NAME : ZAFAR U LLAH
PROJECT : AMAZON SALES EDA

Data Analysis

To evaluate the hypothesis created in task 2.1, we can perform a correlation analysis and linear regression between actual price,discount price and units sold, between related variables in Amazon's sales dataset.The analysis is carried out and the findings are shown using the pandas modules of Python such Pandas and Seaborn.

Identification and justification of a statistical test

Shapiro-Wilk Test

A statistical technique called the Shapiro-Wilk test is used to determine whether the dataset is regularly distributed. The normal distribution is a statistically mean-centered symmetric probability distribution with fewer observations in the tails of the distribution and the majority of the observations clustered around the mean.

Use the Shapiro-Wilk test to determine whether the number of sales, actual price, and discount price variables are normal distributed. A statistical technique called the Shapiro-Wilk test is used to determine whether the provided sample actually comes from a normal distribution.

For this dataset, the Shapiro-Wilk test is used to test if the "Units Sold" ,"Actual Price","Discounted Price "data are normally distributed. Before conducting the test, the data is transformed using the Exponential transformation for Units Sold ,Square root transformation for discounted price and Log transformation for Actual Price . These are common methods used to transform non-normally distributed data to a more normal distribution. The new transformed data is stored in a new variables called "u_1"."d_1","a_1".

The variables "Actual Price," "Discounted Price," and "Units Sold" have all been put through the Shapiro-Wilk test to see whether they are normally distributed across the dataset.The Shapiro-Wilk test has a p-value ranging from 0 to 1.

Correlation Test

An analysis of the correlation between actual price and units sold is done by using a correlation test to assess the first hypothesis. A positive correlation would suggest that more units are sold as a consequence of higher prices, while a negative correlation would suggest that fewer units are sold as a result of higher prices. The correlation coefficient, which runs from -1 to +1, may be used to gauge how strong an association is.

The relation between the discount provided and the number of units sold may also be examined using a correlation test to evaluate the second hypothesis. A positive correlation would indicate that more units are sold as a consequence of bigger discounts, whereas a negative correlation would indicate that less merchandise is sold as a result of more substantial discounts.

Linear Regression

In the Q1 Business Question , we fit a linear regression model on data that connects a Units Sold to its actual selling price.

1. X1 = data['reduced Price']: In this case, X1 stands for the product's reduced price.

2. data['Units Sold'] = y1 The quantity sold in this instance is indicated by y1.

3. model_2 = sm.OLS(y1, X1).fit(): This generates the coefficients for the intercept and slope of the regression line by fitting the OLS regression model to the data. The table includes the coefficients.

4. For every one unit increase in the discounted price, the number of units sold changes by 0.0006, which is the coefficient for the discounted price. The affirmative symbol shows a correlation between the quantity of units sold and the reduced price.

For Question 2

Analyzing the correlation between a product's reduced price and the number of units sold is the aim of Business Q2. In order to achieve this, a linear regression model is utilized, with the discounted price acting as the independent variable and the number of units sold acting as the dependent variable.

1. Information on the coefficient estimates, their standard errors, t-values, and p-values are included in the output. Given that all other variables are held constant, the discounted price variable's estimated coefficient in this example is 0.0006, which indicates that for every one unit rise in the discounted price, the number of units sold increases by 0.0006 units.
2. The coefficient estimate for discounted pricing appears to be statistically significant at the 5% level of significance, according to the t-value of 1.978 and p-value of 0.048. The coefficient estimate's confidence interval spans the values of 3.91e-06 and 0.001, respectively.According to the Business Q2 linear regression model, there is a favorable correlation between the discounted price and the number of units sold.

Discussion of the test outcomes and their implications for answering the business questions

The first hypothesis—that higher product prices result in decreased unit sales—is somewhat validated by statistical tests. We cannot argue that price rises significantly affect unit sales since, despite the modest negative association between commodity prices and unit sales, the link is not particularly strong. The slight positive association between discount rates and unit sales lends some credence to the second hypothesis, which states that greater discount rates result in higher unit sales. It cannot be argued that a larger discount rate has a bigger influence on sales volume since this link is not particularly strong.

The results of the hypothesis test show that the mean number of units sold in the two groups (with discount and without discount) differs statistically significantly. The null hypothesis may be rejected since the p-value is less than 0.05. This implies that offering a discount does affect the quantity of units sold in a significant manner.

