FACULTAD DE INGENIERIA
Universidad de Buenos Aires

Trabajo Práctico 1

[7506/9558] Organización de Datos
Curso 1
Primer cuatrimestre de 2021

Alumnos	Padrón
Elvis, Claros Castro	99879
Lerer, Joaquín	105493
Bach, Alan	91440
Reinaudo, Dante	102848

1. Introducción

El presente informe reúne la documentación de la solución del primer trabajo práctico de la materia Organización de Datos, el cual consiste en desarrollar un análisis exploratorio de un set de datos a partir de los conocimientos adquiridos en la materia. En particular, vamos a realizar el análisis sobre un set de datos con información acerca del impacto del terremoto Gorkha, ocurrido en Nepal durante el año 2015. Especialmente el dataset se enfoca en cómo eran las condiciones de una determinada vivienda y cuál fue su grado de daño luego del accidente.

2. Desarrollo

Para facilitar la lectura, a partir de este momento también se hará referencia al set de datos como “DataFrame”.

2.1. Preprocesamiento de los datos

Con el fin de optimizar el espacio utilizado y facilitar la futura investigación, se procedió a hacer una limpieza de los datos sobre los dos DataFrame involucrados. En principio ambos set de datos se encontraban en archivos .csv (valores separados por comas), disponibles en la web de Driven Data, y cuentan con información de encuestas realizadas por Kathmandu Living Labs y el Central Bureau of Statistics .

2.1.1. Columnas sin información

En primer lugar, se buscó si en los DataFrame existían columnas que no aporten ningún tipo de información o que tuvieran un único valor repetido a lo largo de todas sus filas, con el objetivo de descartarlas, ya que sería indiferente incluirlas e irrelevantes para la investigación. Luego se prosiguió a eliminar los valores nulos que pudieran existir en las columnas. Asombrosamente, en ninguno de los DataFrame analizados se encontraron columnas sin información, ni tampoco columnas con valores nulos.

2.1.2. Análisis del tipo de datos de las columnas

Con el objetivo de administrar el espacio utilizado, se analizó el valor de los tipos de datos de todas las columnas de nuestros dos DataFrame y se realizaron modificaciones sobre estos.

Al leer una columna del tipo entero con pandas, esta se almacena por default con un int64, este tipo de dato ocupa 64 bits para representar valores enteros [1]. Si el valor de los números a guardar es pequeño, se puede optimizar bastante el espacio utilizado empleando un tipo de dato entero que use menos bits para la representación. Se procedió a buscar el valor máximo de cada una de las columnas numéricas, para saber con cuantos bits es posible representarlas.

[1] Dado que el lector de este informe puede desconocer sobre informática y sistemas de numeración, cabe aclarar que en el sistema binario, con un número entero n de bits es posible representar valores de hasta 2^n .

Columna	Valor Máximo	Tipo de dato utilizado para la representación	Tipo de dato suficiente para la representación
building_id	1052934	Int64	Int32
geo_level_1_id	30	Int64	Int8
geo_level_2_id	1427	Int64	Int16
geo_level_3_id	12567	Int64	Int16
count_flours_pre_eq	9	Int64	Int8
age	995	Int64	Int16
area_percentage	100	Int64	Int8
height_percentage	32	Int64	Int8
count_families	9	Int64	Int8

Cuadro 1: Tipo de dato entero suficiente para almacenar la columna

Además, las columnas booleanas, inicialmente se almacenaron con un tipo de dato int64, pero fueron transformadas a un tipo de dato booleano (bool), para mejorar la perfomance de nuestro código.

Por último, las siguiente columnas fueron tratadas como tipo categórico (category):

- damage_grade
- land_surface_condition
- foundation_type
- roof_type
- ground_floor_type
- other_floor_type
- position
- plan_configuration
- legal_ownership_tatus

2.1.3. Limpieza de los datos

Luego de efectuar el análisis previo sobre los dos set de datos, se puede observar cómo se redujo considerablemente el espacio utilizado. Se pasó de trabajar con dos DataFrame, donde uno de ellos tenía un peso de 71.3 Mb, a trabajar con un único DataFrame limpio con un peso de 11.4 Mb. Queda al descubierto la importancia de realizar una limpieza de los datos.

Tras las modificaciones mencionadas, se puede observar que el DataFrame cuenta con 40 columnas y 260601 filas, sumando un total de 10424040 celdas

2.2. Análisis de las variables

Vamos a realizar un análisis exploratorio de nuestro datos con el objetivo de determinar características y variables importantes, descubrir conclusiones interesantes, y analizar la estructura de los mismos.

2.2.1. Grado de Daño (Damage Grade)

Dentro de nuestro set de datos, contamos con la información de 260601 edificios afectados por el sismo Gorkha, de 7.8 grados de magnitud en la escala Richter. Una de las variables proporcionadas, es el grado de daño que sufrió cada edificio luego de la tragedia. Basándose en esta variable de interés, se investigarán las características de cada construcción para tratar de encontrar alguna correlación entre las propiedades de los edificios y su daño sufrido a causa del terremoto.

En primer lugar, se calculó la proporción del grado de daño del total de los 26061 edificios afectados, obteniendo la siguiente tabla:

Grado de Daño	Total de edificios afectados	Porcentaje del total (%)
Low Damage	25124	9.6
Medium Damage	148259	56.9
High Damage	87218	33.5

Cuadro 2: Tipo de dato entero suficiente para almacenar la columna

Para obtener una mejor visualización de la información, se realizó el siguiente gráfico de tortas, también conocido como pie chart .

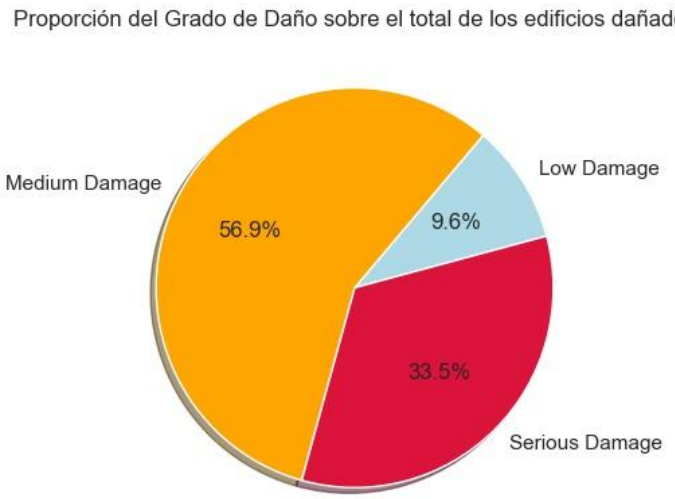

Figura 1: Proporción del Grado de Daño sobre el total de los edificios

Como puede verse en la figura anterior, la mayoría de los edificios sufrió un daño moderado, otro gran porcentaje sufrió un daño alto, mientras que tan solo una pequeña minoría sufrió un daño bajo. En las secciones posteriores vamos a investigar si existen características comunes entre los edificios que sufrieron un mismo grado de daño.

2.2.2. Hipótesis

Una vez analizada la variable de interés e investigado un poco sobre la información que nos provee nuestro Data Frame, vamos a proceder a generarnos preguntas e interrogantes acerca de estos datos. Se investigará si existen características propias de los edificios que les hagan tener un mayor grado de sensibilidad a sufrir daño, con el objetivo de aprender de esta información y en un futuro poder disminuir el grado de destrucción producido por los desastres naturales.

Se indagará en cómo afecta al grado de daño que sufrió el edificio sus materiales de construcción, su antigüedad, su ubicación geográfica, el uso que se le daba al mismo, entre otras características.

2.2.3. Correlación de los datos

La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente. Es una herramienta común para describir relaciones simples sin hacer afirmaciones sobre causa y efecto.

Ahora, se pasará a estudiar el grado de correlación que existe entre las distintas columnas de nuestro set de datos con respecto a la columna Damage Grade. Esto nos permitirá tener un mejor conocimiento de nuestros datos, y nos ayudará a profundizar más en nuestro análisis. Haciendo uso de las herramientas brindadas por las librerías pandas y matplotlib, se obtuvo el siguiente gráfico.

