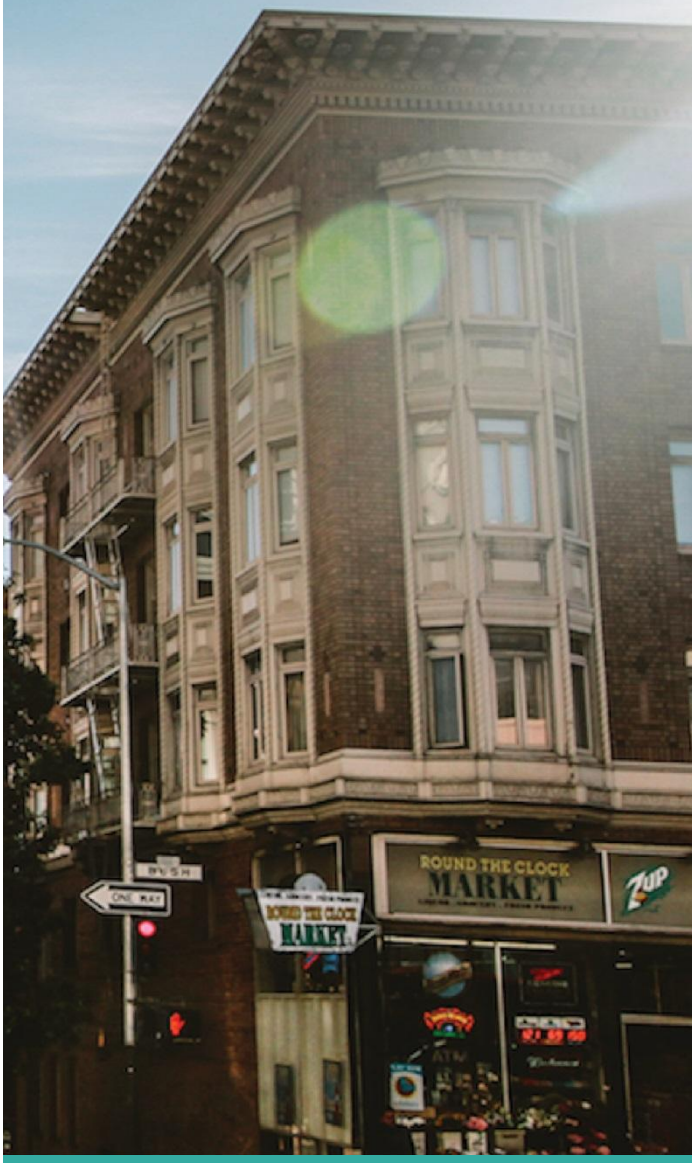
LAB 1: REGRESSION ANALYSIS DIAGNOSE TEST



