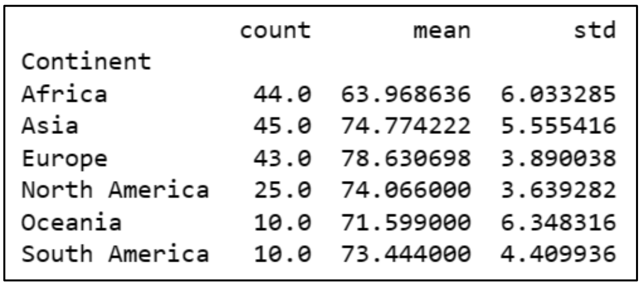
DATA ANALYTICS IN ACTION

PART - A

[Word Count: 749]
‘gxnj57’

Q1a) The average life expectancy stands at approximately 72.67 years.

Q1b) Based on statistics from Figure 1 and the bar graph depicted in Figure 2, it is clear that life expectancy varies most for Oceania, with a standard deviation of about 6.35 years, followed by Africa exhibiting variation of 6.03 years. North America and Europe exhibit lower fluctuations, with variations of 3.64 and 3.89 years, respectively.



```
               count       mean       std
Continent
Africa          44.0  63.968636  6.033285
Asia            45.0  74.774222  5.555416
Europe          43.0  78.630698  3.890038
North America   25.0  74.066000  3.639282
Oceania         10.0  71.599000  6.348316
South America   10.0  73.444000  4.409936
```

Figure 1: Central tendency measures across continents

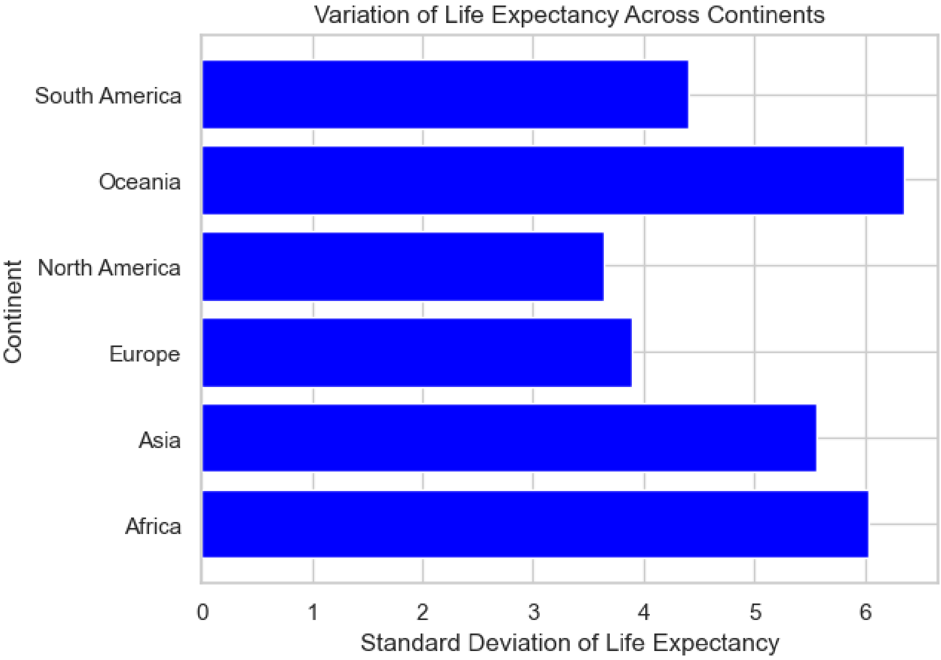


Figure 2: Variation of life expectancy across continents

Q1c) The histogram as seen in Figure 3 depicts a somewhat bell-shaped distribution but with some skewness. A skewness of -0.50 indicates a moderate negative skewness suggesting minor asymmetry in the distribution. The kurtosis of -0.36 suggests that the distribution has a flatter shape as compared to a normal distribution.

Overall, data shows a gentle tendency towards lower values and a relatively less peaked shape since there is a small leftward shift as left skewness suggests that countries with below-average life expectancies are more common.

