adani

Institute of Digital Technology Management

PGDM in Big Data Analytics
Term II (Batch 2023-25)

Course Name: Statistics for Data Analysts

Life Expentancy

Submitted to: Dr. Manjari Mundanad

Submitted by:

Name	Enrolment Number
Dhineshkumar B	20231017
Malhar Kalse	20231033
Pavan Jangid	20231041
Sanjana Detroja	20231050

Dated: 18th January 2024

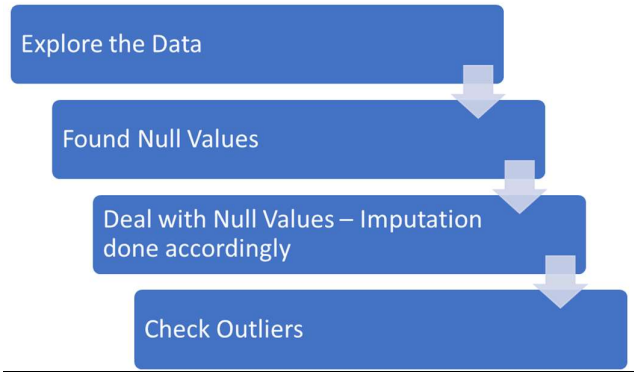
Variable Details

Variable	Variable code	Descriptions	Measurement
Life expectancy	Life expectancy	The average number of years a person is expected to live.	years
Adult Mortality	Adult Mortality	The probability of dying between the ages of 15 and 60 per 1000 population.	Deaths per 1000 population
Alcohol	Alcohol	Alcohol consumption per capita	Liters of pure alcohol per capita
percentage expenditure	percentage expenditure	The percentage of the government's total expenditure on health.	Percentage
BMI	BMI	A measure of body fat based on height and weight	Body Mass Index
Polio	Polio	Immunization coverage against polio	Percentage
Diphtheria / Hepatitis	Diphtheria	Immunization coverage against diphtheria/Hepatitis	Percentage
HIV/AIDS	HIV/AIDS	The prevalence of HIV/AIDS	Percentage
GDP	GDP	The total value of goods and services produced by the country.	US dollars
Income composition of resources	Income composition of resources	The percentage of total income derived from different sources	Percentage
Schooling	Schooling	Average number of years of schooling for adults aged 15 and older	years