OCTOBER 22

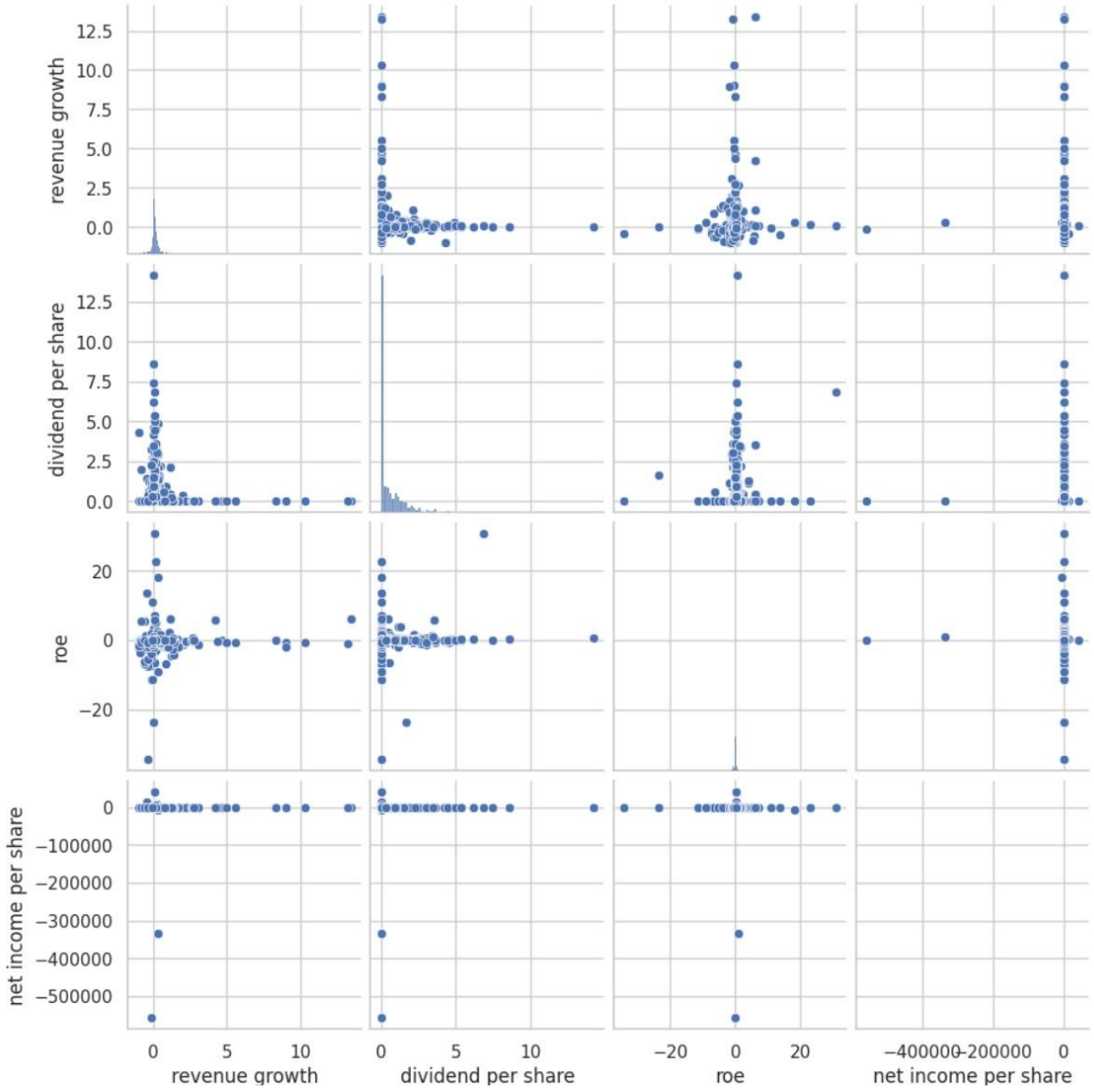
BSAN6070 INTRODUCTION TO MACHINE LEARNING

Authored by: Andres Sebastian Gaibor Heredia

1. **Utilize your dataset and analyze it using the provided code.**

1.1 Descriptive Analytics

This pairplot provides insights into the relationships, spread, and potential outliers among the variables: revenue growth, dividend per share, return on equity and net income per share for 3,998 companies in the year 2018. I chose those independent variables because, in my opinion, they should have an impact on the price of the stock and consequently on the target dependent variable: P/E ratio. Considering system limitations, I had to choose only those four variables to conduct the exploratory data analysis (pairplot) and had to select a subset of 1,000 records instead of the full 3,998 records in the preprocessed dataset. The resulting plot was the following:



Key Observations

Skewed Distributions in Individual Variables

Revenue Growth, Dividend per Share, and Net Income per Share have highly right-skewed distributions with most values concentrated near zero. ROE has a more even spread but still contains outliers.

Potential Outliers

Several extreme values appear in Net Income per Share, with values as low as -500,000, indicating highly negative earnings in some cases.

ROE and Dividend per Share show clustering near zero, with a few extreme positive values.

Relationship Analysis

Revenue Growth vs. Dividend per Share

No strong linear relationship is evident. Most companies seem to have low dividend per share despite variations in revenue growth.

ROE vs. Revenue Growth

Some clustering near 0% ROE, suggesting that many companies have either low profitability or low revenue growth.

A few extreme cases show high negative ROE values, indicating companies with major losses.

Net Income per Share vs. Other Variables

The distribution suggests severe outliers, with companies reporting very large losses.

Most values are concentrated near zero, suggesting that only a few companies generate significantly positive net income per share.

