ANÁLISIS EXPLORATORIO: MORTALIDAD EN UN GRUPO HOSPITALARIO

Roger Sans I Prats

DATA SCIENTIST

ÍNDICE

Carga y limpieza de los datos

Lo que hacemos primero de todo es cargar las librerías necesarias para el desarrollo del informe y cargar los datos del fichero .csv que se nos ha proporcionado en un data frame llamado df, con la configuración establecida para una correcta lectura del mismo. Una vez cargado dicho archivo, procedemos con la limpieza de los datos. En este caso, únicamente se ha requerido de modificar la codificación "." en la columna "Total" para una correcta interpretación de los datos.

Filtrado de los datos

Para la resolución de las cuestiones planteadas, previamente se nos requiere un filtrado de los datos. Este consiste en lo siguiente:

- Sexo: en esta columna hemos seleccionado aquellos datos pertenecientes al grupo "Ambos sexos", ya que corresponde a la suma de los grupos "Hombres" y "Mujeres", y por lo tanto, de incluir también ambos grupos estaríamos duplicando los resultados y aportando una información errónea acerca de los mismos.
- Causas: en esta ocasión, el filtrado que se ha requerido es eliminar una causa que venía con el nombre "001-102 I-XXII.Todas las causas" ya que de incluirla, del mismo modo que nos sucedía en el apartado anterior estaríamos duplicando los resultados.
- Edad: para ésta columna, nos suceden nuevamente los visto en los casos anteriores. Excluimos el grupo de "Todas las edades" para que no se no repita información y caer en posibles errores dentro de nuestro análisis.

Intervalo de edad con más defunciones

Con el fin de mostrar los datos que respondan a esta cuestión, hemos utilizado la función "group_by" para agrupar los diferente grupos de edad posibles dentro del dataset ya filtrado. Una vez hecho este proceso con una función de suma aplicada a la columna "Total", hemos unido todas las causas de defunción en una sola llamada "Total" para que así se muestre el total de defunciones por grupo de edad. Finalmente, para una rápida visualización de los datos hemos añadido un orden descendiente para que así, aquellos grupos de edad con mayor número de defunciones se visualicen más rápidamente en la parte superior.

Una vez explicado dicho proceso, se ha determinado que el grupo con un mayor número de defunciones es de 85 a 89 años, con un total de 81189 defunciones.

Causa con la media de defunciones por año más alta

Si observamos detenidamente esta pregunta, donde se nos pide el cálculo de una media anual, si hemos hecho una observación previa del dataset para ver a que corresponden los datos, nos habremos dado cuenta de que el conjunto de registros que se nos ofrece corresponde a un periodo de tiempo indeterminado, ya que no se ofrece una información a nivel temporal de los registros y, por lo tanto, no podemos deducir si el dataset extraído corresponde a un periodo de tiempo de meses, años, etc. Entonces no tiene mucho sentido responder a esta pregunta porque no se puede calcular una media anual sin saber que registro corresponde a cada espacio temporal distinto. Ahora bien, para brindar una información que podría ser de utilidad, podemos observar la media de defunciones para cada grupo de edad o de sexo. Ambos cálculos nos explicarán un poco mejor como están distribuidas las defunciones según las características demográficas. Primeramente, vamos a abordar el caso de la media para cada grupo de edad, que nos puede dar información al respecto del impacto que tiene cada causa de defunción dentro de un grupo etario. Para ello, volvemos a hacer uso de la función "group_by" pero esta vez pasando dos parámetros por la función; el primero, la columna de "Causas..lista.reducida." y el segundo, la columna "Edad". Entonces tendremos diferentes grupos correspondientes a cada causa de defunción respecto a cada uno de los grupos etarios. A partir de aquí, generamos una nueva columna llamada "Media_defunciones" donde se calculará la media correspondiente a los datos. Una vez explicado dicho procedimiento, de aquí podemos sacar algunas conclusiones como que las Enfermedades del sistema circulatorio presentan unos valores de media considerables a partir de los 80 años, con una media de 13681 entre los 90 y 94 años, y de 13295 entre los 85 y 89. Los tumores también presentan valores bastante elevados a partir de los 75 hasta los 89 años.

