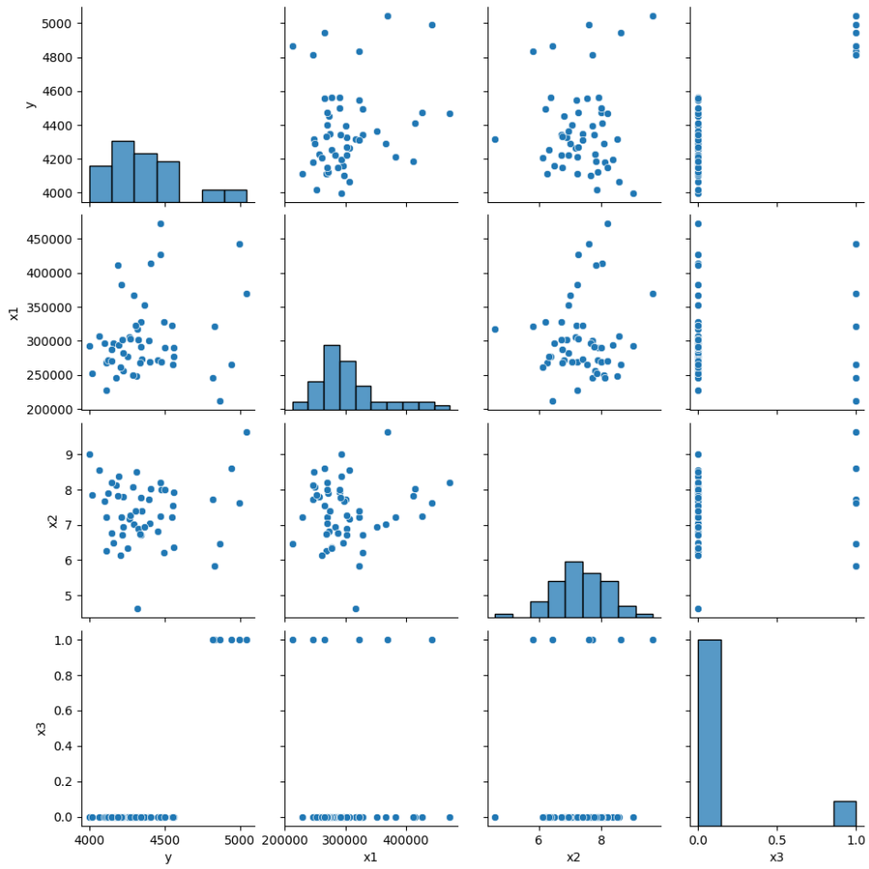
1. (i). Obtain the scatter plot matrix and the correlation matrix. What information do these diagnostic aids provide here?



- **The number of cases shipped (X1)**
- **The indirect costs of the total labor hours as a percentage (X2)**
- **A qualitative predictor called holiday that is coded 1 if the week has a holiday and 0 otherwise (X3)**
- **The total labor hours (Y)**

Correlation Matrix:

	y	x1	x2	x3
y	1	0.207665	0.06003	0.810579
x1	0.207665	1	0.084896	0.045657
x2	0.06003	0.084896	1	0.113371
x3	0.810579	0.045657	0.113371	1

Interpretation:

These values are quite low, indicating that there is no significant pairwise correlation between the predictor variables. This suggests that MULTICOLLINEARITY may not be a major issue based on pairwise correlations alone.

Linear Relationship -

Total Labor Hours to Number of Cases Shipped: **Weak**

Total Labor Hours to Indirect Costs of the total labor hours as a percentage: **Very Weak**

Total Labor Hours to Holiday in the week: **Strong**

2. (ii). Write a multiple regression model to the data for three predictor variables. State the estimate regression function.

The regression model will be:

Y = 4149.8872+0.0008 * (X1 = Number of Cases Shipped) -13.166 * (X2= Indirect Costs of the Total Labor Hours) +623.5545 * (X3=Holiday)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.688
Model:	OLS	Adj. R-squared:	0.669
Method:	Least Squares	F-statistic:	35.34
Date:	Sun, 30 Jun 2024	Prob (F-statistic):	3.32e-12
Time:	20:02:42	Log-Likelihood:	-329.88
No. Observations:	52	AIC:	667.8
Df Residuals:	48	BIC:	675.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4149.8872	195.565	21.220	0.000	3756.677	4543.098
x1	0.0008	0.000	2.159	0.036	5.41e-05	0.002
x2	-13.1660	23.092	-0.570	0.571	-59.595	33.263
x3	623.5545	62.641	9.954	0.000	497.606	749.503

