FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico Nº1: Análisis Exploratorio de Datos

[75.06] Organización de Datos - FIUBA

Autores

Buceta, M. Belen	102121	bbuceta@fi.uba.ar
Companys, Gonzalo Alejo	103026	gcompanys@fi.uba.ar
Di como, Juan Pablo	102889	jpdicomo@fi.uba.ar
Jure, Federico	1234	fedejure@gmail.com

https://github.com/jpdicomo/Tps-Datos/tree/main/TP1

Cátedra: Argerich

Cuatrimestre: Primer Cuatrimestre 2021

Lenguaje elegido: Python

Índice

1. Introducción

El presente informe reúne la documentación correspondiente al primer trabajo práctico de la materia Organización de Datos. El mismo consiste en un análisis exploratorio de datos realizado en Pandas sobre dos sets de datos que contiene información acerca del terremoto que tuvo lugar en Gorkha, Nepal en el año 2015, el cual registró una magnitud de 7.8 en la escala Richter y tuvo su epicentro en la ciudad de Kathmandu..
Los sets de datos pueden obtenerse en:
https://www.drivendata.org/competitions/57/nepal-earthquake.

2. Objetivos

El trabajo tiene los siguientes objetivos:

- Determinar características y variables importantes de los sets de datos provistos.
- Encontrar y establecer correlaciones entre variables.
- Estructurar y guiar el análisis en base a responder las interrogantes planteadas y elaboradas a partir del estudio de la información brindada por los datasets.
- Implementar visualizaciones que muestren a simple vista las respuestas a cada una de las cuestiones desarrolladas.

3. Información Datasets

3.1. Análisis de la estructura de los datos

Lo primero que hacemos es analizar la estructura de los datos, para poder entender con que estamos trabajando.

Contamos con un total de 260601 observaciones con 39 atributos cada una. La descripción de estos últimos puede encontrarse en drata driven, cuyo link fue provisto en la sección introducción.

3.2. Conversión de los tipos de datos

Cuando se leen los archivos con extensión .csv provistos, Pandas infiere el tipo de dato de cada columna. Generalmente esto lleva a utilizar mas memoria de la realmente necesaria. Es por esto que realizamos un análisis individual de cada columna, junto a la descripción de la misma provista en DrivenData, para convertir el tipo de los datos de cada una.

- **Conversión a datos categóricos:** En el dataset provisto tenemos diferentes columnas que presentan valores obfuscados. Ej: la columna 'roof_type"puede tomar los valores n, q, x. No sabemos que representa cada uno de estos valores pero si sabemos que son consistentes dentro de una misma columna. Para estas columnas decidimos tratar los datos como categóricos, dado que los valores que tomar son acotados. Dichas columnas son:
  - land_surface_condition
  - foundation_type
  - rood_type
  - ground_floor_type
  - other_floor_type
  - position
  - plan_configuration

- legal_ownership_status

- **Conversión a datos numéricos:** Para la conversión de datos numéricos lo que hicimos fue analizar las columnas cuyos datos eran valores numéricos (exceptuando aquellas que eran del tipo binario, estas las tratamos en la conversión siguiente) y analizamos si los valores posibles son únicamente enteros, su máximo y mínimo valor. La razón para hacer esto se debe a que, como se dijo antes, pandas infiere el tipo de dato de cada columna al leer el dataset, asignándole en este caso el tipo de dato int64 a nuestras columnas. Esto es un desperdicio dado que para ciertas columnas con 8 bits es suficiente para el rango de valores que presenta dicha columna. Por eso las conversiones que hicimos se basa en la cantidad de bits que se usan para cada uno. A continuación mostramos las columnas que transformas en el formato (columna, tipo de dato viejo, tipo de dato nuevo).
  Nota: Usamos la biblioteca Numpy para los tipos de datos. Abreviamos Numpy como np.

  - ('building_id', int64, np.int32)
  - ('geo_level_id_1', int64, np.int8)
  - ('geo_level_id_2', int64, np.int16)
  - ('geo_level_id_3', int64, np.int16)
  - ('count_floors_pre_eq', int64, np.int16)
  - ('age', int64, np.int16)
  - ('area_percentage', int64, np.int8)
  - ('height_percentage', int64, np.int8)
  - ('count_families', int64, np.int8)

- **Conversión a datos booleanos:** Las columnas que se especifican como tipo binario (posibles valores 1 o 0) determinan una condición de verdad sobre la observación en cuestión. Por lo tanto, decidimos convertir los datos de estas columnas al tipo booleano. Dicha columnas son:

  - has_superstructure_adobe_mud
  - has_superstructure_mud_mortar_stone
  - has_superstructure_stone_flag
  - has_superstructure_cement_mortar_stone
  - has_superstructure_mud_mortar_brick
  - has_superstructure_cement_mortar_brick
  - has_superstructure_timber
  - has_superstructure_bamboo
  - has_superstructure_rc_non_engineered
  - has_superstructure_rc_engineered
  - has_superstructure_other
  - has_secondary_use
  - has_secondary_use_agriculture
  - has_secondary_use_hotel
  - has_secondary_use_rental
  - has_secondary_use_institution
  - has_secondary_use_school
  - has_secondary_use_industry

- has_secondary_use_health_post
- has_secondary_use_gov_office
- has_secondary_use_use_police
- has_secondary_use_other

4. Análisis Introductorio de los features

4.1. Damage Grade

Comenzamos por analizar la distribución de los tipos de daño.

Figura 1: Cantidad de edificios por grado de daño

Podemos ver que el daño a los edificios se concentra entre el grado 2 y el grado 3 alcanzando su máximo en grado 2, con 148259 edificios, seguido por el grado 3, con 87218 y por ultimo el grado 1 con 25124.

4.2. Materiales Usados

. Procedemos a analizar los usos de los materiales en los edificios

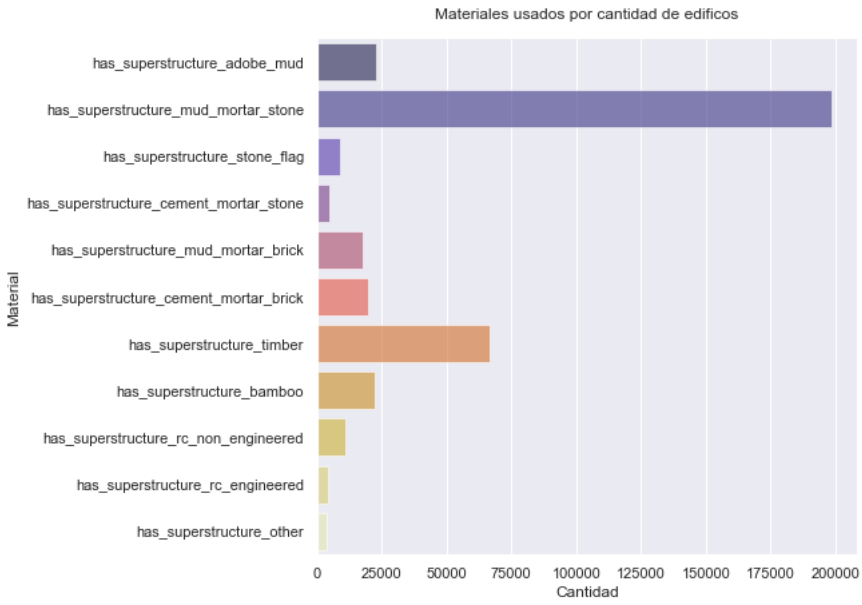

Figura 2: Materiales usados por cantidad de edificios

Esta visualización nos muestra en proporción los materiales utilizados para la construcción de los edificios. Viendo de esta forma que los materiales más utilizados son:

- has_superstructure_mud_mortar_stone
- has_superstructure_timber
- has_superstructure_adobe_mud
- has_superstructure_bamboo
- has_superstructure_mud_mortar_brick
- has_superstructure_cement_mortar_brick.