Conclusion

The pairplot highlights significant outliers, skewed distributions, and weak direct relationships among financial variables. Further investigation, particularly in handling extreme values, will improve predictive modeling and financial insights.

1.2 OLS results

By running an ordinary least squares regression on P/E based on the independent variables: revenue growth, dividend per share, profit margin, return on equity, free cash flow, asset growth, debt growth, R&D expenses, and net income per share we obtain the following output:

Model Summary:

OLS Regression Results

Dep. Variable:	pe ratio	R-squared:	0.002
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.8162
Date:	Sun, 16 Feb 2025	Prob (F-statistic):	0.601
Time:	15:46:59	Log-Likelihood:	-24360.
No. Observations:	3998	AIC:	4.874e+04
Df Residuals:	3988	BIC:	4.880e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	24.0301	1.896	12.676	0.000	20.313	27.747
revenue growth	-0.0019	0.008	-0.225	0.822	-0.018	0.015
dividend per share	0.3822	1.171	0.327	0.744	-1.913	2.677
profit margin	0.0067	0.009	0.716	0.474	-0.012	0.025
roe	-1.287e-06	9.63e-06	-0.134	0.894	-2.02e-05	1.76e-05
free cash flow	3.199e-10	6.04e-10	0.530	0.596	-8.64e-10	1.5e-09
asset growth	-1.0226	2.047	-0.499	0.617	-5.037	2.992
debt growth	-0.0010	0.005	-0.215	0.830	-0.010	0.008
r&d expenses	4.13e-09	2.08e-09	1.985	0.047	5.12e-11	8.21e-09
net income per share	-9.394e-07	3.29e-06	-0.286	0.775	-7.38e-06	5.5e-06

Omnibus:	8755.956	Durbin-Watson:	2.014
Prob(Omnibus):	0.000	Jarque-Bera (JB):	47508333.879
Skew:	19.675	Prob(JB):	0.00
Kurtosis:	535.582	Cond. No.	4.04e+09

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 4.04e+09. This might indicate that there are strong multicollinearity or other numerical problems.

We can analyze the output for OLS based on the following

Model Performance & Fit

R-squared: 0.002: This indicates that the model explains only 0.2% of the variance in the P/E ratio, suggesting very poor explanatory power.

Adjusted R-squared: -0.000: The adjusted R^2 is almost zero, implying that adding independent variables did not improve the model's explanatory power.

F-statistic: 0.8162 with p-value = 0.601: This high p-value suggests that the overall model is not statistically significant.

Conclusion: The model is not effective in explaining variations in the P/E ratio, indicating that other variables (not included in the model) may be better predictors.

Significance

The only significant variable is R&D Expenses (p = 0.047), meaning there is a weak statistical relationship between P/E ratio and R&D expenses.

All other variables have p-values > 0.05, indicating they are not statistically significant predictors of P/E ratio. The Intercept (const = 24.03, p < 0.001) predicts a base P/E ratio of 24.03 when all independent variables are zero.

Potential Model Issues

Omnibus: 8755.956, Prob(Omnibus) = 0.000 → Indicates non-normal residuals, suggesting the model violates normality assumptions.

Jarque-Bera (JB) = 47,508,333 with p = 0.00 → Extremely high skewness and kurtosis (19.675 and 535.582, respectively), indicating severe non-normality in the data.

Conclusion: The model has serious violations of normality and high skewness and kurtosis, making its predictions unreliable.

1.3 Durbin-Watson Statistic

The Durbin-Watson (DW) statistic is used to detect the presence of autocorrelation (serial correlation) in the residuals of a regression model.

Since the DW statistic is 2.0135, it is very close to 2.0, it suggests that there is no significant autocorrelation in the residuals. No systematic pattern in errors, hence the residuals are likely independent.

Therefore, based on DW statistic we can conclude the regression model meets the independent errors assumption required for OLS.