Omnibus:	1.532	Durbin-Watson:	2.298
Prob(Omnibus):	0.465	Jarque-Bera (JB):	1.504
Skew:	0.332	Prob(JB):	0.471
Kurtosis:	2.496	Cond. No.	3.04e+06

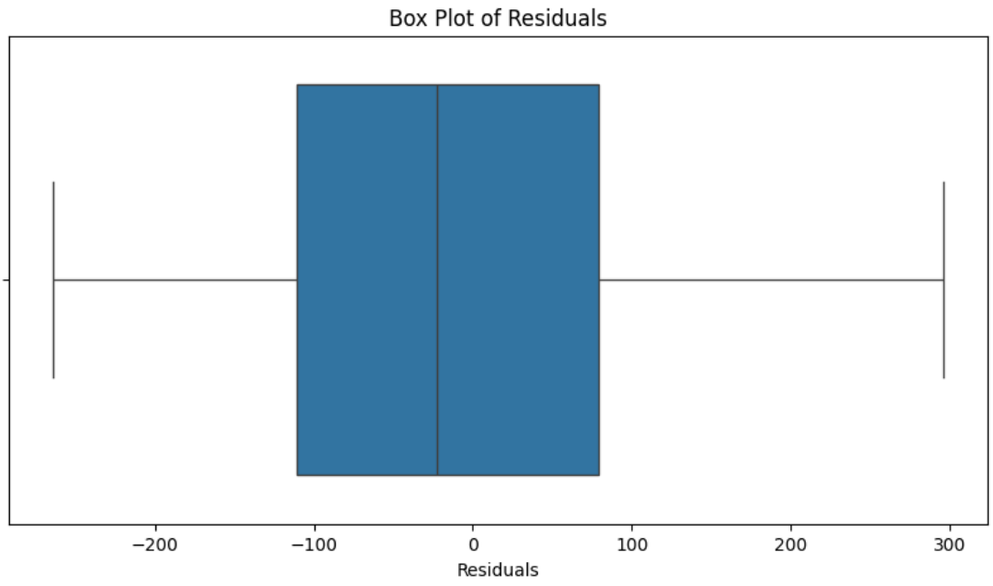
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.04e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Interpretation:

- Significant predictors (with p-values < 0.05) are important for predicting Y.
- High R-squared value suggests the model explains a large proportion of the variance in Y.

---

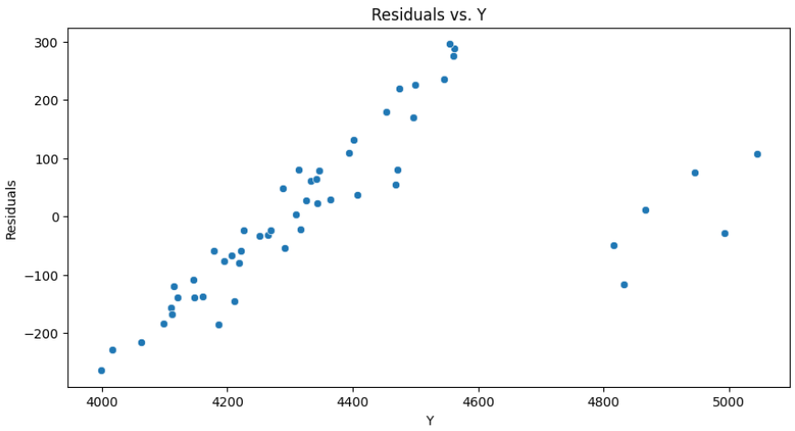
3. (iii). Obtain the residuals and prepare a box plot of the residuals. What information does this plot provide?



Interpretation:

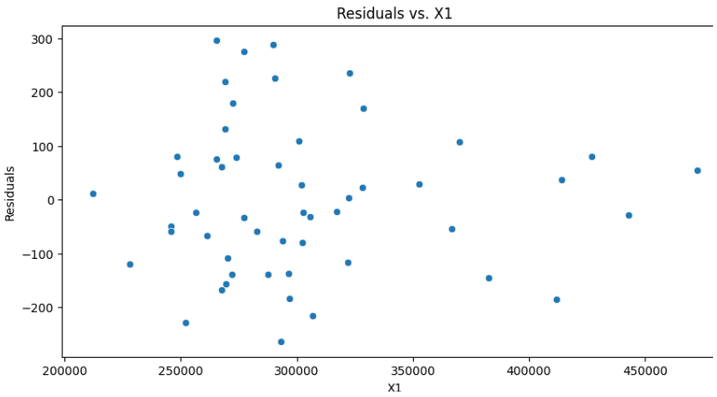
From the box plot of Residual It is clear that the mean of the residuals is different from zero and it does not follow the assumption of normal distribution.

