SUMMER INTERNSHIP PROJECT

**Project Title:** Data Visualisation & Analysis using HIV Dataset.

*Chiranjib Poddar* & *Mandira Mitra*

2nd Year, Biochemistry Honours

Asutosh College, University of Calcutta

Project Guide/ Mentor Name: Adrija Das

Period of Internship: 19th May 2025 - 15th July 2025

**Report submitted to**: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata.

**ABSTRACT**

This project focuses on data visualization and regression analysis using an HIV dataset sourced from Kaggle via DeepChem. The dataset presents country-wise HIV infection data over several years, allowing us to observe trends and patterns in the spread of the virus globally. Using Python, we performed basic data preprocessing, visualizations, and applied regression models to analyze relationships between various features, such as infection rates, time, and regional impact. Regression analysis plays a key role in identifying patterns and making predictions from numerical data. It is also commonly used in bioinformatics, such as in tools like DeepChem and AlphaFold, to predict protein structures and biophysical properties, while our project remains focused on statistical exploration and visual trends in the HIV dataset.

**INTRODUCTION**

Human Immunodeficiency Virus (HIV) is a virus that attacks the body's immune system, specifically the CD4 cells (T cells), which help the immune system fight off infections. If left untreated, HIV can lead to Acquired Immunodeficiency Syndrome (AIDS), a stage where the immune system becomes severely damaged and vulnerable to life-threatening infections and cancers.

One of the major concerns surrounding HIV is that it currently has no permanent cure. While antiretroviral therapy (ART) helps in managing the infection and increasing lifespan, millions of people still suffer from limited access to treatment, especially in low-resource regions. The increasing number of global HIV cases remains a serious public health issue, driven by lack of awareness, social stigma, late diagnosis, and uneven healthcare access — particularly in certain parts of the world.

To better understand and analyze such health challenges, datasets are commonly used. A dataset is a structured collection of data that helps in drawing insights, finding trends, and building models in fields like healthcare, finance, and machine learning. Classic examples include the Boston Housing dataset, which is often used to predict house prices based on certain features. Similarly, in this project, we have used the HIV dataset sourced from Kaggle via DeepChem, which compiles HIV-related case data across different countries and regions over several years. The dataset captures the spread and progression of HIV cases over time, making it valuable for data-driven health analysis.

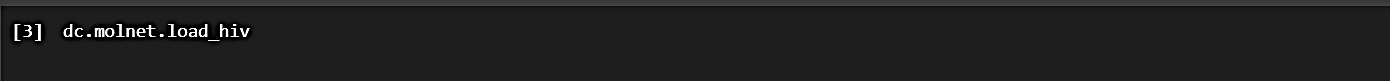
Understanding such data often requires statistical methods like **linear regression**, a predictive modeling technique used to analyze the relationship between a dependent variable and one or more independent variables. In Python, libraries such as scikit-learn make it easy to implement linear regression and visualize results, helping us understand correlations and potential trends in the data. Linear regression is also one of the core techniques used in scientific AI tools like **AlphaFold**, developed by DeepMind. In AlphaFold, regression models are used to predict protein structures by learning from large datasets of amino acid sequences and structural annotations, allowing researchers to estimate physical and functional properties of proteins with high accuracy.

**PROJECT OBJECTIVES**

* To do Exploratory Data Analysis (EDA) of the HIV Dataset
* To construct visual plots from the dataset to help analyze the data more efficiently
* To construct Linear Regression models based on it

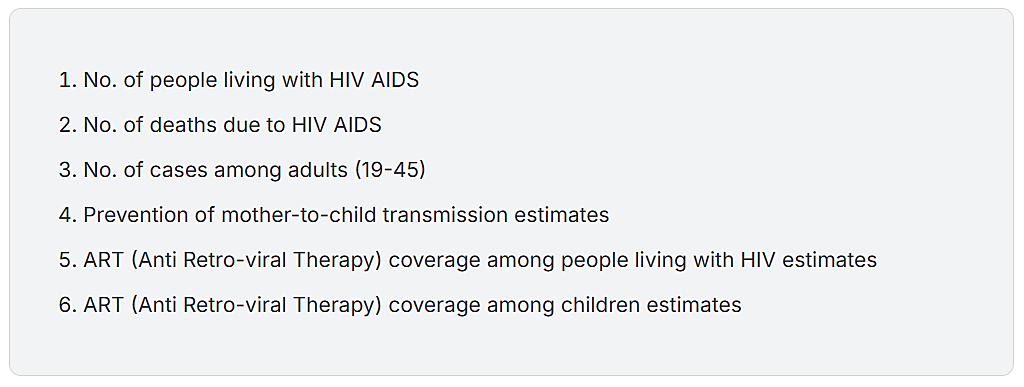
**METHODOLOGY**

1. The HIV dataset was downloaded from Kaggle website as per the instruction from the mentor and loaded in Google Colab, the working platform.

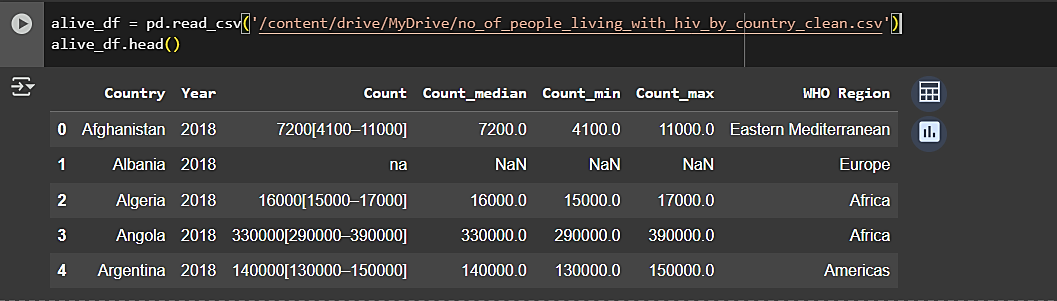


This dataset was originally collected from World Health Organisation (WHO) website and United Nations Educational, Scientific and Cultural Organization (UNESCO) website.

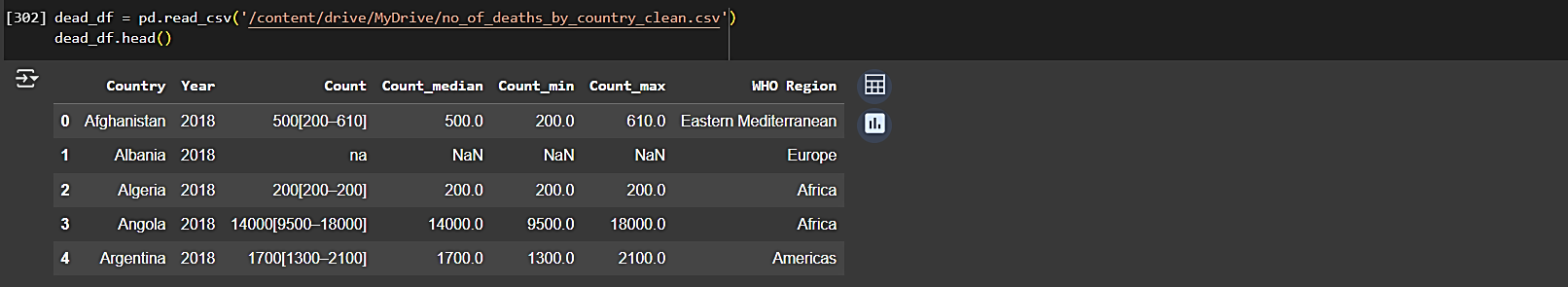
The dataset contains data on:



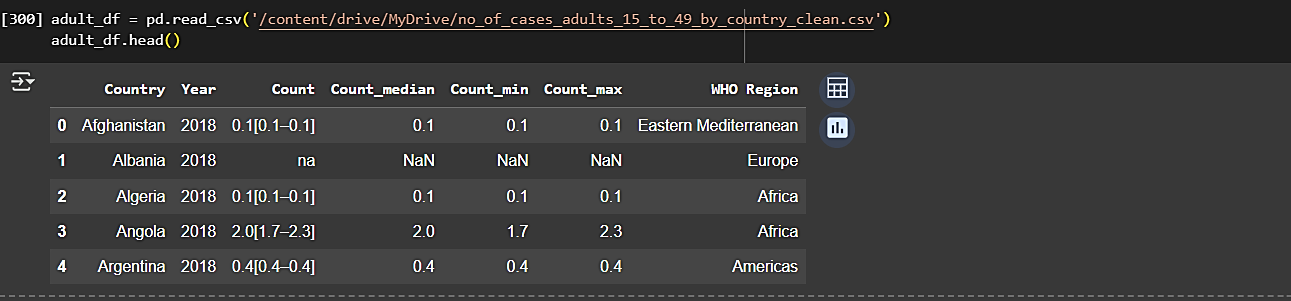
1. Number of people living with HIV stored as alive\_df



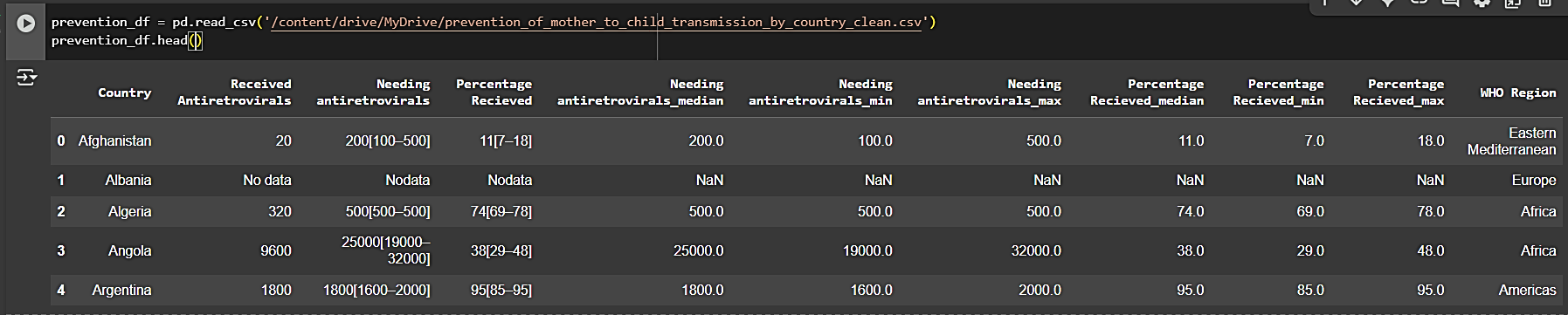
1. Number of deaths due to HIV was stored as dead\_df

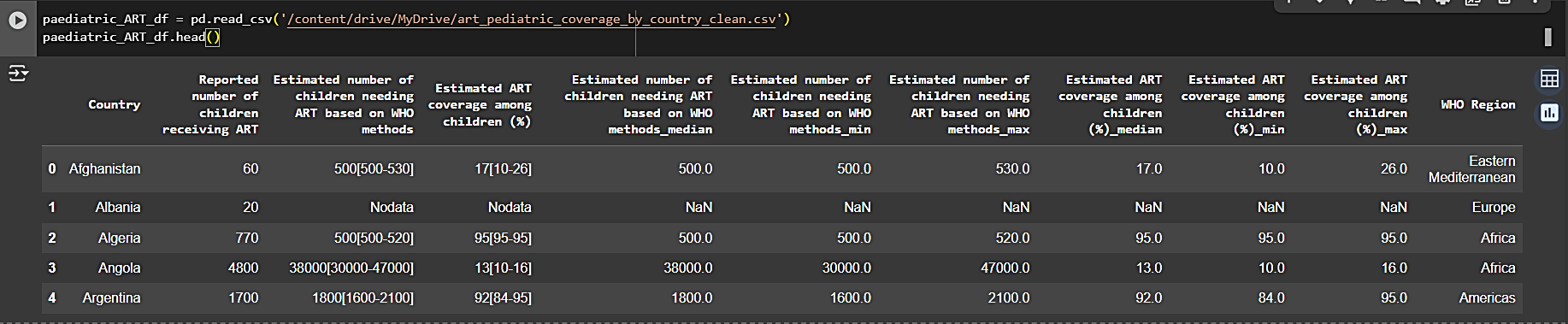


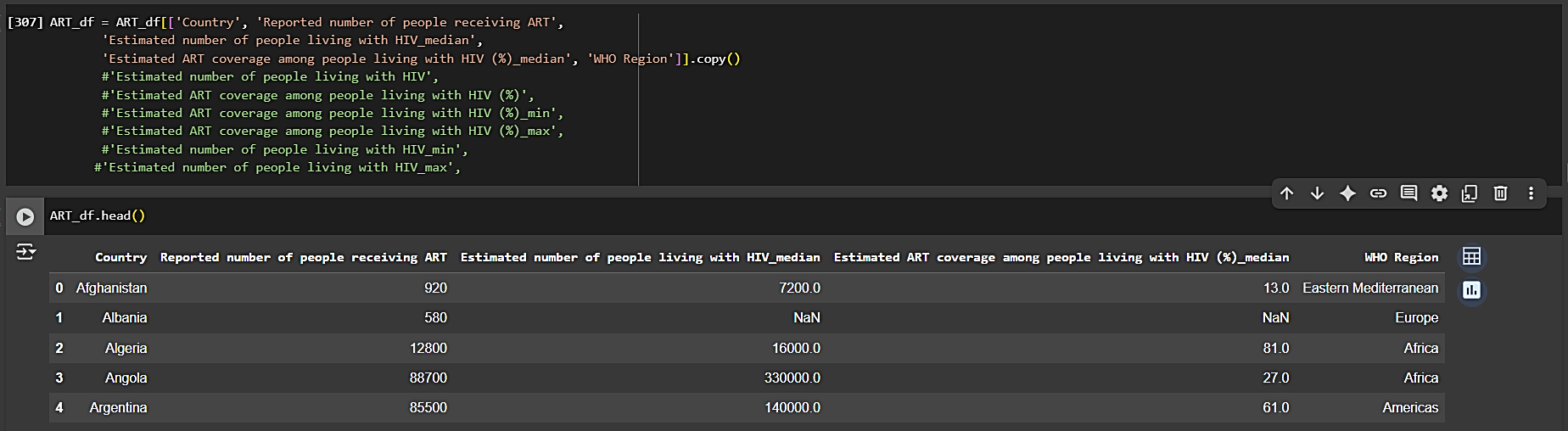
1. Number of cases among adults aged 19 to 45 was stored as adult\_df

