PRESENTACIÓN

Bon día! Os saludo desde Valencia.

Soy Inés Benaches, aspirante a analista de datos Junior, y como parte final de este bootcamp en upgrade hub, y como todos mis compañeros, vengo a presentaros mi proyecto, un análisis demográfico mundial.

El pasado 15 de noviembre se eligió, como fecha simbólica, el momento en que la humanidad ha alcanzado los 8 mil millones de personas.

Al escuchar esta cifra tan abrumadora, hay quien piensa en el término sobrepoblación. Otros lo tachan como una historia de éxito de la humanidad.

Con este análisis no pretendemos responder a esta gran pregunta, pero sí ponernos en contexto mostrándonos **el mundo en datos**, para que cada uno pueda llegar a sus propias conclusiones.

Para coger perspectiva haremos un breve recorrido sobre las estimaciones de población humana desde la prehistoria, veremos un análisis global con datos de Naciones Unidas desde 1950 hasta 2021 y profundizaremos en parámetros estadísticos por país.

Por último veremos cómo se ejecuta una predicción de crecimiento de la población global en un futuro próximo.

HISTÓRICO GLOBAL

Para realizar este gráfico hemos usado un dataset de DataHub que es un compendio de estimaciones de población desde 1 millón de años antes de la Era Común (o año 1), hasta 1995.
Dichas aproximaciones provienen de la media de varias estimaciones independientes, publicadas, entre otros, por del paleolimnólogo Edward Deevey, el polimático Colin Peter McEvedy y Naciones Unidas.

En este gráfico sensible podemos ver que la población total ha permanecido bastante estable a lo largo de los años previos a la era común. Sin embargo asociamos picos de crecimiento a descubrimientos tecnológicos que han mejorado la calidad de vida en materias de refugio, comida y salubridad.

9000 - 6500 AEC: popularización de la agricultura , Ladrillo
4000 AEC : escritura , Arado poco después embarcaciones de vela
1700 AEC: Fundición de hierro
1500 AEC: Monedas
entre 600 y 300 AEC: mecanizar procesos agrícolas, tornillo elevador de agua, molinos,
105: papel

edad media el crecimiento parece detenerse, lo que se suele achacar al oscurantismo religioso de la época
luego llega la era de los grandes inventos

1589: Sir John Harrington inventa el inodoro con depósito.
1712 máquina de vapor
1748 refrigerador moderno
1796 vacunas
1842 Anestésicos
1850 Electricidad
1864 Pasteurización
1928 Penicilina

ANALISIS EDA

Ahora vamos a realizar un Análisis exploratorio del dataset principal, que he obtenido de Naciones Unidas.

Se compone de 64 columnas y 20596 filas. Y contiene, aparte de las diferentes etiquetas de identificación de regiones de estudio, el año de estudio y datos sobre tasas demográficas diversas, como la población total, esperanzas de vida, datos de fallecimientos, tasas de fecundidad, etc.

En primer lugar hemos tratado los datos nulos.No he tenido complicaciones. Faltan etiquetas referentes a regiones globales, como era de esperar, si bien he tenido algún problema con la etiqueta ISO2 de Namibia, que es NA, y como sabemos, pandas identifica NA como un valor nulo al leerlo de un excel o csv. lo hemos cambiado y listo.

Al parecer también había algún dato como '…', que hemos sustituido por '-8'.

Siguiendo la dinámica, hemos convertido todos los tipos a tipo numérico flotante donde se podía, para poder analizar más tarde en una matriz de correlación.

Después de eliminar las variables referentes a etiquetas que hacían falta, he identificado la variable objetivo como total de la población en 1 Julio, ya que todas las mediciones se realizan en esta fecha y he comprobado mediante un test de Shapiro si era Normal. Y efectivamente, la muestra sigue una gaussiana.

Como todavía hay demasiadas variables para observar, he decidido eliminar del estudio exploratorio aquellas que aportan una información lineal con otra y tengan menor correlación con la variable objetivo la Población total.

Un mapa de calor de la correlación entre ellas, me ha permitido llegar a conclusiones variopintas. Aquellas que más me han llamado la atención son:

- El ratio de sexo no interfiere en ninguna otra variable
- Cuanto más alta es la mediana de edad, más baja es la tasa de cambio Natural y la Tasa de nacimiento
- La mediana de la edad de la mujer en el parto esta directamente correlacionada con el sexo del bebé
- Cuando hay más mujeres jóvenes (entre 15 y 19 años) que dan a luz, más mujeres mueren y mayor es la población total.
- La tasa de mortalidad infantil está muy y negativamente correlacionada con el año de estudio
- La densidad de población no está correlacionada con ninguna otra variable

La verdad es que nos da para hacer otro estudio.

MAPAMUNDI

Con este gráfico podemos observar como la densidad de población en 1950 se centralizaba en Europa y Asia, frente a 2021 donde la densidad de población en Europa ha descendido mientras que en Asia y África ha crecido notablemente.

Mientras que si observamos los datos totales de población la cosa cambia. Por ejemplo, Rusia es uno de los países más poblados, pero debido a lo extenso de su territorio, no nos parece significativo en términos de densidad.

TOP 10 PAÍSES

Aquí podemos ver cómo el ranking de los 10 países más poblados del planeta ha ido cambiando desde 1950. Para hacerlo me he inspirado en esta infografía. Realizada por Raul Amorós. nick Routley y Joyce Ma:

https://www.visualcapitalist.com/most-populous-countries-over-50-years/

Las predicciones dicen que India superará a China en 2023.
La mayor tasa de crecimiento observada en los últimos 50 años se encuentra en Nigeria. Se prevé que sobrepase a la población Estadounidense en 2040.
Alemania ha pasado de ser el 8 país más poblado del mundo, al nº 19.