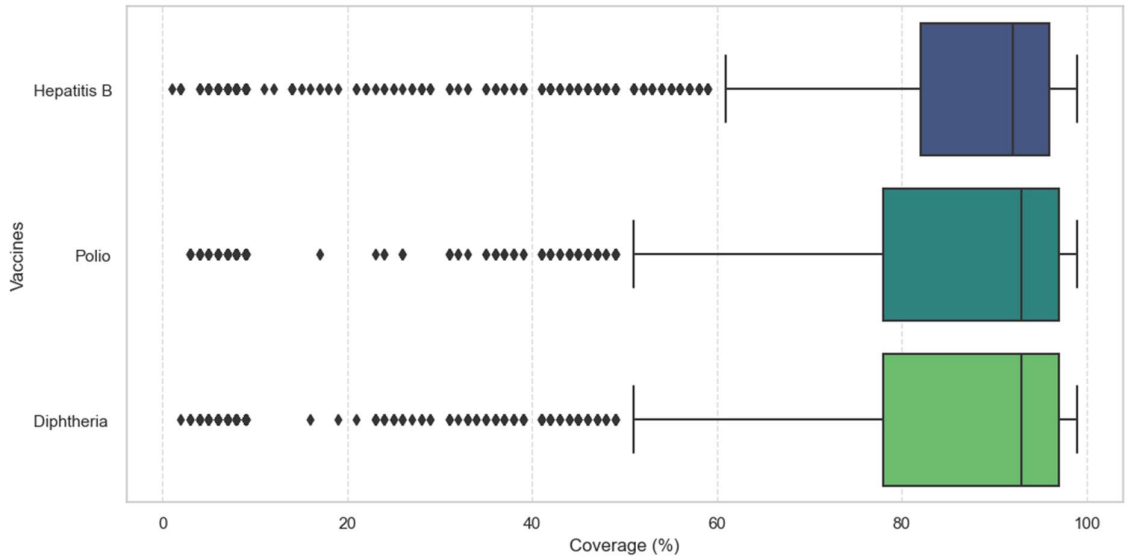
Descriptive statistics

Test	LIFE EXPE	ADULT MO	ALCOHOL	BMI	GDP	DIPHTHERIA	HIV_AIDS	INCOME C..	PERCENTA.	POLIO	SCHOOLING
Mean	69.22433	164.7257	4.546875	38.32125	6611.524	82.322408	1.742103	0.630362	738.2513	82.55019	12.00984
Median	72	144	3.755	43	1766.948	93	0.1	0.677	64.91291	93	12.3
Maximum	89	723	17.87	87.3	119172.7	99	50.6	0.948	19479.91	99	20.7
Minimum	36.3	1	0.01	1	1.68135	2	0.1	0	0	3	0
Std. Dev.	9.50764	124.0862	3.921946	19.92768	13296.6	23.64007	5.077785	0.20514	1987.915	23.35214	3.265139
Skewness	-0.639367	1.177298	0.649246	-0.220478	3.541946	-2.078419	5.393357	-1.211907	4.649676	-2.103789	-0.634727
Kurtosis	2.773314	4.761809	2.374108	1.729071	18.11539	6.592631	37.83061	4.689144	29.52614	6.812043	4.119723
Jarque-Bera	206.4612	1058.67	254.3604	221.5378	34112.19	3695.306	162756	1068.462	96723.15	3946.147	350.7599
Probability	0	0	0	0	0	0	0	0	0	0	0
Sum	203382.8	483964	13358.72	112587.8	19424657	241868.2	5118.3	1852.003	2168982	242532.5	35284.9
Sum sq. Dev.	265490.8	45222131	45175.93	1166319	5.19E+11	1641351	75727.3	123.5956	1.16E+10	1601612	31311.75
Observations	2938	2938	2938	2938	2938	2938	2938	2938	2938	2938	2938
Coe. Of Variation	13.73454	75.32899	86.25586	52.00164	201.1125	28.71645	291.4744	32.54321	269.2735	28.28841	27.1872

