FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

Ánalisis Terremeto en Kathmandu

[7507/9502] Organización de Datos
Curso 1
Primer cuatrimestre 2021

Alumno	Padrón
Mateo Capón Blanquer	104258
Gonzalo Sabatino	104609
Ignacio Iragui	105110
Santiago Pablo Fernandez Caruso	105267

Índice

1. Introducción

El presente informe reúne la documentación de la solución del primer trabajo práctico de la materia Organización de Datos. Este consiste en desarrollar un análisis exploratorio de un set de datos sobre el daño generado por un terremoto en el año 2015 en la ciudad Kathmandu, Nepal.

2. Objetivos

El objetivo principal del trabajo es entender el set de datos y cómo se relacionan las variables para poder comprender su impacto en el daño recibido en los edificios de la región. La idea es explorar las variables más significativas para entender por qué ciertas edificaciones sufren mayores o menores daños. Tomaremos:

- Casos de éxito: Aquellas edificaciones con grado de daño 1.
- Casos de fracasos: Aquellas edificaciones con grado de daño 3.

3. Preguntas e Hipótesis

Para entender el set de datos nos formulamos preguntas e hipótesis a medida que fuimos desarrollando el trabajo, que se resolvieron obteniendo relaciones provechosas de todas las variables. Las preguntas nos sirvieron para poder obtener las variables más relevantes para los grados de daño y los casos de éxito y fracaso. Es una parte más objetiva del análisis, ya que examinamos todas las variables, estudiando, para todos los valores posibles que pueden tomar, qué conjunto generan éxito o fracaso. Podemos relacionar las preguntas con un análisis univariable.

Las hipótesis nos sirvieron para poder entender de forma subjetiva el set de datos, y corroborarlas o no a partir de los análisis objetivos. Podemos relacionar las hipótesis con un análisis multivariable, orientado a responder inquietudes iniciales.

Las preguntas que nos hicimos fueron:

- ¿Qué casos de éxito y fracaso se pueden extrapolar de cada variable?
- Esos casos de éxito y fracaso, ¿Se deben a la presencia de otras variables o se puede afirmar que la variable analizada es de relevancia?
- ¿Qué relaciones entre las variables nos dan las relaciones más provechosas para analizar para los casos extremos?
- Para los casos de fracaso hallados como más importantes, ¿Qué casos de éxito se necesitan para contrarrestarlos y lograr daño 1?

Para resolverlas, en muchos casos utilizaremos los casos más significativos para cada variable (los que más muestras tienen o los que mayor diferencia entre casos de éxito y fracaso presentan).

Las hipótesis formuladas fueron:

- La edad, altura y área son determinantes. A mayor área y menor altura, menor es el daño, y a mayor edad mayor es el daño.
- Los aspectos sociales (cantidad de familias, los usos secundarios y el estado legal de la tierra donde fue construida) no deberían ser significativas, ya que lo significativo son las condiciones en que fueron construidas. En lugar de relacionar, por ejemplo, lugares más carenciados, lo significativo para el análisis es la estructura de la edificación.

4. Estructura del DataFrame

La primer tarea antes de comenzar el análisis de cada variable fue realizar un estudio rápido e inicial de los datos para verificar que estén correctamente cargados y extraer información preliminar que nos permita elaborar un análisis final más desarrollado.

4.1. Verificación de calidad de los datos

En función de controlar la calidad de los datos debemos primero asegurarnos que la carga de los mismos se genere correctamente (los tipos de datos se correspondan a la característica) y no contengan datos nulos. Además, observamos que los datos están cargados en un formato adecuado, al ser cada fila una muestra distinta. Asimismo corroboramos que no hay observaciones repetidas, y existe una relación uno a uno entre ambos archivos para los id de los edificios.

Una vez verificado esto, procedemos a chequear que los datos correspondientes a cada columna sean correctos dentro de su dominio esperado. Es decir, todos los valores dentro de la columna son los esperados para esta.

Primero comparamos las columnas de uso secundario de forma tal que si la columna general está en True haya al menos una columna especifica de uso secundario en True. A su vez controlamos que aquellas variables que estén normalizadas, no superen el 100 % ni estén por debajo del 0. Con respecto a las variables caracterizadas por palabras, vemos que cumplen que los valores que toman son los mismos que fueron presentados en el enunciado. En cuanto a las del tipo numérico verificamos que todas tengan valores mayores a cero (no puede haber una edad o una cantidad de familias menor a cero).

Por último, analizamos si los valores de las zonas geográficas locales son válidos entre si. Es decir, que ningún valor de la segunda sub-región tenga 2 hiper-regiones distintas, y lo mismo con las relaciones entre las terceras y segundas sub-regiones.

4.2. Identificación de valores atípicos

A continuación mostramos un gráfico de barras que indica la cantidad de edificios en función de su edad. Es necesario aclarar que la barra de 995 no se encuentra alineada como los otros datos, ya que si lo hubiéramos puesto como corresponde, el gráfico sería muy extenso.

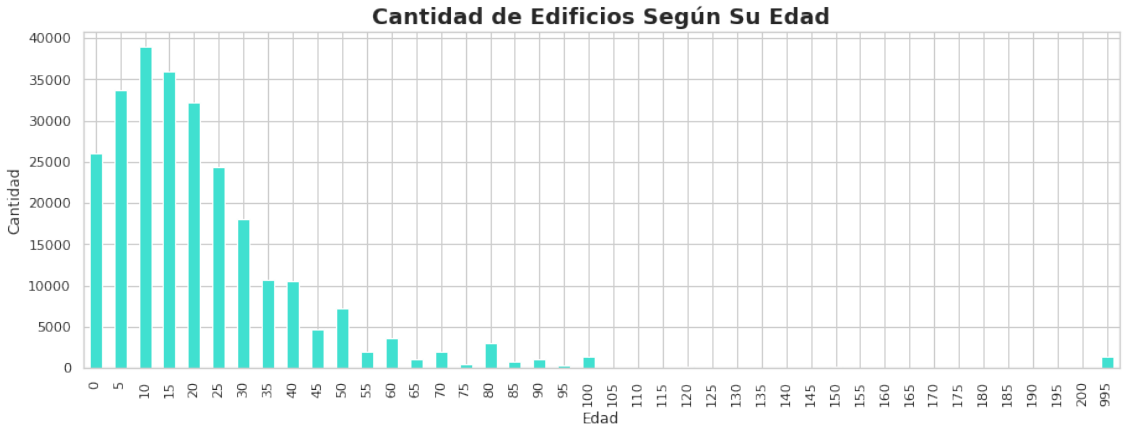

Figura 1: Gráfico de barras de cantidad de edificios según su edad

Podemos apreciar que la mayor parte de las edificaciones datan de los últimos 25 años. Además, registramos un conjunto de edificaciones con edad de 995 años, los cuales notificamos como valores anormales. Esto podría deberse a que no se conoce la edad exacta de una edificación, y por lo

tanto se la agrupa por default en 995, o simplemente podría ser una carga incorrecta de los datos. En nuestro análisis tendremos estos datos en cuenta, sin embargo las separaremos como valores atípicos en algunos casos de estudio particulares en los que buscaremos centrarnos en la gran mayoría de datos relacionados con la edad.

5. Desarrollo

5.1. Análisis de la variable principal: Grado del daño

Para comenzar el análisis, ponemos el foco de atención en la Variable Aleatoria Discreta de respuesta

$$Y \equiv \textit{"Grado del daño recibido"}.$$

Esta es la variable más importante del set de datos. Luego del análisis exploratorio, en el tp2, buscaremos predecirla. En este sentido, realizaremos modelos que estimen su comportamiento conociendo muestras de las demás Variables Aleatorias regresoras o explicativas X_i. Por el momento $0 < i < n = 38$. Sin embargo existe la posibilidad de que combinemos variables o eliminemos otras si es necesario, de modo tal que no podemos afirmar que la cantidad de Variables Aleatorias que tenemos se mantendrá constante. Utilizaremos un set de datos auxiliar donde volcaremos toda la información útil a los casos de éxito y fracaso de cada variable. Esto es así para no eliminar datos del dataset original, que podrían quererse utilizar luego, aún sabiendo que el pedido de los datos puede ser más costoso, ya que en el dataset auxiliar volcaremos solo la información más relevante para el estudio del problema.

Dependiendo del tipo de estudio que querremos explorar, consideraremos a la variable también como categórica.

5.1.1. Descripción y visualización inicial de la variable

Comenzaremos viendo el porcentaje de edificios que presentan daño de tipo 1, 2 y 3.

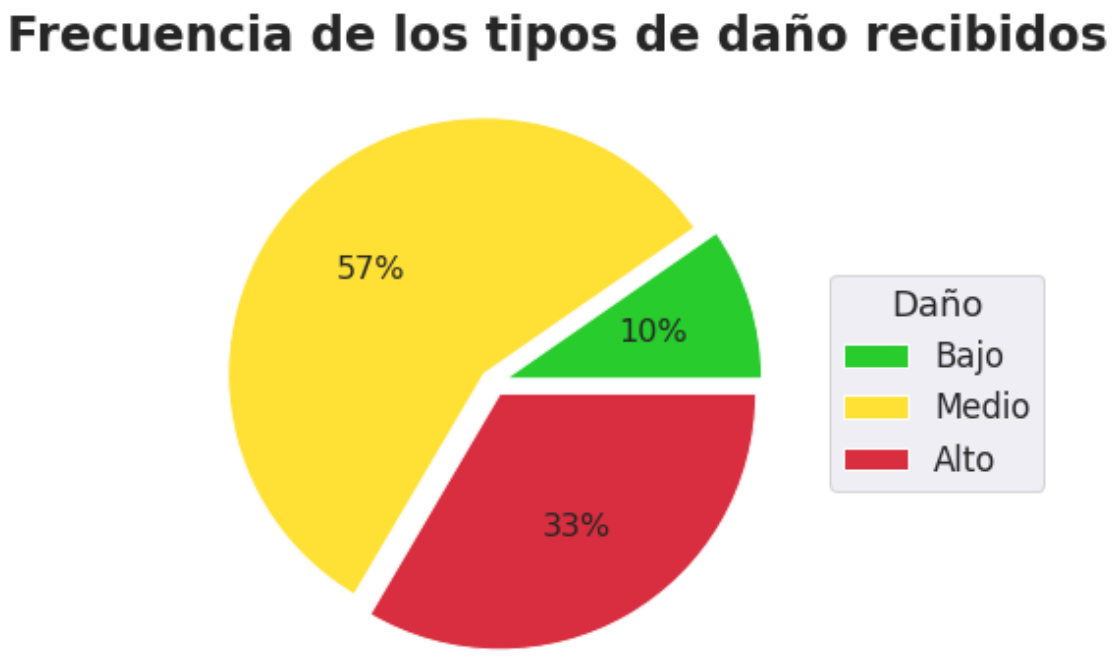

Figura 2: Proporción de los daños recibidos sobre el total de los edificios

A continuación mostramos un gráfico de barras, el cual nos presenta una mirada cuantitativa del daño, complementándose así a la mirada cualitativa dada por el gráfico de tortas. Es esencial

para nuestro análisis tener presentes ambos gráficos en todas las etapas del desarrollo, para poder comparar el efecto de las variables en el comportamiento de la variable daño.

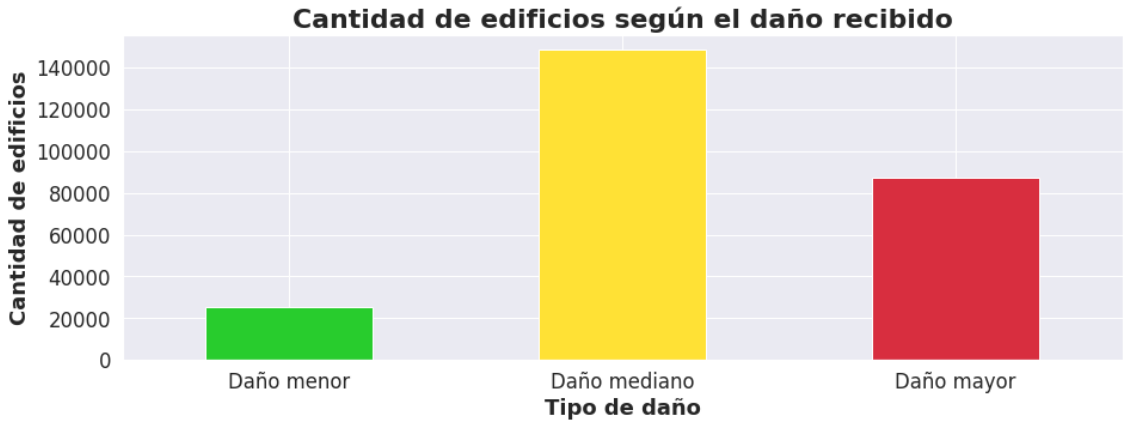

Figura 3: Gráfico de barras de cantidad de edificios según su daño

5.1.2. Correlación con las demás variables

Analizamos como se correlacionan las variables con la variable Y. Separamos con colores de tonalidad las distintas familias de variables aleatorias que se pueden diferenciar según su característica.

1. Zona geográfica (Amarillo): *geo_level_1_id, geo_level_2_id, geo_level_3_id.*
2. Las de tamaño y edad (Azules): *count_floors_pre_eq, age, area_percentage y height_percentage.*
3. Materiales de construcción (Violetas): todas aquellas que comienzan con *has_superstructure.*
4. Cantidad de Familias (Marrón).
5. Uso Secundario (Verdes).

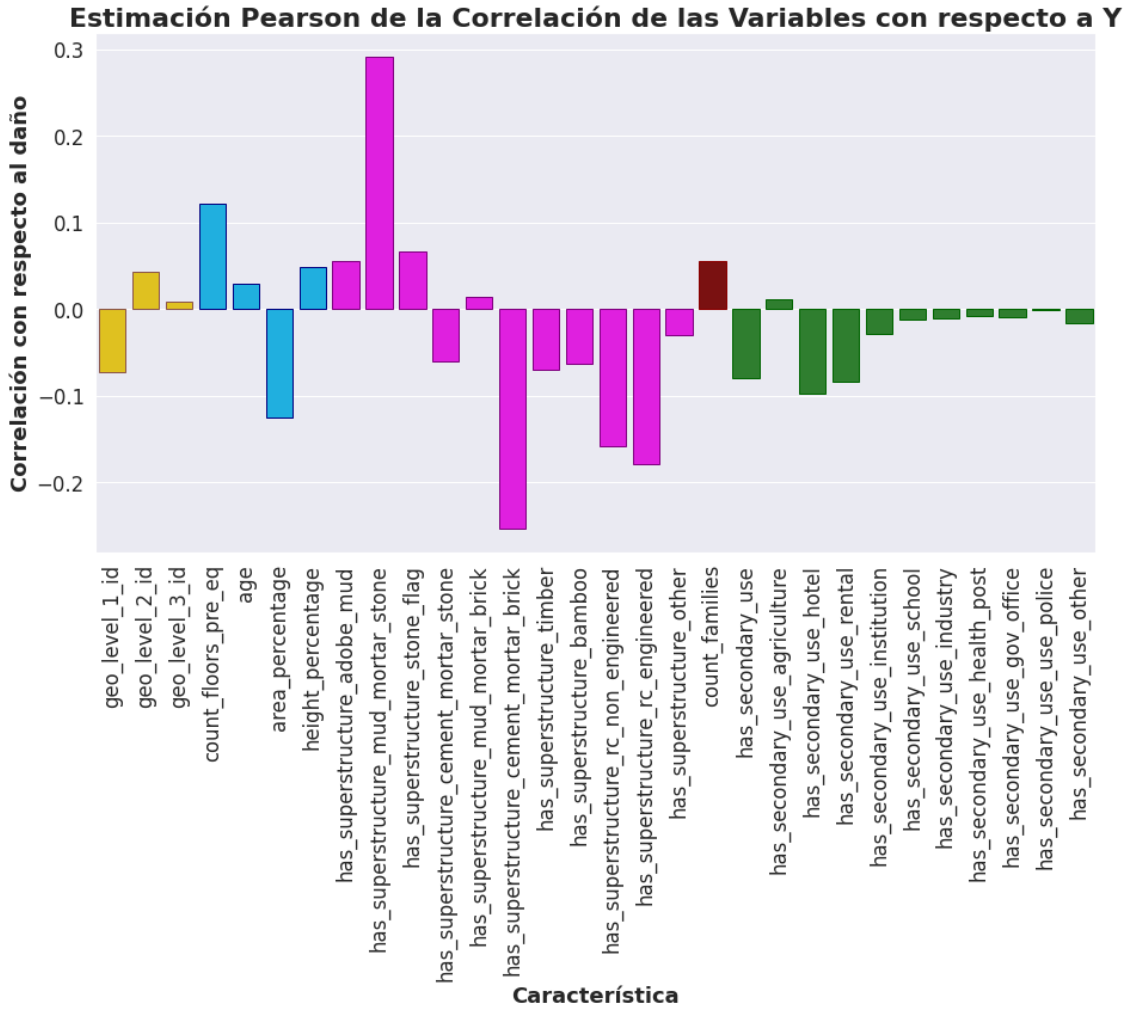

Figura 4: Correlación entre las distintas variables numéricas y *damage_grade*

De este gráfico de correlaciones extraemos ciertas variables que debemos analizar detalladamente, dado que su correlación (en módulo) con la variable importante a analizar es significativa en comparación con las demás variables.

Es cuanto a la variable que indica la cantidad de pisos previo a el terremoto, el tener una correlación positiva y fuerte (en comparación con las demás) era algo esperable dado que a mayor cantidad de pisos, menos estable es una estructura, y por lo tanto, mayor es el riesgo de desastre. Lo mismo ocurre para la variable de estructura de tipo barro y piedra, ya que es sabido que estos materiales no son los más resistentes.

