Exploratory Data Analysis

Bootcamp Data Science – full time

The Bridge, Digital Talent Accelerator

Noviembre 2020

Marcos Díaz Díaz

Índice

1. Elección de tema

¿Cuál es el tema a analizar en este proyecto?

Las distintas causas de fallecimiento en España entre los años 2007 y 2018, ambos inclusive, y su evolución tanto en la distribución geográfica del país como en el tiempo.

¿Por qué se ha elegido esta temática?

Existen diversas razones que me han llevado a decidirme por esta temática particular:

- Tengo perfil docente enfocado a salud: Educación Física.
- Especialidad en investigación e intervención en la enfermedad de Alzheimer.
- Interés en sectores sanitario y socio-sanitario.
- Soy firme defensor del sistema sanitario público.

¿Qué pregunta o preguntas se van a responder?

- ¿Cuáles son las causas principales de fallecimiento en España?
- ¿Cuál ha sido su evolución a lo largo de los años?
- ¿Cómo inciden estas causas en la diversidad de la geografía española?
- ¿Cuáles son las diferencias entre comunidades autónomas?
- ¿Existe relación entre las causas y la renta per cápita?
- ¿Y su relación con la inversión en sanidad?
- ¿Qué diferencia existe entre regiones con rentas altas y bajas?

¿Qué pasos se van a seguir?

- Obtención de datos
- Lectura y limpieza de datos
- Visualización
- Interpretación de gráficos

2. Obtención de datos

Se han utilizado diversas fuentes para obtener la información necesaria para responder a las preguntas planteadas:

Instituto Nacional de Estadística
www.ine.es

Defunciones según la causa de muerte

Sistema Nacional de Salud
http://inclasns.msssi.es

Indicadores clave

Ministerio de Sanidad, Consumo y Bienestar Social
https://www.mscbs.gob.es

Sanidad en datos

Todos los archivos utilizados han sido de tipo *xls* y *xlsx*.

En total se han empleado 6 datasets:

- 2 con datos de fallecimiento por causa, año y comunidad autónoma
- 1 con datos de renta per cápita
- 1 con datos de tasa de pobreza
- 1 con datos de inversión pública por habitante en sanidad
- 1 con datos de número de médicos especialistas por cada 1000 habitantes

En la recogida de datos de fallecimiento por causa, año y comunidad autónoma se han seleccionado las categorías de las causas y no todas las subcategorías, pues no era la intención de este análisis entrar en esos detalles.

3. Lectura y limpieza de datos

Para la lectura y el desarrollo de todos los pasos de la limpieza de datos se ha utilizado Python en el entorno Jupyter Lab de la suite Anaconda.

Los datos leídos en bruto mostraban dataframes desorganizados que arrojaban la información de forma inconexa. Para hacer una lectura correcta he tenido que determinar inicio y fin de lectura de filas.

Para limpiar la información, estructurarla y tener los datos preparados para ser utilizados he empleado principalmente la librería de *pandas*.

- He encontrado dificultades a la hora de manejar ciertas columnas con motivo del tipo de dato que había, de modo que en cada dataset me he asegurado de estar trabajando con el tipo de dato correcto.

- A la hora de juntar los datasets relativos a población, renta per cápita, tasa de pobreza, inversión por habitante y médicos especialistas, realicé una forma poco ortodoxa -pero funcional- inicialmente para añadir sus correspondientes columnas al dataframe donde manejo los datos totales de causas de fallecimiento, pues no sabía realizar el 'merge' de forma correcta. Empleé *numpy* para recoger los valores de cada año y guardarlos ordenados en un array que serían los datos de la nueva columna.

- Para poder realizar de forma correcta el 'merge' mencionado arriba era necesario dar un paso previo: emplear la función 'melt' para darle la estructura correcta a los datasets y poder unirlos al principal.

- Los datos de cada causa de fallecimiento vienen dados en valores absolutos, de modo que de cara a poder comparar unas comunidades autónomas con otras los he adaptado a la población de cada región de cada año particular. El resultado son porcentajes de fallecimiento por cada causa, por cada comunidad autónoma, por cada año, con respecto a la población de ese año. En otras palabras, se muestra qué porcentaje de la población de cada comunidad autónoma fallece por cada causa cada año.

- Los datos de médicos especialistas vienen dados por cada 1000 habitantes, de modo que lo he adaptado a la población particular de cada comunidad autónoma de cada año.

La idea inicial era trabajar con los datos de 2000 a 2018, sin embargo, tuve dificultades para encontrar información de todos estos años relativas a tasa de pobreza, inversión en sanidad por habitante y médicos especialistas. Sólo hallé datos a partir de 2007 y quería darle el mismo rango temporal a todas las gráficas e interpretaciones a posteriori, motivo por el cual opté por trabajar sólo con este rango.

Además, la información de Ceuta y Melilla no estaba completa, de modo que decidí no contemplarlas para mi análisis, quedándome con las 17 comunidad autónomas.

También quería incluir información relativa al género, rangos de edad e inversión particular en sanidad por cada enfermedad/patología, pero la magnitud del proyecto era superior al nivel que actualmente poseo. Esa idea aún la mantengo y quiero desarrollarla a medio plazo.

4. Visualización

Para representar gráficamente mis datos una vez estaban limpios y estructurados he utilizado diferentes librerías en el entorno Jupyter Notebook de la suite Anaconda:

Matplotlib

Seaborn

Plotly

Folium

Según las necesidades de lo que quería representar he empleado una u otra.

