Link of the .ipynb file:
https://colab.research.google.com/drive/10_HAP7_JXvkHqjndYnqH3AWnYWBD6vA7?usp=sharing

Q1. Using UNHDD data, fit a regression model for GNI (Gross National Income) with the partial effects of HDI (Human Development Index), GII (Gender Inequality Index), life expectancy at birth, and population age 15-64 years (in millions).

a) Which effects are significant?
Ans: From the regression output:

- Human Development Index (HDI): The coefficient for HDI is significant ($p < 0.001$) with a positive impact on GNI. This suggests that as HDI increases, GNI per capita also increases.
- Gender Inequality Index (GII): The coefficient for GII is not statistically significant ($p = 0.138$).
- Life expectancy at birth: This variable is not statistically significant ($p = 0.250$).
- Population age 15-64 years (millions): This variable is also not statistically significant ($p = 0.514$).

b) Perform model diagnostics. Mention the assumptions tested, methods used and their corresponding results.

1. Multicollinearity:
    - The Variance Inflation Factor (VIF) results indicate severe multicollinearity among the predictors:
        - Human Development Index (HDI): VIF = 190.54
        - Life expectancy at birth: VIF = 253.64
        - Gender Inequality Index (GII): VIF = 9.69
        - Population age 15-64 years (millions): VIF = 1.07
    - Typically, a VIF above 10 indicates serious multicollinearity. The extremely high VIF values for HDI and life expectancy suggest that these variables are highly correlated, which can distort the regression estimates.

Linearity:

- The assumption of linearity is crucial, meaning the relationship between the predictors and the response variable should be linear.
- The Residuals vs. Fitted Values plot (as seen in the image provided) suggests a potential non-linearity. The pattern observed indicates a curve, which suggests that the linearity assumption may not hold perfectly.

Normality of Residuals:

- The residuals should be approximately normally distributed.
- The Histogram of Residuals indicates that the residuals are skewed to the right, deviating from a normal distribution.
- In Q-Q Plot

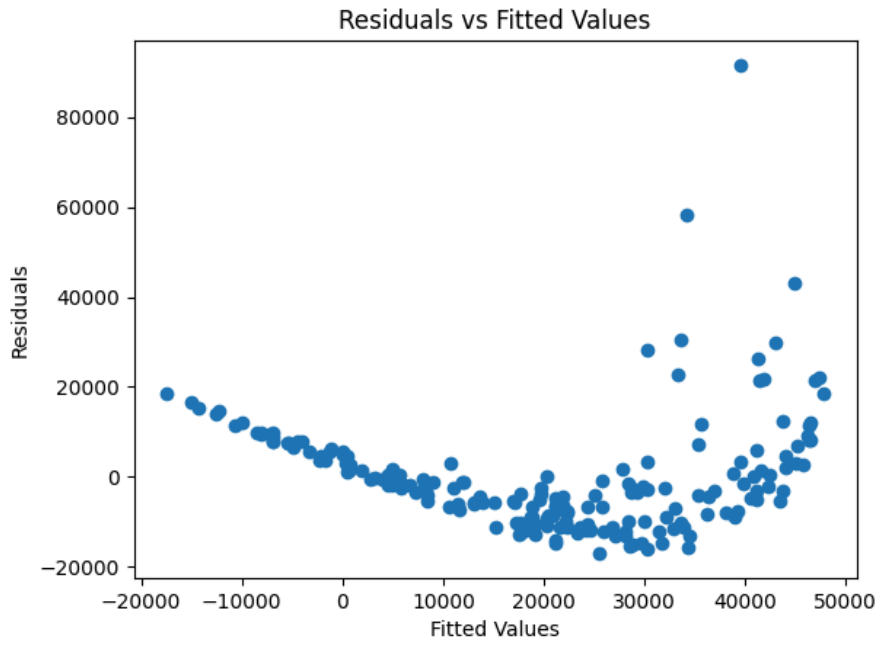
    - The points closely follow the straight line (especially in the middle of the plot), and the residuals are approximately normally distributed
    - Also, the points curve downward at the lower end and upward at the higher end, which indicates a right skew (positive skew)