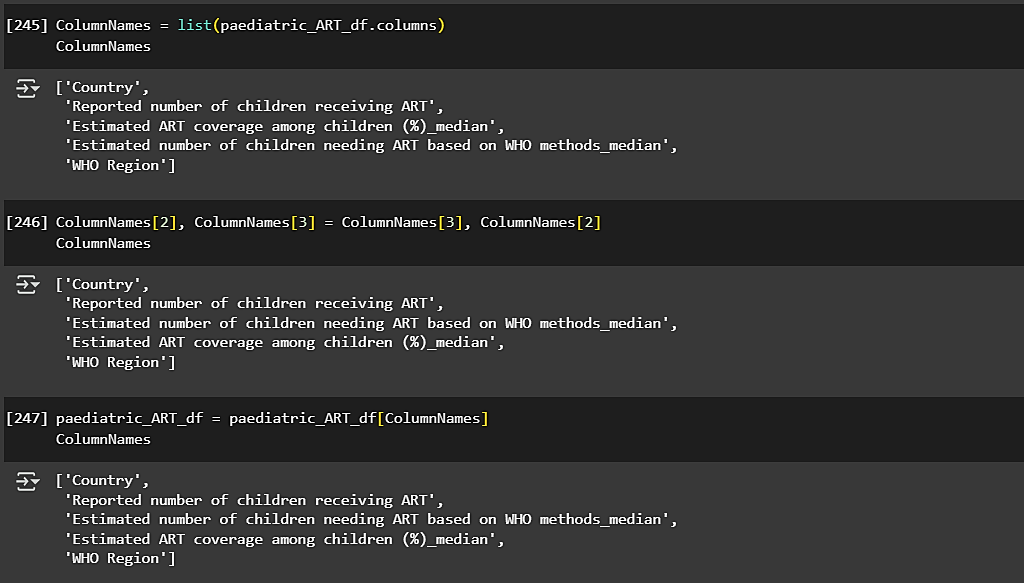
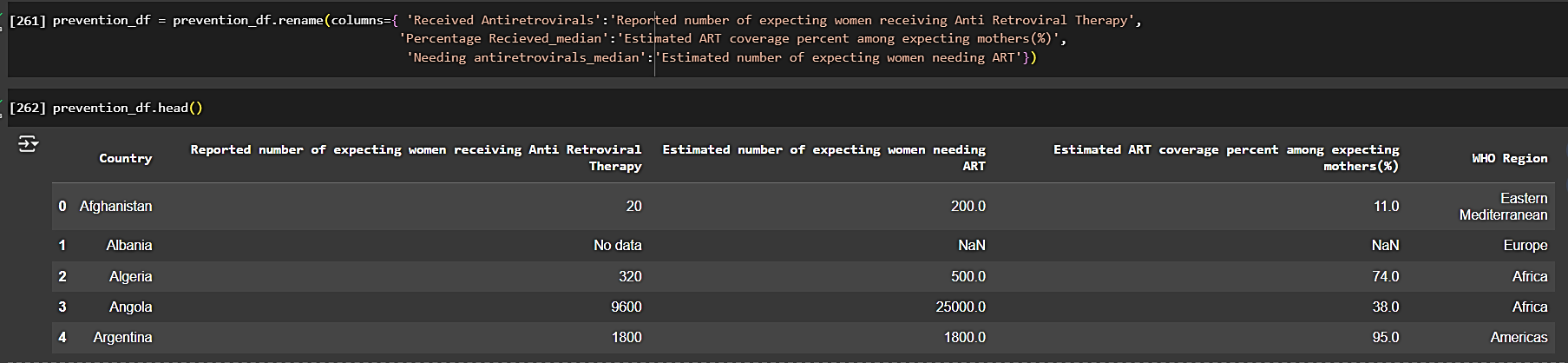


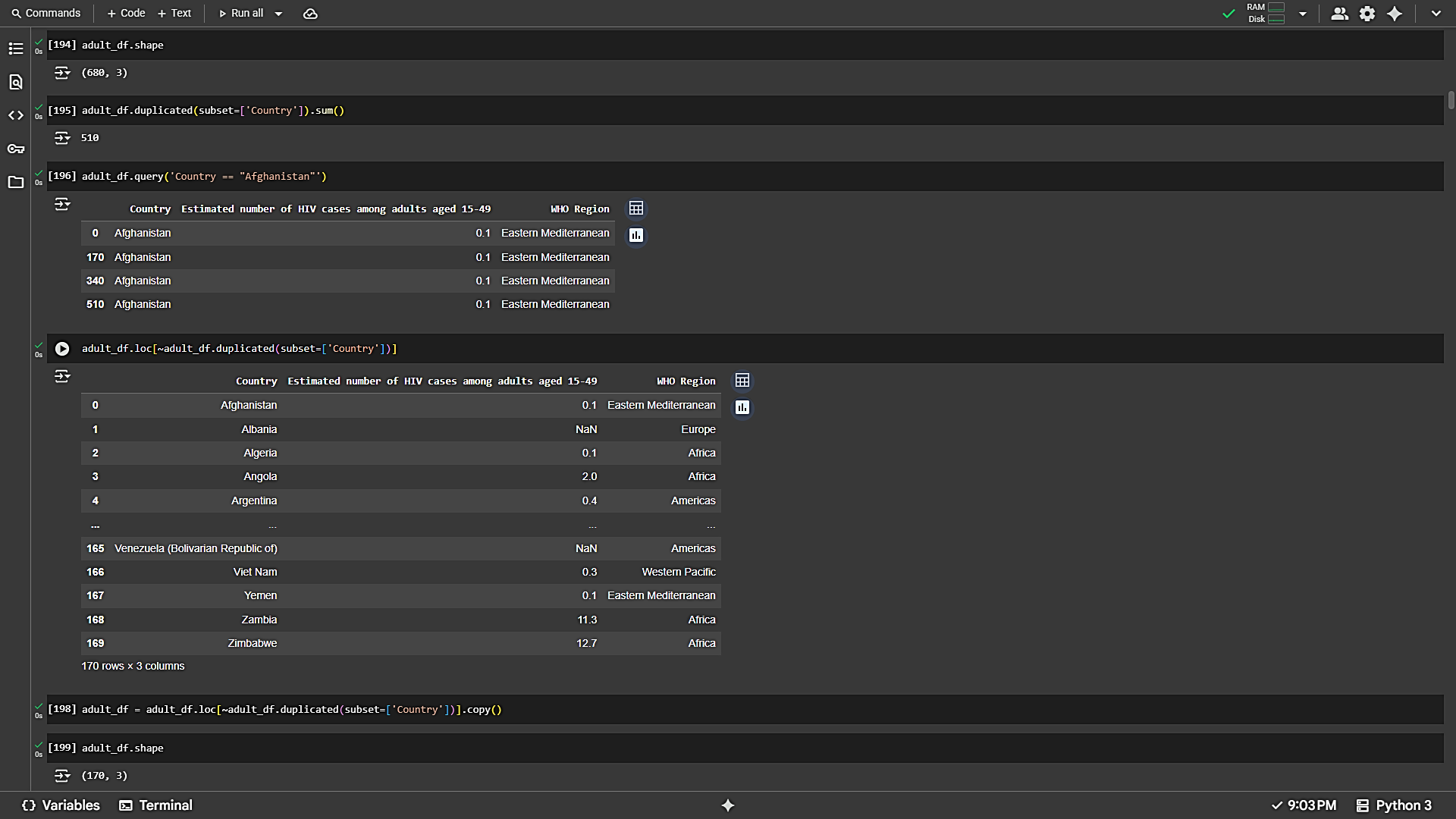
1. Prevention of mother-to-child transmission estimates was stored as prevention\_df



1. ART coverage among people living with HIV estimates was stored as ART\_df
2. ART coverage among children estimates was stored as paediatric\_ART\_df
3. Each metadata was pre-processed separately before merging them into a single data frame. Data cleaning included the following steps:
4. Removal of unnecessary columns



1. Rearranging columns as per requirement
2. Renaming columns to better suit our workflow
3. Removing duplicated rows to condense the data



