Trabajo Práctico N°1: Análisis de datos

(Github: https://github.com/alansuno/organizaciondedatos**)**

Integrantes: López Condori, Gabriel; Hernández, Pablo; Vadell, Federico; Viera, Alan

Introducción

En el año 2015 Nepal fue afectado por el terremoto Gorkha, un sismo que registró una magnitud de 7.8 en la escala Richter y tuvo su epicentro en la ciudad de Kathmandu. Aproximadamente 600,000 estructuras en el centro y pueblos aledaños fueron dañadas o destruidas. Un análisis posterior al sismo llevado por la Comisión Nacional de Planeamiento de Nepal comunicó que la pérdida total económica ocasionada por el terremoto fue de aproximadamente $7 mil millones (USD; NPC, 2015).

El dataset para el presente TP está compuesto de encuestas realizadas por Kathmandu Living Labs y el Central Bureau of Statistics y contiene información sobre el impacto del terremoto, estado de edificaciones y estadísticas sociodemográficas.

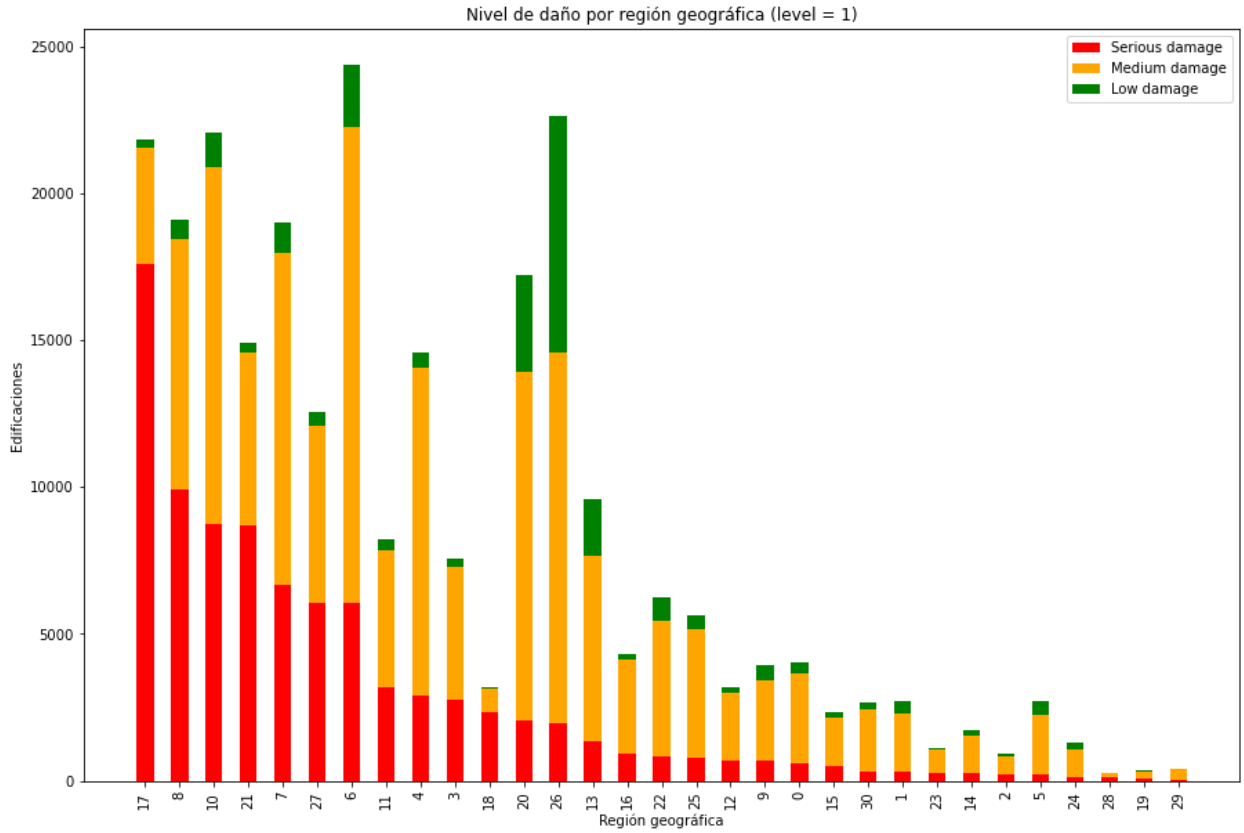
Particularmente el dataset se enfoca en cómo eran las condiciones de una determinada vivienda y cuál fue su grado de daño luego del accidente.