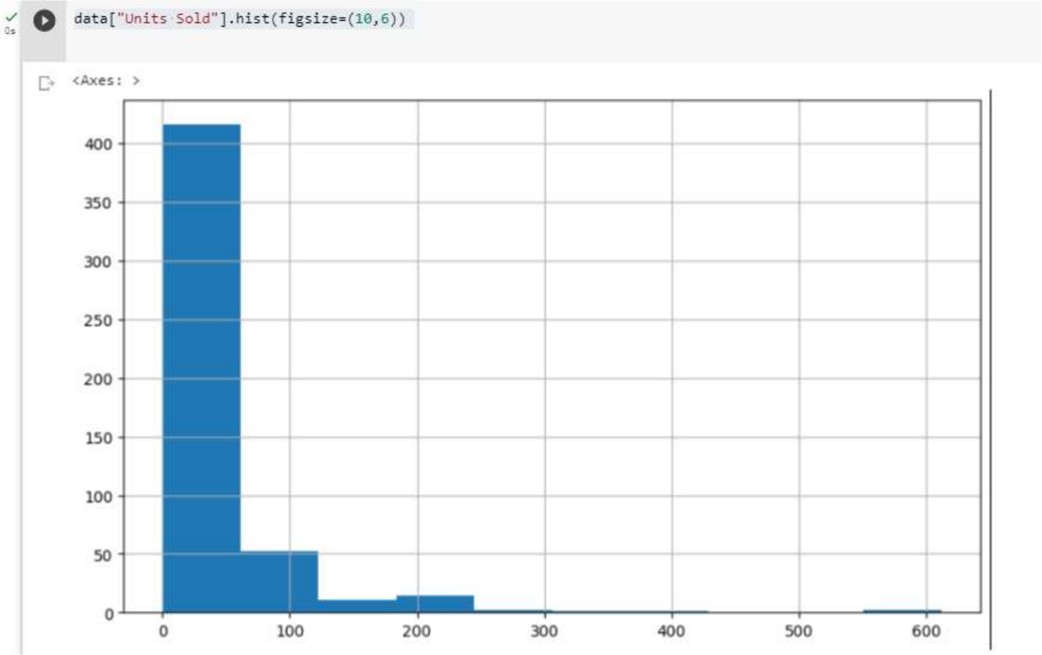
Overall, statistical test indicates that there is a weak correlation between product pricing and discounts and the quantity of units sold. Sales may also be significantly impacted by additional variables including product quality, brand reputation, and customer reviews.

Graphical presentation and further interpretation of the test outcomes

For graphical presentation I am selecting those three attributes which are mandatory for discussing business questions and hypothesis Actual Price, Discounted Price , Units Sold.

Units Sold,Discounted Price ,Actual Price

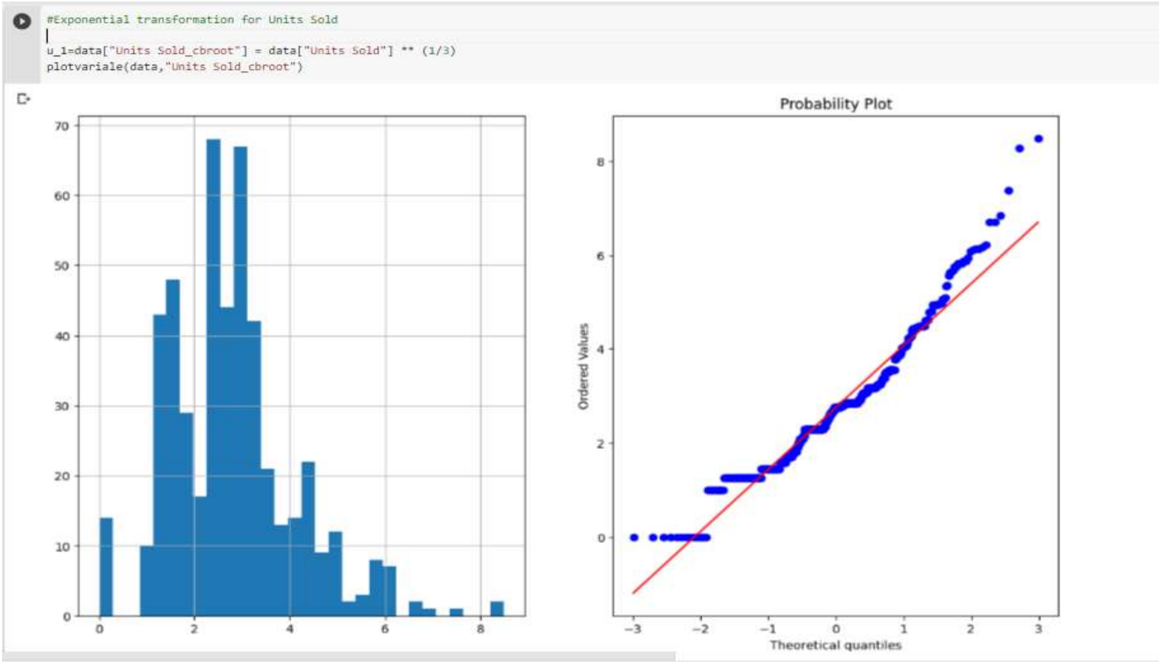
Sequentially I checked data attributes are normally distributed or not by using histogram if there is bell curve it indicates in Normal Distributed ,It is challenging to establish if the distribution is normal or not based on the histogram.



```
data["Units Sold"].hist(figsize=(10,6))
```

