Prueba ADRES Analista de datos

Autor: Juan Pablo Holguín Marín

**Metodología**

Para el desarrollo de la actividad inicialmente se comenzó descargando e instalando las herramientas necesarias, como el gestor de bases de datos SQlite y las diferentes librerías en R.

En primer lugar se inspeccionaron los archivos compartidos y se indagó sobre la naturaleza de los datos allí consignados y los posibles análisis que podrían desarrollarse con los mismos, apuntando a resolver preguntas cómo el nivel de cobertura de prestadores de servicios de salud en diferentes unidades territoriales como municipios, departamentos y regiones.

Una vez comprendido el problema y su solución se preparó el ambiente de trabajo para leer los archivos, crear una conexión desde R hacía una base de datos en SQLite para insertar allí los datos y comenzar la exploración de los mismos a través de consultas SQL ejecutadas y almacenadas en R.

Posteriormente se evaluó el comportamiento de las variables cuantitativas como la superficie de cada municipio y su población. Se evaluaron aspectos como la distribución estadística de los datos y su normalidad por medio de pruebas estadísticas y gráficos.

Una vez se tuvo mayor conocimiento de cada conjunto de datos, se procedió a unirlos a través del código DIVIPOLA, el cual se extrajo de los primeros cinco caracteres de la variable código habilitación en la tabla de Prestadores, y en la de Municipios ya venía incorporado en la variable Depmun. Con esta unión de los archivos se logró identificar cuantos prestadores estaban asociados o prestaban sus servicios en cada unidad territorial, y se exportó el archivo con el cual se construyó una visualización de los hallazgos usando las bondades del software Power BI.

**Resultados**

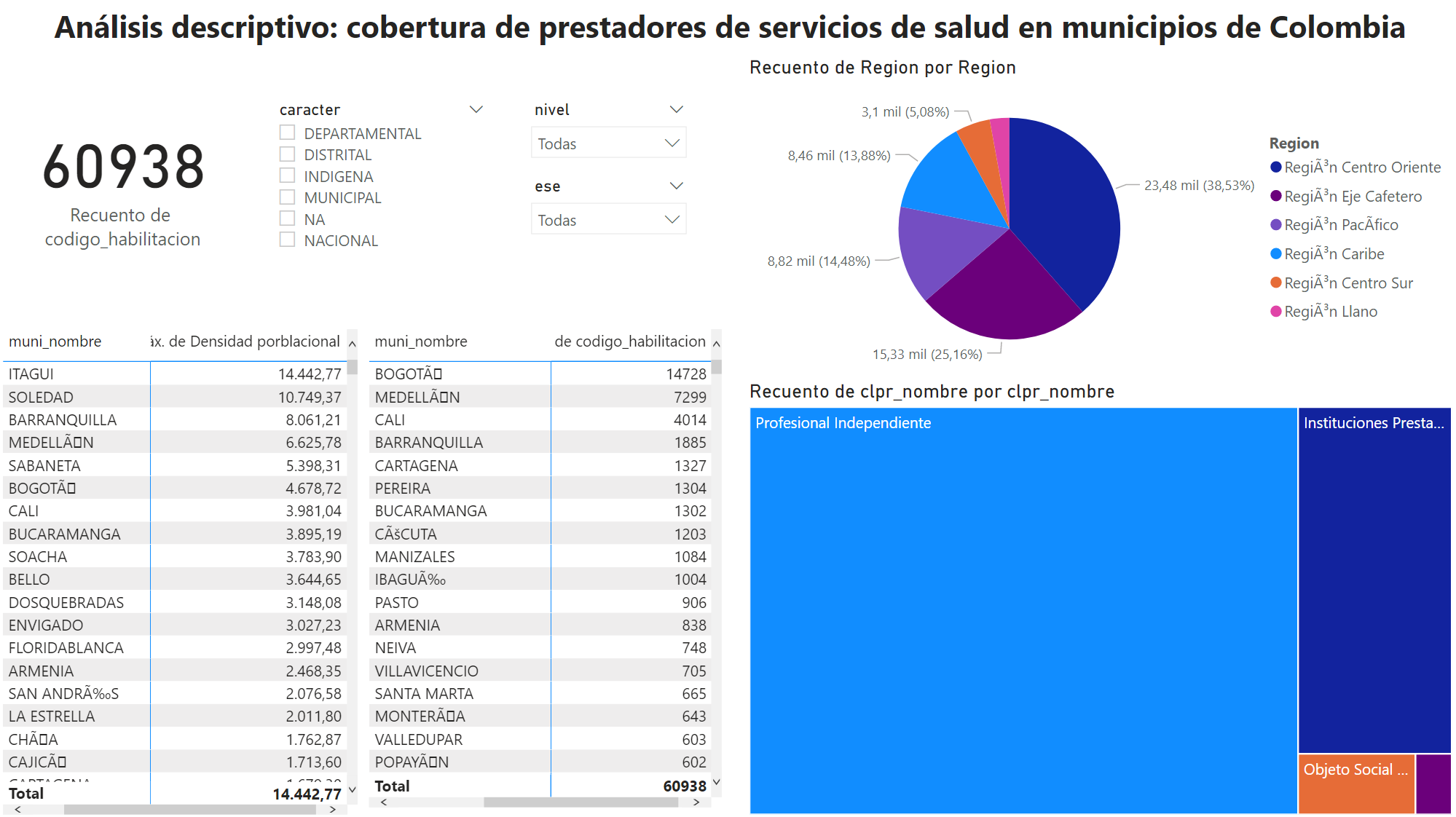
Las primeras consultas de exploración de datos permitieron observar las principales características de los datos y sus problemas más relevantes, comenzando por los caracteres especiales en los nombres de los municipios y departamentos, lo cual planteó un primer desafío de limpieza de los datos. Sin embargo, al explorar a mayor profundidad los mismos, se logró identificar que los primeros cinco caracteres de la columna código\_habilitación conformaban el código DIVIPOLA del municipio y departamento en conjunto, por lo cual se descartó la limpieza de texto en ambas tablas y se optó por acudir al código DIVIPOLA que brindaba mayor integridad de los datos.

