Universidad de Buenos Aires

Facultad de Ingeniería

1er Cuatrimestre de 2021

[75.06] Organización de Datos

Trabajo Práctico 1: Análisis Exploratorio de Datos

Integrantes: *Padrón:*

Tizziana Mazza Reta <tmazzar@fi.uba.ar> 101715
Mauricio Rodríguez Bertella <mrodriguezb@fi.uba.ar> 100624
Lucía Pardo <lpardo@fi.uba.ar> 99999
Nicolás Sanchez <nsanchez@fi.uba.ar> 98960

GitHub: Trabajo práctico 1 - Richter Predictor

Índice

1. Introducción

Para el presente trabajo práctico se busca realizar un análisis exploratorio sobre un dataset que contiene información recolectada luego de un terremoto de 7.8 en la escala Richter ocurrido en Nepal en 2015. Estos datos fueron recolectados por el National Planning Commission Secretariat of Nepal y constituye uno de los datasets post-desastre mayor recolectado.
A lo largo del desarrollo del informe buscaremos plantear preguntas y resolver dichos interrogantes para poder comprender mejor cómo sacar provecho de los datos. Para esto buscaremos relaciones que puedan existir entre las variables, tratando de encontrar correlaciones y dependencias que puedan servir para posterior uso y análisis.

2. Información acerca del Dataset

Para comenzar es importante tener una idea acerca de los datos brindados por el dataset y los posibles valores que éstos pueden tomar. A continuación vemos que los datos se encuentran es dos archivos csv llamados train_values.csv con 38 features que describen a la edificación identificada por un id y train_labels.csv con una variable 'damage_grade' cuyos valores consisten en 3 valores posibles:

- **1** daño menor
- **2** daño medio
- **3** daño mayor

y además contiene el correspondiente id de la edificación. A continuación numeramos las columnas y su contenido del archivo train_values.csv

- **building_id** (tipo: ID): identificador único de la edificación.
- **geo_level_1_id, geo_level_2_id, geo_level_3_id** (tipo: enteros): región geográfica en la cual la edificación existe, desde la más general (level 1) a la más específica (level 3). Valores posibles:

  level 1: 0-30,

  level 2: 0-1427,

  level 3: 0-12567.
- **count_floors_pre_eq** (tipo: entero): número de pisos en la edificación antes del terremoto.
- **age** (tipo: entero): antigüedad de la edificación en años.
- **area_percentage** (tipo: entero): superficie normalizada ocupada por la edificación.
- **height_percentage** (tipo: entero): altura normalizada ocupada por la edificación
- **land_surface_condition** (tipo: categórico): condición de la superficie terrestre donde el edificio fue construido. Valores posibles: n, o, t.
- **foundation_type** (tipo: categórico): tipo de cimientos usados cuando se construyó la edificación. Valores posibles: h, i, r, u, w.

- **roof_type** (tipo: categórico): tipo de techo usado cuando se construyó la edificación. Valores posibles: n, q, x.
- **ground_floor_type** (tipo: categórico): tipo de construcción usado en la planta baja cuando se construyó la edificación. Valores posibles: f, m, v, x, z.
- **other_floor_type**(tipo: categorical): tipo de construcción usado en otros pisos cuando se construyó la edificación (exceptuando el techo). Posibles valores: j, q, s, x.
- **position** (tipo: categórico): orientación de la edificación. Posibles valores: j, o, s, t.
- **plan_configuration** (tipo: categórico): formato de construcción de la edificación (para diseño sísmico). Valores posibles: a, c, d, f, m, n, o, q, s, u.
- **has_superstructure_adobe_mud** (tipo: binario): variable que indica si la edificación fue construida con adobe/barro.
- **has_superstructure_mud_mortar_stone** (tipo: binario): variable que indica si la edificación fue construida con barro - piedra.
- **has_superstructure_stone_flag** (tipo: binario): variable que indica si la edificación fue construida con piedra.
- **has_superstructure_cement_mortar_stone** (tipo: binario): variable que indica si la edificación fue construida con cemento - piedra.
- **has_superstructure_mud_mortar_brick** (tipo: binario): variable que indica si la edificación fue construida con barro - ladrillos.
- **has_superstructure_cement_mortar_brick** (tipo: binario): variable que indica si la edificación fue construida con cemento - ladrillos.
- **has_superstructure_timber** (tipo: binario): variable que indica si la edificación fue construida con Timber (madera específica para la construcción).
- **has_superstructure_bamboo** (tipo: binario): variable que indica si la edificación fue construida con Bambú (caña).
- **has_superstructure_rc_non_engineered** (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado no-diseñado.
- **has_superstructure_rc_engineered** (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado diseñado.
- **has_superstructure_other** (tipo: binario): variable que indica si la edificación fue construida con otro material.
- **legal_ownership_status** (tipo: categórico): estado legal de la tierra donde la edificación fue construida. Valores posibles: a, r, v, w.
- **count_families** (tipo: entero): número de familias que vivían en la edificación.
- **has_secondary_use** (tipo: binario): variable que indica si la edificación era usada con un uso secundario.

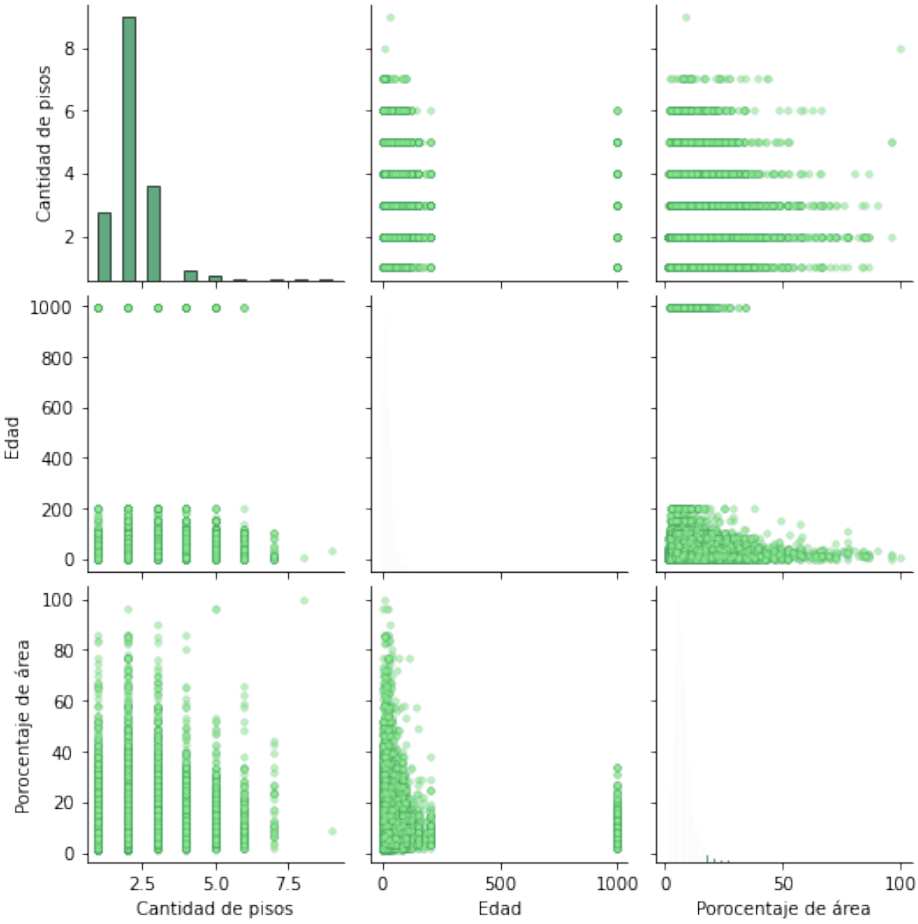
- **has_secondary_use_agriculture** (tipo: binario): variable que indica si la edificación era usada con propósitos de agricultura.
- **has_secondary_use_hotel** (tipo: binario): variable que indica si la edificación era usada como oficina de gobierno
- **has_secondary_use_rental** (tipo: binario): variable que indica si la edificación se alquilaba.
- **has_secondary_use_institution** (tipo: binario): variable que indica si la edificación era usada como sede de una institución.
- **has_secondary_use_school** (tipo: binario): variable que indica si la edificación era usada como escuela.
- **has_secondary_use_industry** (tipo: binario): variable que indica si la edificación era usada con propósitos industriales.
- **has_secondary_use_health_post** (tipo: binario): variable que indica si la edificación era usada como puesto de salud.
- **has_secondary_use_gov_office** (tipo: binario): variable que indica si la edificación era usada como oficina de gobierno.
- **has_secondary_use_use_police** (tipo: binario): variable que indica si la edificación era usada como estación de policía.
- **has_secondary_use_other** (tipo: binario): variable que indica si la edificación era usada con otro uso secundario.

3. Filtración de datos

Para comenzar el análisis verificamos que no haya campos faltantes o nulos. Asimismo verificamos y modificamos los tipos de categorías necesarios que se encuentren generalizados a un tipo de dato más específico para poder facilitar más tarde el uso de los mismos y poder ahorrar recursos. De esta forma se realizó una reducción del tamaño del archivo de 85mb a 14mb.

4. Análisis general

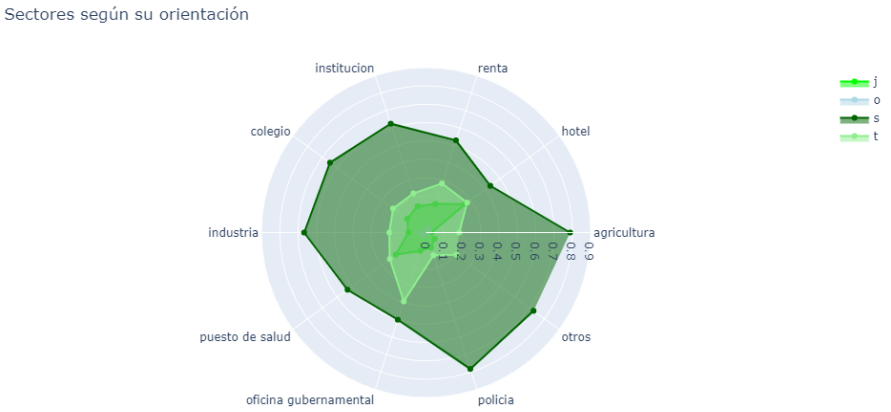
Para comenzar probamos hacer un Matrixplot para ver rápidamente como se relacionan las distintas variables entre sí.

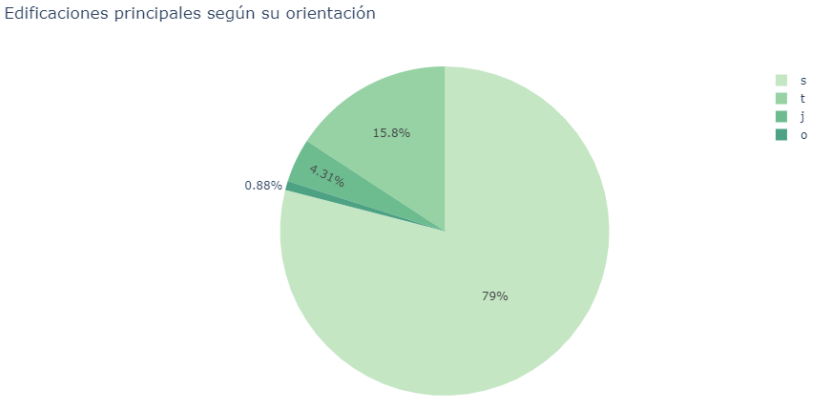


Con esto vimos que la mayoría de las variables categóricas y binarias no generan gráficos útiles. Los que si se destacan son entre las columnas numéricas (”age”,”height_porcentage”, etc).
La metodología inicial que usamos fue una reunión de equipo para generar preguntas iniciales, buscando *insights* interesantes, tratando de analizar qué variable tenía relación con cada una de las otras. Luego en un principio lo qué más capta nuestra atención siendo que hablamos de un terremoto es contrastar las variables contra el valor ”damage_grade”.

