Hugo Andres Sabogal Perez
Natalia Plata Ángel

Taller de programación en R: Taller 3

Facultad de Economía, Universidad de los Andes

Oct 14/2024

Punto 1 (2.0):

1.1. Descarguen de la página del DANE Geoportal DANE- Página de descarga datos geoestadisticos un shapefile a nivel municipal de Colombia. Cargue este shapefile a R utilizando la librería SF y elimine San Andrés y Providencia.

Este punto no requiere explicación, el procedimiento está explícito en el código.

1.2. De la pagina del CEDE descargue el Panel de Características generales de los Municipios. Mantengan las variables que contienen la palabra "pob" y las variables del PIB.

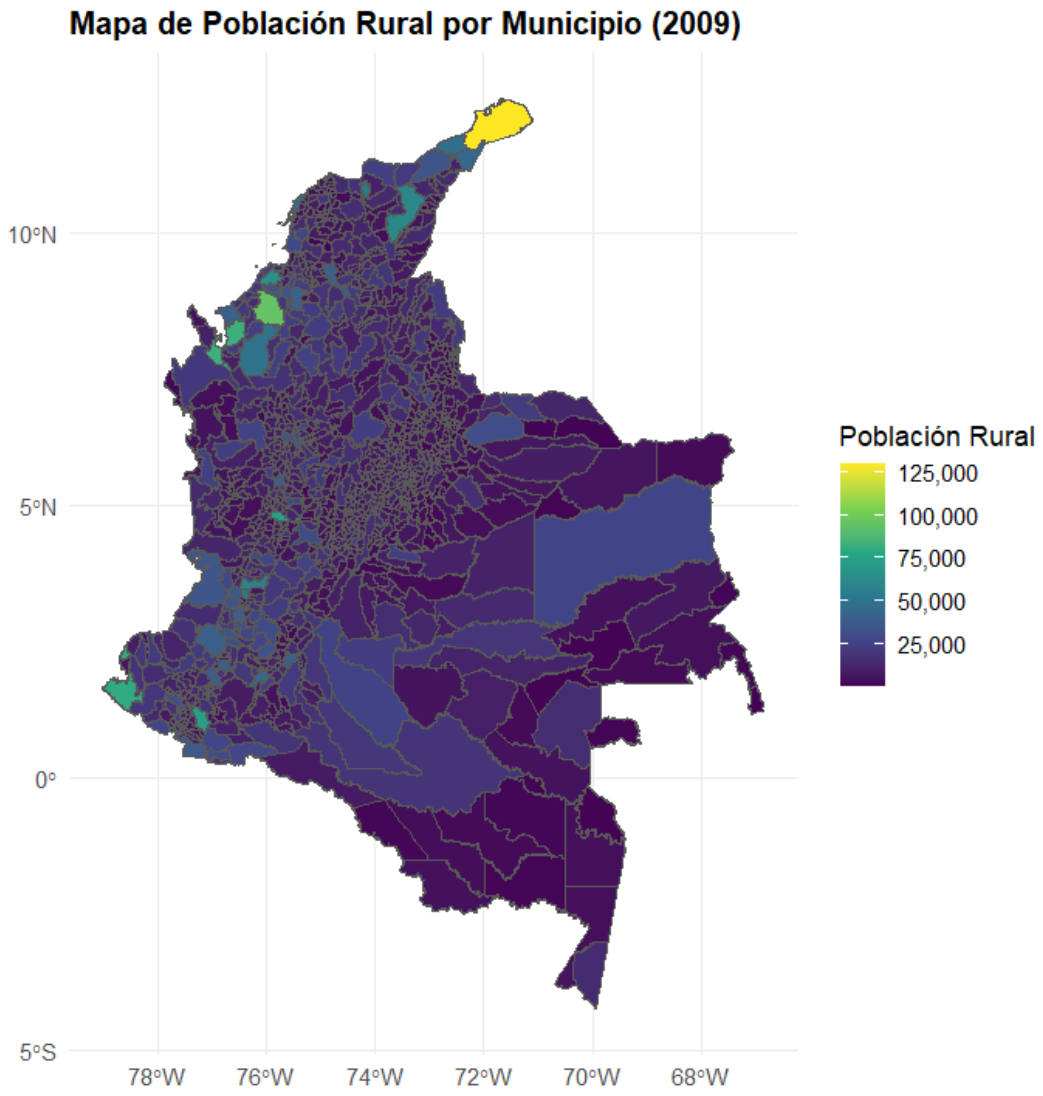
Este punto no requiere explicación, el procedimiento está explícito en el código.