Análisis exploratorio de los datos

Comencemos analizando la relación entre la cantidad de edificaciones a nivel geográfico y el impacto de destrucción ocasionado. La escala para el daño está dada por:

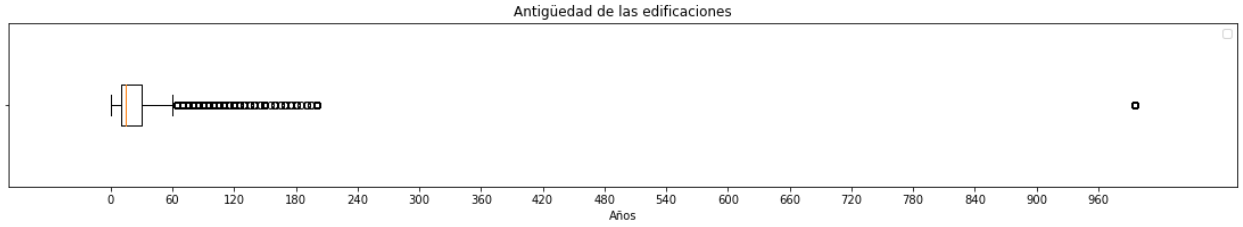
1. Low damage
2. Medium damage
3. Serious damage

Utilizamos la columna **“geo_level_1_id”** correspondiente al nivel más general obteniendo el siguiente resultado:



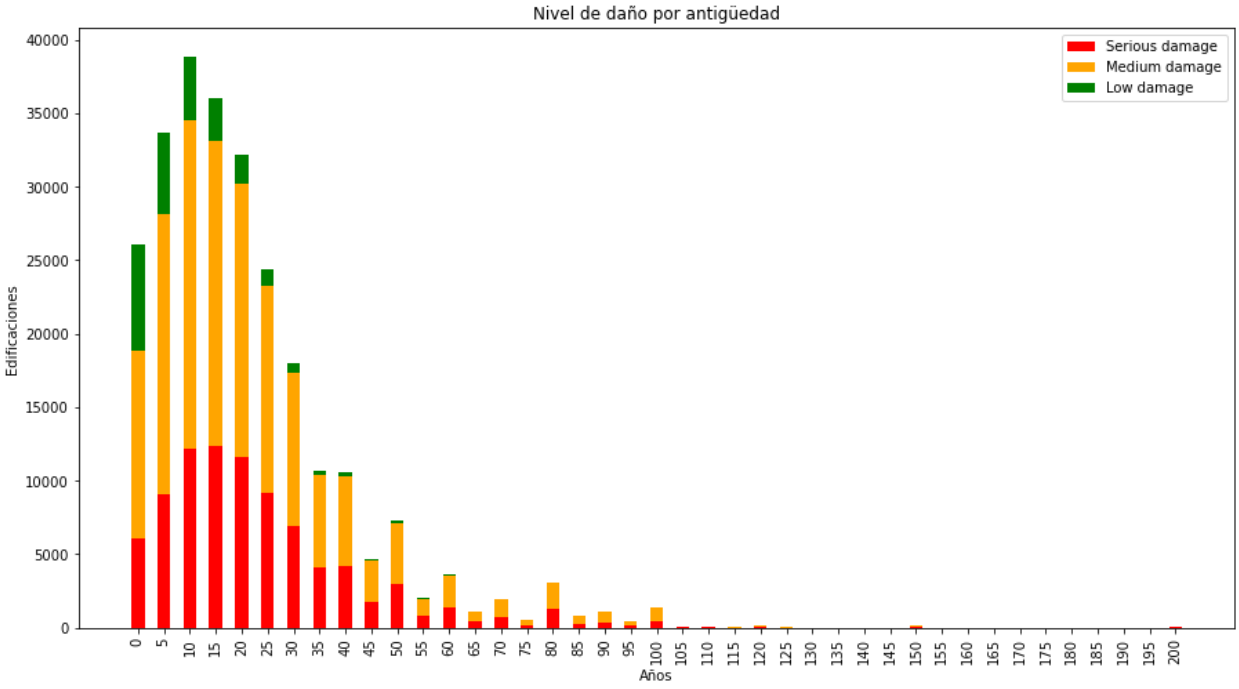
Las regiones fueron ordenadas de mayor a menor por **“Serious damage”**. La región 17 fue la más impactada. Por otro lado, el nivel de impacto no tiene una relación directamente proporcional a la cantidad de edificaciones.