Al construir gráficos para evaluar la normalidad y distribución de las variables cuantitativas, se observó que por ejemplo en variables como la superficie y población de las unidades territoriales, la mayor cantidad de datos se encontraba agrupada en las primeros valores, lo cual puede brindar un indicio de que los datos pueden estar sesgados hacia valores más pequeños y se tienen registros de valores con superficies muy altas, situación similar a lo evidenciado con la variable población. El comportamiento del gráfico de normalidad de la población podría significar que los datos tienen una mayor concentración de observaciones en la parte inferior de la distribución, pero también tienen un número inusualmente alto de valores extremadamente altos en comparación con una distribución normal. Las colas más pesadas en el extremo derecho sugieren una mayor variabilidad o valores atípicos en el lado superior de la distribución (Figura 1).

|  |  |
| --- | --- |
|  |  |
|  |  |

**Figura 1.** Distribución y normalidad de variables cuantitativas.

Por último, se exportó el archivo que relacionaba a cada prestador en cada unidad territorial para ser analizado visualmente en Power BI, en donde se construyó una nueva variable llamada densidad poblacional para determinar el número de personas por área y así permitir realizar un análisis de la cobertura de los prestadores de servicios de salud según la densidad poblacional y no solamente sobre la población o superficie de una unidad territorial. De esta manera se logró evidenciar cuales eran los municipios con mayor densidad poblacional, y analizaer si eran estos mismos quienes presentaron una mayor cobertura por parte de los prestadores que allí operan o no, caracterizando esta cobertura según la región, el carácter y razón social de los prestadores, su nivel de servicio, y si eran ESE o no.



**Figura 2.** Análisis descriptivo de la cobertura de los prestadores en Colombia