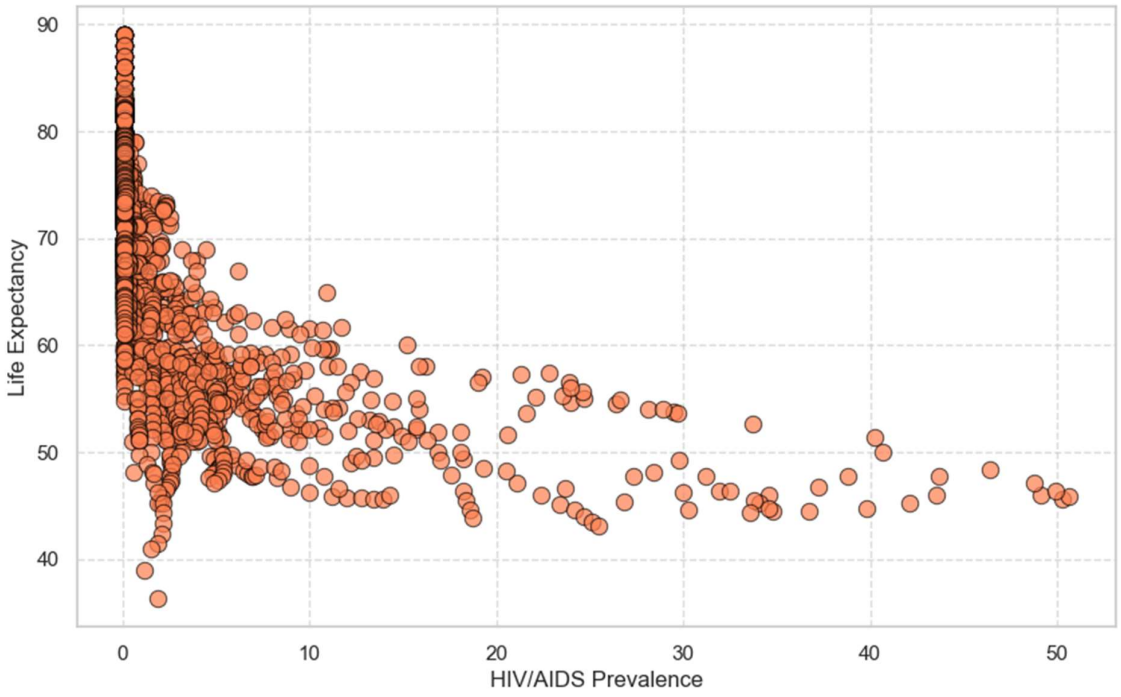
Clean data with Python



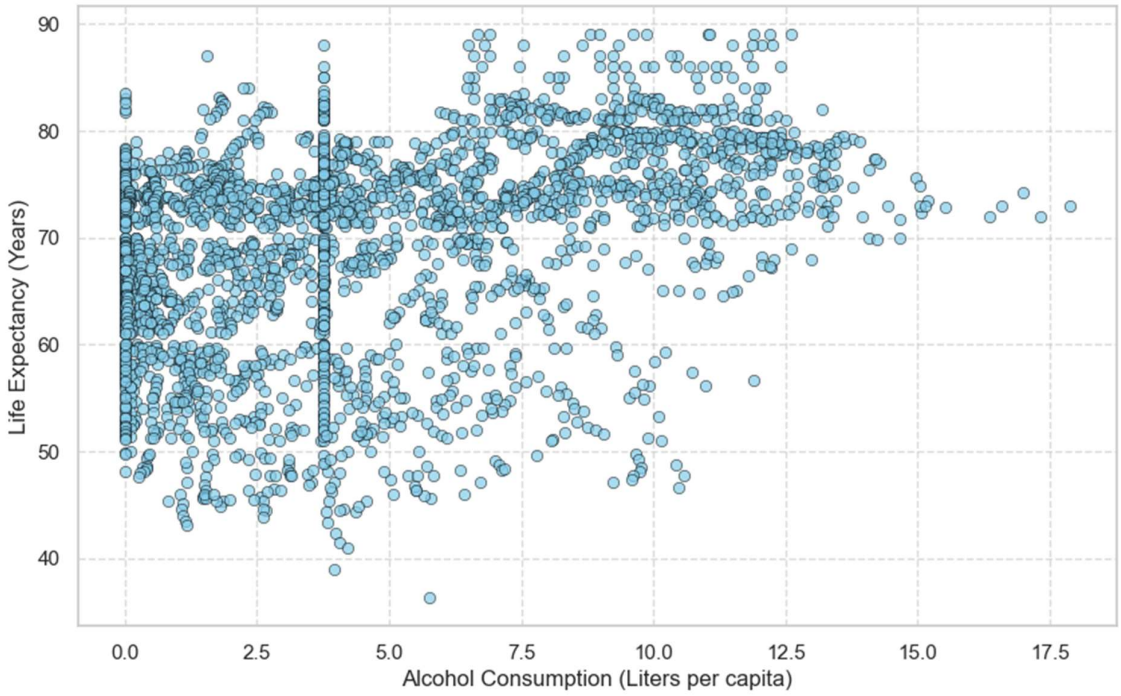
Distribution of Immunization Coverage



