Facultad de ingeniería de la UBA

75.06/95.58 Organización de Datos

Trabajo práctico N°1
Análisis de datos

Primer cuatrimestre de 2021

Análisis exploratorio sobre los datos de Richter's Predictor

Integrante	Padrón	Correo electrónico
Gutiérrez, Matías	92172	magutierrez@fi.uba.ar
Julián Rodrigo Cropano Miranda	103960	jcropano@fi.uba.ar
Alexis Brian, Herrera Aguilar	104639	abherrera@fi.uba.ar
Tomás Yavicoli	97617	tyavicoli@fi.uba.ar

Nombre del grupo: *DataStalkers*
Link al repositorio:
`git@github.com:gutierrezmatias/Organizacion-de-Datos.git`

Índice

Parte I
Introducción

En el presente informe se presentan los resultados obtenidos al realizar un análisis exploratorio sobre los datos de encuestas hechas por Kathmandu Living Labs y el Central Bureau of Statistics para analizar el impacto del terremoto *Gorkha* ocurrido en el año 2015 en Nepal.

Los datos brindados consistieron en dos datasets:

- train values: Con datos que describen a las edificaciónes identificadas por un id.
- train labels: Con datos relacionados al daño recibido por las edificaciones identificadas por su id.

Las siguientes secciones resumen los resultados obtenidos luego de realizar el análisis, junto a las conclusiones obtenidas.

Parte II
Análisis previo y limpieza de datos

1. Introducción

Inicialmente se debe analizar los datos de manera amplia para entender que atributos podemos relacionar entre sí resultando en información relevante. En terminos generales:

- Hay 260601 registros con 38 atributos para cada id del edificio
- Hay columnas relacionadas entre sí como las que indican la superestructura o el uso secundario.

2. Tipo de datos

Analizamos los valores que pueden tomar cada columna y le asignamos un tipo de datos apropiado para que ocupe menos espacio en memoria

Int16: Estas columnas tienen un rango de 0 a 12.567[1]
geo_level_1_id || geo_level_2_id || geo_level_3_id || count_floors_pre_eq || age || area_percentage || height_percentage || count_families ||

Int32: El rango de la columna va de 4 a 1.052.934[2]
Building_id

Category: Dados por la consigna
land_surface_condition || foundation_type || roof_type || ground_floor_type || other_floor_type || position || plan_configuration || legal_ownership_status ||

Bool: Dados por la consigna
Los que empiezan con has_superstructure: adobe_mud || mud_mortar_stone || stone_flag || cement_mortar_stone || mud_mortar_brick || cement_mortar_brick || timber || bamboo || rc_non_engineered || rc_engineered || other ||

[1]Int16 puede almacenar valores de -32,768 a 32,767
[2]Int32 puede almacenar valores de -2,147,483,648 a 2,147,483,647

Los que empiezan con has_secondary_use: has_secondary_use || agriculture || hotel || rental || institution || school || industry || health_post || gov_office || use_police || other ||

3. Reporte detallado de la primera inspección de los datos

Aqui se presentan algunos hallazgos realizados con las distintas columnas del set de datos.

3.1. Respecto a *Materiales de construccion*

- Se observa una preponderancia del mud mortar stone

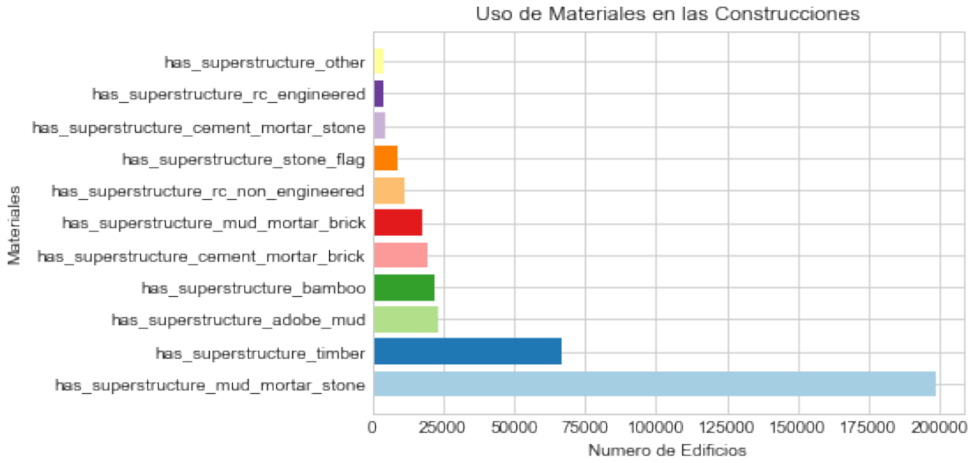


Figura 1: Utilización de Materiales en Edificios

- El mud mortar stone es el mas utilizado entre las edificaciones nuevas con menos de 30 años.

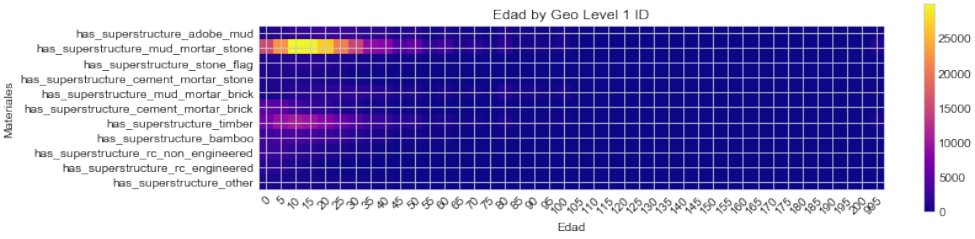


Figura 2: Relación entre materiales y edad de Edificios

- Se puede ver que en ciertas regiones hay mayor concentracion de edificios de mud mortar stone

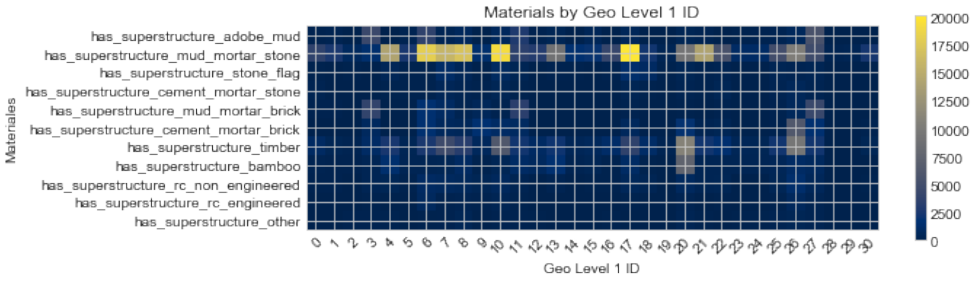


Figura 3: Relacion entre Materiales y Geo level 1 ID

3.2. Respecto a las *Edificaciones*

- La mayoria de los edificios cuentan con dos pisos.

Figura 4: Cantidad de edificios por piso

- La mayoria de los edificios tienen una edad menor a 200 años. Y se puede ver en casi todas las regiones edificios de mas de 900 años. En estos edificios antiguos hay tanto uso secundario como residencial(su mayoria).

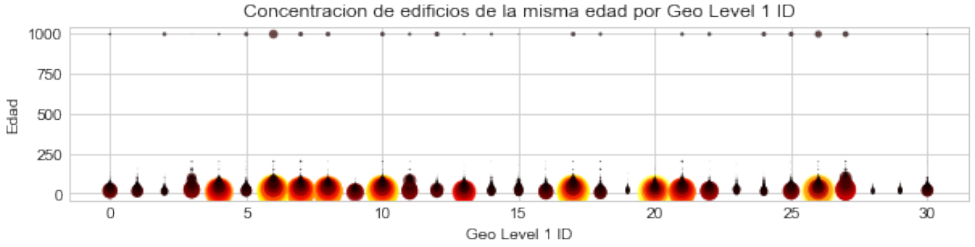


Figura 5: Distribucion de edades por Geo Level 1 ID

- Hay regiones con mayor cantidad de edificaciones nuevas y coinciden con las regiones con mayor concentración de mud mortar stone

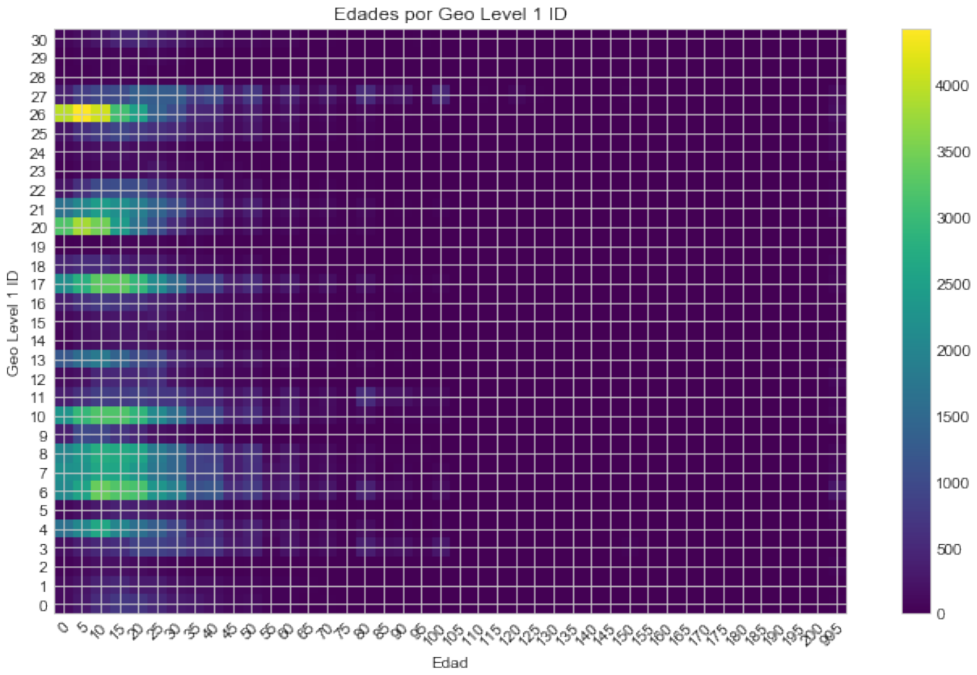


Figura 6: Edades por Geo Level 1 ID

- La proporcion de la superficie edificada es realativamente baja (menor al 30 %)
- La altura de los edificios rondan el 5 % (lo que se condice con la cantidad de pisos mas frecuente)

3.3. Respecto a *La utilizacion de los edificios*

- Se observa que aproximadamente el 11 % de los edificios se usan secundariamente.
- Se utilizan mayormente en agricultura , hoteleria y renta.

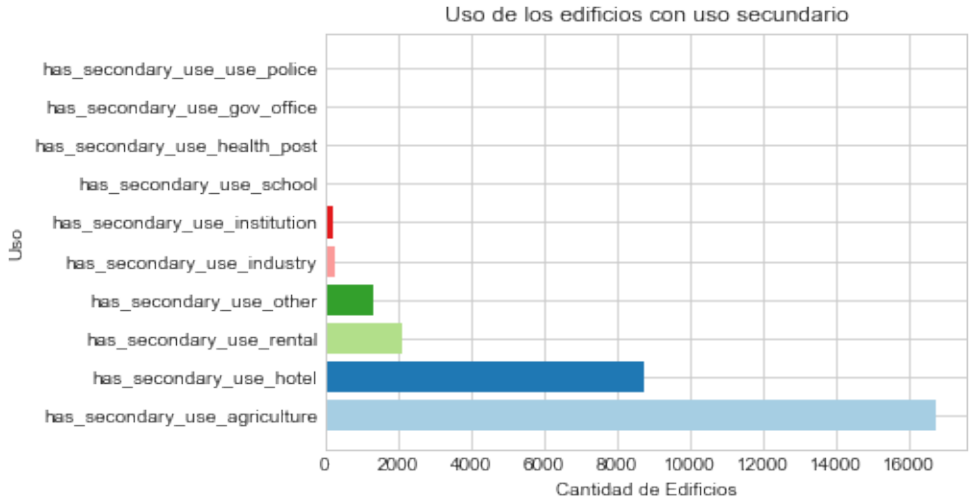


Figura 7: Utilización de edificios con uso secundario

3.4. Respecto a *Los habitantes*

- Se puede ver que la poblacion no esta distribuida de la misma forma en todas las regiones. Hay regiones que poseen muchos habitantes y regiones con muy pocos.Y las regiones mas pobladas coinciden con las regiones con mayor concentracion de mud mortar stone en las edificaciones, que son a su vez tambien son las que tienen mayor concentracion de edificaciones nuevas.