En contraste, la variable estructural de cemento y ladrillos tiene una correlación negativa y fuerte, pues soporta más daños. Otras estructuras que parecieran tener una buena resistencia a sismos son aquellas que fueron construidas con concreto reforzado.

Las variables de uso secundario tienen en general una correlación baja con la variable en cuestión. Sin embargo se puede observar que en general son más resistentes a los sismos que la media. Esto se puede deber a que estas edificaciones suelen pasar por más controles de normas de seguridad.

Es interesante observar también la correlación de la variable principal con las variables categóricas del DataFrame. Al no ser numéricas, no las podemos estudiar en conjunto. Por lo tanto aplicamos el método conocido por la comunidad como One Hot Encoding para generar una columna con contenido binario por cada categoría posible.

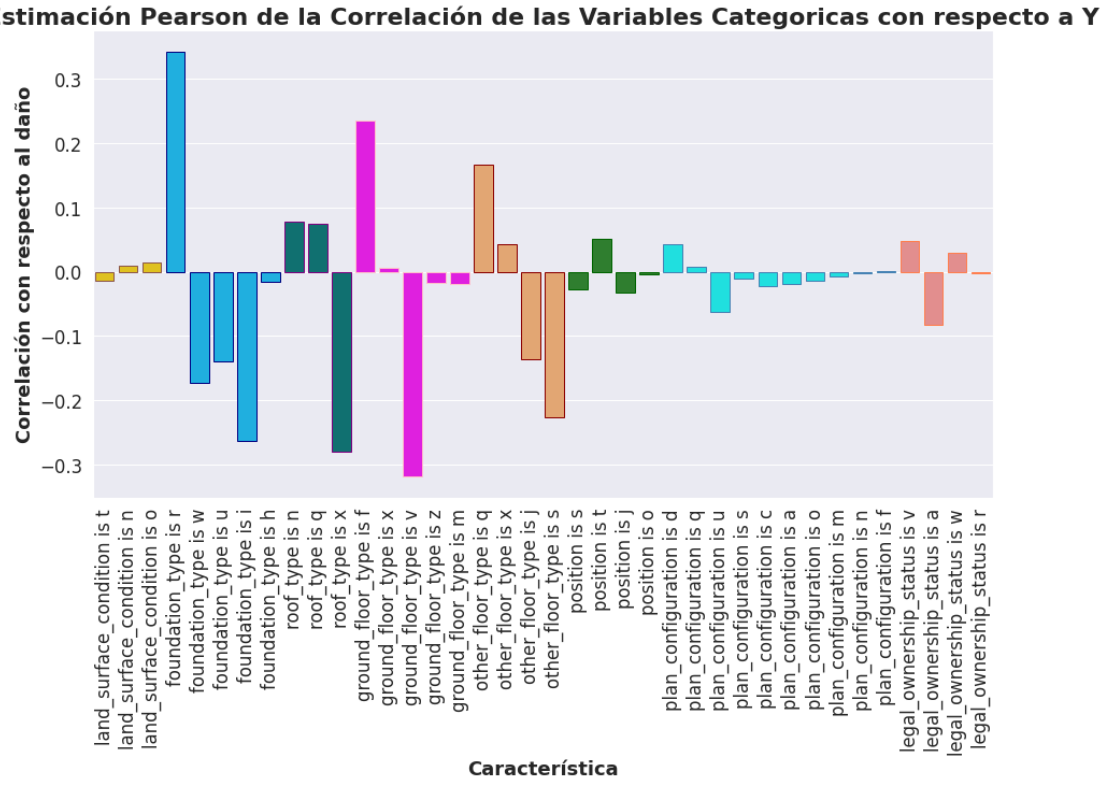

Figura 5: Correlación entre las distintas variables categóricas y *damage_grade*

El presente gráfico nos permite identificar nuevamente variables de gran importancia. En nuestro análisis le tendremos que prestar principalmente mucha atención a las variables del tipo de cimientos usados en la construcción (*foundation_type*), el tipo de techo (*roof_type*), el tipo de cimiento de los pisos (*other_floor_type*) y el de la planta baja en particular (*ground_floor_type*). Estas muestran gran correlación, tanto positiva como negativa, evidenciando así por ejemplo que un tipo de cimientos *r* es más propenso a daños que los otros cimientos.

Por el otro lado, no nos detendremos demasiado a analizar las variables con correlaciones en módulo muy bajas. Estas son: el formato de construcción en la edificación (*plan_configuration*),

la condición de la tierra en su construcción (*land_surface_condition*), la orientación del edificio (*position*) y el estado legal de la tierra (*legal_ownership_status*).

5.2. Visualizaciones genéricas

Previo a nuestra exploración específica sobre cada variable, calculamos y graficamos la matriz de correlaciones para todas las variables aleatorias. Esto es, para las variables que pueden ser consideradas como numéricas hasta cierto punto: las binarias, las numéricas discretas y continuas, y las categóricas caracterizadas por números.

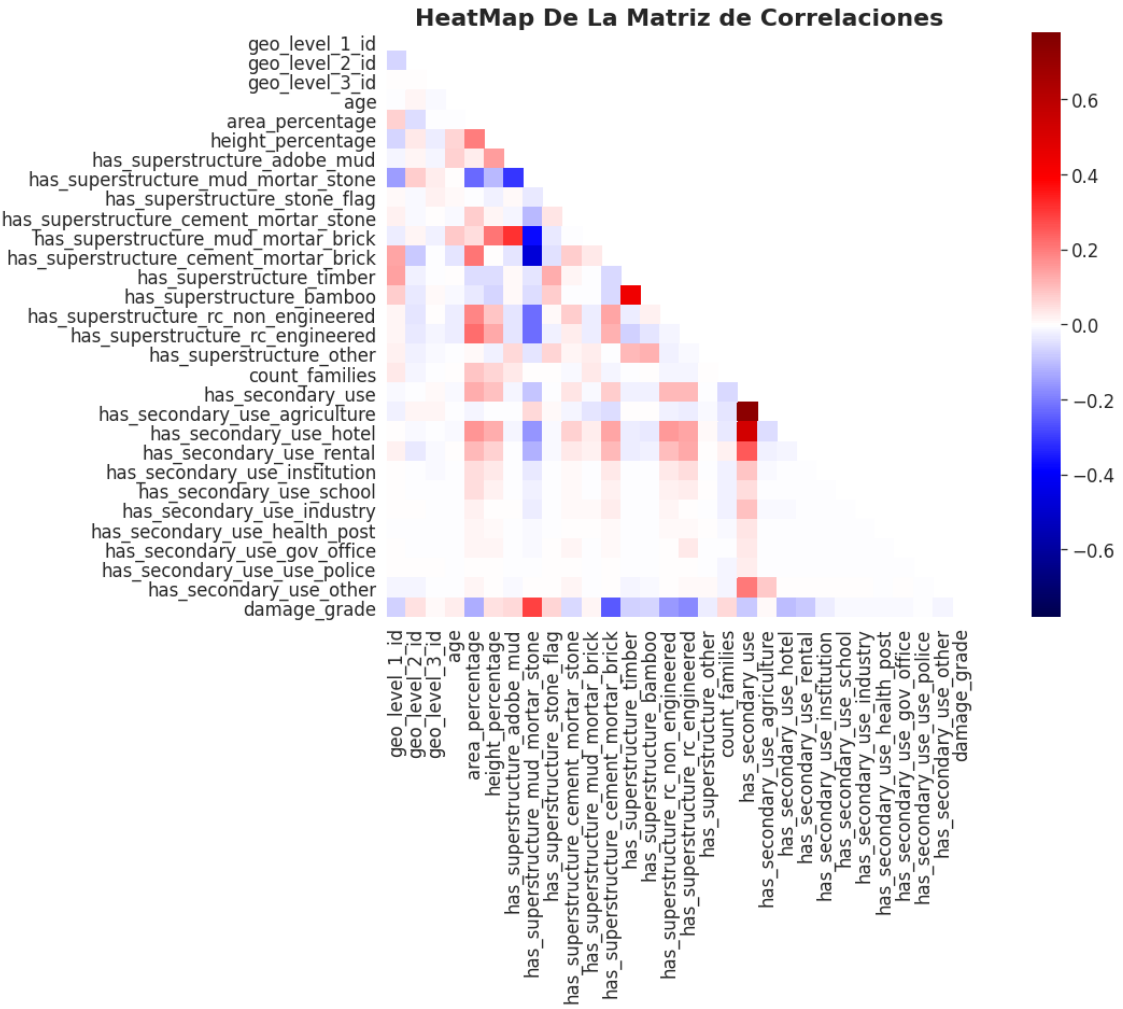

Figura 6: Correlaciones para todas las variables

Este gráfico muestra muchísimas relaciones entre las variables. Intentaremos analizar las cuestiones más significativas. Diferenciando en las Familias mencionadas, se pueden observar claras relaciones entre las mismas.

En primer lugar, para aquellas de Uso Secundario, las correlación es muy cercana a cero (región a la izquierda e inferior del gráfico). Esto se debe a que generalmente si un edificio tiene una función, es muy probable que no tenga otra. Es poco común conocer un colegio que comparta espacios con una fábrica industrial.

El hecho de que la correlación entre el uso secundario genérico y el de agricultura sea muy alta nos indica que la gran mayoría de edificios con uso secundario son usadas con propósitos de agricultura. Esto se verifica al calcular el porcentaje de edificios para usos agrícolas: representan al 57 % de todos los usos secundarios. En segundo lugar hay más hoteles, luego más edificios con propósito de renta, etc., siguiendo los tonos de rojo en la columna de uso secundario (*has_secondary_use*) con respecto a las demás variables de su familia.

Se cumple también un nivel de asociación relativamente alto entre las familias de Materiales de Construcción y las de Uso Secundario. Dejando de lado la variable de uso secundario para agronomía, se observa en la columna de estructura de lodo y piedra (*has_superstructure_mud_mortar_stone*) que la correlación es negativa. Esto es, la mayoría de edificios con uso secundario, no están construidos con barro y piedra, la cual por el color rojizo que toma en la correlación con el grado de daño, indica que es de las estructuras menos resistentes. En contraposición, la correlación es mayor a cero con algunas de las estructuras que menos daño sufrieron: las de cemento y ladrillo y las de estructuras anti-sismo (con concreto reforzado).

Esto pareciera fortalecer nuestra suposición previa: Los edificios con uso secundario suelen tener más controles de normas de seguridad.

Adicionalmente, a excepción de los lugares de uso agrónomo, el volumen de los edificios suele ser mayor. Esto se evidencia por la tonalidad de las columnas de área y altura en relación a la familia de Uso Secundario.

Para la variable que indica una normalización de la altura, se observa naturalmente una relevante relación con la cantidad de pisos. Además, ésta aumenta a medida que aumenta el área del suelo.

Buscaremos demostrar las anteriores afirmaciones y responder parte de las preguntas planteadas mediante un análisis más exhaustivo sobre cada familia de variables de forma individual.

5.3. Visualización de variables numéricas

5.3.1. Altura y cantidad de pisos

Recordando que la variables altura y cantidad de pisos están fuertemente correlacionadas, las analizamos en conjunto, buscando entender gráficamente esta relación mencionada. Veamos en una primera instancia distintos scatter plots de la altura en relación con la cantidad de pisos, condicionando al daño.

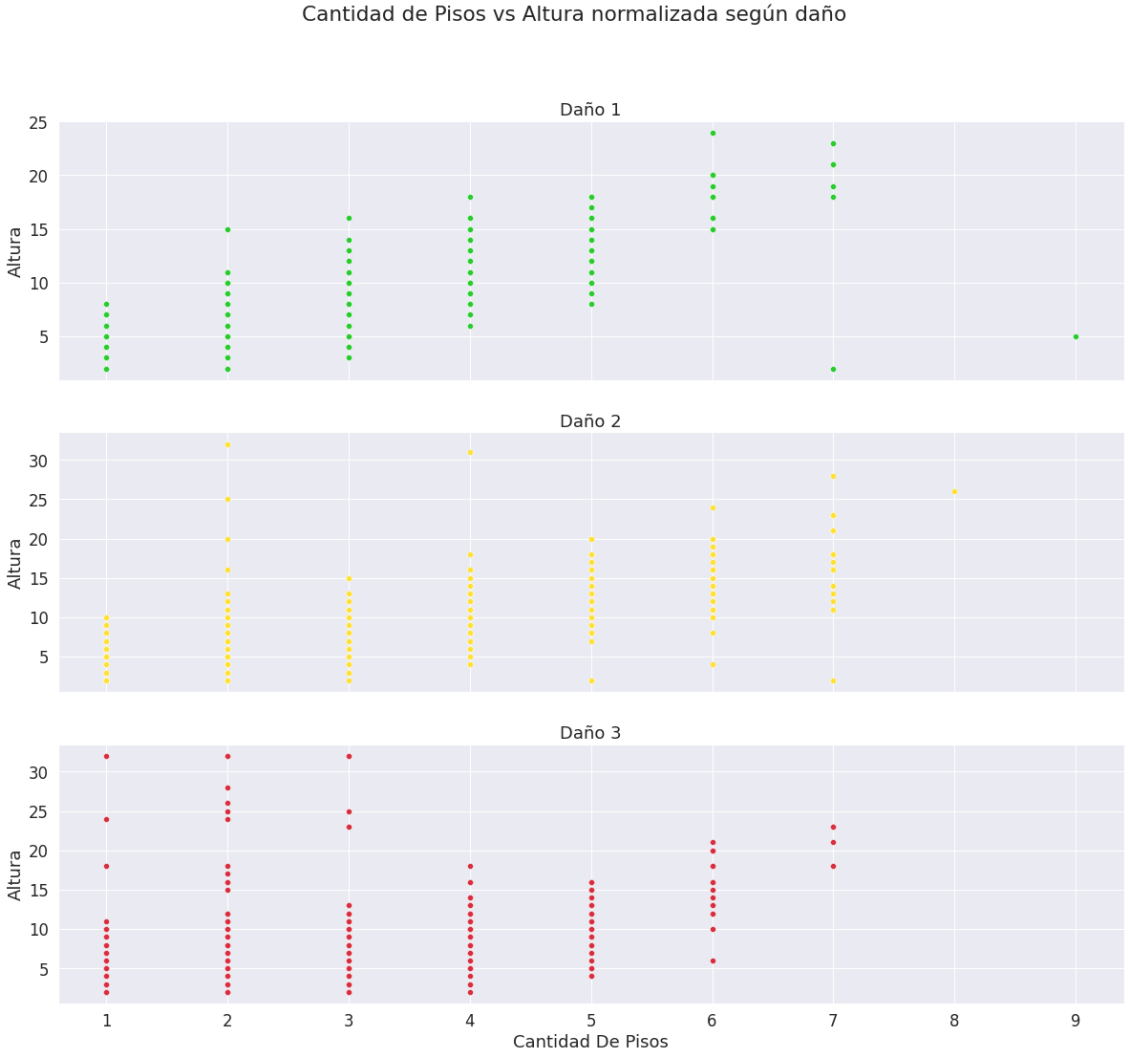

Figura 7: Relación entre la cantidad de pisos y su altura separados según su daño

La primera impresión que nos entrega el gráfico es que los edificios más altos en Kathmandu tienen 9 pisos, lo cual nos indica que en general los edificios no son tan grandes.

Adicionalmente, estos gráficos nos permiten ver a modo muy superficial la correlación que buscábamos mostrar. Para todas las condiciones de daño, se observa un crecimiento de la altura de la columna de puntos, a medida que aumenta la cantidad de pisos. Esta conclusión es totalmente intuitiva: A mayor cantidad de pisos, mayor altura del edificio.

Con el propósito de afianzar esta idea, calculamos la esperanza de la altura condicionada a cada posible valor de la cantidad de pisos con un intervalo de confianza para la misma. El resultado de estos cálculos se observa a continuación.

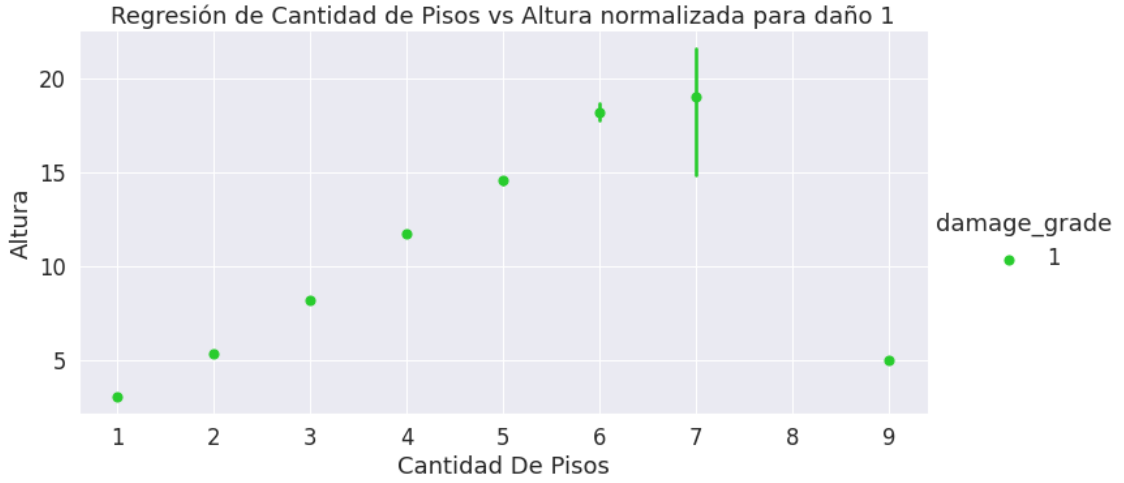

Figura 8: Esperanza de la altura condicionada al daño leve y al valor de la cantidad de pisos con un intervalo de confianza para la misma.