1. All the six metadata were condensed into one data frame(merged\_df), keeping ‘Country’and ‘WHO Region’constant.

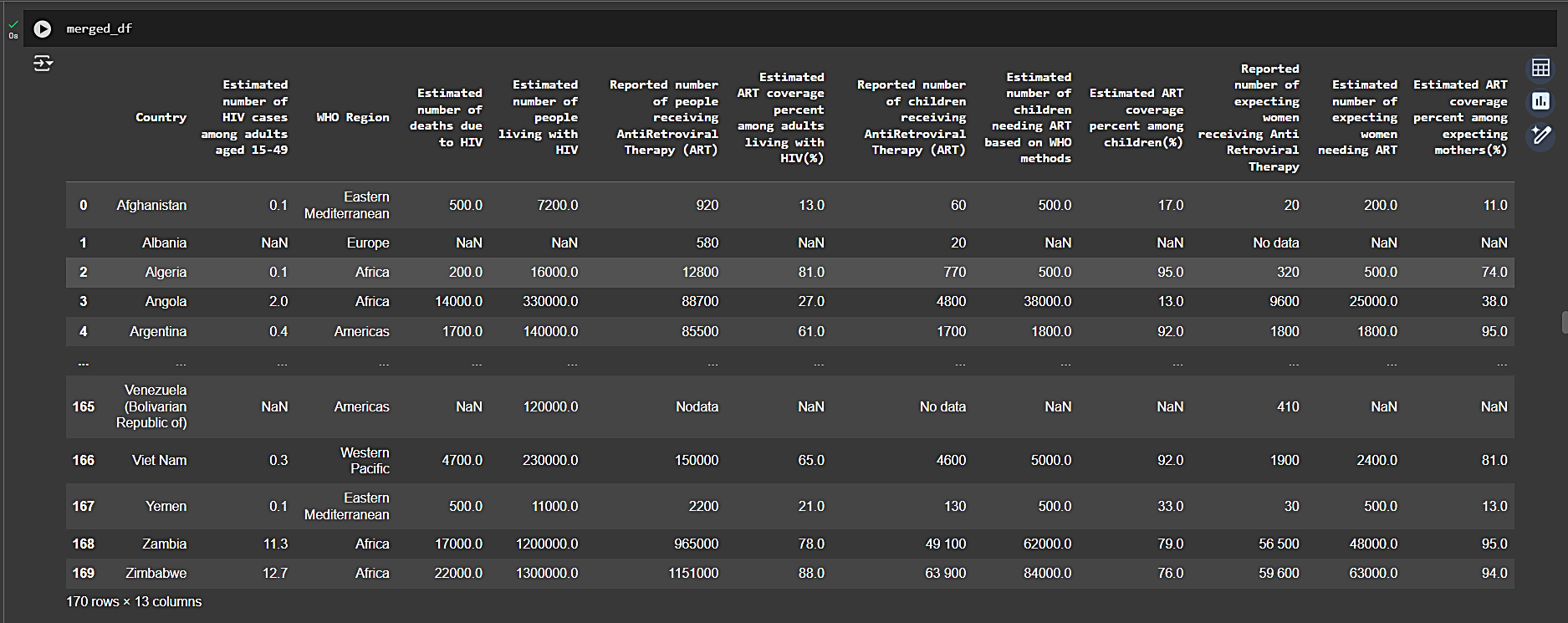
merged\_df = adult\_df.merge(dead\_df, on=['Country','WHORegion'])

.merge(alive\_df, on=['Country','WHO Region'])

.merge(ART\_df, on=['Country','WHO Region','Estimated number of people living with HIV'])

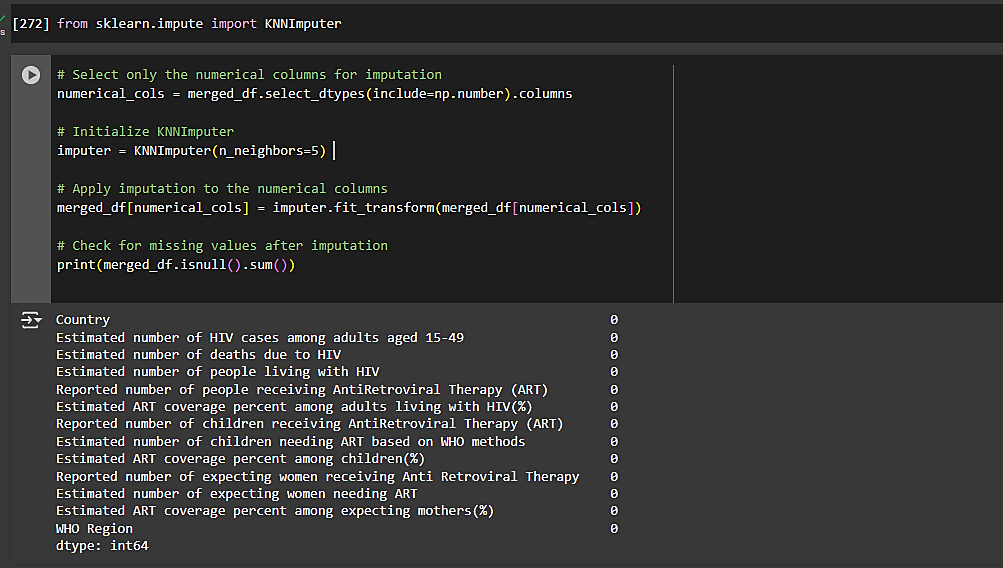
.merge(paediatric\_ART\_df, on=['Country','WHO Region'])

.merge(prevention\_df, on=['Country','WHO Region'])

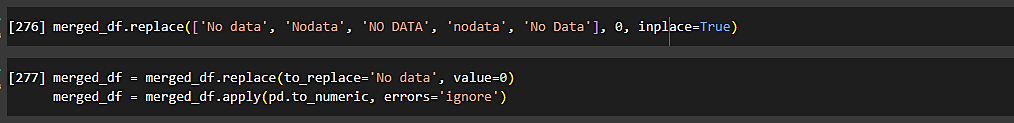


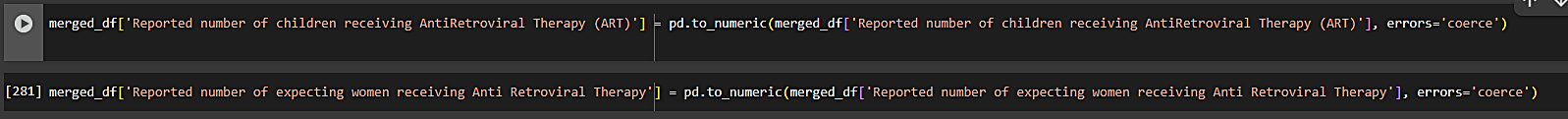
1. The merged\_df was cleaned further.

KNNImputer

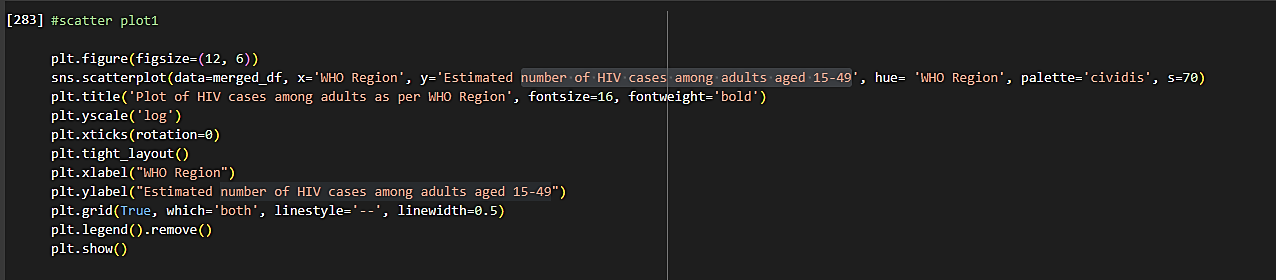
Missing data in the columns were filled using . It is a scikit-learn class used to fill out or predict the missing values in a dataset using n-neighbors (data points) to a data point with a missing value, based on some distance metric, and then using the values of these neighbors to impute the missing value.

Even after imputation, the data frame has entries like ‘No data’ and other string values. All such entries were identified and replaced with ‘0’.