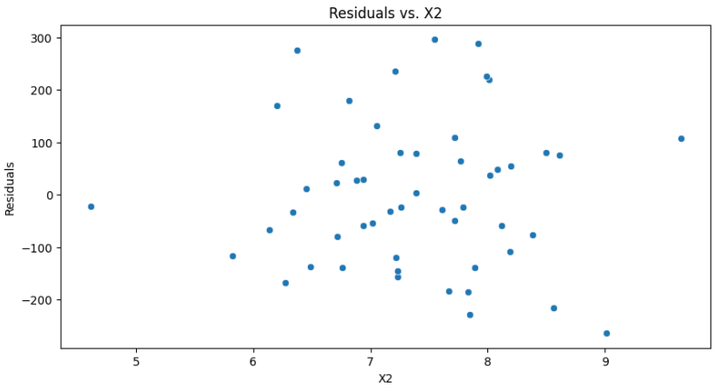
4. (iv). Plot the residuals against Y , X1 , X2 , X3 , and X1 X2 on separate graphs. Also prepare a normal probability plot. Interpret the plots and summarize your findings.

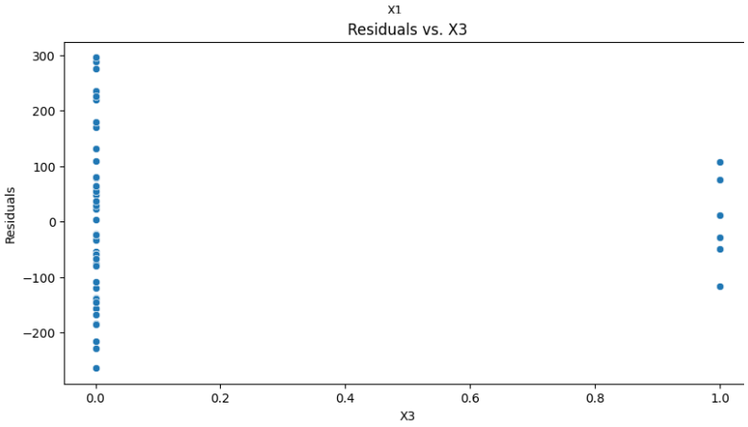


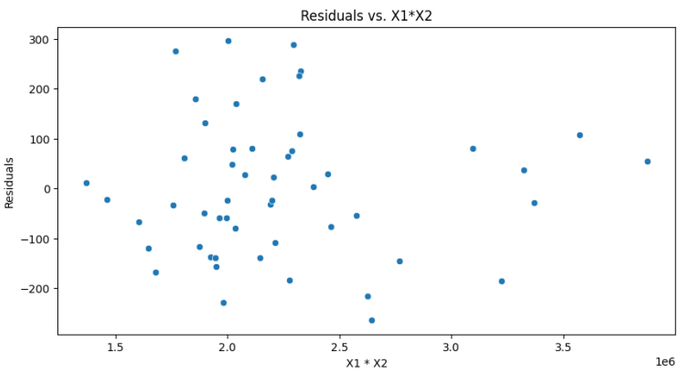
Interpretation:

From the graph, it is clear that residuals are positively related with y = Total Labor Hours





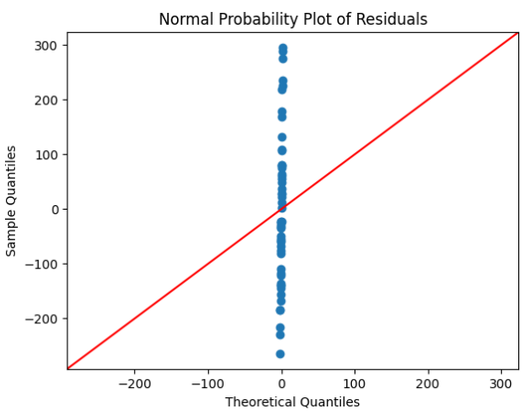




Interpretation:

From the residual plot it is clear that plots are showing nonconstant error variance.