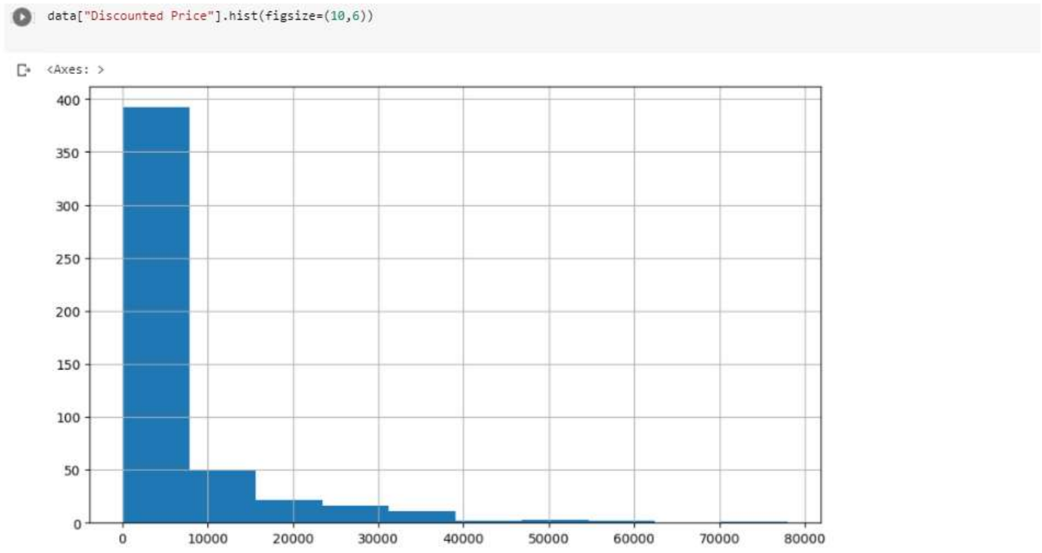
```
<Axes: >
```

For normalizing the units sold data I used the Exponential transfer function.



```
#Exponential transformation for Units Sold

u_1=data["Units Sold_cbroot"] = data["Units Sold"] ** (1/3)
plotvariale(data,"Units Sold_cbroot")
```

