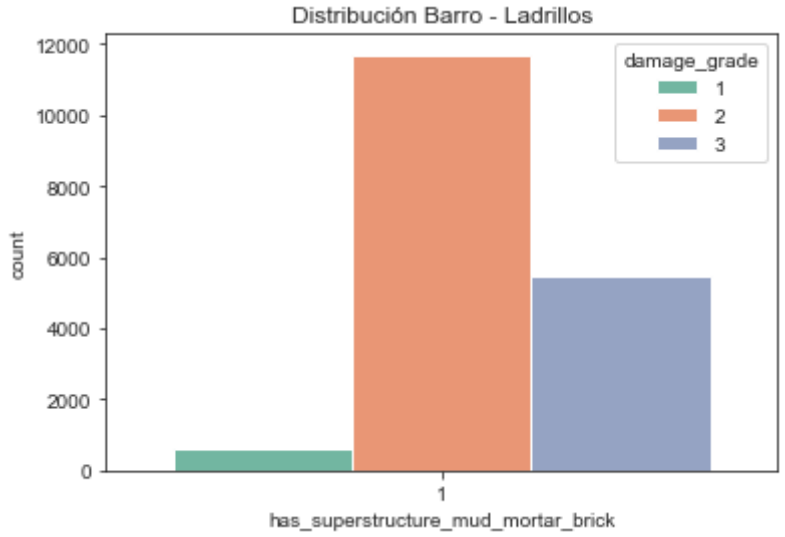
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON CEMENTO - LADRILLOS.

Variable (tipo: binario) que indica si la edificación fue construida con cemento - ladrillos.



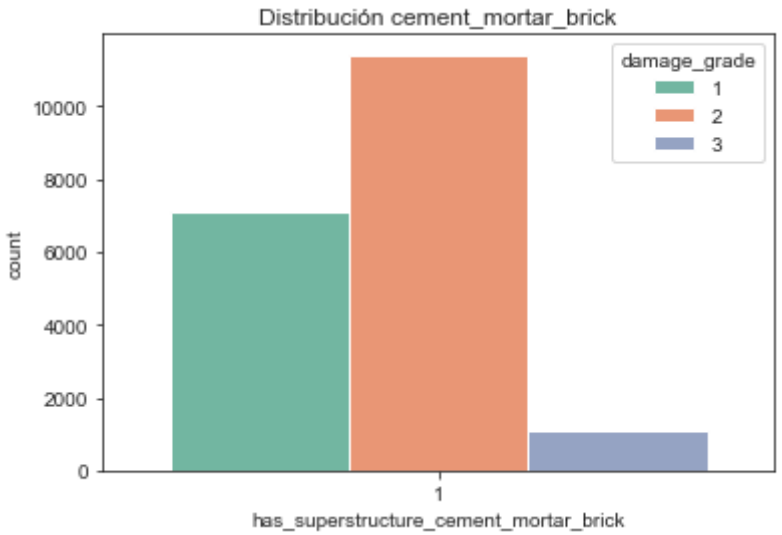
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON BARRO/LADRILLOS.

Variable (tipo: binario) que indica si la edificación fue construida con barro/ladrillos.



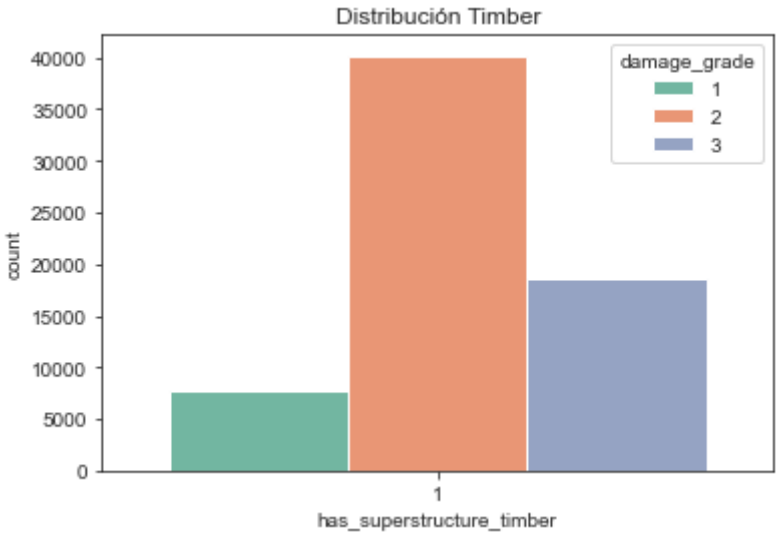
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON CEMENTO LADRILLOS.