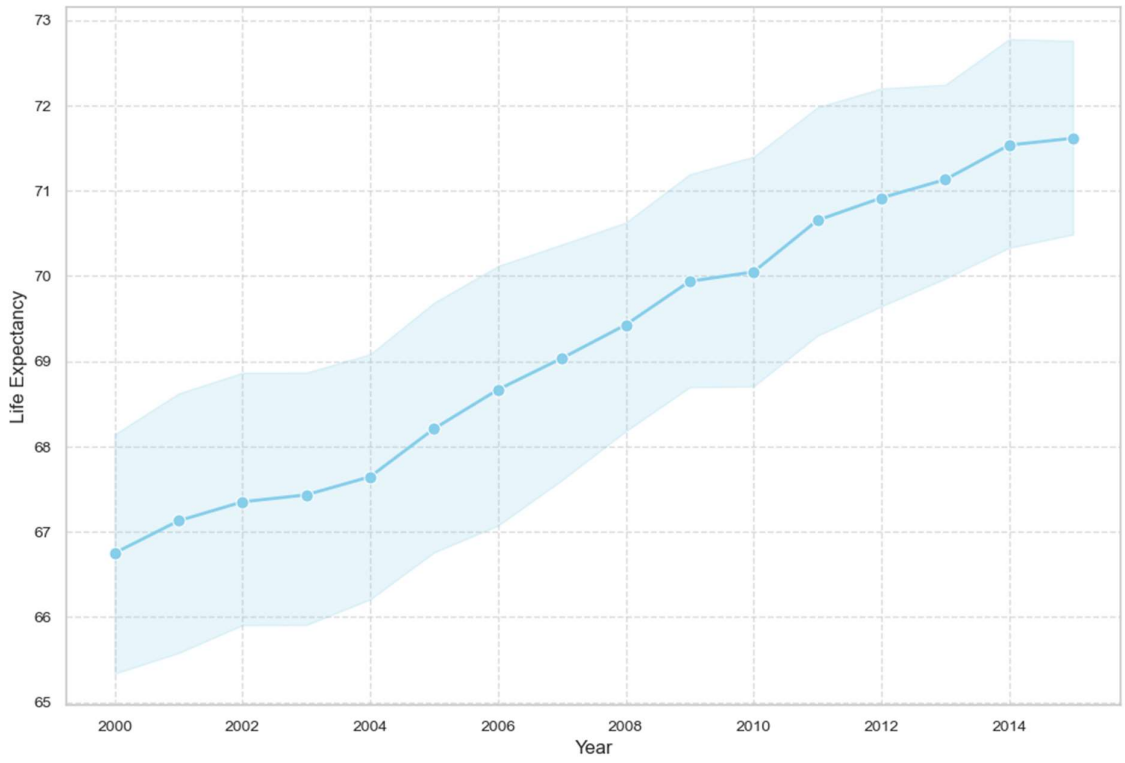
Impact of HIV/AIDS on Life Expectancy



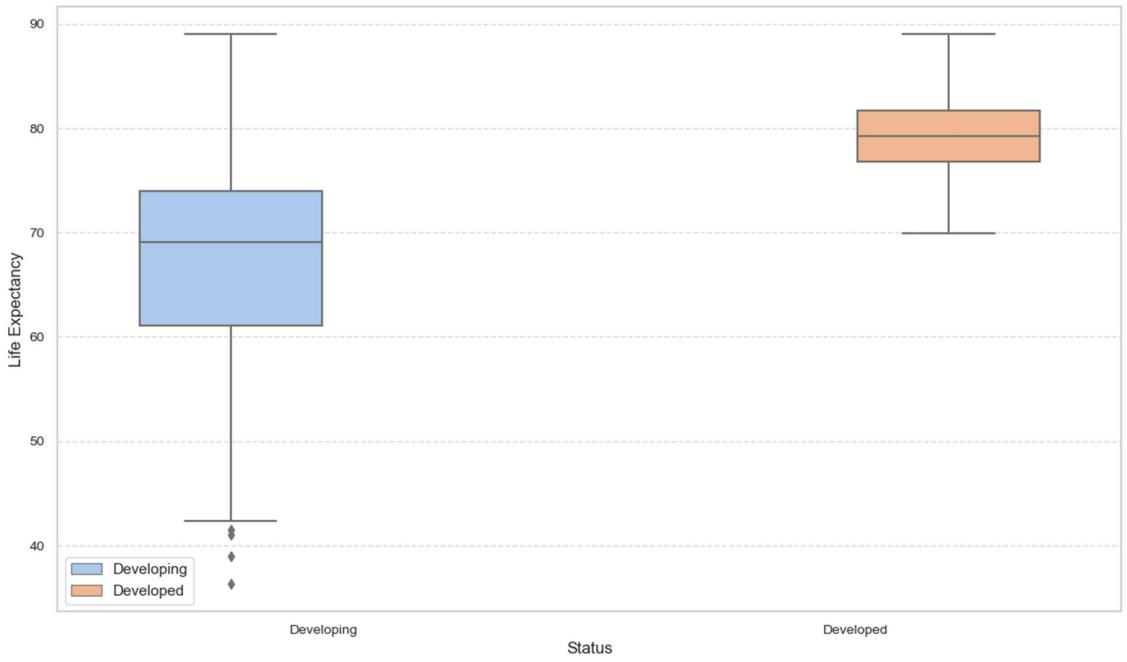
Relationship between Alcohol Consumption and Life Expectancy