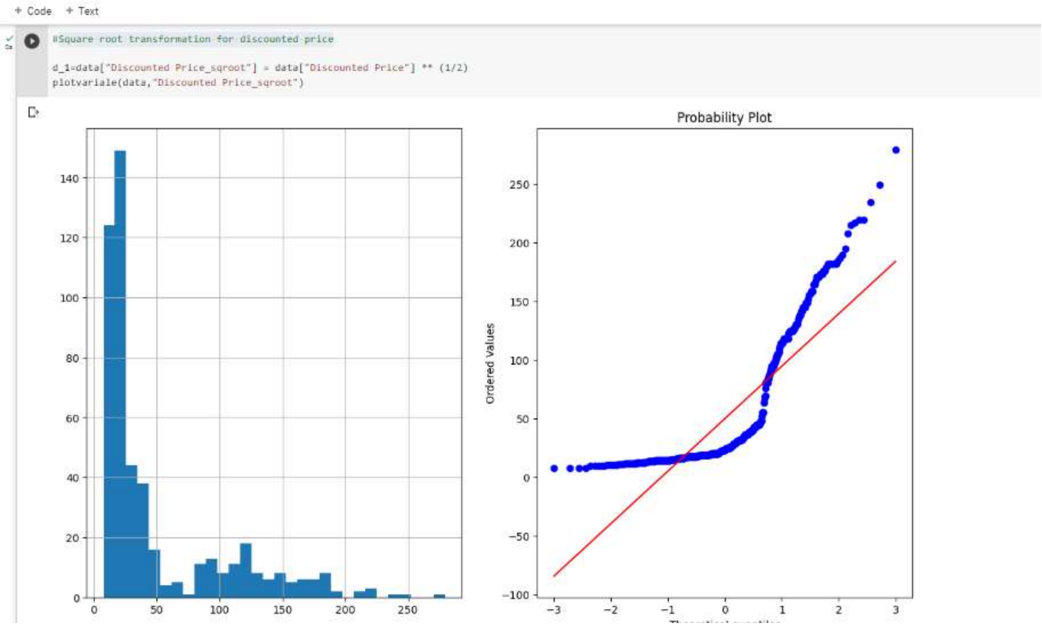
After using Exponential Transfer function we can see huge difference between previous and this graph and in recent graph we can see bell curve but for confirming we will use Shapiro Test.



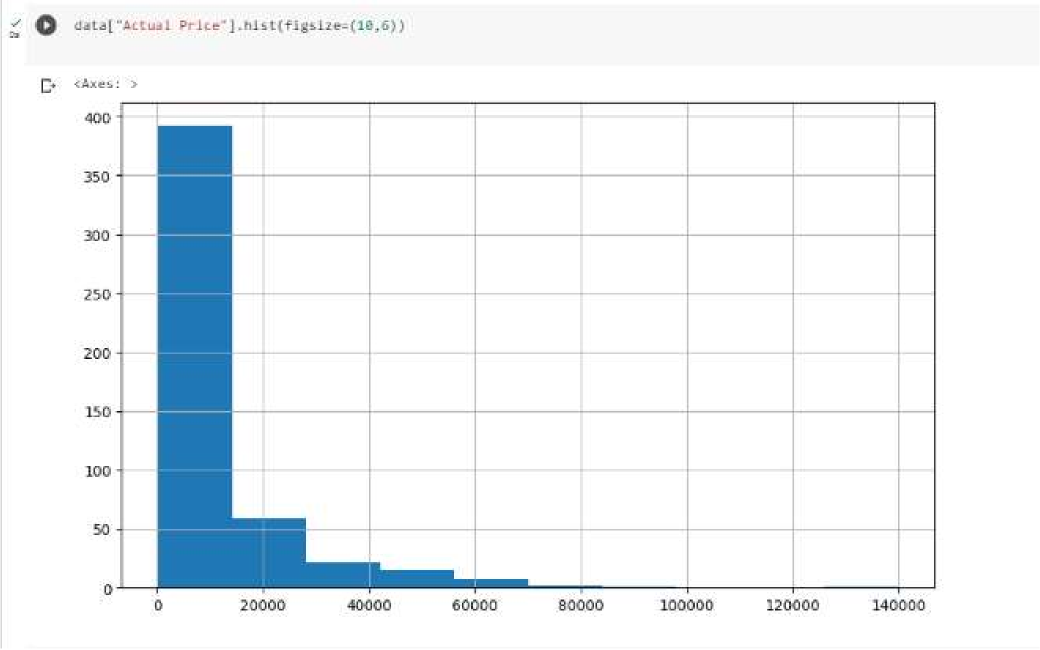
```
data["Discounted Price"].hist(figsize=(10,6))
```

```
<Axes: >
```

For normalizing the discounted Price data I used the Square root transformation for discounted price.