4.3. Planes de configuracion adoptado

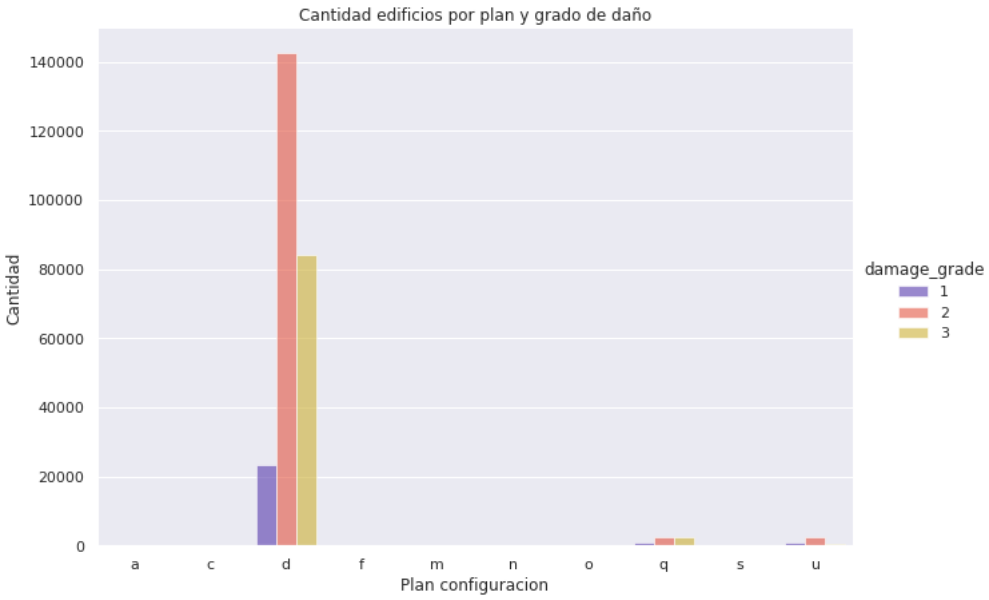

Figura 3: Plan adoptado por cantidad de edificios

El plan adoptado por la mayoría para la construcción de los edificios fue el "d". Representando casi un 96 % del total. Por ende cabe preguntarnos, para el resto de los datos, ¿cual es el aporte de los mismos?¿Como deberíamos tratar al resto de esos datos?

4.4. Correlación entre Features

Procedemos a analizar la relación que existe entre todas las variables. De mas claro a mas oscuro es que tan relacionadas.[es]tán dos variables.

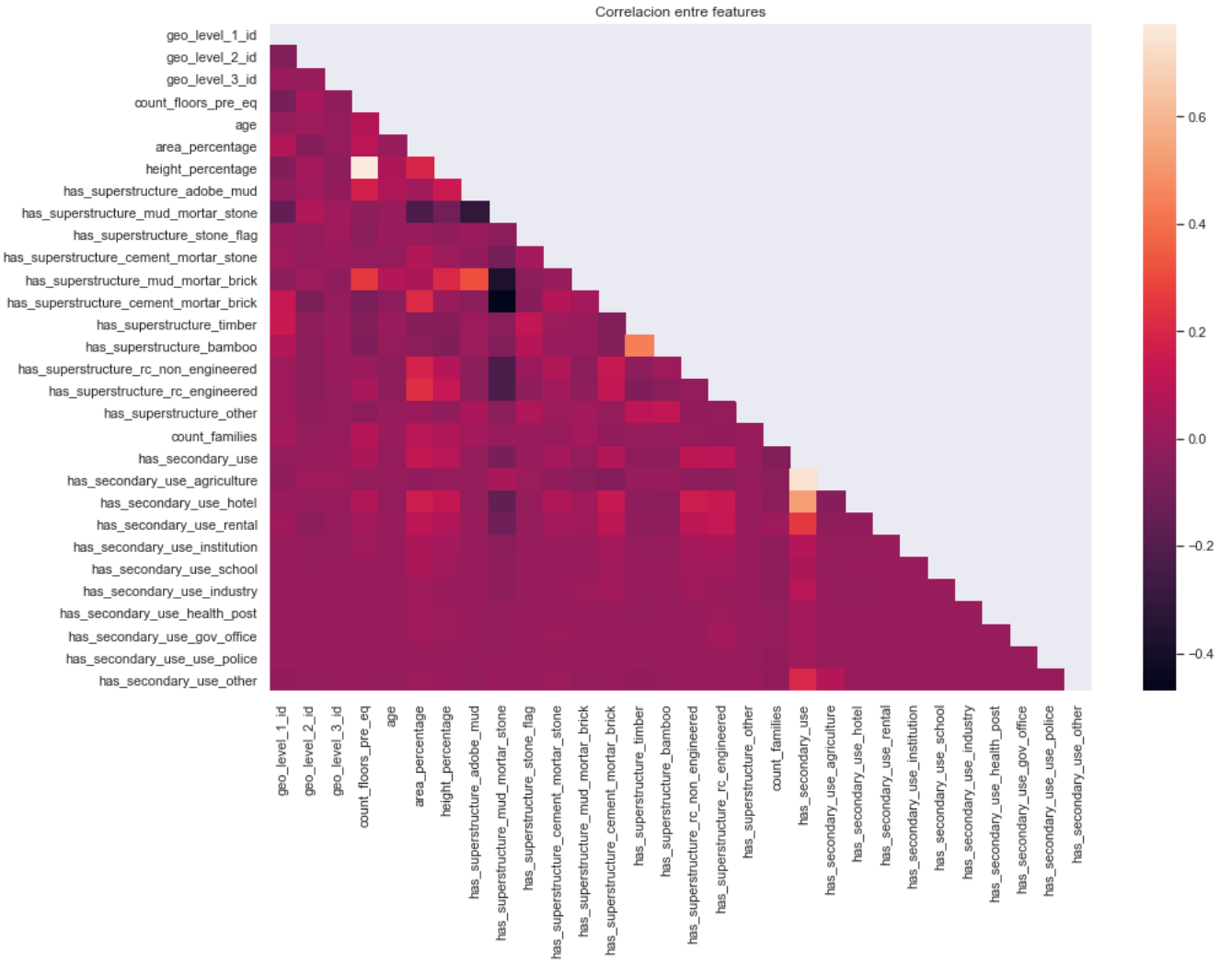

Figura 4: Relación entre todas las variables

4.4.1. Correalciones positivas

Se pueden observar algunos puntos críticos:

- La altura con la cantidad de pisos, tiene sentido que el el coeficiente de correlación sea muy cercano a 1 ya que mientras mas pisos tenga un edificio mas alto será el mismo.
- Otro punto critico que se puede observar es "has_secondary_useçon "has_secondary_use_agriculture", de la cual deducimos que los que tienen un uso secundario son mas probables a tener un uso agricultor.
- De has_superstructure_timber y has_superstructure_bamboo se deduce que si un edificio usa alguno de estos materiales, es mas probable que use el otro.

- Asi mismo de has_superstructure_mud_mortar_brick y has_superstructure_adobe_mud se deduce lo mismo, con menor coeficiente de correlacion.

4.4.2. Correalciones negativas

- Dentro de la eleccion de los materiales a usar, el cement_mortar_brick y el mud_mortar_stone son los que menos se encontraran utilizandose juntos
- Se puede observar que el uso de mud_mortar_stone no es muy popular en combinación, a parte de no ser para nada compatible con cement_mortar_brick, tiene otras malas sinergias como el mud_mortar_brick.
- Otros dos materiales que no parecen soler usarse juntos son mud_mortar_stone con el adobe_mud

4.5. Tierra, Cimientos, Pisos y Techo

Procedemos a analizar el tipo de tierra predominante junto con los cimientos, pisos y techos mas usados:

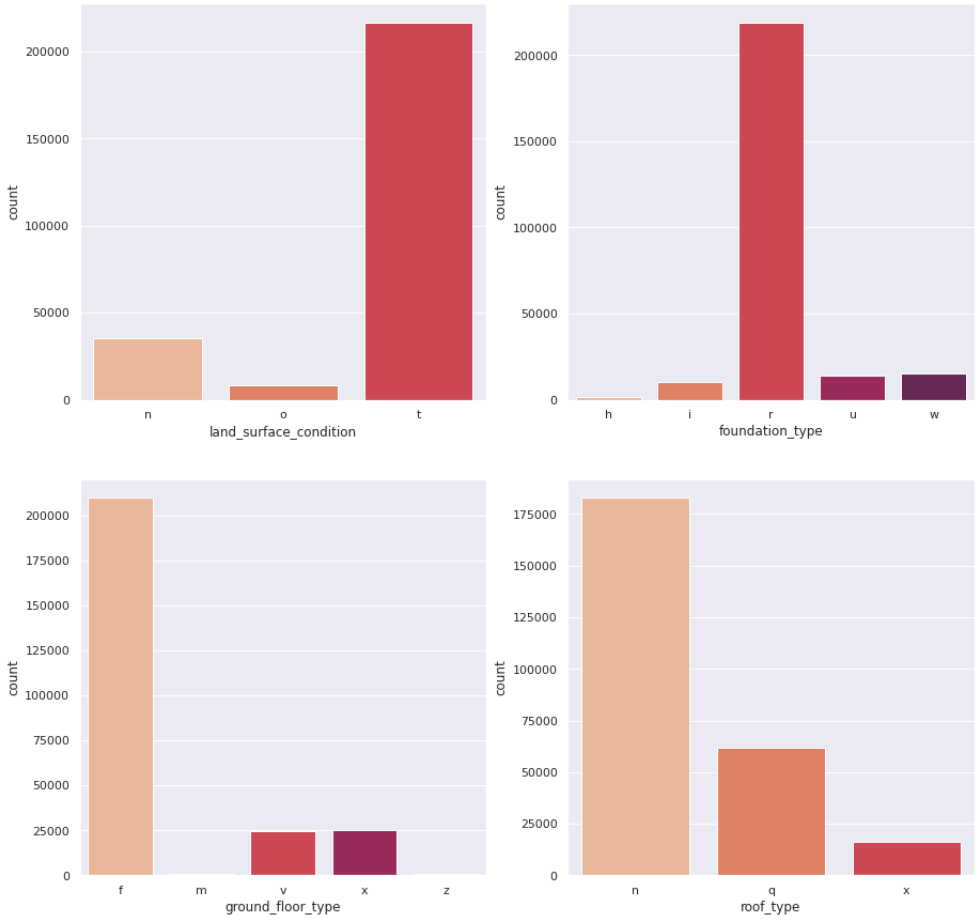

Figura 5: Tipo predominante

De los gráficos podemos deducir fácilmente cuales son los tipos predominantes en cada edificio. Estos serian para cada feature:

- land_surface_condition: t

- foundation_type: r
- ground_floor_type: f
- rood_type: n

En nuestro datos, esta combinación de features particular representa el 45 % de las observaciones. Aproximadamente la mitad de nuestros datos. Además, tomando solo las observaciones con esta combinación, podemos como es el grado de daño para dichas observaciones.