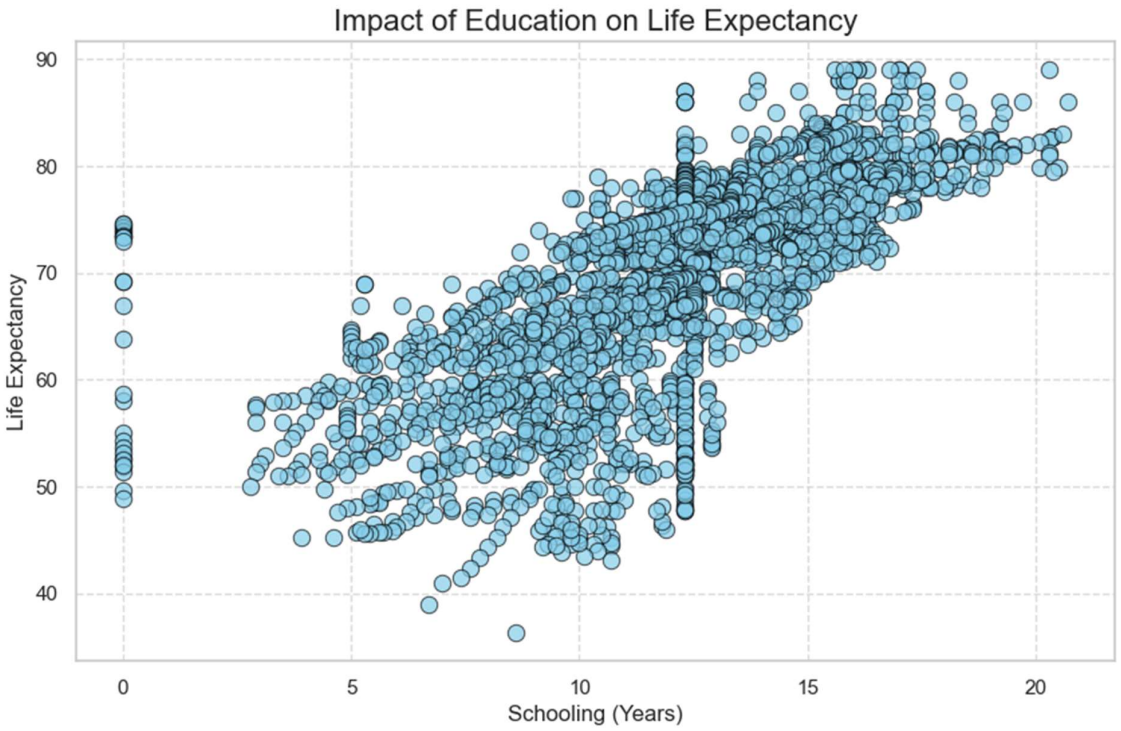
Global Life Expectancy Trends over Time



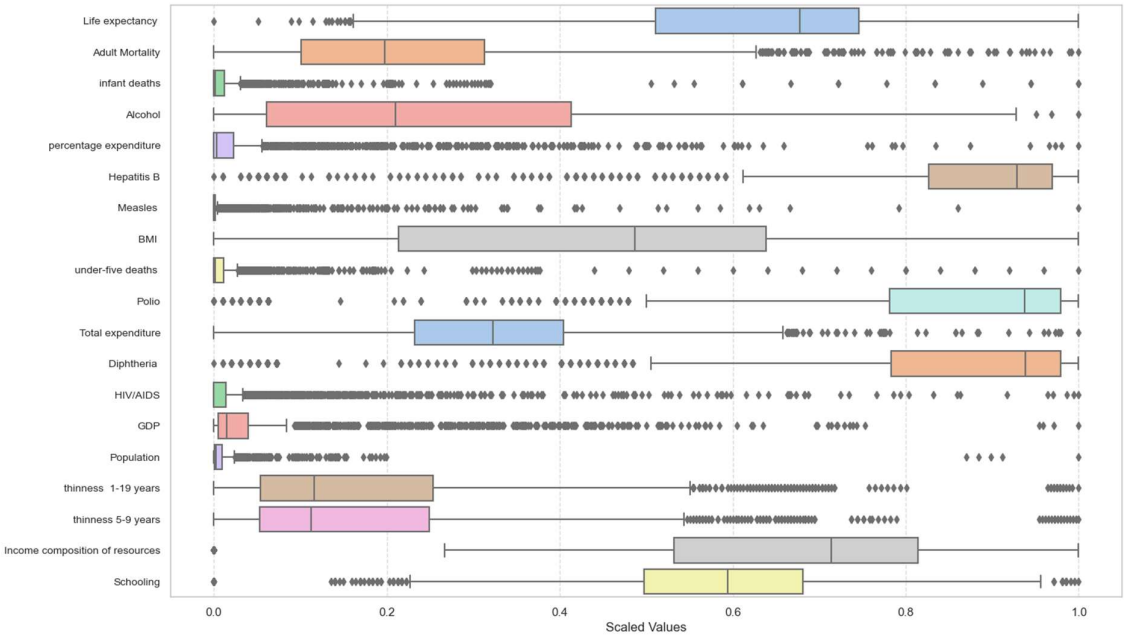
Regional Disparities in Life Expectancy



Impact of Education on Life Expectancy



Box Plots after Scaling



Regression

1. Dependent variable: Life expectancy
2. Independent variable: Adult Mortality, Alcohol, percentage expenditure, BMI, Polio, Diphtheria, HIV/AIDS, GDP, Income composition of resources, Schooling
3. No of independent variable (k): 10
4. No of observation (N): 2938
5. Time period: 15 years (2000 – 2015)

Equation:

$$y = f(x_1, x_2, \dots, x_n)$$

Where,

- $x_1, x_2, \dots, x_n$ *are independent variables*
- y *is dependent variable*

Correlation

Variable	*Life expectancy*
Life expectancy	1
Adult Mortality	-0.696326127
Alcohol	0.389846664
percentage expenditure	0.381791173
BMI	0.559255305
Polio	0.461573775
Diphtheria	0.475418385
HIV/AIDS	-0.556456817
GDP	0.430894571
Income composition of resources	0.68842496
Schooling	0.713738004