Además de los recursos utilizados en clase sobre visualización, me he servido de diferentes webs para tomar referencias:

https://plotly.com/python/

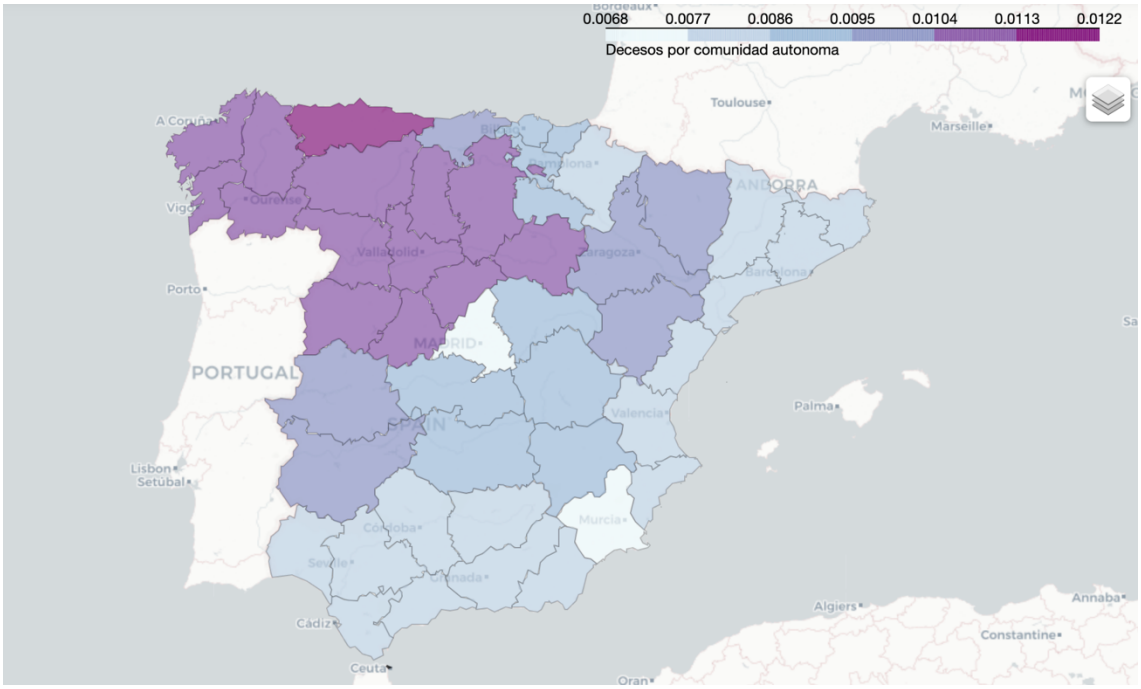
https://python-graph-gallery.com/

https://www.data-to-viz.com/

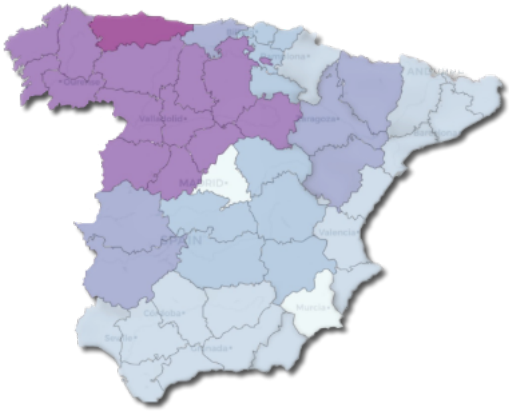
A la hora de utilizar *folium* he encontrado problemas para mostrar las Islas Baleares y las Islas Canarias en el mapa. Sus datos vienen incluidos en el archivo *geojson* que empleo, pero no he logrado visualizarlas. Ambas regiones no se muestran en los mapas que he generado analizando la incidencia de enfermedades por zona geográfica, acotando la visualización a la España peninsular.

Los mapas, una vez generados, los he editado con el fin de eliminar todo aquello que escapa a los límites geográficos de la España peninsular, pues originalmente aparecían Portugal, Francia y las zonas correspondientes a agua.
No conseguí representar únicamente la España peninsular con fondo neutro como se muestra en la siguiente url: https://plotly.com/python/choropleth-maps/

Original



Resultado



5. Interpretación de gráficos

La visualización de los gráficos ha arrojado cosas muy curiosas acerca de diferentes temas, pues en el momento de relacionar los porcentajes de deceso de cada comunidad autónoma con las demás muestra cómo se comporta el índice de mortandad en las diferentes regiones de España por cada causa.

Además, la comparación de los porcentajes de decesos con las variables de renta per cápita, tasa de pobreza, inversión en sanidad por habitante y el número de médicos especialistas, permite hacerse una idea clara sobre qué indicadores están afectando o no a que haya un resultado determinado en una comunidad autónoma.

Algunas de las conclusiones obtenidas a raíz de las visualizaciones son:

- Existe una gran diferencia en % de decesos entre comunidades autónomas.
- Hay enfermedades que tienen mayor prevalencia en ciertas áreas geográficas.
- No hay relación entre decesos y poder adquisitivo (renta y tasa de pobreza).
- La tendencia de inversión en sanidad por ciudadano es ascendente.
- Las diferencias de renta per cápita entre comunidades no justifican sus inversiones en sanidad.

6. Recursos adicionales

Para solución de dudas:
Stackoverflow - https://stackoverflow.com/questions

Para obtención de archivo *geojson* de la geografía española:
Opendatasoft - https://public.opendatasoft.com/

Edición de imagen para mapas:
PhotoScissors - https://photoscissors.com/

Para el control de versiones:
GitHub - https://github.com/marcosdiak