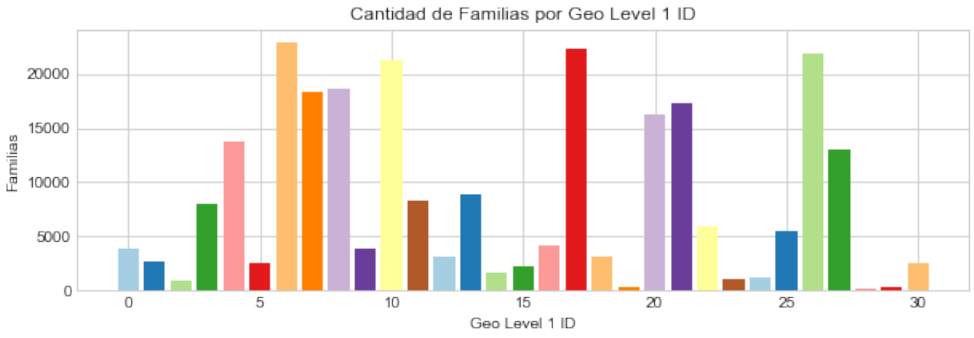


Figura 8: Distribución de la población por Geo Level 1 ID

Parte III
Análisis de datos

1. Relación entre Grado de Daño y Antigüedad

1.1. Antigüedad respecto al Grado de Daño

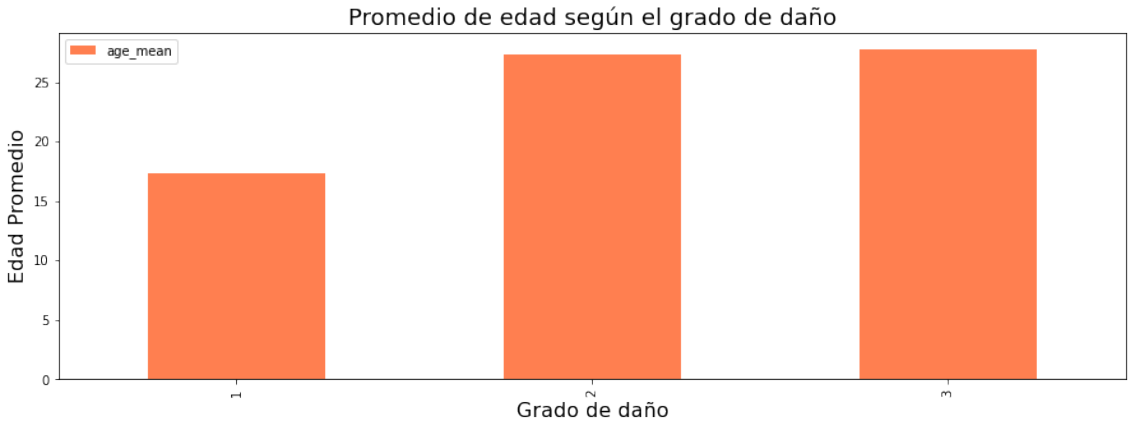


Figura 9: Antigüedad Promedio respecto al Grado de daño.

Tomando como referencia los tres valores que puede tomar el grado de daño, se calculó un promedio aproximado de la edad que tienen los edificios que recibieron tal daño. Como se muestra en el gráfico superior, construcciones de entre 15 y 20 años, fueron más propensas a tener grado de año 1. Una antigüedad de 25 años para arriba, es más propensa a tener un grado de daño 2 o mayor.

1.2. Grado de daño promedio respecto a la Antigüedad

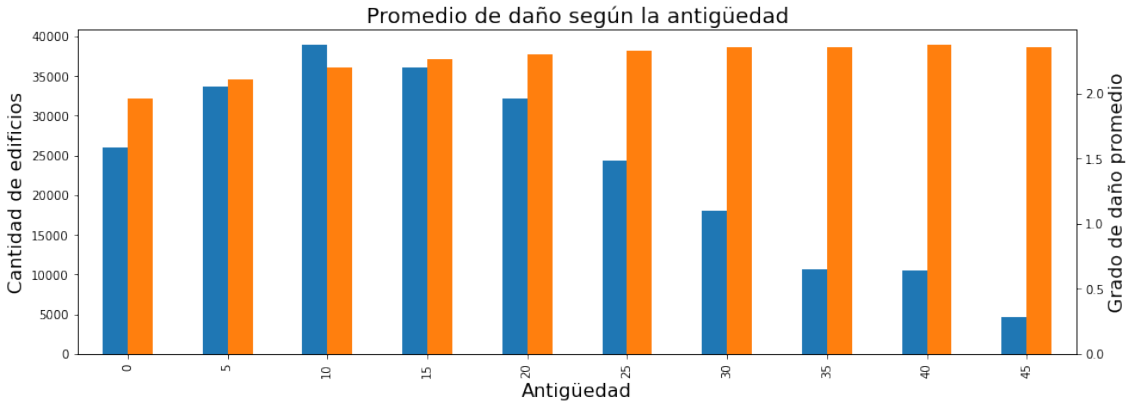


Figura 10: Grado de daño promedio respecto a la Antigüedad.

Verificamos la inversa del análisis realizado anteriormente, tomando esta vez como referencia las edades de forma creciente, queremos ver como como evoluciona el grado de daño promedio entre todos los edificios. Se puede observar que a medida que aumenta la antigüedad, el grado de daño también aumenta.

2. Relación entre Grado de Daño, Antigüedad, Área y Altura del edificio

2.1. Área y Daño promedio según la antigüedad

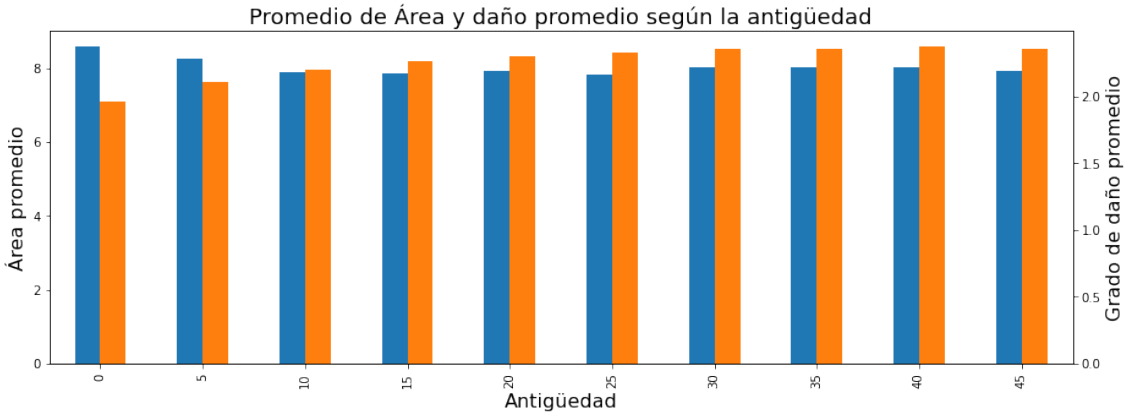


Figura 11: Área y daño promedio según la antigüedad

Para comenzar el análisis nos fijamos la relación entre el área, el grado de daño y la antigüedad de un edificio. Podemos notar que conforme avanzan los años, el área disminuye, y el grado de daño aumenta. En el próximo gráfico podemos observar el por qué.

2.2. Altura y Daño promedio según la antigüedad

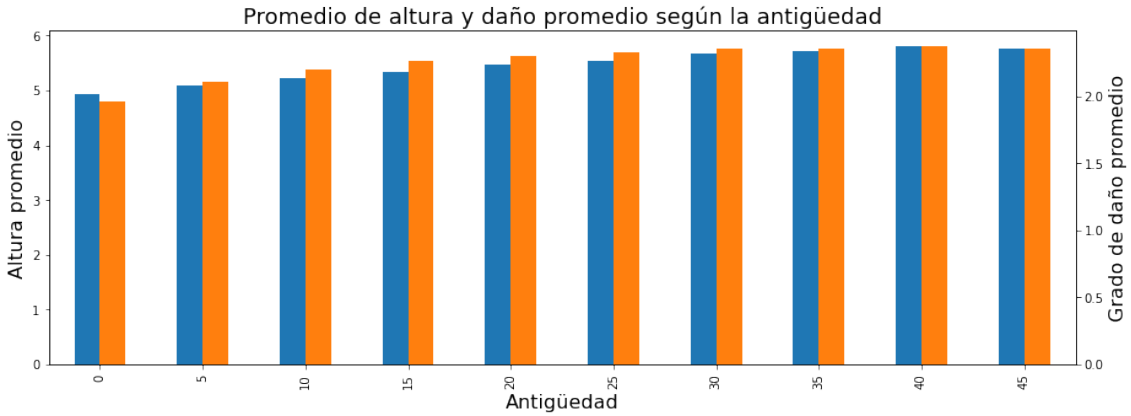


Figura 12: Altura y daño promedio según la antigüedad

Al analizar lo mismo, pero ahora por la altura, vemos también que la altura va aumentando conforme pasan los años, como también el grado de daño. Acompañado del gráfico anterior cuando se analizó el área, los edificios más antiguos tenían áreas menores y alturas más altas, lo que hacía que fueran propensas a grados de daño mayores.

3. Relación entre Grado de Daño, Antigüedad y material de la Súperestructura

Para esta sección se hizo algo similar que para nuestro de antigüedad contra el grado de daño. Esta vez decidimos filtrar aquellos edificios que tenían la súperestructura hecha de los materiales que se mostrarán a continuación, queriendo saber qué clase de influencia tienen sobre la resistencia del edificio, si la hay.