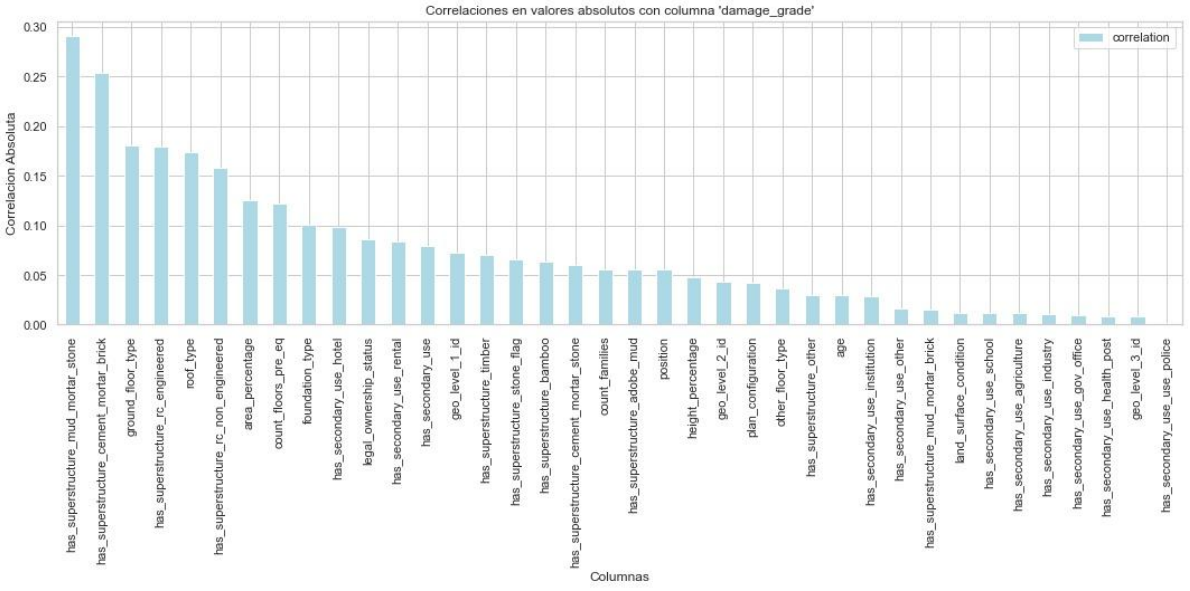

Figura 2: Correlación en valores absolutos de las columnas con Damage Grade

Al observar la figura 2, queda al descubierto cuales son las columnas que guardan más relación lineal con el grado de daño, esto quiere decir que cuando en la segunda parte del trabajo práctico tengamos que hacer predicciones van a tener mayor influencia en los resultados.

Para el caso en que se correlacionan una columna categórica, se definió un nuevo tipo de categoría numérica. Por ejemplo, si la columna tenía categorías u, v y w entonces se transformaron las categorías a 1, 2 y 3. Al no conocerse las categorías, la asignación numérica es **totalmente aleatoria** por lo que únicamente se puede inferir que a un mayor grado de correlación, una o más categorías se pueden relacionar con un grado de daño de edificios.

Las columnas que poseen un grado de correlación mayor a 0.1 son:

colum_name	correlation
has_superstructure_mud_mortar_stone	0.29
has_superstructure_cement_mortar_brick	0.25
ground_floor_type	0.18
has_superstructure_rc_engineered	0.18
roof_type	0.17
has_superstructure_rc_non_engineered	0.16
area_percentage	0.13
count_floors_pre_eq	0.12
foundation_type	0.10
has_secondary_use_hotel	0.10

Cuadro 3: Correlación de las columnas con el grado de daño

Tomando como referencia las 5 correlaciones principales se puede observar una clara relación entre la selección de materiales, el grado de ingeniería, el tipo de suelo y el tipo de suelo seleccionados en relación con la gravedad del estado del edificio afectado por un terremoto.

Por otro lado, relacionando únicamente la variables categóricas **roof_type**, **ground_floor_type**, y **foundation_type** a partir de una simple suma de sus valores aritméticos asociados, se puede observar como existe una correlación inversa entre el puntaje obtenido y el grado de daño. Esto quiere decir que para las categorías con la asignación de valores numéricos más altos, la gravedad disminuye.

2.2.4. Daño por ubicación geográfica

Dentro de nuestro Data Frame, contamos con información sobre la localización geográfica de cada edificio afectado. Esta información está separada en tres columnas

diferentes: “geo_level_1_id”, “geo_level_2_id” y “geo_level_3_id”, que van desde la ubicación más general a la más específica respectivamente. En primera instancia, se buscó graficar en un mapa la distribución de las zonas más afectadas por el sismo, pero desgraciadamente no fue posible realizar un mapeo de esta información con ubicación reales de Nepal.

Para llevar a cabo el análisis, se dividieron a los edificios afectados en 30 regiones, definidas por su código “geo_level_1_id”, y se dejó de lado las otras dos columnas. Se tomaron las 10 regiones con más cantidad de edificios dañados, y se calculó la proporción de grado de daño de cada una de estas. Se obtuvo la siguiente tabla.

Región	Cantidad de edificios con Low Damage	Cantidad de edificios con Medium Damage	Cantidad de edificios con High Damage	Cantidad total de edificios dañados
4	521	11164	2883	14568
6	2108	16222	6051	24381
7	1033	11273	6688	18994
8	654	8513	9913	19080
10	1211	12107	8761	22079
17	285	3913	17615	21813
20	3311	11860	2045	17216
21	322	5857	8710	14889
26	8028	12645	1942	22615
27	465	6007	6060	12532

Cuadro 4: Top 10 edificios más dañados

Mediante la gráfica de un stacked bar plot vamos a ver cuales fueron las 10 regiones con mayor cantidad de edificios dañados y la proporción de grado de daño tuvo cada una.

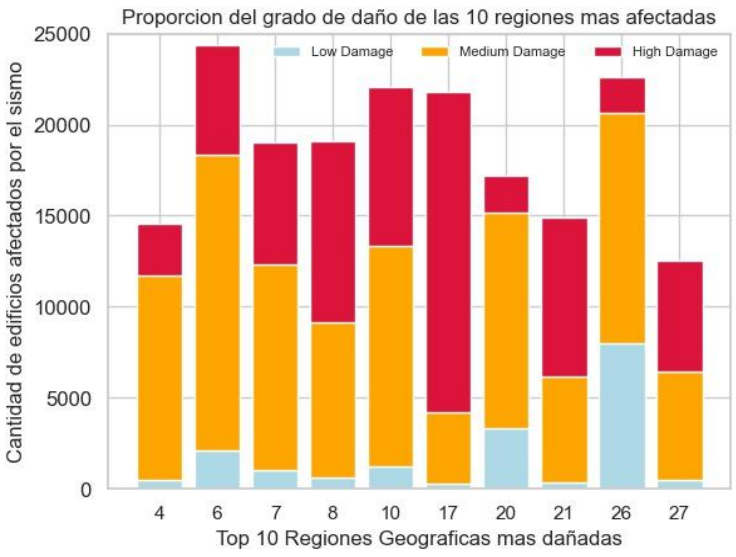

Figura 3: Top 10 regiones más afectadas por el sismo

Del gráfico podemos destacar 3 regiones:

- ➢ Región 6: es la región con mayor cantidad de edificios dañados, y a su vez la que mayor cantidad de edificios con daño moderado.
- ➢ Región 17: es la región con mayor cantidad de edificios don daño crítico y la que menor cantidad con daño bajo presenta.
- ➢ Región 26: si bien es la segunda región con mayor cantidad de edificios dañados, presenta una considerable cantidad de edificios con daño bajo y también, un bajo grado de edificios con daño crítico .

2.2.5. Daño por antigüedad del edificio

En nuestro set de datos, también contamos con información acerca de la edad de los edificios, por lo tanto vamos a analizar cómo afecta la antigüedad del edificio al grado de daño que sufrió. Naturalmente, uno tendería a pensar que cuanto más antiguo es un edificio, mayor probabilidad hay de que este se derrumbe, debido al deterioro ocasionado por el paso del tiempo. Procederemos a investigar la información disponible, para determinar si la afirmación anterior es válida. En primer lugar, realizaremos un gráfico de densidad, también llamado Density Plot, de la columna age de nuestro set de datos, esta visualización es muy útil para analizar la distribución de una variable.

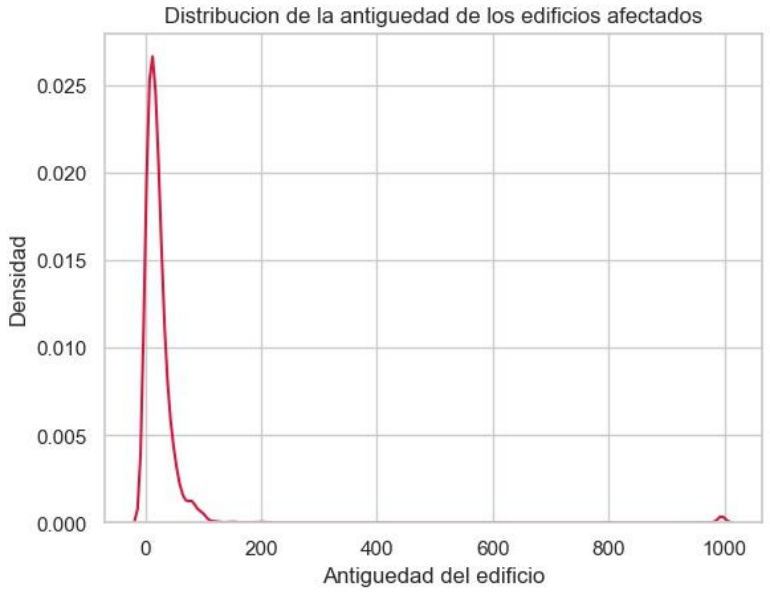

Figura 4: Distribución de la antigüedad de los edificios afectados

Echando un vistazo a la gráfica superior, podemos destacar dos resultados:

- ➢ a. La gran mayoría de los edificios dañados tiene una edad entre los 0 y 100 años aproximadamente.
- ➢ b. La cantidad de edificios afectados con edades entre los 200 y 900 años es prácticamente despreciable o nula.