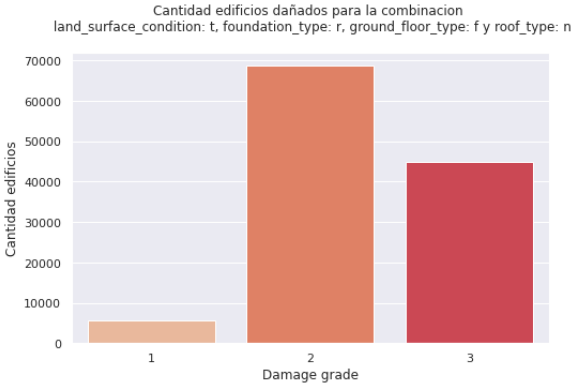

Figura 6: Cantidad de edificios dañados para la combinación de features bajo análisis

El total de observaciones para esta combinación es 119114, siendo el total para cada tipo de daño:

damage_grade	total observaciones
1	5626
2	68605
3	44883

Con esta ultima tabla podemos ver que el 95 % de las observaciones para este caso tienen un grado de daño 2 o 3.

Es por esto que vamos a analizar estos features particulares con los tipos de superestructuras usados en el apartado **7.2 Análisis de combinación de features**.

5. Análisis del efecto de la antigüedad del edificio en el daño causado

5.1. Desarrollo

Este análisis surge a partir de la pregunta: ¿Es la antigüedad un factor relevante para estudiar el daño causado en los edificios? ¿Podemos a partir de ella establecer claras tendencias en el daño producido en las edificaciones?.

Para llegar a las respuestas se comenzó por visualizar, sin ningún tipo de filtrado, un gráfico que muestre la cantidad de edificios por edad de antigüedad.

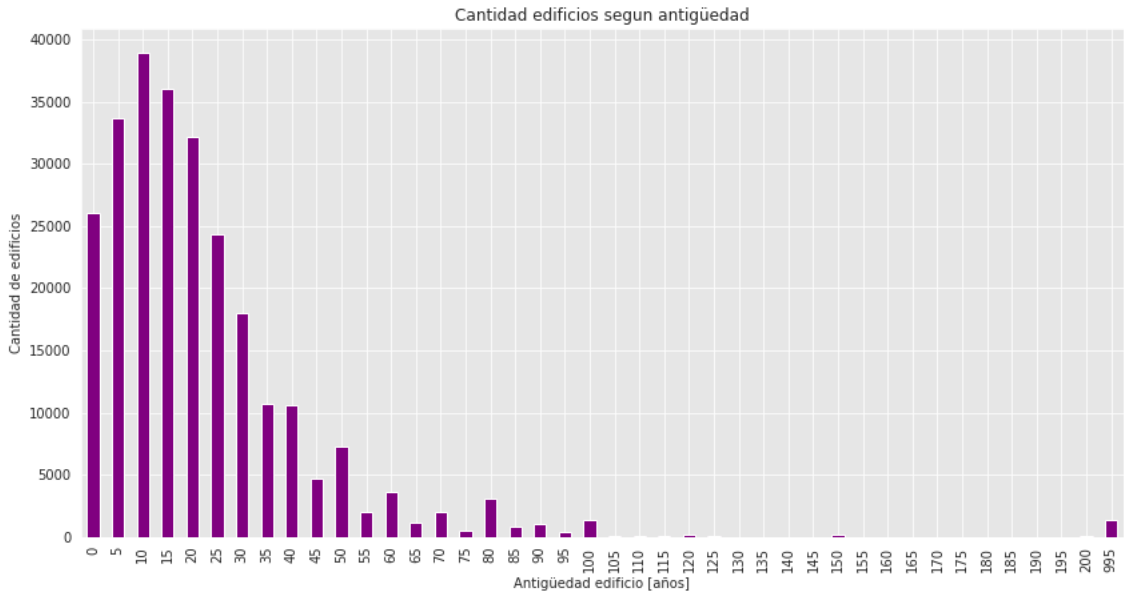

Figura 7: Cantidad de edificios según su antigüedad

A simple vista, pudimos notar como las edificaciones mayores a los cien años de vejéz eran lo suficiente excasas, tal que podían ser despreciadas en el resto del análisis. Para ser más rigurosos, llevamos a cabo un cálculo para determinar qué porcentaje representaban en la muestra. Obteniendo que menos del 1 % de los edificios tenían una antigüedad mayor a cien años.

Por lo tanto, para seguir adelante con el análisis exploratorio del efecto de la misma en el daño causado, consideraremos despreciables a aquellas antigüedades que superen los cien años.

5.2. Resultados

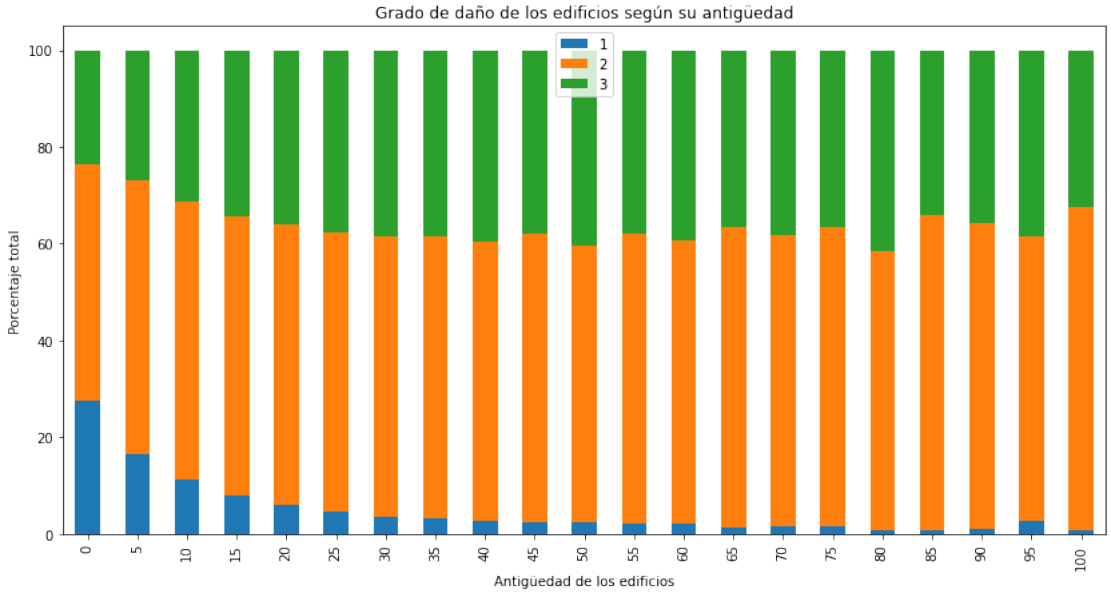

Figura 8: Grado de daño de los edificios según su antigüedad

5.3. Conclusiones

El predominante grado de daño entre todos los casos bajo observación es el correspondiente al daño medio (2).

Sólo los edificios con antigüedad menor a un año, tuvieron un porcentaje de daño del tipo 3 menor al de grado 1.

En los edificios con antigüedad menor a un año, el porcentaje de daño del tipo 3 (Serious damage) es menor al del grado 1 (Low damage). Esta característica sólo se da en este grupo de edificios. En el resto de los casos, el porcentaje de daño del grado tipo 3, supera el del grado 1.

Por último, la tendencia indica que a medida que aumenta la antigüedad, los edificios toleran menos la catástrofe y más daño se ocasiona.

6. Análisis de la altura de la edificaciones

6.1. Desarrollo

Para comenzar con este análisis, cabe primero realizar una visualización que nos de una idea sobre cómo se distribuyen las edificaciones según su altura y antigüedad.