ANÁLISIS POR PAÍSES

Pasamos a un análisis más pormenorizado de las variables referentes a cada país.

La población total de España es, en 2021 de 14 millones y medio de personas, habiendo una densidad de casi 95 personas por kilómetro cuadrado. La mediana de edad es de 43 años, bastante alta, así como también es muy alta la esperanza de vida al nacer. Podemos decir que en España se vive bastante bien.
La tasa de fertilidad, que es lo que se suelen basar las estimaciones de población futura es de 1.28 hijos por mujer, lo que está muy por debajo de la tasa de reemplazo que es de 2.

Se observa en la pirámide de población, una base más delgada, donde corresponde a gente más joven, con una aglomeración en torno desde los 65 a los 40 años, sin distinción aparente por sexos.
Esto nos demuestra que somos una población envejecida. Veremos qué retos nos supone esto en un futuro. La cosa cambia si visualizamos datos de 1950.
Si observamos la tasa de mortalidad segregada por rangos de edad, vemos que el 95% de los fallecimientos corresponden a personas mayores de 60 años.

En la gráfica de crecimiento total de la población, vemos que ha ido descendiendo la cantidad de nacimientos a lo largo de los años, mientras que la mortalidad se mantiene constante, pese al envejecimiento generalizado. Podemos definir que el crecimiento se sostiene gracias a la migración.
Los factores que definen a España como un país de migración es, la calidad de vida, pertenencia a la comunidad europea y la suavidad del clima.

2008 : crisis económica - la gran mayoría de emigrantes son personas que habian migrado y deciden volver a su país de origen.

https://www.elmundo.es/espana/2013/12/10/52a6ef4d61fd3d67268b456f.html

Todas ellas se corresponden con visualizaciones propias de países con mayor desarrollo económico.

Si observamos un país con menor tasa de desarrollo económico, como Nigeria, vemos que las cosas son bastante diferentes.

extra* Emirátos árabes unidos - diferencia entre géneros

ANÁLISIS POR GRUPOS DE OBSERVACIÓN

En este gráfico podemos ver las diferentes segmentaciones de los datos de Naciones Unidas. Contempla diferentes clasificaciones:
Region, desarrollo, renta, grupos regionales, Subregiones, mundo y grupos especiales.

Si analizamos las diferentes regiones, vemos que en Asia y África el hay un crecimiento notable d ela población mientras que en otras zonas, la población se mantiene constante.

Lo mismo podemos observar al estudiar por nivel de desarrollo. Los paises de menor desarrollo tienen un mayor nivel de crecimiento frente a los que pertenecen a tasas más altas.

Si observamos según el nivel de renta, el crecimiento se concentra entre la franja de paises con renta media.

Extra: Podemos analizar también cómo se distribuye la población en diferentes regiones de un mismo continente: Europa

ESTUDIO ARIMA

Para hacer predicciones sobre series temporales sobre la población global del planeta, sospechamos que necesitamos un modelo de predicción multivariable, pero a modo de ejercicio vamos a utilizar el método ARIMA para intentar hacer una predicción de los próximos 20 años, pese a que este modelo está preparado para hacer predicciones a corto plazo, ya que utiliza el sistema de medias móviles.

Al observar el autocorrelograma, podemos identificar una clara autocorrelación de retardo 1, sin estacionalidad aparente.
ARIMA se proyecta sobre una variable que sea estacionaria, o sea, que no proyecte una tendencia. Como nuestros datos claramente lo hacen, hemos de diferenciar la serie (que básicamente, es restar a cada valor el valor anterior) hasta que lo sea. Para saber cuántas veces hemos de diferenciar la serie aplicamos la prueba de Dickey fuller. Nos dice que la diferenciación óptima es de orden 4 (p valor > 0.05) **D**

Para determinar el AR o **P** observamos la autocorrelación parcial diferenciada y buscamos el primer valor con significancia (que esté fuera del sombreado), en este caso es el 2.

Y por último, para determinar MA o **Q**, o el orden del proceso de medias móviles, observamos el gráfico de autocorrelación diferenciada, que tiene que tener forma sinusoidal (+-)
La autocorrelación pinta en cada rezago o lag el coeficiente de correlación de dos vectores (distancia lag)

Los valores de ACF van de -1 a 1, Un valor próximo a 1 indica una gran correlación entre intervalos, si es próximo a -1 la correlación es inversa (los valores de hoy tienden a subir cuando los de ayer bajan), y uno próximo a 0 significa que las columnas comparadas son independientes = no podemos predecirlos (nada nos dice el valor de ayer respecto al que tenemos hoy).

Normalmente se representarán al principio los valores con mayor significancia,(fuera del sombreado) y escogeremos el primer valor que corta esa tendencia.

Una vez ajustados los valores, realizamos la predicción y vemos como las predicciones hacen una forma extraña (parece que los humanos nos convertiremos en antihumanos por el 2050) así que he supuesto que la serie está sobre diferenciada.

Reajustamos los valores y si vemos una predicción que si bien podía ser más realista, dista mucho de las predicciones realizadas por científicos de datos, con otros modelos de predicción, que nos dicen que llegaremos a los 10 mil millones en 2050.

Como nota, decir que he encontrado una página web, que explica muy bien cómo funciona ARIMA e identificar sus parámetros, que apunta a que, debido a la varianza de los datos en el gráfico diferenciado, posiblemente deba hacer una transformación BoxCox. Me lo tomo como nota para seguir practicando^^ y ver si consigue hacer que se aproxime más.

http://enrdados.net/post/series-temporales-con-arima-i/