Actividad 1: Preproceso de los datos

Enunciado

Análisis Estadístico - Semestre 2025.1

Índice

Introducción

En esta actividad realizaremos el análisis exploratorio y el preproceso del conjunto de datos Global Sustainability (https://www.kaggle.com/datasets/truecue/worldsustainabilitydataset/data) el cual agrupa información sobre el desempeño de 173 países en relación con diversos indicadores de sostenibilidad a lo largo de un período de 19 años. El conjunto de datos se ha generado a partir de la fusión de múltiples fuentes internacionales, entre ellas el World Bank DataBank. Incluye un total de 54 variables, que abarcan diferentes dimensiones de la sostenibilidad:

- Económica: Se proporcionan indicadores como el porcentaje de exportaciones del país, producto nacional bruto, crecimiento del ingreso nacional bruto ajustado per cápita, ahorro neto ajustado, etc.
- Social: Contiene indicadores como la esperanza de vida al nacer, índice de Gini (desigualdad de ingresos), educación, etc.
- Ambiental: Indicadores que tienen en cuenta la degradación de recursos naturales y los daños ambientales.

Este conjunto de datos nos ofrece múltiples posibilidades para explorar los indicadores vinculados a los objetivos de desarrollo sostenible (ODS) de las Naciones Unidas. En la actividad, exploraremos estos indicadores y preprocesaremos el conjunto de datos para su posterior análisis.

Debido a la extensión del conjunto de datos, nos centraremos sólo en algunas de las variables del conjunto.

Criterios de verificación y de normalización de las variables:

A continuación se muestran los criterios con los que deben limpiarse los datos del conjunto:

- Revisad la naturaleza de las variables (texto, categórica, numérica). En caso de que el tipo de variable que ha otorgado R no coincida con el tipo que le correspondería, deberá corregirse.
- Los nombres de las columnas deben cambiarse según el fichero "Data Dictionary.xlsx" proporcionado.
- En las variables de tipo numérico, el separador decimal es el punto.
- Los valores perdidos deben tener el valor NA.
- Las variables numéricas con valores NA deben imputarse según las indicaciones que se dan más abajo.
- Para mantener la consistencia con el diccionario de datos, se mantendrán los nombres de las variables en el fichero de datos. Se puede mostrar la información con nombres más comprensibles en las tablas y gráficos si es necesario.

A tener en cuenta para realizar la actividad:

- Se valora la precisión de los términos utilizados (es necesario utilizar de forma precisa la terminología de la estadística).
- Se valora también la concisión de la respuesta. No se trata de hacer explicaciones demasiado largas o documentos muy extensos. Hay que explicar el resultado y argumentar la respuesta a partir de los resultados obtenidos de forma clara y concisa.
- Mostrad en el documento sólo lo imprescindible y que da respuesta a la pregunta planteada. No es necesario que se muestre todo el proceso de exploración que se ha realizado.
- No se pueden realizar listados completos del conjunto de datos en la solución. Esto generaría un documento con cientos de páginas y dificultaría la revisión del texto. Para comprobar las funcionalidades del código sobre los datos, se pueden utilizar las funciones head y tail que sólo muestran unas líneas del archivo de datos.

- Como objetivo secundario, esta actividad pretende desarrollar el conocimiento del lenguaje R, aprovechando al máximo sus características. En este sentido, siempre que sea posible, usad código optimizado y elegante. Por ejemplo, para aplicar un mismo código a distintas variables, usad las funciones de la familia apply o funcionalidades de dplyr. Evitad la inspección visual (manual) siempre que sea posible. Para mostrar las tablas, se puede utilizar la librería kableExtra.
- Para dudas técnicas sobre R y el entorno de programación, tenéis a vuestra disposición el aula Laboratorio R.

Requisitos para la entrega:

- Es necesario entregar el archivo Rmd y el archivo de salida (PDF o html). El archivo de salida debe incluir: el código y el resultado de la ejecución del código (paso a paso).
- Para facilitar la corrección, los archivos deben entregarse por separado en el aula. Es decir, en lugar de un único archivo comprimido, entregad cada archivo por separado.
- Hay que respetar la misma numeración de los apartados que el enunciado.

1 Estructura de los datos

1.1 Diccionario de datos

Cargar los datos del conjunto de datos WorldSustainabilityDataset y del diccionario de datos (Data Dictionary).

A continuación examinad la información que está almacenada en el diccionario, donde se detallan las variables, su código, una breve descripción, a qué ODS (SDG en inglés) pertenecen y la fuente de datos de donde procede.

Para resumir esta información y hacer más comprensible el manejo de los datos a lo largo de la actividad, crearemos dos tablas informativas. En la primera, se deben mostrar las variables y si procede, el ODS asociado. Las columnas de esta tabla deben ser: número de ODS (si procede), código de la variable (indicador) y breve descripción de la variable. Podéis usar la misma descripción que contiene el fichero diccionario. No es necesario traducir ni redactar de nuevo.

En la segunda tabla, mostrad el número de ODS y su descripción. En ambas tablas, ordenad el listado según el número de ODS.