1) Descriptive statistics

	LIFE_EXPEC	ADULT_MO	ALCOHOL	BMI	DIPHTHERIA	HEPATITIS_B	GDP
Mean	69.22493	164.7257	4.546875	38.32125	82.32408	83.02212	6611.524
Median	72.00000	144.0000	3.755000	43.00000	93.00000	92.00000	1766.948
Maximum	89.00000	723.0000	17.87000	87.30000	99.00000	99.00000	119172.7
Minimum	36.30000	1.000000	0.010000	1.000000	2.000000	1.000000	1.681350
Std. Dev.	9.507640	124.0862	3.921946	19.92768	23.64007	22.99698	13296.60
Skewness	-0.639367	1.177298	0.649246	-0.220478	-2.078419	-2.280532	3.541946
Kurtosis	2.773314	4.761809	2.374108	1.729071	6.592631	7.391677	18.11539
Jarque-Bera	206.4612	1058.670	254.3604	221.5378	3695.306	4907.701	34112.19
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	203382.8	483964.0	13358.72	112587.8	241868.2	243919.0	19424657
Sum Sq. Dev.	265490.8	45222131	45175.93	1166319.	1641351.	1553266.	5.19E+11
Observations	2938	2938	2938	2938	2938	2938	2938

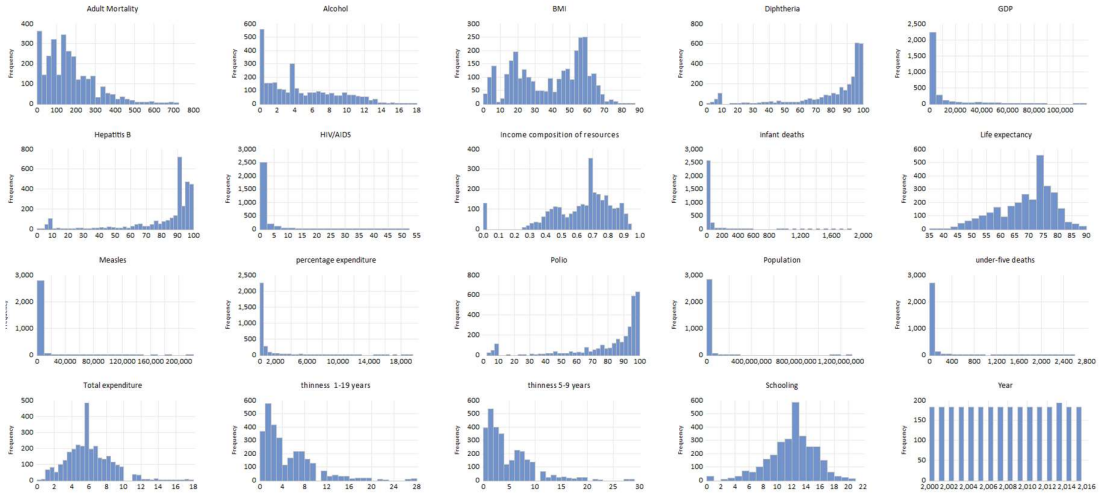
	HIV_AIDS	INCOME_C	INFANT_DE	MEASLES	PERCENTA	POLIO	POPULATION
Mean	1.742103	0.630362	30.30395	2419.592	738.2513	82.55019	12753375
Median	0.100000	0.677000	3.000000	17.00000	64.91291	93.00000	3675929.
Maximum	50.60000	0.948000	1800.000	212183.0	19479.91	99.00000	1.29E+09
Minimum	0.100000	0.000000	0.000000	0.000000	0.000000	3.000000	34.00000
Std. Dev.	5.077785	0.205140	117.9265	11467.27	1987.915	23.35214	53815463
Skewness	5.393357	-1.211907	9.781965	9.436511	4.649676	-2.103789	18.03196
Kurtosis	37.83061	4.689144	118.8433	117.6625	29.52614	6.812043	386.0221
Jarque-Bera	162756.0	1068.462	1689647.	1653075.	96723.15	3946.147	18118470
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	5118.300	1852.003	89033.00	7108762.	2168982.	242532.5	3.75E+10
Sum Sq. Dev.	75727.30	123.5956	40843860	3.86E+11	1.16E+10	1601612.	8.51E+18
Observations	2938	2938	2938	2938	2938	2938	2938

