**Gapminder Dataset explanation**

Table of Contents

[**Introduction** 1](#_Toc99377214)

[**Methodology:** 1](#_Toc99377215)

[**Dataset Explanation:** 2](#_Toc99377216)

[**Data types:** 2](#_Toc99377217)

[**Justification:** 2](#_Toc99377218)

[**Functions:** 3](#_Toc99377219)

[**Figures in this project:** 3](#_Toc99377220)

[Life expectancy trend from 1952-2007: 3](#_Toc99377221)

[Life expectancy plot for continents: 4](#_Toc99377222)

[GPD percapita of continents: 5](#_Toc99377223)

[**Summary:** 6](#_Toc99377224)

**Revision Control**

| **Version** | **Authors** | **Reason for Issue** | **Date** | **Reviewed By** |
| --- | --- | --- | --- | --- |
| 1.0 | Harsha veena | Prepared dataset explanation | 28-Mar-2022 |  |
|  |  |  |  |  |
|  |  |  |  |  |

# **Introduction**

This project is about GDP percapita of different countries. It contains data of different countries population & its life expectancy at birth and GDP percapita of years from 1952 to 2007.

**IDE**: Google colab notebook

**Programming language**: Python

**Variables in dataset:**

* **Population** means population of country.
* **Continent** means in which continent that country located.
* **Year** means in which year data collected.
* **Life\_exp** means birth life expectancy in that country.
* **Gdp\_cap** means GDP percapita of country.
* **Country** means in which country we are calculating GDP.

# **Methodology:**

We downloaded datasets from LMS homework platform.

# **Dataset Explanation:**

We have two datasets named gapminder\_key\_1 and Gapminer\_key\_2. In Gapmider\_key\_1 dataset we have columns named country, year, population, continent. It has 1704 rows and 4 columns. In Gapminder\_key\_2 dataset, we have columns named country, year, Life\_exp, gdp\_cap. It has 1704 rows and 4 columns. When we compare both datasets in to one dataset named Gapminder dataset. It has *1704 rows and 6 columns* named continent, country, year, population, life\_exp, gdp\_cap. GDP is Gross Domestic Product. GDP is average income per person. GDP percapita means GDP per population of country in that year. Life expectancy is on an average age a person excepted to live.

# **Data types:**

Identification of data types can be done by using code gapminder.info().

We got column data types as follows

|  |  |
| --- | --- |
| Country | Object |
| Year | Int64 |
| Population | Int64 |
| Continent | Object |
| Life\_exp | Float64 |
| Gdp\_cap | Float64 |

# **Justification:**

Country: This column has names, and it is a string data type. So, string is nominal, and it is class object. But in python it can’t read string object, so we convert it to category.

Continent: This column has names ant it is a string data types. So, string is nominal data, and it is showing as class object. But in python I can tread string object, so we convert it to category.

Year: these columns have numbers which don’t have intermediate values. So, it is discrete data type, so it shows int64.

Population: This column has numbers which don’t have intermediate values. So, it is discrete datatype, so it shows class int64.

Life\_exp: this column has intermediate values, so it is continuous data type. So, it shows float64.

Gdp\_cap: this column has intermediate values, so it is continuous datatypes. So, it shows class float64.

Datatypes after converting:

|  |  |
| --- | --- |
| Country | Category |
| Continent | Category |
| Year | Int64 |
| Population | Int64 |
| Life\_exp | Float64 |
| Gdp\_cap | Float64 |

# **Functions:**

Here, they have used head(). tail(), describe() functions.

Head():

This function is used to view data in data frame of top 5 rows and all columns. This function is recommended to use where there is large dataset, and it will take time to load all data and we just want to check only few amounts of data. That time we can use this function.

Tail():

This function is used to view data in data frame of last 5 rows and all columns. This function is recommended to use where there is large dataset, and it will take time to load all data and we just want to check only few amounts of data. That time we can use this function.

Describe ():

This function is used to view statistical data of data frame such as to calculate count, maximum value, minimum value, average, standard deviation, 25 percentile, 50 percentile,75 percentile of data. In this dataset to use this function first we did group data by country and year.

# **Figures in this project:**

## Life expectancy trend from 1952-2007:

We created a box plot to visualize trend change between 1952-2007 for life expectancy. In x- axis we have taken year and y-axis we have taken life\_exp variables to plot. From plot we can see in 1952 life-exp median was between 40-50 years range but in 2007 it is between 70-80 years range. We can clearly see median value is increasing from 1952 to 2007. So, life-expectancy is increasing year by year.

Chart, box and whisker chart

Description automatically generated

## Life expectancy plot for continents:

We created a box plot to visualize life expectancy for different continents. We take continent in x-axis and life\_exp in y-axis. Here we can see different continents has different life expectancy. Africa’s boxplot shows its median lies between 40–50-year range and his value is lowest of all median values. So lowest life expectancy is for Africa continent. From this plot we can say highest life expectancy is for Oceania where its median value is between 70-80 years range.Chart, box and whisker chart

Description automatically generated

## GPD percapita of continents:

We plot scatter plot to visualize between GDP percapita in x-axis and life expectancy in y-axis. Here different color labels show different continents. From plot we can say life expectancy increases when GDP percapita increases for some continents like Americans, Asia, Europe. For some continents like Africa life expectancy increased with low GDP itself. Oceania is having life expectancy between 70-80 by increasing GDP. The continent with having highest life expectancy with high GDP is Europe and Asia with 70-80 years range. The continent with lowest life-expectancy with lowest GDP is Africa in between 40-50 years range. As an overall we can say by increasing GDP life expectancy is increasing.

Chart, scatter chart

Description automatically generated

# **Summary:**

In this project we introduced about dataset about variables. We understand about dataset by identifying data types and concert it to proper data types. We describe functions used as head(), tail(),describe(). We described figures in project. We explained box plot of life expectancy year, boxplot of life expectancy by continent and scatter plot of life expectancy and GDP per capita.