5. Análisis de orientación

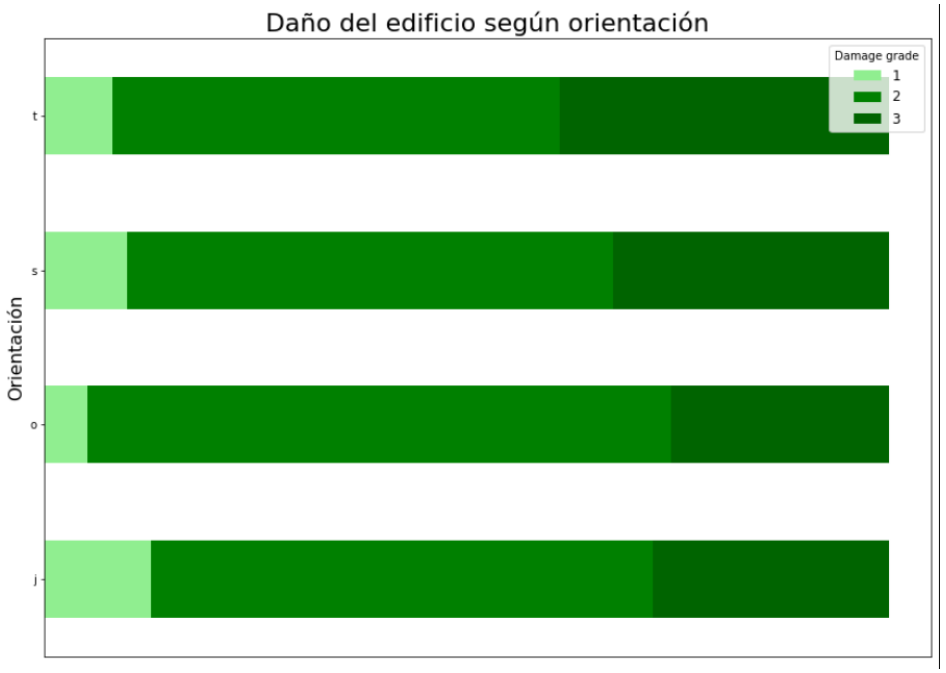
Para analizar el daño que tuvieron las edificaciones según su posicionamiento, primero decidimos ver si había algún tipo de diferenciación en cuanto al posicionamiento de uso principal y uso secundario.





Llama la atención que la mayoría de los edificios tiene posicionamiento "*s*" y que casi ninguno tiene posicionamiento "*o*".

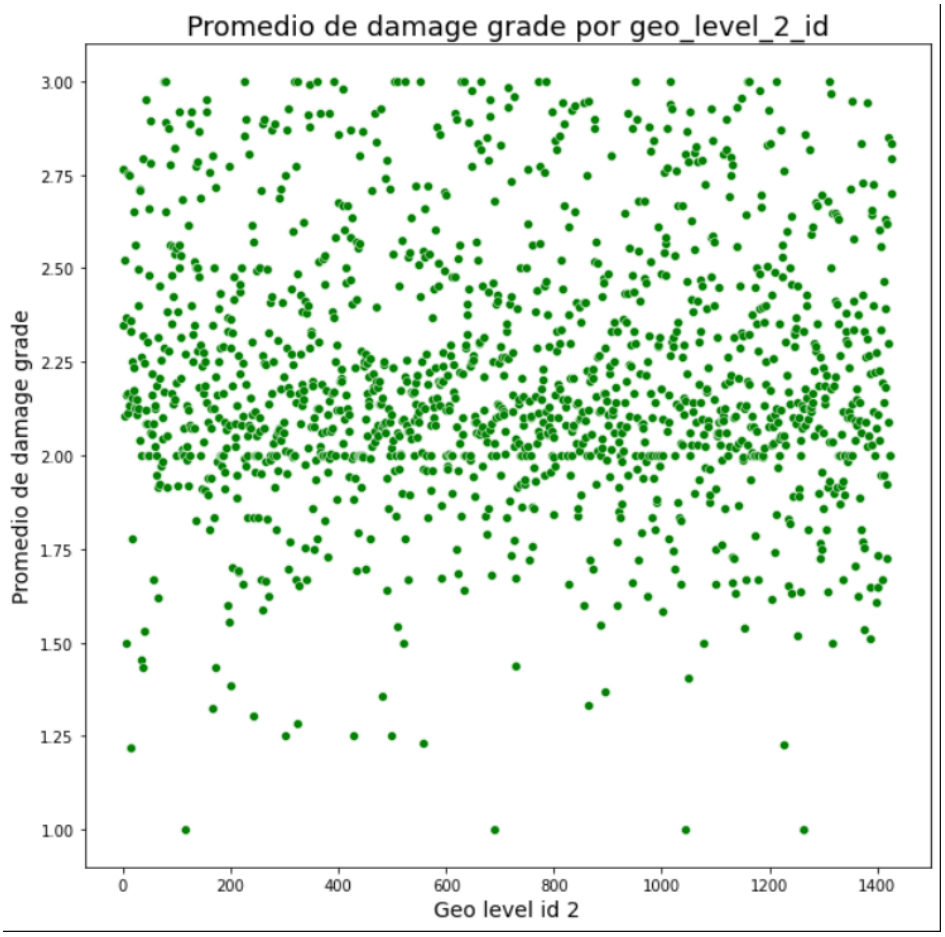
A continuación se muestra la orientación de los edificios con respecto a la cantidad de daño.



En cuanto a esta no se nota ninguna relación en particular, al parecer está el daño distribuido de forma uniforme como era de esperarse.

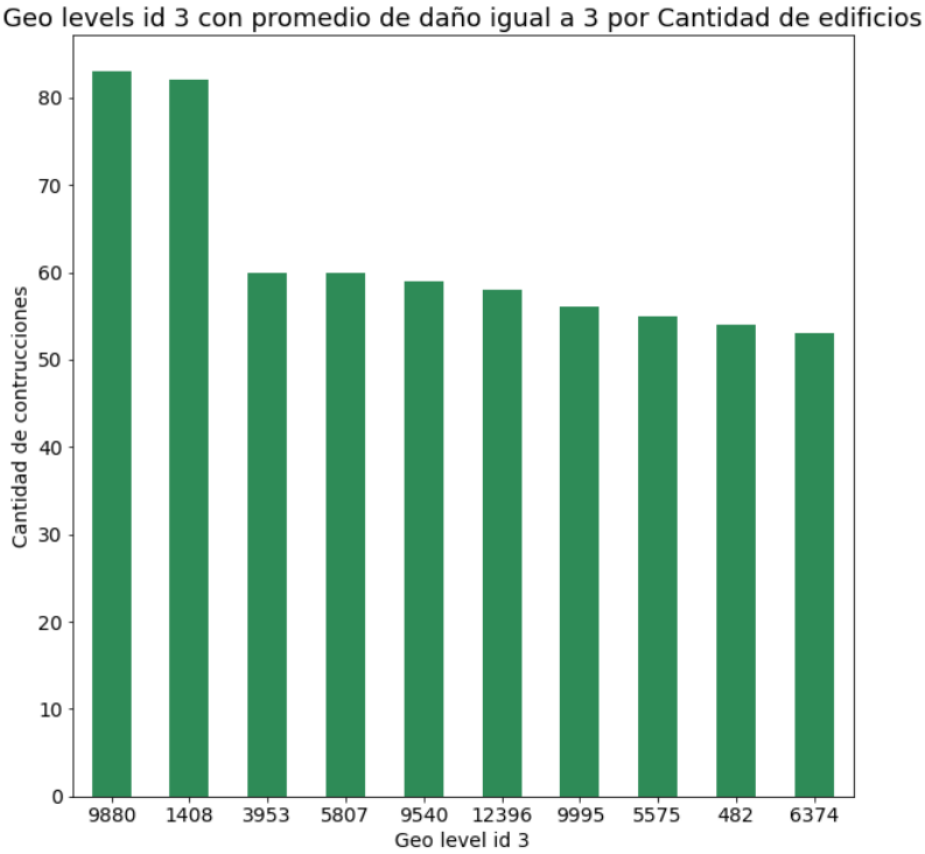
6. Análisis por geolevel

Tuvimos que indagar esta variable en particular y su significado. En base a eso nos preguntamos si podríamos encontrar las zonas más afectadas por el terremoto, pero dado que no necesariamente tienen relación los geo_level entre sí esto no fue posible. Otra relación que buscamos fue el daño promedio por geo_level



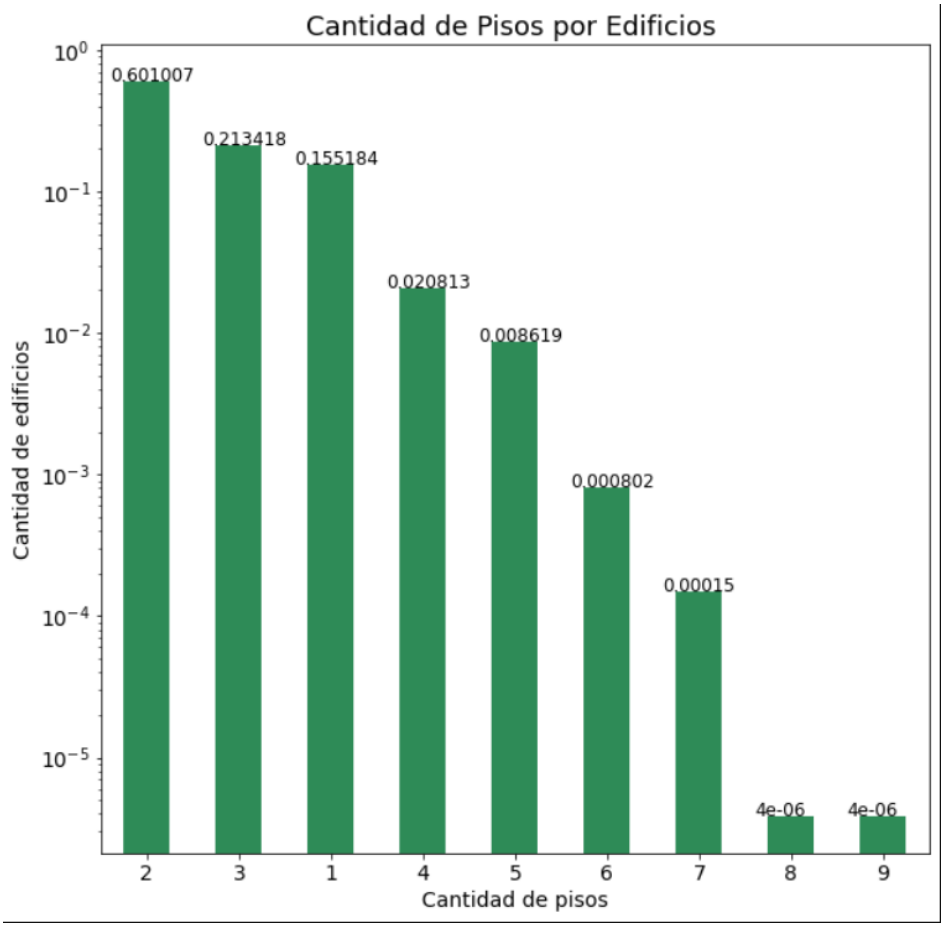
En este gráfico podemos ver que en la mayoría de los geolevel 2 el promedio de daño se encuentra entre 2 y 2.25.

Para ver los lugares más afectados buscamos los geolevel 3 con mayor promedio de daño. Estos son los 10 geolevel 3 con promedio de daño 3 según cuantos edificios tiene cada uno. De esta forma podemos pensar que los que más edificios tienen aún así manteniendo el promedio en 3 son los más perjudicados.