	SCHOOLING	UNDER_FIV	TOTAL_EXP	YEAR	THINNESS__	THINNESS_5
Mean	12.00984	42.03574	5.938190	2007.519	4.821886	4.852144
Median	12.30000	4.000000	5.938190	2008.000	3.300000	3.300000
Maximum	20.70000	2500.000	17.60000	2015.000	27.70000	28.60000
Minimum	0.000000	0.000000	0.370000	2000.000	0.100000	0.100000
Std. Dev.	3.265139	160.4455	2.400274	4.613841	4.397621	4.485854
Skewness	-0.634727	9.490216	0.643592	-0.006406	1.728613	1.794777
Kurtosis	4.119723	112.5641	4.497922	1.786301	7.051398	7.443535
Jarque-Bera	350.7599	1513626.	477.5002	180.3477	3472.500	3994.443
Probability	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
Sum	35284.90	123501.0	17446.40	5898090.	14166.70	14255.60
Sum Sq. Dev.	31311.75	75606527	16920.98	62521.47	56798.84	59100.91
Observations	2938	2938	2938	2938	2938	2938

- Probability of jarque bera is zero in each columns so we can say that there is noi trend in data destribution.
- Some columns are highl;y laptokurtic which denotes that thers is high number of data nearmedian value .
- A smaller standard deviation implies less variability or dispersion in the dataset. The values tend to cluster more closely around the mean.(stanadard deviation is less then mean).

- A larger standard deviation implies greater variability or dispersion in the dataset. The values are more spread out from the mean.(standard deviation is more then mean).
- Minimun and maximum value denotes the range of the data.

1) Data Destribution in the form of graph



- From this above graph we can see the data distribution with use of histogram.
- From this we can say that so many columns is left side skewed.
- Only year data seems equally distributed .

2) Regression analysis

Dependent Variable: LIFE_EXPECTANCY
Method: Least Squares
Date: 01/16/24 Time: 16:24
Sample: 1 2938
Included observations: 2938

Variable	Coefficient	Std. Error	t-Statistic	Prob.
ADULT_MORTALITY	-0.021993	0.000835	-26.33277	0.0000
ALCOHOL	0.187117	0.024102	7.763440	0.0000
BMI	0.060391	0.004855	12.43876	0.0000
GDP	6.87E-05	6.78E-06	10.13421	0.0000
HIV_AIDS	-0.485851	0.018597	-26.12496	0.0000
SCHOOLING	1.021064	0.034263	29.80108	0.0000
UNDER_FIVE_DEATHS	-0.102601	0.006327	-16.21568	0.0000
INFANT_DEATHS	0.134847	0.008581	15.71542	0.0000
C	58.03881	0.429449	135.1471	0.0000

R-squared	0.793331	Mean dependent var	69.22493
Adjusted R-squared	0.792767	S.D. dependent var	9.507640
S.E. of regression	4.328152	Akaike info criterion	5.771217
Sum squared resid	54868.68	Schwarz criterion	5.789553
Log likelihood	-8468.918	Hannan-Quinn criter.	5.777819
F-statistic	1405.429	Durbin-Watson stat	0.648257
Prob(F-statistic)	0.000000		

Dependent variable : Life Expectency

Independent variable: Adult mortality , alcohol , BMI , GDP, HIV/AIDS, schooling, under 5 deaths and infant deaths

Objective: We did this to identify that which factor is affecting life expectency the most

Insights :

- 79% change in dependent variable ia been explained by or independent variables which is good that our selection if varible for this model is good .
- Probability if f stats and all the independent varaible is 0 .so our model and all the variables are significant
- Some of this is affecting negatively to the dependent variable which is under 5 deaths,adult mortality Hiv/aids. If the coefficient is negative, it suggests a negative relationship. An increase in the independent variable is associated with a decrease in the dependent variable.
- If the coefficient is positive, it suggests a positive relationship between the independent variable and the dependent variable. As the independent variable increases, the dependent variable is expected to increase.
- As the coefficient value of GDP is too high so we can say that it affects too much on dependent variable.

Dependent Variable: UNDER_FIVE_DEATHS
Method: Least Squares
Date: 01/17/24 Time: 18:05
Sample: 1 2938
Included observations: 2938

Variable	Coefficient	Std. Error	t-Statistic	Prob.
POLIO	-0.131663	0.010276	-12.81234	0.0000
INFANT_DEATHS	1.351519	0.002035	664.1599	0.0000
C	11.94815	0.892845	13.38211	0.0000

R-squared	0.993626	Mean dependent var	42.03574
Adjusted R-squared	0.993621	S.D. dependent var	160.4455
S.E. of regression	12.81426	Akaike info criterion	7.940015
Sum squared resid	481942.5	Schwarz criterion	7.946127
Log likelihood	-11660.88	Hannan-Quinn criter.	7.942216
F-statistic	228752.0	Durbin-Watson stat	0.321492
Prob(F-statistic)	0.000000		