Variable (tipo: binario) que indica si la edificación fue construida con cemento/ladrillos.



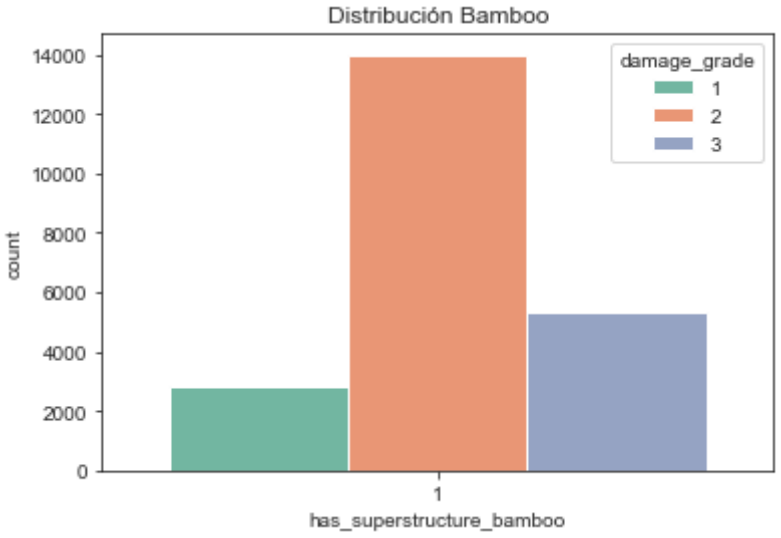
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON TIMBER (MADERA ESPECÍFICA PARA LA CONSTRUCCIÓN).

Variable (tipo: binario) que indica si la edificación fue construida con Timber.



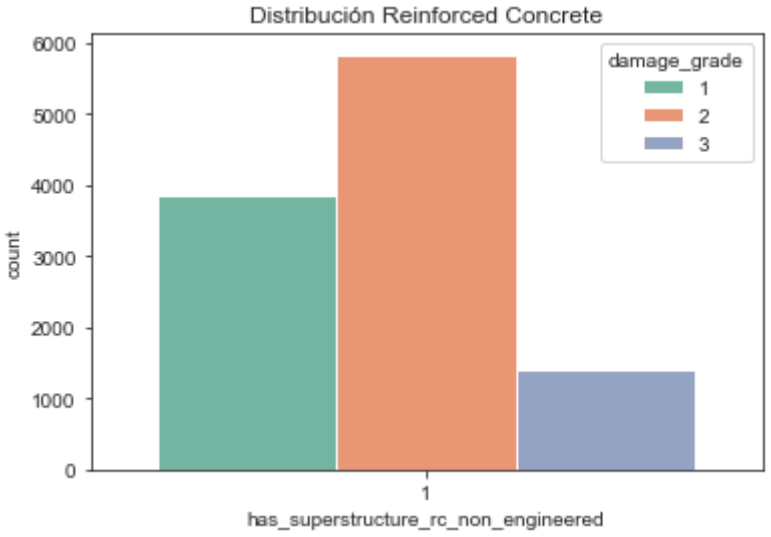
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON BAMBÚ (CAÑA).

Variable (tipo: binario) que indica si la edificación fue construida con Bambú.