Analicemos la distribución de la antigüedad de las edificaciones:



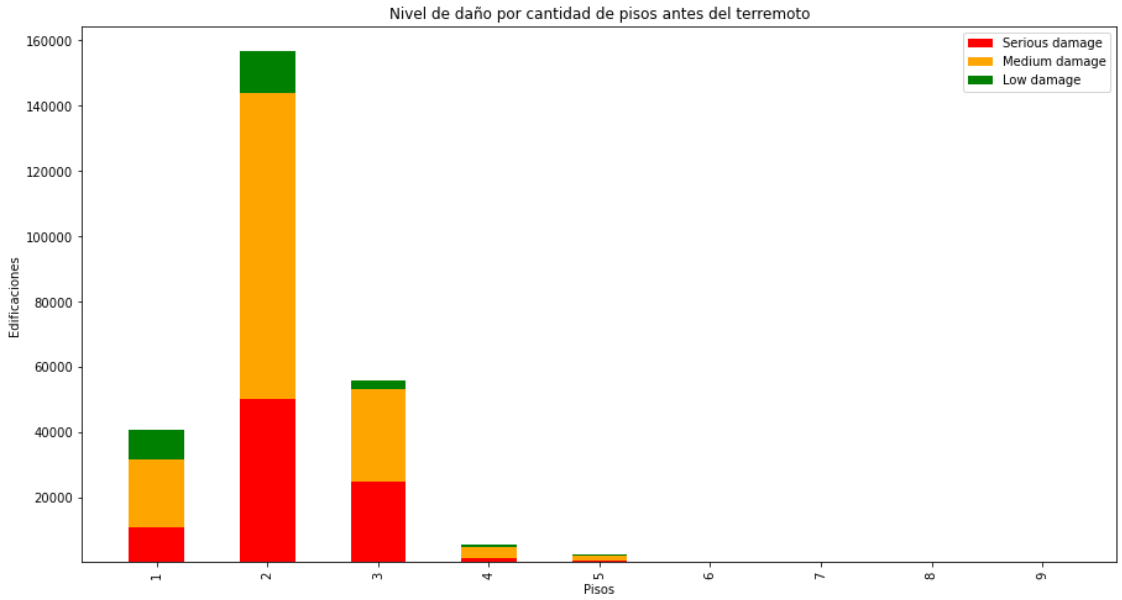
El gráfico muestra que la mayoría de las edificaciones poseen entre 0 y 60 años, encontrando algunos casos aislados entre 60 y 210, y por último casos extremos con más de 960.

Filtremos las edificaciones entre 0 y 240 años, para identificar el nivel de impacto según la antigüedad:



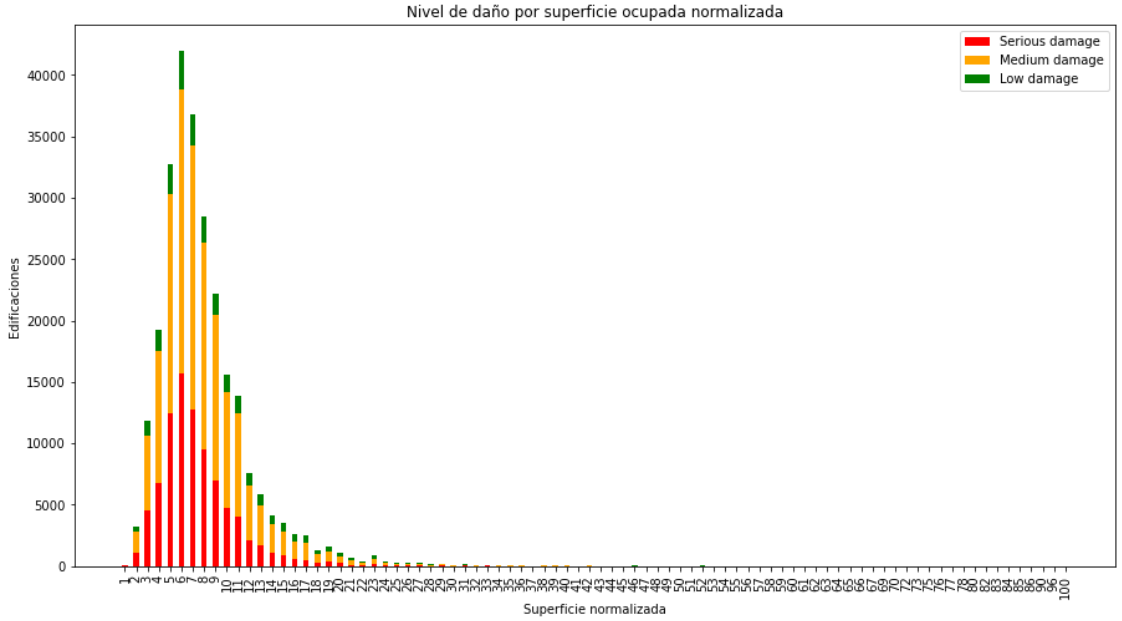
También se puede ver que parece haber una correlación entre el grado de daño y la antigüedad (las edificaciones de más de 35 años casi no tienen Low damage).

Analicemos el impacto del daño respecto a la cantidad de pisos de las edificaciones antes del terremoto:



En este caso el rango de pisos está comprendido entre 1 y 9. El mayor daño se dio en las edificaciones con 2 pisos.

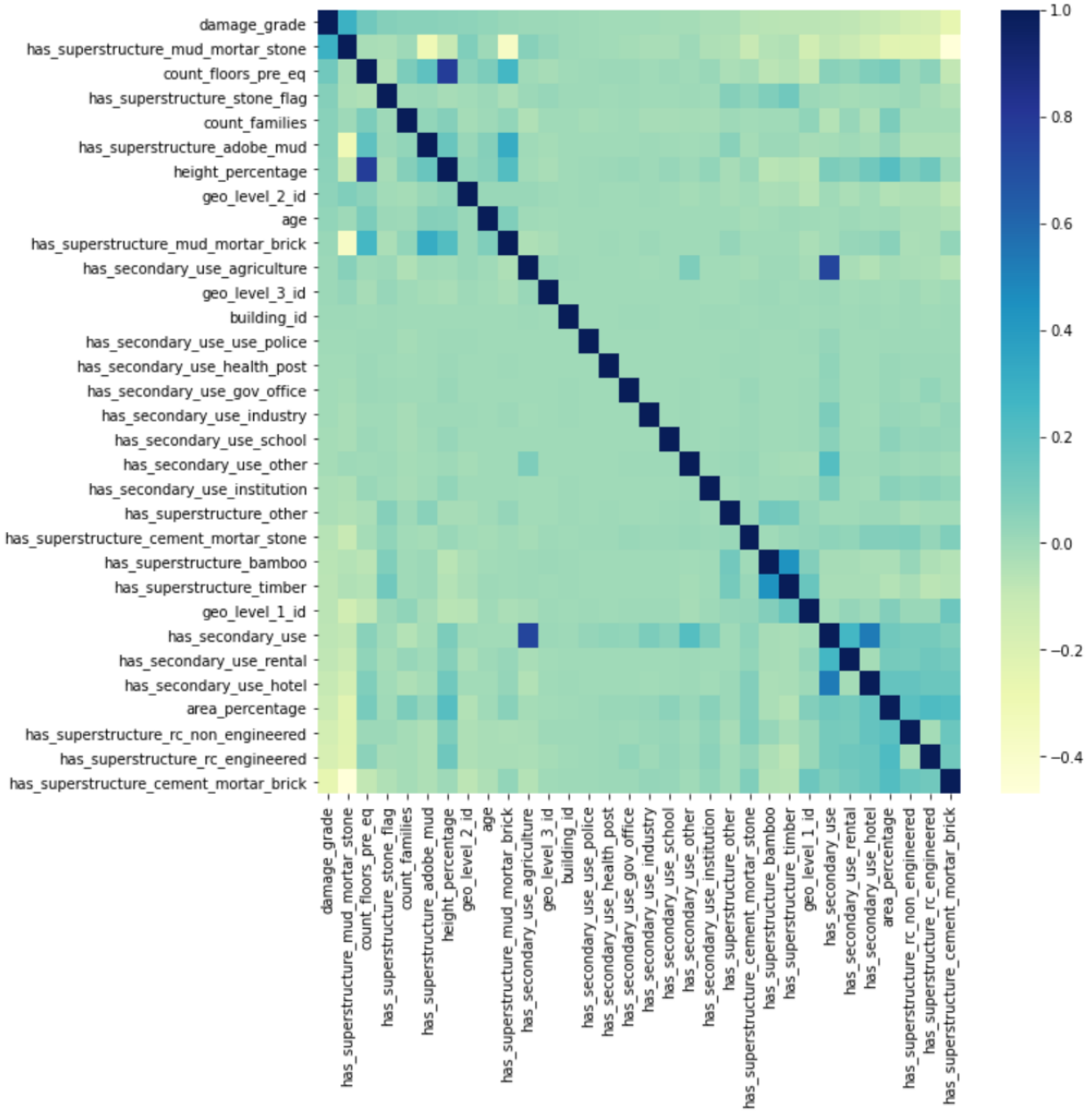
Analicemos el impacto del daño respecto a la superficie ocupada normalizada:



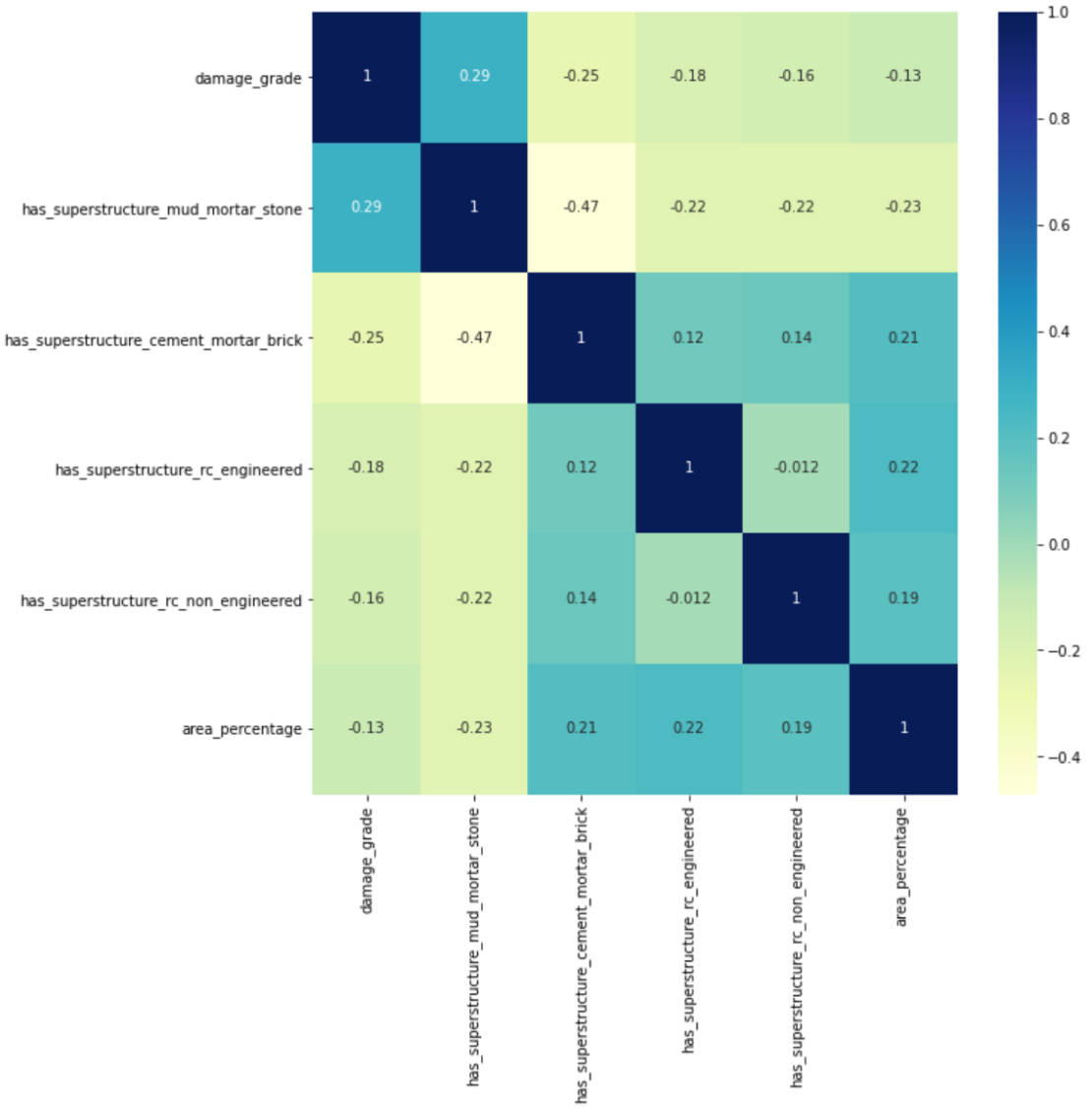
Similar a las características anteriores, en este caso el mayor impacto se dio entre las edificaciones que ocupan entre el 4 y el 10% de la superficie ocupada.