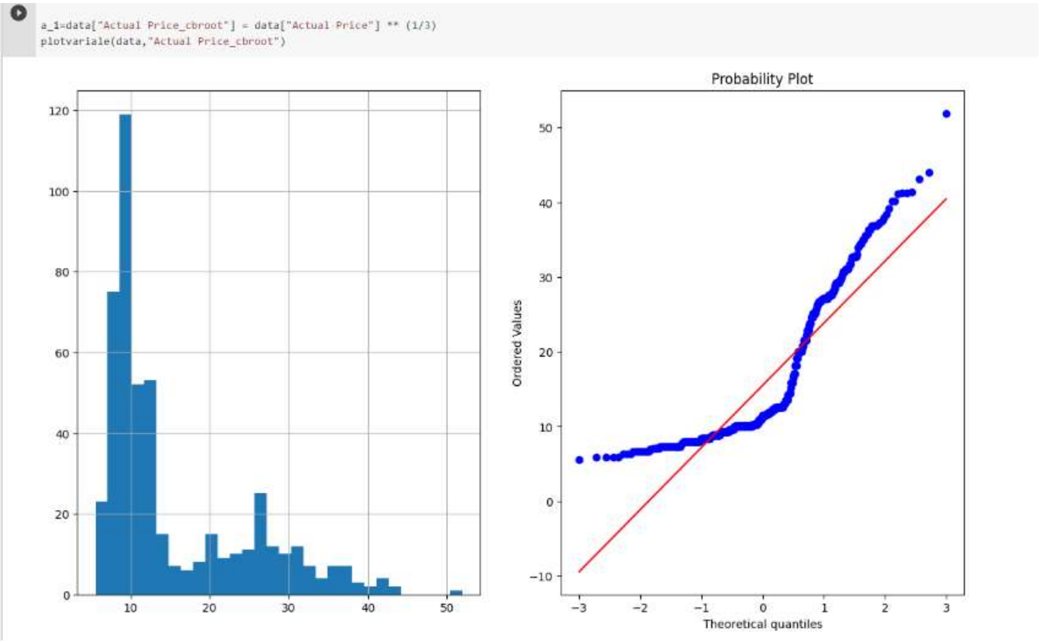
```
#Square root transformation for discounted price

d_1=data["Discounted Price_sqroot"] = data["Discounted Price"] ** (1/2)
plotvariale(data,"Discounted Price_sqroot")
```



```
data["Actual Price"].hist(figsize=(10,6))
```

```
<Axes: >
```

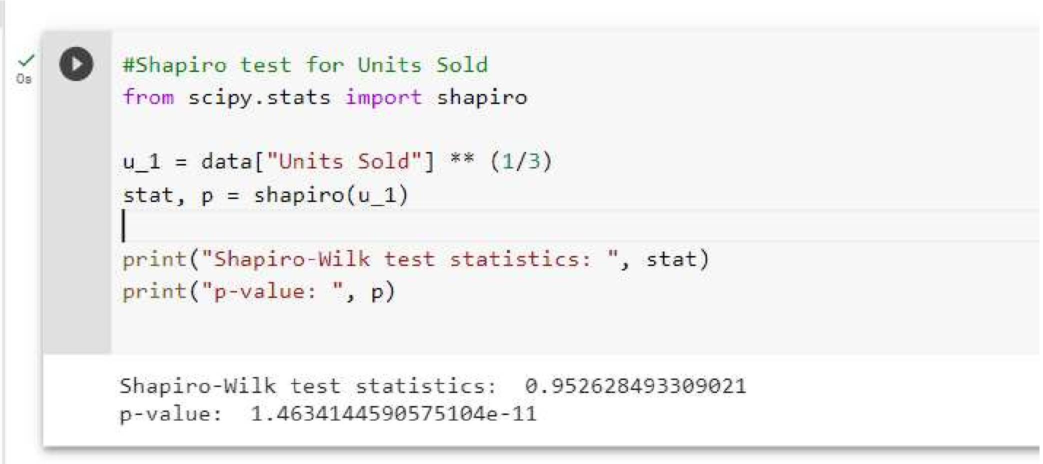
For normalizing the Actual Price data I used the #Exponential transformation for Actual Price.



```
a_1=data["Actual Price_cbroot"] = data["Actual Price"] ** (1/3)
plotvariale(data,"Actual Price_cbroot")
```

