Gap Minder World

Introduction

Gapminder identifies systematic misconceptions about important global trends and proportions and uses reliable data to develop easy to understand teaching materials to rid people of their misconceptions.
Gapminder is an independent Swedish foundation with no political, religious, or economic affiliations.

Questions to be answered:
what is the correlation between Population and GDP per capita?
What is the correlation between life expectancy, GDP and Population?

Gathering Data

Data Gathered from Gapminder site
www.gapminder.org
3 files are gathered as below

- income_per_person_gdppercapita_ppp_inflation_adjusted.csv
- life_expectancy_years.csv
- population_total.csv

and one more file added to them that contains countries with their regions that helped in the data analysis called:
countries.csv

Assessing Data

Data assessed by check them using pandas library in python to know the data I have
I found that each file consists of a column for country and the other columns for years
The columns for years started from 1800 till 2100 and each country have its specific information for indicators on each of the year
Some issues found in data like some duplicated values and some blanks found that had to be handled

Cleaning

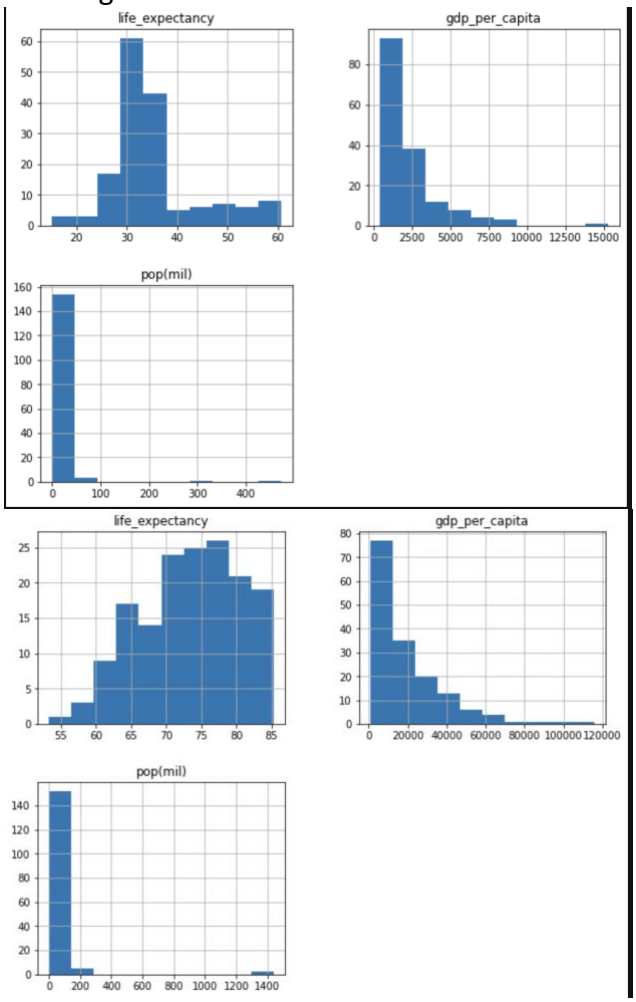
After found the quality and tidiness issue in the data I handled them by using pandas and NumPy libraries in python and after confirmed that data became clean
I make another assessing step to check for any more information that need to be extracted and found that I need each country to have the region or the continent that it's in to help me in analyzing the data so I tried to find a list of countries and regions in the gapminder but couldn't find it so I got it from other source online from the world bank website
www.worldbank.com

Exploratory Data Analysis

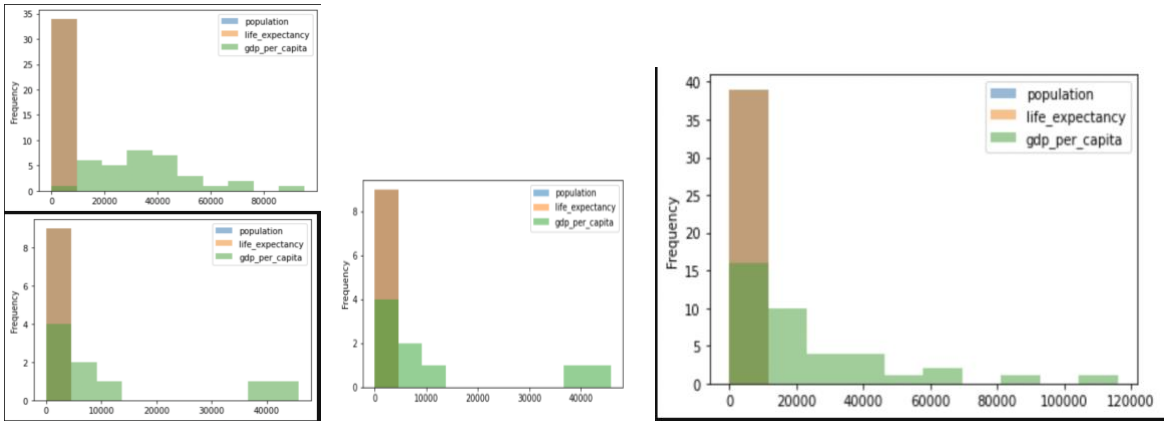
Now that I've trimmed and cleaned the data, I was ready to move on to exploration. Compute statistics and create visualizations with the goal of addressing the research questions that posed in the Introduction section.