Análisis de correlación

Para ganar entendimiento sobre el conjunto de datos se analizó la correlación entre las variables numéricas. Se calculó entonces la matriz de correlación de Pearson, donde se observa que las variables más correlacionadas (linealmente) con “damage_grade” parecen ser las asociadas a la calidad de la construcción. Esto nos llevó a la primera hipótesis.



Se tomaron las variables con mayor correlación absoluta y se muestran en el siguiente gráfico para mayor comodidad.



1. **has_superstructure_mud_mortar_stone** (tipo: binario): variable que indica si la edificación fue construida con barro - piedra.
2. **has_superstructure_cement_mortar_brick** (tipo: binario): variable que indica si la edificación fue construida con cemento - ladrillos.
3. **has_superstructure_rc_engineered** (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado diseñado.
4. **has_superstructure_rc_non_engineered** (tipo: binario): variable que indica si la edificación fue construida con concreto reforzado no-diseñado.

Hipótesis

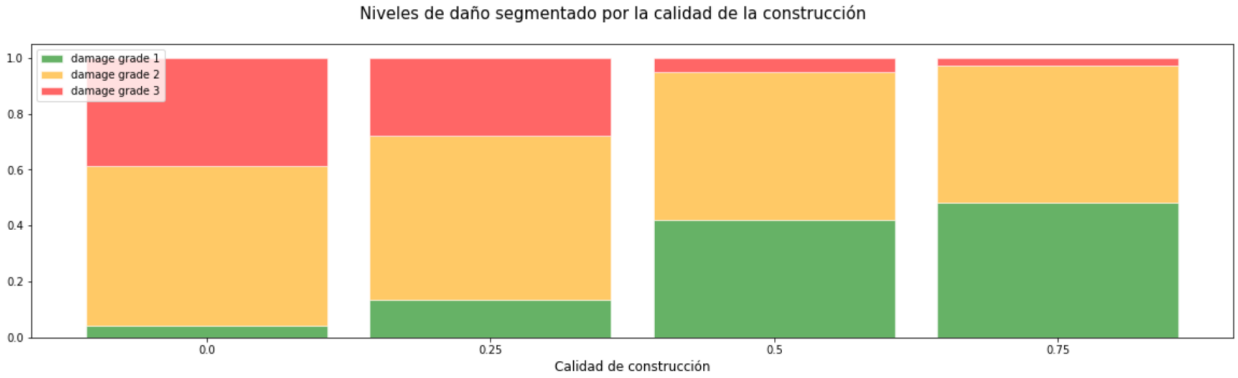
Primer hipótesis

H1: *Las edificaciones de menor calidad sufrieron daños mayores.*

Para probar esta hipótesis se creó una nueva variable que pretende aglomerar las variables de calidad de construcción.