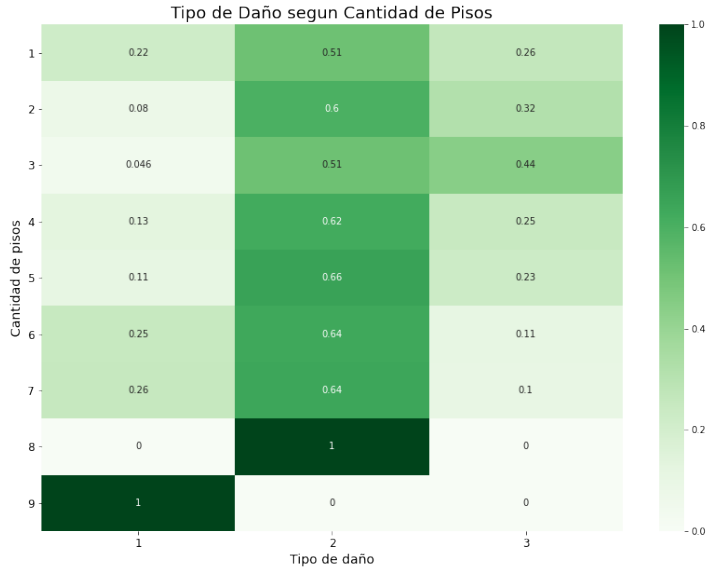


7. Análisis por cantidad de pisos

Analizamos el porcentaje de edificios que hay con cada cantidad de pisos

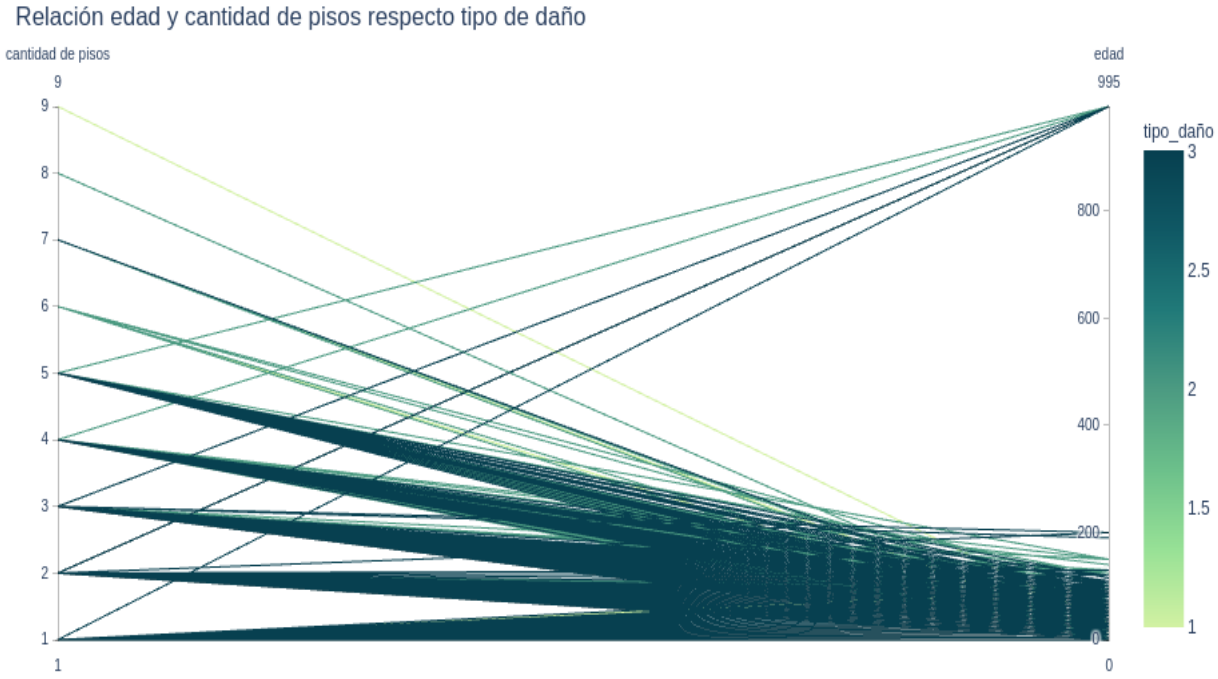


En esta sección lo que uno piensa en primera instancia es que los edificios con menos pisos se verán menos afectados, y que los más altos sufrirán más daños. Pero al parecer, quizás por previsión de que un edificio más alto debe tener una mejor estructura, en los edificios más altos no hay demasiados daños de tipo 3. Incluso en el único edificio de 9 pisos el daño es de tipo 1. En general los más afectados fueron los de 3 pisos.



Además nos preguntamos qué relación podía existir respecto el tipo de edad y la cantidad de pisos construidos, para esto nos pareció que un gráfico de tipo parallel podía

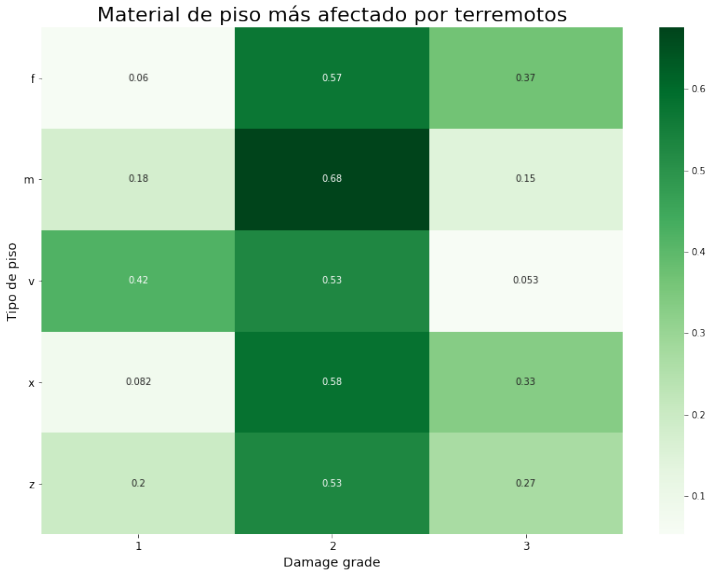
representar correctamente dicha pregunta obteniendo lo siguiente:

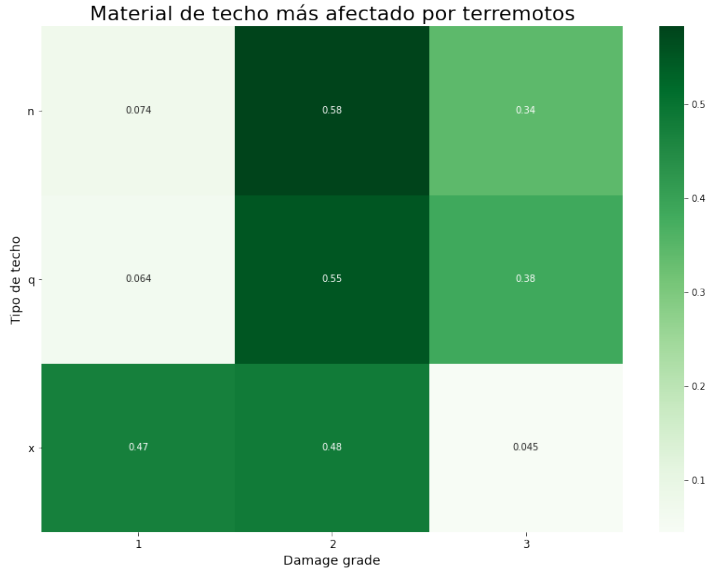


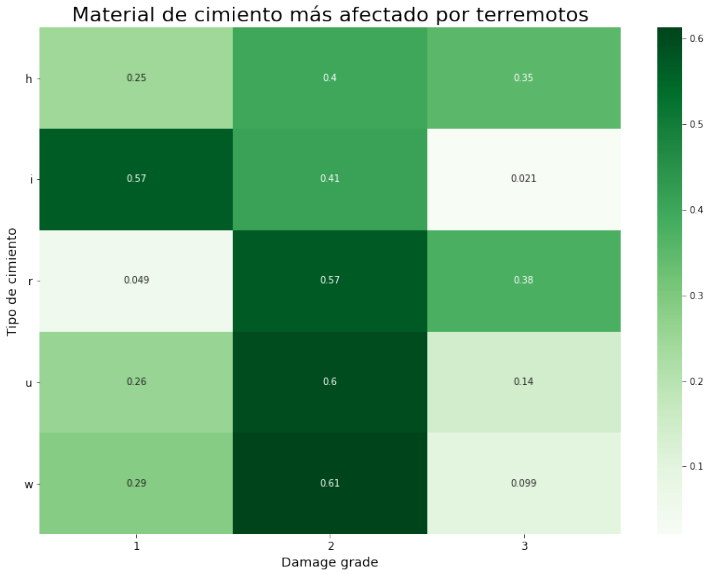
Mostrando que hay una gran diferencia respecto a las edades de los edificios en el dataset, probablemente debido a que una vez que los edificios eran muy antigüos se catalogaban con la edad de 995 sin reparar en edades intermedias. Además podemos observar que los edificios más antigüos poseían una menor cantidad de pisos y respecto a los edificios más nuevos.

8. Análisis por material

Una de las probabilidades con las que pensábamos encontrarnos es encontrar que ciertos materiales son más resistentes que otros, o que ciertas estructuras más antiguas tendrían materiales más viejos que los de hoy en día.
En principio buscamos cuáles fueron los materiales más afectados.

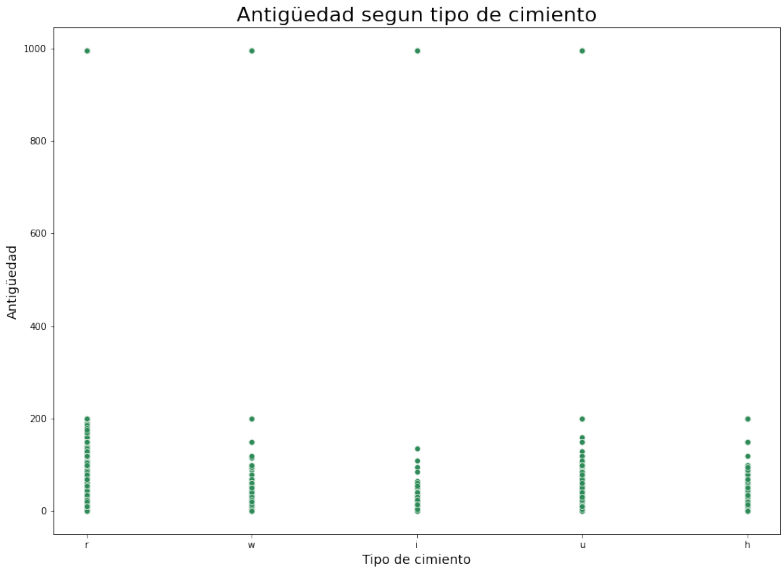

Material de piso más afectado por terremotos
Tipo de piso
f
m
v
x
z
0.06
0.57
0.37
0.18
0.68
0.15
0.42
0.53
0.053
0.082
0.58
0.33
0.2
0.53
0.27
1
2
3
Damage grade
0.6
0.5
0.4
0.3
0.2
0.1

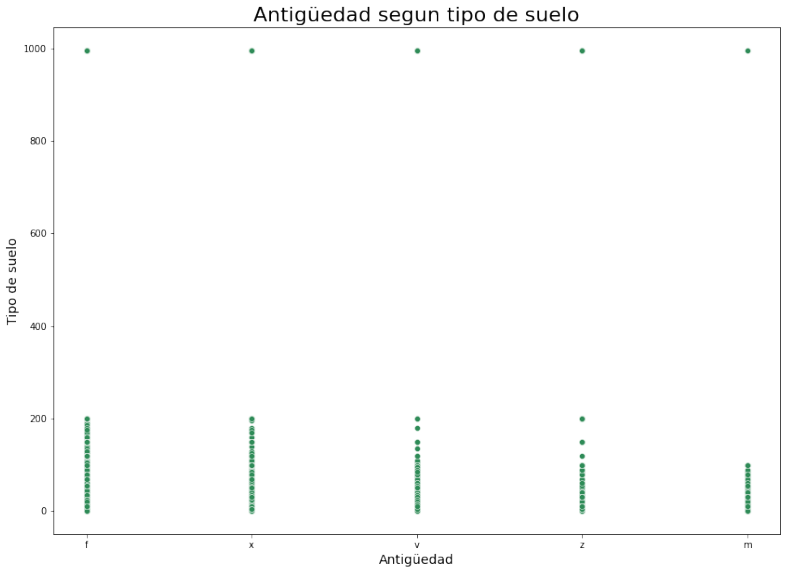

Material de techo más afectado por terremotos
Tipo de techo
n
q
x
0.074
0.58
0.34
0.064
0.55
0.38
0.47
0.48
0.045
1
2
3
Damage grade
0.5
0.4
0.3
0.2
0.1

