DATA ANALYTICS IN ACTION

PART A- Word count: 735

Q1a) The average life expectancy across a diverse range of countries stands at 72.67 years.

Q1b) Among examined continents. Europe demonstrates the highest mean life expectancy of 79.50. Meanwhile, Africa displayed the lowest mean life expectancy among the continents at 64.36 years.

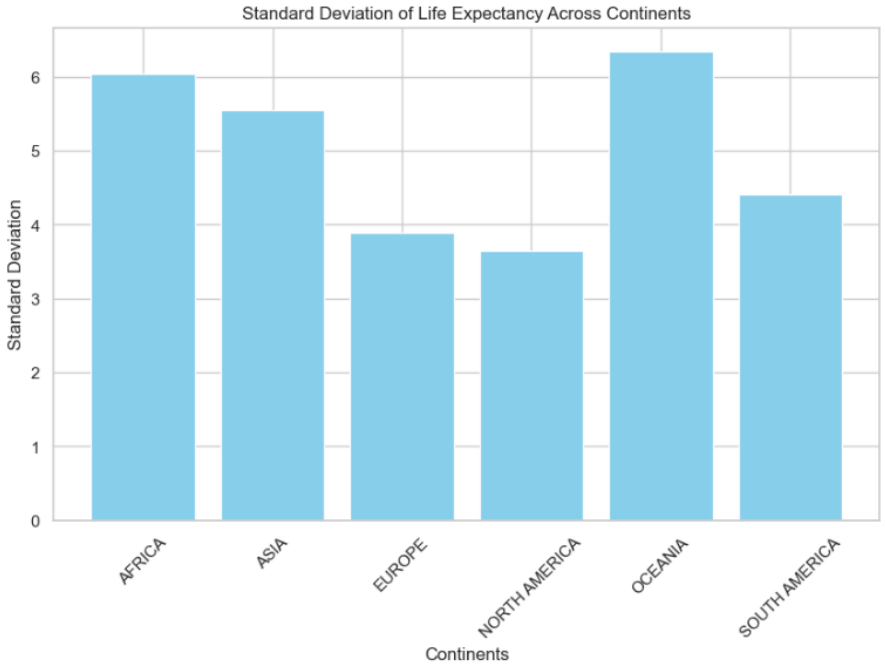

Figure1: This bar chart represents the differences in life expectancy across continents.

INTERPRETATION:

The analysis of life expectancy across continents reveals significant variations. Oceania emerges with the highest variations in life expectancy at 6.34 years. In contrast, North America displays a lower standard deviation of 3.63 years.

Q1c)

ANALYSIS:
A skewness of -0.50 indicates a moderate negative skewness, while not extreme, suggests a minor asymmetry in the distribution. It might indicate that some countries have lower life expectancies, causing a slight deviation from a perfectly symmetric distribution.

A kurtosis of -0.36 implies the data has fewer extreme values than a standard distribution, making it somewhat flatter than the average curve. Overall, the data shows a gentle tendency towards lower values and a relatively less peaked shape

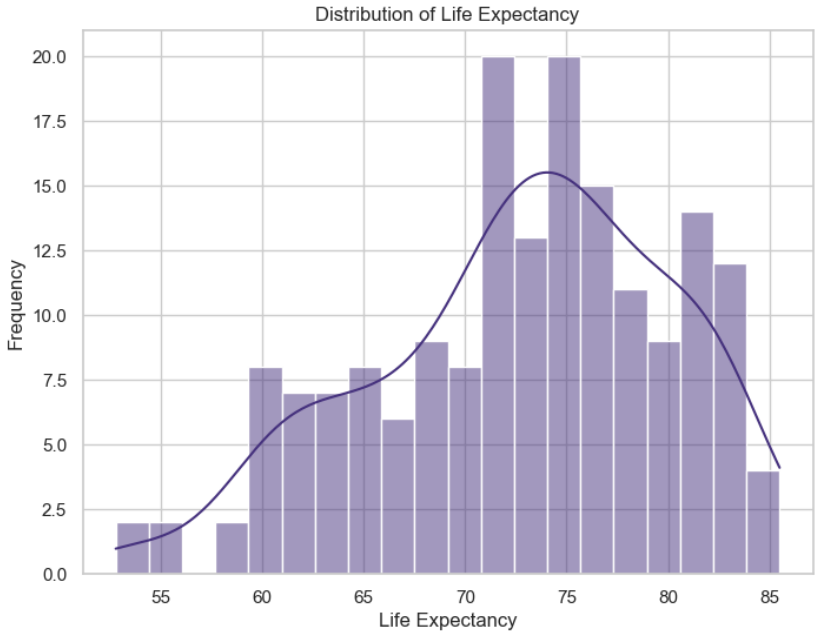

Figure 2: Histogram- Distribution of Life Expectancy

The histogram might display a slight elongation towards the lower values, indicating a subtle shift from perfect symmetry.

Q2) Hypothesis Testing Analysis

ASSUMPTIONS:

- Life expectancy values are normally distributed.
- Significance Level (alpha): Given as 0.01 (99% confidence interval).

HYPOTHESIS:

Null Hypothesis (H0): The average life expectancy is 70 years.
Alternative Hypothesis (H1): The average life expectancy is not 70 years.