3.1. Súperestructura de Adobe/Barro

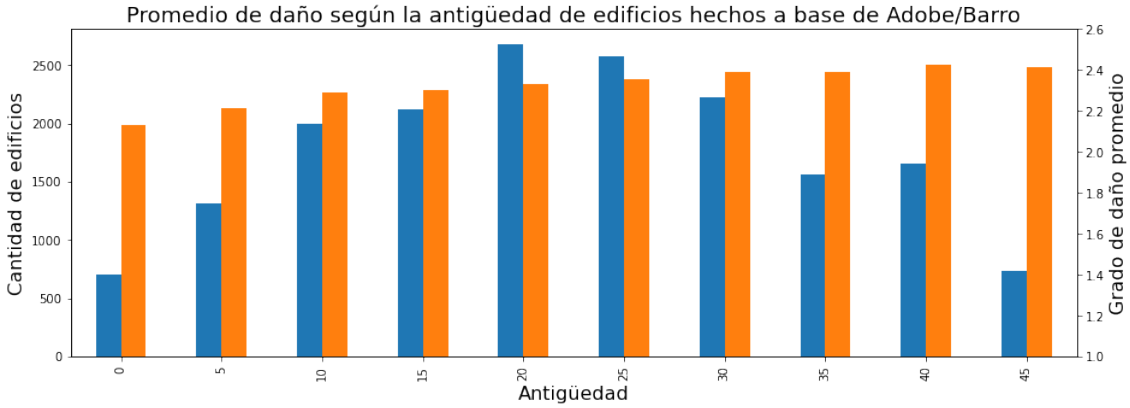


Figura 13: Cantidad de edificios hechos de Adobe/Barro y su grado de daño promedio

3.2. Súperestructura de Barro/Piedra

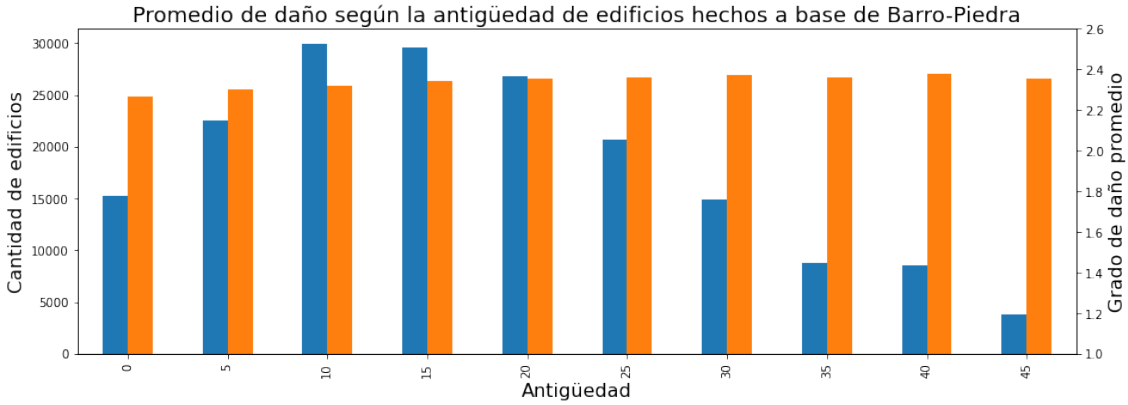


Figura 14: Cantidad de edificios hechos de Barro/Piedra y su grado de daño promedio

3.3. Súperestructura de Piedra

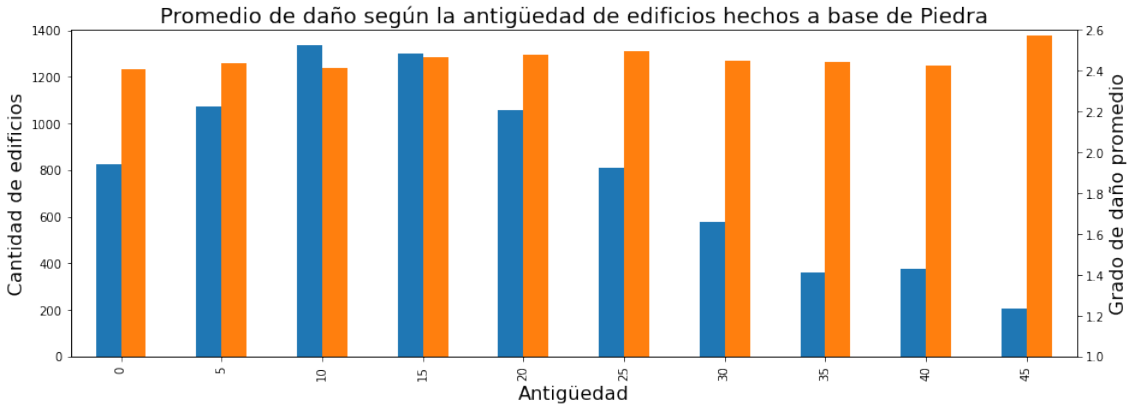


Figura 15: Cantidad de edificios hechos de Piedra y su grado de daño promedio

3.4. Súperestructura de Cemento/Piedra

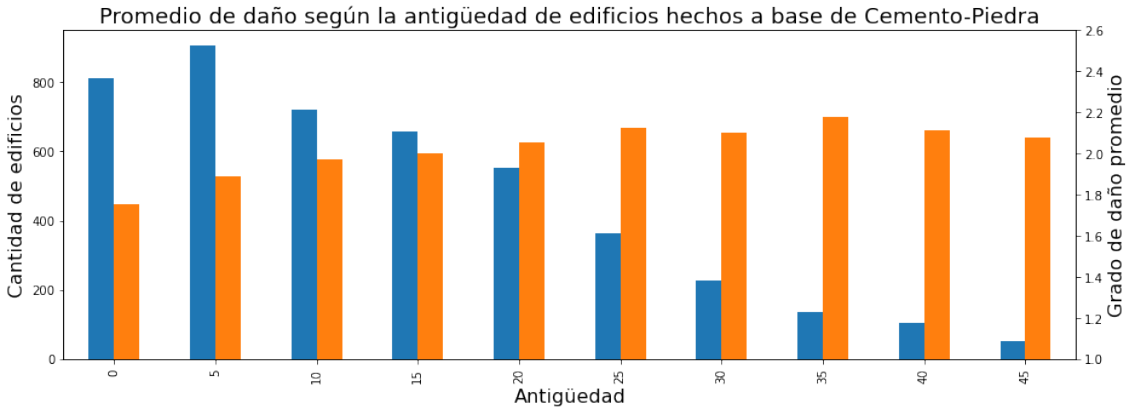


Figura 16: Cantidad de edificios hechos de Cemento/Piedra y su grado de daño promedio

3.5. Súperestructura de Barro/Ladrillo

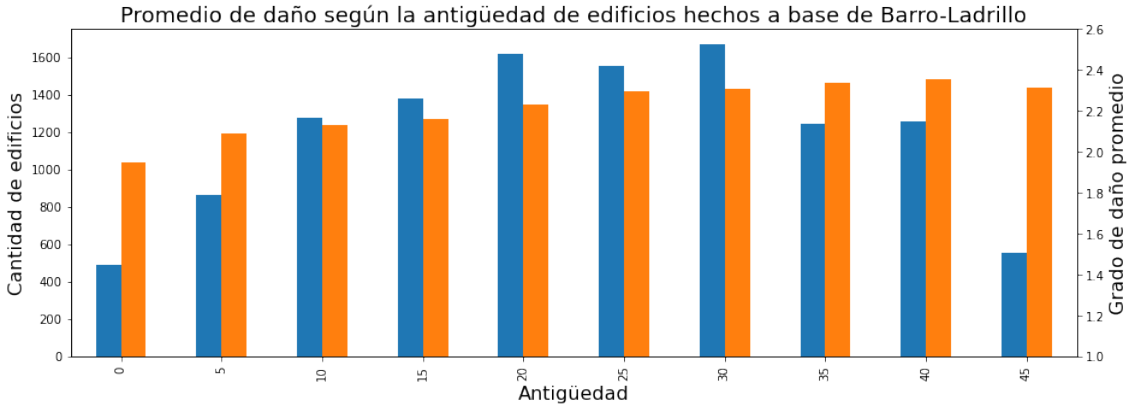


Figura 17: Cantidad de edificios hechos de Barro/Ladrillo y su grado de daño promedio

3.6. Súperestructura de Cemento/Ladrillo

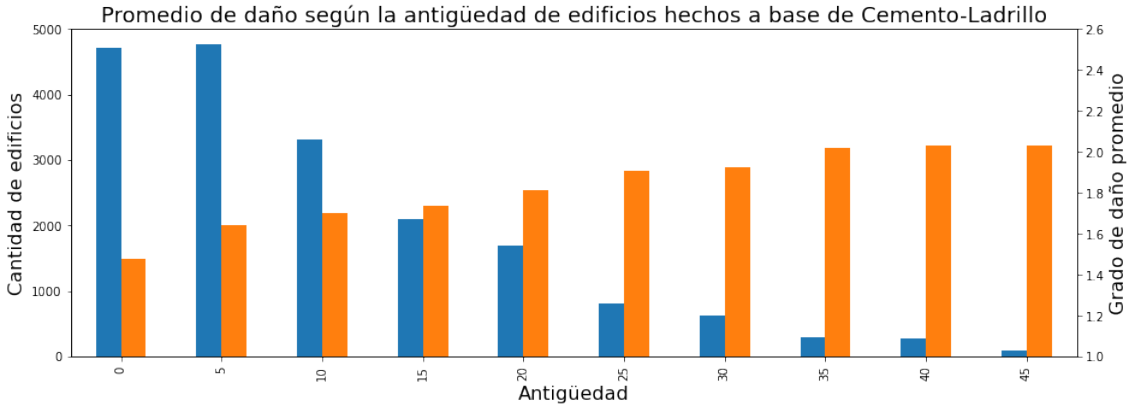


Figura 18: Cantidad de edificios hechos de Cemento/Ladrillo y su grado de daño promedio

3.7. Súperestructura de Madera

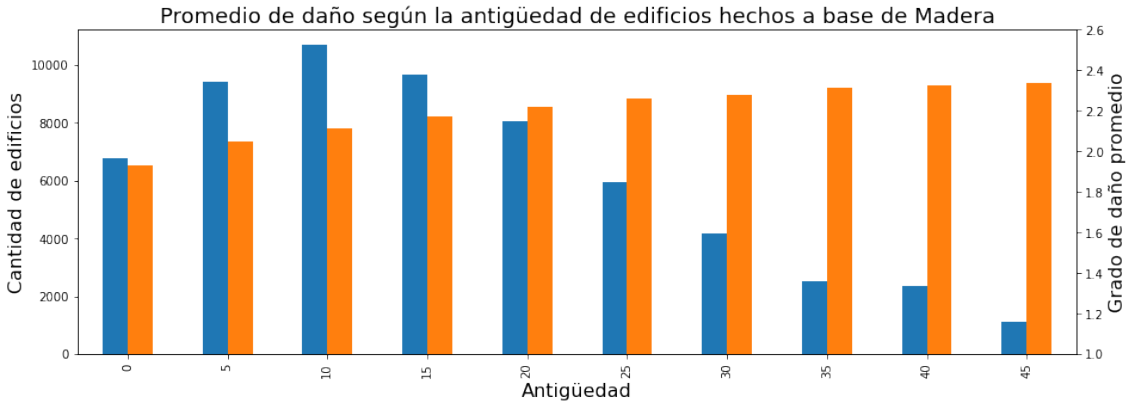


Figura 19: Cantidad de edificios hechos de Madera y su grado de daño promedio

3.8. Súperestructura de Bamboo

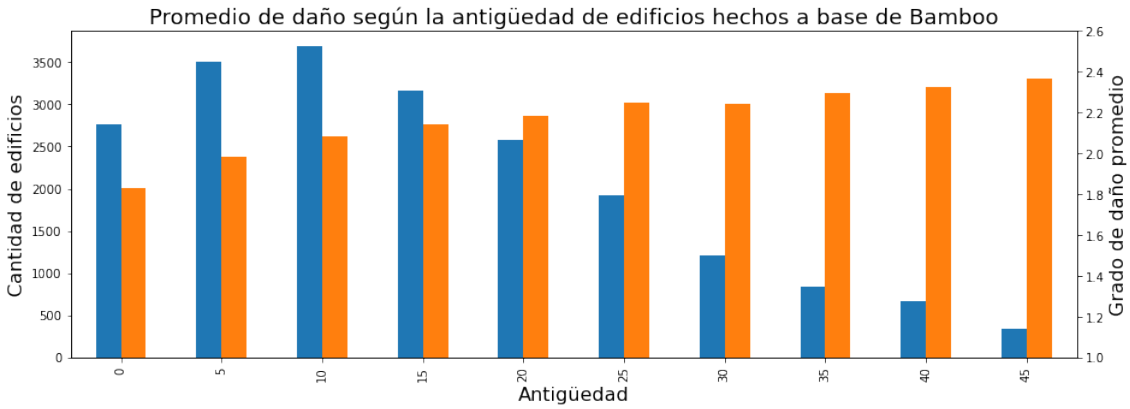


Figura 20: Cantidad de edificios hechos de Bamboo y su grado de daño promedio

3.9. Súperestructura de Concreto Reforzado No Diseñado

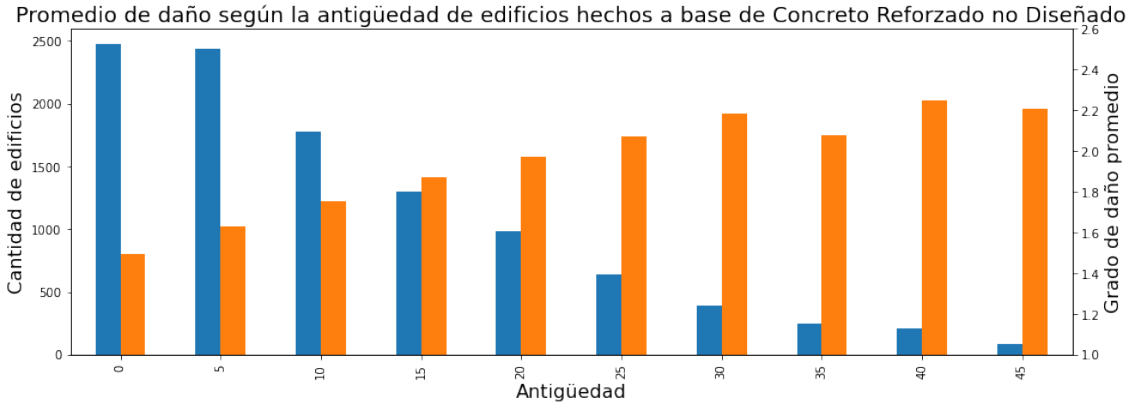


Figura 21: Cantidad de edificios hechos de Concreto Reforzado no Diseñado y su grado de daño promedio