Echando un primer vistazo a la Figura 4, uno estaría tentado a decir que dado que hay mayor proporción de edificios jóvenes entre los afectados, los edificios antiguos son más resistentes, pero esta sería una conclusión errada. De la gráfica sólo podemos afirmar, que contamos con mucha más información sobre edificios modernos, una de las posibles razones sería que en Nepal existen mucha mayor cantidad de edificios modernos, y muy poca cantidad de edificios antiguos.

Ahora pasemos a comparar la distribución de los distintos grados de daño en función de su antigüedad. Para obtener una gráfica más nítida, se restringirá el eje de la antigüedad para edades entre 0 y 100, dado que la mayor proporción de los edificios se encuentra comprendido en este rango etario.

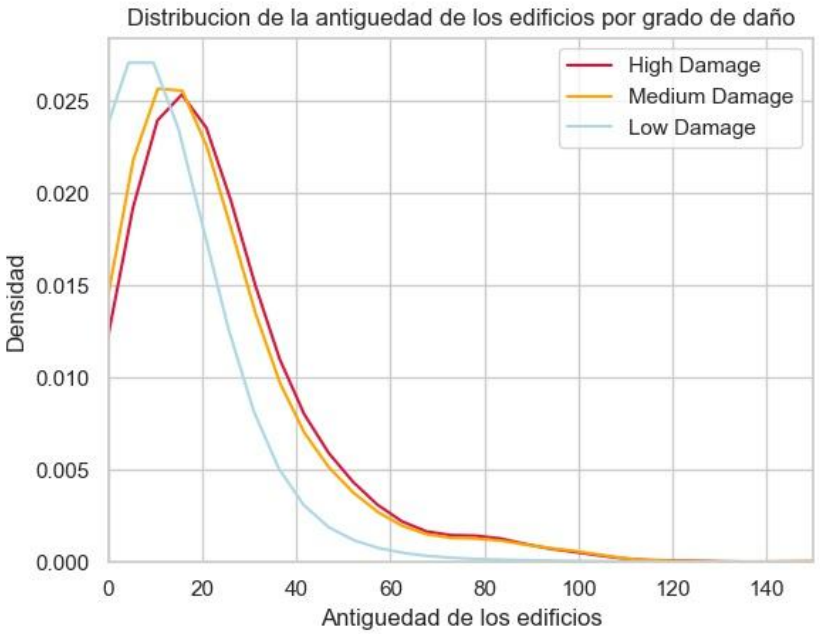

Figura 5: Distribución de la antigüedad de los edificios afectados por grado de daño

De la figura 5, podemos ver que los edificios que sufrieron un daño leve, suelen ser más jóvenes que los edificios que sufrieron daños graves o medios. También se aprecia como los picos de las funciones se encuentran desplazados a lo largo del eje x, para la curva celeste, se alcanza un pico en una antigüedad menor que para la curva naranja, la cual a su vez, alcanza un pico a menor x que la curva roja.

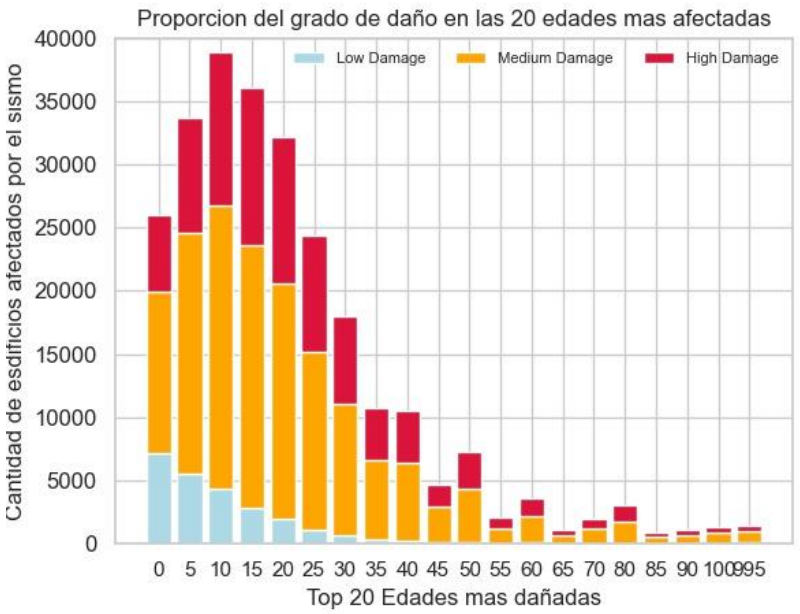

Figura 6: Las 20 edades de edificios más afectadas

Por último, vemos cuáles fueron las edades de los edificios que sufrieron más daños. Vamos a destacar, que para nuestra sorpresa, los edificios con 995 años de antigüedad tienen mayor cantidad de edificios con daño medio que edificios con daño crítico, esto puede deberse a que los edificios antiguos, a diferencia de los modernos, suelen considerarse lugares históricos, y son una gran fuente de turismo, por lo que son refaccionados y muy bien preservados, lo que podría traducirse en un menor daño recibido a la hora de sufrir un desastre natural.

Vamos a finalizar esta sección afirmando que con la información de la antigüedad de los edificios con la que contamos en nuestro Data Frame, es difícil obtener conclusiones concisas, dado que los datos se encuentran muy dispersos, se saltan de edificios de 225 años hasta edificios 995 años, lo que genera una brecha gigante entre los datos, y además, contamos con mayor cantidad de información acerca de los edificios modernos.

2.2.6. Daño por cantidad de pisos del edificio

En esta sección, vamos a investigar si existe alguna relación entre la cantidad de pisos de los edificios y el grado de daño que sufrieron. Primero realizaremos un histograma, para tener una noción de cómo se distribuye nuestra variable cantidad de pisos.

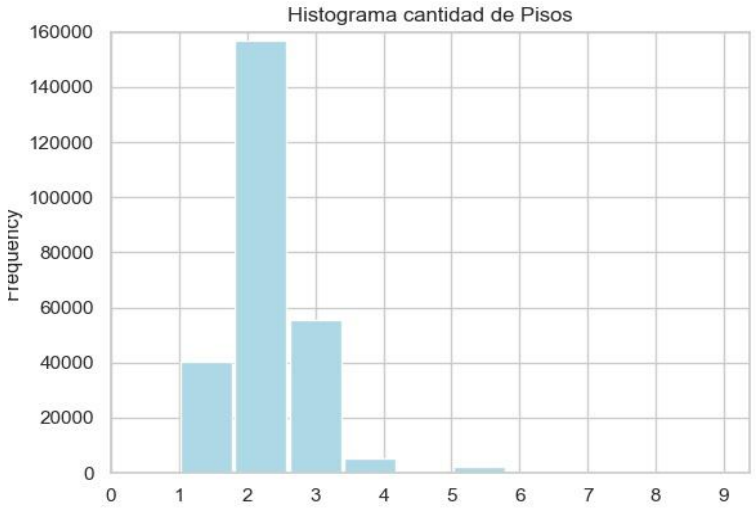

Figura 7: Las 20 edades de edificios más afectadas

Como podemos observar de la figura anterior, la mayoría de los edificios afectados oscila entre uno y tres pisos, por lo que vimos en secciones anteriores, esto no quiere decir que si un edificio tiene esta cantidad de pisos, es más probable que sufra daños, sino que en Nepal las mayoría de las construcciones tienen esta cantidad de pisos, y hay una poca proporción de edificios de gran envergadura. Dado que la gran mayoría de los edificios que estudiamos tienen entre uno y dos pisos, en las próximas gráficas nos centraremos en esta escala para obtener una visualización más nítida.