1.4 Breusch-Pagan test results and Lagrange multiplier statistic

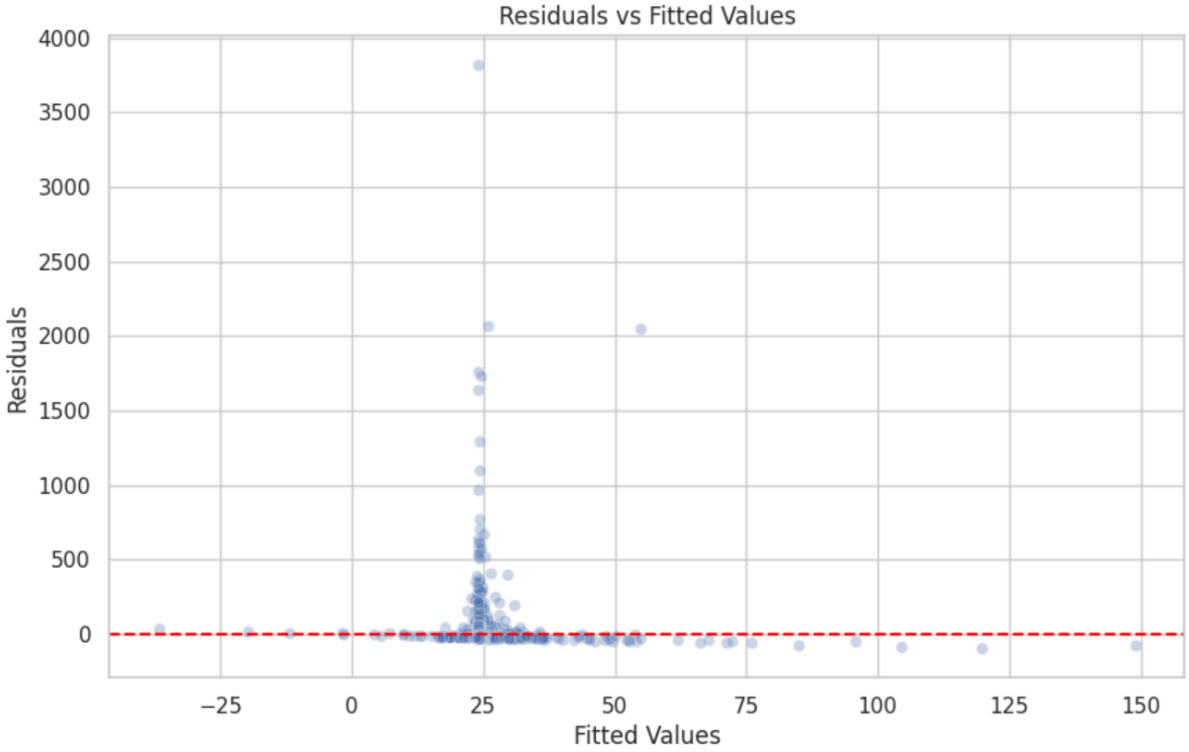
Based on both the residual vs. fitted values plot and the Breusch-Pagan test results, we can draw the following conclusions regarding the assumptions of the OLS regression model.

- ***No Significant Heteroscedasticity (Breusch-Pagan Test):*** The p-value (0.959) from the Breusch-Pagan test indicates that we fail to reject the null hypothesis, meaning no strong evidence of heteroscedasticity.

The low Lagrange Multiplier statistic (3.128) and high p-value (0.959) confirm that heteroscedasticity is not a concern. A high LM statistic suggests strong evidence of heteroscedasticity, while a low LM statistic suggests homoscedasticity (constant variance).

This suggests that the residual variance does not change systematically with fitted values, which aligns with the homoscedasticity assumption.
The assumption of constant variance (homoscedasticity) holds based on statistical testing.

However, we must visually inspect the residual plot for further verification.

