Project Report

Investigating The Relationship of India's Life Expectancy with Income, Death Rate and CO2 emission

Table of Contents

Introduction

In this project I have tried to explore the relationship of life expectancy (which is a statistical measure of the average time a citizen is expected to live) with income (which is GDP of india per person adjusted for differences in purchasing power), death rate (which is the ratio of deaths to the population of a particular area) and CO2 emissions (which is the total amount of CO2 emission in a year) particularly for India as that is where I live. In order to carry out this analysis I have used the updated datasets from the GAPMINDER website.the dataset given on the site is very clean so i had to only reshape to do my particular analysis.

Data Wrangling

- First I loaded the CSV files of the dataset into different dataframes.
- Then I inspected the columns to check that how many years of data is given in each dataset.
- Then I used the **'info'** method of pandas to check the data types of the different columns of the dataset and **if there are any missing values in any column.**
- There were **no missing values** in any column so no cleaning was required there.

- Since i am doing an India specific analysis i extracted the india rows from all the datasets as there are many countries given in the gapminder dataset.
- All the data in these dataset is given in pivotal format which i had to convert to non-pivotal using the ‘melt’ method of pandas.
- Because of the melting all the years are now under one column called year but they **were of type ‘string’ so I converted them to ‘int’.**
- All the datasets had data for different number of years. So to make the analysis process easy i had to trim the datasets to get data for the same years everywhere.
- So I trimmed all the datasets to contain data for years starting from 1960 till 2014.
- Since all data is of the same country **I dropped the ‘country’ column from every dataset.**
- Then I merged all the datasets into a single one using an ‘outer merge’ on the column ‘year’.
- Then I saved the cleaned dataframe in a new csv called ‘combined_dataset.csv’.

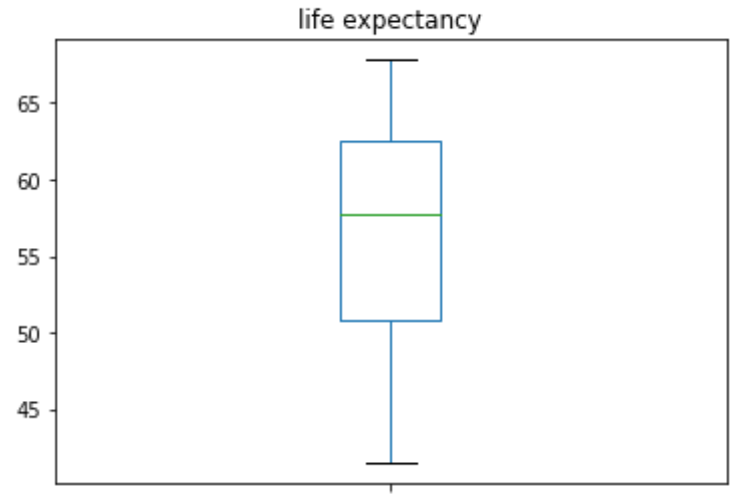
Exploratory Data Analysis

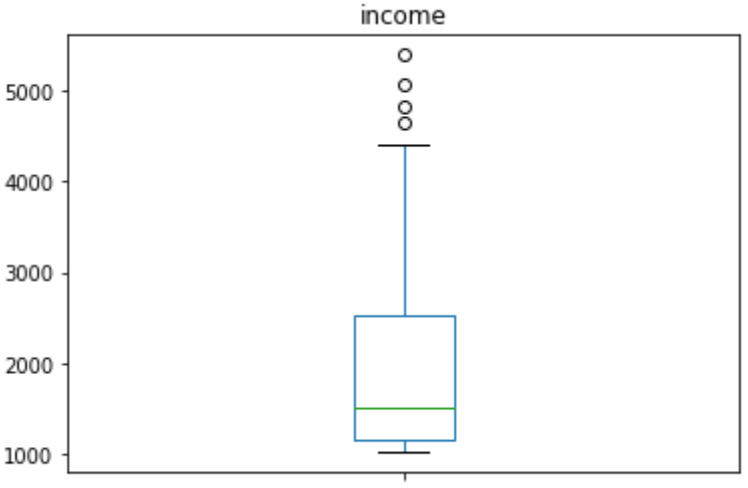
Questions to be answered:

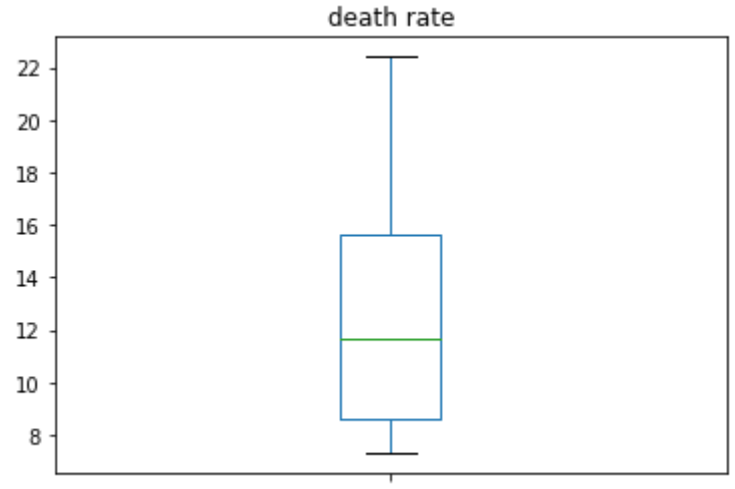
- How does income affect life expectancy in india?
- What is the correlation between life expectancy and death rate in india?
- Is CO2 emission a significant factor to predict life expectancy?
- How has life expectancy , income, death rate and CO2 emission changed over the years?

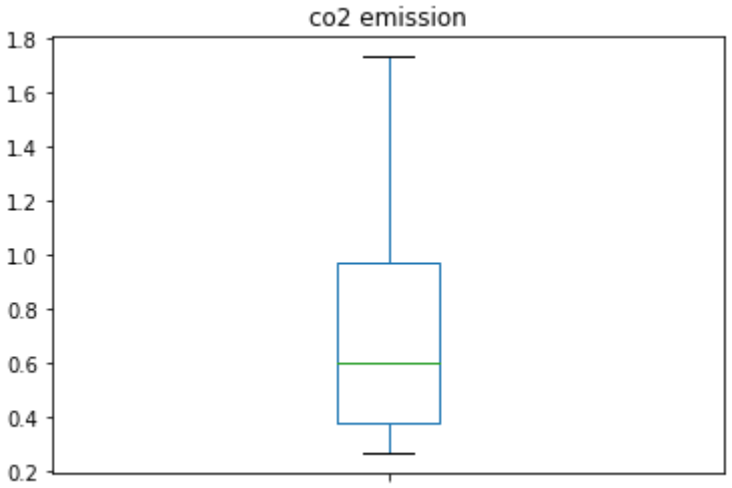
Initial exploration:

- I created box plots for every feature to check for any bias.


life expectancy
65
60
55
50
45

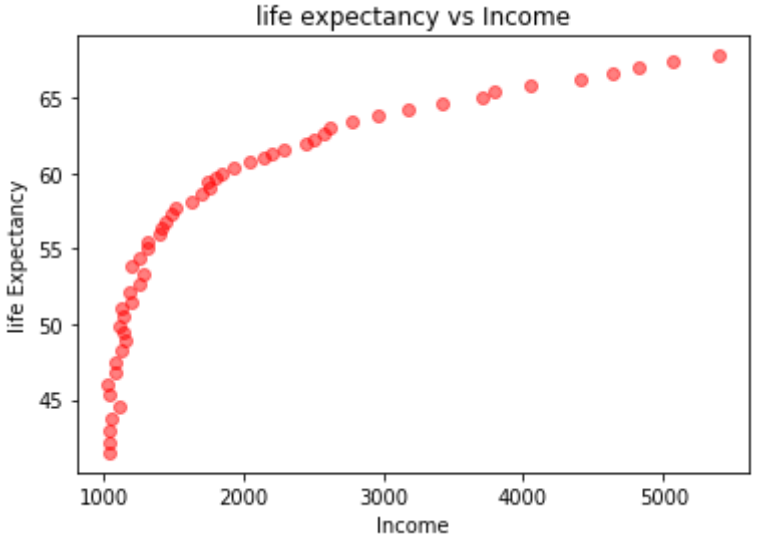

income
5000
4000
3000
2000
1000


death rate
22
20
18
16
14
12
10
8



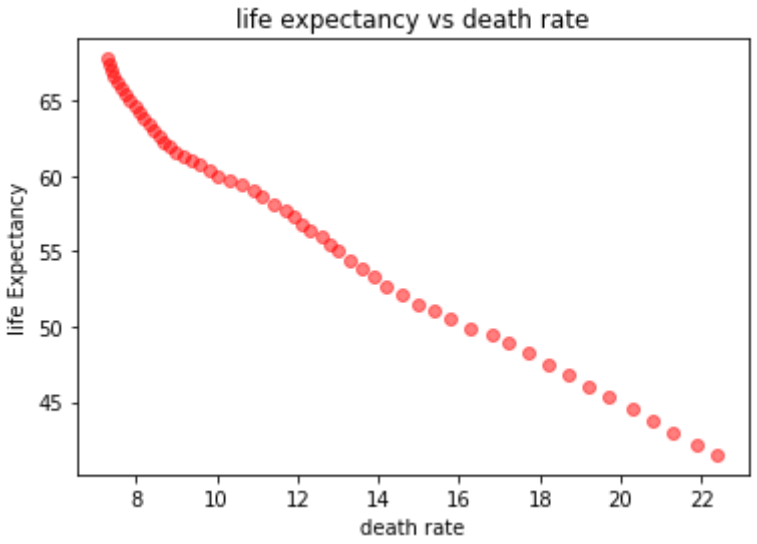
- We can see that income has some outliers on the higher side and most of the distribution is on poorer side.
- Death rate and CO2 emission are biased on the right.