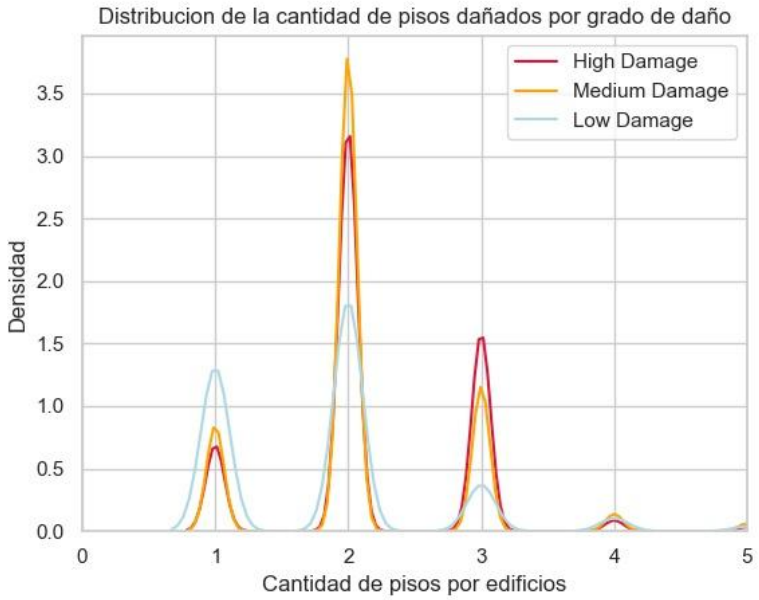

Figura 8: Distribución de la cantidad de pisos dañados por grado de daño

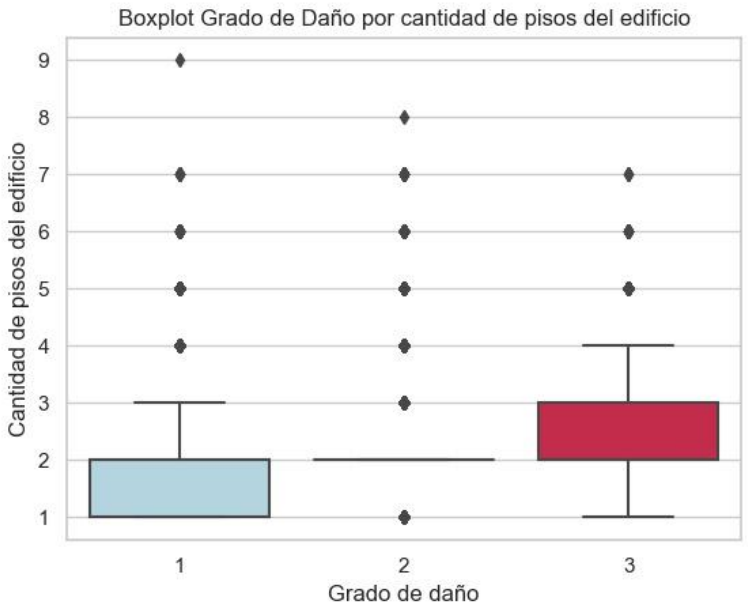

Figura 9: Box plot de la cantidad de pisos de los edificios afectados (siendo 1: Low Damage, 2 : Medium Damage y 3: High Damage)

Ambas visualizaciones, son de gran utilidad para comparar grupos categóricos de información. De la figura 7, vemos que la mayor proporción de edificios con daño crítico se distribuye en edificios 2 o 3 pisos, los que sufrieron un daño medio en su mayoría son edificios de 2 pisos, y la distribución de los edificios con daño bajo se concentran en construcciones de un solo piso. Por lo tanto, pareciera ser que los edificios bajos son más propensos a sufrir un daño leve, mientras que los edificios altos tienden a sufrir un mayor impacto.

2.2.7. Daño por altura del edificio

En esta nueva sección, se buscará si existe alguna relación entre la altura de los edificios y el grado de daño que sufrieron. Primero realizaremos un density plot, para tener una noción de cómo se distribuye nuestra variable altura del edificio.

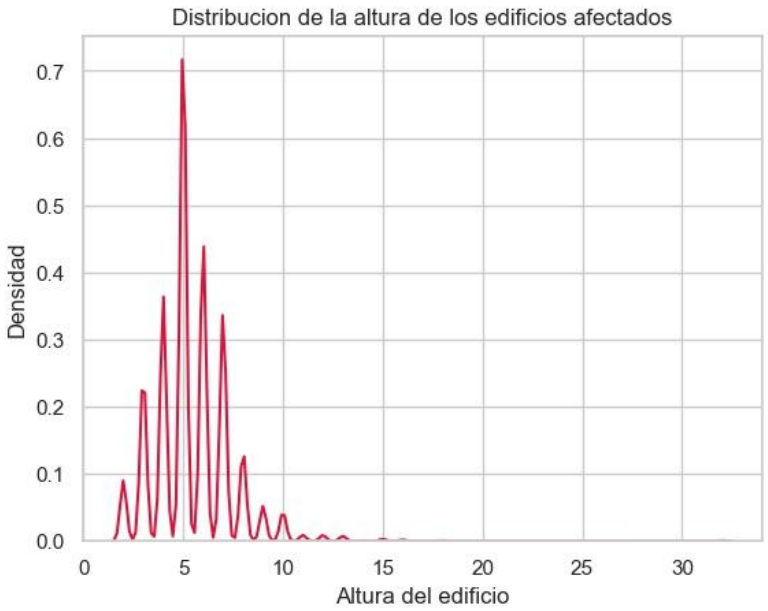

Figura 10: Distribución de la altura de los edificios afectados

Como vemos en la gráfica, la gran mayoría de los edificios en Nepal no supera los 10 metros de altura, y tiene una gran proporción de construcciones que oscilan entre los 3 y 7 metros. Ahora compararemos la distribución de los diferentes grados de daño en función de la altura.

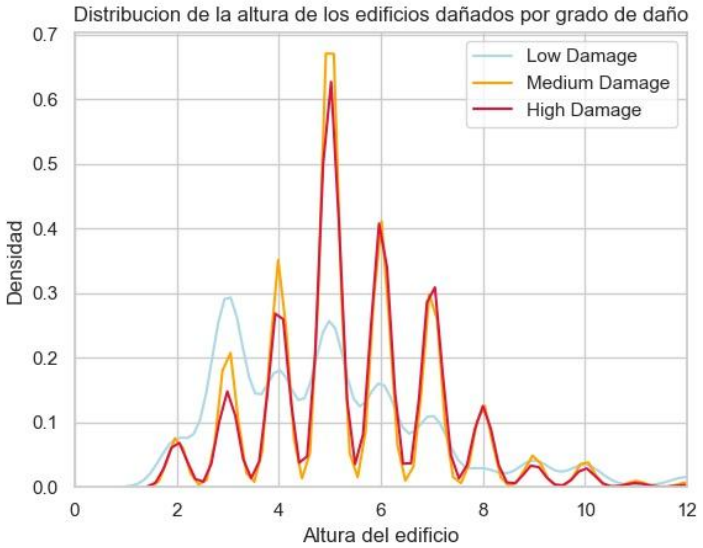

Figura 11: Distribución de la altura de los edificios por grado de daño

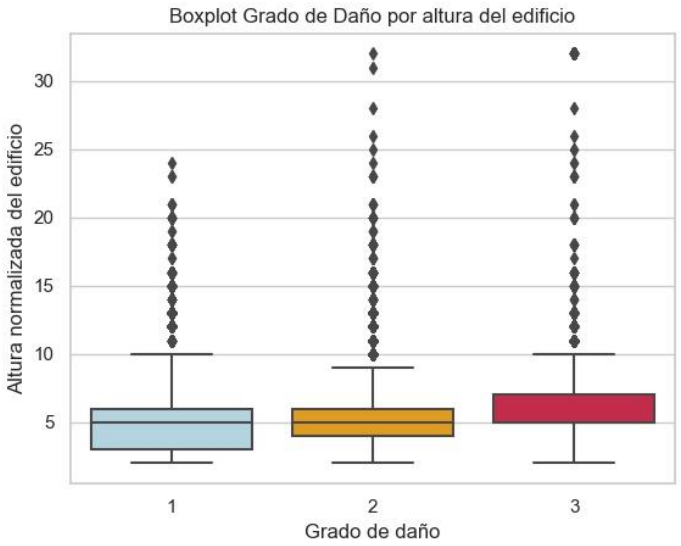

Figura 12: Box plot de la altura de los edificios afectados (siendo 1: Low Damage, 2 : Medium Damage y 3: High Damage)

De las dos gráficas posteriores, podemos observar la tendencia de los edificios de pequeña altura a sufrir un grado menor de daño, mientras que los edificios mayores a 4 metros suelen sufrir un grado medio o crítico.