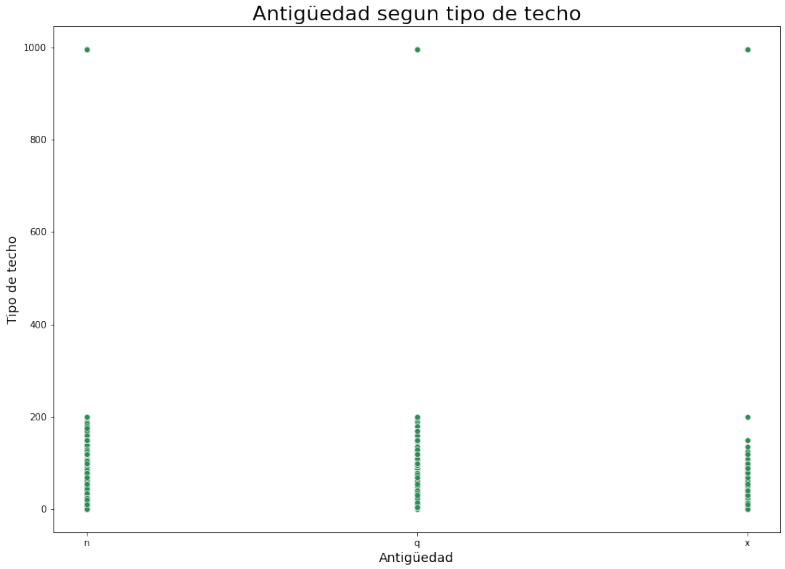


Se puede ver que en particular hay 3 materiales específicos que recibieron mayor cantidad de daño por gran diferencia, son: para el techo el material "q", para los cimientos el material "r" y para los pisos el material "f".
Podemos ver también como es razonable que los edificios más desprotegidos (más afectados) son los que casi no reciben daño de tipo 1, y viceversa, los edificios más protegidos son los que más reciben daño de tipo 1 y casi nada de tipo 3.

En cuanto a la antigüedad con respecto a los tipos de materiales, encontramos que de lo contrario a lo que suponíamos, parece estar distribuido de forma uniforme el uso de los materiales. No hay ningún material que sobresalga en alguna antigüedad en particular.



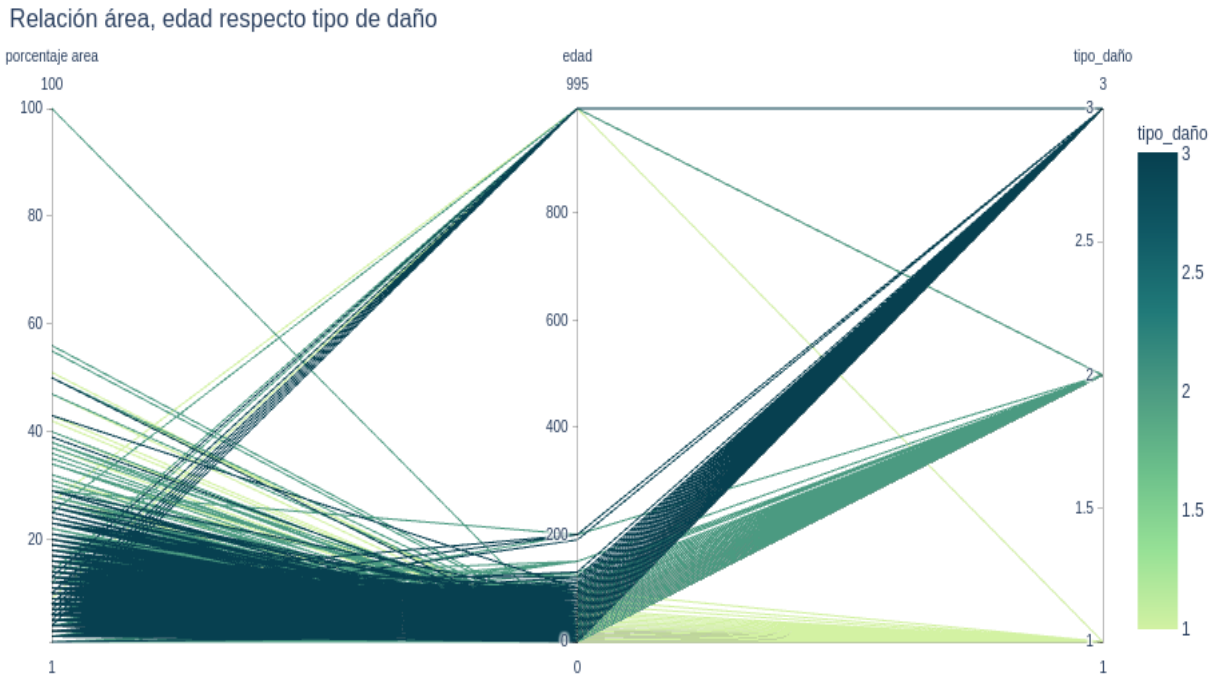




9. Análisis por área y edad

En esta sección analizamos las diferentes variables presentadas en relación al área y la edad de las edificaciones.

En el siguiente gráfico podemos ver, además de que la gran mayoría de los edificios tienen como máximo de 200 años, es que los que tienen mayor edad tienen áreas muy pequeñas en comparación a los mas nuevos, que tienen porcentaje de área mas variado. Las edificaciones mas antiguas no tienen ni la mitad del área mas grande.

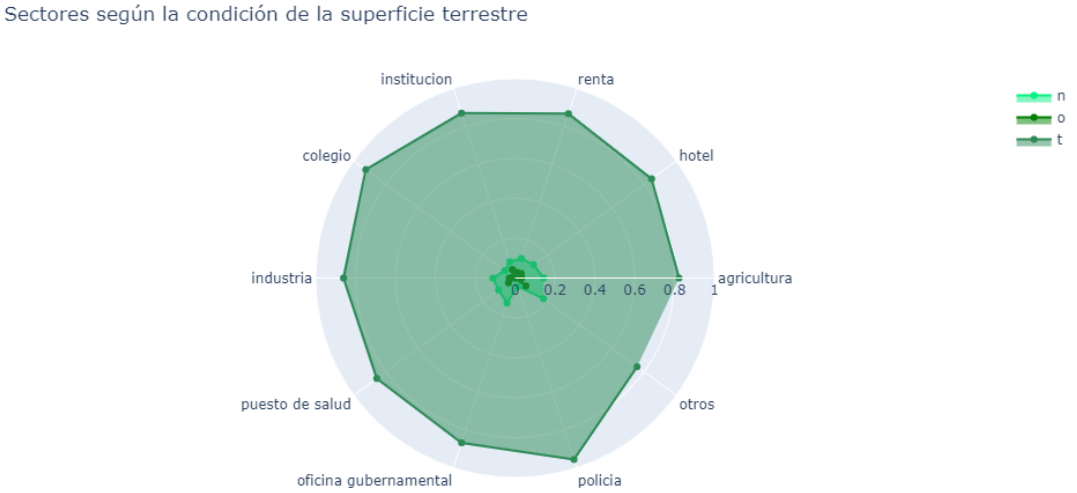


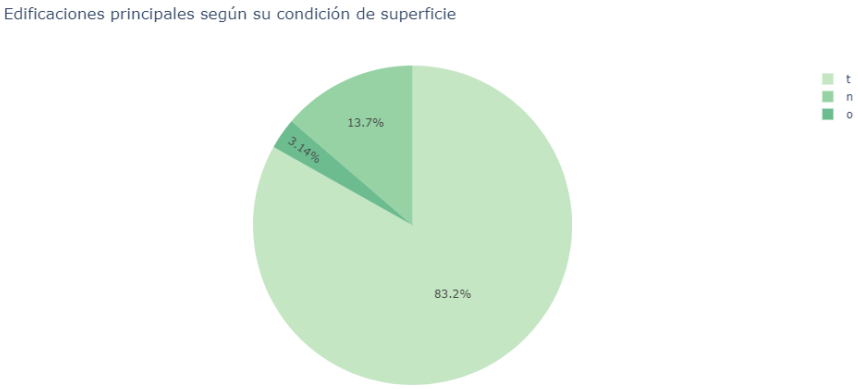
Por último, el gráfico muestra el tipo de daño que tuvo cada edificio mediante una escala de colores. Se puede apreciar que los de menor porcentaje de área ocupada sufrieron un gran tipo de daño por parte del terremoto, ya que tienen el color más oscuro de todos, el cual representa un gran tipo de daño. Los que están representados como de 995 años, o sea los más antiguos, tienen, en su mayoría, sus líneas del color más claro, lo que significa un menor daño.

10. Análisis de estructura de edificación

Una de las dudas es saber si las edificaciones de uso principal y de segundo uso difieren en cuanto a su construcción. Queremos ver si fueron construidas con menor resistencia a la hora de haber un terremoto. A continuación se mostrarán gráficos que relacionan los materiales de la edificación con los edificios.

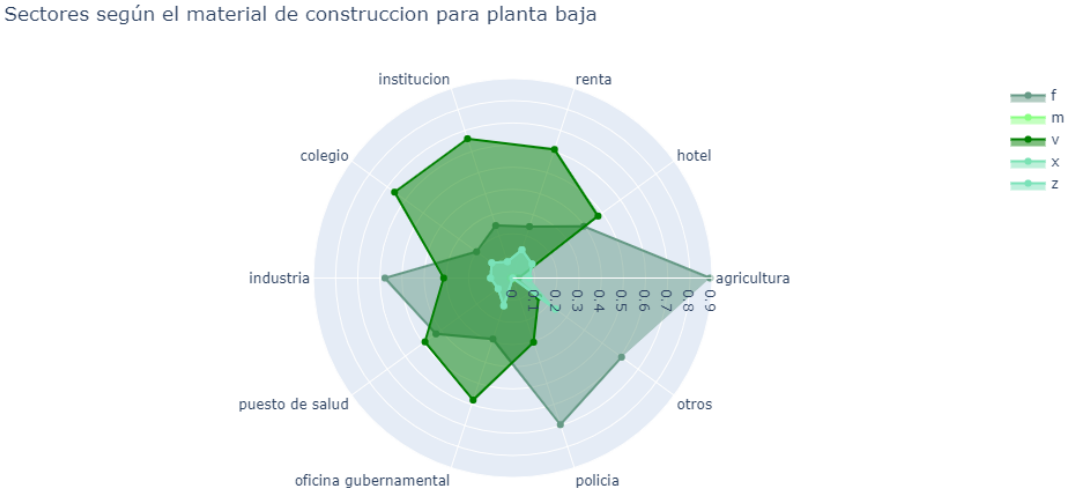
Edificaciones según su condición de superficie terrestre.

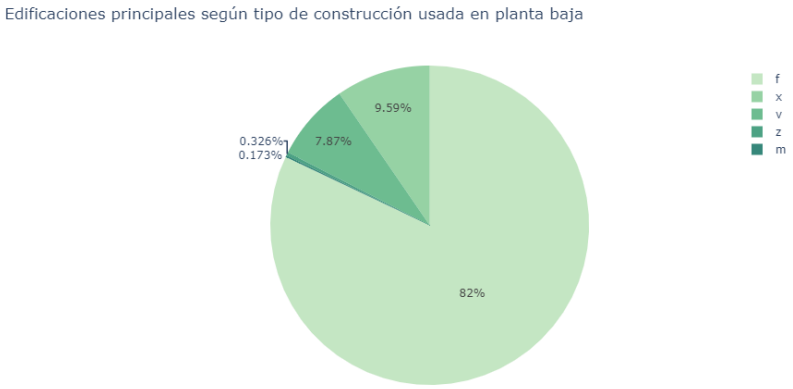




La condición de la superficie terrestre en las edificaciones presentan en su mayoría un tipo n, sin discriminar el uso.