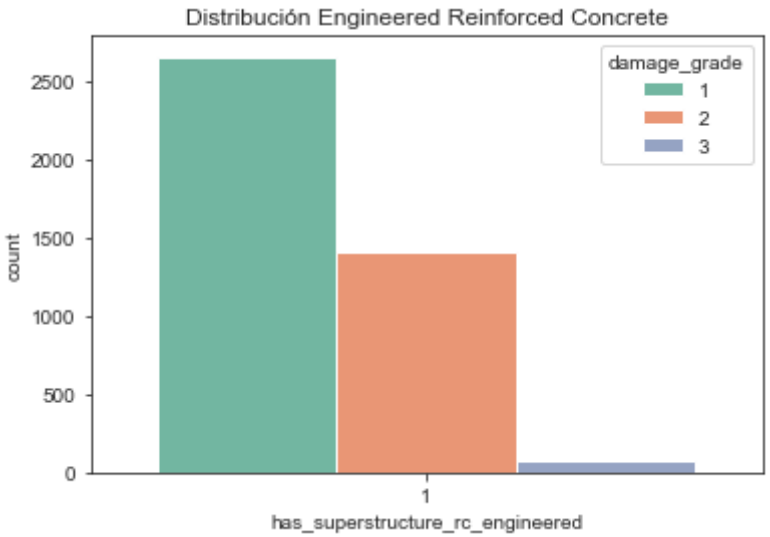
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON CONCRETO REFORZADO NO-DISEÑADO.

Variable (tipo: binario) que indica si la edificación fue construida con concreto reforzado no-diseñado.



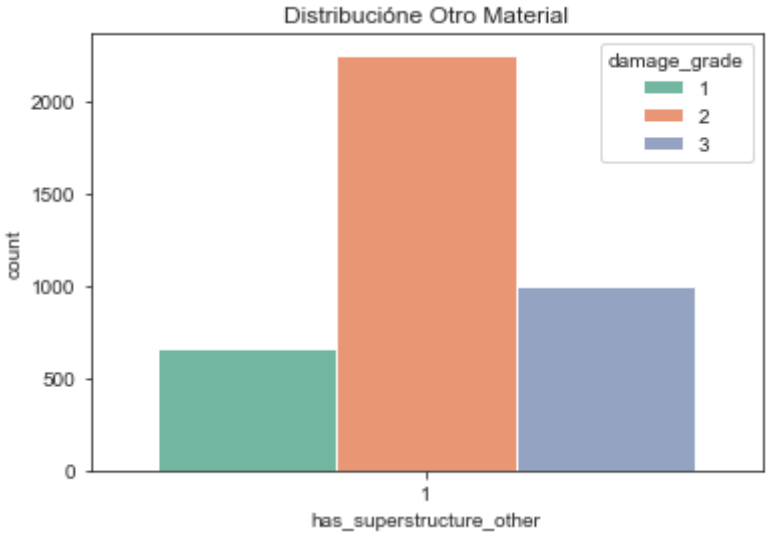
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON CONCRETO REFORZADO DISEÑADO.

Variable (tipo: binario) que indica si la edificación fue construida con concreto reforzado diseñado.



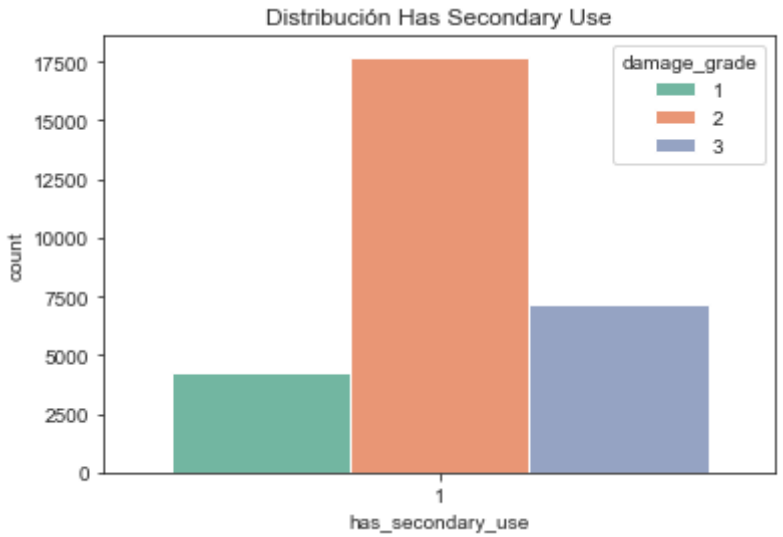
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN FUE CONSTRUIDA CON OTRO MATERIAL.