2.2.8. Daño por área de pisos del edificio

Ahora pasaremos analizar si existe alguna relación entre el area de los edificios y el grado de daño que sufrieron. Primero, como hicimos en las secciones previas, realizaremos un density plot, para tener una noción de cómo se distribuye nuestra variable área del edificio.

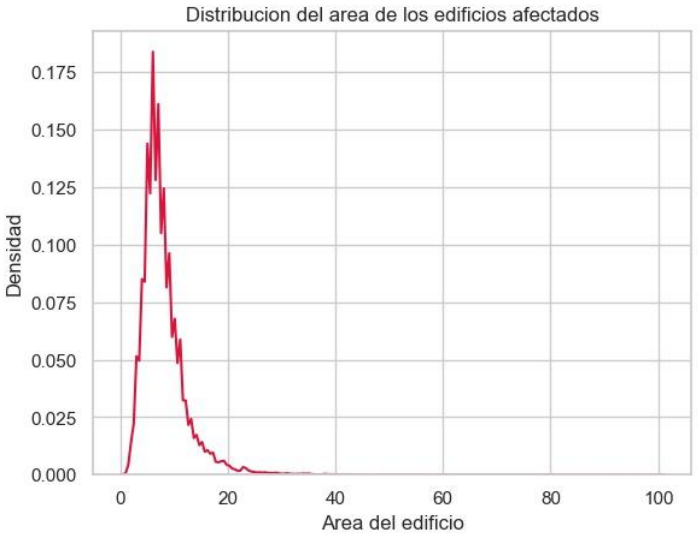

Figura 13:Densidad del área de los edificios afectados

Como se aprecia en la gráfica, en Nepal, la mayoría de los edificios tiene un area comprendida entre los 5 y 15 metros cuadrados, y existen pocas construcciones que

superen los 20 metros cuadrados.Ahora compararemos la distribución de los diferentes grados de daño en función del área de los edificios.

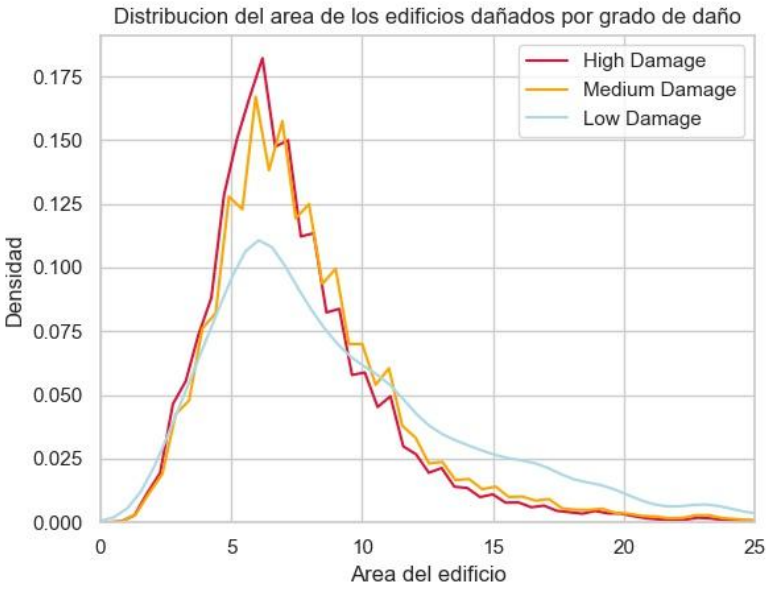

Figura 14: Distribución del área de los edificios afectados

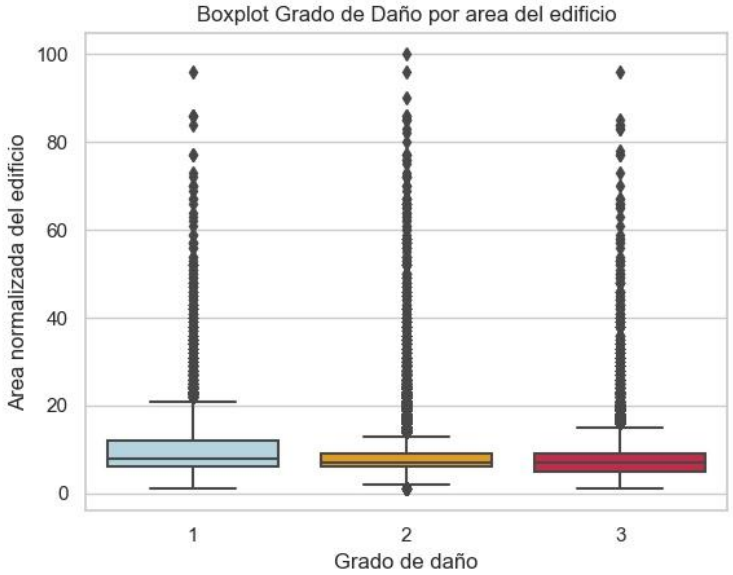

Figura 15: Box plot del área de los edificios afectados (siendo 1: Low Damage, 2 : Medium Damage y 3: High Damage)

De las gráficas de densidad superior, vemos que los tres grados posibles de daños comparten bastante área en común, dado que los edificios de Nepal no suelen variar mucho en su área. Es interesante observar, en la figura 14 , que el área comprendida por la curva de la distribución de los edificios de grado de daño leve, para áreas mayores es superior a la de las curvas de daño medio y alto.

2.2.9. Daño según condición de superficie

Para entender el análisis de la columna “land_surface_condition” hay que tener en cuenta que esta variable es de tipo “categórico”, eso quiere decir que solo puede tomar un valor dentro de los conocidos.
Esta columna analiza la condición de la superficie terrestre donde el edificio fue construido. En este caso las distintas variables posibles son n, o y t. Al no saber qué significa cada valor no se puede analizar y sacar muchas conclusiones, sin embarazo podes llegar a resultados interesantes, que se procederá a explicar en esta sección.

Lo primero que vamos a analizar es la distribución de esta columna, es decir, cual es el porcentaje de cada tipo de superficie sobre el total.

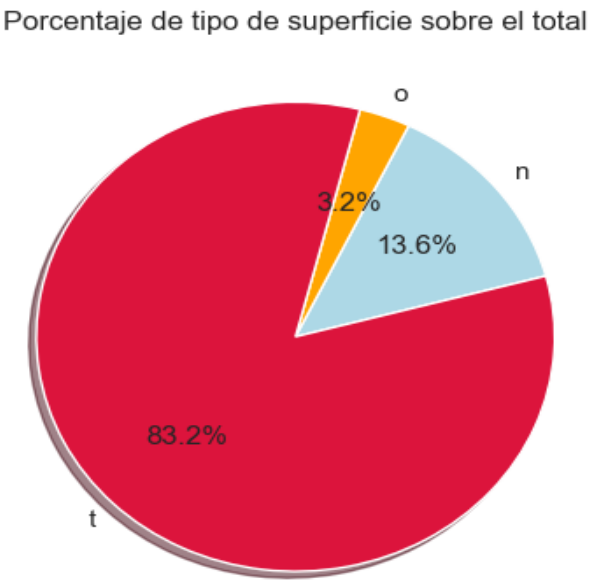

Figura 16: Porcentaje de tipo de superficie sobre el total

Como se puede observar en la figura 16, la mayoría de los edificios se encuentran en el tipo de superficie terrestre “t”, un porcentaje menor en el de tipo “n” y un porcentaje mucho menor en el “o”.
Debido a esto, vamos a analizar más en profundidad los edificios cuya superficie terrestre es “t”, para ver de esta manera qué tipo de daño sufrieron estos edificios y ver si podemos sacar alguna conclusión.

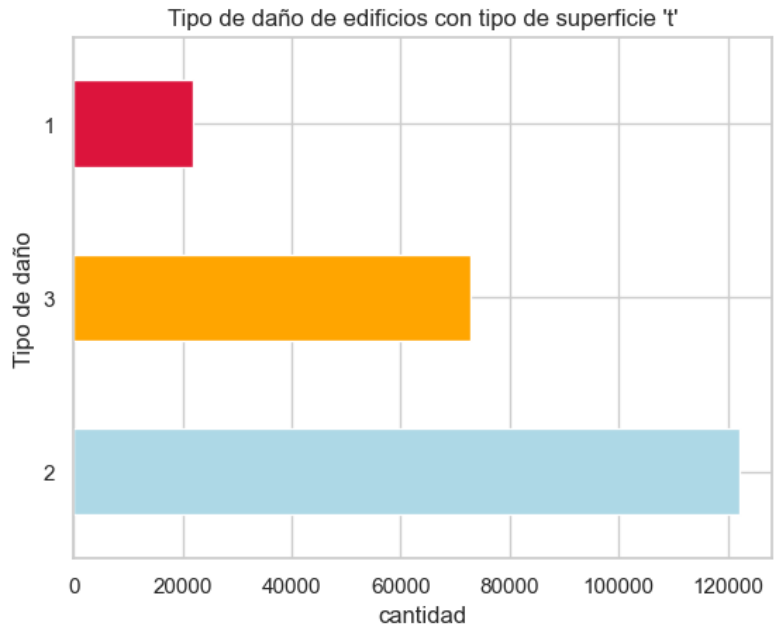

Figura 17: Tipo de daño para los edificios con superficie “t”

Como podemos ver en la figura 17, el mayor daño que sufrieron los edificios en esta superficie fue daño medio, seguido por el daño grave. Alrededor de 120.000 edificios sufrieron algunos daños y aproximadamente 70.000 edificios sufrieron daños graves. Por otro lado, solo 20.000 edificios (un 10% del total) sufrieron daños leves.

