PRACTICA 2: LIMPIEZA Y VALIDACIÓN DE LOS DATOS.

Carlos Lavado Mahia, Dionisio González Jiménez

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset elegido originalmente de kaggle, *"country_profile_data.csv"*. consiste en la muestra de varios indicadores macroeconómicos (GDP, tasa de crecimiento GDP, GDP per cápita, agricultura, industria, servicios, empleados en varios sectores, desempleados, entre otros) a nivel global durante el año 2017.

En particular, queremos encontrar patrones mundiales para explicar cómo influyen diferentes factores sobre el Producto Interior Bruto de un país. Por ejemplo, quisiéramos ver cómo contribuye cada uno de los sectores económicos al PIB por cápita.

Por otro lado, también nos gustaría emplear los datos poblacionales para poder obtener conclusiones demográficas a través de las distintas regiones del mundo, comparándolas con los indicadores económicos.

En definitiva, las principales preguntas que queremos responder son las siguientes: ¿Cómo influyen los factores económicos y demográficos sobre la riqueza de un país? ¿Puede esto ser estudiado a través de técnicas estadísticas?

2. Integración y selección de los datos de interés a analizar.

- **Integración:** preguntar a Diego -> según el material de estudio, integración consiste en la fusión de dos datasets. ¿Debemos seleccionar otro dataset, o simplemente decir que la usuaria de Kaggle ha sido la que ha realizado este proceso?

- **Selección:** el proceso de selección ha consistido en analizar qué columnas del dataset eran interesantes para realizar posteriormente nuestros análisis debido a que tenían muchas columnas que no las hemos considerado relevantes para llevar a cabo nuestros estudios.

  Las columnas elegidas en el proceso de selección han sido las siguientes:

  - **Country**→ nombre del país.
  - **Region**→ nombre de la región.
  - **Population in thousands**→ población en miles de personas.
  - **Population density**→ población por kilómetro cuadrado.
  - **GDP**→ Producto Interior Bruto (PIB).
  - **GDP growth rate**→ tasa de crecimiento del PIB.
  - **GDP per capita**→ PIB per cápita.
  - **Economy: Agriculture**→ porcentaje que aporta el sector agricultura al Valor Agregado Bruto (VAB).

- **Economy: Industry**→ porcentaje que aporta el sector industrial al VAB.
- **Economy: Services and others activities**→ porcentaje que aporta el sector servicios y otras actividades al VAB.
- **Employment: Agriculture**→ porcentaje de empleados en agricultura.
- **Employment: Economy**→ porcentaje de empleados en el sector industrial.
- **Employment: Services and others activities**→ porcentaje de empleados en el sector servicios y otras actividades.
- **Population growth rate**→ tasa de crecimiento de la población (media anual).
- **Urban population growth rate**→ tasa de crecimiento de la población urbana (media anual).
- **Education: Government expenditure**→ gasto público en educación (medido en porcentaje).

Posteriormente, hemos transformado varias columnas de datos, ya que varias de las variables del dataset estaban en formato ***character***; por lo que han sido convertidas a formato numérico para facilitar sus análisis posteriores.

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Tras una revisión de las columnas de datos seleccionados, se ha descubierto que **muchas disponían del valor -99 en descontadas ocasiones**. Dada la repetición de este suceso, se ha llegado a la conclusión de que estos campos han sido marcados por la usuaria original de Kaggle al tratarse de valores perdidos.

Nuestra solución para arreglar estos valores en las columnas numéricas, ha sido el de asignar, en cada campo, la **media de observaciones correspondientes a la región del dato**. De este modo, por ejemplo, para tratar el valor perdido de crecimiento de población anual en Bielorrusia, se ha computado la media de crecimientos poblacionales en Europa Oriental en 2017, asignándose este valor al campo.

3.2. Identificación y tratamiento de valores extremos.

Los valores extremos han sido identificados a través de la función *box-plot* incorporada en R, en concreto a través del parámetro *$out*.

Sin embargo, lo que hemos observado mayoritariamente, ha sido que **los valores extremos eran razonables**, por lo que no hemos considerado necesaria su modificación. Por ejemplo, el archipiélago Tokelau presentaba un valor del 0% de población urbana, mientras que a Macao (China) se le asignaba uno del 100%. Tras unas comprobaciones manuales a través de nuestros navegadores, vimos que estos son datos oficiales, a pesar de considerarse obviamente valores extremos por la función *box-plot*.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

En primer lugar, queremos realizar una **regresión lineal múltiple**, para observar cómo influye cada uno de los tres sectores económicos (agricultura, industria y servicios) sobre el PIB por cápita. Para ello, seleccionaremos tanto la columna del PIB por cápita, como las seis columnas específicas relacionadas con los sectores económicos (las 3 con el peso sobre el VAB de cada sector, y las 3 con el porcentaje de empleo de cada sector).