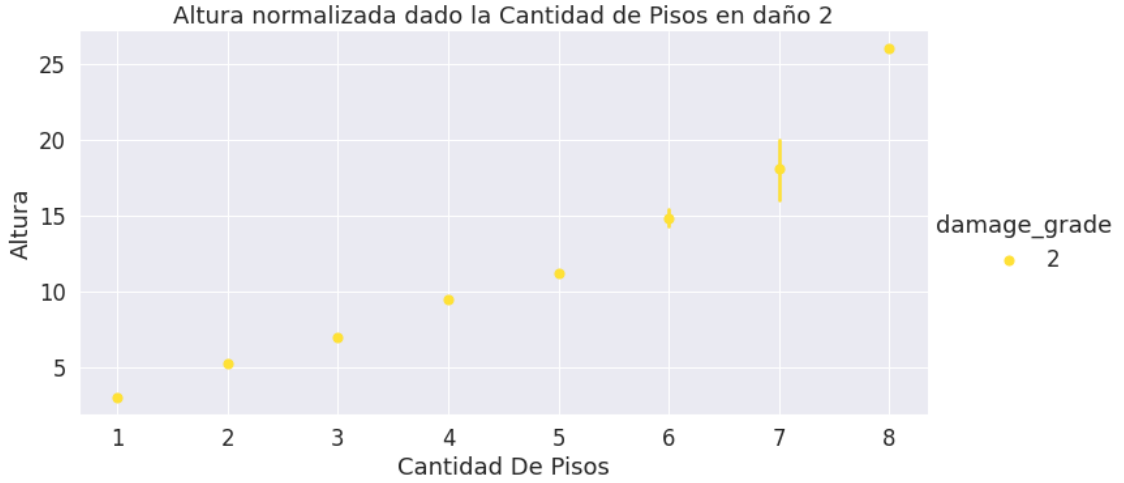

Figura 9: Esperanza de la altura condicionada al daño medio y al valor de la cantidad de pisos con un intervalo de confianza para la misma

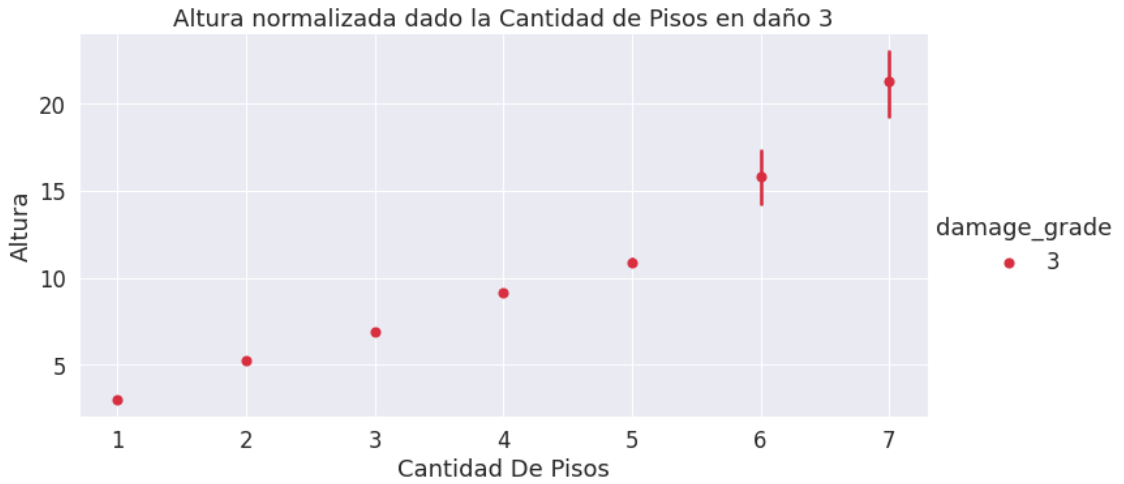

Figura 10: Esperanza de la altura condicionada al daño grave y al valor de la cantidad de pisos con un intervalo de confianza para la misma

Debido a la gran cantidad de datos para edificios que tienen entre 1 y 5 pisos, los intervalos de confianza son tan pequeños que son inapreciables. Para estos se puede observar una relación lineal casi perfecta entre las esperanzas. Esto nos podría permitir en una futura búsqueda del modelo, reducir la dimensión del set de datos combinando ambas variables.

En contraposición, hay pocos datos con más de 5 pisos, lo cual genera que haya intervalos de confianza tan grandes que no se puede afirmar nada sobre la media condicionada.

5.3.2. Edad de los edificios

Observamos cómo evoluciona la cantidad de edificios en base a los años en las edificaciones. Para este gráfico filtramos los valores atípicos de edad, los cuales analizaremos en un apartado separado.

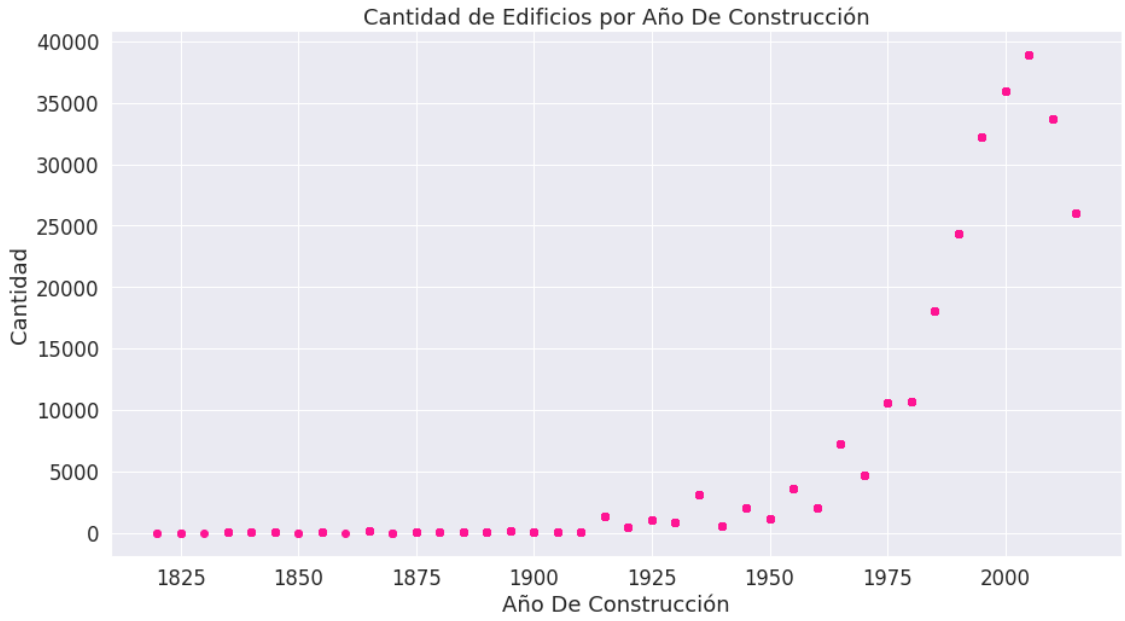

Figura 11: Avance de cantidad de edificaciones por año

Del gráfico apreciamos que la mayor cantidad de edificios se concentran en el rango de los últimos 50 años con un pico claro en 2005. Se advierte que solo hay datos en años múltiplos de 5.

Con el objetivo de observar cómo afecta la edad de un edificio al daño recibido, condicionamos a la variable aleatoria edad, cada daño. En este caso tomamos al daño como una variable categórica. A pesar de que nuestro propósito principal es el de observar el comportamiento del daño condicionado a las demás variables (Y|X), este gráfico nos permite observar la condición inversa X|Y, la cual por Bayes, nos dará también información sobre el comportamiento de Y condicionado.

Figura 12: Cantidad de edificios agrupados por edad separados por tipo de daño

Es notable que cuanto mayor es la cantidad de años de la edificación, mayor es el daño que sufren. Por ejemplo, para los edificios con edad 0, hay más construcciones con daños leves que con daños graves, mientras que para los edificios de más de 15 años de edad, esta situación se invierte fuertemente; y ya para aquellos con más de 35 años de edad, la cantidad de edificios con daño leve es casi insignificativa.

Dado que en un primer momento quitamos las edificaciones de mas de 995 años, las examinamos por separado.

Figura 13: Cantidad de edificios de 995 años por tipo de daño

Podemos apreciar que para los edificios con supuestamente 995 años en este registro, la mayor cantidad de casos es de daño medio. Además, que el daño máximo duplica (aproximadamente) al

daño menor. Sin embargo, la diferencia de proporciones no es tan grande como la que se observa en el anterior gráfico para las estructuras de más de 15 años. Esto nos sugiere que los datos para la edad 995 no son de edificios tan antiguos, sino que pareciera más probable que estos valores indican una carga incorrecta de los datos.

5.3.3. Superficie ocupada por la edificación

En un primer análisis de los datos notamos que el porcentaje de área no es un continuo, al tomar valores en los naturales menores o iguales a 100. De igual modo, al ser un porcentaje, los valores deberían ser continuos; y al haber muchas clases distintas analizamos al dato como un continuo y graficamos su densidad de probabilidad condicionada al daño grave (en rojo), y al daño leve (en verde).

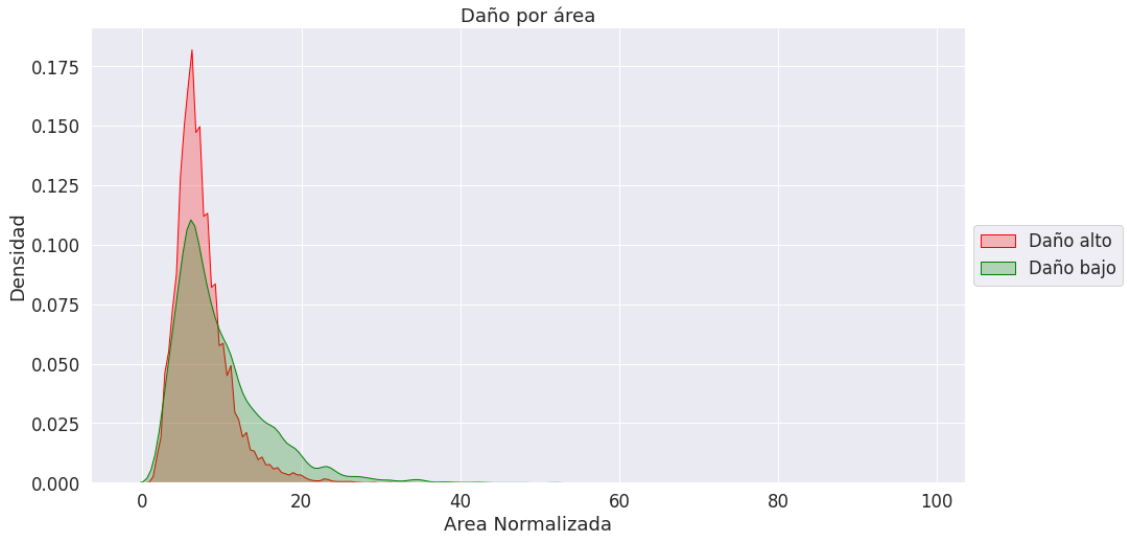

Figura 14: Densidad de probabilidad del área condicionada al tipo de daño.

Observamos que hay una mayor concentración de daño alto en las áreas mas pequeñas, y que el daño menor supera al mayor (casos de éxito) en las áreas mayores. En otras palabras, la curva para el área condicionado a que el daño fue menor tiene mayor dispersión que la otra curva, indicando que dentro de los edificios con mayor área, la mayoría recibieron un impacto menor. Este efecto sobre la dispersión se observa claramente en los siguientes BoxPlots.

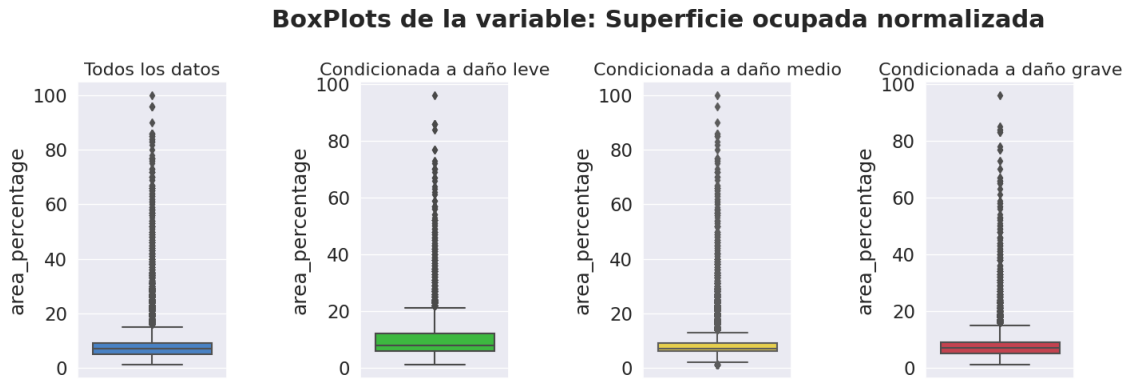

Figura 15: BoxPlot de la distribución del área condicionando a los distintos daños.

Siendo que no se puede apreciar tan claramente la dispersión, a causa de los valores atípicos, filtramos el 0.3 % de los datos (generando un zoom sobre los BoxPlots).

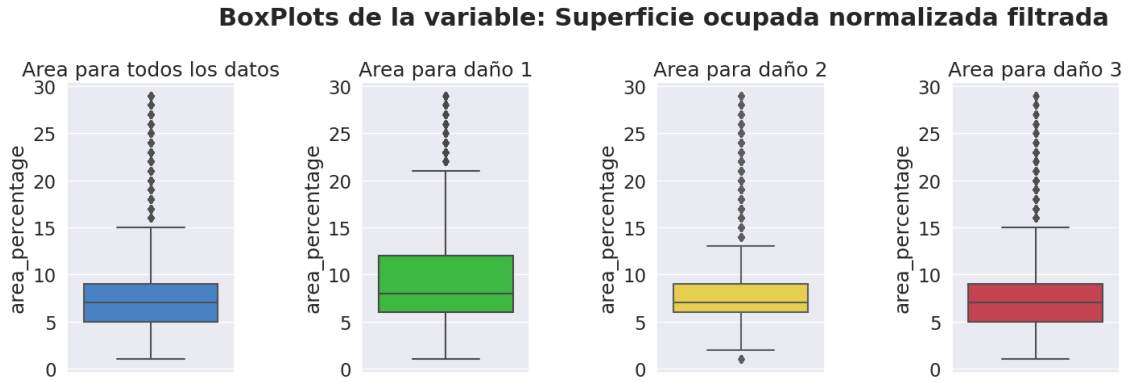

Figura 16: BoxPlot de la distribución del área filtrada condicionando a los distintos daños.

Se ve muy claro en estos gráficos como la región intercuantil para el área condicionado al daño leve es significativamente mayor, mostrando así cómo se modifica la distribución de esta variable aleatoria, si se agrega información sobre el tipo de daño recibido.

5.3.4. Relaciones de las variables numéricas

Una vez examinada la distribución de estas variables, procedemos a investigar sobre las relaciones entre ellas, y principalmente sobre su efecto en el tipo de daño. Con este fin, condicionamos al daño nuevamente y graficamos todos los datos como puntos para relacionar las distintas variables numéricas.

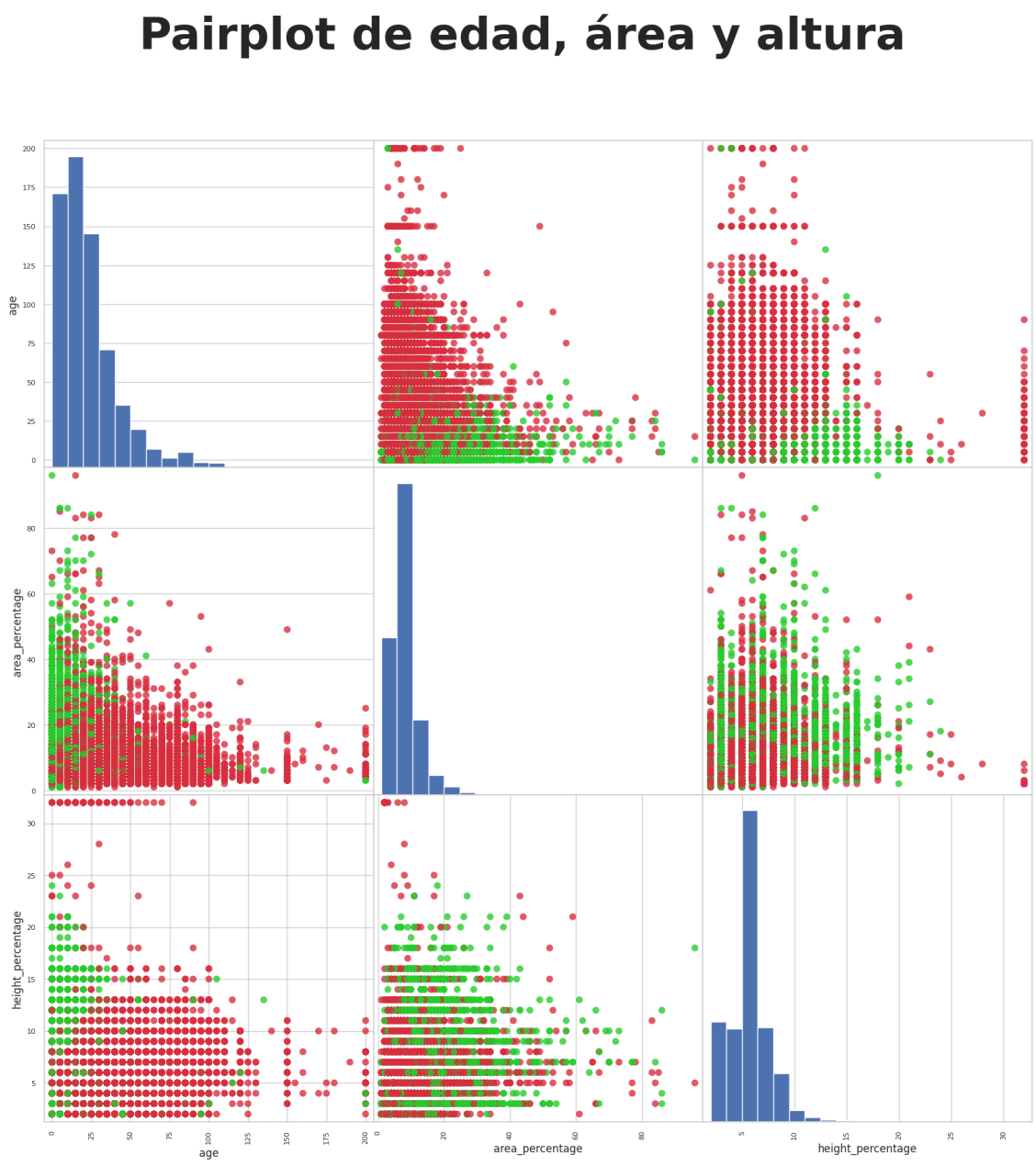

Figura 17: Matriz de gráficos de dispersión e histogramas para las variables de edad, área y altura.