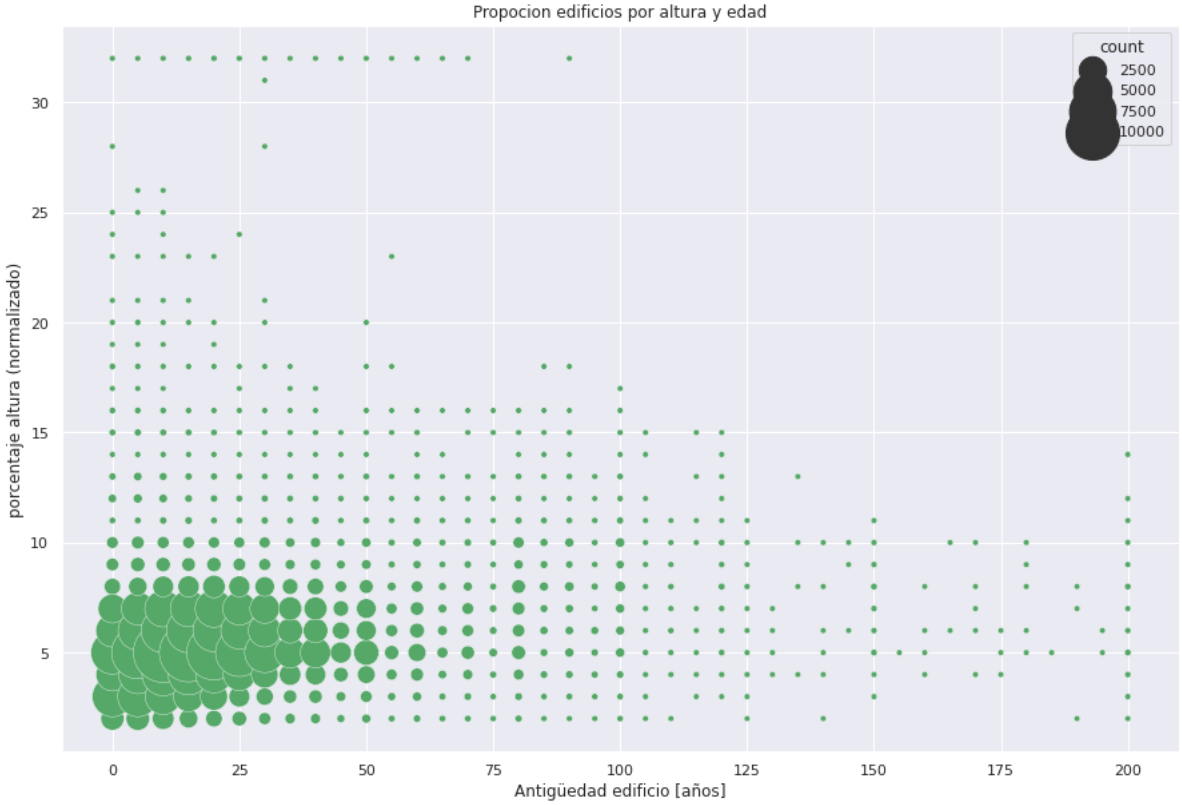

Figura 9: Proporción de edificios según su antigüedad y altura

A partir de ella, podemos resaltar que los edificios de antigüedad menor a cincuenta años muestran una gran concentración de construcciones que poseen una altura preponderante menor al 10 % (normalizado), en especial, alta aglomeración en las alturas del 5 % (normalizado). Esto indica además que la gran aglomeración de las construcciones tienen una antigüedad menor a los cincuenta años, dejando como minoría a las construcciones de mayor edad. Asimismo, a medida que los edificios aumentan su antigüedad, la tendencia en la densidad de la altura se ve más disipada pues la cantidad de edificaciones se reduce. Sin embargo, se hace notar que no hay variación en cuanto a las alturas construidas para las edifciaciones.

Siguiendo con el análisis y utilizando la información recolectada por el gráfico realizado anteriormente, nos interesa conocer cómo se distribuyó el daño ocasionado por la catástrofe natural, según las alturas de las edificaciones. Con la motivación de responder las interrogantes: ¿Se aprecia que ciertas alturas tuvieron un daño mayor a otras? ¿Existe un patrón entre el daño ocasionado y la altura que tenía el edificio? ¿Son las alturas más altas las más afectadas o las más bajas?.

6.2. Resultados

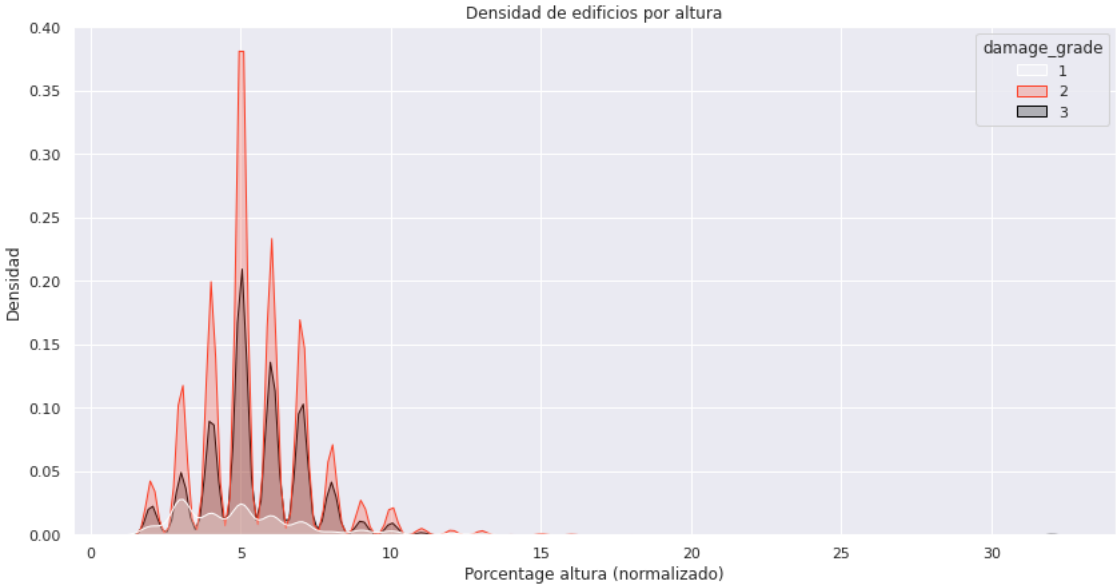

Figura 10: Densidad de edificios por altura

6.3. Conclusión

Se obtiene a partir de los resultados que para los tipos de daño 2 y 3, las densidades de las alturas siguen el mismo patrón, con una diferencia entre picos. En general, el daño sufrido por los edificios, independientemente de la altura, fue de grado medio a alto. De manera que podemos sostener que no se observa un daño superior o menor para determinadas alturas. Con esto también se concluye que no existe una evidencia que indique las alturas más altas sufrieron un daño mayor a otras más bajas, o viceversa. A su vez, con la información recolectada en el principio del análisis, conocemos la distribución de las alturas según su antigüedad. Esto hace que podamos desarrollar ciertas deducciones. En el gráfico 10 se observa que la altura de 5 % (normalizado) constituye la mayor densidad de todas, información que ya conocíamos a partir del gráfico 9 también sabemos que la mayor proporción de estos edificios tienen una antigüedad menor a cincuenta años, es decir, no son edificios de larga antigüedad y además su altura no es de las más altas.

Todo esto nos puede dar lugar a desarrollar una hipótesis sobre alguna de las razones por las cuales éste tipo de edificaciones relativamente jóvenes fueron tan agraviadas por la catástrofe. Físicamente, se puede decir que cuando cierta estructura se mueve en resonancia con un terremoto, las probabilidades de que la construcción sufra un daño grave es considerablemente alta. Sin embargo, se hace hincapié en que esto último resulta solamente una hipótesis sobre alguna de las razones por las cuales las edificaciones de ésta altura sufrieron un tipo de daño tan alto.

7. Análisis de la superestructura

En esta sección nos dedicaremos a analizar la relación que tienen los distintos tipos de superestructura en relación al daño recibido por los edificios.

7.1. Uso de las distintas superestructuras

Procedemos a analizar la cantidad de edificios que usaron cierta superestructura y qué porcentaje de daño recibió en relación al total.

