# Análisis Exploratorio del Dataset GDSC en R / Exploratory Analysis of the GDSC Dataset in R

## 📌 Descripción del Proyecto / Project Description

Este repositorio contiene un análisis exploratorio de datos biomédicos utilizando el dataset **GDSC (Genomics of Drug Sensitivity in Cancer)**.  
El objetivo fue integrar datos clínicos, moleculares y farmacológicos de líneas celulares cancerosas para preparar un dataset consolidado que pueda utilizarse en futuras tareas de modelado predictivo en oncología.

This repository contains an exploratory biomedical data analysis using the **GDSC (Genomics of Drug Sensitivity in Cancer)** dataset.  
The goal was to integrate clinical, molecular, and pharmacological data from cancer cell lines to create a consolidated dataset for future predictive modeling tasks in oncology.

## 🧪 Dataset Utilizado / Dataset Used

* **Drug Sensitivity**: LN\_IC50, AUC, RMSE, Z-Score
* **Molecular Data**: Mutaciones, expresión génica, alteraciones en número de copias
* **Clinical Data**: Tipo tumoral (TCGA), descriptores tisulares

Los datos fueron obtenidos del sitio oficial del proyecto GDSC:  
Data sourced from the official GDSC portal: <https://www.cancerrxgene.org>

## 🛠️ Herramientas / Tools Used

* Lenguaje / Language: **R**
* Principales librerías utilizadas / Key R packages:
  + tidyverse
  + readxl
  + ggplot2
  + naniar
  + dplyr

## 🚀 Cómo ejecutar el script / How to Run the Script

1. Clonar este repositorio / Clone this repository:

git clone https://github.com/usuario/tu-repo.git

cd tu-repo

1. Abrir el script project\_gdsc\_analysis.R en RStudio o desde la terminal de R.
2. Ejecutar el código paso a paso o correr todo el archivo para generar visualizaciones y preparar el dataset integrado.

Asegurarse de colocar los archivos de datos (.csv y .xlsx) en la misma carpeta o actualizar las rutas dentro del script.  
Make sure the data files (.csv and .xlsx) are in the same directory or update the file paths in the script.

## 📊 Contenido del Script / Script Contents

* Carga de archivos CSV y Excel
* Análisis de métricas de sensibilidad farmacológica (LN\_IC50, AUC, RMSE, Z-Score)
* Visualizaciones agrupadas por grupos de drogas
* Limpieza de datos y tratamiento de valores faltantes con naniar
* Análisis de distribución de variables categóricas y numéricas
* Integración final de datasets clínicos, moleculares y farmacológicos

## 👩‍💻 Autora / Author

Este proyecto fue realizado por Soledad Varady.  
This project was developed by Soledad Varady.

## 📂 Licencia / License

MIT License