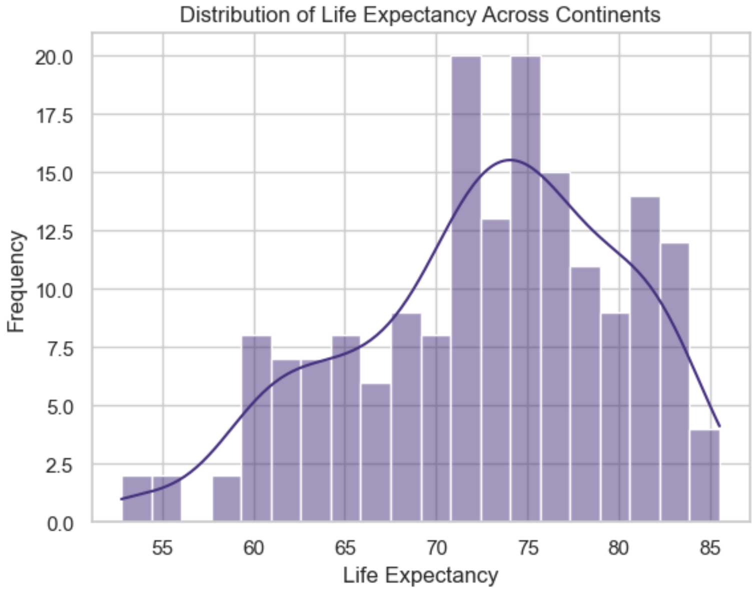


Figure 3: Distribution of Life Expectancy

Q2. Hypothesis Testing

Assumptions:

- Life Expectancy values are normally distributed.
- Significance Level (α) = 0.01 (99% confidence interval).

Hypothesis:

(Ho): The average life expectancy is 70 years.
(H_a): The average life expectancy is not 70 years.

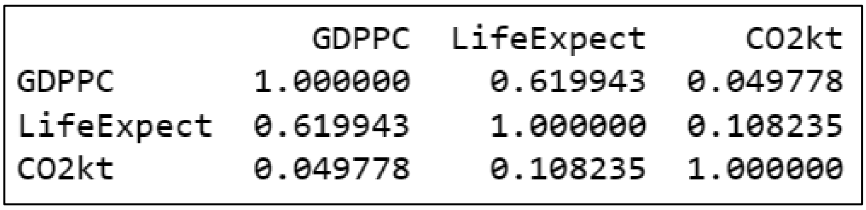
Results (Using a one-sample t-test):

- T-statistic: 4.83
- P-value: $3.01*10^{-6}$
- Significant differences between the sample mean and the estimated population mean are shown by the large t-statistic of 4.83. P-value of nearly zero indicates a strong rejection of the null hypothesis.

Conclusion:

We reject the null hypothesis since the p-value is smaller than the significance level (α=0.01). This indicates that, contrary to what the internal records state, the data offers enough proof to draw the conclusion that the average life expectancy is not 70 years.

Q3a) Relationship between life expectancy and GDPPC of countries.



```
               GDPPC  LifeExpect     CO2kt
GDPPC       1.000000    0.619943  0.049778
LifeExpect  0.619943    1.000000  0.108235
CO2kt       0.049778    0.108235  1.000000
```

Figure 4: Correlation matrix

As seen in Figure 4, there is a very strong correlation between GDPPC and life expectancy. The relationship between GDPPC and CO2Kt isn't as strong (lower value of the correlation coefficient).