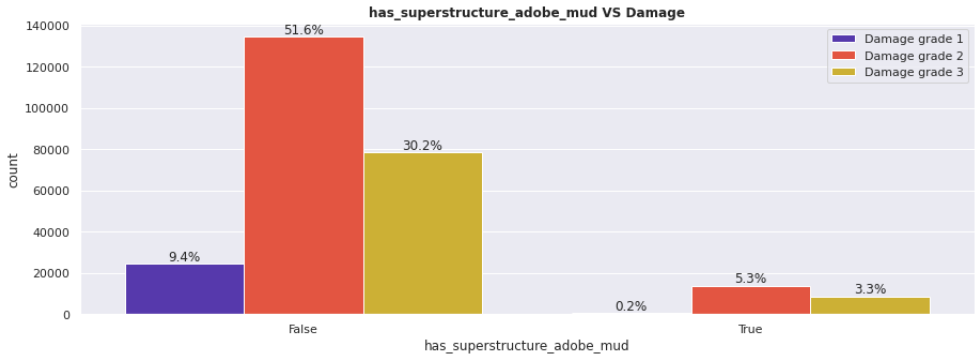

Figura 11: Cantidad de edificios que usan Adobe Mud

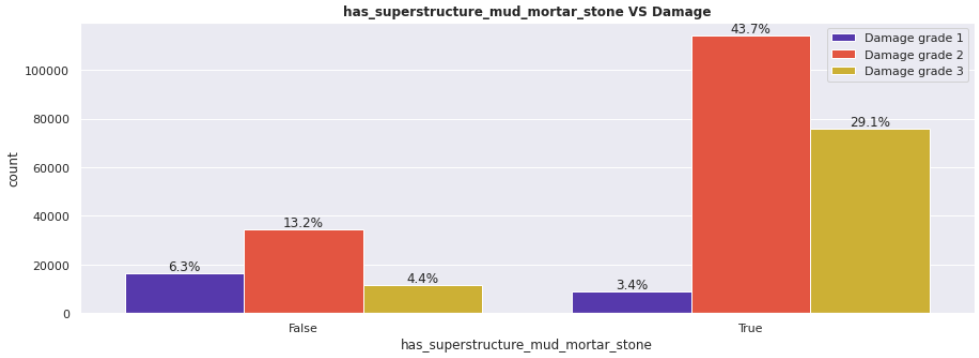

Figura 12: Cantidad de edificios que usan Mud Mortar Stone

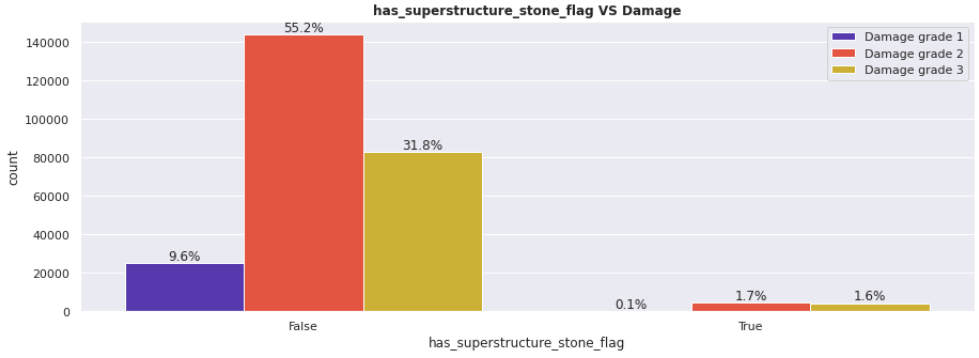

Figura 13: Cantidad de edificios que usan Stone Flag

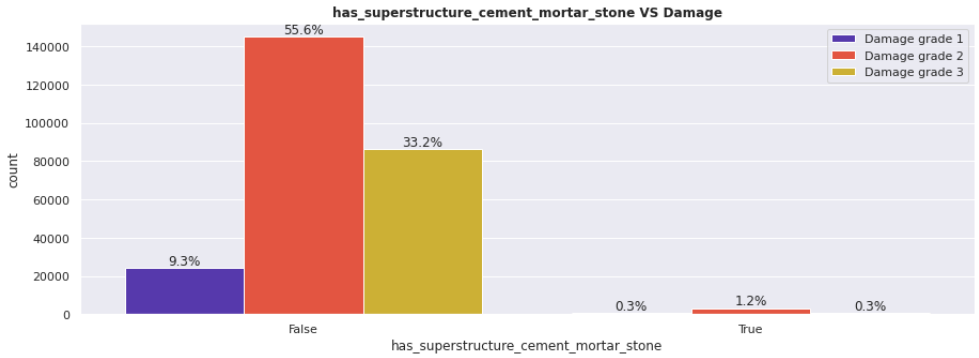

Figura 14: Cantidad de edificios que usan Cement Mortar Stone

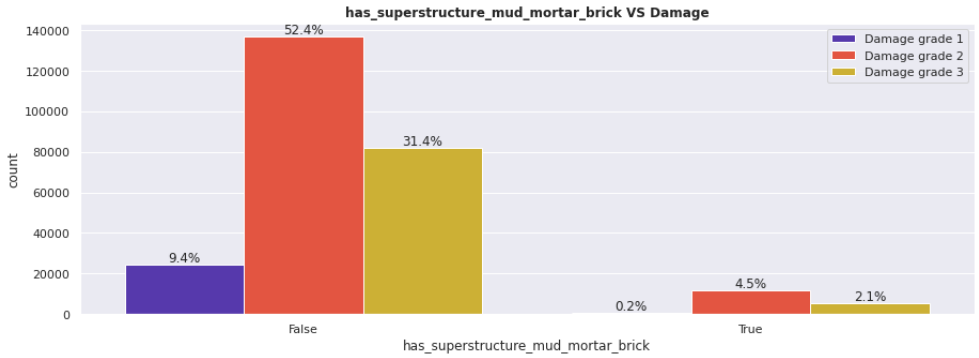

Figura 15: Cantidad de edificios que usan Mud Mortar Brick

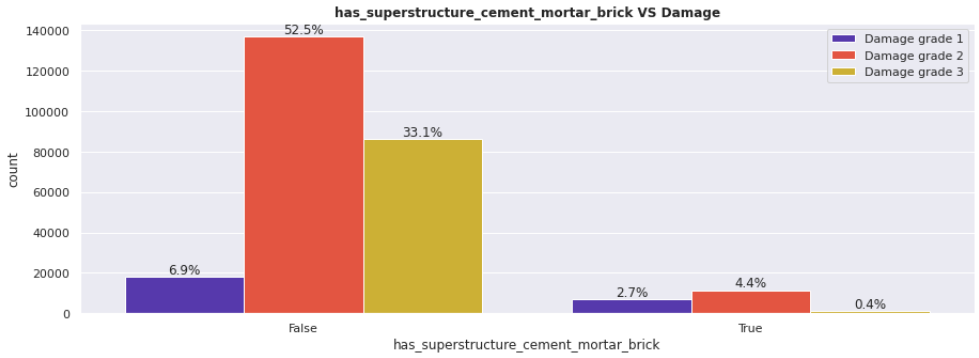

Figura 16: Cantidad de edificios que usan Cement Mortar Brick

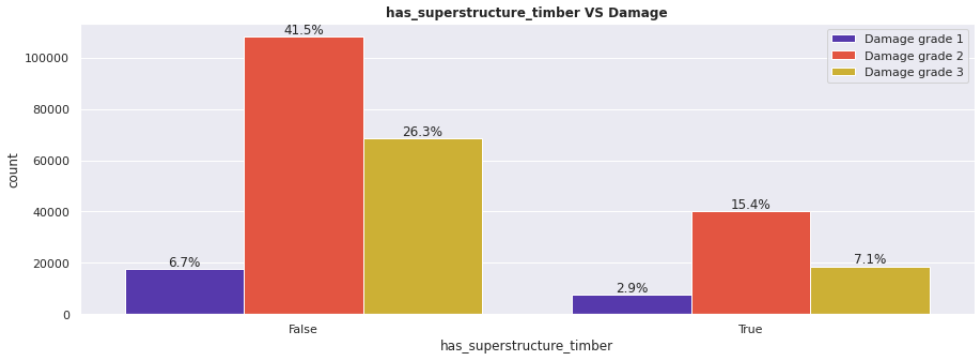

Figura 17: Cantidad de edificios que usan Timber

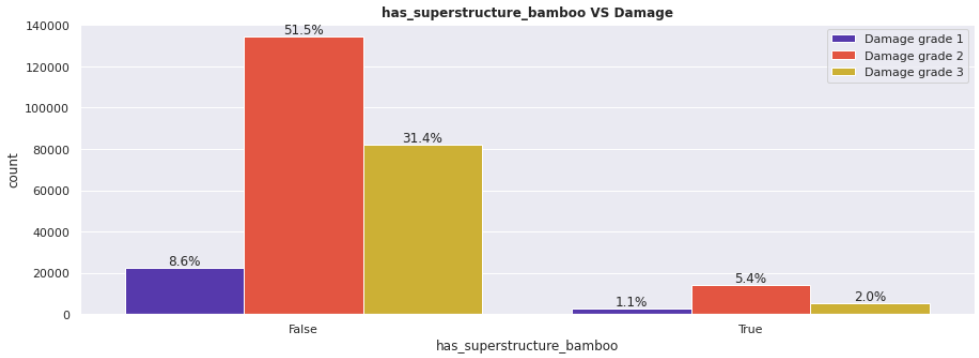

Figura 18: Cantidad de edificios que usan Bamboo

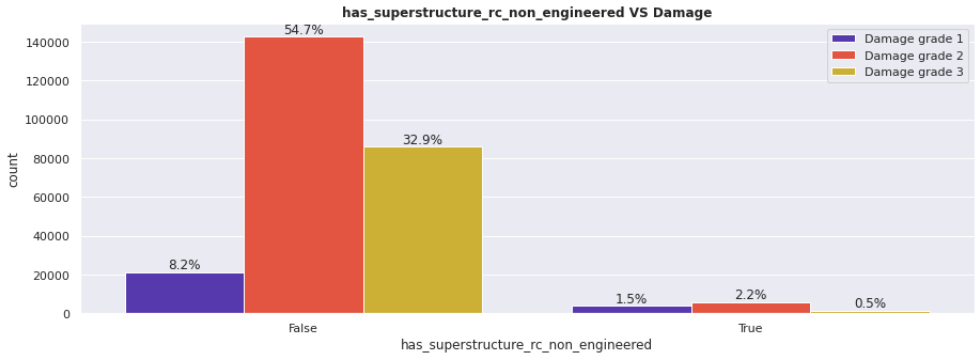

Figura 19: Cantidad de edificios que usan RC Non Engineered

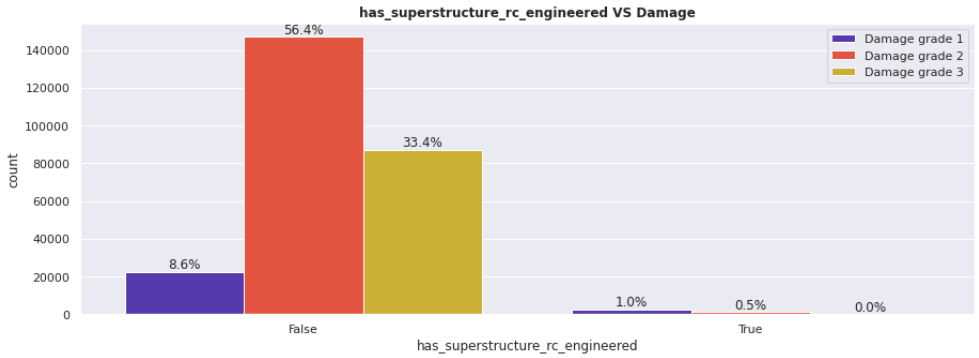

Figura 20: Cantidad de edificios que usan RC Engineered

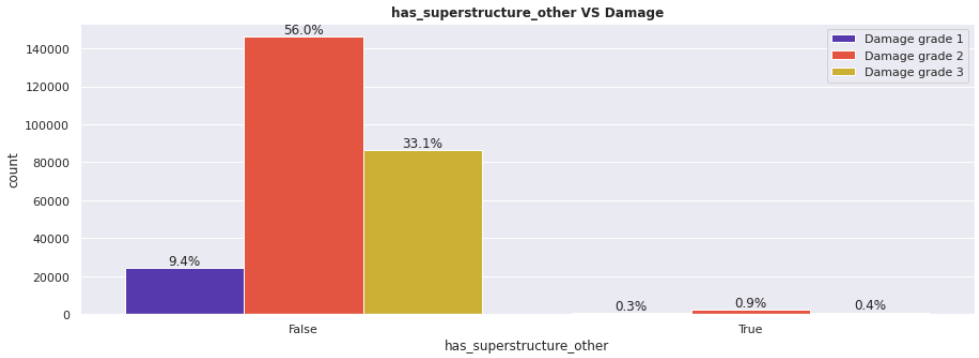

Figura 21: Cantidad de edificios que usan Other