La nueva variable la llamamos “calidad_construcción” y se construyó sumando las variables 2, 3, 4 y restando la variable 1. Se eligió sumar las variables que tienen correlación negativa con “damage_grade” y restar la única que tiene correlación negativa. Por último se aplicó una normalización “min max” de manera de llevar la variable al intervalo [0,1].

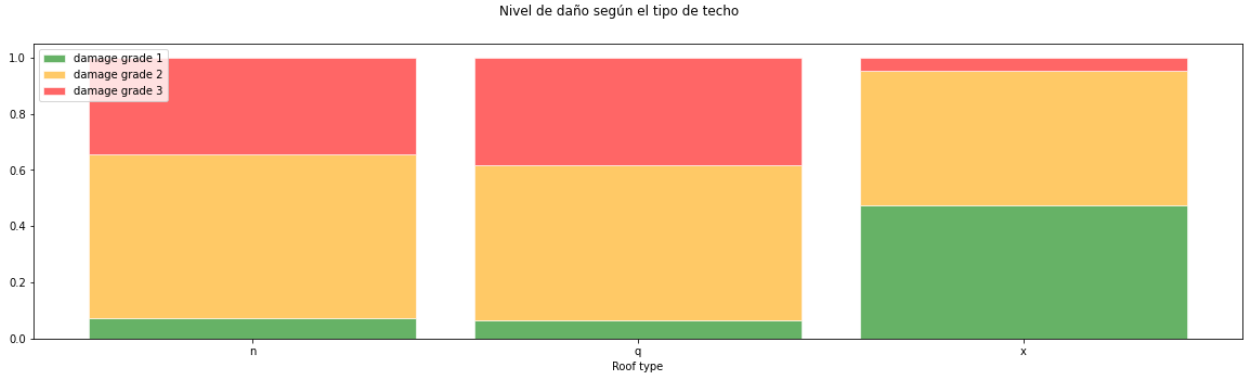
El siguiente gráfico muestra las distribuciones de las proporciones obtenidas de los distintos niveles de daño, donde se ve una clara correlación negativa entre el nivel de daño y la calidad de construcción.



Esto parece indicar que las construcciones de menor calidad efectivamente sufrieron daños significativamente mayores que las de mayor calidad.

Segunda hipótesis

Se analizó también la correlación con variables categóricas y se encontró que la de mayor impacto es “roof_type”. Como el nombre sugiere, esta variable informa acerca del tipo de techo usado en la construcción y presenta tres posibles valores ofuscados, “n”, “q” y “x”.



El gráfico muestra que el tipo de techo no parece tener gran implicancia entre las categorías “n” y “q” pero sí se ve una distribución completamente distinta del nivel de daño en el tercer tipo de techo “x”.

Esto nos lleva a pensar la segunda hipótesis de la siguiente manera:

H2: ***El tipo de techo está fuertemente relacionado con la calidad de la construcción. En especial el tipo de techo “x” parece estar asociado a una calidad de construcción muy superior.***

Si fuera cierta la hipótesis sería esperable encontrar:

- Una cantidad significativamente menor de construcciones con tipo de techo “x”.
- Un promedio de “calidad_construcción” significativamente mayor en la categoría “x” y similar entre las otras dos categorías.
- Percentiles de “calidad_construcción” más elevados en la categoría “x”

Para probar esta hipótesis analizamos la distribución de la nueva variable “calidad_construcción” condicionada a los tres tipos distintos de techos.

roof_type	count	mean	std	min	25%	50%	75%	max
n	182842.0	0.063793	0.136012	0.0	0.0	0.0	0.0	0.75
q	61576.0	0.065707	0.133879	0.0	0.0	0.0	0.0	0.75
x	16183.0	0.525752	0.152904	0.0	0.5	0.5	0.5	0.75

Para ello calculamos algunos estimadores de la distribución y se corroboran lo que se esperaba encontrar si la hipótesis fuera cierta.