3.10. Súperestructura de Concreto Reforzado Diseñado

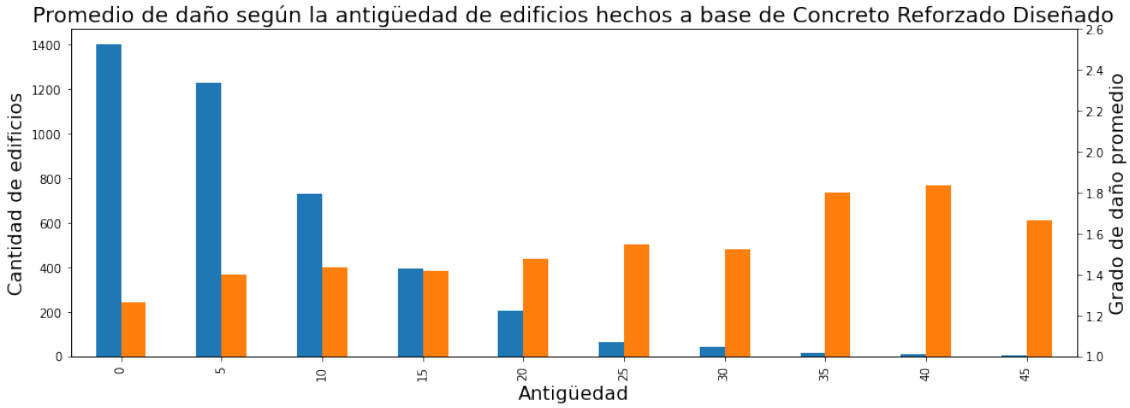


Figura 22: Cantidad de edificios hechos de Concreto Reforzado Diseñado y su grado de daño promedio

3.11. Súperestructura de Otros materiales

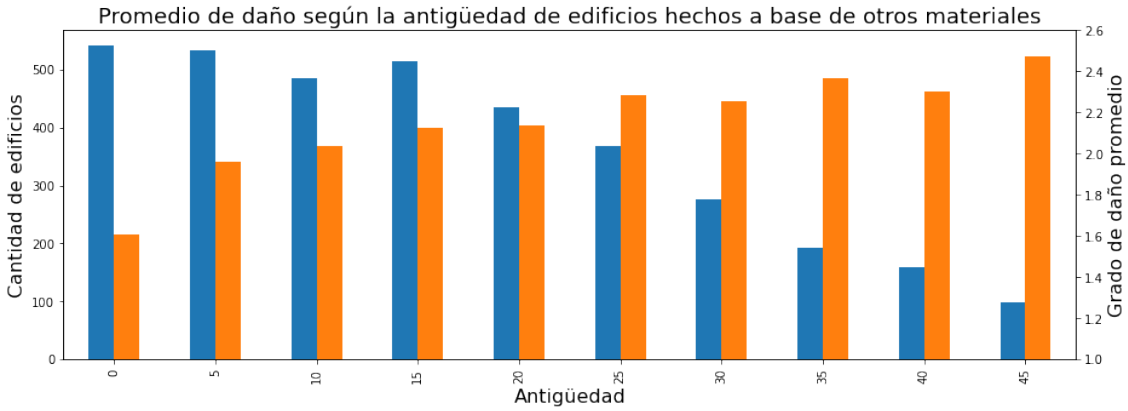


Figura 23: Cantidad de edificios hechos de otros materiales y su grado de daño promedio

Como podemos observar, la persistencia de un edificio a lo largo del tiempo varía dependiendo del material que se usa en su súperestructura. Por supuesto, como se notó anteriormente, a medida que la antigüedad aumenta, también la vuelve menos resistente. Sin embargo, se puede notar claramente que hay materiales que no solo son más resistentes que otros, sino que su deterioro al pasar los años es mucho menor. Por ejemplo, aquellos edificios que cuentan con súperestructuras de piedra, tienen el mayor grado de daño de todos, tanto nuevas como antiguas. Mientras que aquellas que son de Concreto Reforzado Diseñado, tienen el menor grado de daño.

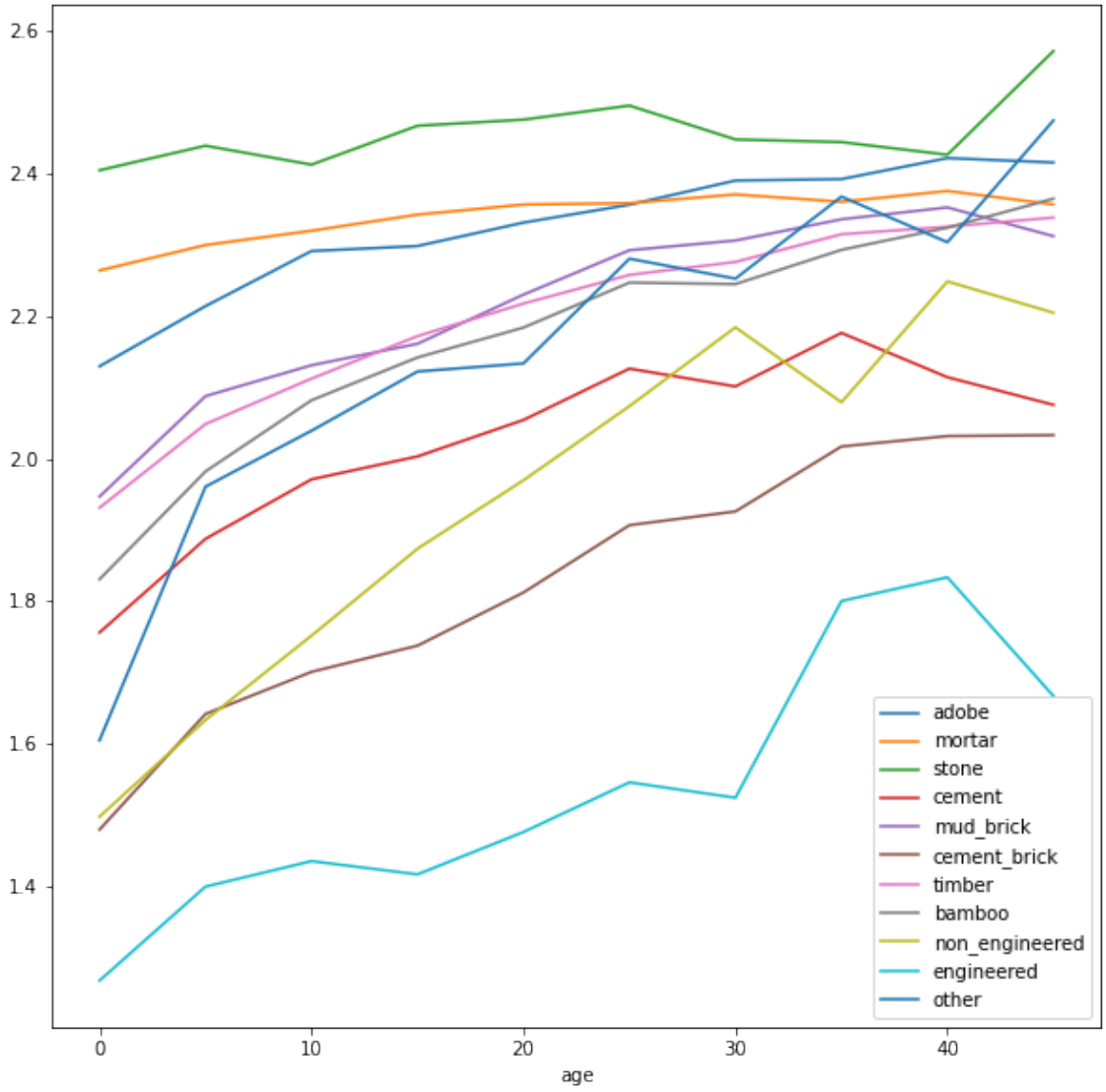
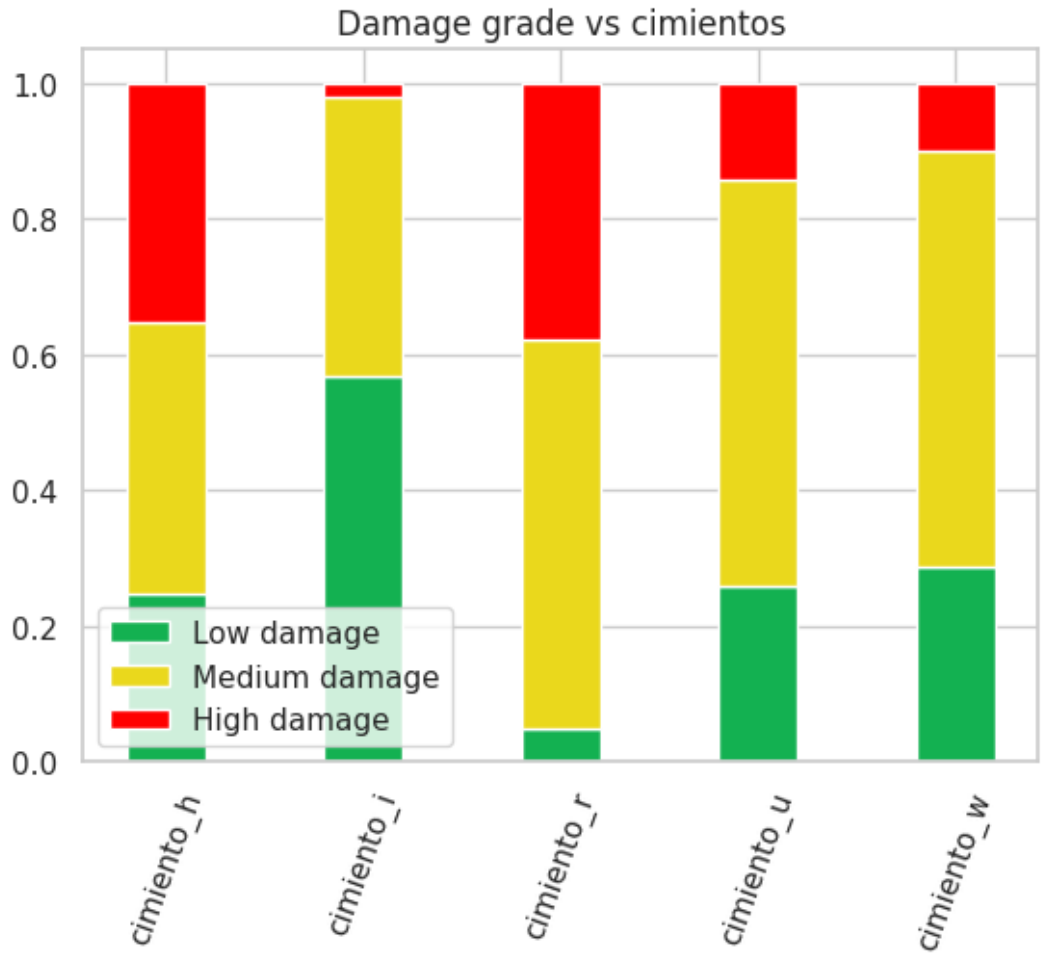