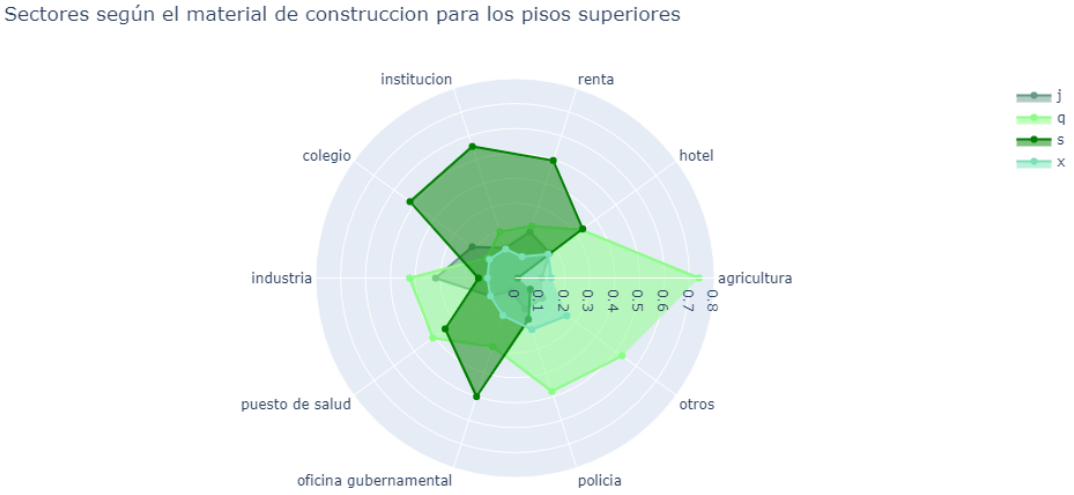
Edificaciones según el tipo de construcción ultilizado en la planta baja.





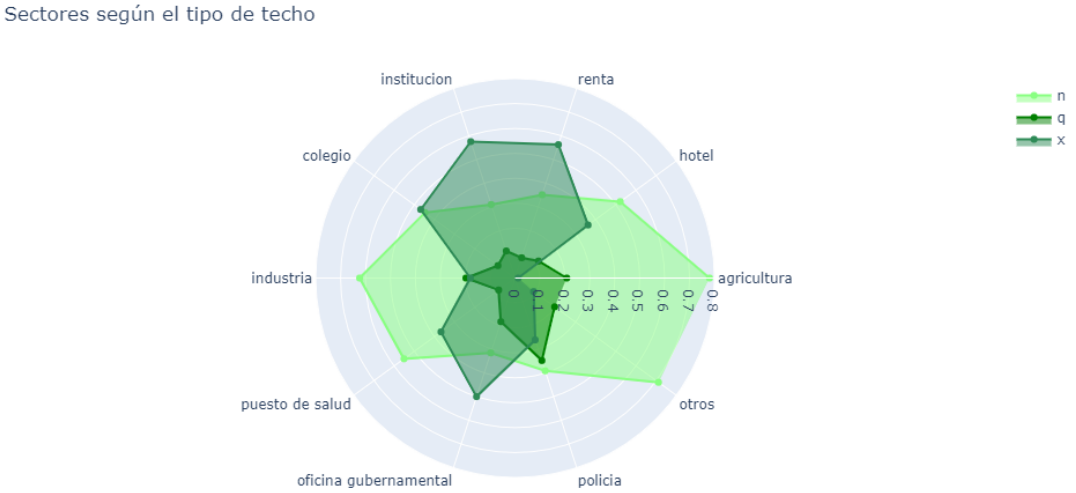
En la planta baja de los secundarios, además de utilizarse el tipo de construcción f, predomina bastante el tipo v, cosa que no sucede en las edificaciones de uso principal.

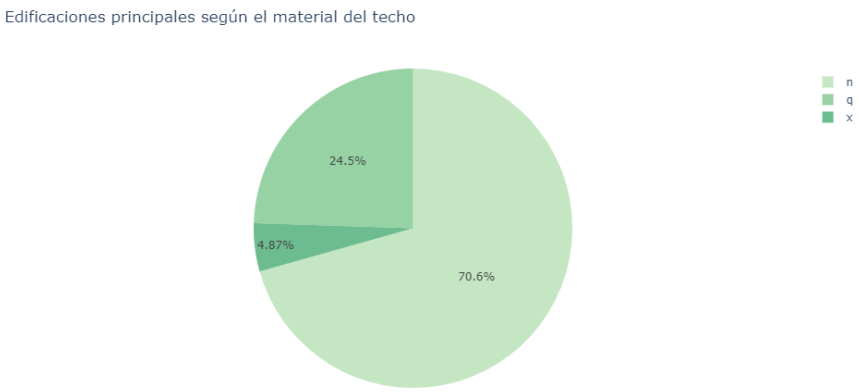
Edificaciones según el tipo de construcción utilizado en los pisos superiores.



En el tipo de material utilizado para los pisos superiores, en los de uso secundarios se observa una tendencia al tipo q y s, lo cual no ocurre en los principales, ya que s es el que menos se utiliza para este sector.

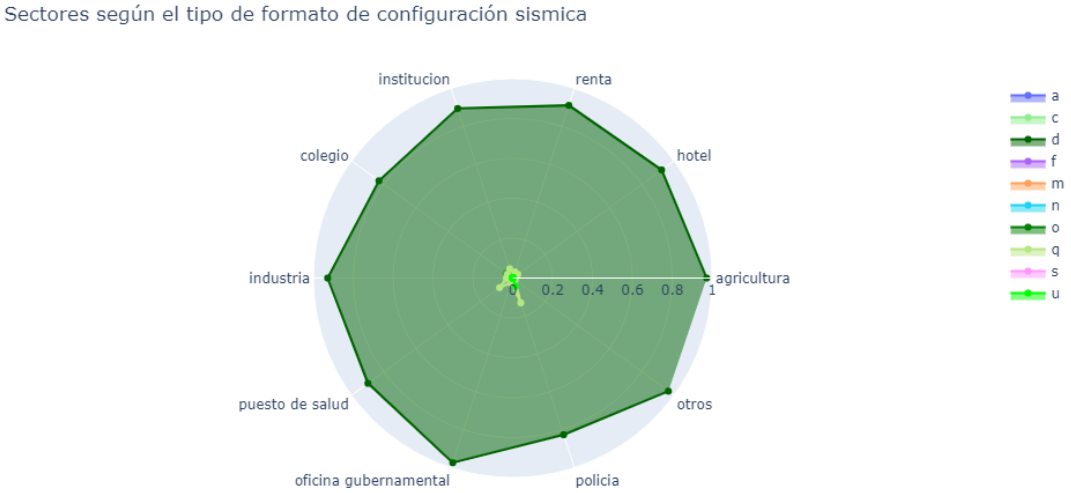
Edificaciones según el tipo de construcción utilizado en el techo.

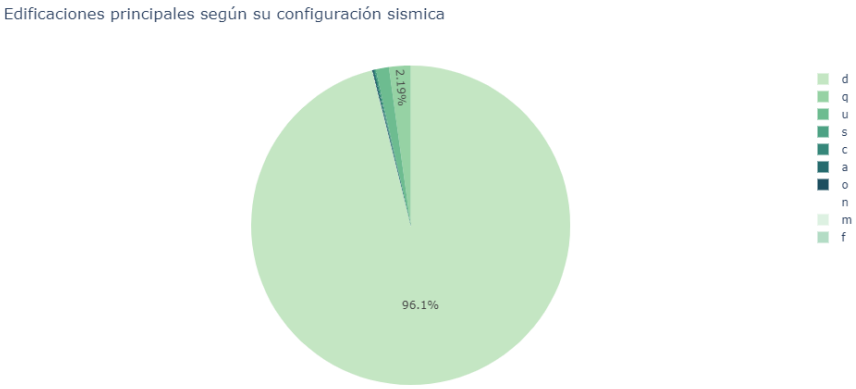




Lo mismo sucede con el tipo de construcción de los techos de ambos. Se ve, indudablemente que el tipo n aparece con más frecuencia en las construcciones, pero en las de segundo uso le sigue los de tipo x, y en los principales es el menos usado.

Edificaciones según el tipo de formato de configuración para diseño sismico utilizado.

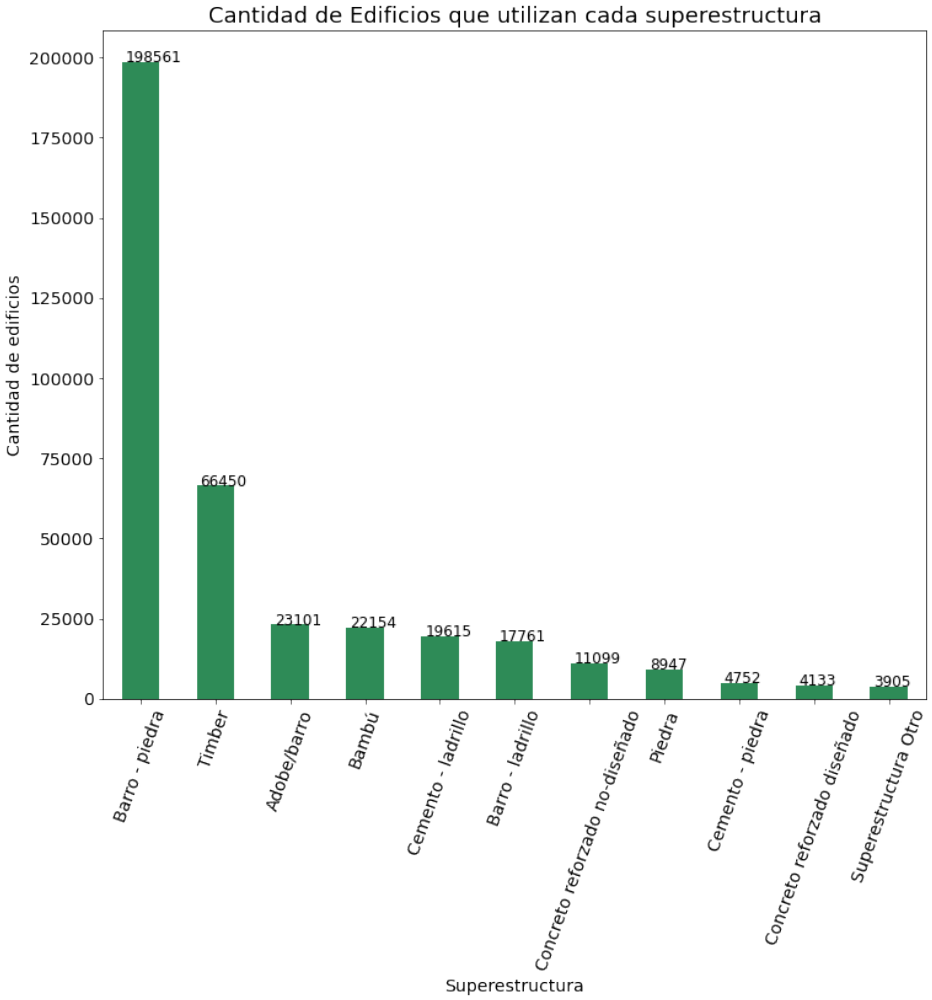




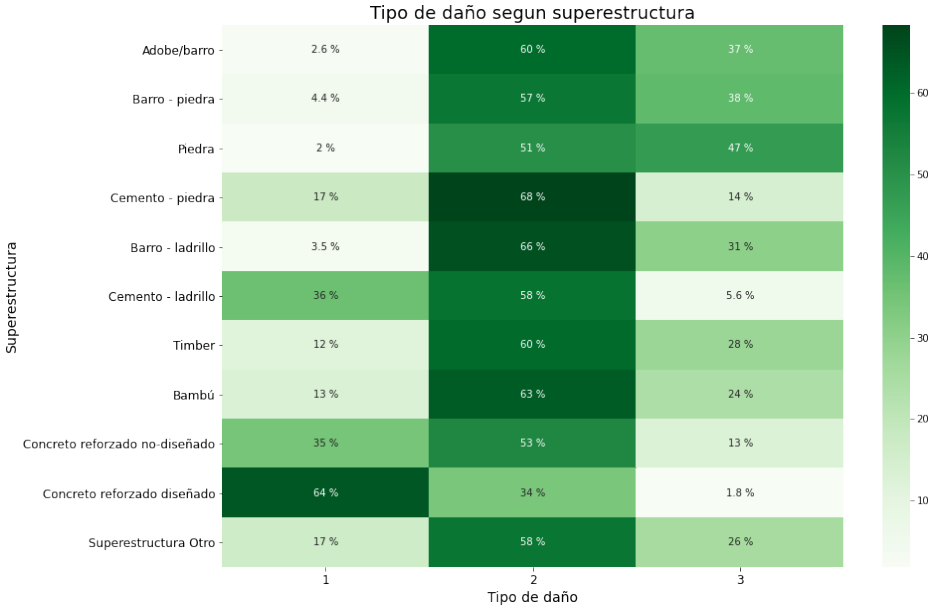
La gran mayoría de edificaciones del país esta compuesta por una configuración sísmica de tipo *d*.
Como se puede apreciar, las construcciones no difieren mucho, ya sean de uso principal como de uso secundario.