De estos gráficos observamos la cantidad de edificios que usaron o no dicha superestructura. También está detallado el porcentaje con respecto al total de edificios que recibieron cada grado de daño. De este análisis podemos entender como están presentados los datos del uso de la superestructura en el dataset.

7.2. Análisis de combinación de features

Teniendo en cuenta el análisis de los tipos predominantes de tierra, cimientos, pisos y techos detallado en la sección 4.5 en la Figura 5, procedemos a hacer el siguiente análisis.
Primero creamos un nuevo feature llamado *superstructure_types* en el cual almacenamos todos los tipos de superestructuras que utiliza cada observación.
Luego, dado el tipo de tierra,cimientos,piso y techo, se busco la combinación de estos y superestructuras que mas edificios dañados tiene, agrupados por tipo de daño.
Comenzamos con las combinaciones del tipo (superestructura usada + tipo techo + tipo piso base).

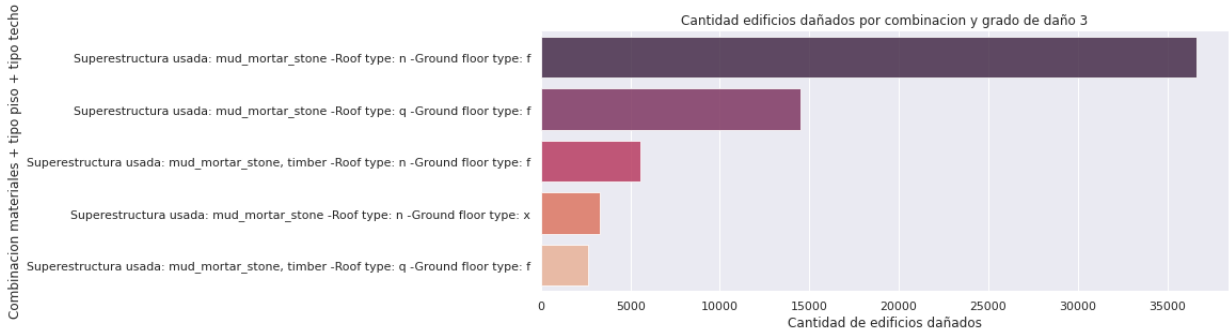

Figura 22: Combinaciones para daño de grado 3

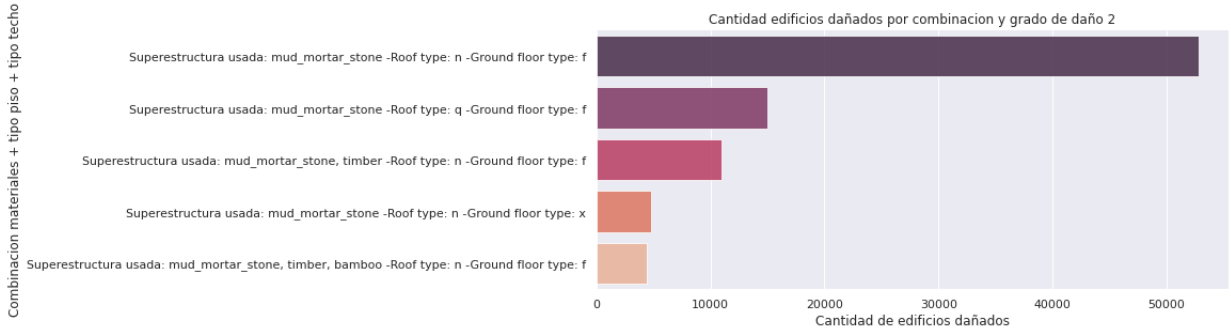

Figura 23: Combinaciones para daño de grado 2

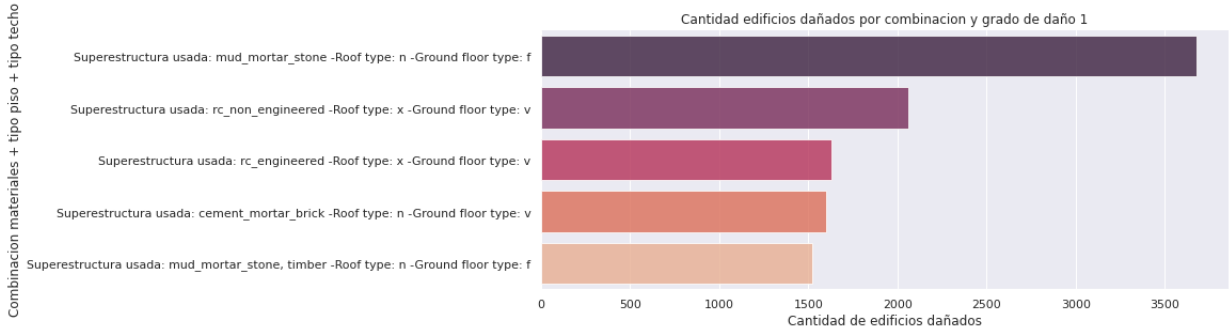

Figura 24: Combinaciones para daño de grado 1

A continuación procedemos con las combinaciones del tipo (superestructura usada, condición superficie, tipo cimiento)