En segundo caso, el observar la media de defunciones para cada sexo nos puede ayudar a identificar si existen diferencias importantes entre hombres y mujeres en términos de mortalidad para estas diferentes causas, y si el análisis realizado estuviese orientado a las diferencias de género sería un primer cálculo imprescindible para tener una primera representación de los datos con la cual seguir trabajando posteriormente. Para ello, previamente hemos modificado el filtrado para separar el grupo de ambos sexos en uno de hombres y otro de mujeres, a partir de aquí, igual que en el apartado anterior, hemos utilizado la función "group_by" para agrupar cada causa de defunción en cada uno de los sexos diferentes y posteriormente generado una nueva columna llamada "Media_defunciones" para calcular la media de los datos. A partir de aquí, en una primera observación podemos ver como para el caso de las Enfermedades del sistema circulatorio los hombres presentan valores de media muy

similares al de las mujeres, y por lo tanto podemos indicar que esta causa incide de una manera similar o parecida en ambos sexos. Sin embargo, cogiendo el ejemplo de los tumores vemos como los valores que corresponden a la media en hombres es superior al de las mujeres.

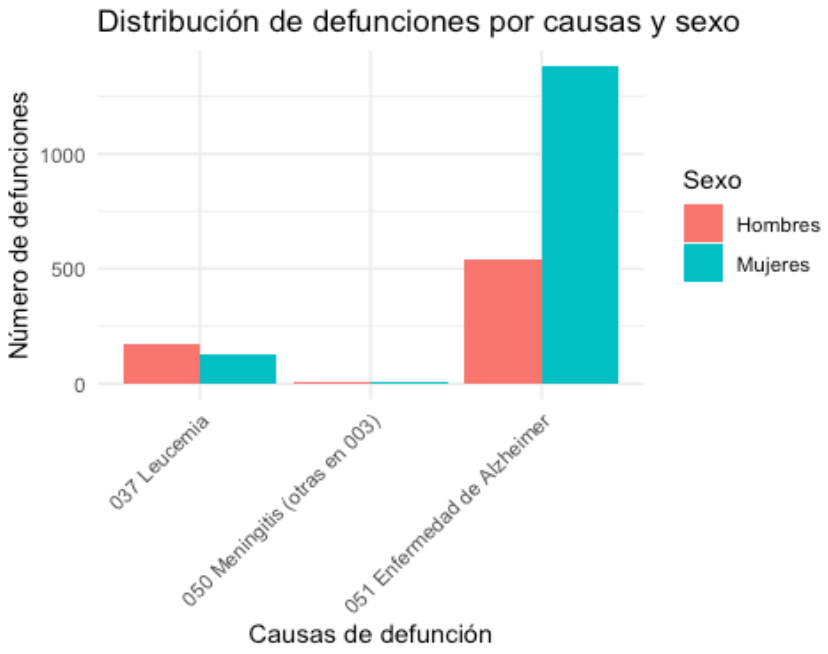
Segundo filtrado de datos

Para la resolución de las cuestiones planteadas en la actividad, previamente se nos requiere un filtrado de los datos distinto al presentado hasta ahora. Éste consiste en lo siguiente: - Sexo: en ésta columna hemos eliminado aquellos datos pertenecientes al grupo "Ambos sexos", ya que corresponde a la suma de los grupos "Hombres" y "Mujeres", por lo tanto estaríamos duplicando los resultados y aportando una información errónea acerca de los mismos. - Causas: en esta ocasión, el filtrado nos viene dado para acotar la información a aquella relevante para un posterior uso, y del largo listado de causas de defunción se nos requieren únicamente "Leucemia", "Meningitis" y "Enfermedad de Alzheimer". - Edad: para ésta columna, nos suceden los ambos casos mostrados en las columnas anteriores. Primeramente, excluimos el grupo de "Todas las edades" para que no se no repita información y caer en posibles errores dentro de nuestro análisis. Y en segundo caso, se nos requiere que no mostremos el grupo "Menores de 1 año" para brindar una información más relevante para su posterior uso.

Visualización 1. Distribución de las defunciones por sexo y causas.