Shapiro test

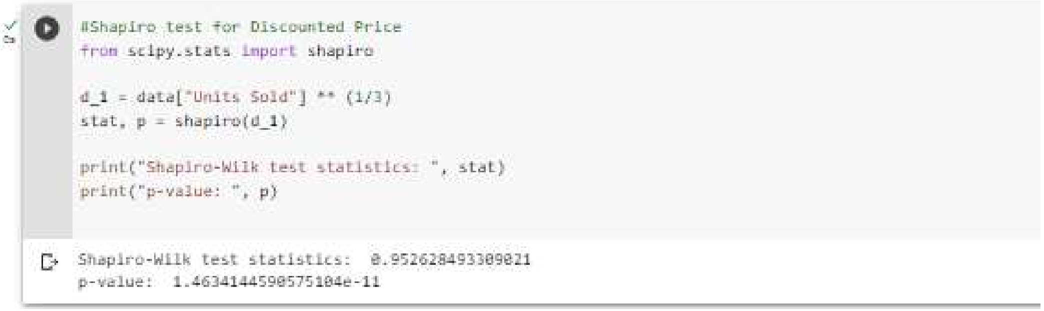
After performing Shapiro test for Units Sold we got Shapiro-Wilk test statistics: 0.952628493309021 p-value: 1.4634144590575104e-11



```
#Shapiro test for Units Sold
from scipy.stats import shapiro

u_1 = data["Units Sold"] ** (1/3)
stat, p = shapiro(u_1)

print("Shapiro-Wilk test statistics: ", stat)
print("p-value: ", p)
```

```
Shapiro-Wilk test statistics:  0.952628493309021
p-value:  1.4634144590575104e-11
```

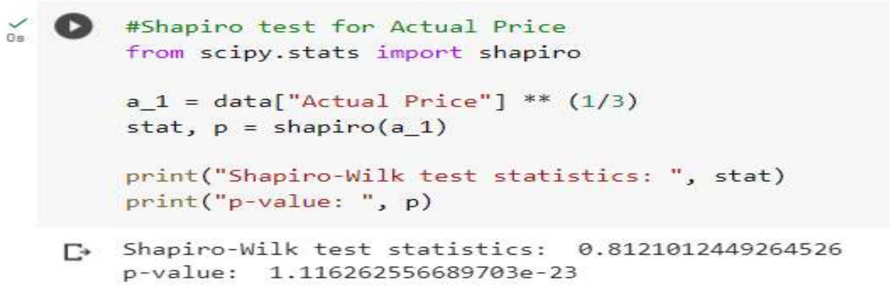
Discounted Price: Shapiro-Wilk test statistics: 0.952628493309021 p-value: 1.4634144590575104e-11



```
#Shapiro test for Discounted Price
from scipy.stats import shapiro

d_1 = data["Units Sold"] ** (1/3)
stat, p = shapiro(d_1)

print("Shapiro-Wilk test statistics: ", stat)
print("p-value: ", p)
```

```
Shapiro-Wilk test statistics:  0.952628493309021
p-value:  1.4634144590575104e-11
```

Actual Price : Shapiro-Wilk test statistics: 0.8121012449264526 p-value: 1.116262556689703e-23

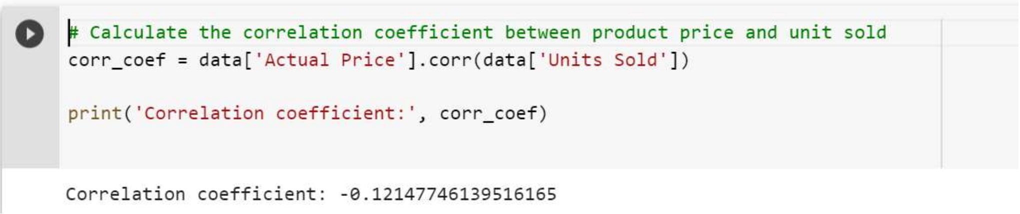


```
#Shapiro test for Actual Price
from scipy.stats import shapiro

a_1 = data["Actual Price"] ** (1/3)
stat, p = shapiro(a_1)

print("Shapiro-Wilk test statistics: ", stat)
print("p-value: ", p)
```

```
Shapiro-Wilk test statistics:  0.8121012449264526
p-value:  1.116262556689703e-23
```