11. Análisis por superestructura

En nuestro set de datos podemos ver que materiales fueron utilizados para su construcción (superestructura). Es interesante preguntarnos si el daño que sufrió la edificación se ve afectado por los materiales utilizados para su construcción. En un primer análisis podemos ver que las edificaciones pueden tener mas de un tipo de superestructura, es decir, varios materiales utilizados para su construcción, donde los mas frecuentes son los siguientes:



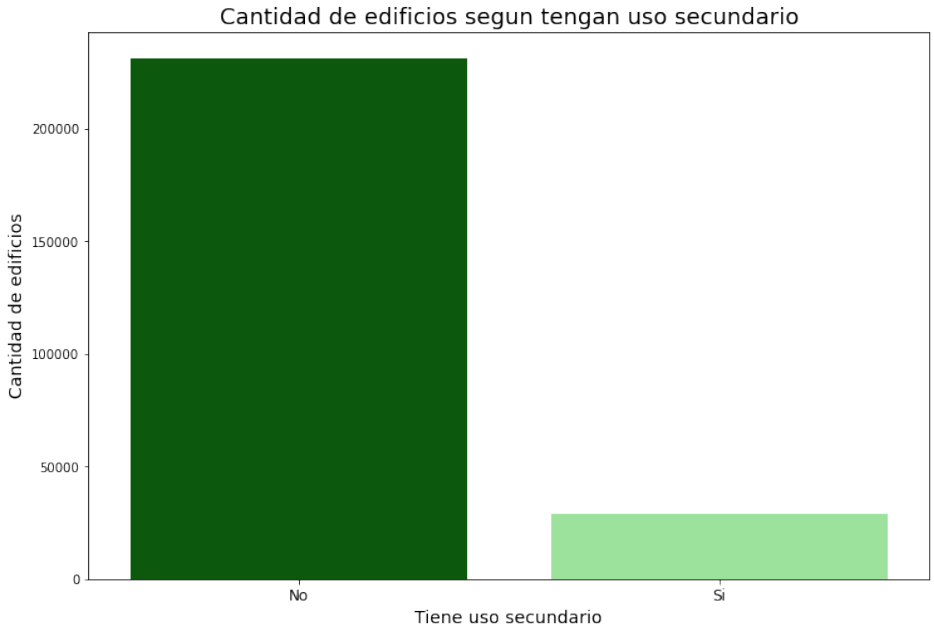
Veamos como se comportan respecto al daño del terremoto:



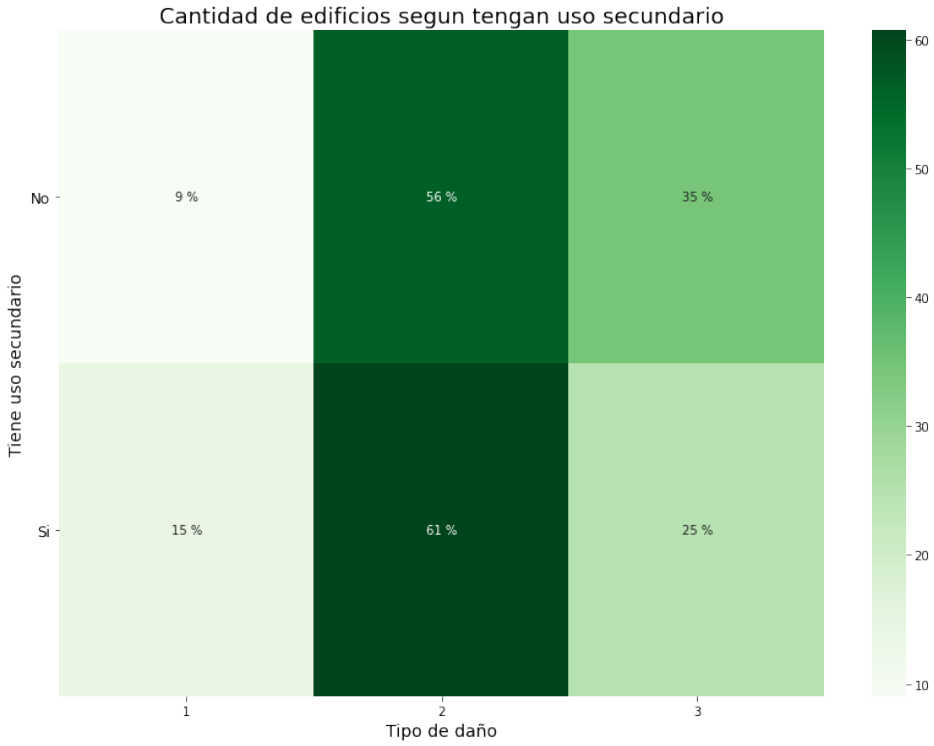
Podemos ver que efectivamente los materiales utilizados para la construcción de cada edificación son una variable de interés a la hora de determinar el daño sufrido a causa del terremoto. Edificaciones que utilizaron materiales especiales y mas resistentes como por ejemplo el concreto reforzado diseñado, sufrieron daños leves en su mayoría, en cambio las edificaciones que utilizaron materiales mas rudimentarios como el barro y la piedra, sufrieron mayor daño.

12. Análisis por uso secundario

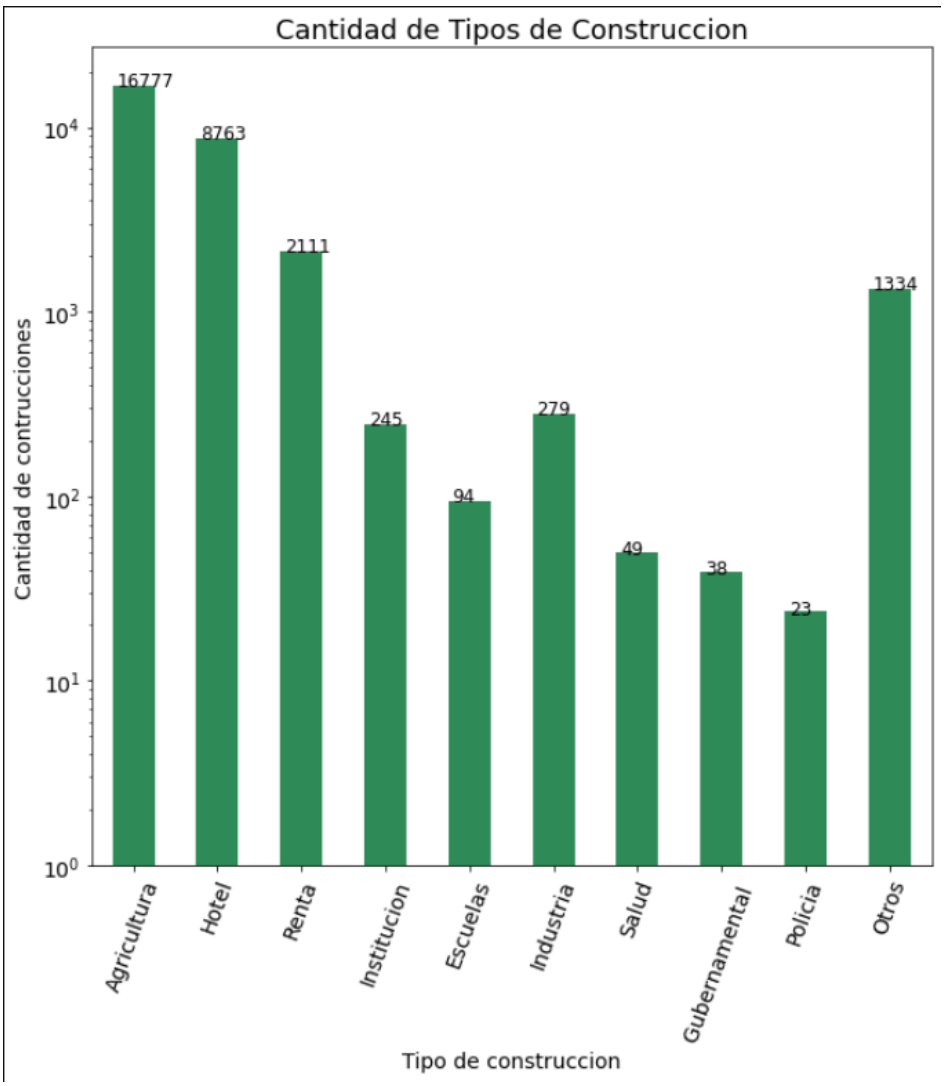
Podemos también analizar si una edificación es utilizada con algun tipo de uso secundario (una oficina gubernamental, por ejemplo). Podríamos pensar que una edificación utilizada como hospital, oficina de gobierno, etc., tiene una mejor planificación y/o construcción y a su vez ser mas resistentes a los daños del terremoto.
A diferencia de lo que pasaba con las superestructuras, no todas las edificaciones tienen algun tipo de uso secundario. En el siguiente gráfico podemos ver la proporción de edificaciones que tienen o no un tipo de uso secundario:



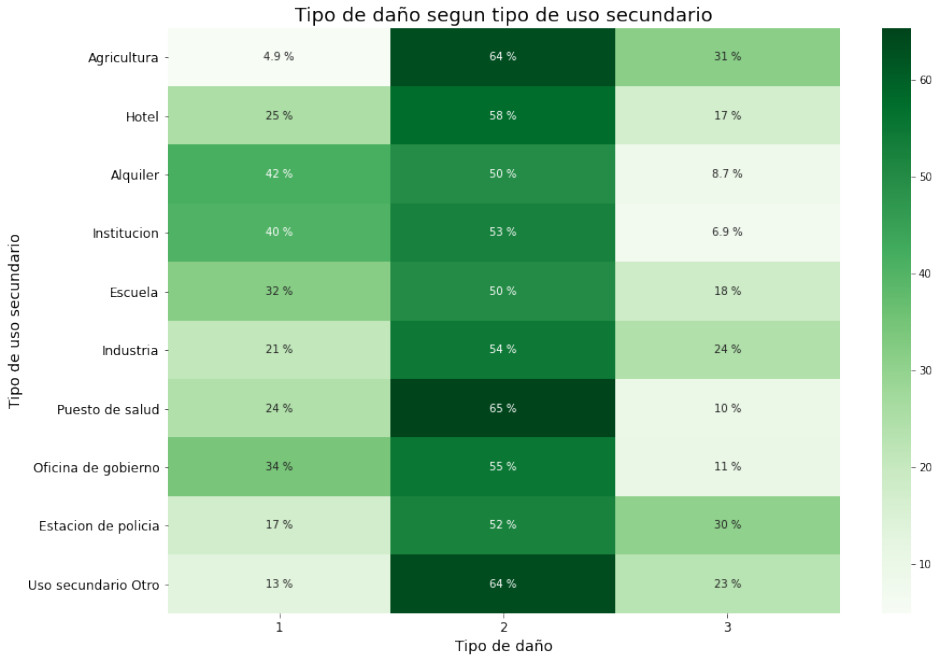
Como era de esperar, son mayor cantidad las edificaciones que no tienen un uso secundario, aunque la cantidad de edificaciones que si cumplen un uso secundario es suficiente para considerar interesante ahondar en su analisis.
Veamos como se comportan respecto del daño sufrido por el terremoto:



Vemos que hay una ligera tendencia a que las edificaciones que tienen algún tipo de uso secundario son mas resistentes a los daños, como sospechábamos antes del análisis. Veamos ahora que sucede con los distintos tipos de uso secundario:

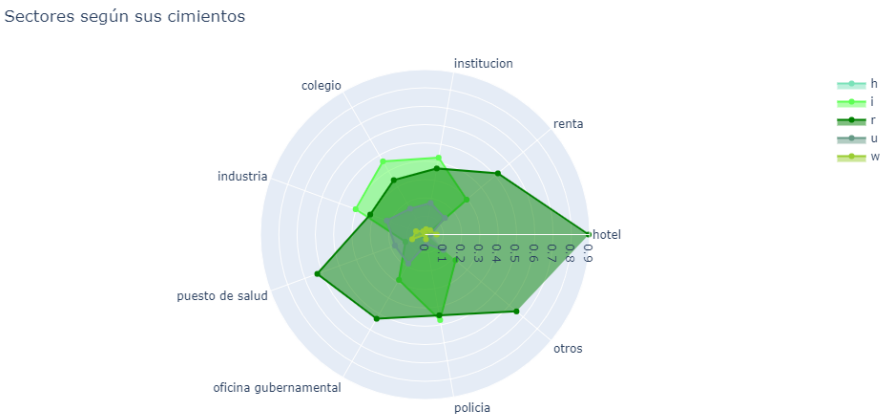


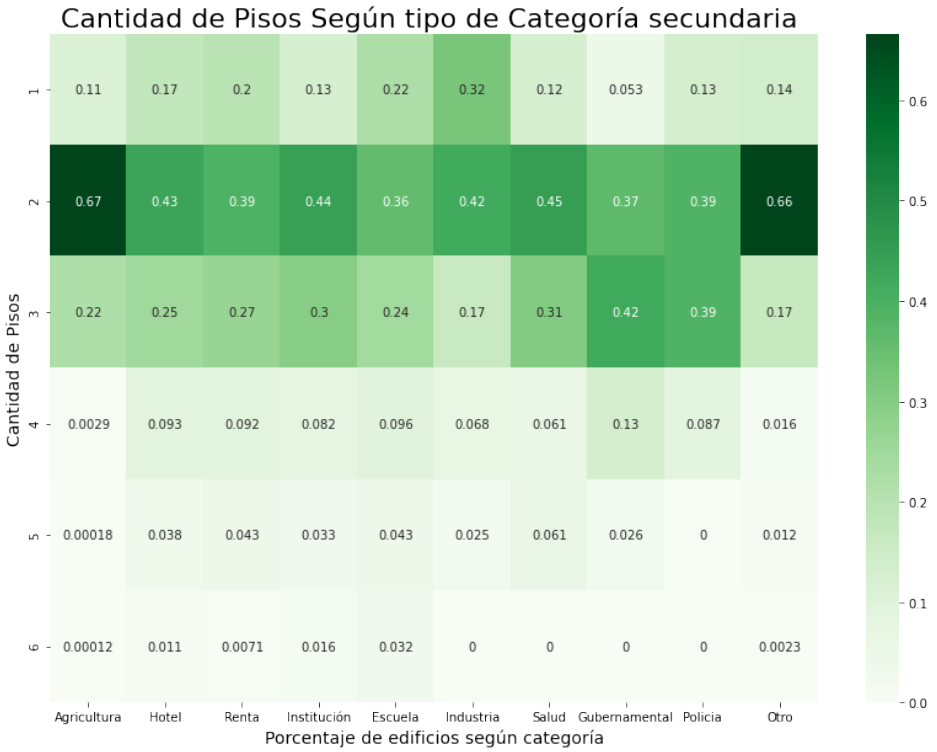
Podemos observar que hay una mayoría de edificios dedicados a la agricultura, seguidos por una gran cantidad de hoteles y por útlimo observamos que hay muchos tipos no catalogados o variados bajo la categoría 'otros'.
Teniendo esto en cuenta proseguimos viendo qué tipo de categoría fue la que más daño sufrió:

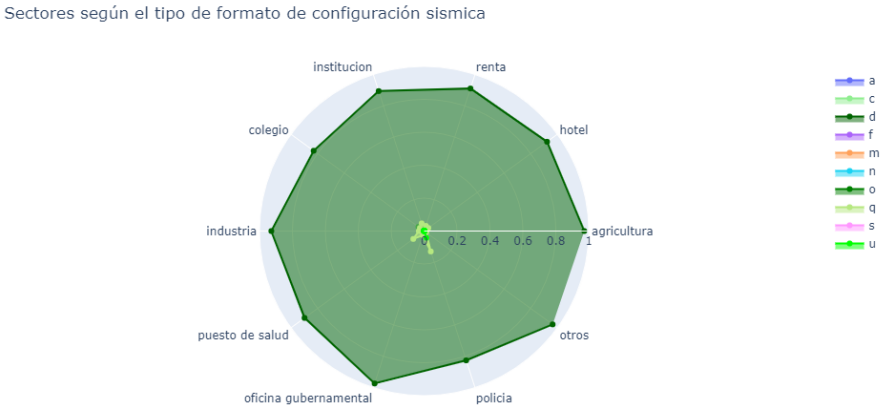


Podemos ver que las edificaciones dedicadas a la agricultura son las mas dañadas y las de tipo institucional, gubernamental son las mas resistentes.

A partir de este análisis nos preguntamos si el tipo de construcción en base a la categoría así como el tipo de terreno en el que fue construido en general según área puedo haber sido el causante de la razón de éstos tipos de daño obteniendo lo siguiente:

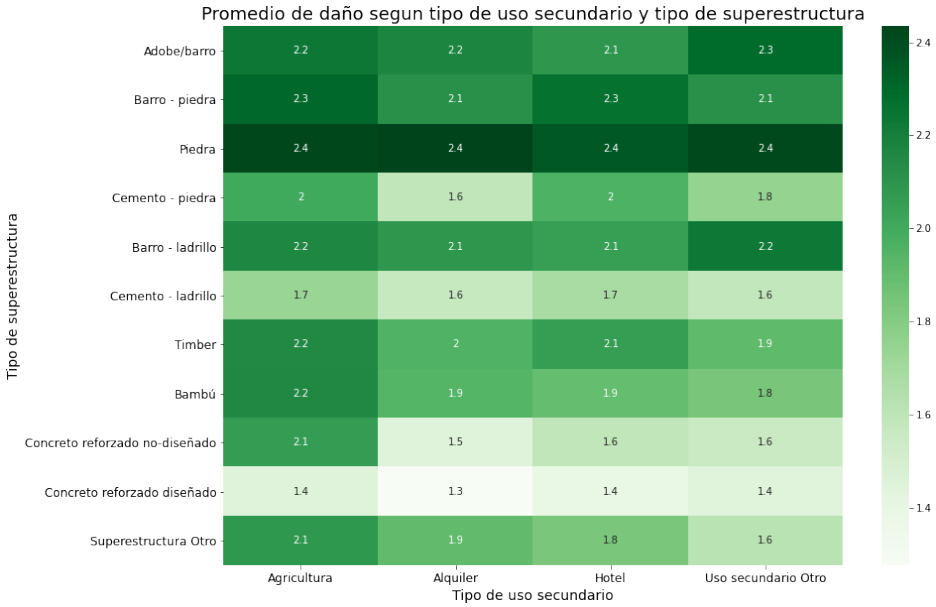






Con estos gráficos podemos rescatar que a mayor cantidad de pisos sufrieron menor cantidad de daño (asimismo observar que por la poca cantidad de pisos de mayor altura representaba muy poco porcentaje y fueron filtrados). Además la mayoría de las categorías están compuestas por el tipo "d" de configuración sísmica mientras que hay un poco más de diversidad de tipo de cimiento respecto a la categoría.

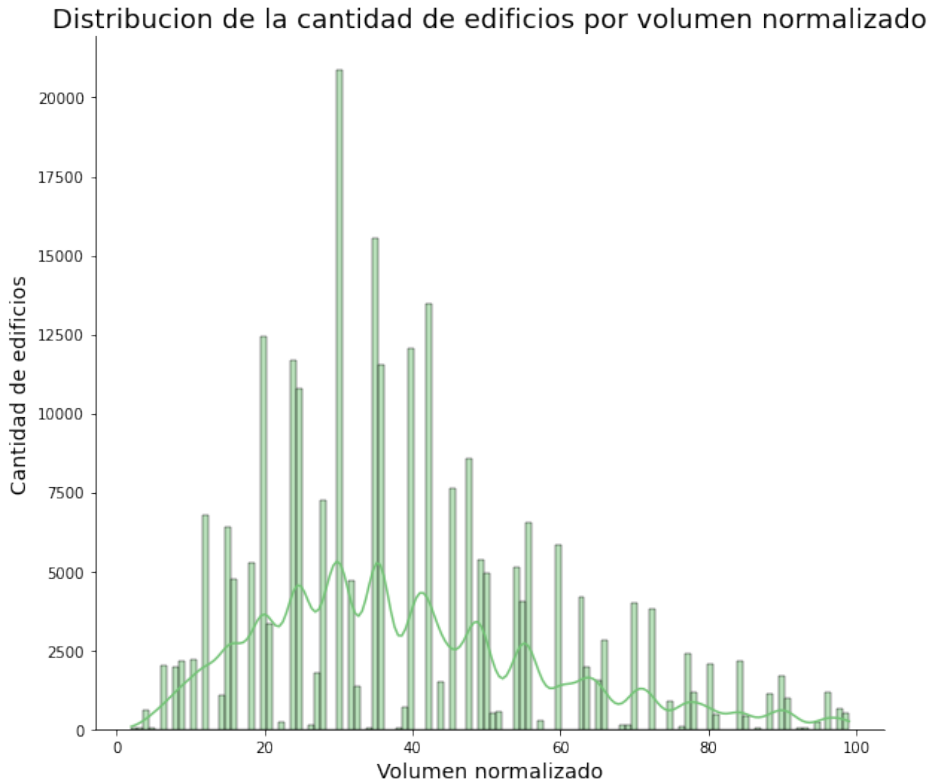
Otro aspecto interesante a analizar, podría ser analizar el daño que sufrieron las edificaciones de cada tipo de uso secundario según los materiales con las que fueron construidas, ya que por ejemplo la categoría alquiler podría contener edificaciones nuevas, bien diseñadas y con materiales especiales pero también edificaciones mas viejas, con materiales mas baratos. Veamos para las categorías de uso secundario mas frecuentes en nuestros datos, que es lo que sucede:



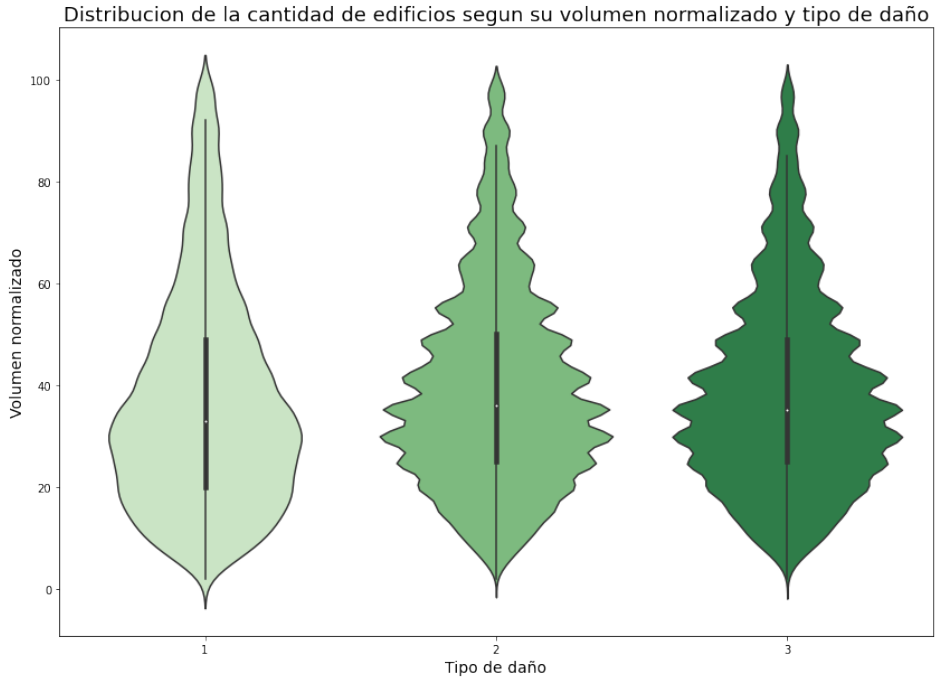
Efectivamente discriminar cada tipo de uso secundario según su material da mas información, ya que el promedio de daño varía bastante según los materiales utilizados.