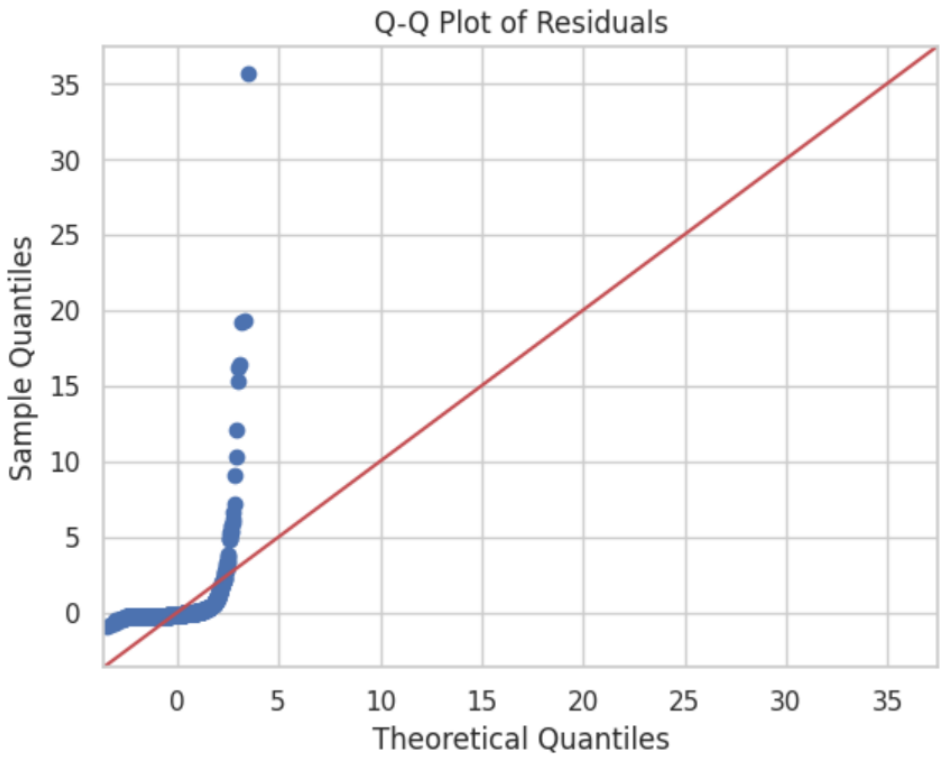


By inspecting the Residuals vs. Fitted Plot we find that although the Breusch-Pagan test indicates no formal heteroscedasticity, the residual plot shows irregularities:
Clustered residuals near 0 with a few extreme outliers reaching values above 2,000-4,000. Most fitted values are concentrated near 25, with extreme spread for residuals at that point.
The spread of residuals does not seem completely random, which could suggest model misspecification or influential outliers.
The extreme residual values suggest influential data points that may be distorting the regression.
These could be:

- Extreme values in the dependent variable (P/E Ratio)
- Influential observations that need further investigation
- Non-linearity, which may require transformations (e.g., log transformation).

1.4 Shapiro-Wilk test statistic

The Shapiro-Wilk test is used to assess whether the residuals of a regression model follow a normal distribution, which is a key assumption in Ordinary Least Squares (OLS) regression. The Q-Q plot visually compares the residuals' distribution to a normal distribution.



Shapiro-Wilk Test Statistic: 0.1557 (Very low)
p-value: 4.13×10^{-86} (Extremely small)
Q-Q Plot: Residuals do not align with the red theoretical normal line.
Clear evidence of heavy right skew (long tail of residuals).

The p-value is much smaller than 0.05, meaning we **reject the null hypothesis of normality.**
The residuals are highly non-normal, indicating that OLS model assumptions are violated.

Non-normal residuals can lead to biased hypothesis tests and confidence intervals.
Standard t-tests and p-values for coefficients may be unreliable.
The Durbin-Watson test was acceptable, so serial correlation is not the issue—rather, outliers or skewness are causing problems.

1.5 Variance Inflation Factor

The Variance Inflation Factor (VIF) is used to detect multicollinearity among independent variables in a regression model. High VIF values indicate strong correlations between predictors, which can distort coefficient estimates and lead to unreliable statistical inferences.

All VIF values are very low, ranging from 1.000 to 1.248. Hence, none of the independent variables are highly correlated.

Conclusion:
There is no evidence of multicollinearity, meaning:

- The independent variables are not redundant.
- The OLS estimates remain stable and reliable.
- There is no need to remove or combine variables.

This suggests that poor model performance (low R^2, high residual variance) is not due to multicollinearity.
The model's predictive power is likely limited by other factors (e.g., non-linearity, outliers, or omitted variables).

2. **What is the purpose of the diagnostic test? Why are we conducting it?**

Regression diagnostics are performed to validate model OLS assumptions, detect potential issues, and ensure that the regression model produces reliable and interpretable results.