Variable (tipo: binario) que indica si la edificación fue construida con otro material.



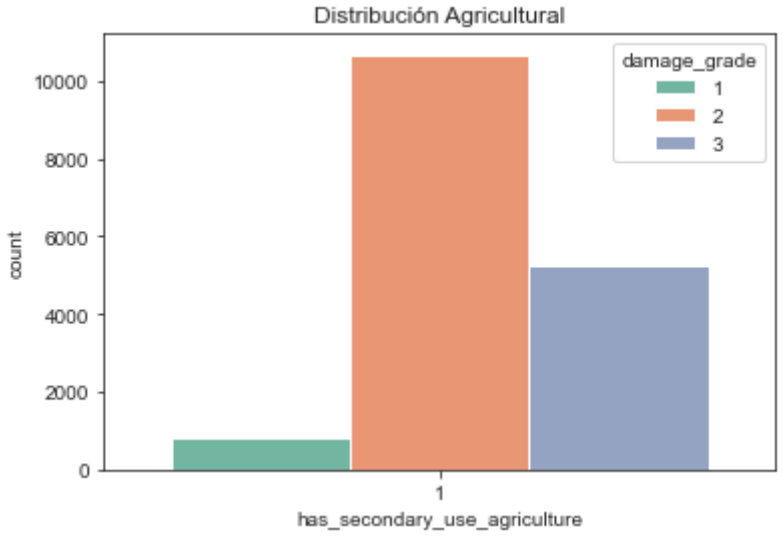
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA CON UN USO SECUNDARIO.

Variable (tipo: binario) que indica si la edificación era usada con un uso secundario.



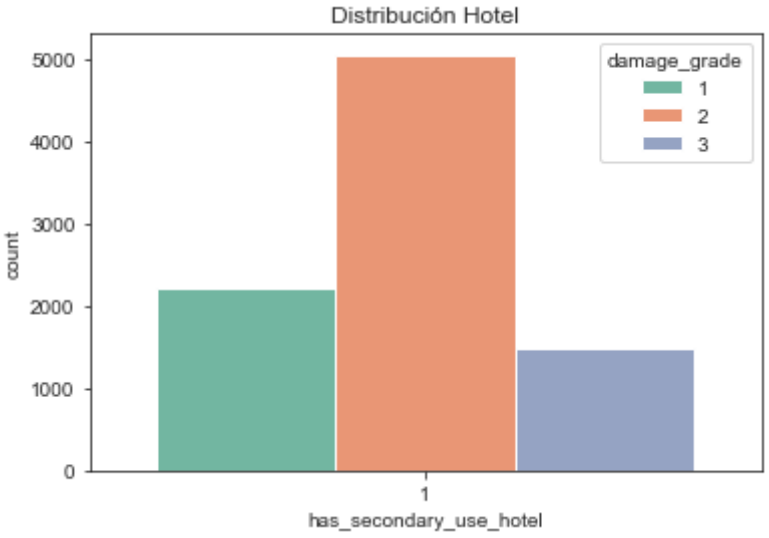
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA CON PROPÓSITOS DE AGRICULTURA.

Variable (tipo: binario) que indica si la edificación era usada con propósitos de agricultura.

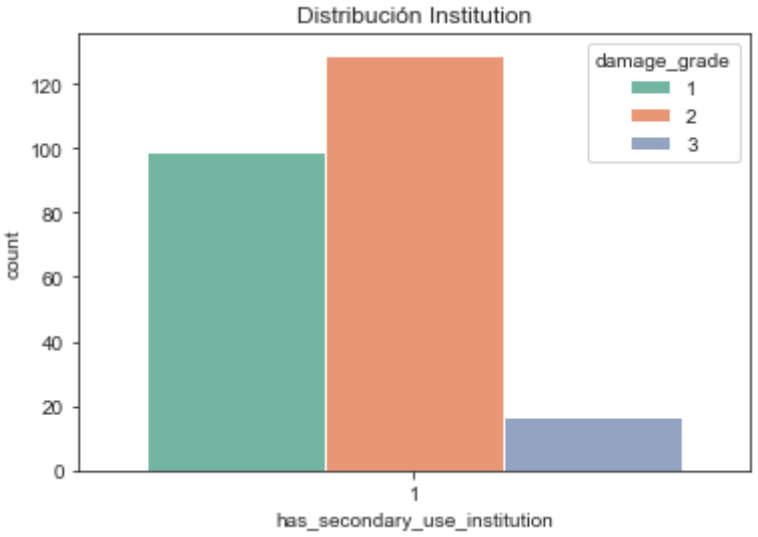


DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA COMO HOTEL.