Figura 24: Promedio de daño de los materiales a lo largo del tiempo

4. Relacion entre diseño de construccion y Grado de Daño

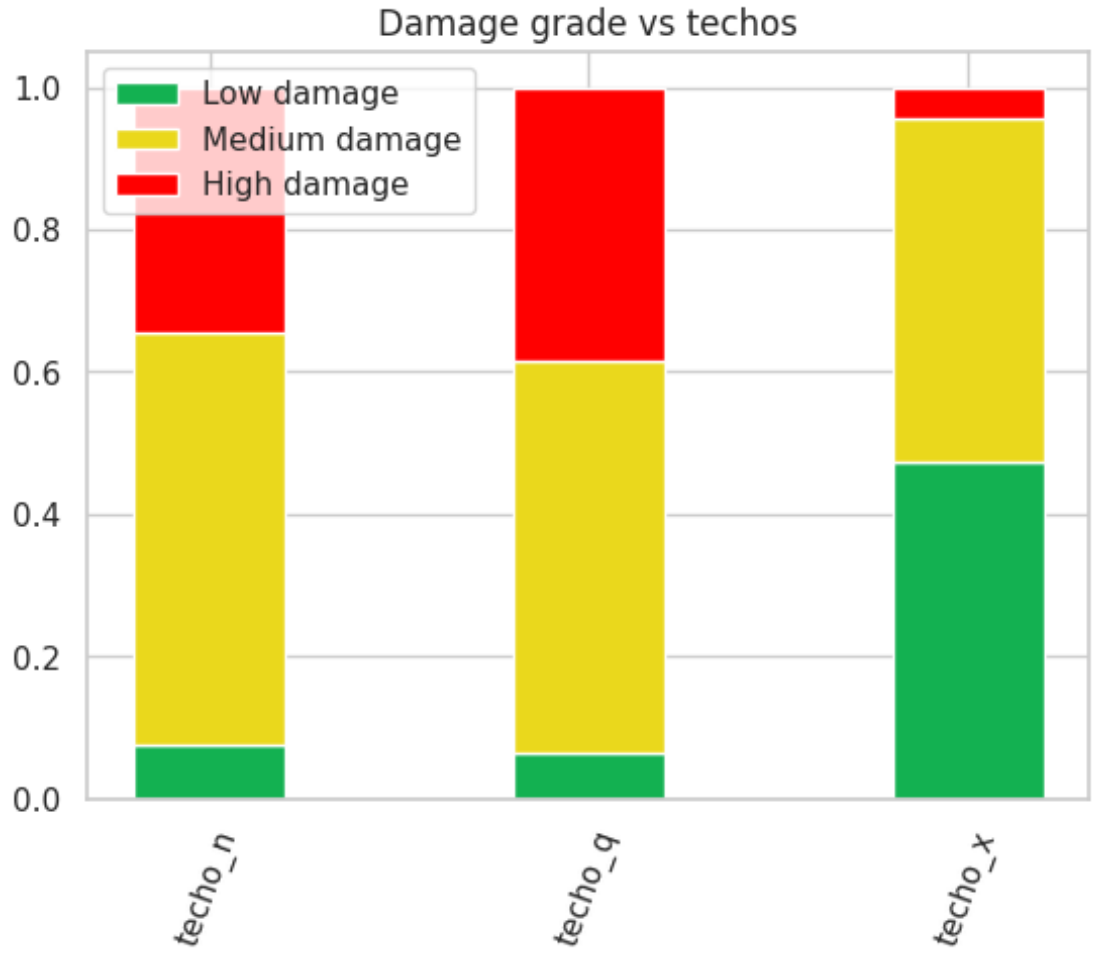
Aqui veremos como los distintos tipos de diseño de edificios se comportaron frente al terremoto, sin tener en consideracion cuales son los mas frecuentes ni la localizacion geografica.

4.1. Tipo de cimientos usados cuando se construyó vs Daño



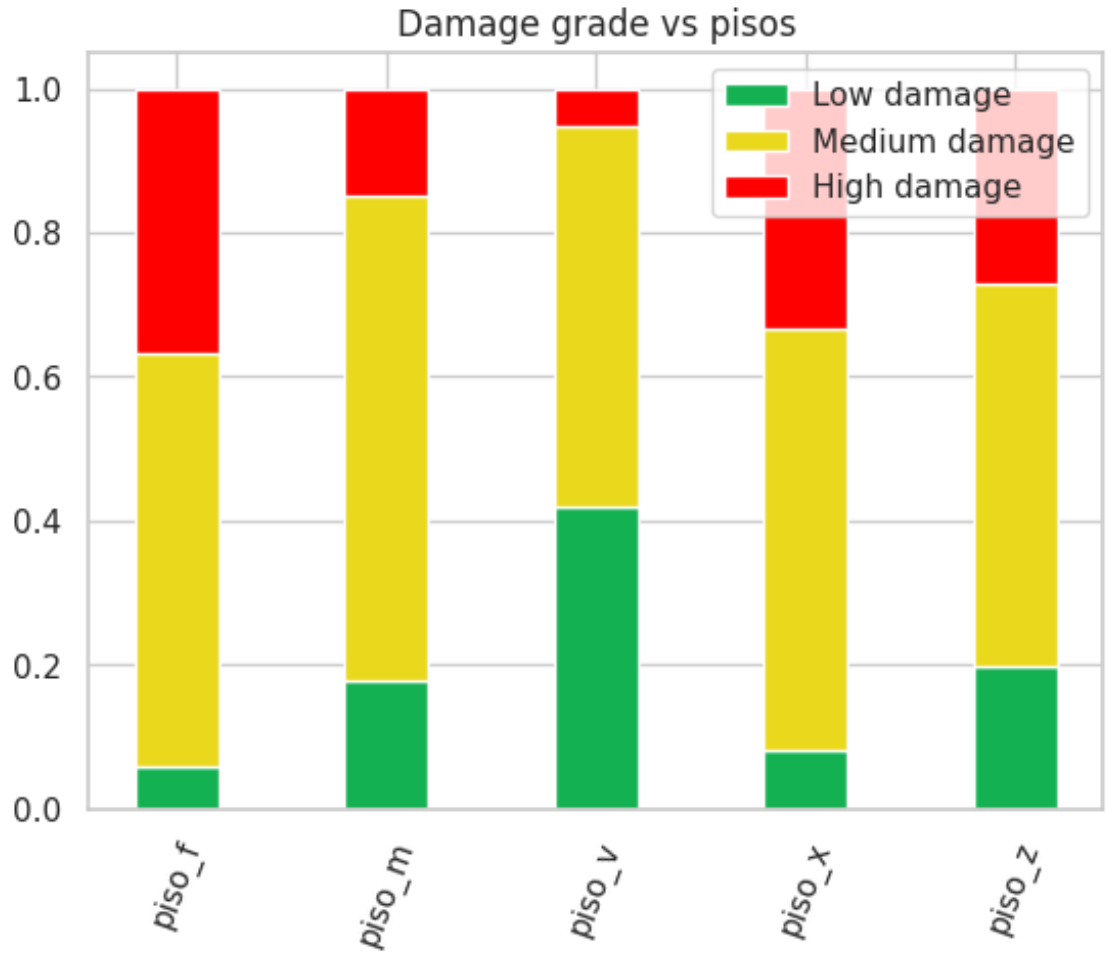
El cimiento mas resistente para la construccion pareceria ser el del tipo i, por ser el que menor daño recibio en total. Por otro lado el mas dañado es el tipo r.

4.2. Tipo de techos usados cuando se construyó vs Daño



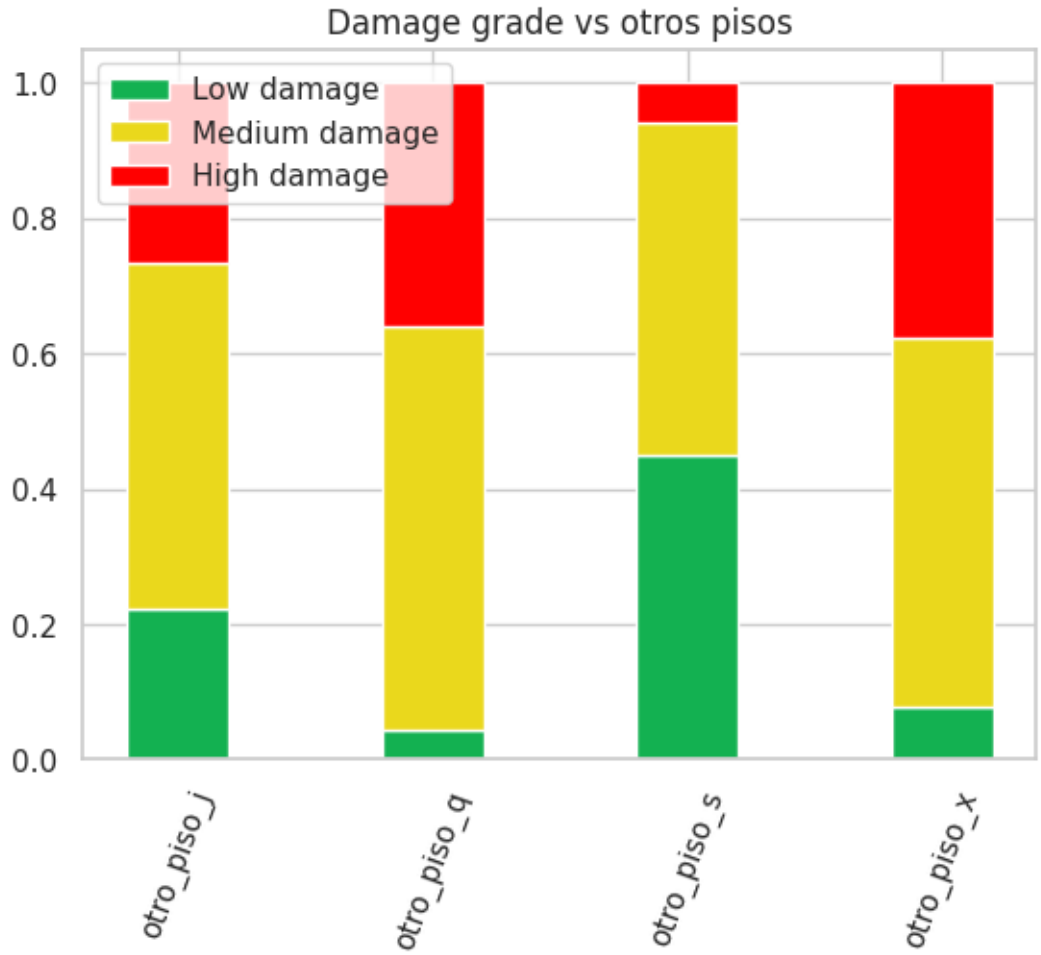
Los techos de tipo x son los que sufrieron menor daño en comparacion a los otros dos tipos de techos utilizados para la construccion.