13. Análisis por volumen normalizado

En nuestros datos contamos con la altura y el área normalizadas de cada edificación, con esto podemos obtener el volumen normalizado de cada edificación (resultante de la multiplicación del área y la altura). Es interesante preguntarnos si los daños sufridos de cada edificación a causa del terremoto tiene que ver con el volumen que ocupa la misma.
Veamos el volumen de las edificaciones:



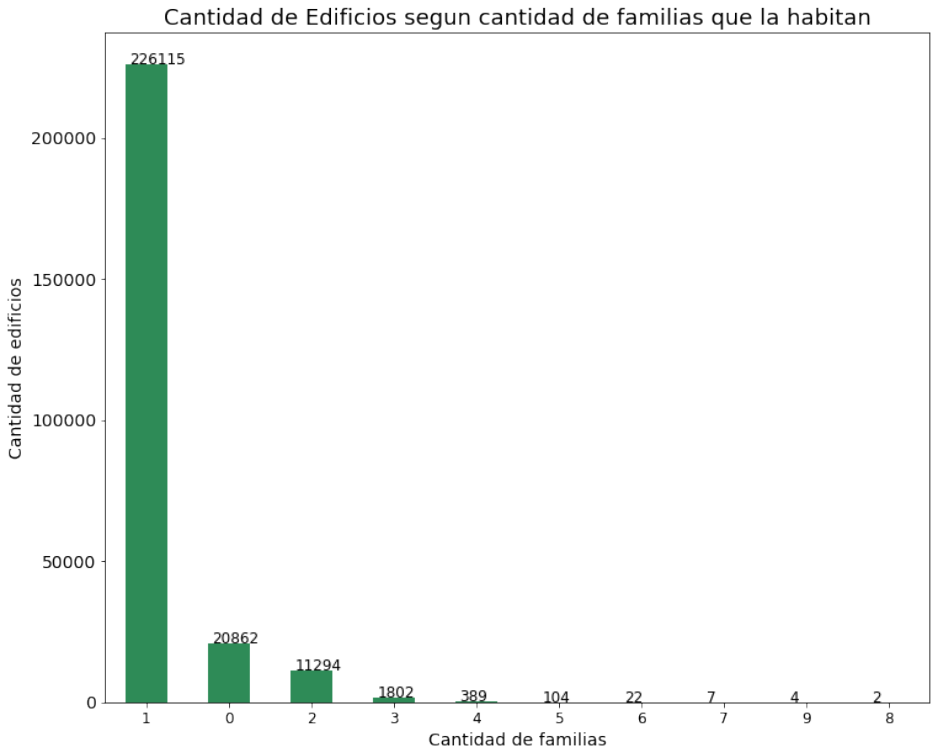
Vemos que la mayoría de las edificaciones están entre 20 y 50, veamos como se comportan respecto al tipo de daño sufrido:



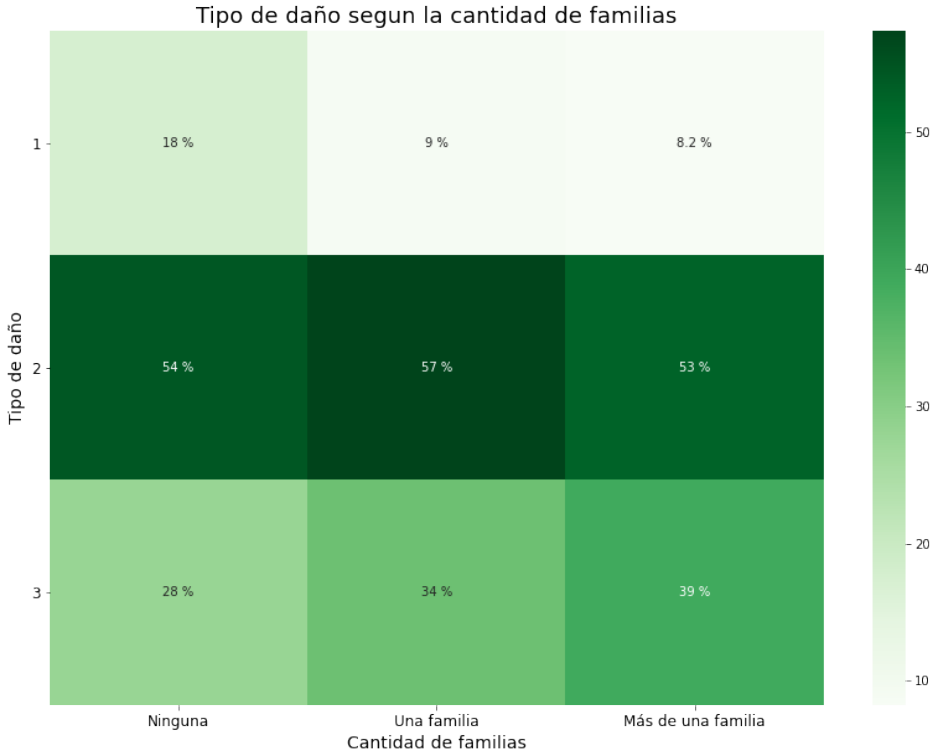
Vemos que no hay una diferencia notable entre las distribuciones para cada tipo de daño, lo que sentencia que los daños no están muy relacionados con el volumen que ocupa la edificación.

14. Análisis por cantidad de familias

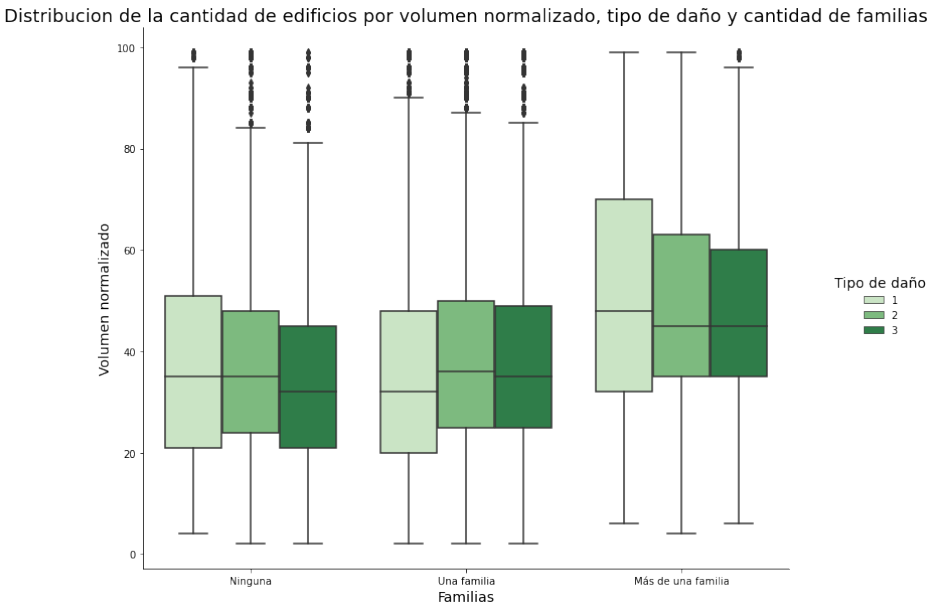
Otra característica que tenemos a disposición, es la cantidad de familias que habitan en la edificación. La cantidad de familias que habitan cada edificación podría estar relacionado con el daño sufrido, ya que mayor cantidad de familias indican edificaciones mas grandes. En un primer análisis vemos que la cantidad de familias que habitan, van de 0 (edificación vacía) a 9.



Vemos que a medida que aumenta la cantidad de familias, menos edificaciones tenemos, por lo que decidimos juntar las edificaciones con mas de 1 familia en una sola categoría. Veamos como se comportan estas categorías respecto del daño sufrido:



Vemos que a medida que aumenta la cantidad de familias que la habitan, mayor es el daño sufrido a causa del terremoto. Veamos si es por una cuestión de volumen de la edificación, como nos habíamos preguntado en un principio:



Vemos nuevamente que el volumen no juega un papel importante a la hora de determinar el daño sufrido, ya que en las 3 categorías de familia, para los 3 tipos de daño tenemos distribuciones similares. Este gráfico también nos ayuda a confirmar lo que inferíamos inicialmente, que las edificaciones con mayor cantidad de familias poseen un mayor volumen.

15. Conclusión

Para comenzar nos resulta importante recalcar que si bien consideramos que realizamos un análisis extensivo de los datos brindados, habiendo tenido datos que se pudieran extrapolar a un mapa real de Nepal, teniendo así latitud o longitud por ejemplo, hubiera podido generarse un análisis más avanzado de la localización del terremoto y las zonas más afectadas, por ejemplo, buscar el epicentro. Otro comentario a agregar es que la gran cantidad de datos categóricos y binarios limita el tipo de gráficos que se pueden utilizar ya que una gran parte están pensados para variables continuas. Tras visualizar los datos brindados a medida que ahondamos más en las observaciones constatamos que las distribuciones de los datos son bastante uniformes.

- Respecto al sector principal verificamos que la mayoría correspondía a un sector residencial. En cuanto a los sectores secundarios pudimos observar que el sector de agricultura es el más utilizado y que a su vez el más dañado dado que en su mayoría estaba compuesto por materiales como piedra y adobe y no tanto por materiales como concreto (reforzado diseñado).

- Respecto a la orientación algo llamativo a rescatar fue que respecto los 200000 edificios, 600 únicamente estaban orientados en 'o' mientras que había en cambio muchos con orientación 's'. Una teoría podría ser que esto se debe a que es un terreno montañoso, y el tipo de bioma llevaba a esto.

- Analizando el volumen de los edificios, es decir el área por la altura, el tipo de daño no cambia, mientras que al analizar la altura y área si.

- Como era previsto, los materiales más preparados y robustos fueron los que menor daño recibieron y a su vez los de peor material fueron los más afectados.

- Observamos que no necesariamente a mayor cantidad de pisos implicaba una mayor cantidad de daño.
- Respecto al análisis de familias obtuvimos que a mayor cantidad, mas grave es el daño que sufrieron las edificaciones que habitan, pero que no dependía del volumen que éstas ocupaban debido a que el volumen crecía respecto a cuanto más aumentaba la cantidad de familias.
- Los materiales de construcción no difieren a pesar de la antigüedad, están distribuidos de manera uniforme contrario a lo que podíamos pensar en cuanto a cambios tecnológicos en materiales.
- No hubo discriminación en el tipo de techo, plantas superiores y planta baja con los que fueron construidos los edificios en cuanto a su uso, ya sea de uso principal o de uso secundario.
- Las edificaciones de uso secundario fueron mas resistentes al daño del terremoto, respecto de las que no.

Por último queremos mencionar que más allá de este análisis y los *insights* que obtuvimos a lo largo del desarrollo generamos algunas visualizaciones más que decidimos no incluir debido a que no consideramos que aportara más información relevante a las preguntas planteadas.