- Now coming to our first question I have created a scatter plot between life expectancy and income to see if there is any correlation.



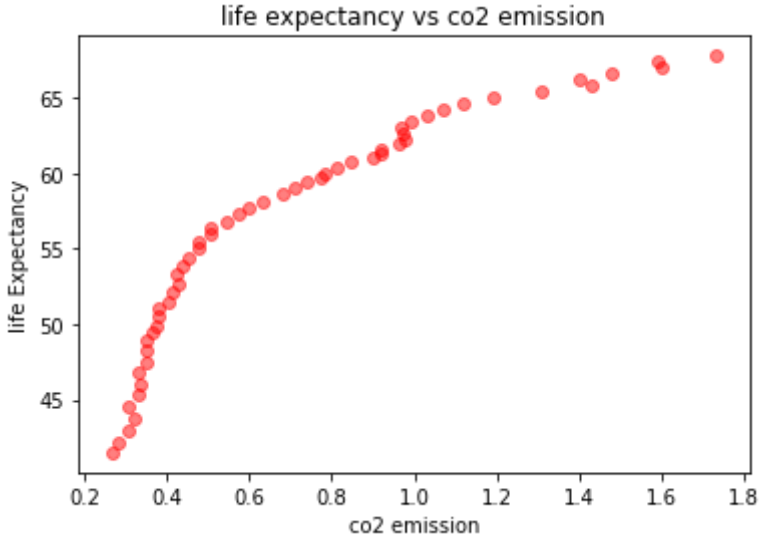
- This plot shows us that there is a positive correlation between income and life expectancy.
- In this plot we can see that the rate of growth for life expectancy is very high for the income range 1000$ to 2000$.

- For the second question I have created a scatter plot between death rate and life expectancy.



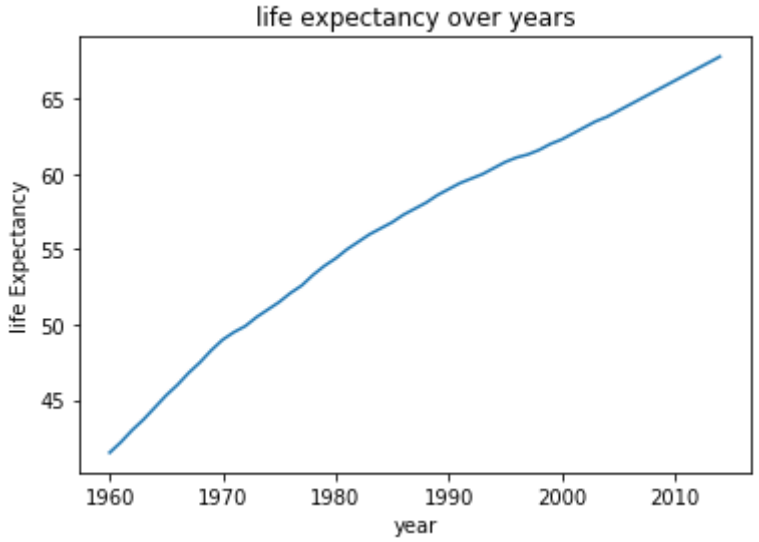
- We can tell by seeing this plot that there is a negative correlation between life expectancy and death rate and the rate of change is very consistent.

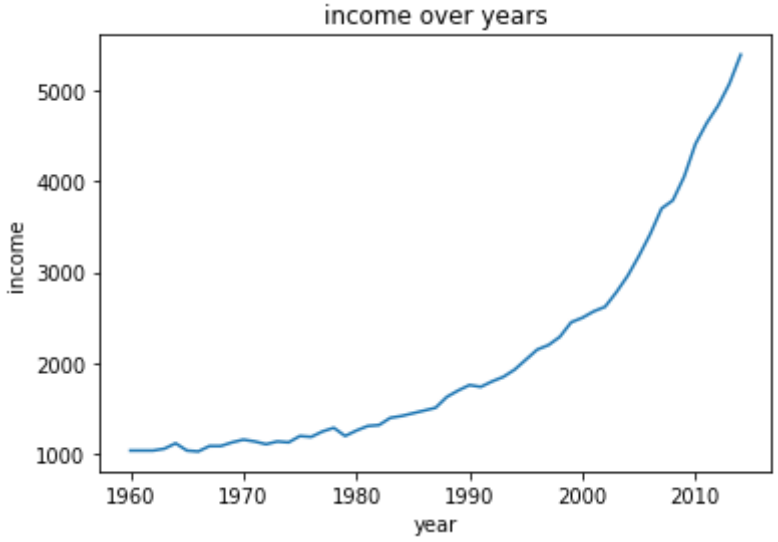
- Now for the third question I created a scatter plot for CO2 emission vs life expectancy.