De este gráfico se pueden extraer varias conclusiones, pero primero debemos entender ciertos aspectos del mismo para poder analizarlo. En base a los 3 histogramas notamos como los 3 datos tienen una gran concentración de datos para los casos de menor valor. Como consecuencia de esto, en los scatter plots se observa una gran concentración de puntos en la esquina inferior izquierda y casi ninguno en la superior derecha. A pesar de que ninguno de los datos es continuo con tipo de dato float, nos permitimos graficar histogramas de las tres variables, suponinendo un cuasi-continuo. Si las muestras fueran precisas, los datos verdaderos tanto para la edad, como para el área y para la altura, deberían pertenecer a los números reales.

Una vez visto esto, podemos proceder a analizar cómo, dentro de la distribución de puntos, se distribuyen los daños. Comenzamos viendo que en el gráfico de comparación entre el área y la

altura, los puntos están muy dispersos sin permitirnos llegar a una conclusión clara.

Los dos gráficos restantes los analizaremos juntos ya que vemos cómo su distribución de colores es similar. A mayor edad los puntos rojos aumentan y tanto para la altura como para el área, a mayor valor en los edificios construidos hace poco tiempo, más verde. Con esto concluimos que los edificios modernos están más preparados para combatir los terremotos. Esto podemos intuir que esta unido a un avance en las técnicas de construcción, ya que notamos como la altura los edificios tuvo un gran crecimiento en los últimos años.

El hecho de que a mayor altura pareciera haber más casos de daño leve en relación a los de daño grave nos llama principalmente la atención. En un principio habíamos mostrado que la correlación de la altura con el daño era positiva, ¿por qué pareciera que esto es al revés? La respuesta que le encontramos inicialmente a esta pregunta es que hay poca cantidad de datos con altura mayor a diez, tal como se observa en el histograma, pero esta pregunta trataremos de profundizarla en un análisis posterior.

5.4. Visualización de variables Binarias

En esta etapa analizaremos dos de las familias ya mencionadas: las de Materiales y las de Uso Secundario.

5.4.1. Familia de Materiales de Construcción

En este gráfico mostramos la distribución de probabilidades estimada para cada variable de esta familia, condicionando a si se cumple la correspondiente relación.

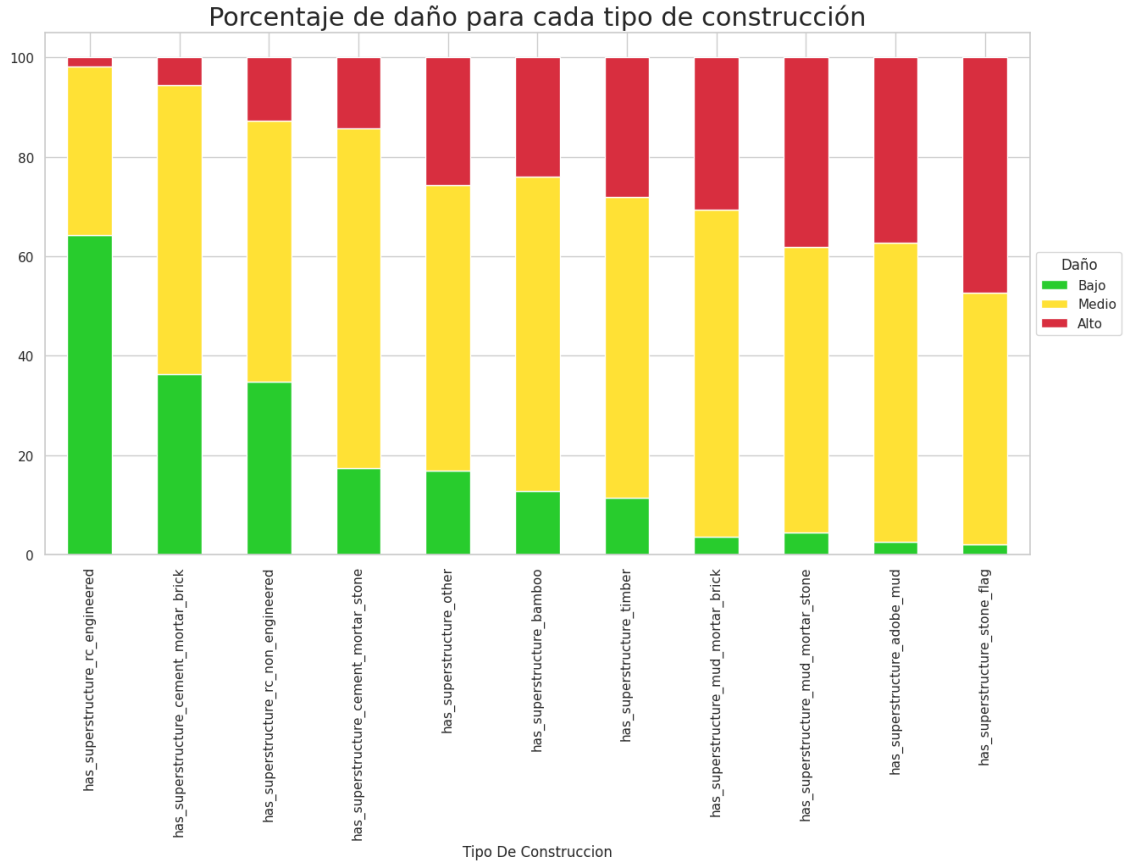

Figura 18: Análisis cualitativo del daño según el tipo de construcción

Este primer gráfico sobre las distintas condiciones nos permite extraer ciertas variables que claramente tienen una distribución muy distante del promedio general. Tal como era de esperarse la condición de tener una esctructura anti sismos, tanto ingenieril, como no ingenieril, generan un impacto positivo sobre la resistencia de la construcción.

En contraste, las estructuras hechas con piedra; piedra y ladrillos; y adobe y barro tienen un efecto claramente negativo sobre el impacto del terremoto.

Evaluamos con el siguiente gráfico las estructuras más representativas en cuánto a la cantidad de muestras en el set de datos.

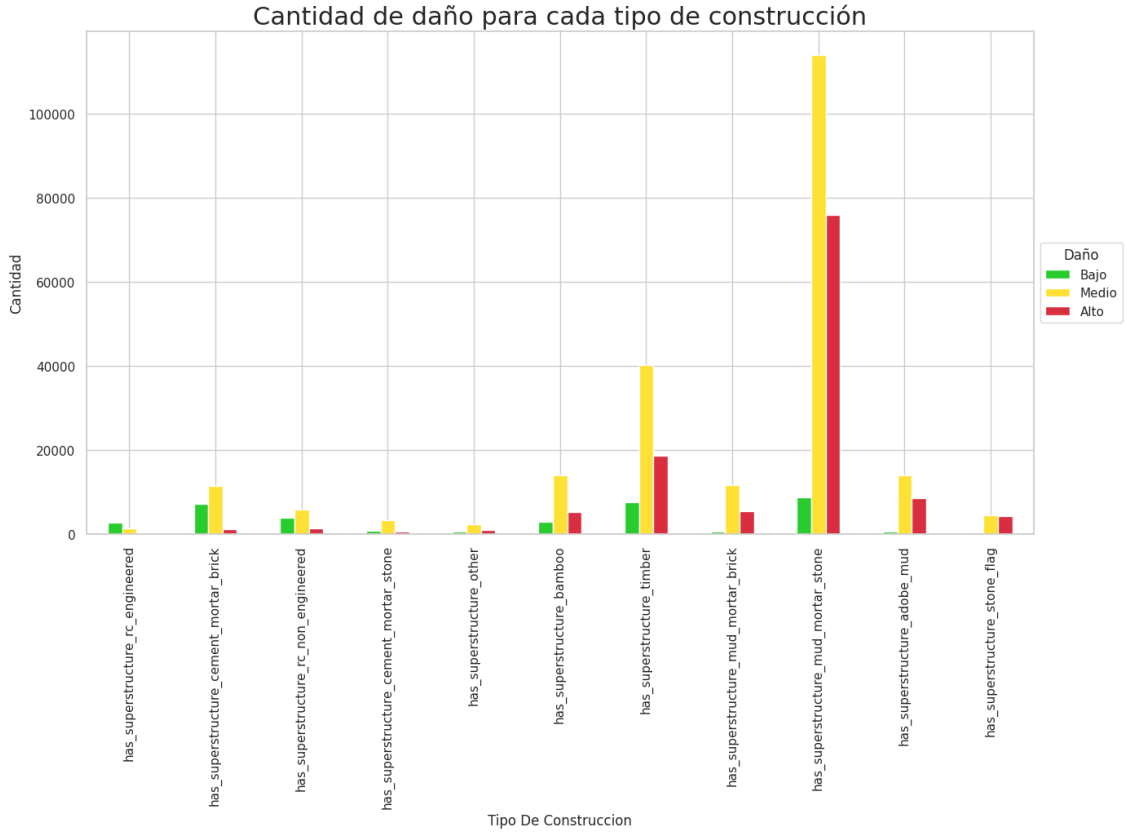

Figura 19: Análisis cuantitativo del daño según el tipo de construcción

Habiendo observado fijamente cada una de estas representaciones, concluimos:

- Todas las variables son representativas.
- Las estructuras que fracasan son: la madera; barro, mortero y ladrillos; adobe y barro; y piedra.
- Los casos de éxito son: las estructuras anti-sismos tanto ingenieriles como no ingenieriles y las estructuras de cemento y ladrillo.

5.4.2. Familia de Uso Secundario

A modo general, generamos un gráfico similar al previo mostrando el impacto del daño sobre los distintos usos secundarios.

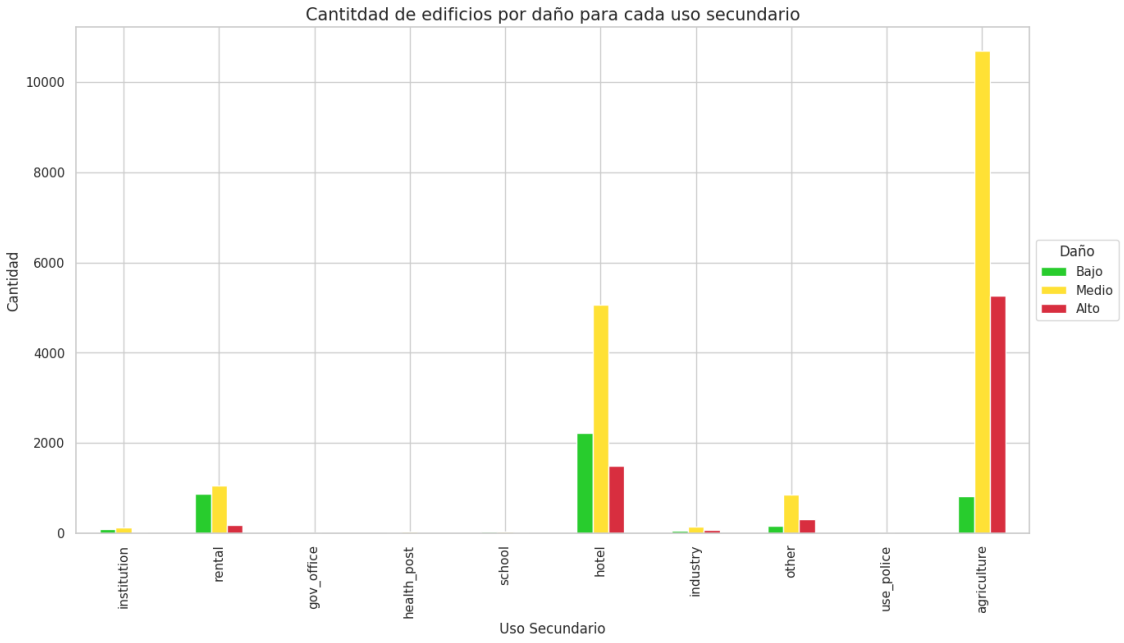

Figura 20: Análisis cuantitativo del daño según el uso secundario

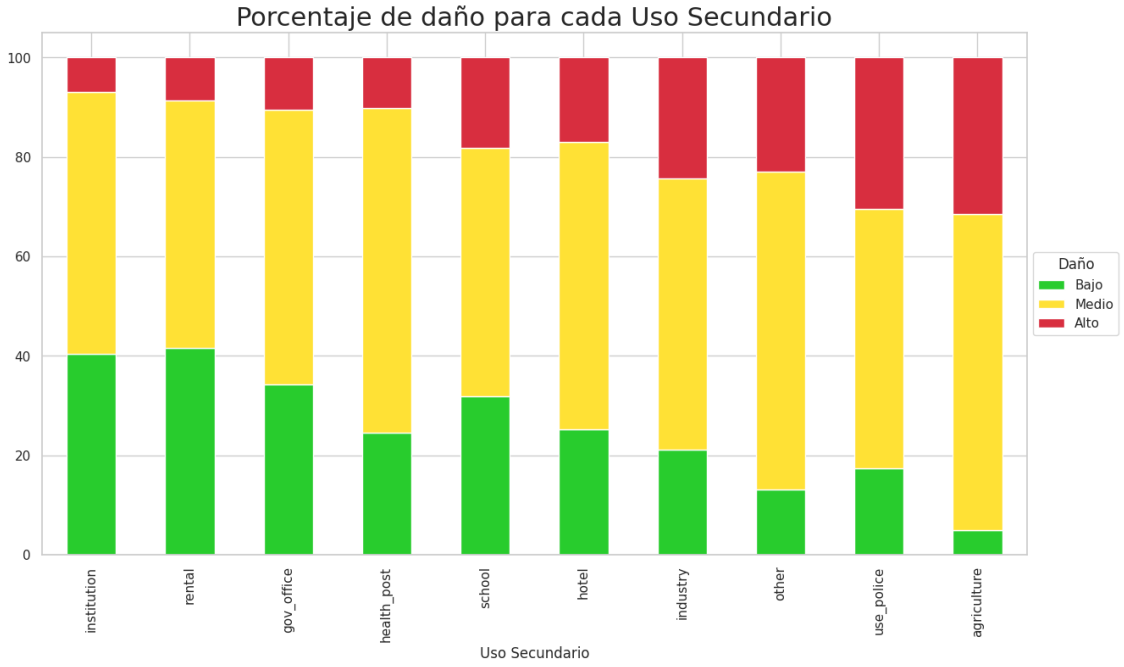

Figura 21: Análisis cualitativo del daño según el uso secundario

Se pueden observar muchas variables que cuando cumplen su condición, mejoran la probabilidad de que el daño sea menor. Entre estas se encuentran los hoteles, los edificios de renta, las instituciones, los colegios, las industrias y las oficinas de gobierno. Sin embargo, en el primer gráfico cuantitativo se observa que los edificios que cumplen tener un uso secundario representan a un porcentaje muy disminuido sobre la población total. Los únicos que podemos considerar como significativos son los de renta, los de hotel, agricultura y los usos otros.

A partir de los gráficos podemos observar que en las edificaciones que tienen algún tipo de uso secundario, el porcentaje de daño mayor es 10 % menor que en las que no tiene uso secundario.

Analizando las relaciones mostradas por los presentes gráficos llegamos a las siguientes conclusiones:

- Las usos secundarios significativos son: los de agricultura, de hotelería, de renta y la variable que incluye a los tipos que no son caracterizables en una de las otras columnas.
- El caso de fracaso es el uso para la agricultura.
- Los casos de éxito son los de hotelería y renta.

5.5. Visualización de variables categóricas

En esta sección intentaremos comprender todas las variables categóricas puras (solo se pueden considerar como categóricas y de ninguna manera como numéricas) del set de datos. En este sentido, graficaremos para cada variable que la cantidad de categorías lo permita, un gráfico cualitativo y otro cuantitativo de los datos.

5.5.1. Tipo de cimientos usados en planta baja

En primer lugar, para la variable del tipo de cimientos usados en la planta baja (*ground_floor_type*) podemos apreciar que el valor *f* es el que más se repite con mucha diferencia por sobre el resto.