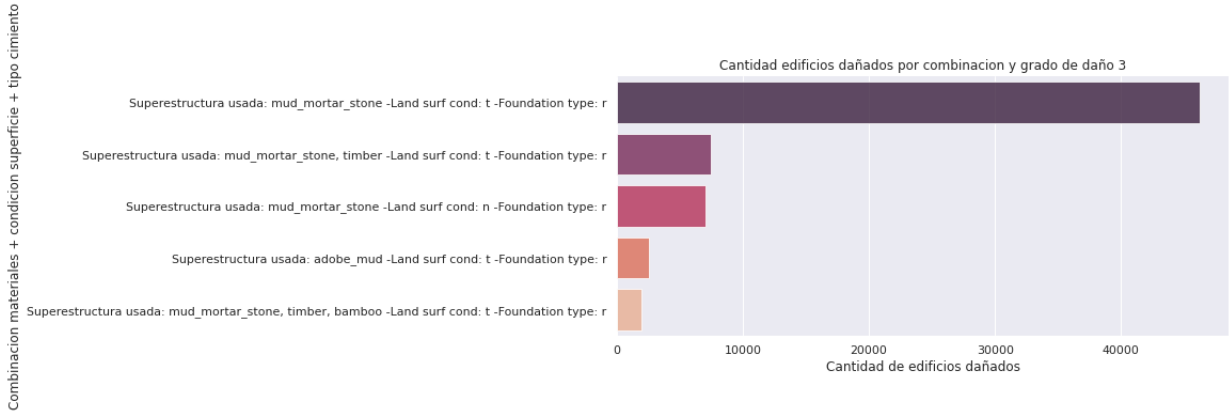

Figura 25: Combinaciones para daño de grado 3

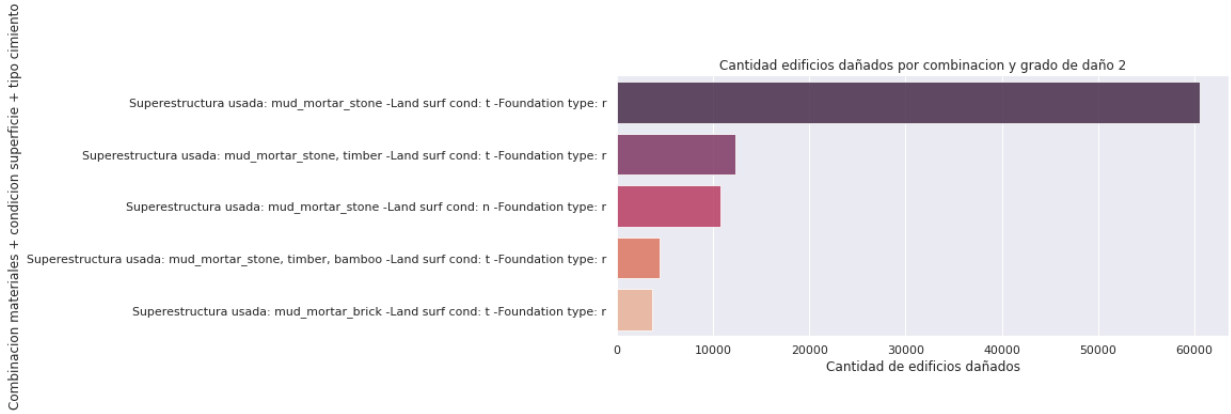

Figura 26: Combinaciones para daño de grado 2

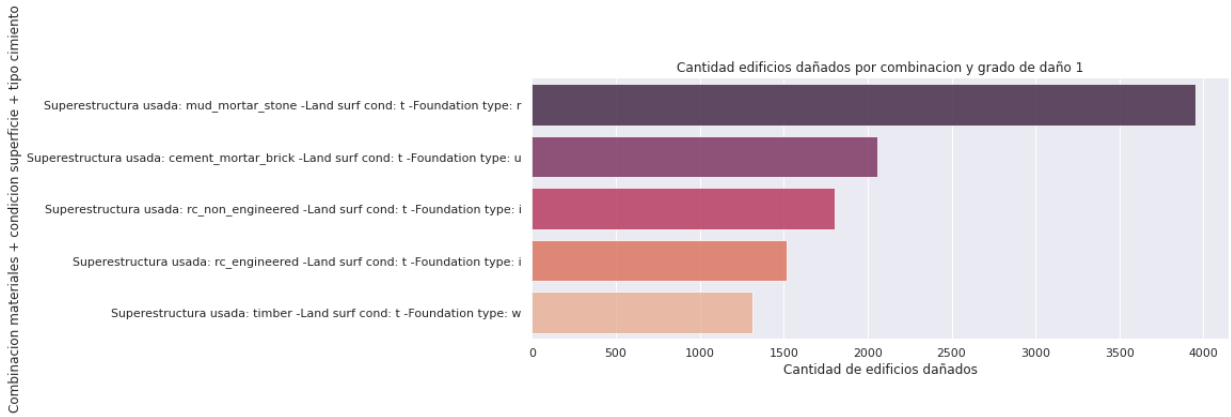

Figura 27: Combinaciones para daño de grado 1

7.2.1. Conclusiones

Se puede concluir de los datos analizados que las combinaciones que hacen que los edificios reciban un mayor daño son las detalladas en la figura 22 como un top 5. Asimismo las combinaciones que minimizan el daño son las presentadas en la figura 24.

7.3. Combinaciones Tierra, Cimientos y Piso vs. Daño

En esta sección nos proponemos analizar las distintas combinaciones entre el Tipo de tierra, cimientos y piso, para ver cuales son las que mas o menos daño recibieron.

Para este análisis, tomamos la variable daño como una variable continua, ya que deducimos que el grado de daño se refiere a la condición física del edificio luego del sismo, teniendo en cuenta que el grado 1 es el menor daño físico y el grado 3 el mayor, por lo que recurrimos a calcular el promedio del daño recibido por combinación y eso lo utilizamos como nueva variable en nuestro análisis.

Para evitar el problema de la ecuación más peligrosa (ecuacion de Moivre), recurrimos a filtrar los datos en relación a si la cantidad de ocurrencias de una misma combinación no superaba el promedio del total no la considerábamos relevante.

Aclaración: El feature damage_grade presente en train_labels es una variable ordinal (la podemos considerar categórica). Sin embargo, para este caso, la consideramos como una variable numérica.

Los mapeos entre los valores obfuscados de cada feature a valores numericos para poder representarlos mediante un plot de parallel coordinates es el detallado en la siguiente tabla.

Valor numerico	land_surface_condition	foundation_type	ground_floor_type
1	t	r	f
2	o	w	x
3	n	i	v
4	-	u	z
5	-	h	m

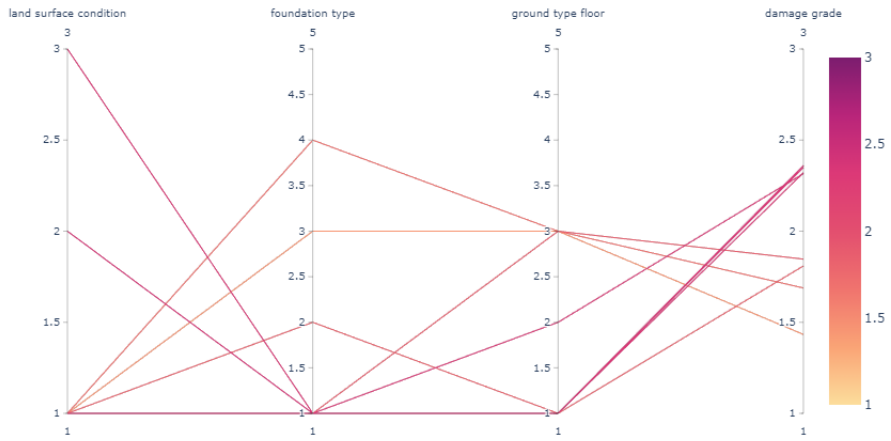

Figura 28: Combinaciones

En el gráfico 28 podemos apreciar las combinaciones mas relevantes y tu tendencia de daño.

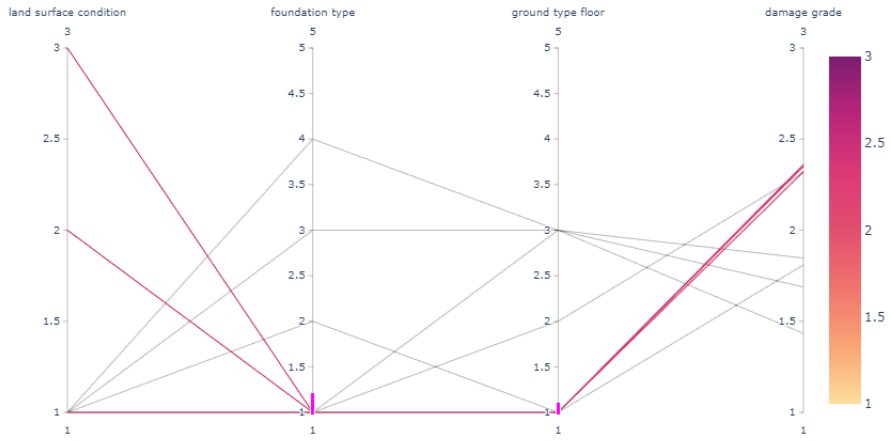

Figura 29: Independencia de los cimientos y los pisos

Algo que se puede ver a simple vista, es que independientes de el tipo de Tierra, los cimientos de tipo r y los pisos de tipo f, tienden en promedio a un daño cercano a los 2.7, que nos indica una tendencia a recibir daño alto.