Con esta información podemos formar una hipótesis y sospechar que la superficie “t” no resulta ser muy segura. Para ver la validez vamos a analizar los tipos de daños de las otras superficies y compararlos.

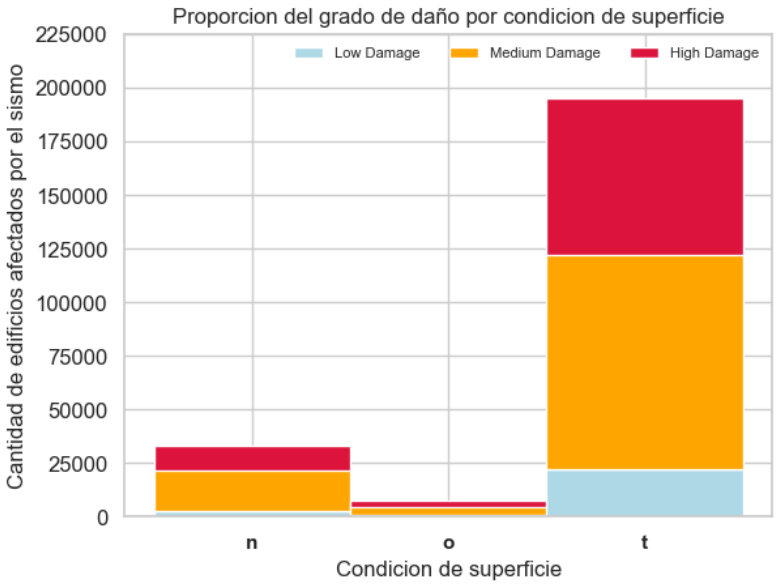

Figura 18: Cantidad de edificios por tipo de superficie terrestre y tipo de daño

Al analizar la figura 18 podemos ver que, si bien la superficie “t” es la superficie que más cantidad de edificios fueron construidos. Usando también el cuadro 2 de la sección 2.2.1 podemos ver que solamente un 9.6% de los edificios sufrieron daños mínimos, esto se puede entender como que el 90.4% de los edificios sufrieron daños medios o graves.

Al tener en cuenta esta información es bastante coherente con lo que se puede observar en la figura 17. Si por ejemplo tuviésemos un porcentaje muy grande de edificios con daño 1, se podría decir que la superficie de tipo “t” resultaba ser más segura, pero los datos obtenidos hacen que la hipótesis resulta falsa, y no podemos sacar ninguna conclusión a priori.

2.2.10. Daño según tipo de estructura

Para dar un poco más de sentido y simplificar la información que contienen las columnas “has_superstructure_…” se toma convención de que las estructuras que contengan barro de contener el valor uno en dicha posición en la nueva columna (material_de_construccion) pasen a denominarse barro, de igual modo para esta nueva columna se coloca cemento para las posiciones donde hay un uno en las columnas donde interviene cemento como material de construcción y usando el mismo proceder ya descrito la columna material_de_construccion también contiene madera, no_clasificados, otro y piedra y se obtuvo la cuadro 5 que es la siguiente:

Material de construcción	Cantidad total de edificios dañados
Barro	222380
Cemento	18043
Madera	10336
No Clasificados	6819
Otros	349
Piedra	2674

Cuadro 5: Totalidades de cada material de construcción.

Sabíamos que Nepal no es una potencia mundial, pero nos llevamos una sorpresa al ver que 85% de la propiedades son de barro, o interviene barro en su construcción. Esto deja en evidencia las carencias económicas del país y explica porque el terremoto generó un desastre de tal magnitud.

2.2.11. Daño según formato de construcción de la edificación

En esta sección analizaremos el formato de construcción de la edificación para diseño sísmico. Esta columna es categórica y los valores posibles que puede tomar son: a, c, d, f, m, n, o, q, s y u.

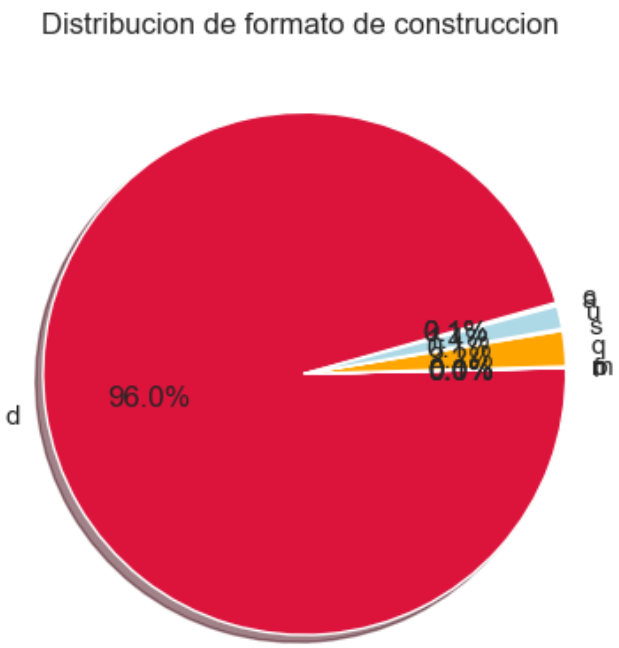

Figura 19: Porcentaje de formato de construcción sobre el total

Como se puede ver en la figura 19, un 96% de los edificios contienen el formato de construcción tipo “d”. Debido a esto vamos a centrar nuestro análisis en esta configuración, para ver que repercusiones puede tener sobre el tipo de daño que recibieron estos edificios.

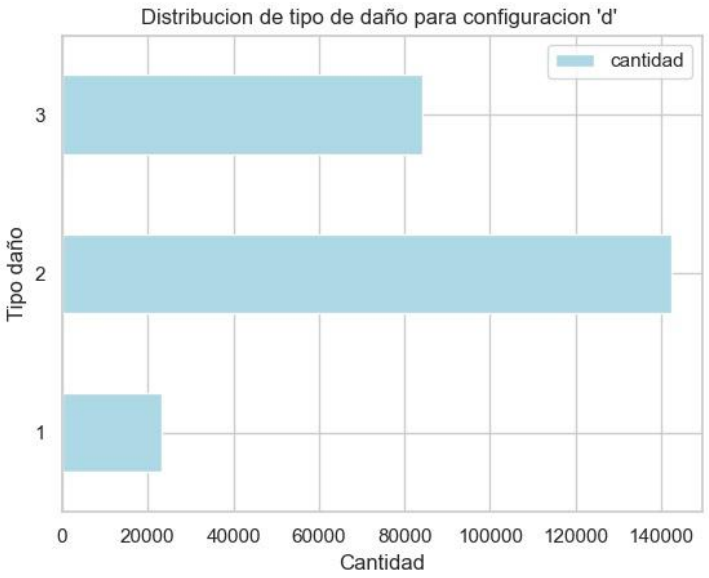

Figura 20: Tipo de daño para los edificios con configuración “d”

En esta figura podemos ver como el tipo de daño medio sigue siendo el más prominente, seguido por el daño grave y el daño leve.

Como no llegamos a alguna conclusión interesante, procederemos a analizar el tipo de superficie para esta configuración en particular.

Figura 21: Distribución del tipo de superficie para el formato de configuración D

Al observar la figura 20 vemos una particularidad, queda la misma proporción que la figura 13. Al ver esto decidimos analizar un poco más los valores para ver si había algún error de nuestra parte.

Primero analizamos la cantidad de edificios por condición de superficie, con un total de 260601 elementos, es decir todos los edificios, tenemos esta proporción.

Condición de superficie	Cantidad de edificios dañados	Porcentaje de edificios dañados
n	35528	13.6%
o	8316	3.2%
t	216757	83.2%

Cuadro 6: cantidad de edificios por condición de superficie

Recordar que en este cuadro estamos analizando sólo los edificios que tienen configuración tipo “d”, es por eso que no tenemos 260.601 de valores, ya que no todos tienen esa configuración. En este caso tenemos un total de 250072 elementos.