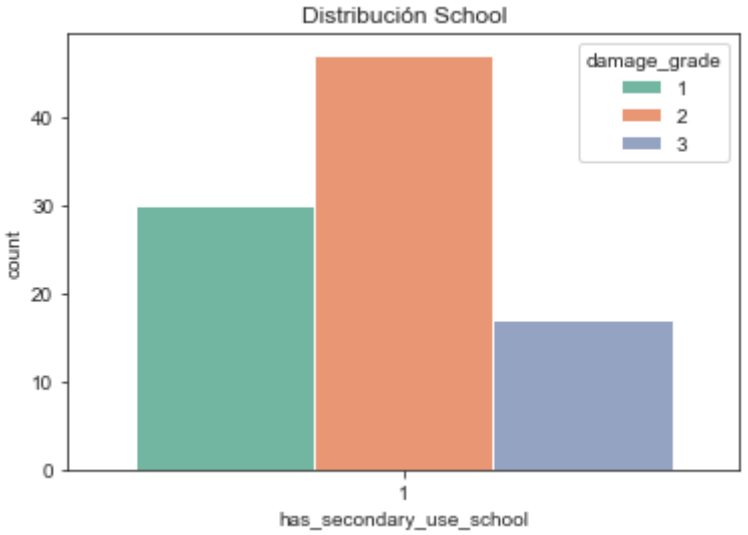
Variable (tipo: binario) que indica si la edificación era usada como Hotel.



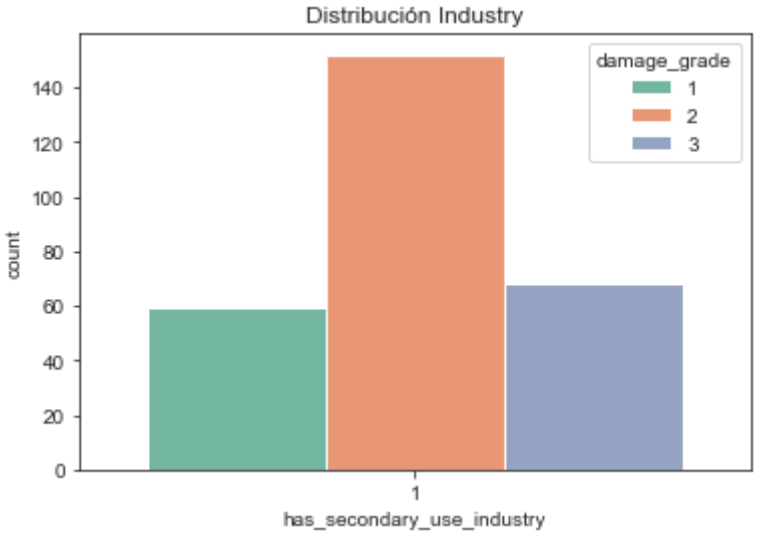
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN SE USABA COMO INSTITUCIÓN



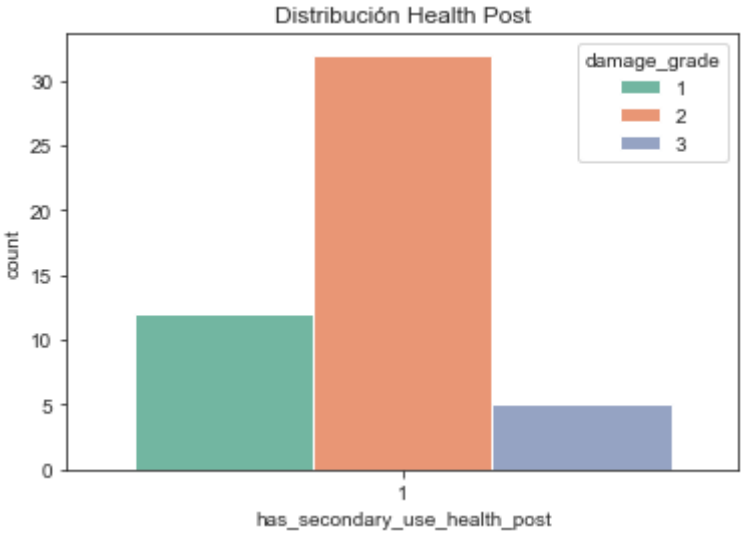
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA COMO ESCUELA.