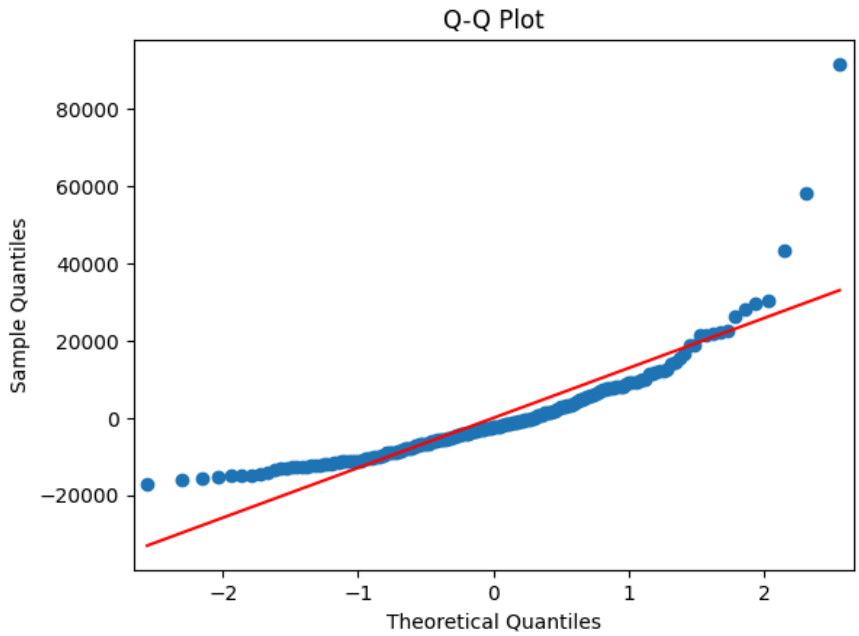
Homoscedasticity:

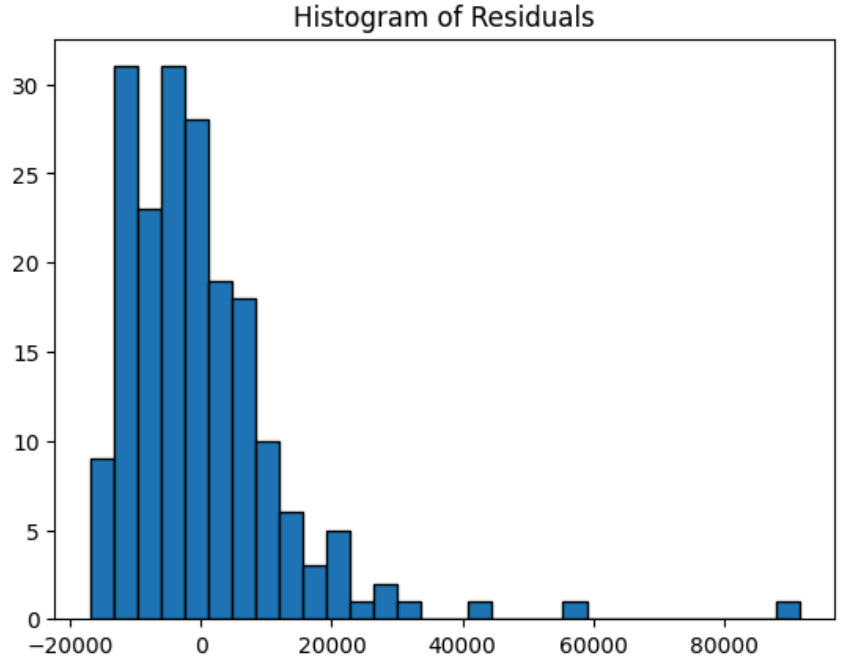
- Homoscedasticity means that the variance of residuals should be constant across all levels of the independent variables.
- The Residuals vs. Fitted Values plot shows a funnel shape, indicating heteroscedasticity (non-constant variance).

Autocorrelation:

- The Durbin-Watson statistic is 1.539, which is within an acceptable range, suggesting that there is no severe autocorrelation in the residuals.







c) Interpret the regression coefficients.

- Intercept (const): The intercept value of −2.948×10^4 suggests that holding all predictors constant, the average GNI per capita would be approximately −29480. However, this is not statistically significant.
- Human Development Index (HDI): The coefficient for HDI is 1.126×10^5, indicating that a one-unit increase in HDI is associated with an increase of approximately $112,600 in GNI per capita, holding other variables constant. This is significant and shows a strong positive relationship.