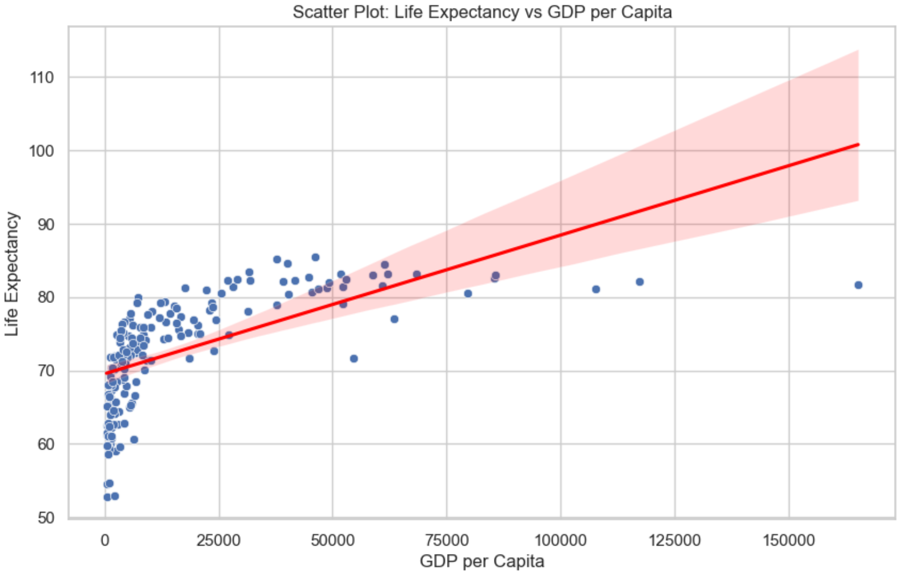
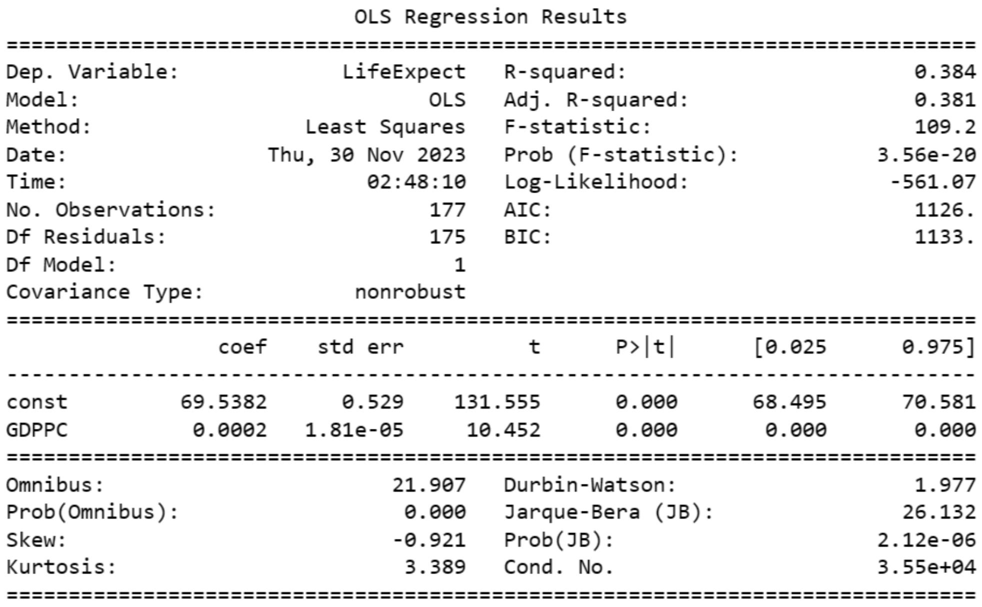


Figure 5: Scatter plot

The scatter plot as seen in Figure 5 visualizes the relationship between life expectancy at birth and GDPPC for different countries. The trend observed suggests that higher GDPPC is associated with higher life expectancy, indicating a positive correlation.

3b) Simple Linear Regression

We take life expectancy as the dependent variable and GDPPC as the independent variable. This is because of the correlation coefficient matrix suggesting the strongest relationship between them. The output of the linear regression looks as given in Figure 6:



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:             LifeExpect   R-squared:                       0.384
Model:                            OLS   Adj. R-squared:                  0.381
Method:                 Least Squares   F-statistic:                     109.2
Date:                Thu, 30 Nov 2023   Prob (F-statistic):           3.56e-20
Time:                        02:48:10   Log-Likelihood:                -561.07
No. Observations:                 177   AIC:                             1126.
Df Residuals:                     175   BIC:                             1133.
Df Model:                           1                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         69.5382      0.529    131.555      0.000      68.495      70.581
GDPPC          0.0002   1.81e-05     10.452      0.000       0.000       0.000
==============================================================================
Omnibus:                       21.907   Durbin-Watson:                   1.977
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               26.132
Skew:                          -0.921   Prob(JB):                     2.12e-06
Kurtosis:                       3.389   Cond. No.                     3.55e+04
==============================================================================
```

Figure 6: Simple Linear Regression

Interpretation:

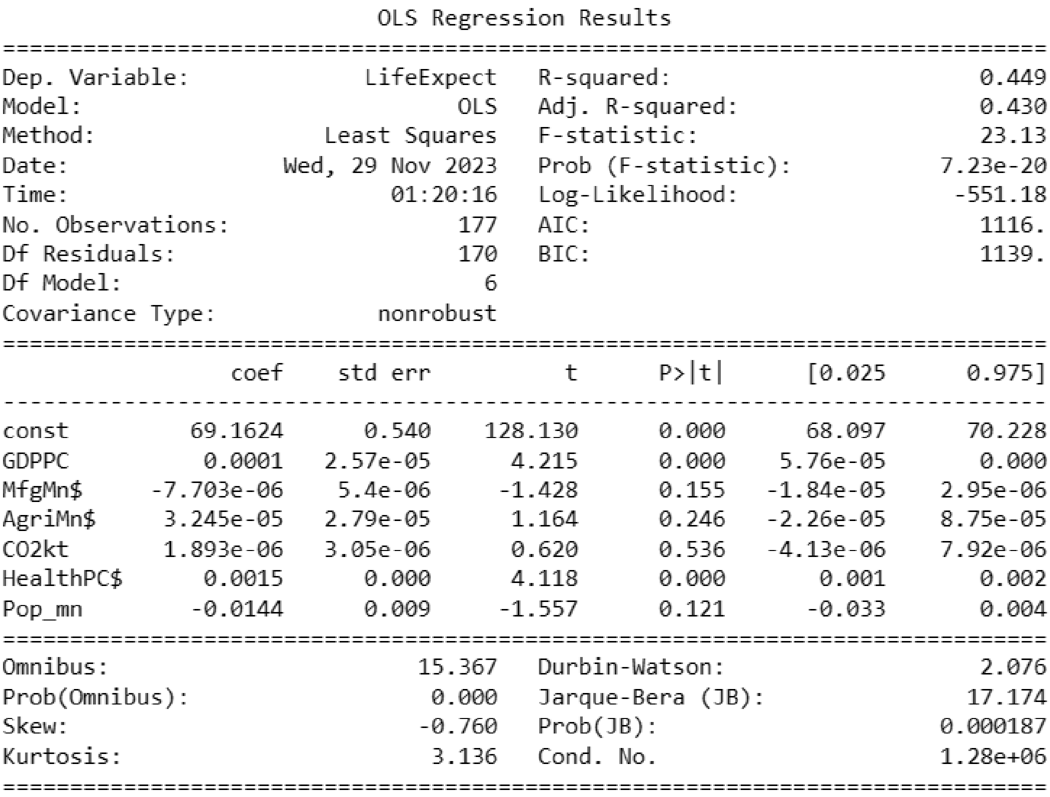
- According to the regression coefficient of 0.0002 (as seen in Figure 6), life expectancy is predicted to rise by approximately 0.0002 units for every unit increase in GDPPC.
- Given that GDP per capita is likely measured in dollars, this suggests that for every additional thousand dollars in GDPPC, life expectancy increases by roughly 0.2 years.
- The variation in GDPPC alone accounts for about 38.4% of the variation in Life Expectancy.

Summary:

A significant correlation between GDPPC and Life Expectancy is demonstrated by the regression model. A one-unit increase in GDPPC correlates with a small increase in Life Expectancy. To put it simply for the client: the model indicates higher life expectancies in nations with greater GDPPC.

Q4) Multiple Regression

Here, we take life expectancy as the dependent variable and various independent variables as seen in Figure 7.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             LifeExpect   R-squared:                       0.449
Model:                            OLS   Adj. R-squared:                  0.430
Method:                 Least Squares   F-statistic:                     23.13
Date:                Wed, 29 Nov 2023   Prob (F-statistic):           7.23e-20
Time:                        01:20:16   Log-Likelihood:                -551.18
No. Observations:                 177   AIC:                             1116.
Df Residuals:                     170   BIC:                             1139.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         69.1624      0.540    128.130      0.000      68.097      70.228
GDPPC          0.0001   2.57e-05      4.215      0.000    5.76e-05       0.000
MfgMn$     -7.703e-06    5.4e-06     -1.428      0.155   -1.84e-05    2.95e-06
AgriMn$     3.245e-05   2.79e-05      1.164      0.246   -2.26e-05    8.75e-05
CO2kt       1.893e-06   3.05e-06      0.620      0.536   -4.13e-06    7.92e-06
HealthPC$      0.0015      0.000      4.118      0.000       0.001       0.002
Pop_mn        -0.0144      0.009     -1.557      0.121      -0.033       0.004
==============================================================================
Omnibus:                       15.367   Durbin-Watson:                   2.076
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               17.174
Skew:                          -0.760   Prob(JB):                     0.000187
Kurtosis:                       3.136   Cond. No.                     1.28e+06
==============================================================================
```

Figure 7: Multiple Linear Regression

Interpretation:

- R-squared: This indicates that the model accounts for approximately 44.9% of the variation in life expectancy.
- Adjusted R-squared: After adjusting for the number of factors, a moderate 43% of the variation in life expectancy is explained.
- The standard error of estimate varies for variables when compared to our coefficients, but the fluctuation is not great enough to conclude that our model is inaccurate.

Coefficients Interpretation:

- There is a slight rise in life expectancy with every unit increase in GDPPC and HealthPC$.
- GDPPC coefficient: implies a 0.0001 increase in Life Expectancy per unit rise. HealthPC$ coefficient: suggests a 0.0015 increase per unit rise in health expenditure. Positive, statistically significant coefficients are present for both.
- The lack of statistical significance in other independent variables implies no significant linear association with life expectancy in this model.

Prediction Accuracy:

- The model explains nearly 45% of the differences in life expectancy across different countries. This indicates a moderate level of prediction accuracy.
- Further analysis or adding new variables could improve prediction accuracy.