Condición de superficie	Cantidad de edificios dañados	Porcentaje de edificios dañados
n	34071	13.6%
o	7962	3.2%
t	208039	83.2%

Cuadro 7: cantidad de edificios con configuración “d” por condición de superficie

Extrañamente llegamos al mismo porcentaje que el cuadro 6, pero analizando la configuración ”d” en específico. Este resultado es peculiar, ya que si bien los valores se van a encontrar en el mismo rango, debido a que es la distribución de la condición de la superficie, hace dudar de si es una coincidencia o si tal vez los datos fueron generados de alguna manera arbitraria, dejando estas coincidencias a lo largo del dataset.

2.2.12 Daño según tipo de construcción usado en la planta baja

Procederemos a analizar otra columna del tipo categórico, el “ground_floor_type”, el cual indica el tipo de construcción usado en la planta baja cuando se construyó la edificación. Dentro de los valores posibles se encuentran: f, m, v, x, z.

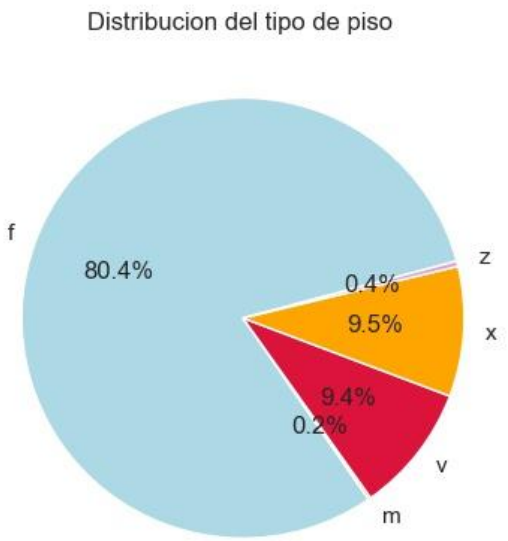

Fig 22: distribución del tipo de piso

Como podemos observar en el gráfico, el tipo de piso “f” es el de mayor proporción y es el que tomaremos para ver si podemos sacar alguna conclusión interesante.

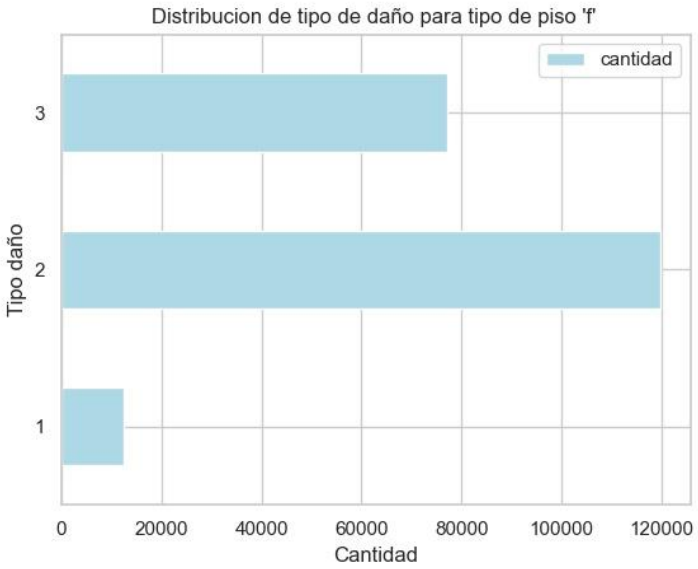

Figura 23: Tipo de daño por tipo de piso

Nuevamente vemos el mismo patrón, el tipo categórico más importante de cierta columna suele ser una mayoría muy grande que al analizar comparte casi todas las características con los otros categóricos. Más abajo realizamos un comentario sobre esto más en específico.

2.2.13 Daño según tipo de techo

Analizaremos en esta sección la columna ”roof_type”, la cual indica el tipo de techo usado cuando se construyó la edificación. Sus valores posibles son: n, q, x.

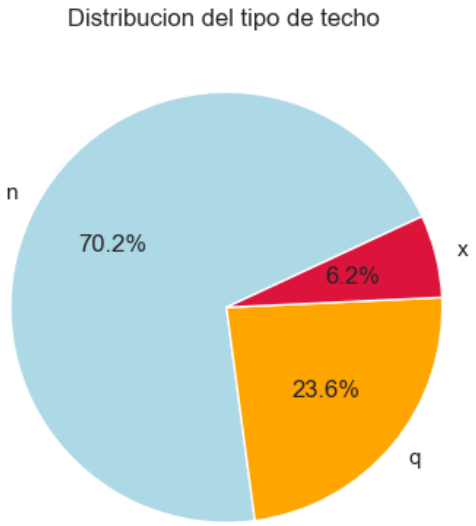

Figura 24: Distribución del tipo de techo

Como podemos ver en el gráfico, el tipo de techo “n” es el que la mayoría de los edificios poseen. Debido a esto vamos a analizar el tipo de daño que recibieron estos edificios y ver si podemos sacar algo interesante.

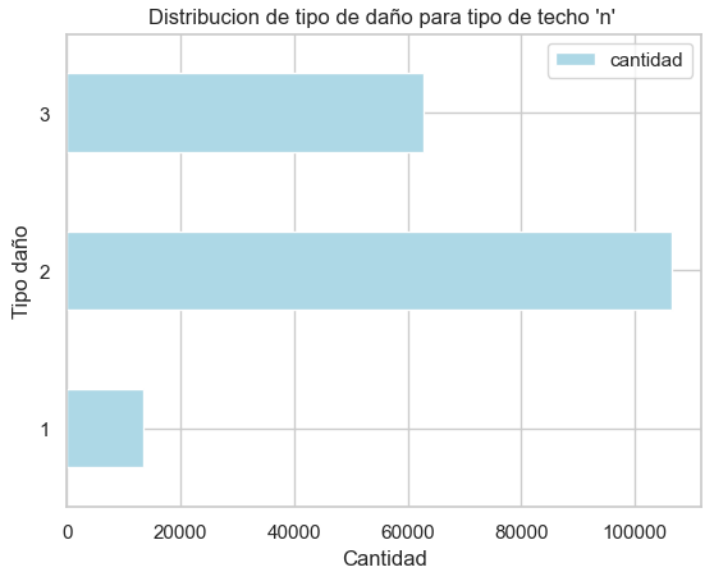

Figura 25: Distribución de tipo de daño para tipo de techo 'n'

Como es recurrente en el análisis de las columnas categóricas, se mantiene la misma distribución de siempre. El tipo de daño mediano con un porcentaje de alrededor del 58%, el daño grave con un porcentaje del 34% y el resto en daños leves.

Intentamos cruzar conclusiones entre los datos categóricos pero terminaron siendo inconclusos.

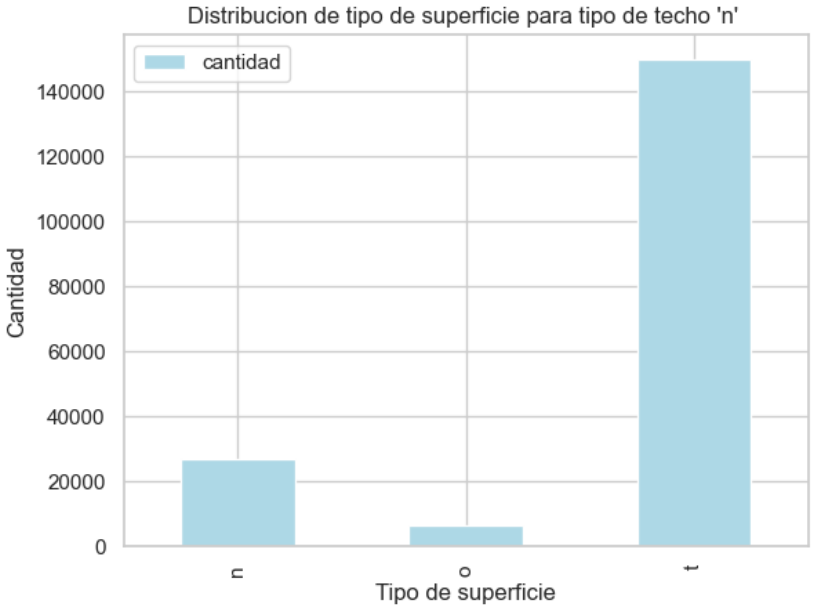

Figura 26: Distribución de tipo de superficie para tipo de techo 'n'