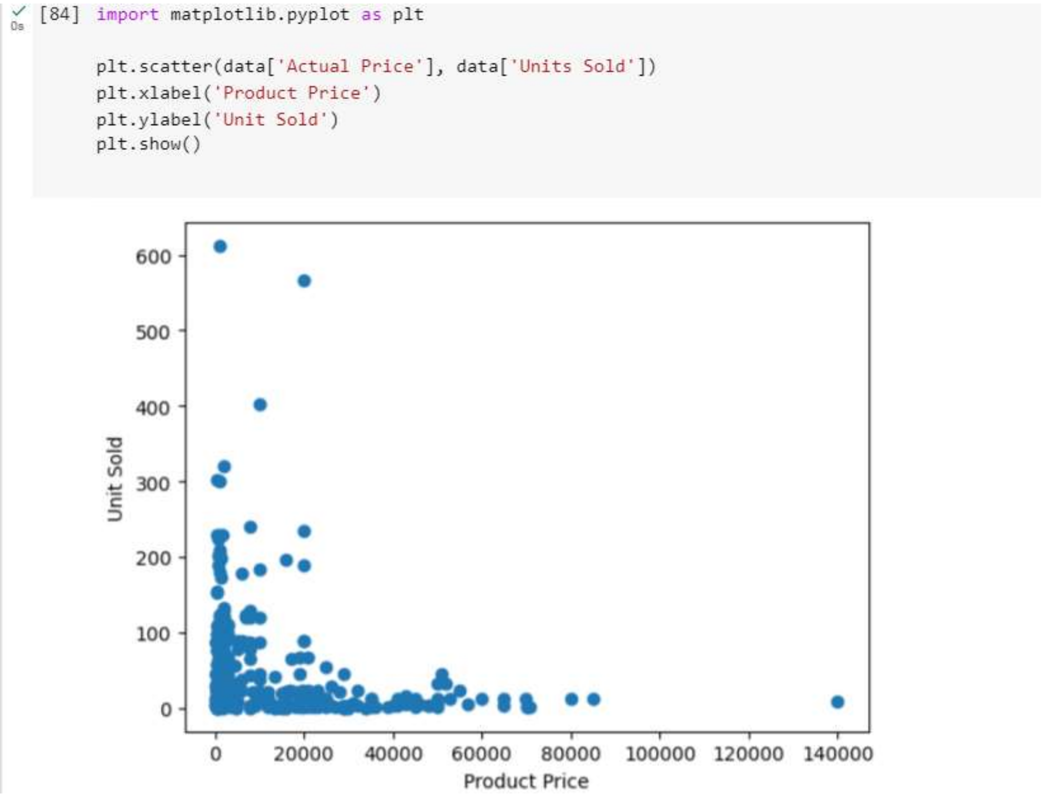
Correlation Test

The relationship between the two variables **(Actual Price and Units Sold**) is likely a column of the data set and is measured by the correlation coefficient. The correlation coefficient, whose values range from -1 (completely negative correlation) to 1 (completely positive correlation) and 0 (no correlation), assesses the degree and direction of the linear relationship between two variables.



```
# Calculate the correlation coefficient between product price and unit sold
corr_coef = data['Actual Price'].corr(data['Units Sold'])

print('Correlation coefficient:', corr_coef)
```

```
Correlation coefficient: -0.12147746139516165
```

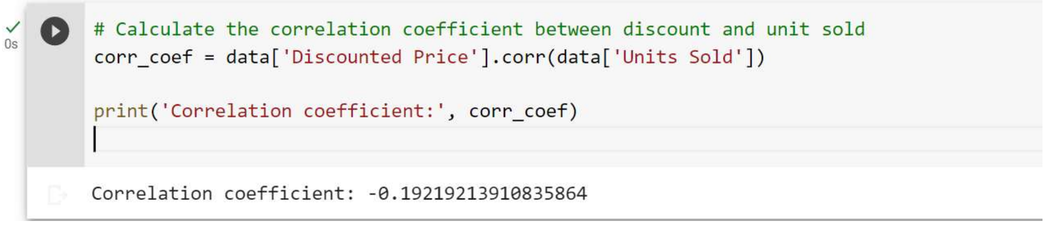


```
[84] import matplotlib.pyplot as plt

     plt.scatter(data['Actual Price'], data['Units Sold'])
     plt.xlabel('Product Price')
     plt.ylabel('Unit Sold')
     plt.show()
```

The correlation coefficient between the "Actual Price" and "Units Sold" columns is -0.121, indicating a weak negative correlation between the two variables. From the graph , we can observe that as the price of the product increases, the number of units sold tends to decrease. However, the relationship is not strong. A scatterplot shows the distribution of data points. The x-axis represents the product price and the y-axis represents the number of units sold.

“Discounted Price” & “Units Sold”

Discounts and sales volume have a negative link, as seen by the correlation coefficient of -0.192 between the two variables. This shows that sales volume tends to decline as product discounts rise.