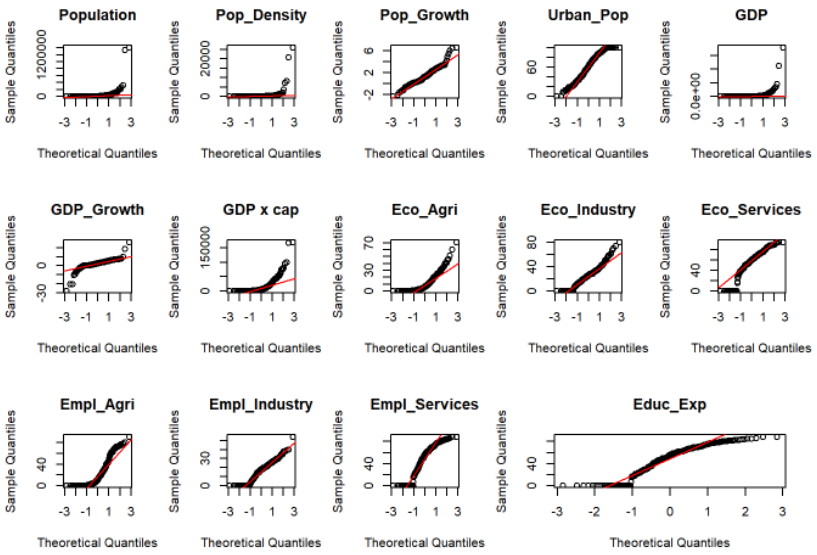
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

<u>Comprobación de normalidad</u>

Todas las columnas seleccionadas al inicio de la actividad, han sido sometidas a una inspección visual para valorar si siguen una distribución normal.

En primer lugar, se han realizado gráficos de densidad, para comprobar si la forma representada podría asimilarse a la que caracteriza a una distribución normal (montaña simétrica). A través de este análisis visual, hemos comprobado que muchos gráficos presentaban cierta asimetría. Sin embargo, podía entreverse como los valores extremos que habíamos dejado al considerar razonables estaban contribuyendo a esto significativamente.

Por ello, hemos procedido a realizar QQPlots, con tal de investigar si la mayoría de puntos de cada columna se alinean con la línea que representa la distribución normal. A través de esta inspección, hemos podido comprobar como, en la mayoría de casos, se trataba únicamente de puntos en los extremos de la distribución que se salían del patrón.



Por ello, y considerando que nuestra muestra es de hecho de 229 observaciones (superando con creces el umbral delimitado por el Teorema Central del Límite), hemos llegado a la conclusión de que nuestros datos siguen una distribución normal.

Preguntar a Diego si él consideraría que, por ejemplo, GDP_xcap sigue una distribución normal. De momento, hemos aplicado sqrt para disminuir la skewness en la regresión lineal.

Hemos prescindido de emplear tests como el de Shapiro-Wilk, ya que comprobamos que presenta resultados sesgados al ser muy estricto con su comprobación de la normalidad.

Comprobación de homocedasticidad

- **Regresión Lineal**
  La homocedasticidad para este análisis, ha sido comprobada a través de los gráficos Residuals vs Fitted del modelo de regresión lineal. En estos gráficos, se puede observar una distribución más o menos heterogénea, aunque con un mayor volumen de puntos en el extremo derecho.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Regresión Lineal

A través de la regresión llevada a cabo mediante la función *lm()*, se ha descubierto que las siguientes variables influyen en la variabilidad del PIB por cápita: peso económico de la industria, peso económico de los servicios, así como empleados en agricultura, en industria y en servicios. Únicamente el peso económico de la agricultura.

El coeficiente de correlación R del modelo, indica que se explica un 60% de variabilidad del PIB por cápita. Se trata, por tanto, de una regresión fuerte. En caso de eliminar el impacto provocado por la adición de nuevas covariables, el coeficiente de correlación R ajustado sigue siendo fuerte, de prácticamente el 59%.

```
##
## Call:
## lm(formula = gdp_xcap ~ eco_agri + eco_industry + eco_services +
##     empl_agri + empl_industry + empl_services, data = new_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -96.291 -29.951  -5.848  20.807 264.512
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.9316    11.2567   0.260  0.79477
## eco_agri       -0.6105     0.4564  -1.338  0.18239
## eco_industry    1.1494     0.2621   4.386 1.78e-05 ***
## eco_services    1.6664     0.1563  10.665  < 2e-16 ***
## empl_agri      -1.2033     0.2405  -5.003 1.15e-06 ***
## empl_industry  -1.3389     0.4626  -2.894  0.00418 **
## empl_services   0.4256     0.1775   2.398  0.01732 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 49.24 on 222 degrees of freedom
## Multiple R-squared:  0.6001, Adjusted R-squared:  0.5893
## F-statistic: 55.53 on 6 and 222 DF,  p-value: < 2.2e-16
```

Modelo

^GDP x capita = 2,9316 - 0,6105eco_agri + 1,1494 eco_industry + 1,6664 eco_services - 1,2033empl_agri - 1,3389 empl_industry + 0,4256empl_services

5. Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.