**Memoria Proyecto EDA - Competencias educativas: relación con el mercado laboral y situación económica en el mundo**

1. **Limpieza de datos**

Lo primero que hice fue cargar un dataset de la OIT (Organización Internacional del Trabajo) bastante completo con un montón de variables del mercado laboral para cada país. La limpieza y ordenación de este dataset acabó tomando la mayor parte de mi tiempo (de forma bastante inútil como explicaré más adelante).

Este dataset tenía una forma que no me gustaba, ya que cada fila era una combinación variable + país, cuando lo que yo pretendía que cada fila fuera un país y cada columna una variable, con la dificultad añadida de que quería un multiindex de columnas con valor total, valor para hombres y valor para mujeres de cada variable. Con ayuda de Ander, un melt y un pivot acabé sacando esto.

Después, cargué dos datasets del World Bank, uno con los datos de PISA y otro con un par de variables de índole económica como los ingresos per capita y el coeficiente Gini y les hice un merge. Añadí un tercer dataset con la población e hice el merge con el dataframe que había resultado estos dos. Eliminé los países con una alta proporción de nulos (debido a que no participaban en el Informe PISA, aquí se ve bien:

Chart, histogram

Description automatically generated Text

Description automatically generated

Por último, para hacer mi merge entre el dataframe con los datos del Worldbank y el de la OIT necesitaba claves comunes, pues muchos países no tienen el mismo nombre en función de la fuente (Corea vs República de Corea, Vietnam vs Viet Nam… cosas así.

Aquí hice un poco de trampa: exporté un Series con los nombrez de los países del df de la OIT y otro con los nombres y códigos del World Bank (en este venían codigos), hice un VLOOKUP para rellenar los códigos de los países del Worldbank para los nombres del df de la OIT y el resto los rellené a mano hasta tener las mismas claves para los mismo países.

Desde aquí, subí esto como dataframe, le hice merge con el de la OIT y de esa forma incorporé las claves del World Bank en el dataframe de la OIT, lo que hizo facilitó el merge entre los datos de la OIT y los del World Bank.

Limpié este dataframe de columnas irrelevantes, lo ordené y mediante una mezcla de bucles y copy’paste le saqué el nombre de sus columnas para crear un multiindice en función de si la columna se refería a hombres, mujeres o al total. Así pude acto seguido separar el dataframe en 3: uno con las variables generales (dftotales, en el que hago gran parte de mi análisis), otro con las referidas a mujeres y otro con las de los hombres.

Durante el curso del proyecto creé unas cuantas variables nuevas para el dataframe para facilitar el análisis de los resultados PISA:

-Un índice ponderado de nota media PISA que recogiera los 3 exámenes (Mates, Lectura, Ciencia), que se compondría solo de Mates y Ciencia en el caso de España puesto que los resultados de lectura para nuestro país no fueron publicados por [irregularidades](https://www.oecd.org/pisa/data/PISA2018Spain_final.pdf) -tema interesante.

-Variables de diferencias porcentuales entre percentiles por país del estilo:

(P95-P5)/P5

-Variables de diferencias porcentuales entre medias por género, del estilo:

(Media hombres mates-Media mujeres mates)/ Media mujeres mates

1. **Visualizaciones:**

-Dash con:

-Mapamundi de mi muestra por ingresos

-Mapamundi de mi muestra por resultados PISA

-Barchart con resultados PISA por país de menor a mayor

-Boxplots distrubuciones de notas medias por exámenes (mates/lectura/ciencia) de PISA

-Correlaciones entre las notas medias de estos exámenes

-Scatterplot ingresos vs resultados PISA con color = continente

-Jointplot ingresos vs resultados PISA

-Heatmap: diversas variables PISA vs ingresos

-Heatmap: correlaciones más fuertes para media PISA

-Scatterplot desigualdad económica vs desigualdad educativa

-Histograma distribución diferencias educativas por género y por asignatura

-Barplot con diferencia media

-Heatmap correlación diferencias educativas por género

1. **Comentario:**

El trabajo me ha costado bastante porque he tardado muchísimo en elegir el tema (y de hecho lo he cambiado varias veces) he gastado días y días en hacer limpiezas y reorganizaciones de datos completamente innecesarias (la base de datos del World Bank me hubiera permitido sacar todas las variables que quería de una vez).

A penas he tenido tiempo para analizar cosas y pintar gráficos más currados o hacer un dashboard más elaborado y si bien es cierto que noto que he machacado el Pandas, me sigo notando bastante verde en comparación respecto al manejo de las librerías de visualización.

Al haber pasado tanto tiempo limpiando y puliendo la forma de los datos, se me han ido perdiendo algunas piecitas de código por el camino y noto que es muy mejorable en cuanto a presentación.

En resumen, pese a que he disfrutado partes del trabajo y he aprendido cositas (aunque no lo suficiente para las horas que le he echado), no estoy satisfecho con lo que he hecho y ya tengo ganas de que llegue el proyecto de ML para enfrentarme de nuevo a otro EDA, espero que de forma más eficiente.

1. **Fuentes de datos:**

* [**https://databank.worldbank.org/home**](https://databank.worldbank.org/home)
* [**https://ilostat.ilo.org/data/country-profiles/**](https://ilostat.ilo.org/data/country-profiles/)