En este apartado, se nos pide que realicemos una visualización de los resultados en forma de diagrama de barras, donde se pueda observar la distribución que siguen los datos respecto ambos sexos y las diferentes causas de defunción. También se especifica que cada barra debe representar una causa de defunción y se debe dividir en grupos de sexo. Para efectuar dicha visualización hemos hecho uso de la librería "ggplot2", y seleccionando el dataframe ya filtrado nuevamente, hemos pasado que parámetros queremos graficar en el eje de las X y en el eje de las Y, en esta caso como queremos graficar un diagrama de barras que nos muestre el número de defunciones por causa, la primera variable va a ir al eje de las Y, y la segunda al de las X. Además vamos añadir mediante el comando "fill" la agrupación por sexo que nos pide el enunciado para que así se distinga una columna para los hombres y otra para las mujeres. Hasta ahora en nuestro código estábamos haciendo referencia a el contenido de nuestro gráfico. En la siguiente línea, vamos a hacer referencia a la geometría de las barras del diagrama, con el comando "stat ="identity"" lo que hacemos es dotar de proporcionalidad la altura de la barra

respecto los valores de la columna total representadas en el eje de las Y. Con el “position = dodge”” separamos cada barra en dos para que así se muestre una barra coloreada para los hombres y otra de distinto color para las mujeres. Si seguimos en la próxima línea con la etiqueta “labs” vamos a personalizar todo lo que son los títulos del gráfico tanto el general como el de cada eje y de la leyenda, aquí hemos titulado con cada demanda del enunciado. Y por último la etiqueta “theme_minimal” hace referencia al diseño del gráfico y se puede cambiar a gusto del analista, así como la línea inferior donde para obtener una mejor visualización hemos aplicado un comando para rotar ligeramente las etiquetas del eje de las X y aportar una mayor legibilidad al gráfico. Dejando de banda aspectos técnicos, si nos centramos en analizar el diagrama, prácticamente no encontramos una diferencia significativa entre hombres y mujeres para la Meningitis. Sin embargo, en el caso del Alzheimer existe una diferencia bastante remarcable respecto hombres y mujeres ya que la columna correspondiente al sexo femenino es bastante más alta que la que corresponde al masculino. Del mismo modo sucede con la Leucemia pero de una manera mucho más discreta donde la columna correspondiente a los hombres sobresale ligeramente por encima de la de las mujeres.



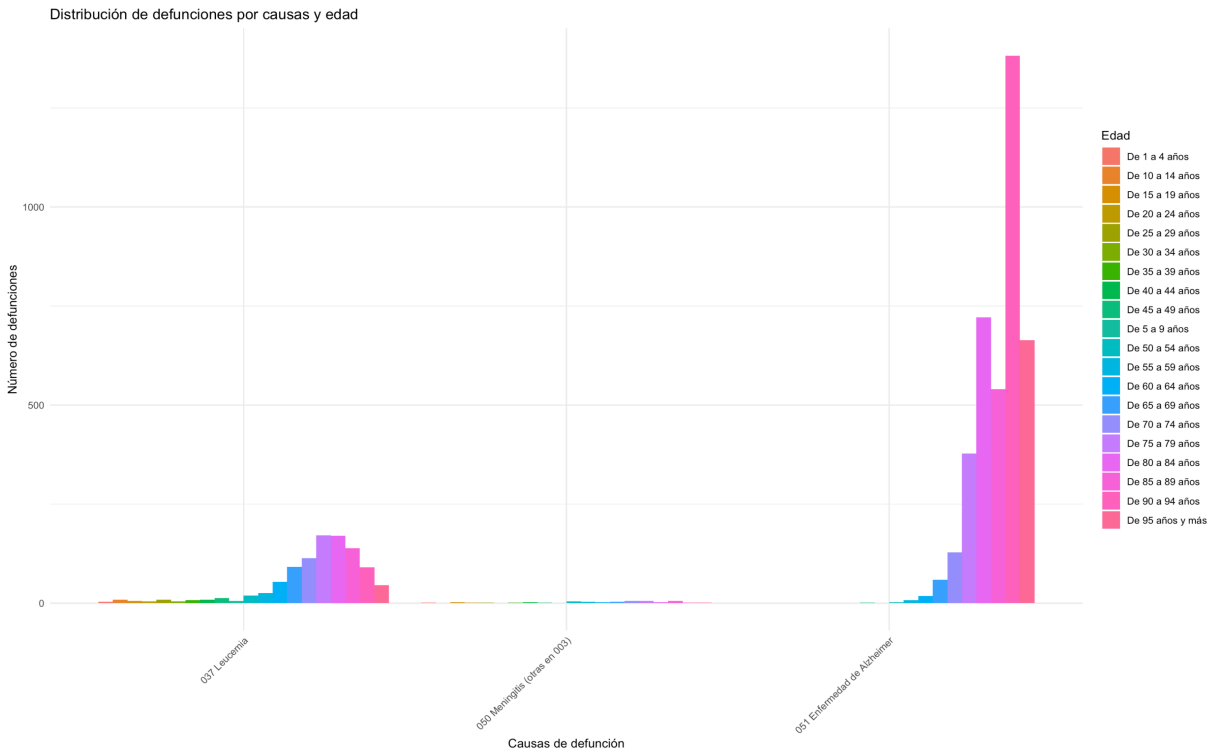
Gráfica 1: Distribución de defunciones por causa y sexo