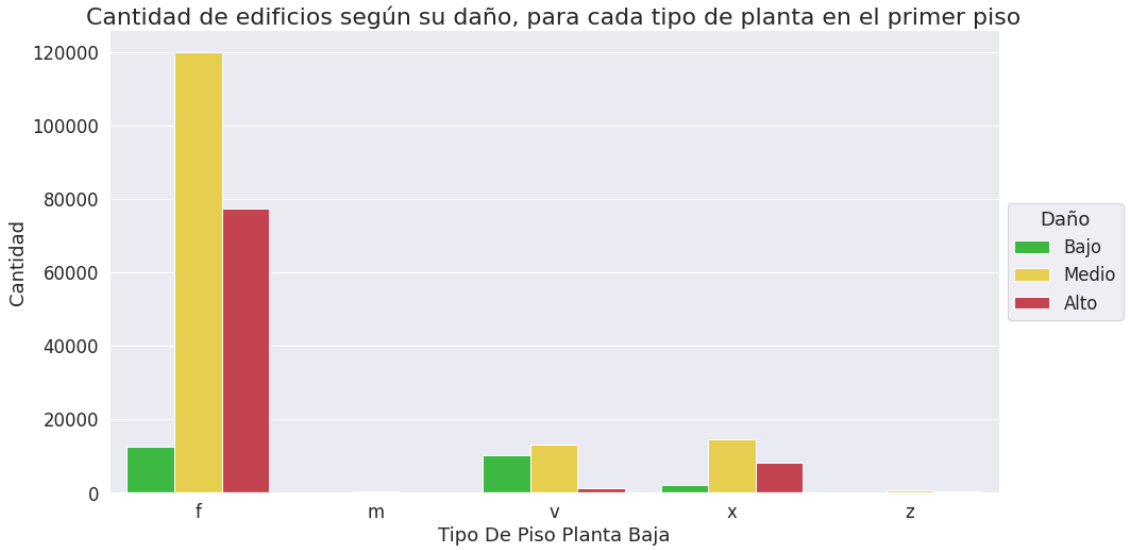

Figura 22: Análisis cuantitativo del daño según el tipo de piso de planta baja

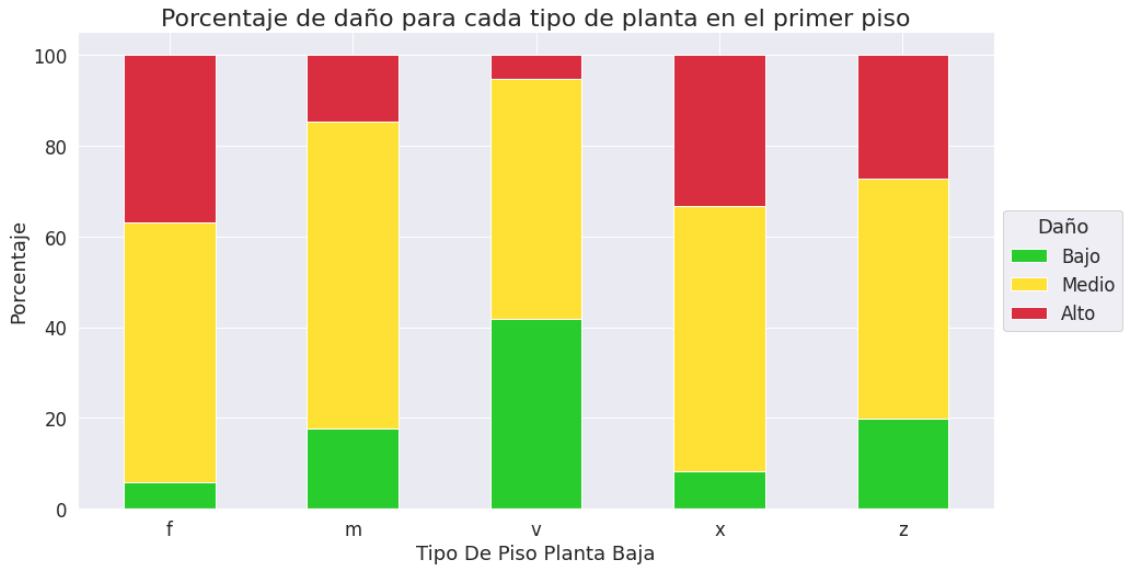

Figura 23: Análisis cualitativo del daño según el tipo de piso de planta baja

Estos dos gráficos nos permiten concluir lo siguiente.

- f es un caso de fracaso, y es muy representativa.
- x y v son dos casos de éxito, y son representativas.
- m y z no son representativas.

5.5.2. Tipo de cimientos usados en otras plantas

De igual modo para el tipo de pisos usados en otras plantas, se observa que hay un único valor que tiene la gran mayoría de los datos: q. Sin embargo, en este caso, hay al menos 10000 datos de cada categoría.

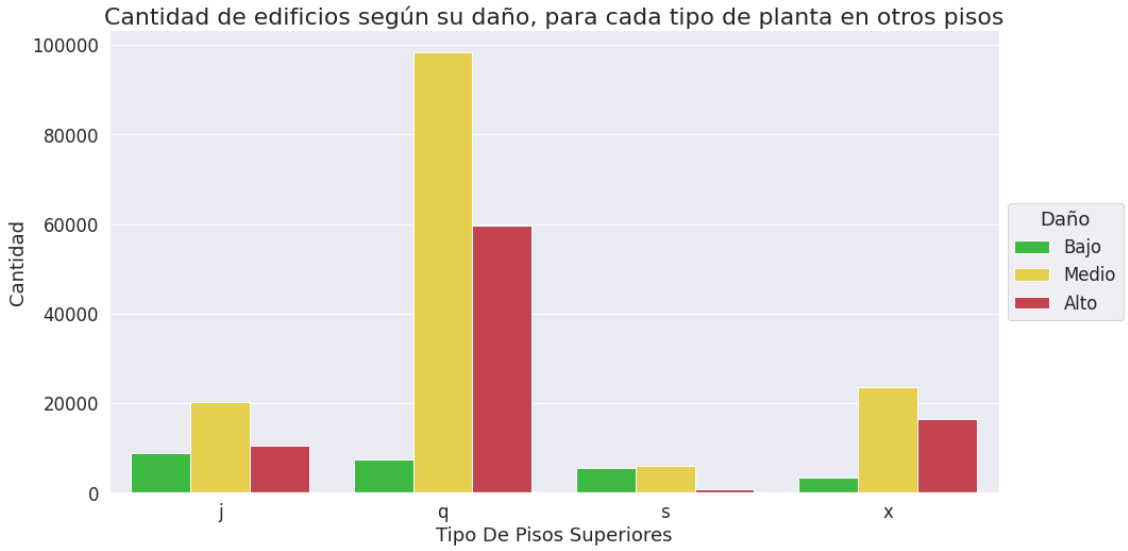

Figura 24: Análisis cuantitativo del daño según el tipo de piso de los pisos superiores

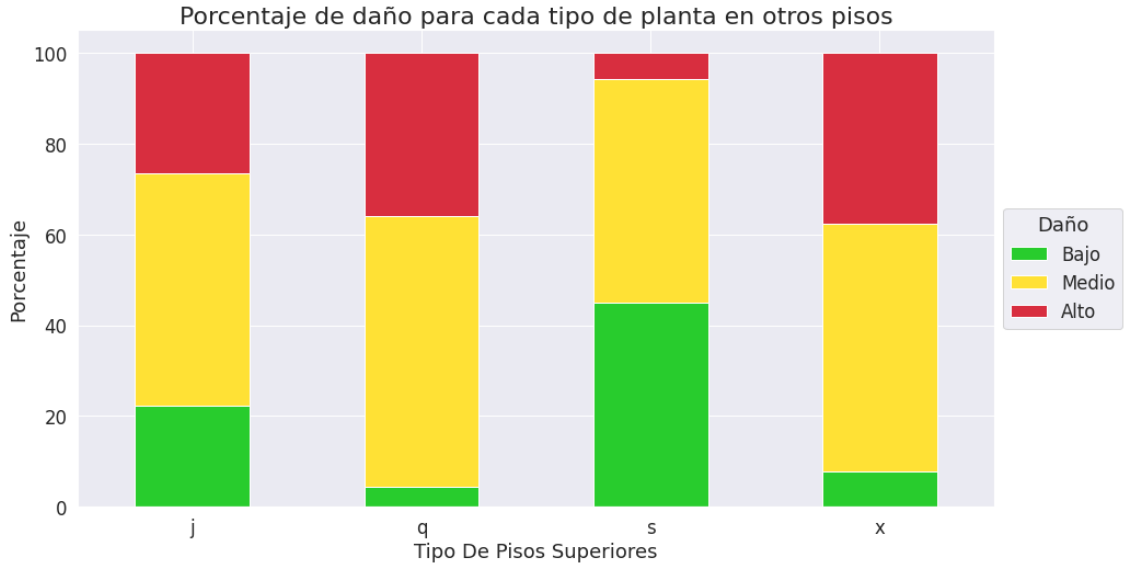

Figura 25: Análisis cualitativo del daño según el tipo de piso de los pisos superiores

Analizando ambos gráficos se puede concluir:

- Todas las variables son representativas.
- q y x son los casos de fracaso.

- s es el caso de éxito.
- j no podemos concluir nada.

5.5.3. Orientación del edificio

Analizamos el porcentaje de daño y también la cantidad de edificios por daño para cada valor posible de la orientación.

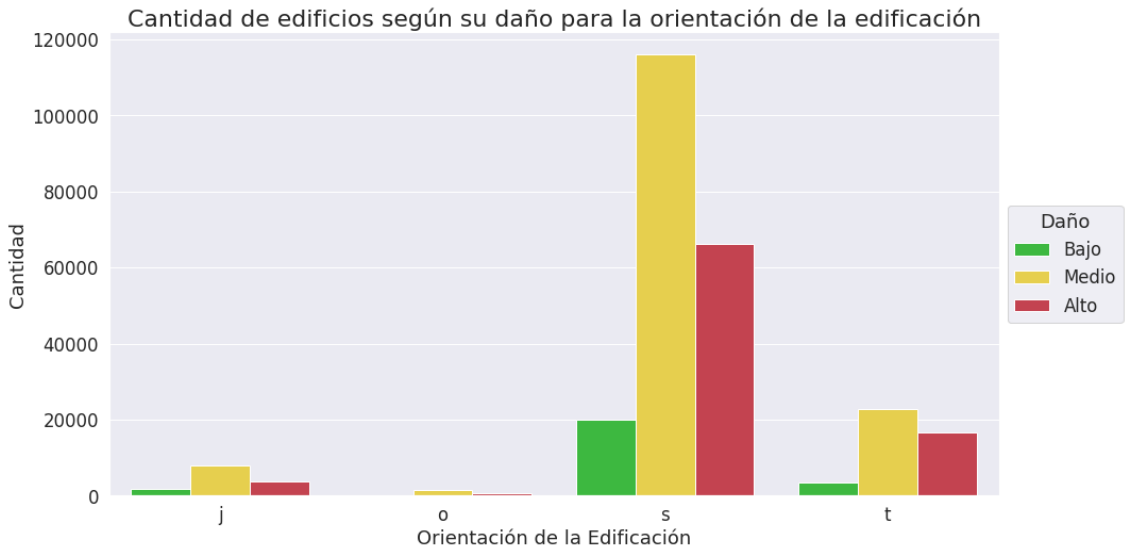

Figura 26: Análisis cuantitativo del daño según la orientación del edificio

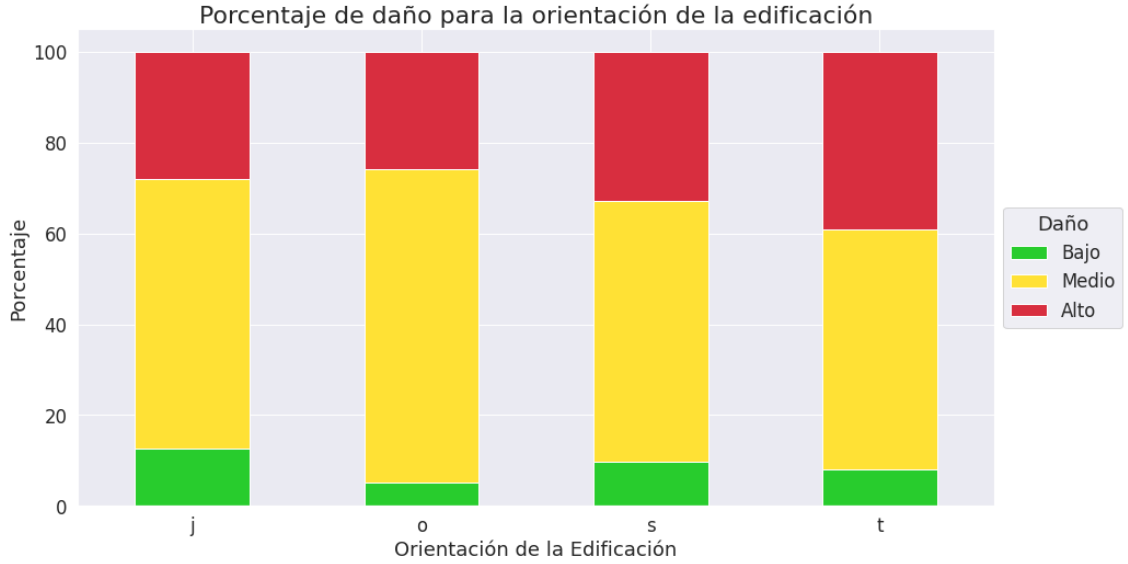

Figura 27: Análisis cualitativo del daño según la orientación del edificio

Se observa nuevamente que una sola categoría, en este caso la *s*, tiene la gran mayoría de los datos. Sin embargo, tal como fue anticipado cuando se analizó la correlación de cada una de estas categorías con la variable Y, la orientación no pareciera generar un gran cambio en la distribución de las probabilidades.

5.5.4. Formato de construcción de la edificación

En el siguiente HeatMap, el cual se complementa con los otros dos gráficos de barras cualitativos y cuantitavios respectivamente, muestra cómo se comporta la variable en cuestión en relación al daño. Para la generación del mismo se filtraron directamente aquellas categorías con menos de 100 muestras, evitándonos cometer el error más peligroso de la historia.

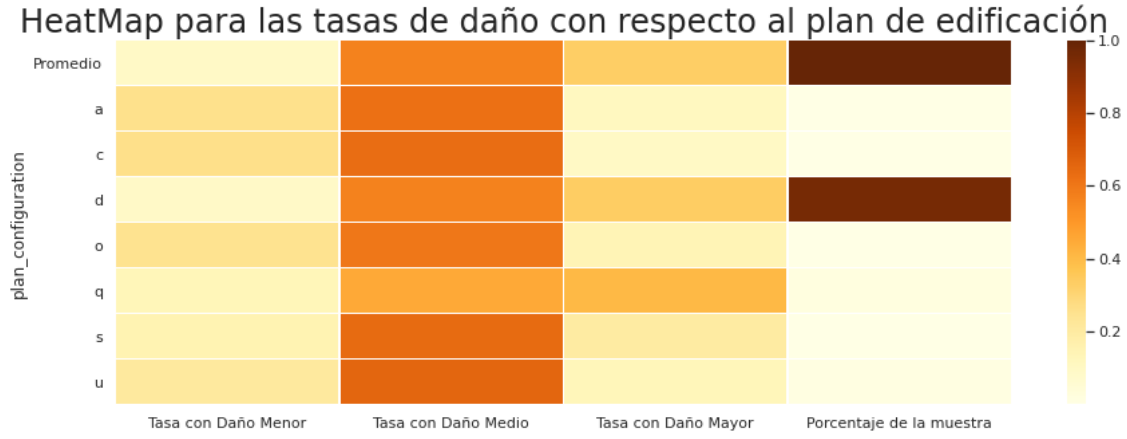

Figura 28: HeatMap para la distribución de probabilidades de cada categoría del formato de construcción.

Se observa claramente cómo las mejores estructuras antisismos que protegen de un daño total son las del tipo *a* y *c*, seguidas por el tipo *u* (colores más claros en la tercera columna). El tipo *o* pareciera tener una buena estructura también. Por el otro lado, las estructuras del tipo *q* y *d* tienen una tasa de destrucción fuerte mayor o igual a la media.