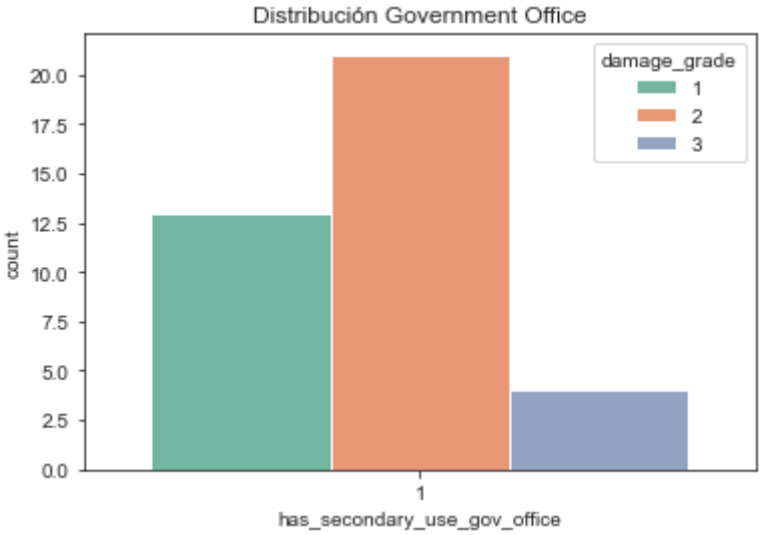
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA CON PROPÓSITOS INDUSTRIALES.



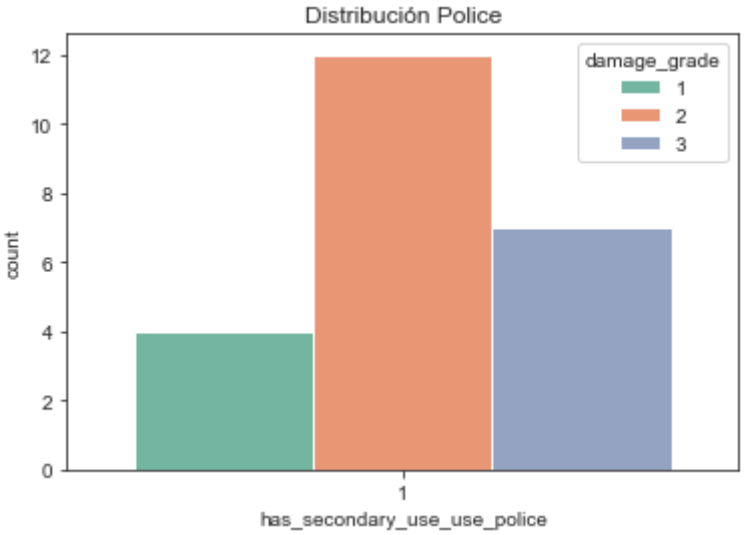
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA COMO PUESTO DE SALUD.