Fuente: Elaboración propia

Visualización 2. Distribución de las defunciones por edad y causas

En la segunda visualización, se nos pide que observemos los resultados mediante un diagrama de barras donde se muestre la distribución que siguen los datos respecto a los distintos grupos de edad y las diferentes causas de defunción. También se especifica que cada barra debe

representar una causa de defunción y se debe dividir en grupos de sexo. En cuanto a la construcción del gráfico, el proceso ha sido exactamente idéntico al realizado en la visualización anterior pero cambiando la división de cada columna en lugar de por sexo, por la edad. En cuanto al gráfico en si, observamos que el Alzheimer, de manera significativa, despunta de manera general como la causa con mayor número de defunciones prácticamente en todos los grupos de edad, exceptuando aquellos grupos correspondientes a edades no tan avanzadas, aproximadamente por debajo de los 70 años, donde la Leucemia también tiene una incidencia parecida en cuanto a la mortalidad que genera en dichos grupos de edad. Para la Meningitis se observa una continuidad en los datos bastante regular en cada uno de los grupos de edad.



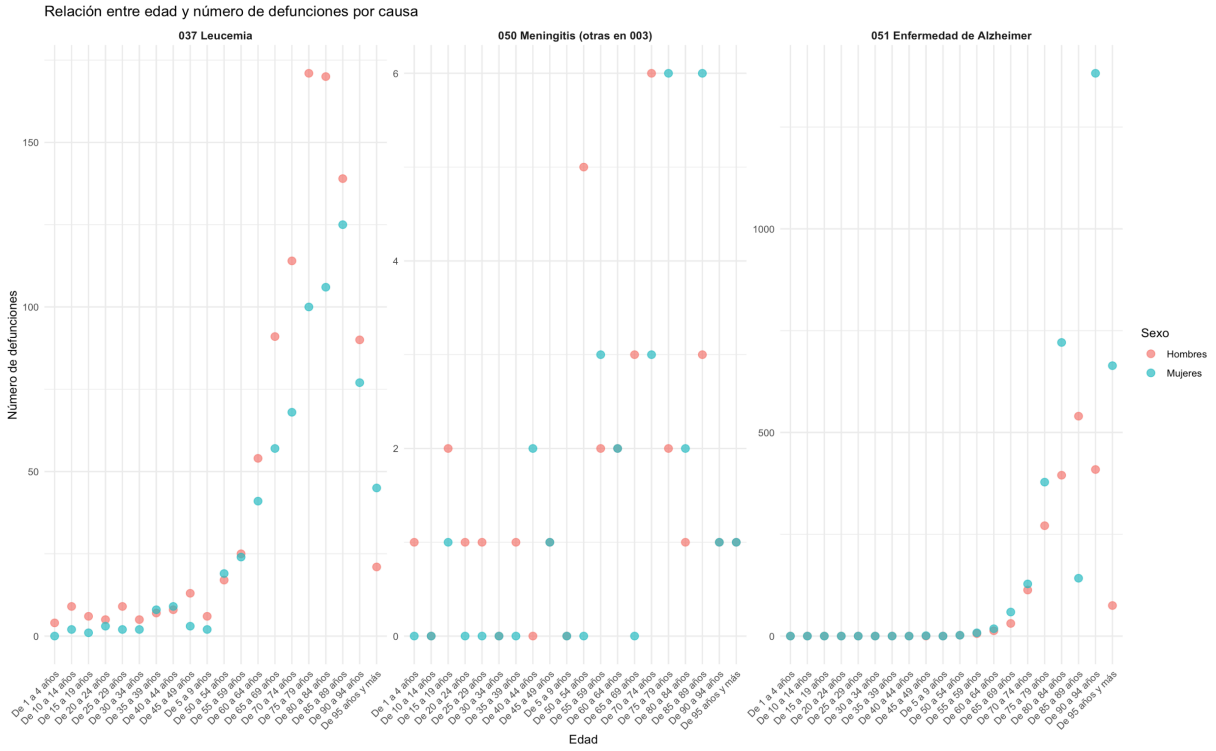
Gráfica 2: Distribución de defunciones por causa y edad