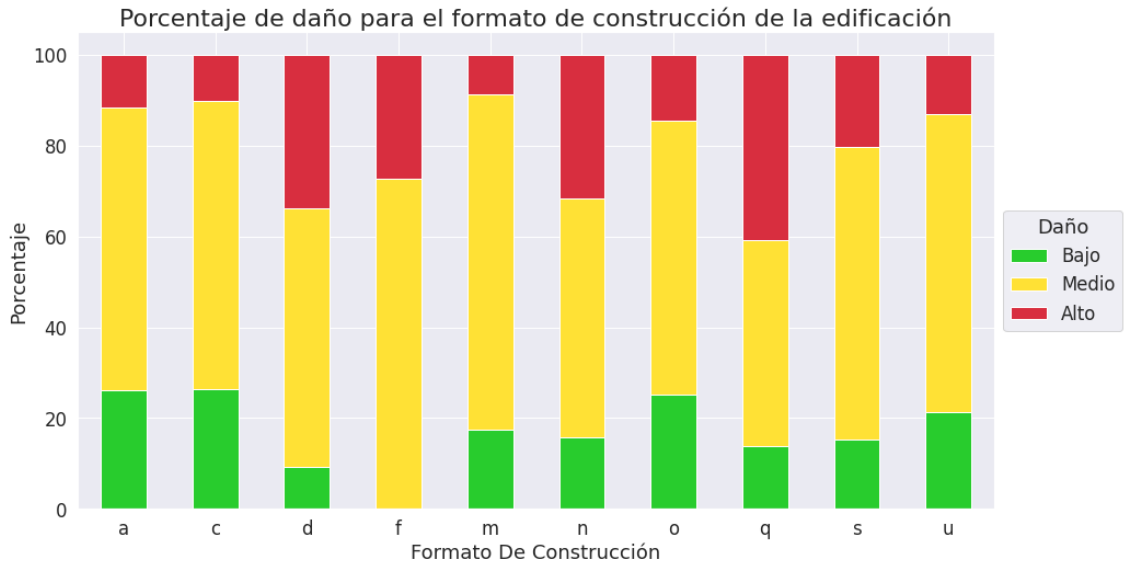

Figura 29: Análisis cualitativo del daño según el formato de construcción

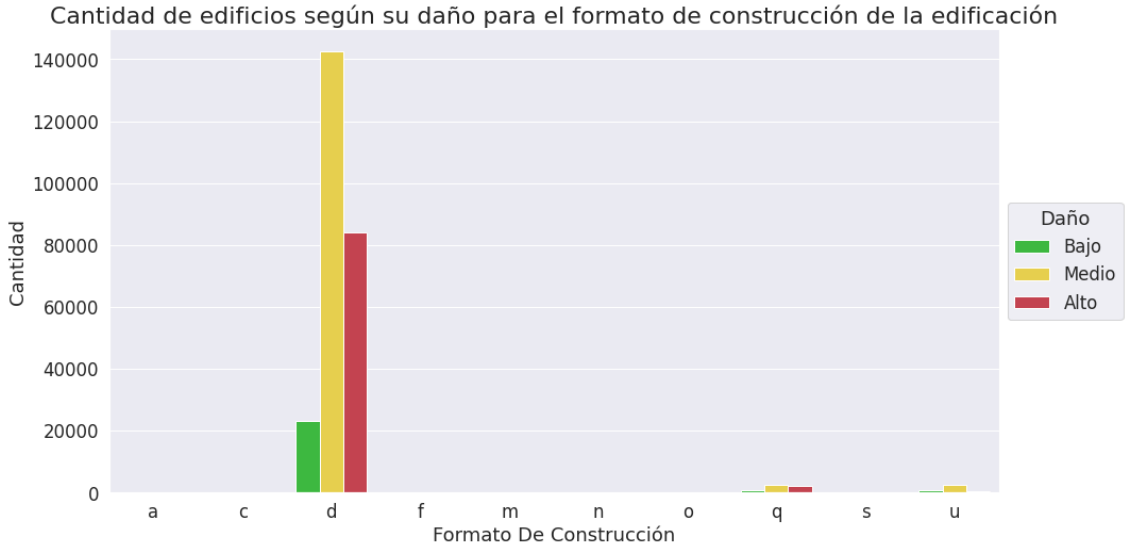

Figura 30: Análisis cuantitativo del daño según el formato de construcción

Vale la pena mencionar que el tipo d es el más representativo, al contar con un 96 % de los datos totales. Es entonces coherente que su distribución de tasas sea muy similar a la del promedio total.

A pesar de haber observado ciertos comportamientos favorables o desfavorables, no podemos asegurar que haya casos de éxito y fracaso claros, al tener todos una distribución bastante similar.

5.5.5. Tipo de techo usado en la construcción de la edificación

Por otro lado, el tipo de techo tiene datos significativos para todas sus clases.

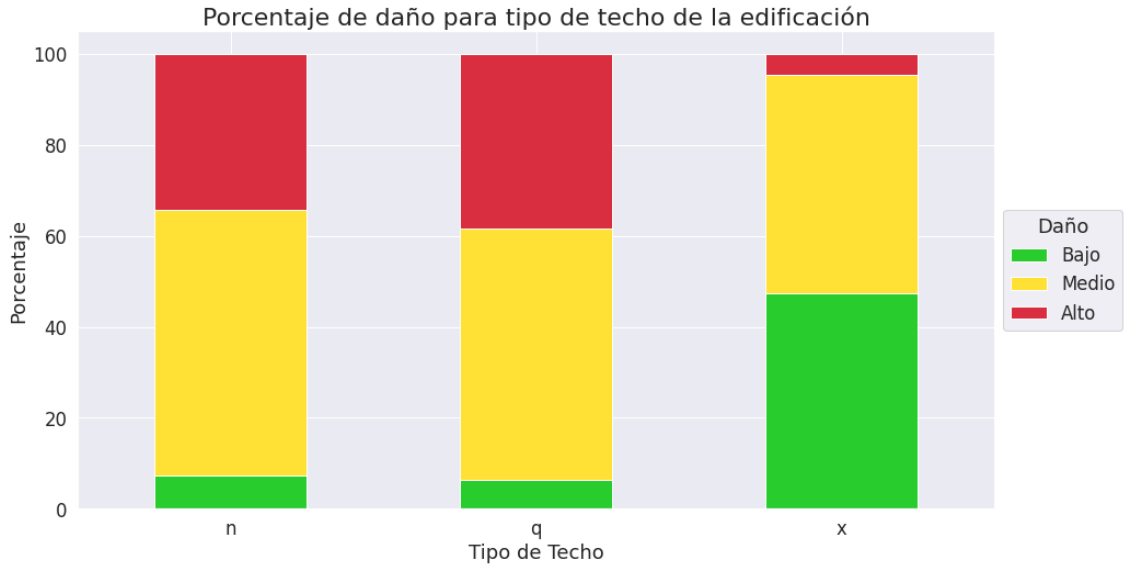

Figura 31: Análisis cualitativo del daño según el tipo de techo.

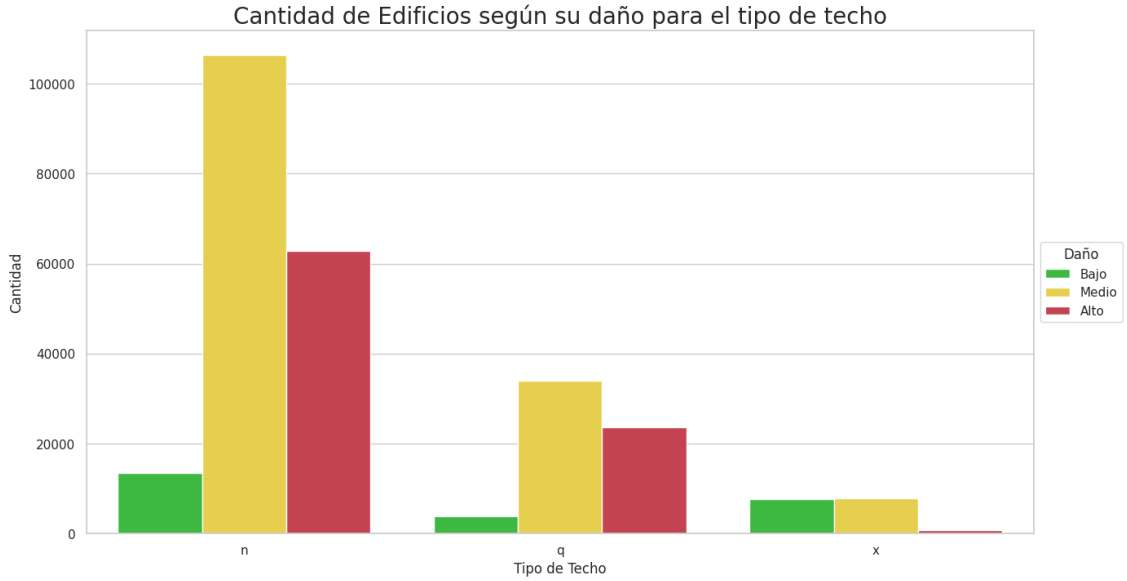

Figura 32: Análisis cuantitativo del daño según el tipo de techo.

Observando los presentes gráficos llegamos a las siguientes conclusiones:

- Todas las variables son representativas.
- El caso de éxito es x.
- Los casos de fracaso son n y q.

5.5.6. Tipo de cimientos usados cuando se construyó la edificación

A continuación realizamos el mismo análisis para el tipo de cimientos.

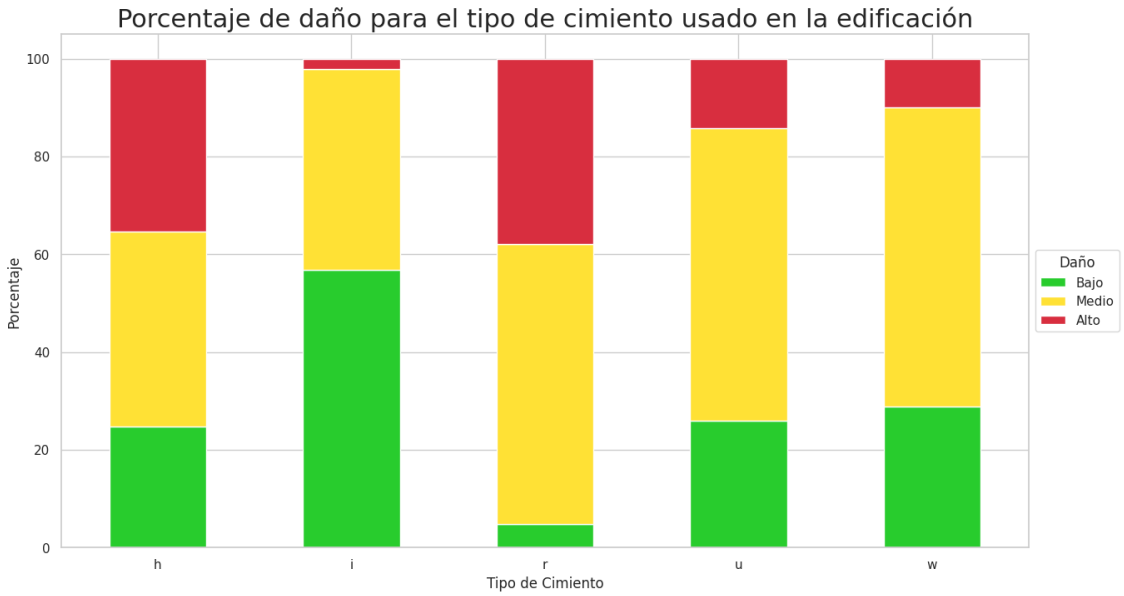

Figura 33: Análisis cualitativo del daño según el tipo de cimiento

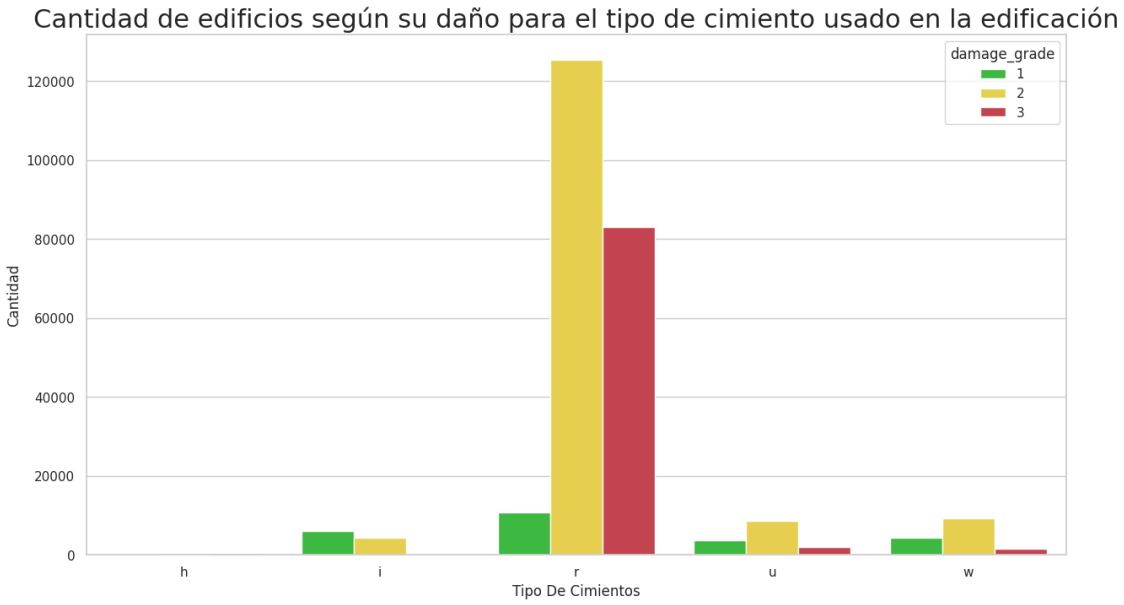

Figura 34: Análisis cuantitativo del daño según el tipo de cimiento

Los gráficos nos permiten afirmar:

- h no es una variable representativa.

- r es un caso de fracaso.
- i, u y w son casos de éxito.

5.5.7. Condición de la tierra de la edificación en su construcción

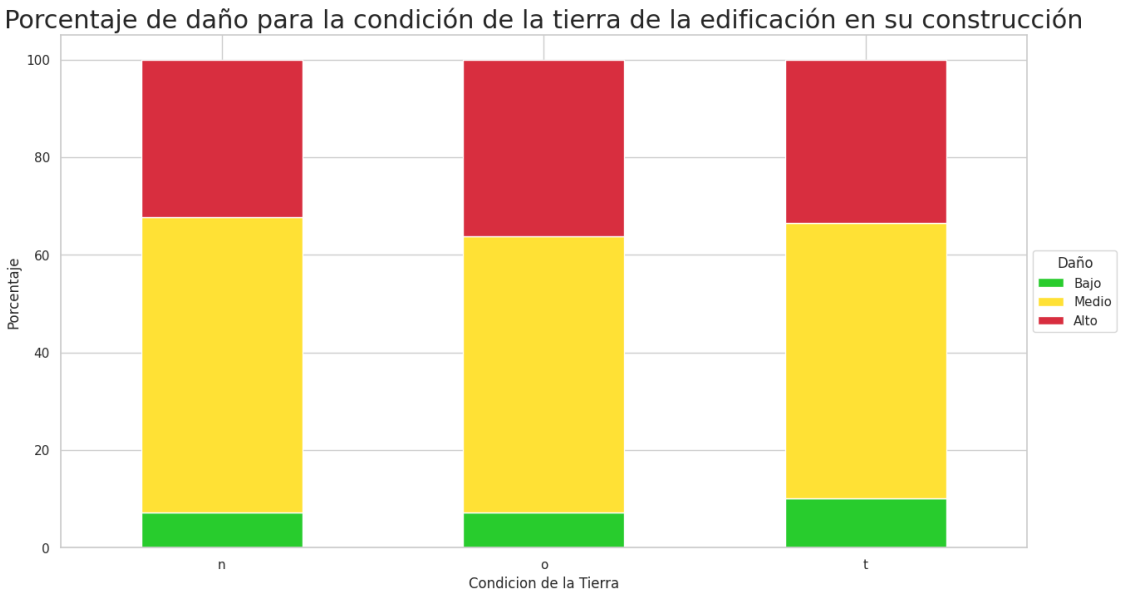

Figura 35: Análisis cualitativo del daño según la condición de la tierra

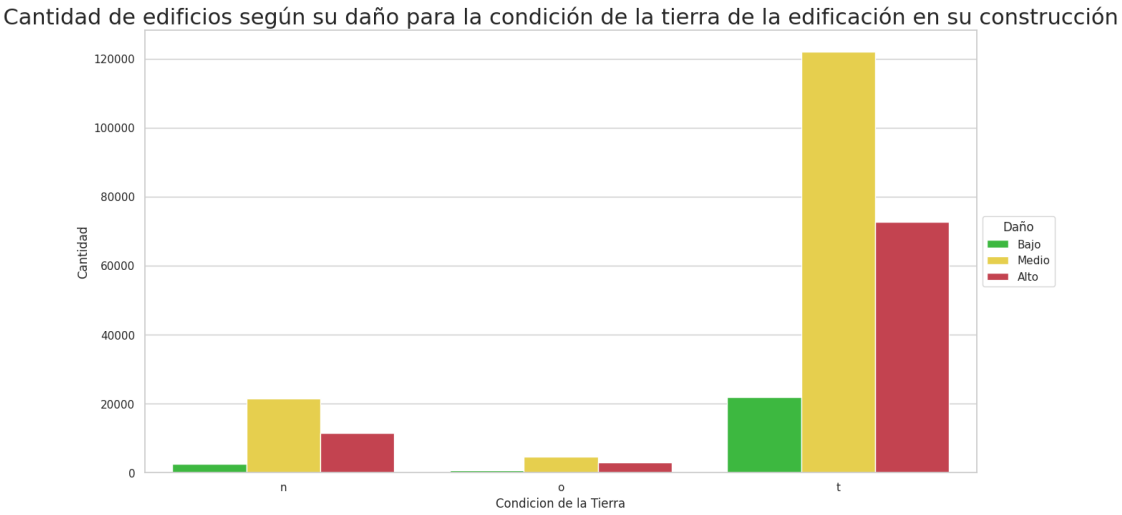

Figura 36: Análisis cuantitativo del daño según la condición de la tierra

Concluimos:

- Todas las variables son representativas, pero ninguna modifica el promedio general del daño.
- A priori, no las incluimos entre las columnas más importantes.