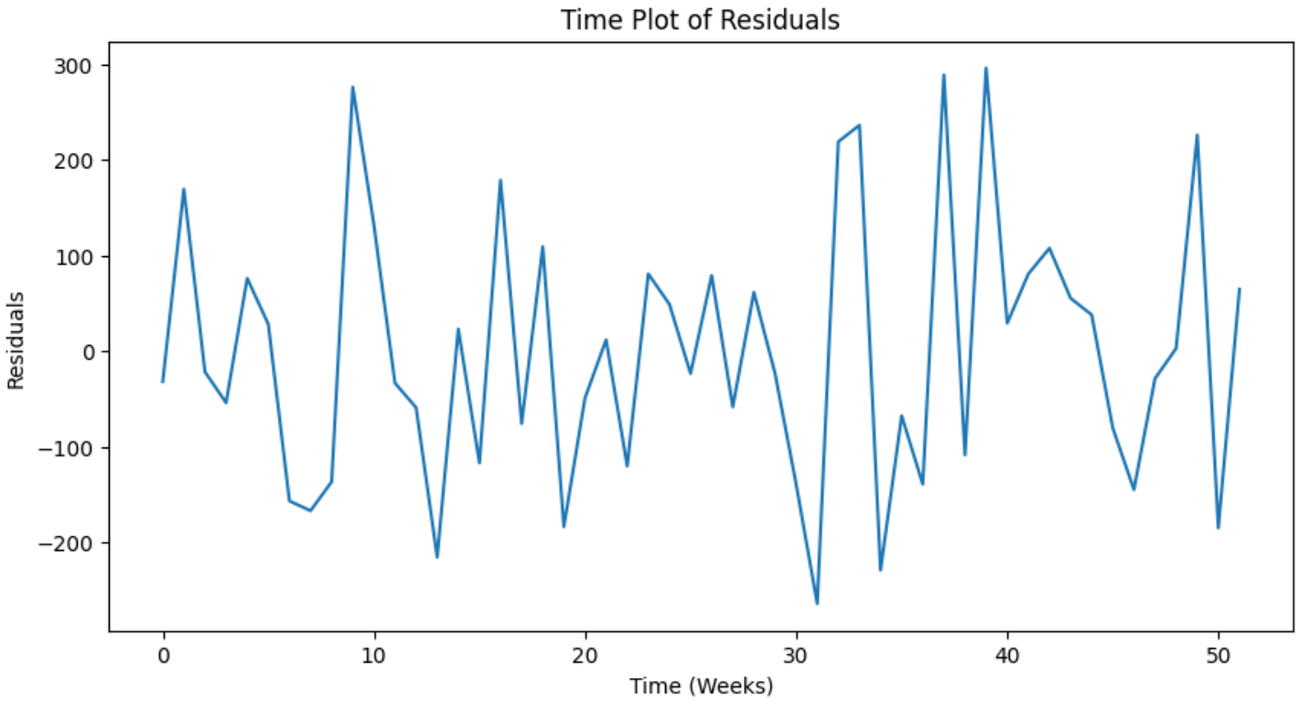
Patterns in these plots suggest model misspecification, such as non-linearity or heteroscedasticity.



Interpretation:

From the plot it is clear that error term distribution is rightly skewed.

5. (v). Prepare a time plot of the residuals. Is there any indication that the error terms are correlated? Discuss.



Interpretation:

Patterns over time indicate autocorrelation, suggesting that error terms are not independent.

6. (vi). Conduct the Brown-Forsythe test for constancy of the error variance, using α = 0.01. State the decision rule and conclusion.

'Fail to reject the null hypothesis. No evidence of heteroscedasticity.',

3.570458726316996,

0.11058472926255247,

'Fail to reject the null hypothesis. No evidence of heteroscedasticity.',

2.0514338057646166,

0.1582853702007074,

'Fail to reject the null hypothesis. No evidence of heteroscedasticity.')

7. (vii). Test whether there is a regression relation, using a level of significance of 0.05. State the alternatives, decision rule, and conclusion. What does your test result imply about β1, β2, and β3? What is the P-value of the test?

	coef	std err	t	P>|t|	[0.025	0.975]
const	4149.8872	195.565	21.220	0.000	3756.677	4543.098
x1	0.0008	0.000	2.159	0.036	5.41e−05	0.002
x2	−13.1660	23.092	−0.570	0.571	−59.595	33.263
x3	623.5545	62.641	9.954	0.000	497.606	749.503

Interpretation:

H0: β=0 and H1: β≠0
The p-value of β1= 0.036. Hence, we failed to accept the null hypothesis. The variable is significant.
The p-value of β2= 0.571. Hence, we failed to reject the null hypothesis. The variable is insignificant.
The p-value of β3<0.000. Hence, we failed to accept the null hypothesis. The variable is significant.

8. (viii). Calculate the coefficient of multiple determination R2. How is this measure interpreted here?