Tercer hipótesis

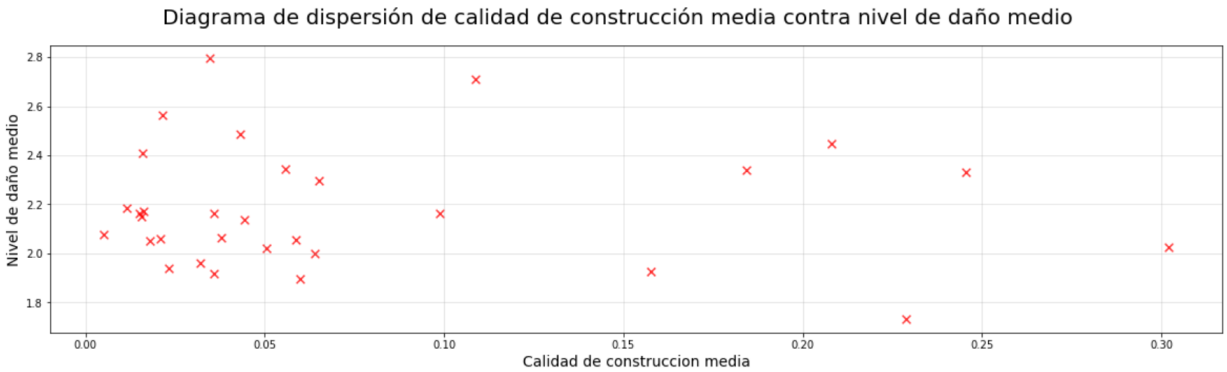
También se analizó la variable “geo_level_1_id” que toma 31 valores posibles. Si bien los valores son numéricos, no podemos afirmar que exista algún orden en esta variable y por lo tanto se considerará como una variable categórica.

Al graficar los niveles de daño promedio en las distintas regiones podemos observar que algunas fueron más castigadas que otras. Esta parte del informe pretende entender a qué se debe esta diferencia en los niveles de daño.

Hasta el momento, la calidad de la construcción parece ser determinante en el nivel de daño y se quiere analizar si esta variable de ubicación también está relacionada con esto. La tercer hipótesis que planteamos entonces es:

H3: *Las regiones que más fueron afectadas presentan menor promedio de calidad de construcción.*

Para poner a prueba esta hipótesis se agruparon los datos por región y se calcularon las medias de ambas variables en cada grupo. Para mejor interpretación se graficaron en un diagrama de dispersión donde cada cruz representa una región distinta de Nepal.



En el gráfico no se ve ninguna tendencia marcada, sin embargo es notable que la región que presenta más daño promedio (~2.8) tiene una media de calidad de construcción muy baja mientras que la región que tiene menor daño promedio (~1.7) tiene una media de calidad de construcción bastante elevada.

Para acompañar la interpretación visual, calculamos el coeficiente de correlación de Pearson entre ambas variables y nos da -0.047216 que si bien es negativo como indica la intuición, es demasiado cercano a cero como para ser significativo. Esto puede indicar que no hay relación lineal entre ambas o que existe alguna otra variable en juego que no estamos teniendo en cuenta. En este caso, consideramos que es muy probable que haya alguna otra variable de peso que no estamos considerando como por ejemplo la proximidad al epicentro del terremoto. Lamentablemente no tenemos una variable que nos indique esto último así no puede ponerse a prueba esa hipótesis.

Conclusiones

En una primera instancia parecería que muchas de las variables no aportan información respecto al daño (por ejemplo, todas las columnas que determinan si la edificación tenía un uso secundario).

Por otro lado, rescatamos dos ejes principales que son los materiales de construcción y en menor medida la ubicación en la que se encuentran las edificaciones. Resultó provechoso englobar los materiales de construcción en una nueva variable “calidad de construcción” y descartar aquellas que detectamos independientes para simplificar el análisis centrándonos en las variables más explicativas. También nos permitió encontrar una relación más fuerte, antes no tan clara, con la calidad de construcción, región y grado de daño que presentaron las edificaciones.