4.3. Tipo de construcción usado en la planta baja vs Daño



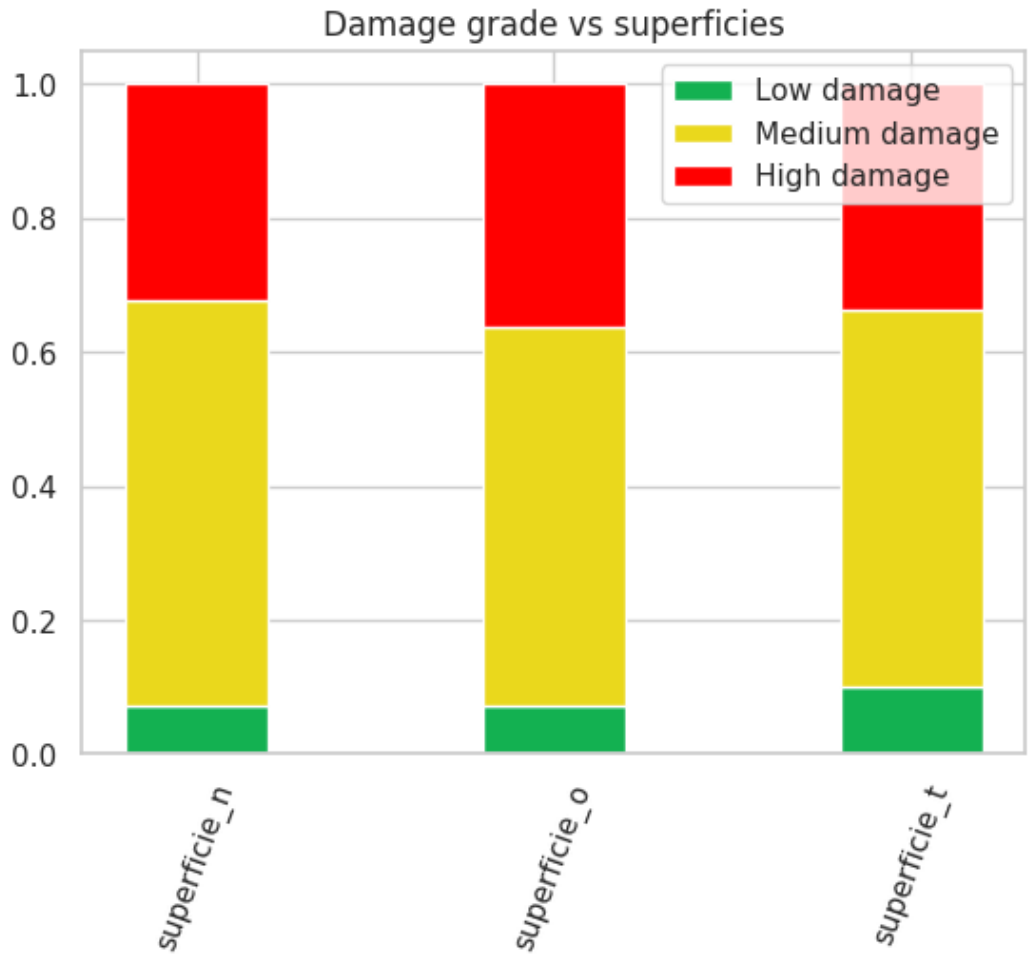
Se puede observar que el mejor tipo usado para la construccion de planta baja es v. En cuanto a el peor no queda del todo claro.

4.4. Tipo de construcción usado en otros pisos vs Daño



Los tipos de construccion s y j tuvieron un buen desempeño contra el terremoto, pero no se puede decir lo mismo del los q y x.

4.5. Condición de la superficie terrestre vs Daño



Claramente vemos que la superficie terreste donde fue construido el edificio no es un parametro importante frente al daño del terremoto. Los tres tipos de superficie tiene un comportamiento similar.

5. Relacion entre utilizacion de los edificios y zona geografica

5.1. Cantidad de edificios con otra utilidad

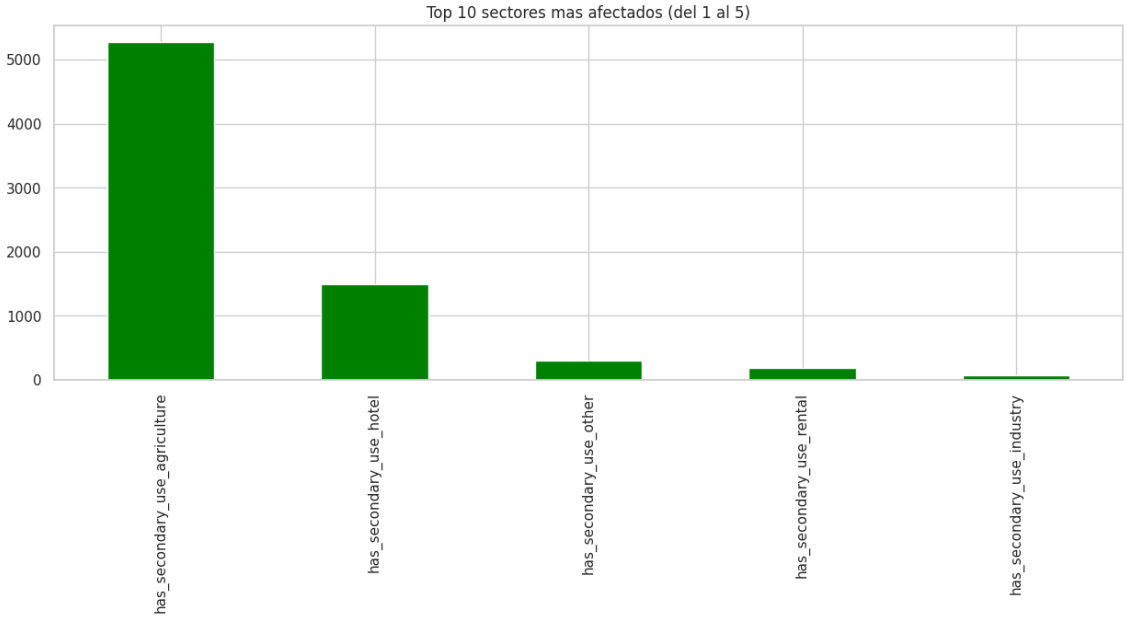


Figura 25: Top 5 utilizaciones secundarias

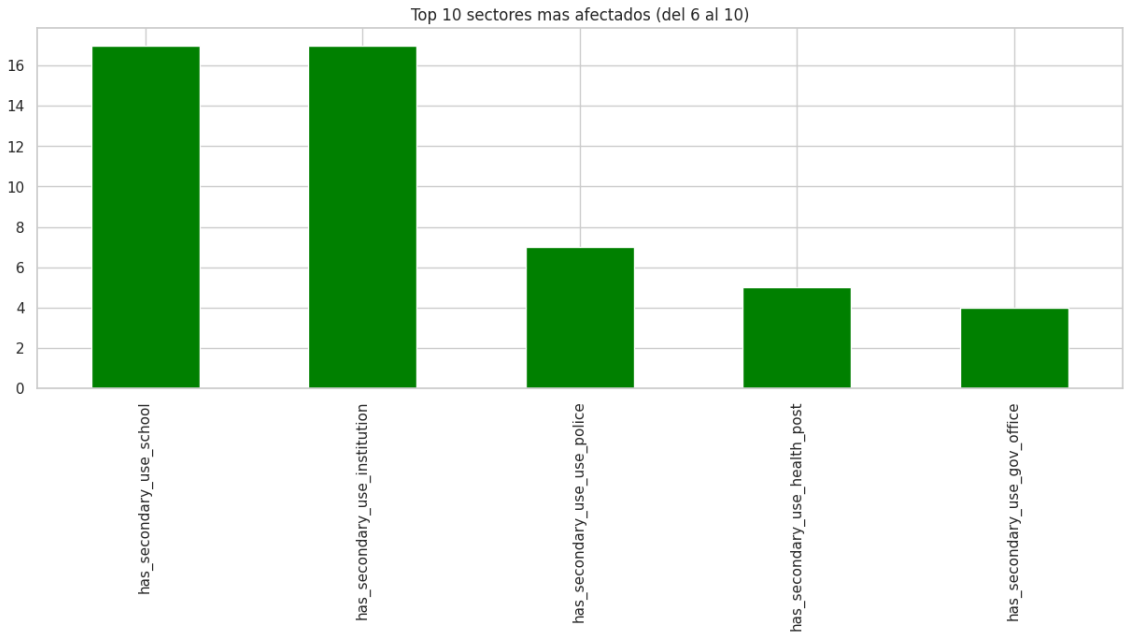
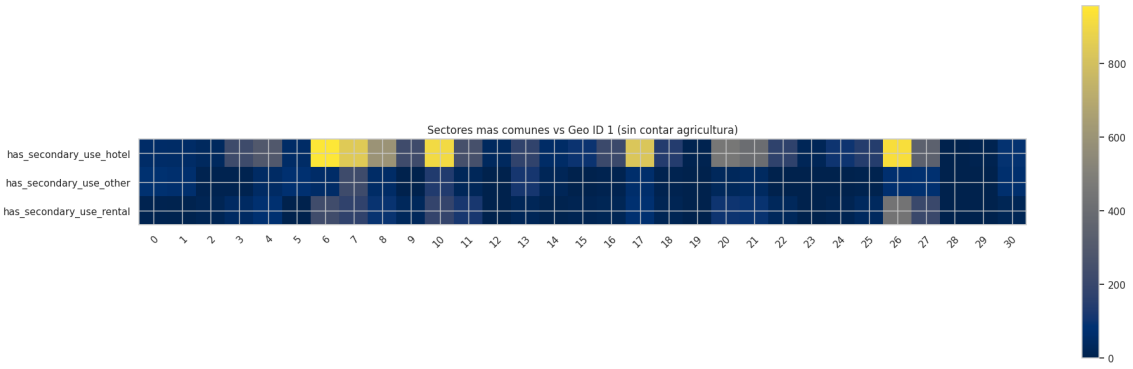
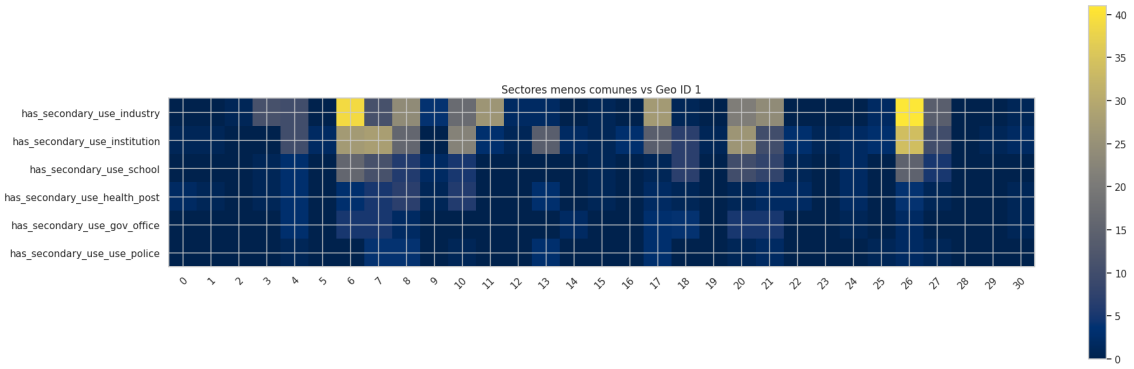


Figura 26: Top 5 utilizaciones secundarias menos frecuentes

Queremos enfocarnos en ese pequeño porcentaje de aquellos edificios que tienen otra utilidad. Podemos notar, como dijimos previamente, que la mayoria estan enfocados en la agricultura y por gran diferencia. Es por esto que para los proximos graficos no los tenemos en cuenta.





En estos dos heatmap podemos diferenciar los distintos sectores y su distribucion segun zona geografica. Notamos que en las zonas mas pobladas hay mayor cantidad de edificios con otra utilidad, como era de esperar. Tambien vemos que hay zonas donde no hay algunos tipos de servicios especificos, como por ejemplo el geo ID 2 no cuenta con edificios policiales, ni de salud, ni escuelas.

