Global Tuberculosis Dataset Exploratory Data Analysis (EDA)

A report by Bamundaga Aloyzius Lubowa.

Abstract

The EDA focused on **Bivariate Data Analysis** of the **Income level of an economy** in which data was collected and the **incidence per 100,000 people**. The questions asked helped us narrow down one of the factors that could be influencing Tuberculosis incidence globally.

Questions

What are the required libraries and How do we add the required libraries to our working environment?

We introduce our Python libraries namely **pandas** for "*dataset manipulation"* and **numpy** for "*calculation functions"* using the "import" function.

How do we add the dataset to the environment?

We access the data set from its storage on my git-hub repository using the **pandas.read_csv** function from the pandas library

How many data points and features do we have in our Global_Incidence dataset?

The rows represent the data points while the columns represent the features. This makes **4784 data points** and **14 features**.

How many data points and features do we have in our Africa dataset?

The rows represent the data points while the columns represent the features. This makes **1208 data points** and **8 features**.

How do we distinguish numerical and categorical features?

The **.dtypes** function from pandas library helps us to "*note the data types in order to distinguish numerals from Alphabetical letters"*.

How do we know the features in the dataset?

The **list(data.columns)** function "*creates a list of the elements in the first row thus the features"*. Let's consider Global_Incidence and Africa datasets below.

How do we get a summary of the data point-to-feature relationship?

We can utilize the **.head** or **.tail** function to "*return the first and last data points respectively"* as arranged in the dataset.

In case I wanted to see the trend of the incidence in Uganda from 2000 to 2022, what do I use?

We search through the feature named GEO_NAME_SHORT for the tag Uganda using the **.loc** function.

What is the incidence trend for Low-income countries?

The incidence trend over the period 2000 to 2022 can be acquired by using the .loc function in the GEO_NAME_SHORT column to look for 'Low-income-countries'. The incidence drops from an **average of 274 to 181 cases per 100,000 people**.

What is the incidence trend for High-income countries?

The incidence trend over the period 2000 to 2022 can be acquired by using the .loc function in the GEO_NAME_SHORT column to look for 'High-income countries'. The incidence drops from an **average of 21 to 8.9 cases per 100,000 people**.

What is the incidence trend for Lower middle-income countries?

The incidence trend over the period 2000 to 2022 can be acquired by using the .loc function in the GEO_NAME_SHORT column to look for 'Lower-middle-income countries'. The incidence drops from an **average of 291 to 206 cases per 100,000 people.**

What is the incidence trend for Upper-middle-income countries?

The incidence trend over the period 2000 to 2022 can be acquired by using the .loc function in the GEO_NAME_SHORT column to look for 'Upper-middle-income countries'. The incidence drops from an **average of 134 to 95 cases per 100,000 people.**

Findings

- The Global Incidence data set is divided into sub-data sets like Africa, Income level, Europe etc
- The Global Incidence data set has 4784 data points and 14 features.
- Sub data sets have various features and data points for example Africa has 1208 data points and 8 features.
- The features are both numerical like incidence per 100,000 people and categorical like income levels.
- The incidence trend in Uganda from 2000 to 2022 varies from 276 to 198 cases per 100,000 people per year.
- The incidence trend in Low-income economies varies from an average of 274 to 181 cases per 100,000 people.
- Low-middle-income economies vary from an average of 291 to 206 cases per 100,000 people.
- Upper -middle-income economies vary from an average of 134 to 95 cases per 100,000 people.
- High-income economies from an average of 21 to 8.9 cases per 100,000 people.

Conclusion

The **TB incidence** trend shows **inverse proportionality** with **income levels** across the globe.
Middle-income countries show a **disruption** between the **upper-middle** and **lower-middle-income** countries. There could be another factor or feature at play that can be revealed with **further exploratory data analysis**.

References

[1] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3691414/
[2] https://data.who.int/indicators/i/13B4226/C288D13