First, I started by look at one variable at a time, and then follow it up by looking at relationships between variables.

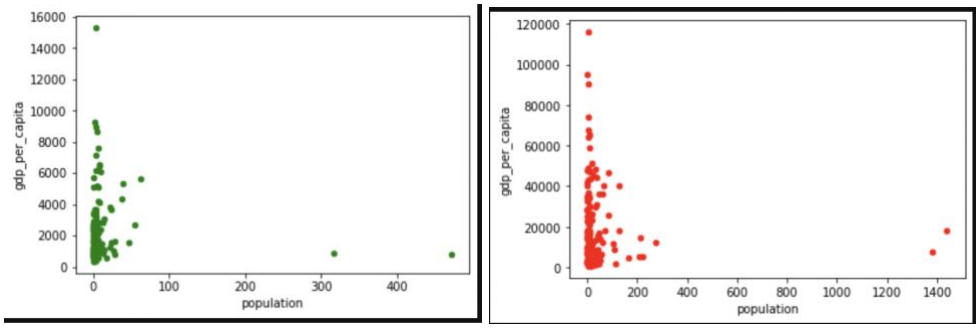
These are visualization for each of the 2 years I chooses to analyze which are 1920 and 2020 Showing each indicator distrubition.



Then I plotted the 3 indicators together with the regions as below



Then made some scatter plots to check correlation between GDP and population through 1920 and 2020 to answer the first question

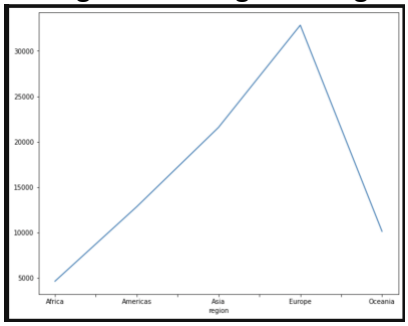


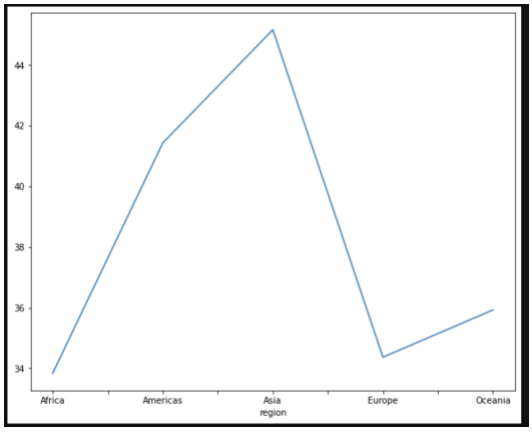
what is the correlation between Population and GDP per capita?

as showing in the scatter plot the gdp and pop changed significantly from 1920 to 2020 that showing the relationship between the population and gdp as the more the population increasing the more decreasing happing in the gdp as the income per person become lower and lower

then started to answer the second question and to answer it I had to make some region analytics and see correlation between each region and the indicators

and got some visualization for them as below
this figure showing each region and gdp change happen in it





And this figure showing the region and change that happen life expectancy in it

2nd question What is the correlations between life expectancy, GDP and Population?

- as showing on the above histograms the distribution of the life expectancy became more left skewed from 1920 to 2020
- and for gdp per capita its changed slightly and the distribution more flattered
- for the population the skewed is changed more to right skewed
- what we can get from it that the more the gdp is increased the more the life expectancy is increased
- and the more the population is increased that slightly impacting gdp per capita
- below visualization confirmed it

Limitation

The whole dataset combined of many indicators like gdp and life expectancy so I had to decide which one to take in order to carry out this project.
Next I have problem with that data have a lot of years which makes the analysis hard and need to pick just few years to analyze
The data does not contain information about the region for each country so I had to get it from other sources.

Conclusions

The goal for project to analyze a dataset that got from gapminder which contains different indicators
I choosed 3 main indicators which are gdp per capita, population and life expectancy and I make some comparison between them and found some correlations between different indicators and how they impacting each other's