1.3. **Realicen un merge del shapefile de los municipios con este dataframe, revisen dos cosas. Primero, que en ambas bases los municipios estén identificados con el mismo número -revisen los 0´s al inicio-. Segundo, que para la base resultante esté la geometría de cada unidad de nivel -municipio, depto, etc.- para cada año**.

Este punto no requiere explicación, el procedimiento está explícito en el código.

1.4. Realicen una visualización de un mapa a nivel municipal con la simbología de la población rural por municipio para el año 2009. Añadan todos los elementos estéticos que permitan una gráfica autocontenida y bien presentada. Hagan este mapa interactivo usando la librería ggplotly.

Hugo Andres Sabogal Perez
Natalia Plata Ángel



Este mapa es una representación del interactivo, el cual se puede encontrar al correr el código.

1.5. Interpreten la gráfica y cuenten una historia acorde a la misma.

Esta gráfica muestra la distribución de la población rural en Colombia por municipio en el año 2009. Cada municipio está representado con un color que indica el tamaño de su población rural, según una escala en la que el morado representa municipios con menor población rural y el amarillo aquellos con mayor población rural. La mayoría de los municipios se presentan en tonos morados y azul oscuro, lo que indica una baja densidad de población rural en gran parte del territorio colombiano. Sin embargo, existen algunas áreas en verde y amarillo en el norte y el oeste del país, las cuales sugieren mayores concentraciones de población rural.

Hugo Andres Sabogal Perez
Natalia Plata Ángel

Para el año 2009, Colombia presentaba bajas densidades de población rural en la mayoría del país, en gran parte debido a que, durante la segunda mitad del siglo XX, muchas personas se trasladaron a las zonas urbanas en busca de nuevas oportunidades. Además, otros grupos fueron desplazados forzosamente por la violencia y la presencia de actores armados al margen de la ley, lo cual contribuyó a la concentración de la población en áreas urbanas y a una disminución de la densidad en las zonas rurales.

Por otra parte, se observa que la mayoría de los municipios con mayor cantidad de población rural están ubicados en la periferia de Colombia, cerca de las fronteras del país. Esto podría deberse a que Colombia, al ser un país con un gobierno centralizado, históricamente ha tenido dificultades para mantener el control en todo el territorio. En consecuencia, es probable que en los municipios periféricos haya menor presencia del Estado y menos acceso a servicios públicos, lo que hace que la población sea más dispersa y menos concentrada en centros urbanos.

Punto 2 (3.0):

2.1. Vayan a la página All products | Books to Scrape - Sandbox la cual tiene información sobre libros en internet y su precio.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.2. Utilizando herramientas de web-scrapping, estructuren un dataframe donde las filas sea cada uno de estos libros y las columnas sean el respectivo título y el precio.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.3. Ejecuten un procesamiento del texto presente en los títulos. Para esto, remuevan las stop-words, pasen todo el texto a minúsculas y eliminen los caracteres especiales y los números.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.4. Guarden esta base de datos como datos_limpios.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.5. A partir de la base datos_limpios generen una base que se llame count la cual contiene una fila con el n-grama definido y una columna con la frecuencia de veces que aparece ese n-grama.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.6. Realicen una nube de palabras que permita entender cuáles son los n-gramas más frecuentes en los títulos de los libros. Interpreten este resultado.

Hugo Andres Sabogal Perez
Natalia Plata Ángel

Esta nube de palabras representa las palabras de los títulos de la página “books to scrap” .Las palabras más frecuentes incluyen los terminos "vol", "life", "love", "girl" y "world", sugiriendo temas variados, desde historias de vida y amor y libros con diversa cantidad de volúmenes. También se observan referencias a términos relacionados con la ciencia, historia, y cocina, lo que indica una gran diversidad de géneros. Algunas palabras como "guide", "trilogy", y "chronicles" sugieren colecciones o series, mientras que títulos como "potter" y "recipes" aluden a sagas o temas específicos.

2.7. Por otra parte, a partir de la base de datos_limpios, realicen una matriz de term frequency TF, la cual tiene en las filas cada uno de los libros y en las columnas cada una de las palabras presentes en el corpus, el valor de esta columna es el número de veces que aparece el n-grama.