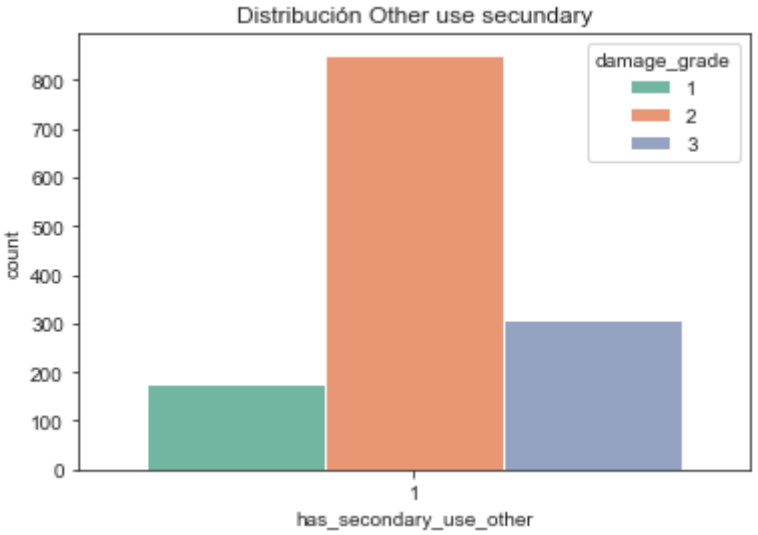
DISTRIBUCIÓN QUE INDICA SI LA EDIFICACIÓN ERA USADA COMO OFICINA DE GOBIERNO.



Distribución que indica si la edificación era usada como estación de policía.

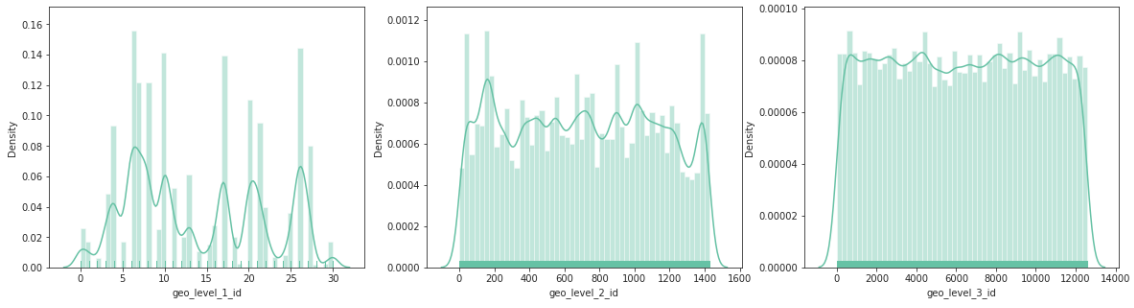


Distribución que indica si la edificación era usada con otro uso secundario.



5.10 Distribución de las zonas geográficas

Podemos observar que el nivel 2 y 3 se acerca a una distribución uniforme.



Teniendo en cuenta que los niveles 2 y 3 proporcionan ya menos información lo más adecuado para trabajar será el nivel 1. Con este nivel se podrían hacer consideraciones de daño más generales y será más fácil modelar un predictor.

6 CONCLUSIÓN

En este análisis exploratorio se pudieron identificar las variables y su incidencia en el daño. Las información más destacada es la distribución intrínseca de cada uno de los categóricos y binarios. Si se toman los binarios con mayor número de casos y los categóricos con mayor números de casos se puede cubrir estadísticamente la mayoría de los edificios. Con esta información y haciendo un cruce con los datos de edad y geográficos se espera obtener la suficiente información para poder modelar el nivel de daño que podría sufrir un edificio cualquiera.

Como conclusiones preliminares:

-Hay algunos tipos de edificaciones cuyas distribuciones de daño son distintas a la media como los edificios gubernamentales cuyo daño grave es menos frecuente que la media.

-Los índices geográficos de nivel 2 y 3 pierden información por su alto nivel de detalle pero el índice de nivel 1 puede servir para hacer una discriminacion adecuada.

-Dentro de las variables categóricas hay algunas que tienen una gran cantidad de casos y cuya distribución podría utilizarse como variable para ayudar en el modelado. Por otro lado las categorías que no tienen muchos casos suelen tener muy pocos y su capacidad estadística sumada a que no sabemos que significan, lo cual impide que tengamos un criterio humano para darles particular importancia, hace que sean datos que no deberíamos tener en cuenta.

-En la geometría del edificio la incidencia de casos puede verse discriminada con respecto a la altura de la edificación.

-Los materiales tienen en algunos casos distribuciones distintas a la media, en los casos en los que la cantidad de casos sea suficiente podría ser otra variable a utilizar en la búsqueda de predicciones.

Todo esto es un estudio preliminar que se buscará completar cuando se realice el predictor.