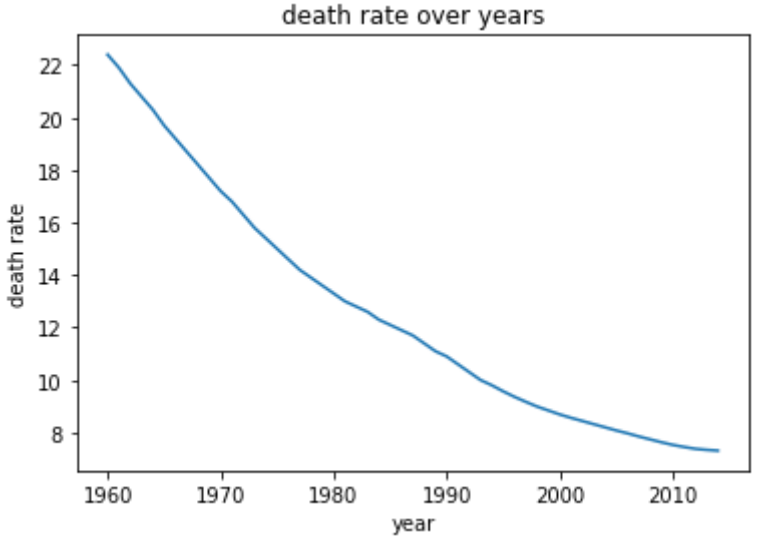


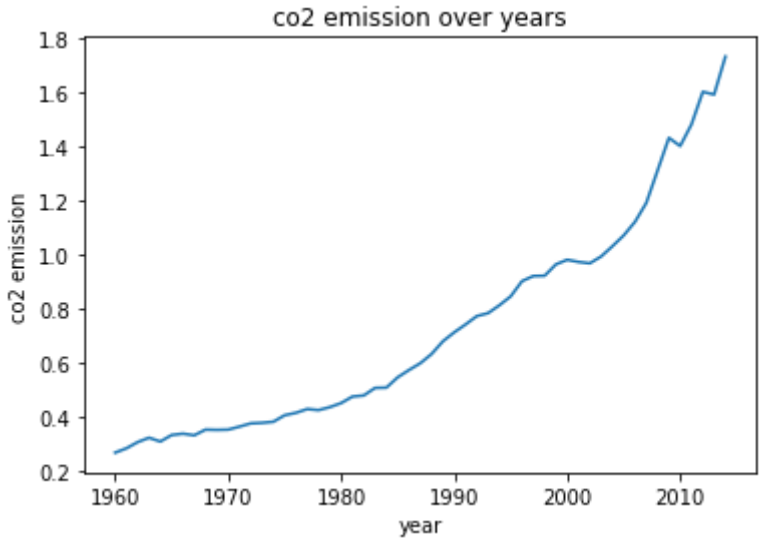
- This plot shows us that there is a positive correlation between these factors but we know that CO2 is not good for our health and continual exposure to it would reduce life expectancy. Because of this we can't use this to predict life expectancy.

- For the 4th question I created line charts for life expectancy, income, death rate and CO2 emission vs years(1960 - 2014).


life expectancy over years
life Expectancy
65
60
55
50
45
1960
1970
1980
1990
2000
2010
year


income over years
income
5000
4000
3000
2000
1000
1960
1970
1980
1990
2000
2010
year


death rate over years
death rate
22
20
18
16
14
12
10
8
1960
1970
1980
1990
2000
2010
year



- These plots tell us that life expectancy has increased consistently over the years in India.
- Income increased very gradually up until 1990 but then increased at a faster rate in India.
- Death rate has decreased at a consistent rate over the years in India.
- CO2 emissions have increased in India with minor fluctuations over the years.

Conclusions

- Since there is a positive correlation between income and life expectancy, higher the income higher the life expectancy will be.
- There is a negative correlation between death rate and life expectancy.
- There is a positive correlation between CO2 emission and life expectancy
- Life expectancy, income, CO2 emission have increased over the years in India and death rate has decreased over the years.

Limitations

These results are limited to this analysis as thorough statistical tests are not done and we saw that there was a bias in the Income column in the bar chart towards higher income and there were also some outliers in the income dataset so the conclusions here may not be 100% accurate. At the time I did this analysis I had very little knowledge of statistics so the analysis may not be that thorough. And to make the analysis easier I had to trim the no of years from each dataset so there were only 50 years worth of data to analyse from, so the conclusions might be inaccurate for a longer time period.