Este punto no requiere explicación, sin embargo, vale la pena incluir una aclaración. Este punto se realizó en el código antes del punto 2.5. Esto se debe a que en la clase del viernes Daniel aclaró que había un problema con el orden de los puntos, ya que era necesario tener el term frecuency de las palabras en los títulos antes de poder hacer la nube de palabras.

2.8. Añadan a esta base la columna de precio del libro.

Este punto no requiere explicación, el procedimiento está explícito en el código

Hugo Andres Sabogal Perez
Natalia Plata Ángel

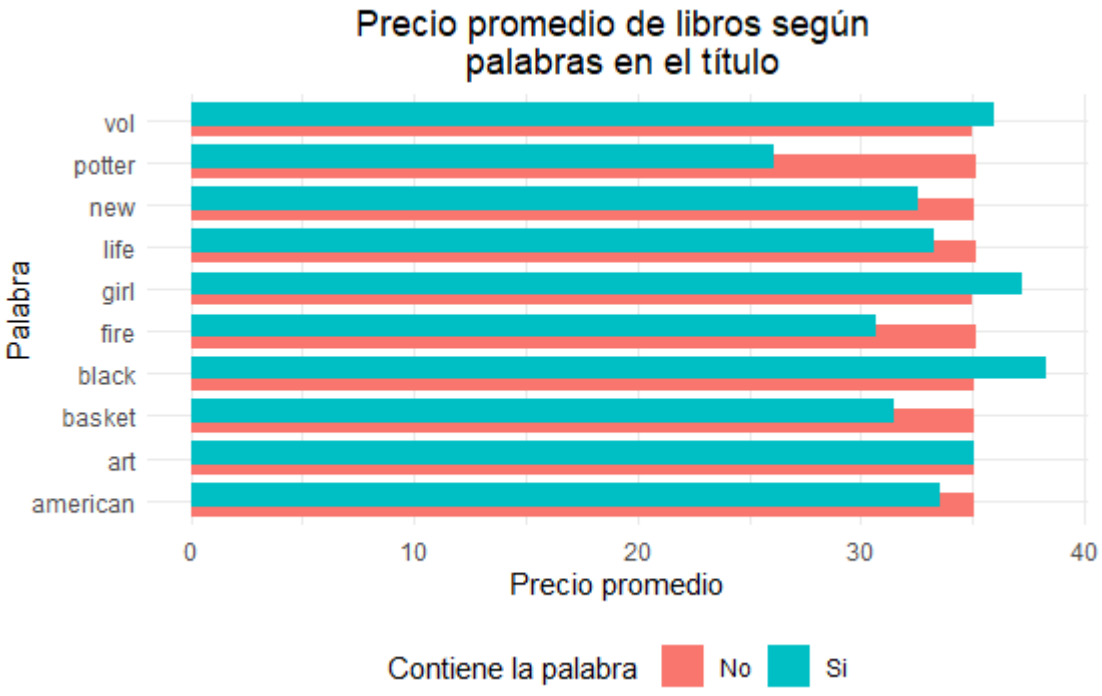
2.9. Elijan 10 palabras (las que quieran), que en esta matriz estarán en las columnas, y reemplacen la frecuencia por un 1 si aparece más de una vez o un 0 si no aparece.

Para seleccionar las 10 palabras primero seleccionamos las a las palabras que tienen una frecuencia mayor a 10 y luego escogimos 10 palabras aleatorias dentro de esta muestra.

2.10. Utilizando la función group_by() calculen para estas palabras el precio de los libros cuando estas palabras aparecen y cuando no. Es decir, cuando la variable toma el valor 1 vs 0.

Este punto no requiere explicación, el procedimiento está explícito en el código.

2.11. Presenten una visualización en forma de gráfico de barras tal que en el eje Y tenga la palabra y en el eje X el valor del libro. Añadan a la gráfica una coloración por si aparece o no aparece en el título (si la es 0 o 1). Interpreten la gráfica ¿Cuáles palabras elegidas parecen estar asociadas positivamente con el precio del libro?



Esta gráfica muestra el precio promedio de libros según si contienen ciertas palabras en el título. Cada barra representa una palabra específica y el color indica si los libros contienen esa palabra (en color turquesa) o no (en color salmón). Las palabras que parecen estar asociadas positivamente con el precio de los libros son: “girl”, “black” y “vol”, ya que en promedio el precio de los libros que contienen esas palabras es mayor que el precio de los libros que no las contienen.