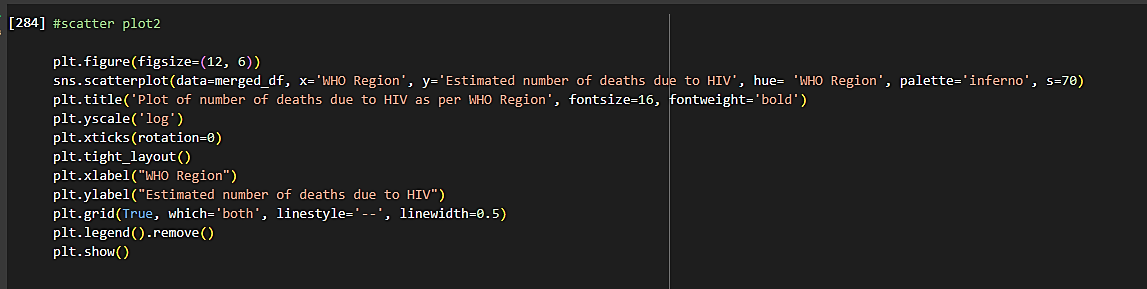




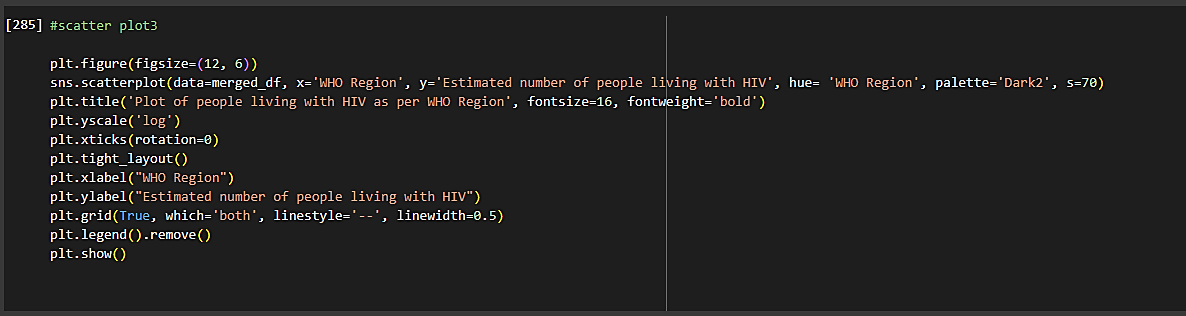
1. The merged data frame was then used to draw visual plots such as scatter plots, heatmap and line plotusing matplotlib and seaborn, the python data visualisation libraries.
2. #scatterplot1shows the variance of ‘number of HIV cases among adults aged 15-49’(on y-axis) with ‘WHO region’(on x-axis).



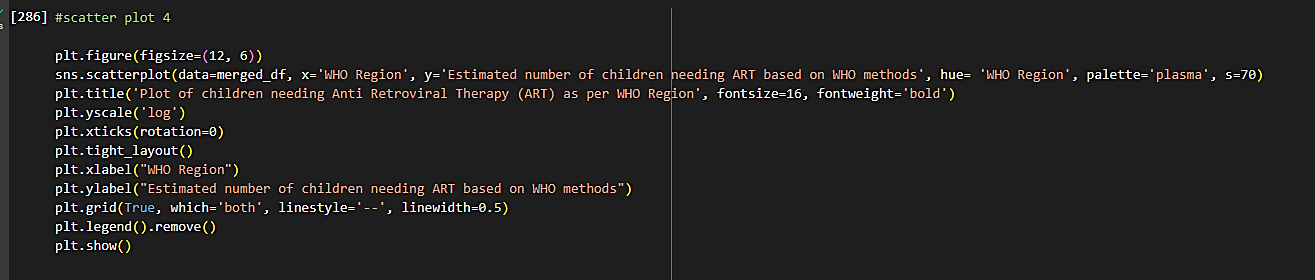
1. #scatterplot2shows the variance of ‘number of deaths due to HIV’(on y-axis) with ‘WHO region’(on x-axis).



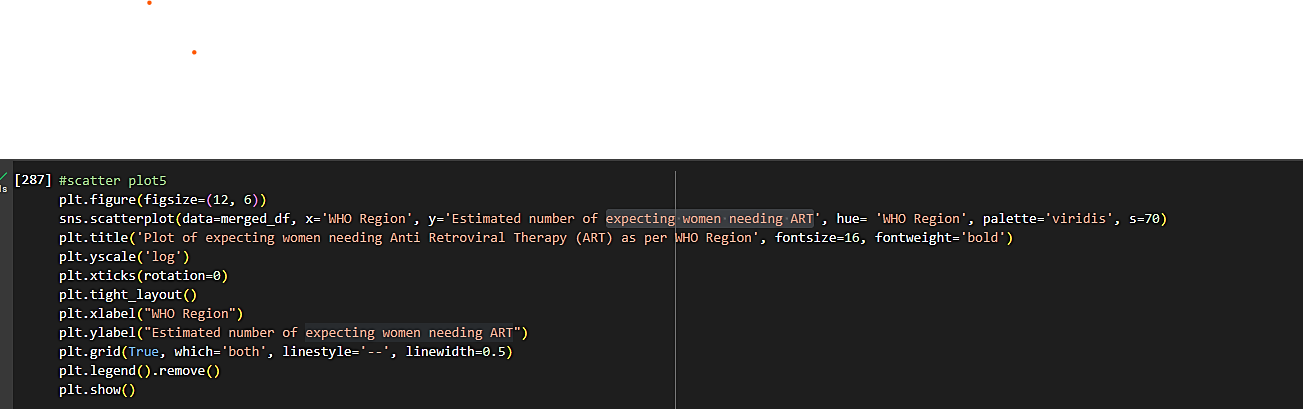
1. #scatterplot3shows the variance of ‘number of people living with HIV’(on y-axis) with ‘WHO region’(on x-axis).



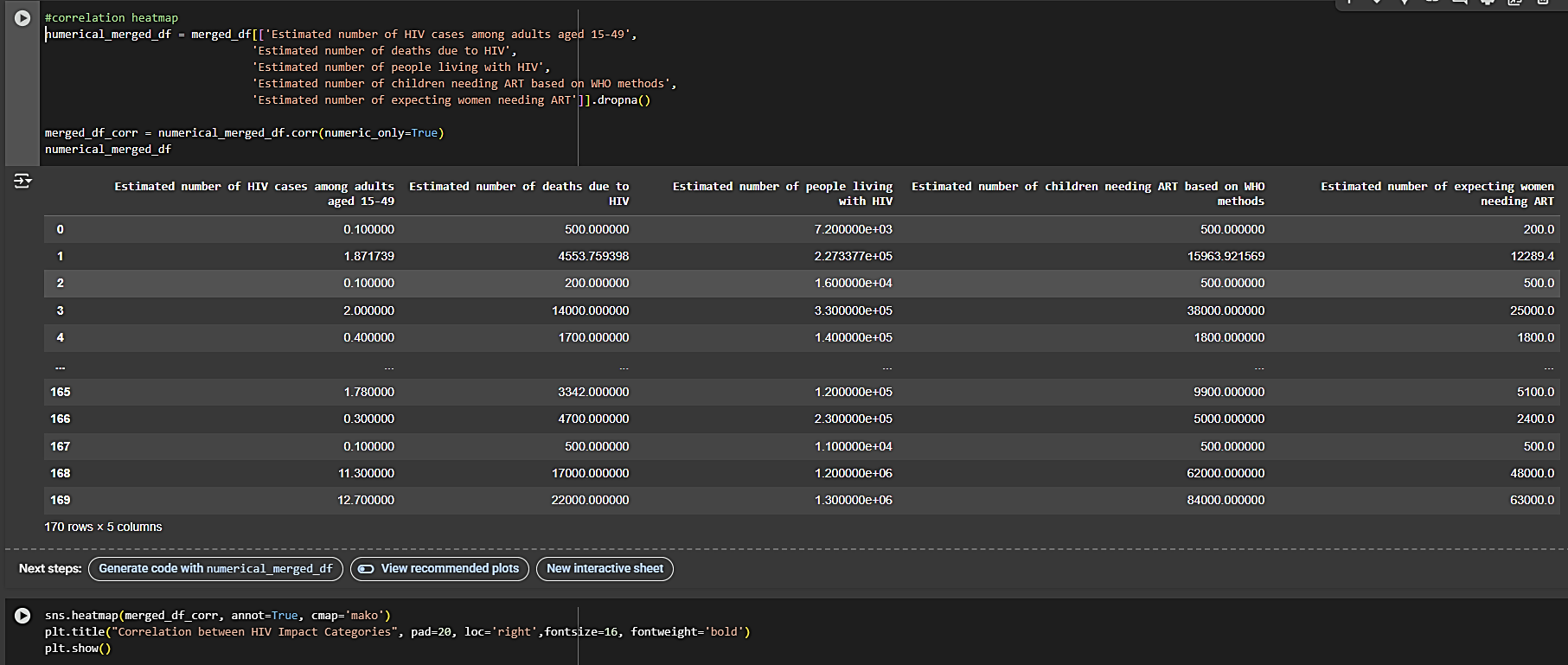
1. #scatterplot4shows the variance of ‘number of children needing ART based on WHO methods’(on y-axis) with ‘WHO region’(on x-axis).