- Gender Inequality Index (GII): The coefficient is −1.513×10^4, suggesting that as GII increases by one unit, GNI per capita decreases by about $15,130, but this effect is not statistically significant.
- Life Expectancy at Birth: The coefficient is −361.79, suggesting a slight decrease in GNI per capita for each additional year of life expectancy, but this effect is not significant.
- Population age 15-64 years: The coefficient is −6.13, indicating a small decrease in GNI per capita for each million increase in the population aged 15-64 years, but this effect is also not significant.

Conclusion:

- Significant Predictor: Only the Human Development Index (HDI) is a significant predictor of GNI per capita.
- Model Diagnostics: There are potential issues with multicollinearity, non-linearity, non-normality of residuals, and heteroscedasticity, which suggest that the model may need adjustments or that a different modelling approach might be necessary.
- Coefficient Interpretation: HDI has a positive and significant impact on GNI per capita, while other variables do not show statistically significant effects.

Q2. In the above regression model, replace HDI with HDI groups (all other predictors have to be retained in the model). Use “low” as the reference category for HDI groups.
a) How many categories of HDI groups are present in the dataset?
Ans: HDI groups in your dataset are:

- High
- Low
- Medium
- Very High
- hdi_group

b) Interpret the regression coefficients of HDI groups. Are they in tune with the results from question 1?
Ans: Model 1 (using “Low” as the reference category):

- HDI Group Coefficients:
    - HDI Group Codes (categorical coding): 5195.9720 (this value represents the overall impact of the HDI group on GNI per capita across different levels, with "Low" as the reference).
    - Gender Inequality Index: -3.09e+04
    - Life Expectancy: 1085.0250
    - Population age 15-64 years: -3.5932

In this model, the HDI Group Codes show a positive relationship with GNI per capita, meaning that as you move to higher HDI groups, the GNI per capita increases. This is consistent with expectations since higher HDI usually correlates with higher income.

Comparison with Question 1:

The regression results from Question 1 showed that HDI was a significant predictor of GNI per capita, with a large positive coefficient. This is consistent with the finding that higher HDI groups (e.g., "Very High" HDI) have higher GNI per capita compared to the "Low" reference category.

c) Change the reference group to “Very High” and fit the model. Interpret the regression coefficients of the categorical predictor. What are your conclusions about the model?
Ans:
Model 2 (using “Very High” as the reference category):

- Coefficients of Interest:
    - HDI Group_High: -1.295e+04
    - HDI Group_Low: -1.377e+04
    - HDI Group_Medium: -1.581e+04
    - HDI Group_Very high: 1.239e+04 (This is the reference group, so no coefficient is listed in the model, but it influences the other categories).

Interpretation:

- Negative Coefficients for Other Groups: All the coefficients for the other HDI groups (High, Low, Medium) are negative, indicating that countries in these groups have significantly lower GNI per capita compared to countries in the "Very High" HDI group.
- HDI Group_hdi_group: This coefficient is negative but not statistically significant, suggesting it may not be a meaningful category or may need further inspection or correction.

Conclusion:

- The coefficients of the HDI groups confirm that GNI per capita tends to be higher in countries with a "Very High" HDI compared to those in the "High," "Medium," and "Low" HDI groups. This aligns with the expected results from Question 1, where HDI had a significant positive effect on GNI per capita.
- Changing the reference category to "Very High" simply shifts the perspective, but the overall pattern remains consistent: higher HDI groups correlate with higher GNI per capita.

Q3. In the regression model from question 2, include an interaction term between HDI groups and GII.
a) Is the interaction significant?
Ans: To determine if the interaction terms are significant, look at the p-values of the interaction terms in the regression output. The p-values indicate the significance of each coefficient:

- Low_GII: p-value = 0.922
- Medium_GII: p-value = 0.701
- VeryHigh_GII: p-value = 0.137

In this case:

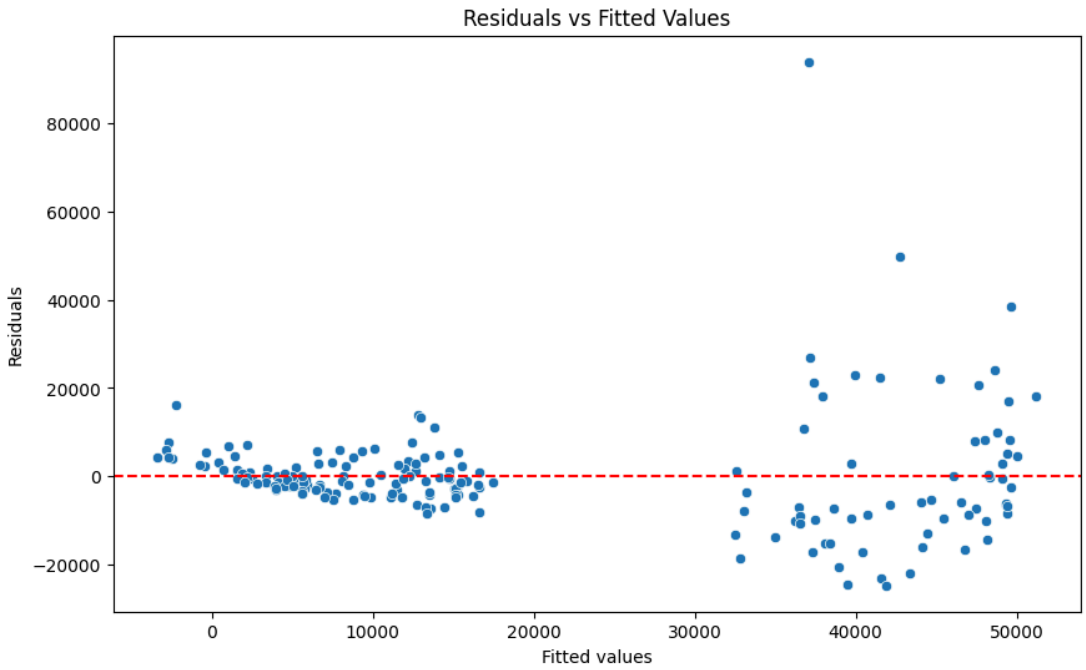
- Low_GII and Medium_GII are not significant (p-values > 0.05).
- VeryHigh_GII is not significant either (p-value = 0.137), but it is closer to the typical threshold of 0.10.

So, based on these p-values, the interaction terms between HDI groups and GII are not statistically significant.

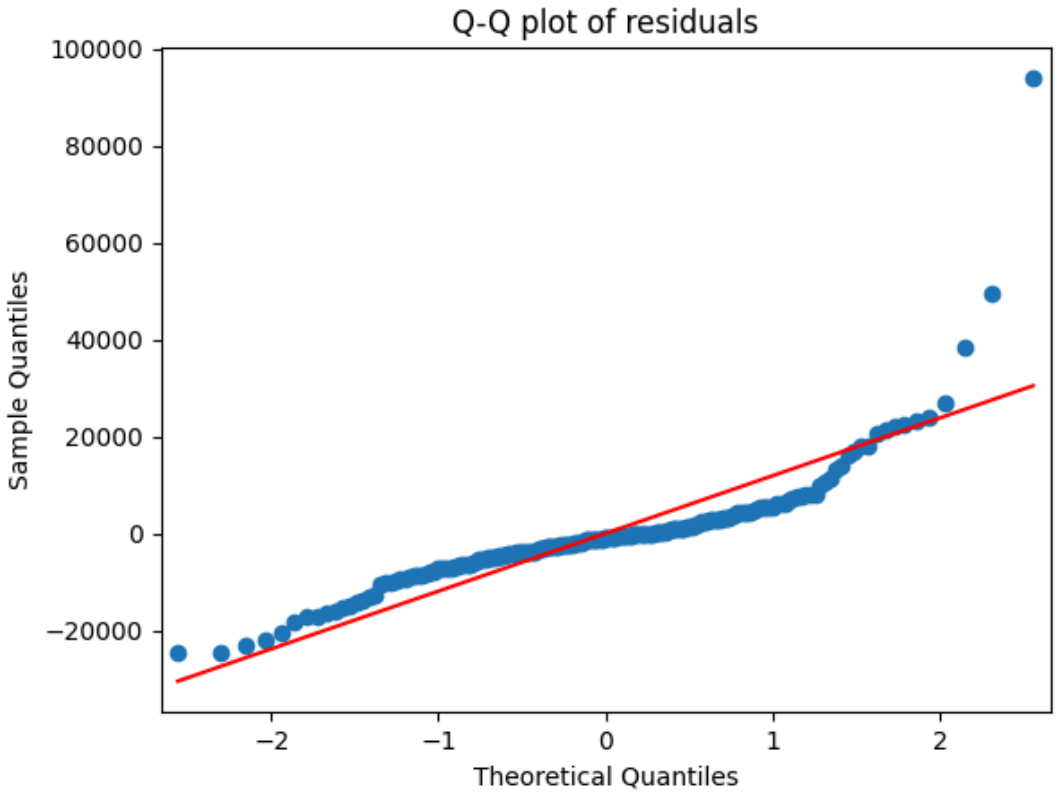
b) Interpret the interaction effects.
Ans: Since the interaction terms are not significant, their effect on the dependent variable (GNI per capita) is not statistically different from zero. However, for the sake of interpretation:

- Low_GII: Represents the interaction effect between the Low HDI group and GII. A coefficient of 2549.15 means that for each unit increase in GII, the GNI per capita increases by approximately 2549.15, but this effect is not statistically significant.
- Medium_GII: Represents the interaction effect between the Medium HDI group and GII. A coefficient of 11540.00 means that for each unit increase in GII, the GNI per capita increases by approximately 11540.00, but this effect is also not statistically significant.
- VeryHigh_GII: Represents the interaction effect between the Very High HDI group and GII. A coefficient of -36500.00 means that for each unit increase in GII, the GNI per capita decreases by approximately 36500.00, which is also not statistically significant.

c) Perform model diagnostics and report their results.



Ans:



From the model summary, you have some key diagnostic metrics:

- R-squared: 0.683
    - Indicates that approximately 68.3% of the variance in GNI per capita is explained by the model.
- Adjusted R-squared: 0.667

    - Adjusted for the number of predictors in the model.
- F-statistic: 43.02 with a p-value of 2.38e-40
    - Indicates that the model is statistically significant overall.
    - The model is statistically significant overall, meaning that at least one of the predictors in the model is significantly associated with the Gross National Income (GNI) per capita.
    - In simpler terms, the model as a whole does a good job of explaining the variation in GNI per capita, even though individual predictors (like the interaction terms) are not significant.
- Omnibus: 167.782 with a p-value of 0.000
    - Suggests that the residuals are not normally distributed.
- Durbin-Watson: 2.032
    - Close to 2, indicating no significant autocorrelation in the residuals.
- Jarque-Bera (JB): 3767.296 with a p-value of 0.00
    - Indicates that the residuals are not normally distributed.
- Condition Number: 5.56e+03
    - A large value indicates potential multicollinearity issues.

Analysis of VIFs

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases due to collinearity. High multicollinearity can inflate the standard errors of the coefficients and make it difficult to assess the individual effect of each predictor.
VIF Thresholds:

- VIF > 10: Indicates high multicollinearity.
- VIF > 5: Indicates moderate multicollinearity.

Results Interpretation:

	Variable	VIF
0	const	559.848211
1	Gender Inequality Index	15.364918
2	Life expectancy at birth (years)	4.470789
3	Population age 15-64 years (millions)	1.051875
4	HDI Group_Low	27.840276
5	HDI Group_Medium	32.683724
6	HDI Group_Very high	17.944267
7	Low_GII	39.006839
8	Medium_GII	40.051531
9	VeryHigh_GII	7.713681

High VIF values for the Gender Inequality Index, HDI Group_Low, HDI Group_Medium, Low_GII, and Medium_GII suggest multicollinearity. These predictors are highly correlated with each other, which could distort the coefficient estimates.

Residuals vs. Fitted Values Plot:

- The plot shows a non-random pattern, particularly at higher fitted values, which may indicate issues with heteroscedasticity or model misspecification.

Q-Q Plot for Residuals:

- The Q-Q plot suggests that the residuals deviate from normality, particularly in the tails. This indicates that the assumption of normality of residuals may not hold, which could affect the validity of statistical tests and confidence intervals.

d) What are your conclusions about the regression model?

1. Multicollinearity: The presence of high VIF values suggests significant multicollinearity in the model, which could compromise the interpretability of the regression coefficients.
2. Model Fit: Although the model explains a reasonable proportion of the variance (R-squared = 0.683), the issues with multicollinearity and non-normality of residuals suggest that the model may not be the best fit for the data.
3. Interaction Terms: The interaction terms are not significant, suggesting that the relationship between GII and GNI per capita does not differ significantly across different HDI groups.
4. Diagnostics: The residual diagnostics indicate potential issues with the model assumptions, particularly normality and homoscedasticity, which could affect the reliability of the model's estimates and inferences.