6. Daños del terremoto según gravedad

Queremos saber como ha afectado el terremoto a las familias, de manera que podamos entender que impacto ha tenido en la sociedad. Sin embargo, tenemos datos de edificios y no de personas. Veamos si existe una relación directa

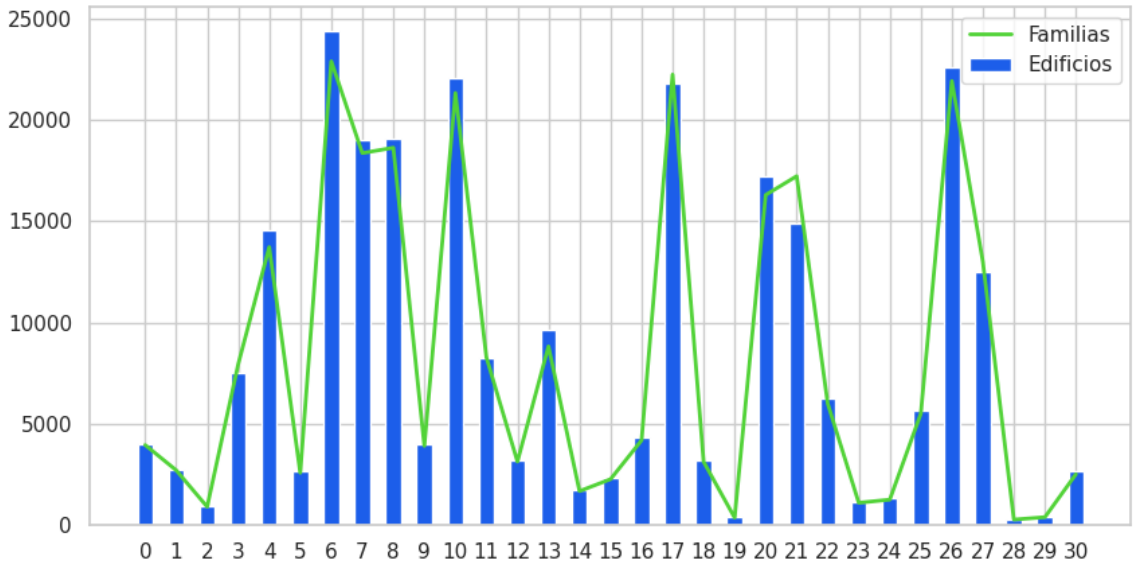


Figura 27: Cantidad de familias vs cantidad de edificios por geo_level_id_1

Efectivamente, la cantidad de familias es proporcional a la cantidad de edificios. Por lo tanto podemos entender que cuando hablamos de daños en edificios también estamos viendo como le afectaron a las personas.

6.1. Localización

Antes de continuar con los daños, queremos mostrar que tan representativo es el geo_level_id_1 para mostrar los daños por región.

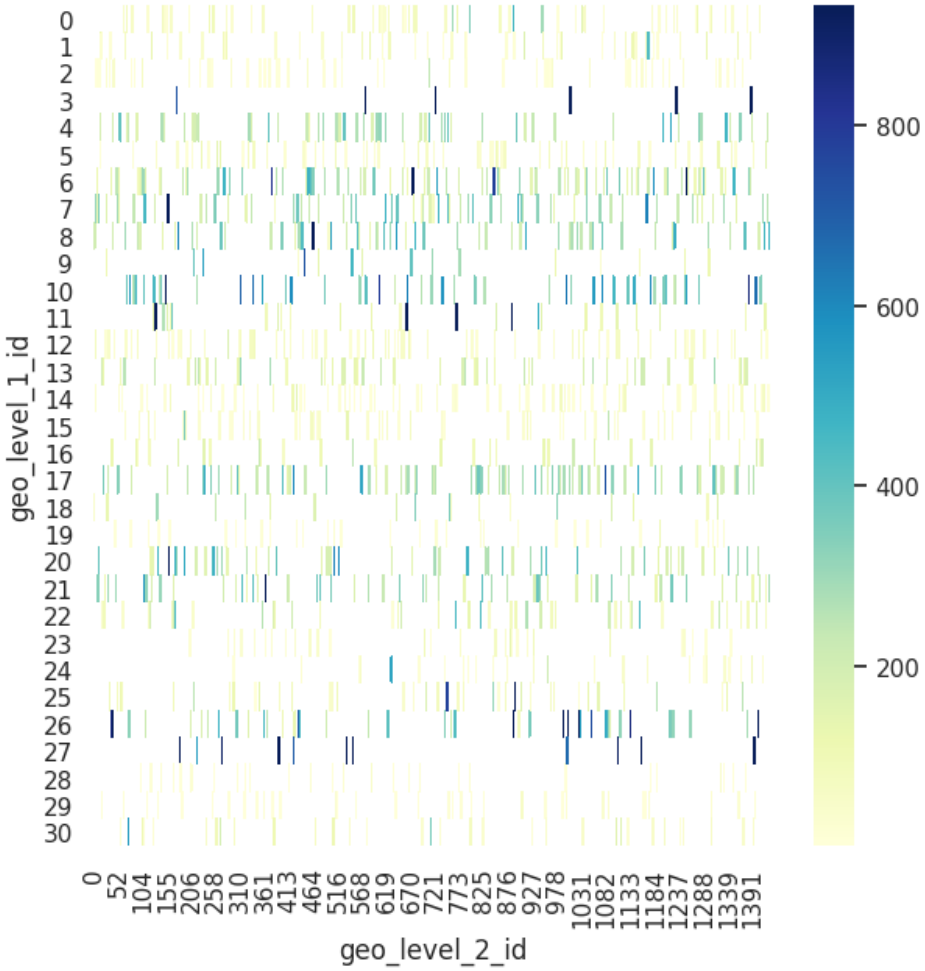


Figura 28: geo_level_id_1 y geo_level_id_2

- La mayoría de las regiones se distribuyen de manera uniforme en sus regiones
- Casos particulares donde no sucede esto, regiones: 3, 11 y 27.

Por lo tanto de aquí en más, utilizaremos geo_level_id_1 como eje de ubicación

6.2. Daños graves, medios y leves

Vamos a tratar de visualizar como ha afectado el terremoto a las distintas regiones del país.

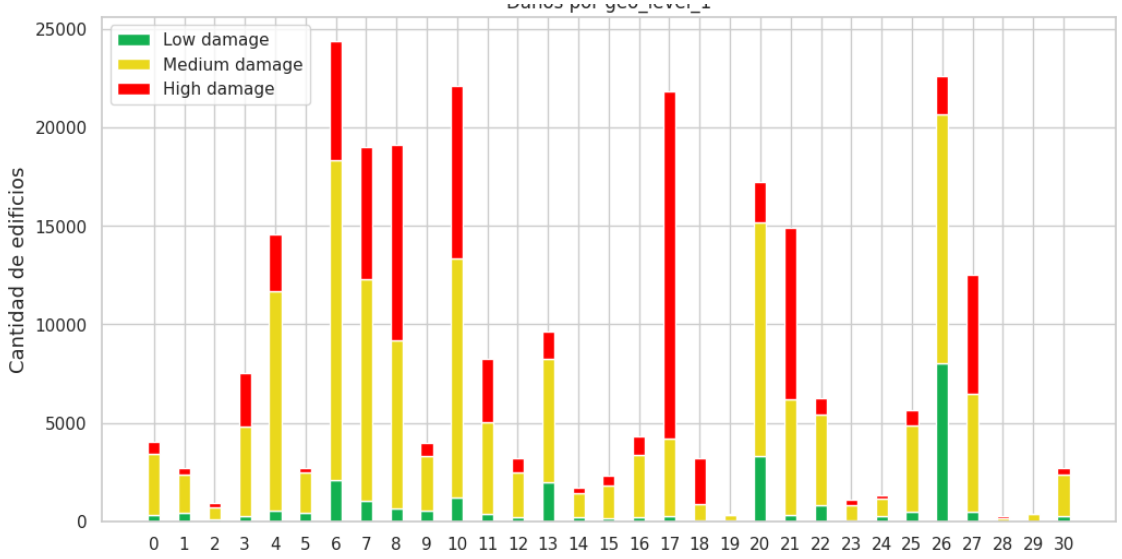


Figura 29: Daños en cada geo_level_id_1

La región que más recibió daños graves es la 17 por amplia diferencia. Le siguen de la 6 a la 10 y la 21. Ordenemos por daño recibido a ver si nuestras afirmaciones son correctas

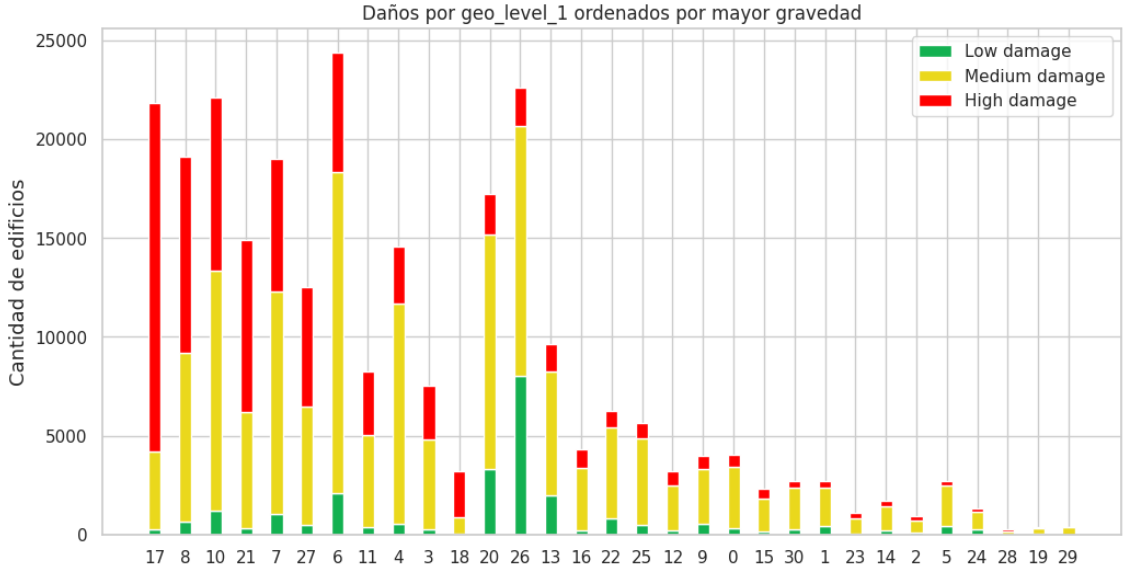


Figura 30: Daños en cada geo_level_id_1 ordenados por daños graves

Las preguntas que nos podemos hacer son las siguientes:

- ¿Aquellas regiones que más cantidad de daños graves tuvieron, están más cerca del epicentro?
- ¿Puede que esto sea fruto de que simplemente haya más cantidad de edificios en esa zona?