Durbin-Watson Statistic (Test for Autocorrelation)

- Detects autocorrelation (serial correlation) in residuals.
- Measures if error terms are independent or exhibit patterns over time.

Why Conduct It:

- OLS Assumption: **Residuals must be independent.**
- If autocorrelation exists:
    - Standard errors may be biased.
    - Hypothesis tests may produce misleading p-values.

Breusch-Pagan Test (Test for Heteroscedasticity)

- Detects heteroscedasticity (unequal variance of residuals).
- Checks if the variance of errors changes with predictor values.

Why Conduct It:

- OLS Assumption: **Homoscedasticity** (constant variance of residuals).
- If heteroscedasticity exists:
    - Standard errors are unreliable, affecting t-tests and confidence intervals.
    - Solution: Robust standard errors or transformations (e.g., log transformation).

Shapiro-Wilk Test (Test for Normality of Residuals)

- Tests if residuals follow a normal distribution.
- Normal residuals ensure valid hypothesis testing (t-tests, confidence intervals, and p-values).

Why Conduct It:

- OLS Assumption: **Residuals must be normally distributed.**
- If residuals are non-normal:
    - Inference becomes unreliable (biased standard errors, incorrect hypothesis testing).

Variance Inflation Factor (VIF - Test for Multicollinearity)

- Measures multicollinearity (high correlation between independent variables).
- High multicollinearity distorts coefficient estimates, making them unstable.

Why Conduct It:

- OLS Assumption: **No Multicollinearity**.
- If multicollinearity exists:
    - Regression coefficients fluctuate widely when adding/removing variables.
    - P-values and standard errors become unreliable.

3. **Is your model robust and rigorous? Why?**

Based on the results of all diagnostic tests, the Ordinary Least Squares (OLS) regression model is NOT robust or rigorous. There are several major issues that undermine its validity and reliability.

Weaknesses:

1. Residuals are not Normally Distributed (Shapiro-Wilk p < 1e-86)
    - The OLS assumption of normality is violated, meaning standard errors and p-values are unreliable.
    - This could be due to severe outliers.
2. Residuals vs. Fitted Plot Shows Outliers
    - The model is being distorted by extreme values, reducing the accuracy of coefficient estimates.
    - Influential points could be overstating or understating relationships between variables.
3. Extremely Low R-squared (0.002)
    - The model explains only 0.2% of the variance in P/E ratio.
    - Independent variables do not contribute meaningfully to explaining the target variable.
    - Possible causes:
        - The true relationship may not be linear.
        - Important variables may be missing.

Strengths:

1. No Autocorrelation (Durbin-Watson ≈ 2.0)
    - Residuals are independent, meaning no time-series bias.
2. No Multicollinearity (Low VIF)
    - Predictor variables are not redundant, so the regression does not suffer from unstable coefficients.

Conclusion: OLS Model is not robust.
The OLS regression model fails to meet key assumptions (normality, meaningful R^2, presence of outliers), making it unreliable for inference or prediction. However, the absence of multicollinearity and heteroscedasticity means the model structure is stable, but its explanatory power is almost zero.

4. What actions can we take if the variable does not meet the regression assumptions?

Linearity: If the relationship between independent and dependent variables is not linear, we could consider applying the following:

- Explore non-linear relationships.
- Consider adding interaction terms or new explanatory variables that may be relevant.

Independence: Residuals are not correlated. If this assumption is violated, we could consider:

- Using models like ARIMA.

Homoscedasticity: Constant variance of residuals. To address a violation of this assumption we can:

- Investigate and remove outliers:
- Compute Cook's Distance to identify influential points.
- Use robust regression methods like Huber Regression to reduce outlier influence.

Normality of Errors: If the distribution of errors is not normal, we can:

- Apply a log transformation to the dependent variable (P/E ratio) if it is skewed.
- Use Quantile Regression or Generalized Least Squares (GLS).

No Multicollinearity: If the explanatory variables show correlation, we will have to address this by:

- Removing highly correlated variables
- Using Principal Component Analysis (PCA).

Finally, if OLS is not suitable then we could explore alternative modeling approaches, such as:

- Decision Trees or Random Forests that might capture relationships better if they are non-linear.
- Machine Learning models (e.g., Gradient Boosting, Neural Networks) could identify patterns missed by OLS.