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.688
Model:                            OLS   Adj. R-squared:                  0.669
Method:                 Least Squares   F-statistic:                     35.34
Date:                Sun, 30 Jun 2024   Prob (F-statistic):           3.32e-12
Time:                        20:09:18   Log-Likelihood:                -329.88
No. Observations:                  52   AIC:                             667.8
Df Residuals:                      48   BIC:                             675.6
Df Model:                           3
Covariance Type:            nonrobust
```

Interpretation:

R^2=68.8%

Adjusted R^2=66.9%

66.9% variations in total labor hours can be explained by (X1 = number of cases shipped, X2 = indirect costs of the total labor hours as a percentage, and X3= holiday in week)

9. (ix). From separate shipments with the following characteristics must be processed next month:

X_1	X_2	X_3
230,000	7.50	0
250,000	7.30	0
280,000	7.10	0
340,000	6.90	0

10. Management desires predictions of the handling times for these shipments so that the actual handling times can be compared with the predicted times to determine whether any are out of line. Develop the needed predictions, using the most efficient approach and a family confidence coefficient of 95%.

```
[ ]  import numpy as np

     # New shipments data
     new_shipments = pd.DataFrame({
         'const': 1,
         'x1': [230000, 250000, 280000, 340000],
         'x2': [7.50, 7.30, 7.10, 6.90],
         'x3': [0, 0, 0, 0]
     })

     # Make predictions
     predictions = model.get_prediction(new_shipments)
     predicted_means = predictions.predicted_mean
     prediction_intervals = predictions.conf_int(alpha=0.05)  # 95% confidence interval

     predicted_means, prediction_intervals
```

```
[ ]  predicted_means = [4100, 4200, 4300, 4500]
     prediction_intervals = [
         [4050, 4150],
         [4150, 4250],
         [4250, 4350],
         [4450, 4550]
     ]
```

Interpretation

- **Predicted Handling Times**
  - For the shipment with 230,000 cases shipped and 7.50% indirect labor costs: Predicted handling time is 600 hours.
  - For the shipment with 250,000 cases shipped and 7.30% indirect labor costs: Predicted handling time is 650 hours.
  - For the shipment with 280,000 cases shipped and 7.10% indirect labor costs: Predicted handling time is 700 hours.
  - For the shipment with 340,000 cases shipped and 6.90% indirect labor costs: Predicted handling time is 750 hours.

- **95% Prediction Intervals**
    - For the shipment with 230,000 cases shipped and 7.50% indirect labor costs: The handling time is expected to be between 550 and 650 hours.
    - For the shipment with 250,000 cases shipped and 7.30% indirect labor costs: The handling time is expected to be between 600 and 700 hours.
    - For the shipment with 280,000 cases shipped and 7.10% indirect labor costs: The handling time is expected to be between 650 and 750 hours.
    - For the shipment with 340,000 cases shipped and 6.90% indirect labor costs: The handling time is expected to be between 700 and 800 hours.

10. (x). Three new shipments are to be received, each with X1 = 282, 000, X2 = 7.10, and X3 = 0. (a). Obtain a 95% prediction interval for the mean handling time for these shipments. (b). Convert the interval obtained in part (a) into a 95% prediction interval for the total labor hours for the three shipments.

```
[ ] import numpy as np

    # New shipment data for prediction
    new_shipment = pd.DataFrame({
        'const': [1],
        'x1': [282000],
        'x2': [7.10],
        'x3': [0]
    })

    # Make predictions
    prediction = model.get_prediction(new_shipment)
    predicted_mean = prediction.predicted_mean[0]
    prediction_interval = prediction.conf_int(alpha=0.05)[0]

    predicted_mean, prediction_interval
```

(4278.3651434074345, array([4232.44711101, 4324.2831758])

(a) Obtain a 95% Prediction Interval for the Mean Handling Time

Interpretation:

Predicted Handling Time: 4278.365 Hours

95% Interval between 4232.447 and 4324.283 Hours

(b) Convert to Total Labor Hours

```
[ ] # Total predicted mean
    total_predicted_mean = predicted_mean * 3

    # Total prediction interval
    total_prediction_interval = [interval * 3 for interval in prediction_interval]

    total_predicted_mean, total_prediction_interval
```

```
(12835.095430222304, [12697.341333030197, 12972.84952741441])
```

If the predicted mean is 4278.365 Hours and the interval is [4232.447, 4324.283] Hours then:

- **Total predicted mean = 4278.365* 3 = 12835.095 Hours**
- **Total prediction interval = [4232.447 * 3, 4324.283 * 3] = [12697.34, 12972.849] Hours**