Estas preguntas están relacionadas, veamos que pasa si ordenamos de manera porcentual respecto de la cantidad de edificios de la region

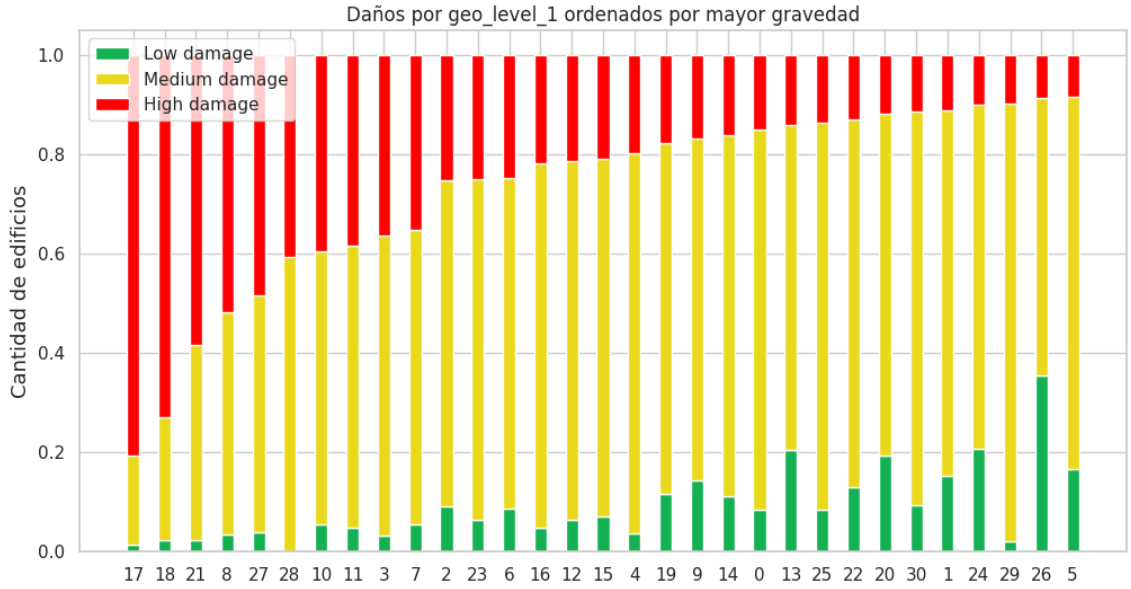


Figura 31: Edificios ordenados por daños graves respecto del total de edificios de la region por geo_level_id_1

Aca se puede ver que por ejemplo la región 18 fue bastante afectada respecto de la cantidad de edificios en la región. Esto no se veía en el grafico anterior. También podemos ver que los daños leves apenas llegan al 10 % para las primeras diez regiones ordenadas de esta forma.
Ya sabemos en que zonas se encuentran las mas afectadas en terminos de cantidad de edificios. Veamos que podemos averiguar de estas zonas.

6.3. Estructuras en las zonas más afectadas

Tomemos las 10 regiones mas dañadas gravemente según la cantidad de edificios. (Pues ya hicimos un análisis de las estructuras por zona geográfica)

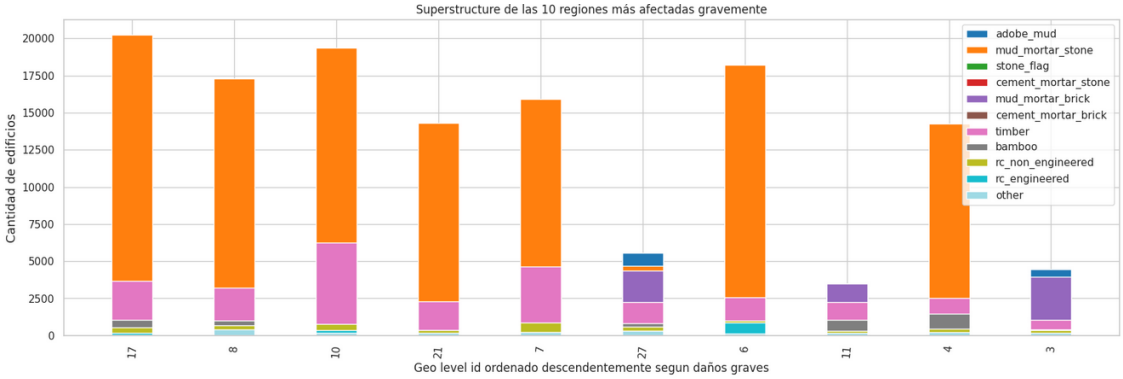


Figura 32: Superstructure de las 10 regiones más afectadas gravemente

- En todas estas regiones tienen como material más usado al 'mud_mortar_stone', salvo las regiones 27,11 y 3.
- En las regiones 27,11 y 3 el material más usado es el 'mud_mortar_brick'

Lo importante de esta visualización es como se distribuye el material según va decreciendo en cantidad de afectados gravemente. Sin embargo no se ve una tendencia respecto a esto.

6.4. Altura y zonas afectadas

Queremos saber si la altura está relacionada con los daños recibidos por los edificios. Realizemos este análisis con los primeros 10 'height_percentage', pues los demás rompen la escala.

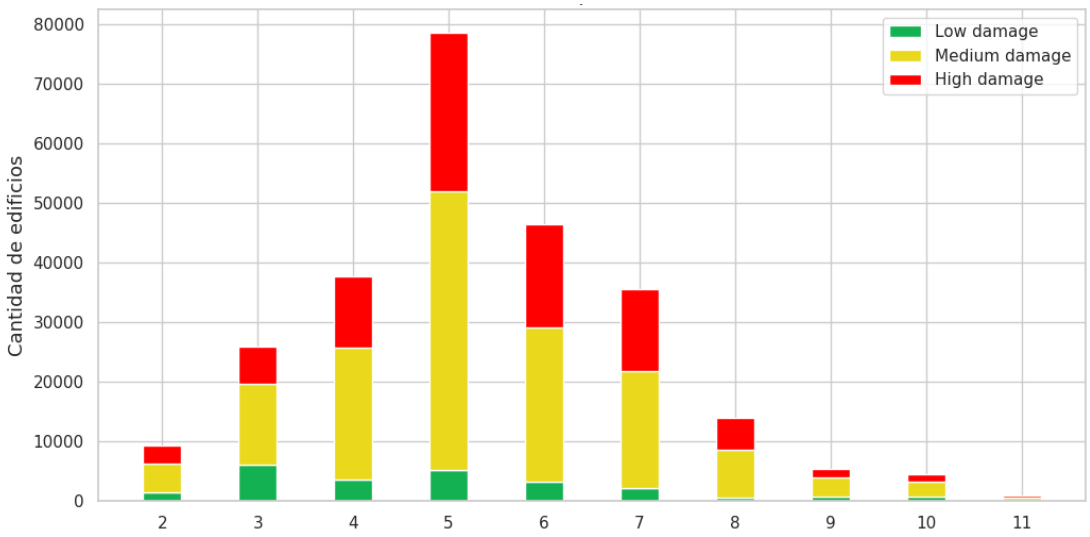


Figura 33: Daños por altura (primeros 10 'height_percentage')

A simple vista parece ser que de manera porcentual crece el daño grave cuanto mas alto es el edificio, verifiquemos esto.

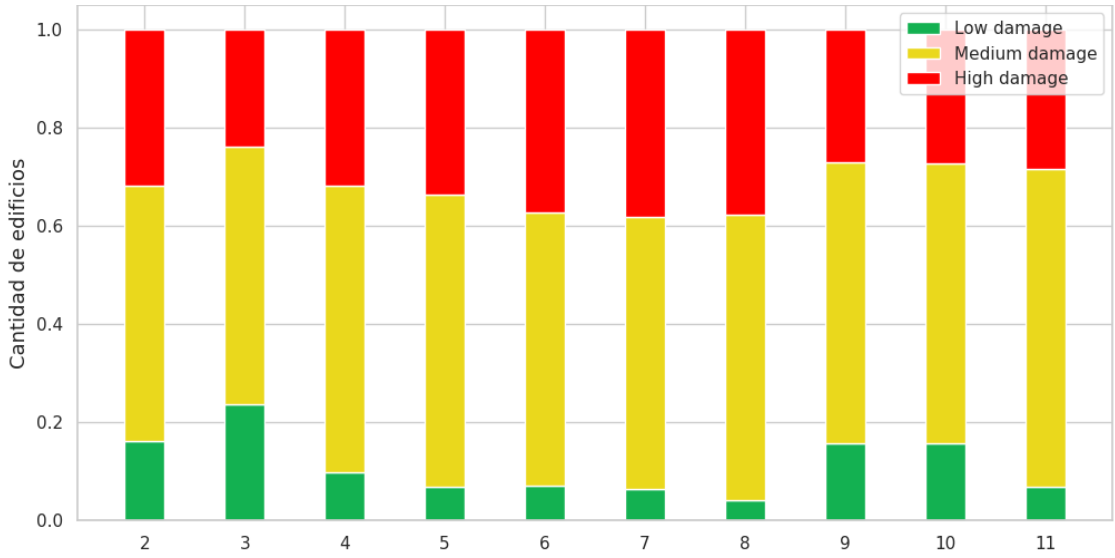


Figura 34: Daños por altura respecto del total de edificios de esa altura

Nuestra visión no es del todo errada pues se ve como crece hasta casi un 40 % los daños graves (altura 6 y 7) pero luego se ve como vuelve al 25 %.

Sin embargo puede estar siendo afectada por la cantidad de edificios.

6.5. Pisos y zonas afectadas

Normalmente cuando ocurren terremotos se piensa en los edificios más altos derrumbados. Veamos que daños sufrieron los edificios con más plantas. Debería ser similar al análisis de altura. Comprobémoslo

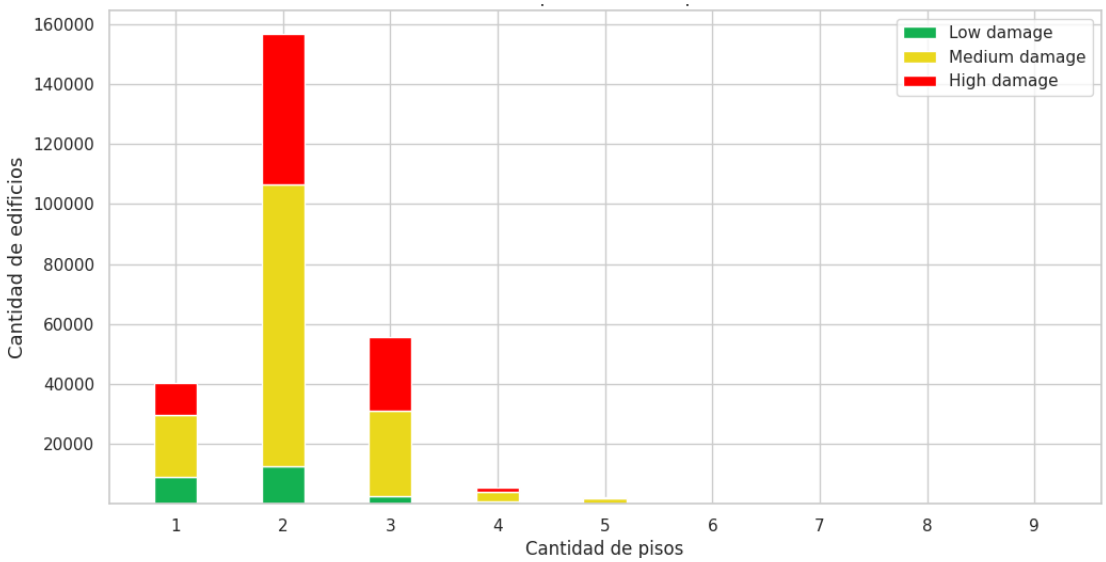


Figura 35: Daños por cantidad de pisos

No se ve una tendencia marcada con la cantidad de pisos ya que la cantidad de edificios con 4 pisos o mas es muy poco representativa

Parte IV
Conclusiones

Luego de analizar el set de datos, se pudo llegar a las siguientes conclusiones:

No aporta tratar de hacer un analisis de daño en relacion con uso secundario de los edificios ya que solo se trata de un 11 % del total, ademas la gran mayoria es de uso secundario agricola. Tampoco aporta al analisis el relacionar daño con los tipos de construccion sin tener en cuenta las posiciones geograficas donde fue mas fuerte el terremoto, ni cual es el mas frecuente en los edificios.

Si se tuviera que predecir qué edificios serían los menos afectados, serían aquellos que cuentan con las siguientes características:

- Muy poca antigüedad (Véase Figura 10)
- Mayor área (Figura 11)
- Menor altura (Figura 12 y 33)
- Súperestructura de Concreto Reforzado Diseñado (Figura 22 y 24, para compararla con el resto), el cual cuenta con el menor desgaste en el tiempo.
- Esar ubicado en la región 29 (geo_level_id_1) [Figura 29 y 30]

Como caso contrario, si hay que predecir qué edificios serían los más afectados cuentan con estas características:

- Mayor antigüedad
- Menor área
- Mayor altura
- Súperestructura de Piedra
- Estar ubicado en la región 17 (geo_level_id_1)