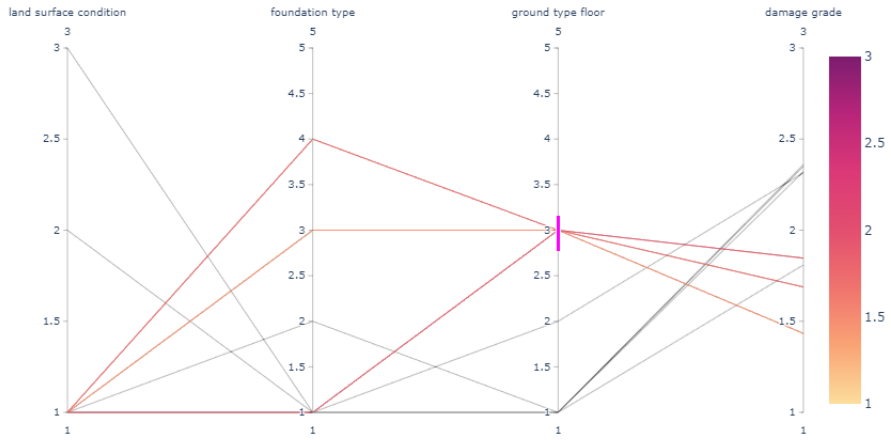

Figura 30: Independencia de los pisos

Independientemente del tipo de cimiento se puede observar que el tipo de piso v tiene una tendencia a daños bajos. Teniendo en cuenta que los datos relevantes de la tierra son solo del tipo t.

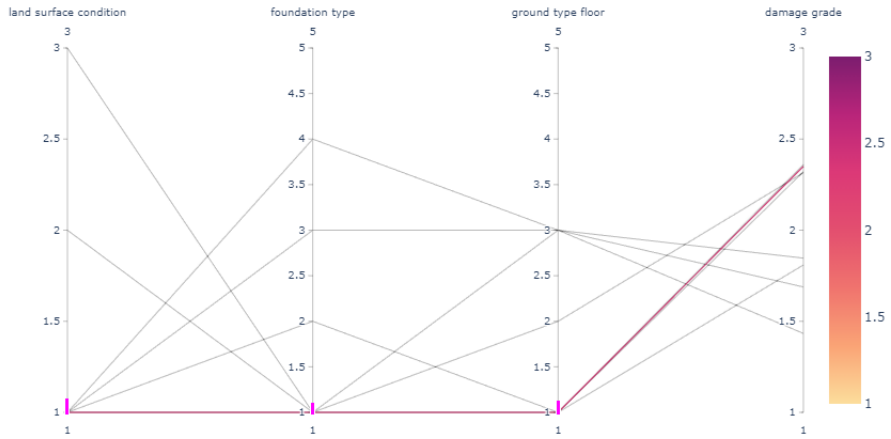

Figura 31: Combinación con mas daño

Se puede observar y corroborar que la combinación con mas daño es la t, r, f que concuerda con todos los datos analizados previamente.

8. Análisis de las regiones

8.1. Desarrollo

En el siguiente análisis, primero se distingue en el dataset proveído la especificación de tres columnas que nos indican la región geográfica en la cual existe un determinado edificio. En especial, tomamos interés particular de la columna geo_level_1_id que representa las regiones más grandes -un total de treinta regiones-. Lo que buscamos analizar es la proporción de los edificios según su región cuyo grado de daño fue 3.

8.2. Resultados

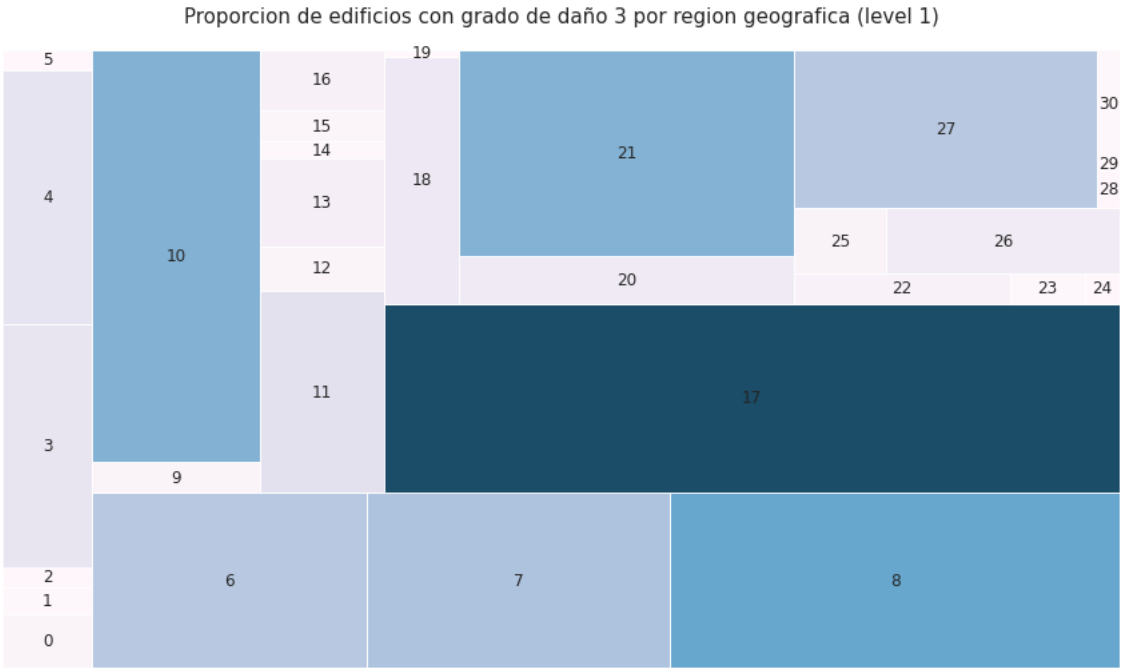

Figura 32: Proporción de edificios con grado de daño 3 por region geografica (level 1)

8.3. Conclusión

A simple vista, podemos notar como las regiones cuya intensidad de color es fuerte, indican que aquella es una de las zonas que sufrió un daño del tipo 3. En especial, podemos destacar que las regiones (17, 8, 21, 10, 27, 6, 7), son las que mayor proporción de daño del tipo 3 concentran. En la misma linea, esta visualización nos lleva a suponer que el epicentro del terremoto pudo haber ocurrido en la región 17. Debido a que de todas las expuestas, es la zona representada con el color más oscuro. Asimismo, teniendo en cuenta los colores del treemap, podríamos deducir que estas regiones que se presentan están próximas unas a otras.

9. Insights

9.1. Análisis Introductorio

En esta sección pudimos observar como se comportaban las distintas variables,y tuvimos una clara idea de como los datos estaban distribuidos notando que estos están desbalanceados.

9.2. Features importantes

- **land_surface_condition (lsc), foundation_type (ft), ground_floor_type (gft) y roof_type (rt)**

  Descubrimos que contábamos con una combinación predominante de estos features, siendo la misma : lsc: t, ft: r, gft: f y rt: n, y para dicha combinación tenemos que el 95 % de las observaciones presentan grado de daño del tipo 2 y 3.
  Otra combinación de interés es la lsc: t, ft: i, gft: v, ya que es la que presenta una tendencia a daños menores como se muestra en la figura 28

- **has_superstructure**
  Se puede observar que la superestructura predominante es la mud_mortar_stone siendo esta la única que llega a aparecer en el análisis de combinación de superestructuras con el tipo de suelo y techo por un lado, y por el otro con el tipo de tierra y cimiento.

  Se podría analizar la distribución de las regiones con respecto al epicentro dadas las edificaciones que presentan todos los tipo de de daño con combinaciones muy similares de superestructura teniendo como supuesto que mientras mas cerca del epicentro del sismo esté el edificio mas daño recibirá el mismo.

- **geo_level_1** Algo interesante que se deduce del análisis de esta feature, es la hipótesis de que la región 17 puede llegar a ser el epicentro del sismo dado que presenta la mayor tendencia a daño 3. Luego en base las secciones detalladas en la Figura 32 se pueden deducir las regiones mas cercanas al epicentro.

- **age** Una observación interesante es que el daño de grado 1 recibido por un edificio se comporta aproximadamente como una variable exponencial
  Algo que está bueno mostrar también es que se puede corroborar que en general un edifico mas antiguo tolera menos el daño de la catástrofe.