5.5.8. Estado legal de la tierra

Por último, presentamos los gráficos para el estado legal de la tierra

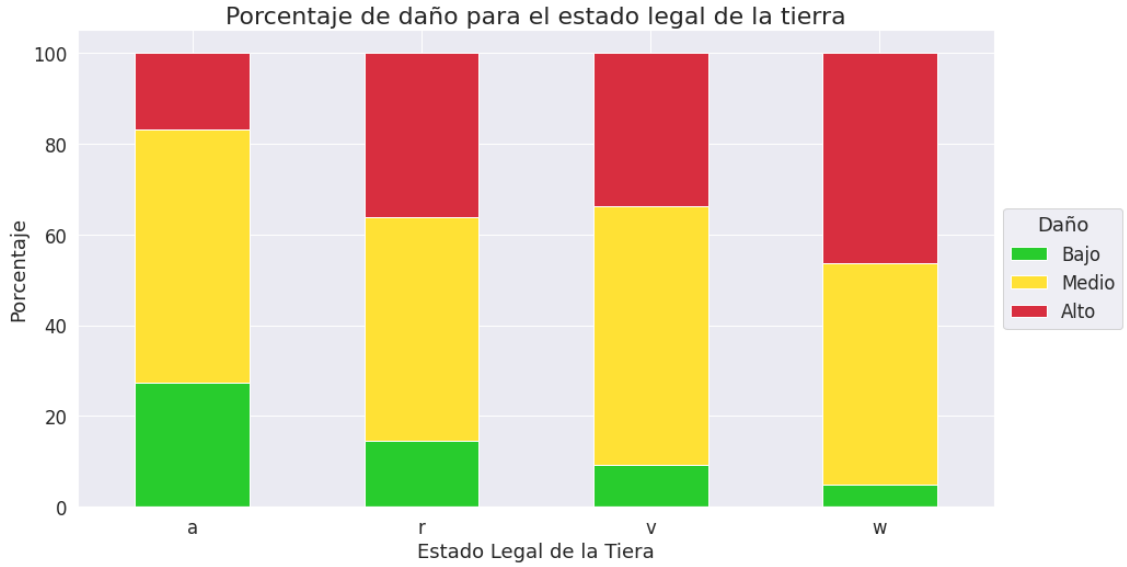

Figura 37: Análisis cualitativo del daño según el estado legal de la tierra

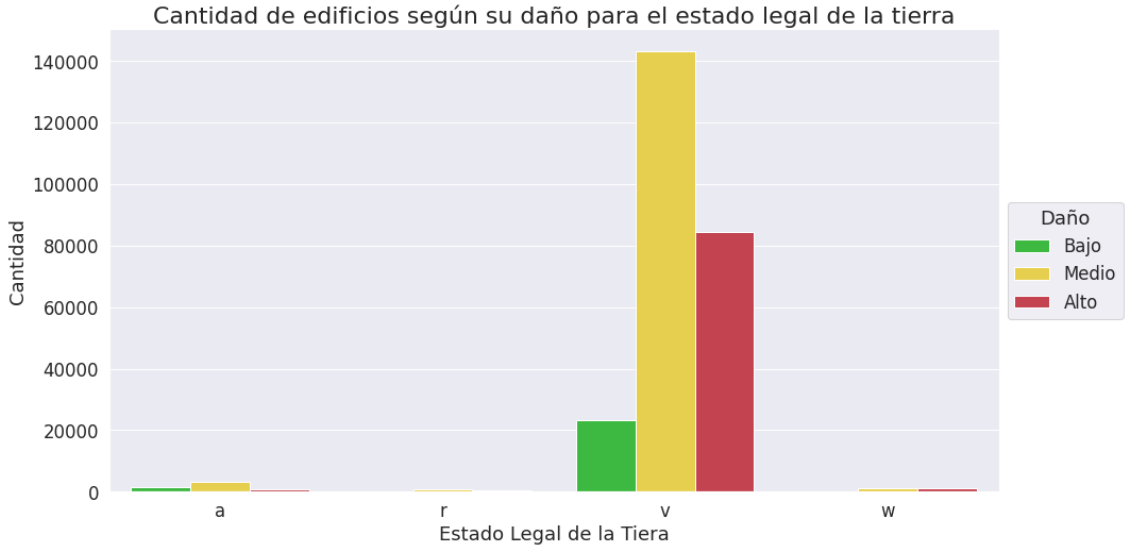

Figura 38: Análisis cuantitativo del daño según el estado legal de la tierra

Podemos observar que la única variable representativa es v y tal como lo predijimos en nuestras hipótesis, no modifica el promedio general de daño.

Concluimos entonces:

- No hay variable representativa que modifique el promedio general de daño.
- A priori, no tendremos en cuenta esta columna.

5.5.9. Zona Geográfica

La presente variable la consideramos como categórica, y no como numérica, ya que al generar los gráficos de correlaciones y buscando otras relaciones, no encontramos ningún ordenamiento entre las mismas, ni cierto patrón. En una primera instancia observamos que todas las regiones tienen una cantidad de datos bastante significativa. Por lo tanto, observamos la distribución de probabilidades de daño aproximada a partir de la muestra para cada uno de estos daños.

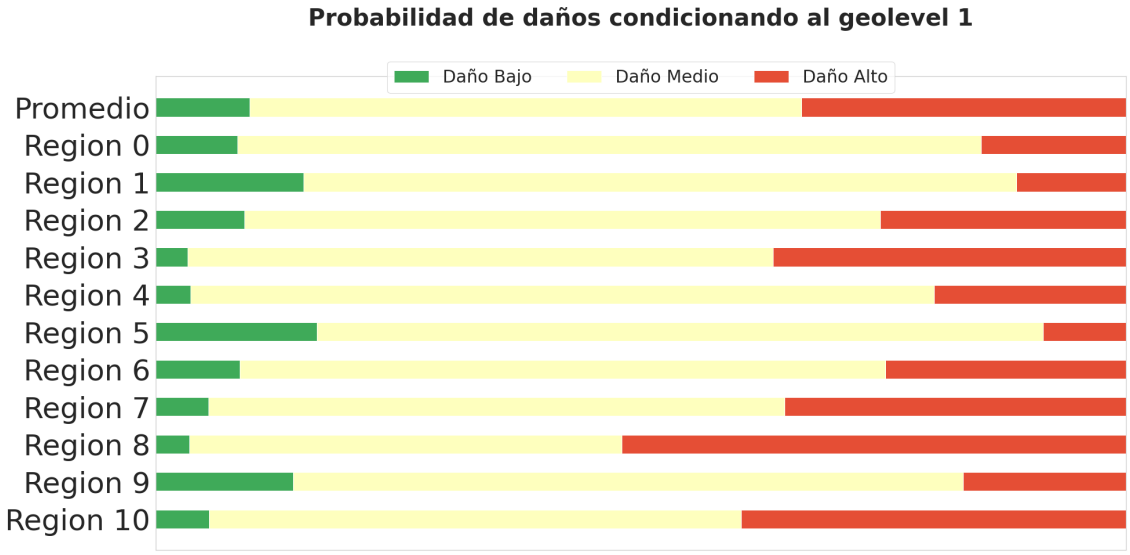

Figura 39: Análisis cualitativo del daño para las primeras 10 regiones de *geo_level 1*

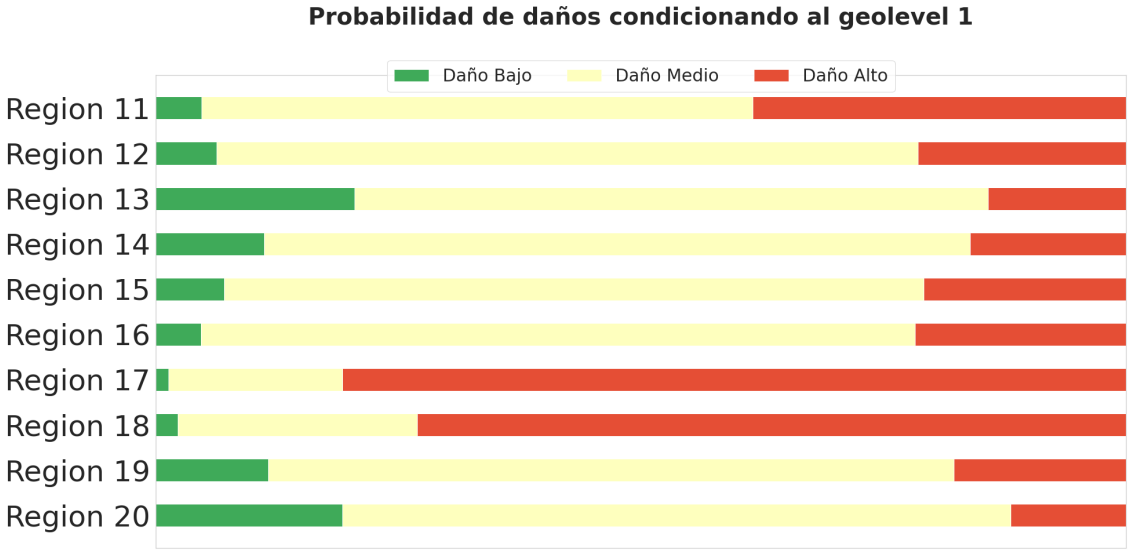

Figura 40: Análisis cualitativo del daño para las 10 regiones intermedias de *geo_level 1*

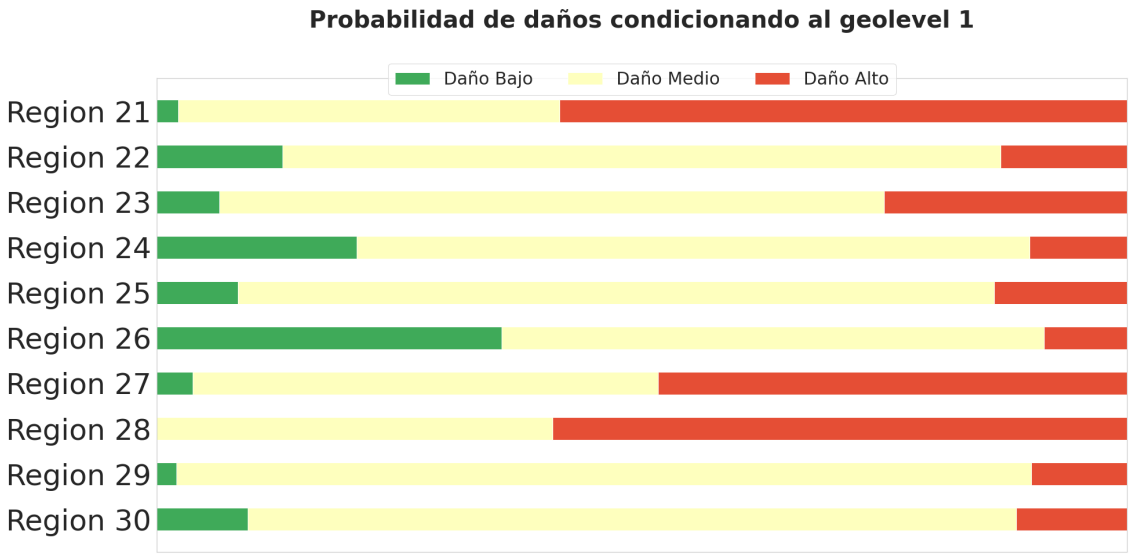

Figura 41: Análisis cualitativo del daño para las ultimas 10 regiones de *geo_level 1*

Este gráfico nos hace considerar a las regiones como una de las variables más relevantes para un futuro modelo. Se observa mucha variabilidad entre los datos, y por lo tanto suponemos una fuerte dependencia entre la zona geográfica y el daño recibido.

5.6. Conclusiones iniciales

Dedicamos esta sección a referirnos sobre los análisis realizados en la visualización individual de cada familia de variables.

Lo que concluimos en cada caso sirvió para generarnos una columna en nuestro dataset auxiliar para casos de éxito y/o fracaso de cada variable (si es que presentan) y para unificar columnas con una fuerte relación.

El resumen de las primeras conclusiones son:

- *Altura de los edificios* ->Unificamos la altura con la cantidad de pisos por su fuerte correlación. Una primera idea que obtuvimos es que a mayor altura, menor es el daño.
- *Edad* ->Mientras más longevos son, mayor es el daño. Para los edificios con 995 años pudimos apreciar que se deben a un error de carga en los datos.
- *Porcentaje de área* ->A mayor área, menor es el daño
- *ground_floor_type* ->f caso de fracaso. x y v caso de éxito. El resto no los tomamos.
- *other_floor_type* ->s es caso de éxito. q y x son caso de fracaso.
- *position* ->f y s son fracaso. No hay caso de éxito.
- *plan_configuration* ->d es la única representativa y es un caso de fracaso claro.
- *legal_ownership_status* ->Nada es representativo para observar un cambio en el daño.
- *foundation_type* ->r es fracaso. i, u y w son casos de éxito. *item land_surface_condition* ->Nada es representativo.
- *has_secondary_use* ->*has_secondary_use_agriculture* es fracaso. *has_secondary_use_hotel*, *has_secondary_use_rental* es éxito.

- *has_ superstructure ->has_ superstructure_ timber*, *has_ superstructure_ mud_ mortar_ brick*, *has_ superstructure_ adobe_ mud*, *has_ superstructure_ stone_ flag* son de fracaso. *has_ superstructure_ rc_ engin*
  has_ superstructure_ cement_ mortar_ brick, *has_ superstructure_ rc_ non_ engineered* son de éxito.

6. Análisis de preguntas e hipótesis planteadas

En esta sección vamos a estudiar las preguntas e hipótesis que nos planteamos. A su vez, nos pueden ir surgiendo nuevas inquietudes que pueden terminar en relaciones más provechosas que las iniciales.

6.1. Diferencias entre los casos de éxitos y fracaso según su ubicación

En función de relacionar las variables categóricas con la zona geográfica, tomamos por separado las tres zonas con los casos más extremos. Estas son: la región 17, con el mayor porcentaje de daño grave; la región 30, con mayor porcentaje de daño medio; y la región 26 con el mayor porcentaje de daño leve. En rigor de verdad, la 29 tiene mayor porcentaje de daño medio. Sin embargo, la descartamos para este análisis al no ser tan significativa en cuanto a la cantidad de datos.

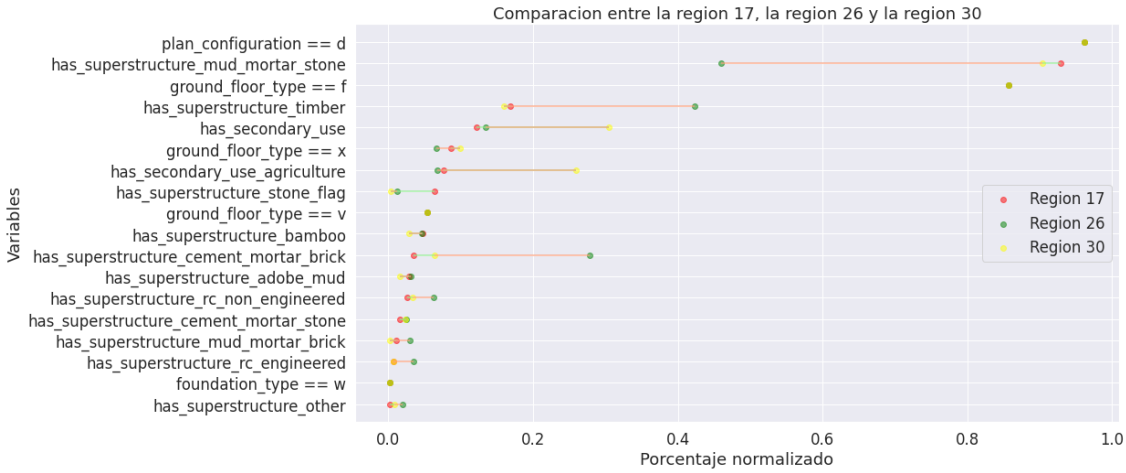

Figura 42: Analisis variables binarias segun geo_level_1_id

El presente gráfico nos permite comparar la distribución de variables binarias con respecto a la región en la que están: la 26 (verde) es la región con mayor porcentaje de edificios con daño menor; la 17 (rojo), con daño mayor; y la 30 (amarillo) con daño medio. En rigor de verdad, la 29 tiene mayor porcentaje de daño medio. Sin embargo, la descartamos para este análisis al no ser tan significativa en cuanto a la cantidad de datos.

La categoría en la que se observa mayor diferencia entre las regiones es la estructura de barro y piedra. Es previsible que la zona con mayor daño, tenga mayor cantidad de edificios en esta categoría. En contraposición, para la zona 26 hay una mucho mayor proporción de estructuras construidas con cemento y ladrillos.

Es interesante observar también que no siempre la región 30 está entre medio de las otras dos, como podría esperarse para variables que son significativas. Particularmente la zona 30 tiene una gran cantidad de espacios de agricultura (y por lo tanto mayor cantidad de espacios con uso secundario), en comparación con las otras.

Al haber muy pocos datos que cumplan las condiciones para las ultimas filas del gráfico, los puntos suelen estar muy cerca.

Conclusión: Podemos apreciar que la verdadera diferencia reside en la variable *has_ superstructure*, que ya empezamos a notar que es un caso especial. Además, habíamos notado que esta variable no tenia casos de éxito apreciables, por lo tanto es aún más provechosa para nuestro análisis.

6.2. Análisis de los casos de éxito para los edificios más altos

Queremos descomponer los casos de éxito para los edificios más altos, ya que por hipótesis entendemos que la altura tendría que influir negativamente, por lo tanto puede haber otro factor que condicione la conclusión del apartado de la altura.

Comenzaremos estudiando cómo se relacionan los casos exitosos de la altura con la variable *foundation_type*

Figura 43: Análisis del efecto de la variable cantidad de pisos sobre el daño según cimientos

Apreciamos una relación fuerte entre los casos de éxito de la altura y *foundation_type*, por lo tanto seguiremos analizando para más variables si pasa lo mismo.

Con el riesgo de sonar repetitivos, realizaremos varios gráficos similares ya que estos nos permitieron una evolución de cómo entendimos estas variables.

Figura 44: Análisis del efecto de la variable cantidad de pisos sobre el daño según materiales

Figura 45: Análisis del efecto de la variable cantidad de pisos sobre el daño según tipo de piso en planta baja

Figura 46: Análisis del efecto de la variable cantidad de pisos sobre el daño según tipo de piso en planta baja

Figura 47: Análisis del efecto de la variable cantidad de pisos sobre el daño según su uso secundario

Vemos una gran similitud de todos los gráficos que realizamos, por lo tanto aquí nos detendremos y examinaremos la correlación entre los casos exitosos.