1.2 Fichero de datos

A continuación, prepararemos el fichero de datos para su posterior manipulación. Para ello, cambiad el nombre de las columnas según los códigos establecidos en el diccionario de datos. Seguid la misma nomenclatura que en el diccionario de datos. Por ejemplo, la variable esperanza de vida tiene el código SP.DYN.LE00.IN. En la medida de lo posible, realizad este proceso de forma automática o semi-automática.

Renombrar las columnas pertinentes con los nombres siguientes: Country, CountryCode, Regime, Income, Region. Al finalizar el proceso, mostrad el listado con los nombres de las columnas.

2 Tipos de datos y posibles inconsistencias

Revisar los tipos de los datos y el formato de los mismos. En el caso de inconsistencias en el formato, corregidlas. Seguid las instrucciones que se detallan a continuación.

2.1 Variables cuantitativas

Revisad los valores de las variables cuantitativas y aplicad las transformaciones de formato necesarias. Si hay valores vacíos, se deben substituir por NA. Si se encuentran valores extremos o erróneos o NAs en alguna variable cuantitativa, lo trataremos en apartados posteriores.

2.2 Variables cualitativas

Revisad los valores de las variables cualitativas y aplicad las transformaciones de formato necesarias. Si hay valores vacíos, se deben sustituir por los valores correctos o si no se conocen, por NA. Para mejor estructuración de este apartado, tratad cada variable en una subsección distinta.

2.2.1 Países y códigos de países

2.2.2 Continente

2.2.3 Régimen

2.2.4 Región

2.2.5 Income

3 Valores extremos

Examinamos los valores extremos de algunas variables de interés. En concreto, el índice GINI y las emisiones de gases de efecto invernadero. Para hacer el código más comprensible podéis usar los nombres GINI y GHE respectivamente en los gráficos y tablas.

3.1 Desigualdad (GINI)

Visualizad la distribución del índice GINI entre los distintos países a lo largo de los años. Interpretad el gráfico. Observad si existen valores erróneos o extremos que deban ser procesados. Si lo consideráis oportuno, aplicad una transformación sobre los valores extremos. Podéis escoger la transformación de datos que consideréis más adecuada.

3.2 Emisiones de gas de efecto invernadero

Revisad los valores extremos en la variable GreenHouse Gas Emissions (emisiones de gases de efecto invernadero). De forma análoga al caso anterior, visualizad un gráfico que muestre la distribución de emisiones de gas a lo largo de los años. Interpretad el gráfico y reflexionad sobre la necesidad de aplicar transformaciones. Si procede, realizad las transformaciones oportunas.

A continuación, usaremos la detección de valores extremos para visualizar una lista de los países más contaminantes. Escoged el año 2018 y mostrar la lista de países con emisiones de gases extremadamente grandes. Visualizad esta lista incluyendo el nombre del país y el valor de la variable de mayor a menor emisión.

4 Correlaciones

4.1 Matriz de correlaciones en indicadores de pobreza

Calculad las correlaciones entre las variables correspondientes a los ODS 1,2,3,4 (solo SE.PRM.UNER.ZS), 10 y GDP (producto interno bruto). Mostrad una matriz con el resultado e interpretad.

Nota: para visualizar la matriz, podéis cambiar el nombre de las variables para facilitar la comprensión del lector.

4.2 Correlaciones con esperanza de vida

Examinad específicamente las variables que presentan más correlación con la esperanza de vida. Mostrad una tabla con las variables ordenadas de más a menos correlación. Interpretad.

5 Imputación

A modo ilustrativo, realizaremos la imputación de NAs para la variable esperanza de vida. Revisad en qué años esta variable presenta valores perdidos.

Como veréis el año 2000 no tiene datos. En este caso, realizaremos la imputación con los datos del 2001. Para el resto de años, imputad el valor de la variable a partir de la media de los vecinos más cercanos, usando las cinco variables más correlacionadas con esperanza de vida.

Nota: Podéis usar la función kNN de la librería VIM.

6 Tabla resumen

Calcular las medidas de tendencia central y dispersión (robustas y no robustas) de las variables cuantitativas siguientes: pobreza (SI_POV_DAY1), GINI y esperanza de vida. Se presentarán dos tablas, una con las medidas de tendencia central y otra tabla con las medidas de dispersión. En las medidas de tendencia central, mostrad la media y mediana. En cuanto a medidas de dispersión, mostrad la desviación estándard y la desviación absoluta respecto de la mediana. Debido a la gran variabilidad entre los datos, agrupad los datos por regiones y mostrad tan solo los datos del último año disponibles. En las tablas, podéis usar nombres para las variables que sean más comprensibles.

Nota: Realizad este cálculo de forma automatizada, usando funcionalidades de la familia *apply* o *dplyr*.

Puntuación de la actividad

- Apartado 1 (10%)
- Apartado 2 (20%)
- Apartado 3 (20%)
- Apartado 4 (20%)
- Apartado 5 (10%)
- Apartado 6 (10%)
- Calidad del informe dinámico (10%)