Fuente: Elaboración propia

Visualización 3. Relación entre edad y defunciones por causa

En la última visualización, se nos pide representar en un gráfico de dispersión la relación entre la edad y el número de defunciones por causa, separando en distintos puntos cada sexo. Como en este caso nuevamente nos encontramos con una visualización distinta, voy a realizar

primeramente una pequeña explicación técnica respecto a la elaboración del gráfico solo en aquellos aspectos distintos a los anteriores. La primera línea de código observamos que sigue una estructura muy parecida a la utilizada anteriormente seleccionando datos y distribuyéndolos en el gráfico según lo solicitado en el enunciado. En la segunda línea, vamos a hacer referencia a los puntos del gráfico, ajustando la transparencia de los puntos mediante "alpha" y el tamaño mediante "size" para mejorar la visibilidad e interpretación del gráfico. Con la etiqueta "facet_wrap" lo que hacemos es dividir en tres paneles distintos el gráfico para así separar las tres causas de defunción diferentes y obtener un diseño más limpio, en esta línea de código también hemos aplicado con "scales ="free_y"" que cada panel tenga su propia escala para que así se vea un diseño proporcionado ya que en este caso nos estamos centrando más en que patrones o tendencias podemos observar en la mortalidad relacionadas con la edad. Y, a partir de aquí, el código prosigue idéntico seleccionando las etiquetas de los títulos y perfilando algunos detalles para una mejor presentación y legibilidad del diagrama. En cuanto a este diagrama, como bien he comentado anteriormente, tiene la intención de diagnosticar posibles patrones o tendencias en cuanto a la mortalidad en los diferentes grupos de edad. Para la Leucemia, podemos observar cómo tanto para hombres como mujeres durante la primera mitad de los grupos de edad aunque las mujeres tienen mayor número de defunciones, esta diferencia es muy ligera pero se mantiene prácticamente a lo largo de todos los grupos de edad. También observamos que esta enfermedad cuando tiene una mayor incidencia respecto a los otros grupos de edad es ente los 60 y los 89 años para ambos sexos. Centrando la mirada en la Meningitis observamos como la distribución no sigue una uniformidad tan marcada como en las otras dos, sin embargo podemos destacar un ligero incremento entre los 60 y los 89 años de manera muy genérica pero si hiciéramos un estudio más afondo de dicha distribución probablemente se podría observar que dichas diferencias no son significativas y, por lo tanto, no nos explicarían ningún tipo de tendencia al respecto. Por último, el Alzheimer si que sigue una distribución muy marcada donde en los grupos de edad más jóvenes es inexistente pero a medida que superamos los 60 años empieza a crecer de manera significativa alcanzando un pico para los hombres entre los 85 y 89 años y, en las mujeres, entre los 90 y 94 años. De esta última observación y de las anteriores donde en grupos de edad más avanzados hay una mayor mortalidad por parte de las mujeres, se puede sacar una ligera conclusión que debería ser contrastada pero que ya nos indica una posible diferencia significativa entre ambos sexos, es que las mujeres tienen una esperanza de vida mayor que los hombres.



Gráfica 3: Distribución de defunciones por causa y edad

Fuente: Elaboración propia