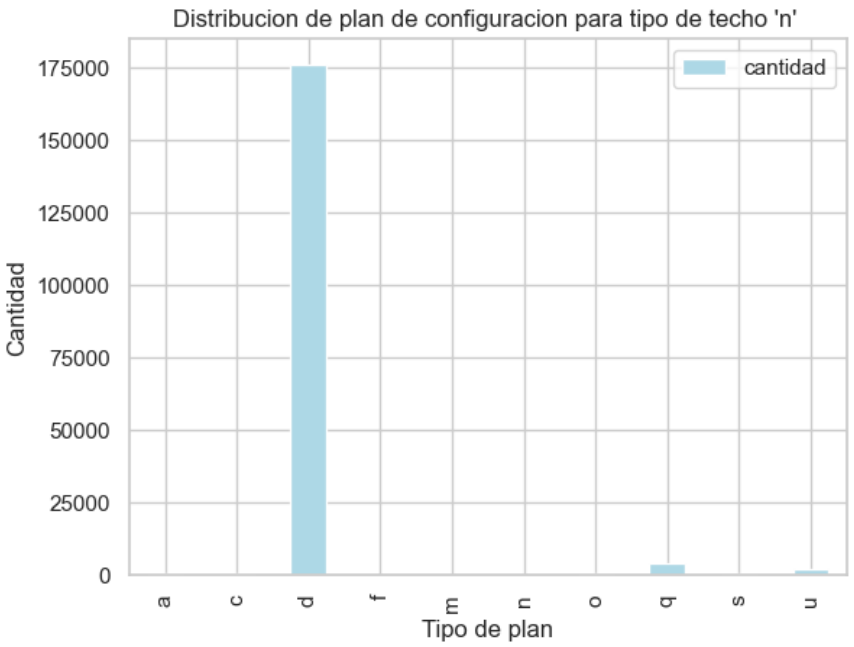

Figura 27: Distribución de plan de configuración para tipo de techo 'n'

Como se puede ver en la comparación entre distintas columnas categóricas siempre se llega a la misma conclusión, la gran mayoría comparte los mismos tipos de datos categóricos y tienen una distribución del tipo de daño casi igual.

Dando por resultado inconcluso el análisis de estas columnas.

2.2.14 Daño según tipo de material de construcción y antigüedad de la propiedad

Si clasificamos las propiedades bajo el criterio de nuevas si tienen de 0 a 18 años de antigüedad, semi-antigua si tiene de 19 a 60 años y antiguas de 61 en adelante (hasta 995).

Antigüedad	Cantidad
Nuevos	134644
Semi-antiguos	113458
Antiguos	12499

Cuadro 8: Antigüedades y su cantidad.

Ahora analizaremos cómo varía el grado de daño en función de la antigüedad y el material de construcción.

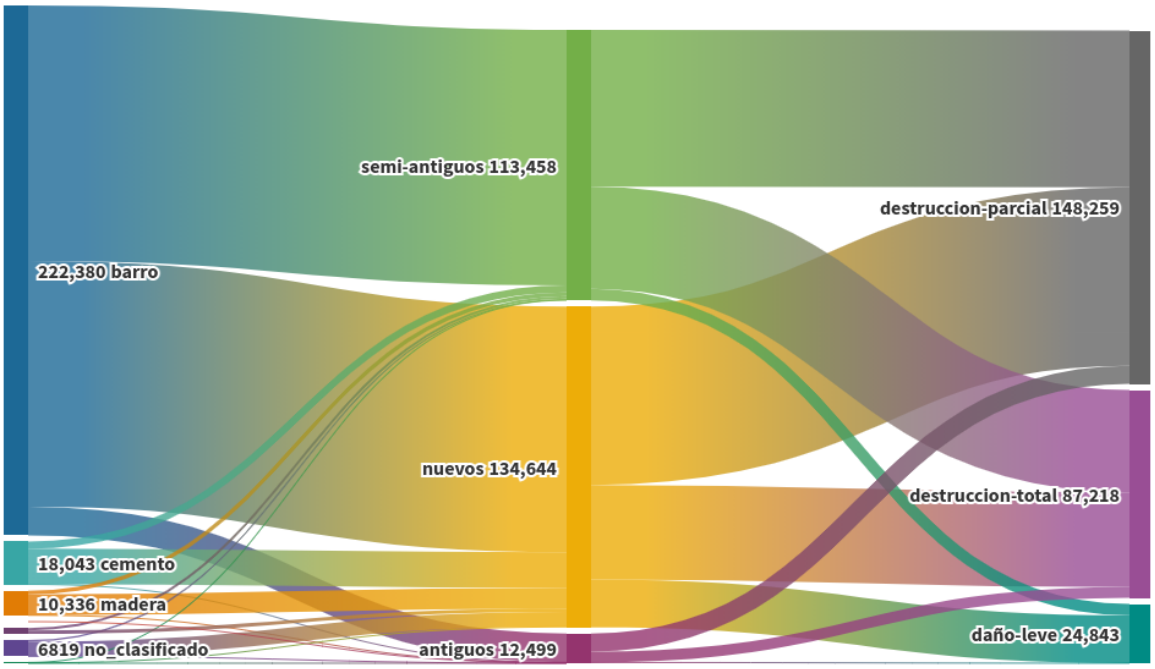

Figura 28: Relación de destrucción según su antigüedad y material de construcción

Nepal tiene muchísima más propiedades nuevas que antiguas, pero estas construcciones tienen en común que utilizan barro como material de construcción, esto definitivamente es una mala señal ya que las propiedades que utilizan barro son las más dañadas, como se aprecia claramente en la Figura 28. .

3. Conclusiones

En conclusión, se cree que con el presente informe se ha logrado desarrollar interesantes aspectos de la base de datos con la que se trabajó. Tras analizar su estructura, relaciones, composición y propiedades, fue posible familiarizarse con la información con la que estábamos teniendo contacto, permitiéndonos adentrarnos en ella.

Aunque no se han analizado todas las columnas, se considera que se hizo un buen trabajo seleccionando aquellas que eran más destacables a partir de su relevancia relacionando el grado de daño de las estructuras edilicias y los distintos factores analizados sobre dichas estructuras. Como continuación del trabajo realizado, se podría terminar de estudiar todas las columnas, pero más importante aún, se podría encontrar aún más vínculos como los hallados en múltiples secciones del informe.

Consideramos que el trabajo aquí realizado, puede ser de mucha utilidad para predecir y prevenir el daño que pueden causar futuras catástrofes, con el objetivo de disminuir el grado de devastación que estas pudieran generar.

Al analizar las distintas columnas se puede observar una clara relación entre la selección de materiales, el grado de ingeniería, el tipo de suelo y el tipo de techo comparados con la gravedad del estado del edificio afectado por un terremoto.

Por último, consideramos que hubiera sido de gran utilidad contar con los valores reales de las columnas categóricas para llegar a una conclusión más interesante y aprender aún más acerca de las características arquitectónicas de Nepal. Para las columnas con tipo de dato categóricos, encontramos con que siempre tenemos cierto valor que es el que la mayoría de los edificios contienen (como es el “t” para la condición de superficie, o la “d” para el formato de construcción de la edificación). Si bien podemos establecer que ciertos valores categóricos tienen una mayor o menor relevancia a la hora de inferir el grado de daño, la realidad es que no se pueden obtener conclusiones certeras ya que la falta de referencias entorpece el análisis de los datos.

1. Código Fuente

El código fuente se encuentra disponible en:
https://github.com/DatosOrga2021/orga_datos_1c_2021

2. Librerías Externas

2.1. Pandas

Pandas es una librería para análisis y manipulación de datos. Fue la principal herramienta utilizada para este proyecto, siendo la base de todos los análisis aquí presentados.

2.2. Numpy

Numpy es una librería para análisis numérico muy conocida. Dado su practicalidad estuvo involucrada de fondo en muchas operaciones que realizamos, así como también forma parte de otras de las librerías usadas.

2.3. Matplotlib

Matplotlib es una librería para creación de visualizaciones estáticas y animadas. Fue la principal herramienta utilizada para crear gráficos, partiendo de nuestro DataFrame que era modificado para llegar a información que le cargaríamos a los gráficos.

2.4. SeaBorn

Seaborn es una librería para visualización de datos estadísticos basada en Matplotlib. Fue utilizada para la creación de algunas visualizaciones más complejas, como los HeatMaps.

Referencias

[1] Federal Democratic Republic of Nepal https://en.wikipedia.org/wiki/Nepal

[2] Hunter, J. (2007). Matplotlib: A 2D graphics environment. Matplotlib: Python plotting. Re- trieved from https://app.flourish.studio/visualisation/4347256/edit.

[3] NumPy. Numpy.org. (2019). Retrieved from https://numpy.org/.

[4] The pandas development team. (2020). pandas - Python Data Analysis Library. Pan- das.pydata.org. Retrieved from https://pandas.pydata.org/.

[5] Florecer https://app.flourish.studio/visualisation/4347256/edit?

[6] Terremoto de Nepal de abril del 2015 https://es.wikipedia.org/wiki/Terremoto_de_Nepal_de_abril_de_2015