1. #scatterplot5shows the variance of ‘number of expecting women needing ART’(on y-axis) with ‘WHO region’(on x-axis).



1. #correlation heatmap shows the correlation between different population categories affected by HIV.



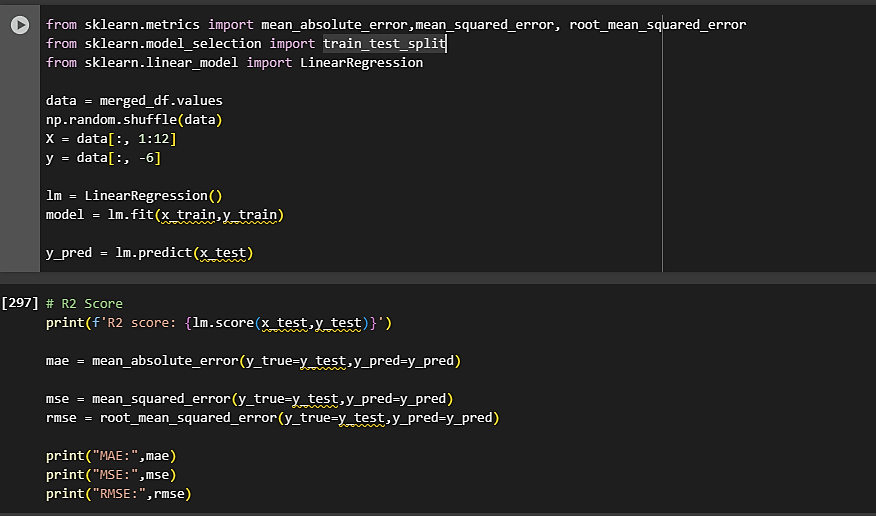
1. #line plot shows the which category of the population is affected most by HIV.



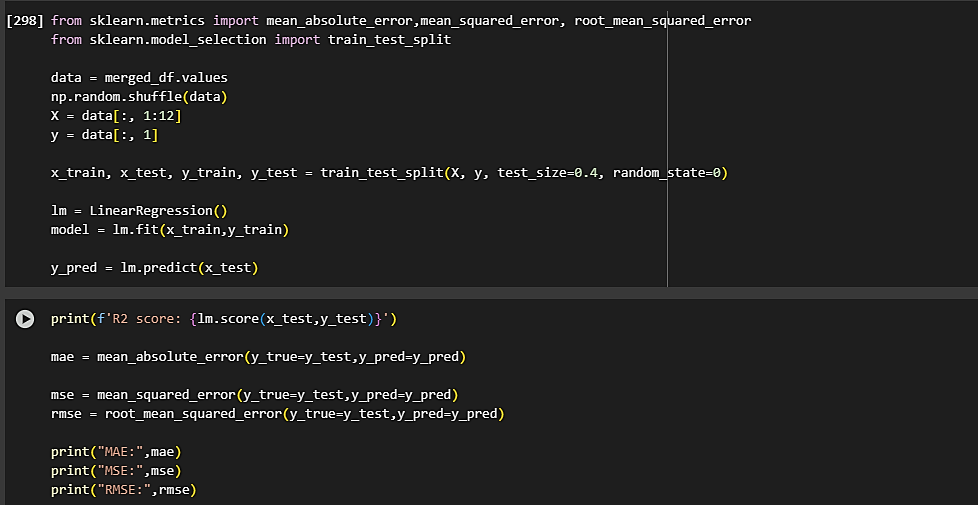
1. Lastly, LinearRegression() was done for analytical modelling. Regression analysis, the learning motive of this project is used by bioinformatics software such as AlphaFold to predict a protein’s biophysical properties and other functions.

The merged\_df was split twice into training data and test data, taking a different testing data each time.

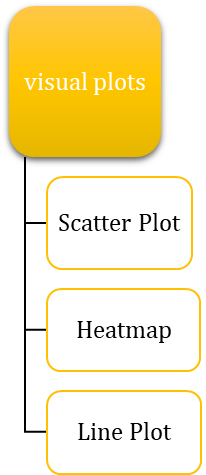
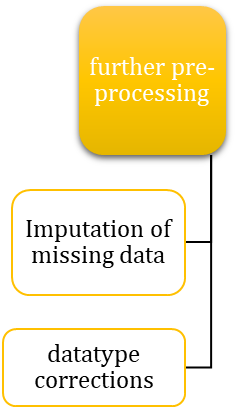
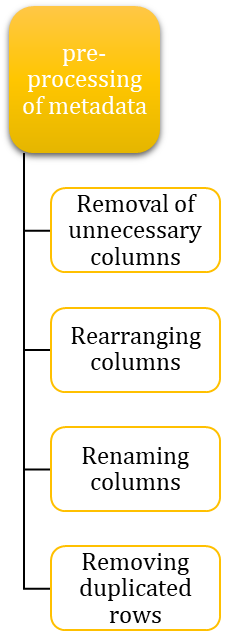
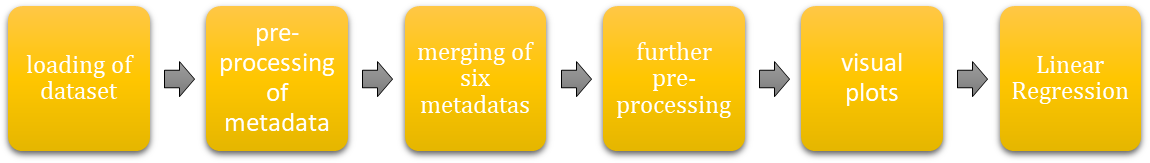
In the first model, the dependant variable or y-variable taken was the ‘Estimated number of children needing ART based on WHO methods’feature. Rest of the columns were the independent variable or x-variable.



Similarly, in the second model, y was taken ‘Estimated number of HIV cases among adults aged 15-49’ while the rest of the features were x.



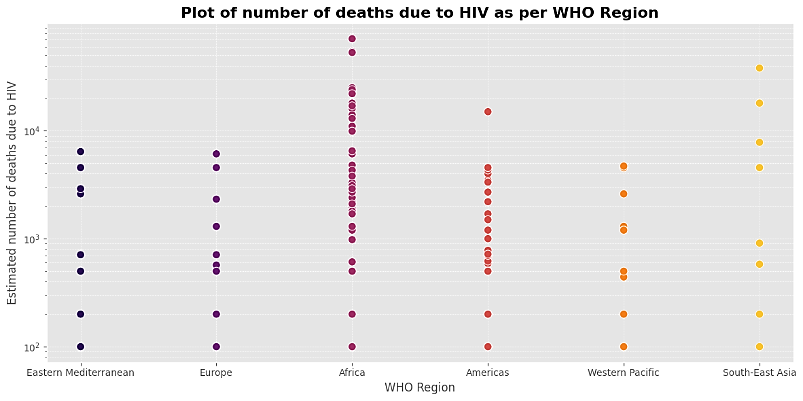
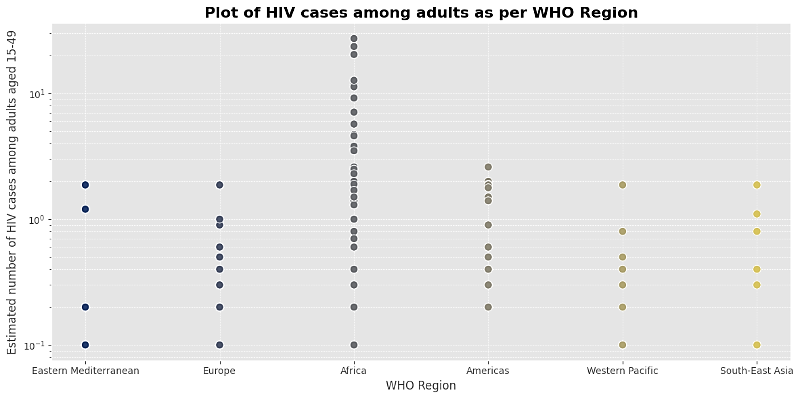
The flowchart below shows all the steps taken to process and analyse the data:

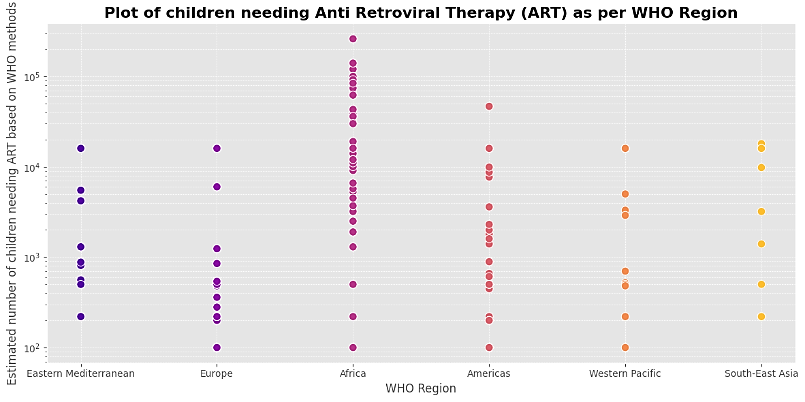
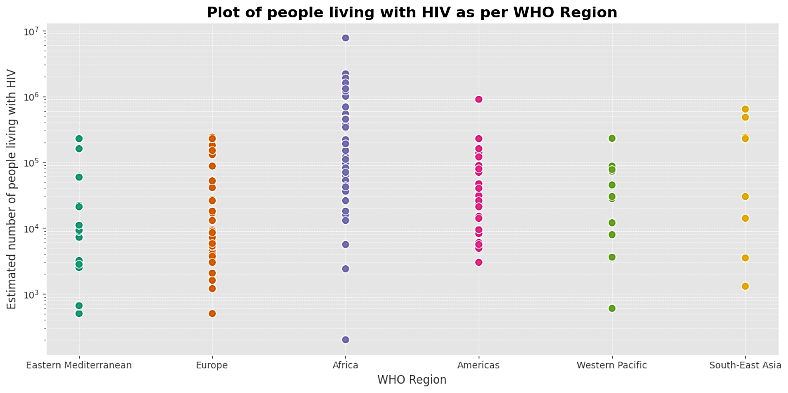


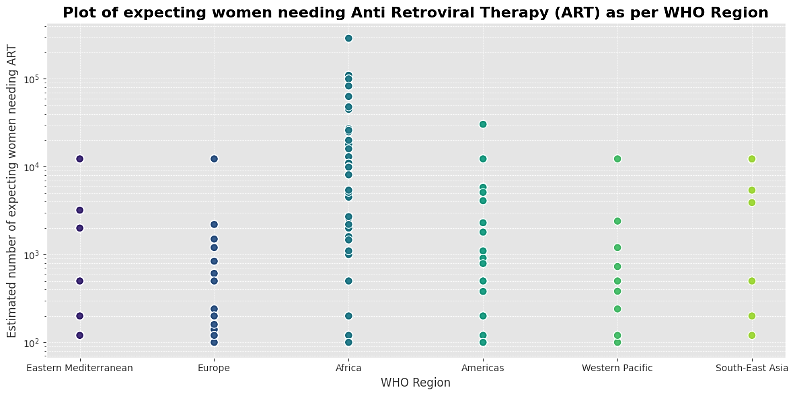
GitHub link to the code of the above program: <https://github.com/IDEASTIHGroup3A/HIV-dataset-analysis>

**DATA ANALYSIS AND RESULTS**

#scatterplotsmade are:

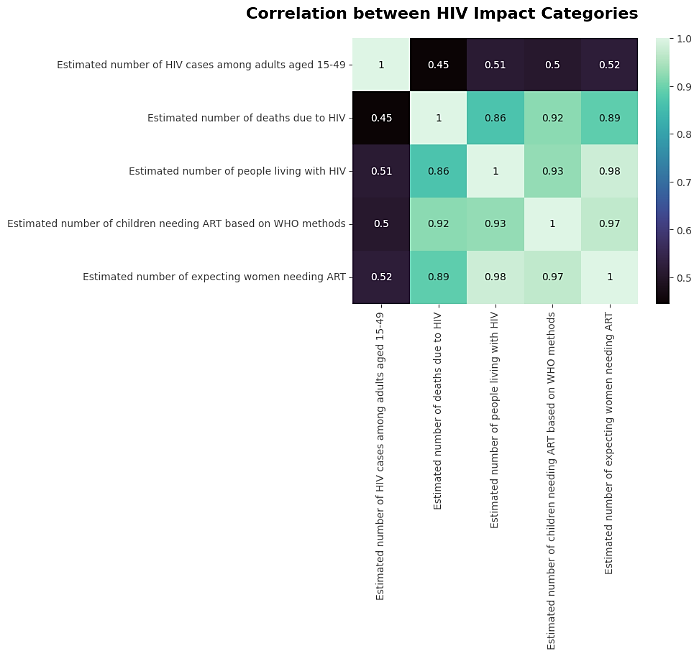






In all the scatterplots drawn, it can be seen that the continent of Africa hasmost number of cases in all categories. It is followed by Americas in most categories and often South-East Asia. Westen Pacific has the least number of reported HIV cases.

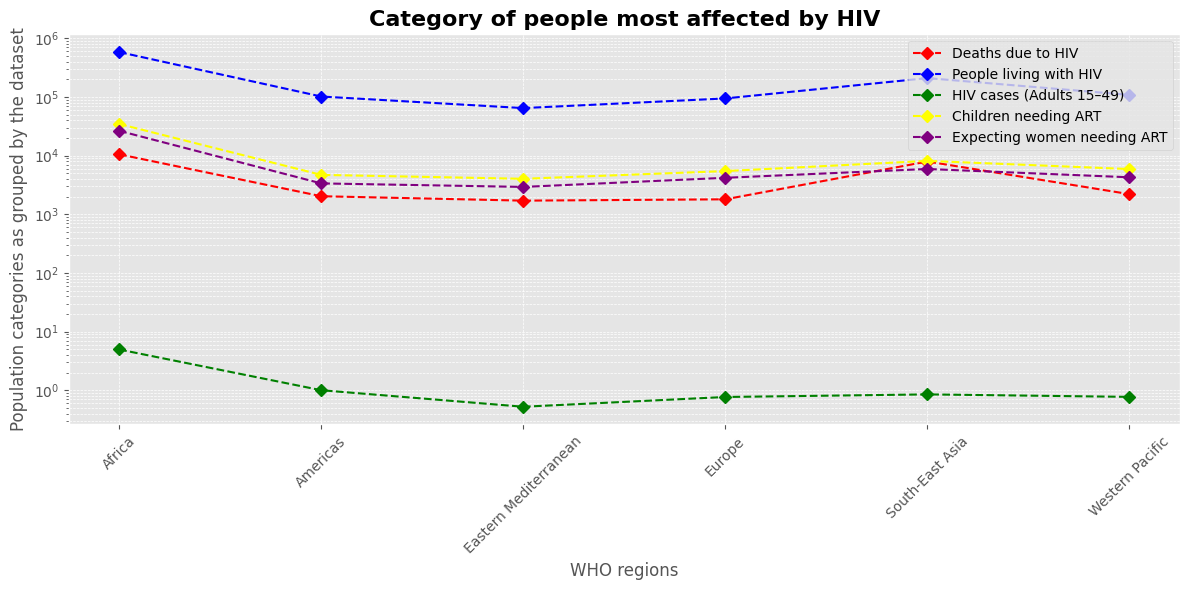
#correlation heatmap showing correlation between different population categories are:

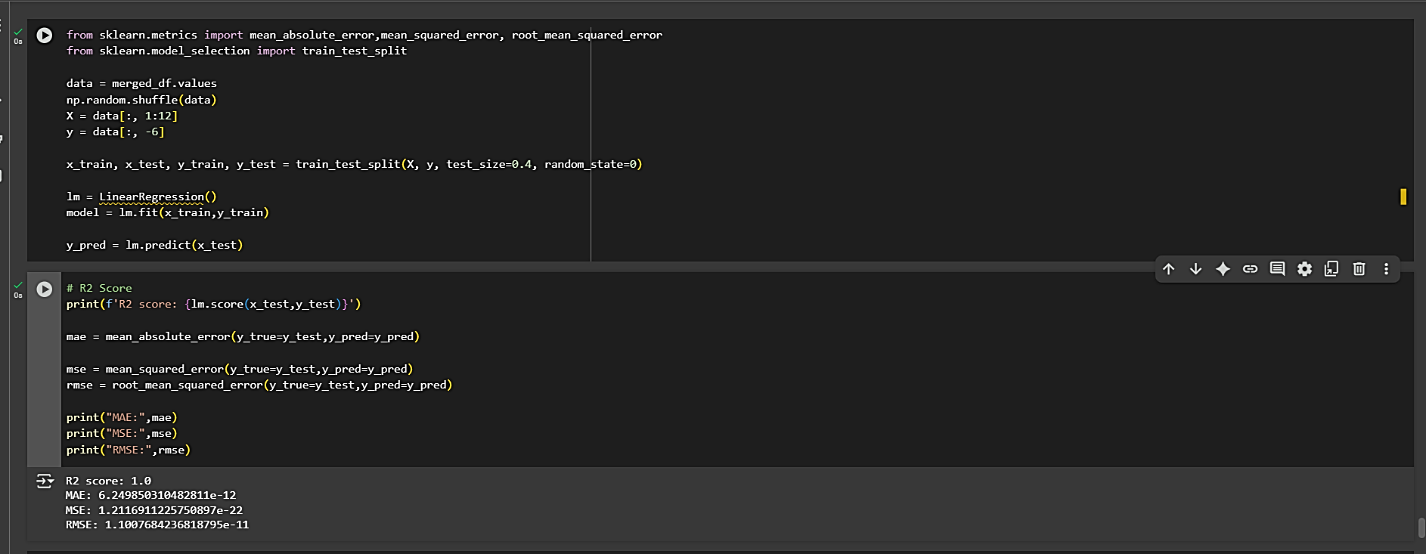


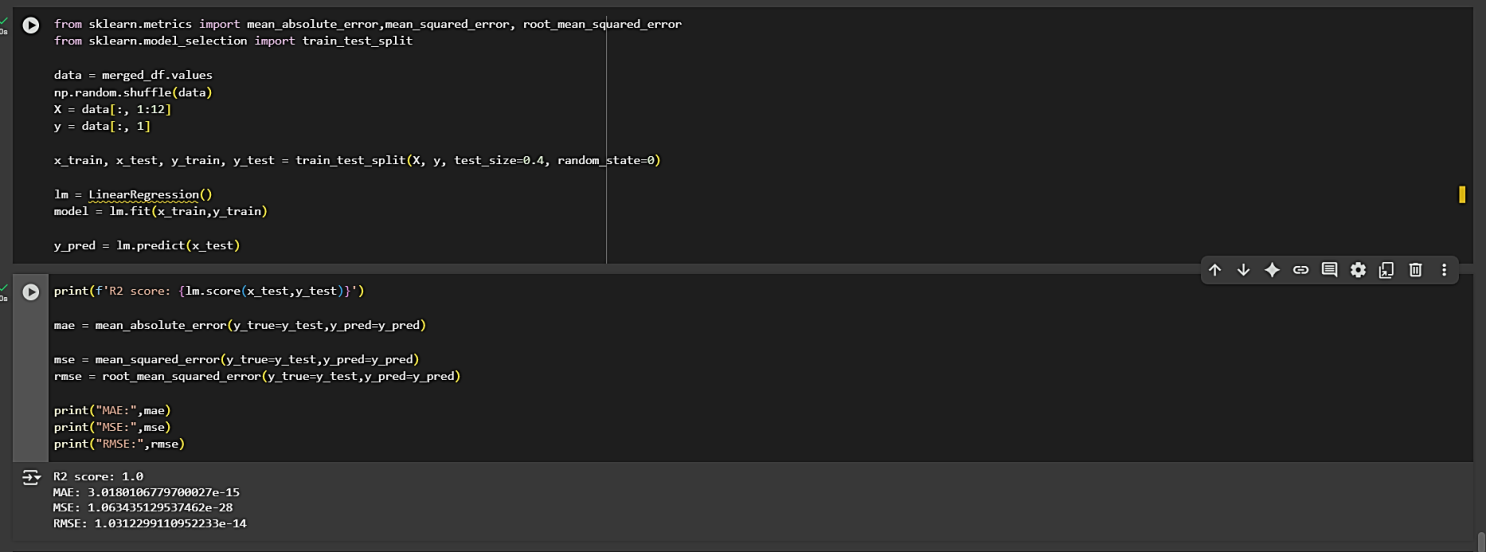
#line plot shows which category of the population is most affected by HIV

Of the meta datasets, maximum number belong to the category, ‘Estimated number of people living with HIV’

Minimum number belong to the category, ‘Estimated Number of HIV cases in adults aged 15 to 49’



The values returned on running the Linear Regression models are:



**CONCLUSION**

Through this project we have learnt how to handle a dataset and the processes we can run in order to better understand and analyze a dataset. The HIV Dataset used in this project sheds light upon the growing concern in the increase in number of reported cases of individuals infected by HIV. As analyzed from the data, we can promptly say that the African continent has a greater number of cases than the rest of the World Health Organization (WHO) regions. Even though Anti-Retroviral Therapy exists as a cure, it isn’t always effective and works best only when detected in the early stages. Thus, there is scope for more drug therapies based on better targets and biomarkers. One of the waysthis can be done by synthesizing novel proteins to that can precisely detect viral proteins and induce metabolic pathways to destroy the viral parts even in later stages of the syndrome. AlphaFold, a groundbreaking AI model is the best way to design such proteins. AlphaFold collects data from the Protein Data Bank (PDB) and predicts structural and functional properties of new proteins. To do so, it uses regression analysis models. Thus, the primary aim of this project was to learn Regression models so that, in future the knowledge can be used for *de novo* protein synthesis and serve the greater cause.

**APPENDICES**

1. HIV dataset taken from Kaggle:

<https://www.kaggle.com/datasets/imdevskp/hiv-aids-dataset>

1. Information about AlphaFold taken from EMBL-EBI website:

<https://www.ebi.ac.uk/training/online/courses/alphafold/inputs-and-outputs/a-high-level-overview/>

1. DeepChem and related commands were learnt from tutorials given on:

<https://deepchem.io/>

1. YouTube tutorials were referred to, for learning the process of Exploratory Data Analysis:

<https://youtu.be/xi0vhXFPegw?si=LtBG4DUybFrLSnMC>

<https://youtu.be/Liv6eeb1VfE?si=LoIP5O8y5eUGWFWX>

1. Youtube tutorial referred to for understanding visual plots for Python:

<https://youtu.be/BGg349k6BCs?si=YdMpfL8z3DFOUBZB>

1. Imputation commands learnt from scikit-learn website:

<https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html>

1. GitHub link for the code written for the project:

<https://github.com/IDEASTIHGroup3A/HIV-dataset-analysis/blob/main/HIV%20dataset%20analysis.ipynb>

1. Link to this document uploaded on GitHub:

<https://github.com/IDEASTIHGroup3A/HIV-dataset-analysis>