Dependent variable : Under 5 Deaths

Independent variable: Polio and Infant Deaths

Objective: We did this to identify that which factor is affecting Under 5 Deaths

- Model Fit:
- The R-squared value of 0.9936 indicates that the model explains a very high proportion (99.36%) of the variance in the number of under-five deaths.
- The adjusted R-squared value is also very high at 0.9936, which suggests that the model is not overfitting the data.
- The F-statistic is highly significant (p-value = 0.0000), which further supports the conclusion that the model is a good fit for the data.
- Coefficients:
- The coefficient for polio is negative and statistically significant (p-value = 0.0000). This means that for every one unit increase in polio, the number of under-five deaths is expected to decrease by 0.13 units, on average, holding infant deaths constant.
- The coefficient for infant deaths is positive and statistically significant (p-value = 0.0000). This means that for every one unit increase in infant deaths, the number of under-five deaths is expected to increase by 1.35 units, on average, holding polio constant.
- The coefficient for the squared term of infant deaths is also positive and statistically significant (p-value = 0.0000). This indicates that the relationship between infant deaths and under-five deaths is not linear, but rather curvilinear. The positive coefficient suggests that the rate of increase in under-five deaths slows down as infant deaths increase.
- Other Statistics:
- The standard errors of the coefficients are all relatively small, which suggests that the estimates are precise.
- The Durbin-Watson statistic is 0.321492, which is within the range of normality (1.5 to 2.5), suggesting that there is no autocorrelation in the errors.
- Overall, the regression model appears to be a good fit for the data and provides evidence that both polio and infant deaths are significantly associated with the number

of under-five deaths. The model also suggests that the relationship between infant deaths and under-five deaths is not linear.

Dependent Variable: GDP
Method: Least Squares
Date: 01/17/24 Time: 18:09
Sample: 1 2938
Included observations: 2938

Variable	Coefficient	Std. Error	t-Statistic	Prob.
TOTAL_EXPENDITURE	618.2236	101.5969	6.085064	0.0000
C	2940.395	650.7077	4.518765	0.0000

R-squared	0.012455	Mean dependent var	6611.524
Adjusted R-squared	0.012118	S.D. dependent var	13296.60
S.E. of regression	13215.79	Akaike info criterion	21.81689
Sum squared resid	5.13E+11	Schwarz criterion	21.82097
Log likelihood	-32047.02	Hannan-Quinn criter.	21.81836
F-statistic	37.02801	Durbin-Watson stat	0.851472
Prob(F-statistic)	0.000000		

Dependent variable : GDP

Independent variable: Total Expenditure

Objective: We did this to identify that which factor is affecting GDP

Insights :

- 1.2% change in dependent variable ia been explained by or independent variables which is bad but we can say that all over model is good because probabilty of f stat is significant..
- Probability if f stats and all the independent varaible is 0 .so our model and all the variables are significant
- Coefficient of total expoenditure is affecting negatively to dependent variable.

Dependent Variable: TOTAL_EXPENDITURE
Method: Least Squares
Date: 01/17/24 Time: 19:23
Sample: 1 2938
Included observations: 2938

Variable	Coefficient	Std. Error	t-Statistic	Prob.
SCHOOLING	0.189613	0.022094	8.582034	0.0000
INCOME_COMPOSITION_OF_RESOU...	-0.724849	0.352369	-2.057073	0.0398
GDP	4.90E-06	3.65E-06	1.339868	0.1804
C	4.085511	0.173830	23.50285	0.0000

R-squared	0.050267	Mean dependent var	5.938190
Adjusted R-squared	0.049296	S.D. dependent var	2.400274
S.E. of regression	2.340364	Akaike info criterion	4.539851
Sum squared resid	16070.41	Schwarz criterion	4.548000
Log likelihood	-6665.040	Hannan-Quinn criter.	4.542785
F-statistic	51.76342	Durbin-Watson stat	0.647775
Prob(F-statistic)	0.000000		

Dependent variable :Total Expenditure

Independent variable: Schooling,income composition of resources and GDP

Objective: We did this to identify that which factor is affecting Total Expenditure

Insights :

- 5 % change in dependent variable ia been explained by or independent variables which is bad but we can say that all over model is good because probabilty of f stat is significant.
- Probability if f stats and all the independent varaible is 0 .so our model and all the variables are significant
- Coefficient of income composition of resources is affecting negatively to dependent variable.
- GDP is highly affecting dependent variable as the magnitude of coefficeint is very high