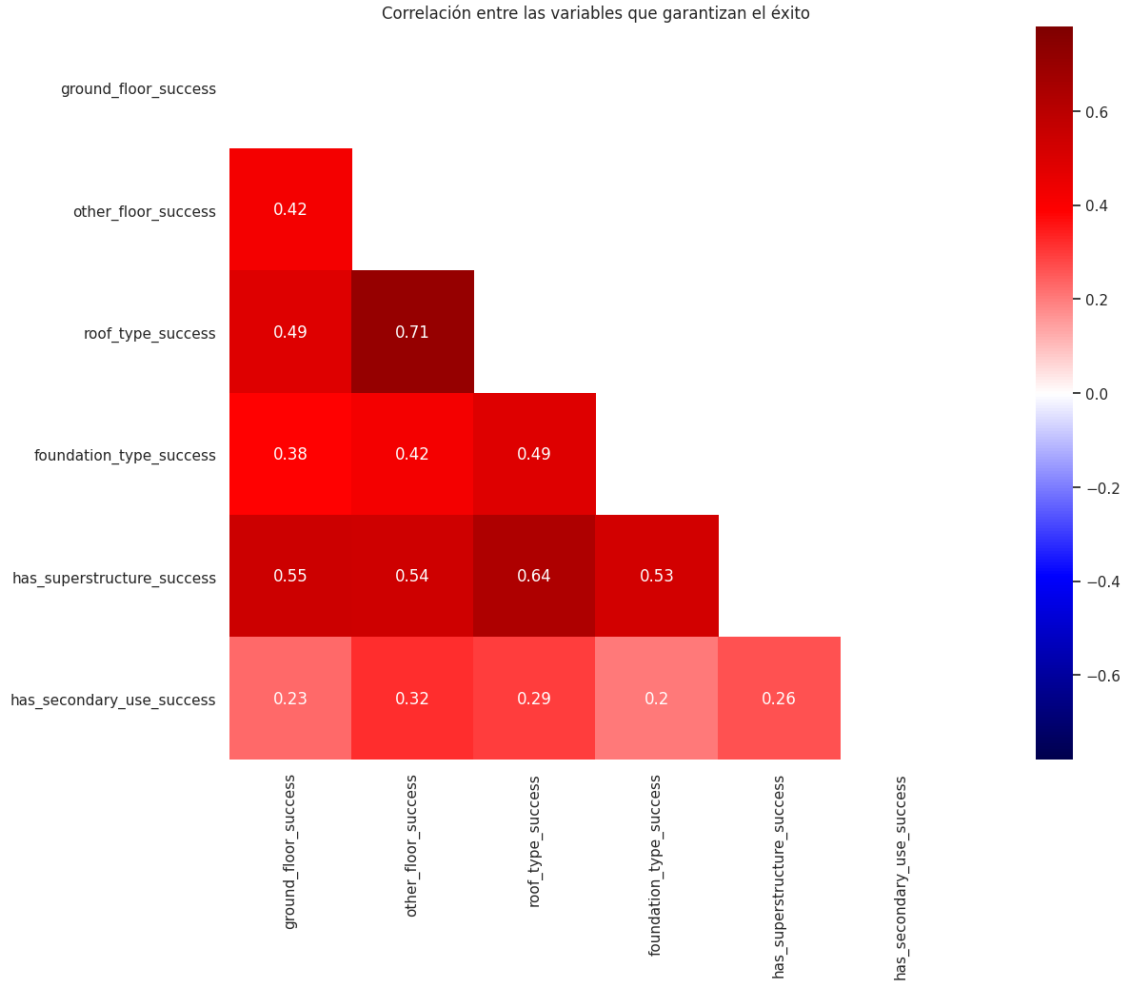

Figura 48: Correlación entre casos exitosos

Apreciamos una fuerte correlación entre la mayor parte de los casos de éxito. Es decir, es muy probable que si una edificación tiene un caso de éxito, tiene algún caso de éxito más.

Esto se corrobora con la correlación Pearson estimada más arriba.

Examinando lo mismo para los casos de fracaso:

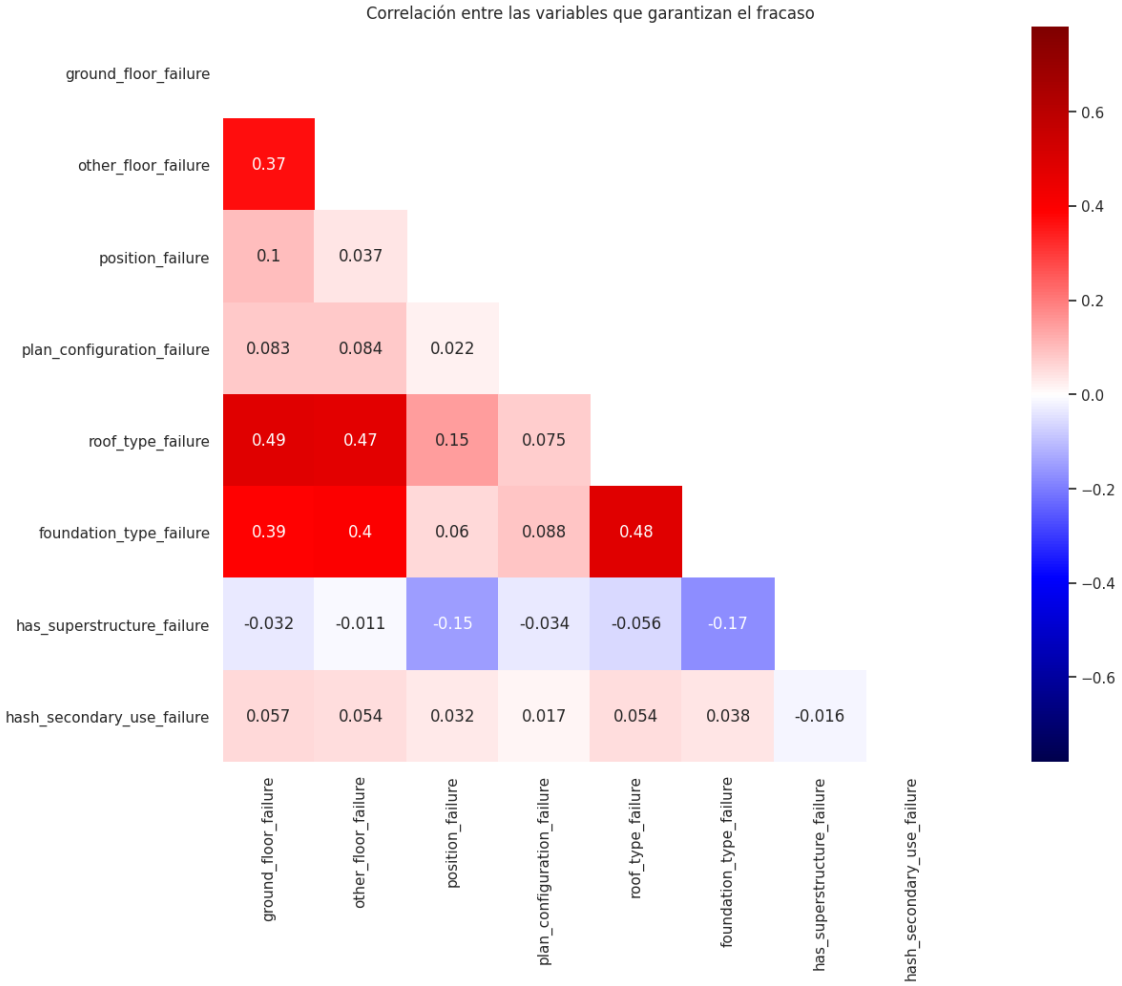

Figura 49: Correlación entre casos de fracaso

Podemos apreciar que los casos de fracaso no presentan una correlación general, pero si hay una correlación muy fuerte, particularmente sacando el apartado de *has_superstructure_failure*.

Para el caso de éxito, habíamos concluido que *has_superstructure* no tenía variables representativas, por lo tanto no aparece en el análisis de los casos exitosos. Esto podría o no indicar que, quitando *has_superstructure*, el tener condiciones favorables o desfavorables para un daño menor viene dado por tener una serie de variables.

Conclusiones:

- Tener solo una variable de éxito en un apartado (por ejemplo, tener x en *ground_floor_type*) no garantiza el éxito, sino que viene garantizado por enlazar más de una de estas condiciones.
- Se aprecia lo mismo para el caso fallido, a excepción de *has_superstructure*. Es decir, los materiales de construcción son muy importantes, ya que estos podrían por sí solos garantizar el fracaso ante un terremoto.
- A partir de esto, nos podemos independizar de las variables que parecen relativas pero que no hacen a la composición de los edificios (años, altura, área, estado legal de la tierra, etc).

6.3. Independizándonos de los aspectos sociales

Una de las hipótesis fuertes que nos planteamos al inicio del trabajo es si el aspecto social tiene influencia sobre el daño que sufren las edificaciones.

Nos pareció lógico que lo social no tenga un papel preponderante, sino que sean las condiciones en que están las edificaciones (que pueden originarse por aspectos sociales).

Al responder las preguntas de las variables más importantes para el grado de daño, pudimos independizarnos de todos los aspectos sociales, exceptuando la cantidad de familias que tiene una edificación.

Por lo tanto, estudiaremos si existe alguna relación, para poder realizar conclusiones más profundas sobre los aspectos sociales.

Figura 50: Análisis del efecto de la variable cantidad de familia sobre el daño en base a si contiene una variable de fracaso

El gráfico superior nos dice que para cada grupo de cantidades de familias (1, 2, 3, etc), hay familias que viven en una edificación con una condición propensa al fracaso ante un sismo. En los gráficos inferiores apreciamos que cada vez que solamente las se llega a un daño alto cuando las

condiciones de fracaso son positivas.

Estas conclusiones tienen coherencia, ya que una familia vive en una edificación y este puede o no caerse independientemente de la familia.

Por lo tanto, podemos independizarnos de los aspectos sociales.

Conclusión:

- Los aspectos sociales, tales como la cantidad de familias en las edificaciones, los usos secundarios (que figuran como importantes, pero muestran correlación fuerte con todas las variables que garantizan fracaso), o la condición legal en la tierra no son importantes por sí solas, sino que son importantes en un contexto. Por ejemplo, una edificación que presente una elevada cantidad de familias para un área pequeña y cantidad de pisos/altura pequeña puede apreciarse como una edificación carente, donde se puede construir sin planificación o uso de materiales especiales, por lo tanto pueden ser más propensas a sufrir mayores daños.

6.4. Entendiendo la variable *has_ superstructure*

Nos queda entender la única variable que no se correlaciona de forma fuerte con los demás casos de fracaso.

Lo que haremos es examinar para qué casos las edificaciones dan como afirmativo los casos fallidos de *has_ superstructure*, pero dan negativos en el resto de los casos de fracaso. Esto lo realizaremos para entender qué tiene que tener una edificación como casos de éxito para garantizar que el daño sea 1.

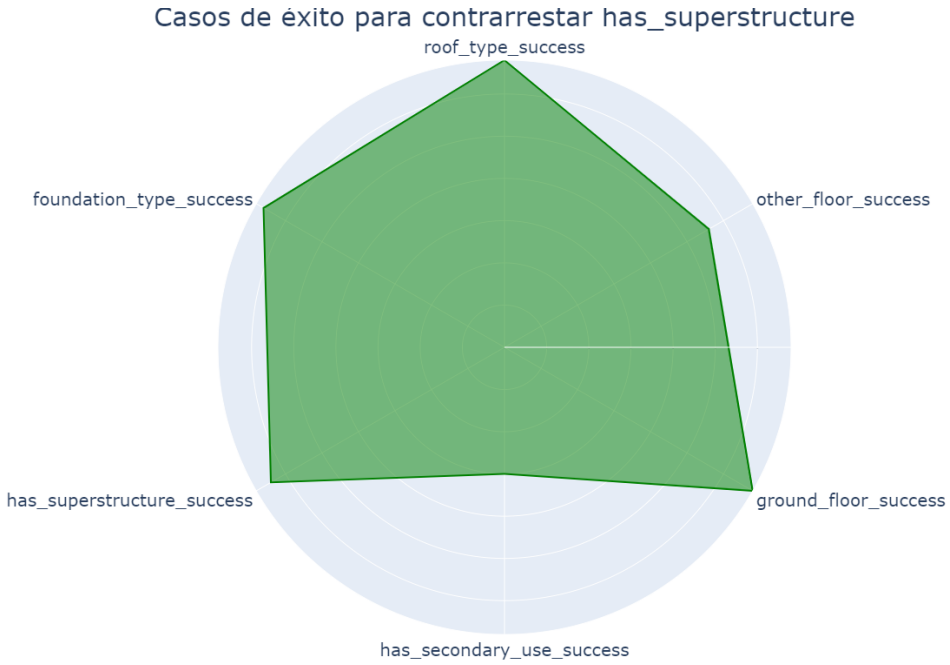

Figura 51: Radar de las variables con éxito para has_superstructure_failure

Podemos concluir que todas las componentes de éxito deben estar en un estado alto en promedio, para contrarrestar el efecto negativo de *has_ superstructure*.

Conclusión:

- Para contrarrestar un material de edificación mal elegido ante un terremoto, una edificación debería presentar prácticamente todo el resto de las variables significativas como un caso de éxito, sino el grado del daño será mayor que 1.

7. Conclusiones

Aquí realizaremos un resumen de todo lo entendido a lo largo del trabajo, ya que este es el objetivo central del mismo.

El DataFrame consistía en datos obtenidos a partir de un terremoto. Por lo tanto, todo nuestro análisis se centró en evaluar qué variables nos proporcionan un menor daño, así poder, en un trabajo posterior, predecir y armar modelos basados en estas conclusiones.

Nuestras inquietudes iniciales vinieron por parte de variables que pensamos como más importantes, como la edad, altura y área.

A medida que avanzamos con los análisis individuales pudimos notar que hay ciertas variables que tienen incidencia en el factor de éxito y/o fracaso que puede tener una edificación. Luego, a medida que nos surgieron preguntas e inquietudes, pudimos independizarnos de algunas de estas variables, unificar ciertos casos, relacionarlas, etc.

El resumen de toda esta lógica se plasmó en las respuestas a las preguntas e hipótesis que nos formulamos antes y durante el trabajo, estas son:

7.1. Respuestas a preguntas:

- ¿Qué casos de éxito y fracaso se pueden extrapolar de cada variable?

Pudimos responder todas estas preguntas analizando individualmente estas variables, el análisis se encuentra en esta sección (click para ir).
Concluimos casos de éxito y fracaso para **cada variable** (importante esto para la siguiente pregunta).

- Esos casos de éxito y fracaso, ¿Se deben a la presencia de otras variables o se puede afirmar que la variable analizada es de relevancia?
  Esta pregunta la pudimos empezar a responder en esta sección (click para ir).
  Concluimos que hay solo ciertas variables que pesan a la hora de concluir éxito y fracaso para una edificación, y que el resto es dependiente de ellas. Entre estas variables se encuentran el material con el que fue construida, los cimientos utilizados al momento de la construcción, los pisos utilizados (tanto en planta baja como en el resto) y el tipo de techo.

- ¿Qué relaciones entre las variables nos dan las relaciones más provechosas para analizar para los casos extremos? y, para los casos de fracaso hallados como más importantes, ¿Qué casos de éxito se necesitan para contrarrestarlos y lograr daño 1?
  Estas también fueron resueltas en esta sección (click para ir).
  El análisis para estos casos es muy parecido al anterior, ya que pudimos explorar, por ejemplo cuando observamos cómo contrarrestar *has_superstructure* con estado de fracaso (click para ir).
  La importancia radicó siempre en entender que para garantizar realmente el éxito se tienen que tener varias componentes, y en cambio para el fracaso, solo con tener ciertas de ellas (como es el caso del material de la edificación) ya basta para que, si no se contrarresta correctamente, la edificación sufra daños graves.

7.2. Respuestas a hipótesis:

- La edad, altura y área son determinantes. A mayor área y menor altura, menor es el daño, y a mayor edad mayor es el daño.
  Pudimos apreciar que una idea inicial coherente (es lógico pensar que una edificación con mayor edad puede recibir mayor daño ante un terremoto) quedaron desmentidas al analizar que si estas presentan las componentes que llevan a éxito, entonces tienen muchas posibilidades de enfrentar un terremoto y recibir daños menores.
  Este fue el punto de inflexión en el trabajo, ya que pudimos independizarnos de todas las variables que no afectan realmente a la edificación a la hora de afrontar un terremoto. Por lo

tanto, fue una hipótesis que, para responderla, el análisis derivó en un mejor entendimiento de los datos.

- Los aspectos sociales (cantidad de familias, los usos secundarios y el estado legal de la tierra donde fue construida) no deberían ser significativas, ya que lo significativo son las condiciones en que fueron construidas. En lugar de relacionar, por ejemplo, lugares más carenciados, lo significativo para el análisis es la estructura de la edificación.
  Esto lo resolvimos esta sección (click para ir).
  Pudimos obtener como conclusión que los factores sociales pueden afectar la planificación de la edificación, pero que lógicamente por sí solos no pueden afectar a cómo va a ser el daño que reciba la construcción.
  Es decir, que haya un colegio en una zona carenciada no determina por sí solo que la edificación reciba daños graves. Pero que, al haber quizás menos planificación, la edificación usada para tales fines puede tener materiales, tipos de cimientos, pisos y techos que pueden no ser idóneos, por lo tanto la edificación tendrá mayores posibilidades de recibir daños.

8. Repositorio de Github

Link al repositorio: https://github.com/gsabatino9/OrganizacionDeDatos