INTERPRETATION:

1. T-Statistic Significance:
   The t-statistic of 4.83 indicates that the sample mean is significantly different from the hypothesized mean of 70 years.
2. P-Value Interpretation:
   With a p-value far below the chosen significance level of 0.01, it suggests that observing a sample mean as extreme as 72.67 years, assuming the population mean is 70 years, is highly unlikely (less than 0.01% chance).
3. Confidence Interval:
   As the confidence interval is not specified, but a 99% confidence interval is generally wide, the sample mean of 72.67 years is likely significantly higher than 70 years, supporting the rejection of the null hypothesis.

CONCLUSION:

Hence, we reject the Null Hypothesis. The p-value being lower than the significance level indicates strong evidence against the null hypothesis.

Q3a) Relationship between Life expectancy at birth and the GDP per capita of countries.

As can be seen from figure 3, there is a very strong correlation between life expectancy and GDP per capita.

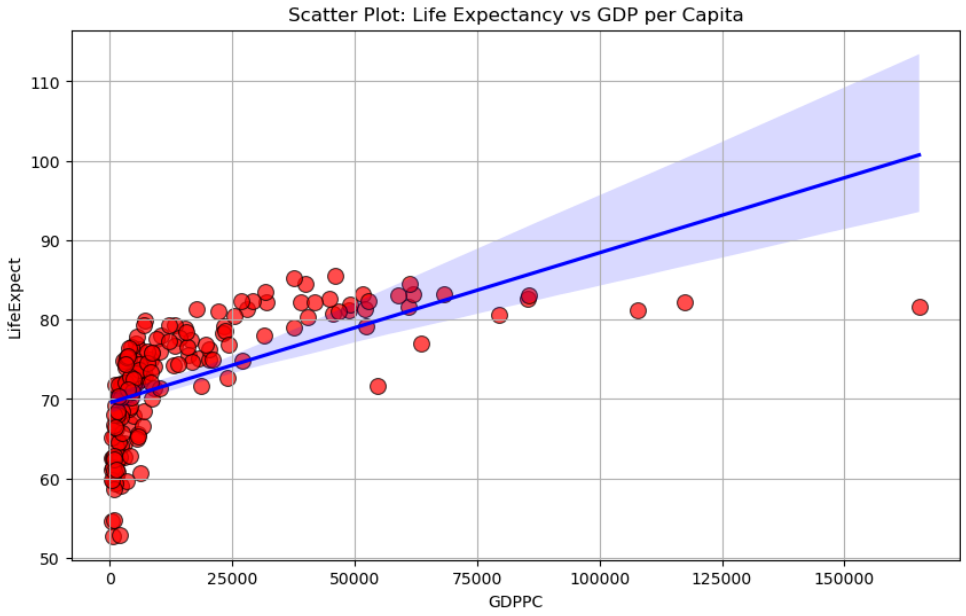

Figure 3: Scatterplot- GDP per capita vs Life expectancy

Q3b) Simple Linear Regression Model

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:	Wed, 29 Nov 2023	Prob (F-statistic):	3.56e-20
Time:	23:25:59	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 4: Simple Linear Regression- OLS Regression model

In this case, we take Life expectancy as the dependent variable and GDP per capita as the independent variable.

INTERPRETATION:

•The model suggests that for every unit increase in GDPPC, there is a predicted increase in Life Expectancy by 0.0002 units.

•Approximately 38.4% of the variation in Life Expectancy can be explained by the variation in GDPPC alone.

•Both GDPPC and the constant term (intercept) significantly contribute to predicting Life Expectancy.

•There might be potential issues with the model's assumptions, such as residuals not perfectly following a normal distribution and mild autocorrelation in residuals.

SUMMARY:

The regression model shows a statistically significant relationship between GDPPC and Life Expectancy. A one-unit increase in GDPPC correlates with a small increase in Life Expectancy.

Q4) MULTIPLE REGRESSION MODEL

In this case, we take Life Expectancy as the dependent variable and various other independent variables. As can be seen from multiple linear regression output,

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:	Tue, 28 Nov 2023	Prob (F-statistic):	7.23e-20
Time:	15:25:35	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
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
CO2kt	1.893e-06	3.05e-06	0.620	0.536	-4.13e-06	7.92e-06
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

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 5: Multiple Linear Regression

INTERPRETATION:

- R-squared and Adjusted R-squared: The R-squared value of 0.449 indicates that approximately 44.9% of the variability in life expectancy can be explained by the included independent variables.
- The adjusted R-squared (0.430) suggests the model accounts for a decent portion of variance while penalizing for additional predictors.
- Standard Error of Estimate: the average deviation of the observed values from the predicted values by the model. A lower value indicates better predictive accuracy.

- Regression Coefficients:
  A unit increase in GDPPC and HealthPC$ is associated with a small increase in life expectancy. Both have statistically significant positive coefficients.
  The other independent variable coefficients lack statistical significance, suggesting these variables might not have a substantial linear relationship with life expectancy in this model.

- Prediction Accuracy: The model exhibits moderate predictive ability. However, caution should be exercised due to insignificant coefficients for certain predictors, indicating other influential factors might be missing. Further analysis or inclusion of additional variables may enhance predictive accuracy.

REFERENCE:

1. Simplilearn. (n.d.). *Hypothesis Testing in Statistics*. Retrieved from https://www.simplilearn.com/tutorials/statistics-tutorial/hypothesis-testing-in-statistics

2. Corporate Finance Institute. (n.d.). *Multiple Linear Regression.* Retrieved from https://corporatefinanceinstitute.com/resources/data-science/multiple-linear-regression/