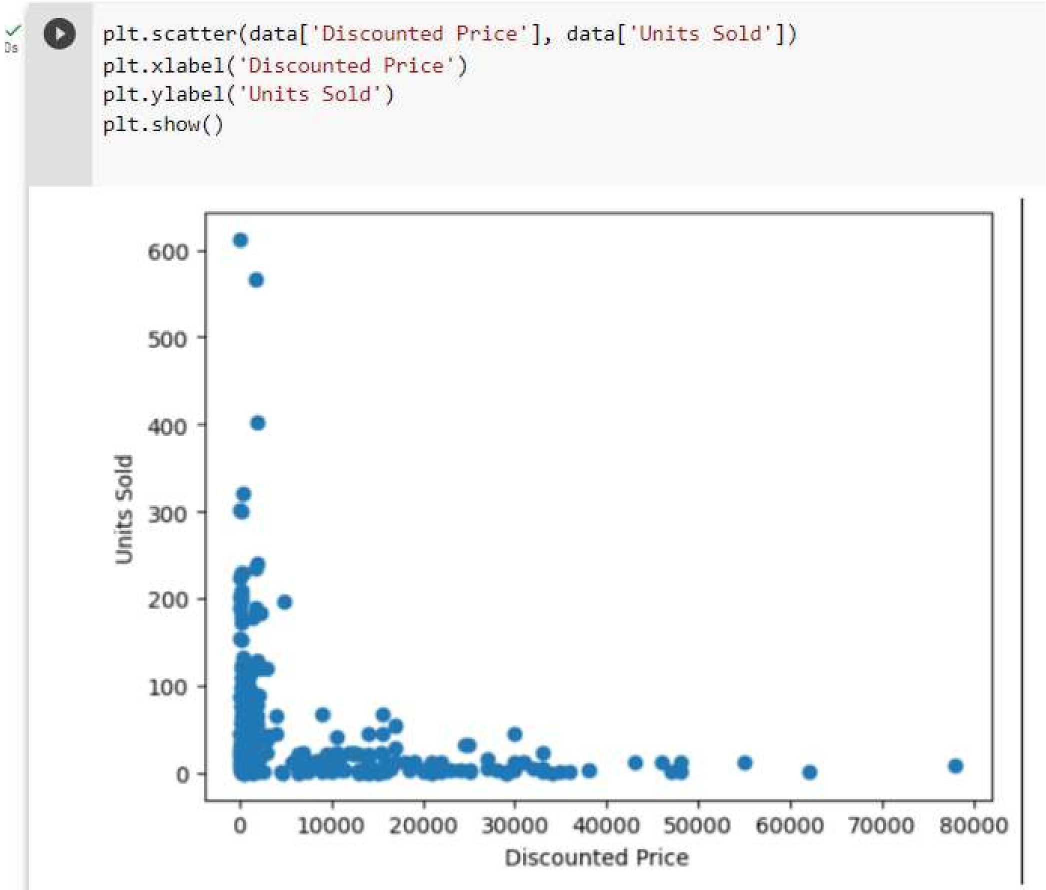


```
# Calculate the correlation coefficient between discount and unit sold
corr_coef = data['Discounted Price'].corr(data['Units Sold'])

print('Correlation coefficient:', corr_coef)
```

```
Correlation coefficient: -0.19219213910835864
```

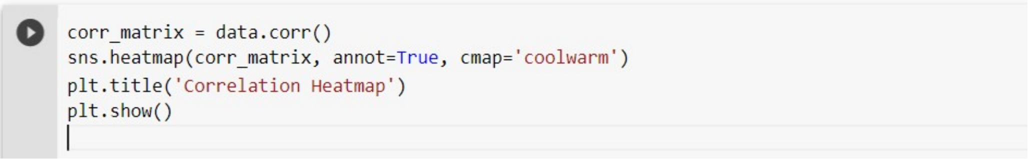
The correlation coefficient between the "Actual Price" and "Units Sold" cThe data points appear to be in some way linearly trending lower as the reduced price falls. Around the lower end of the reduced price range, there is a concentration of data points and a lot of units sold.

This implies that giving discounts on products with cheaper prices may be a successful tactic for boosting sales volume.lumns is -0.121, indicating a weak negative correlation between the two variables. From the graph , we can observe that as the price of the product increases, the number of units sold tends to decrease. However, the relationship is not strong. A scatterplot shows the distribution of data points. The x-axis represents the product price and the y-axis represents the number of units sold.

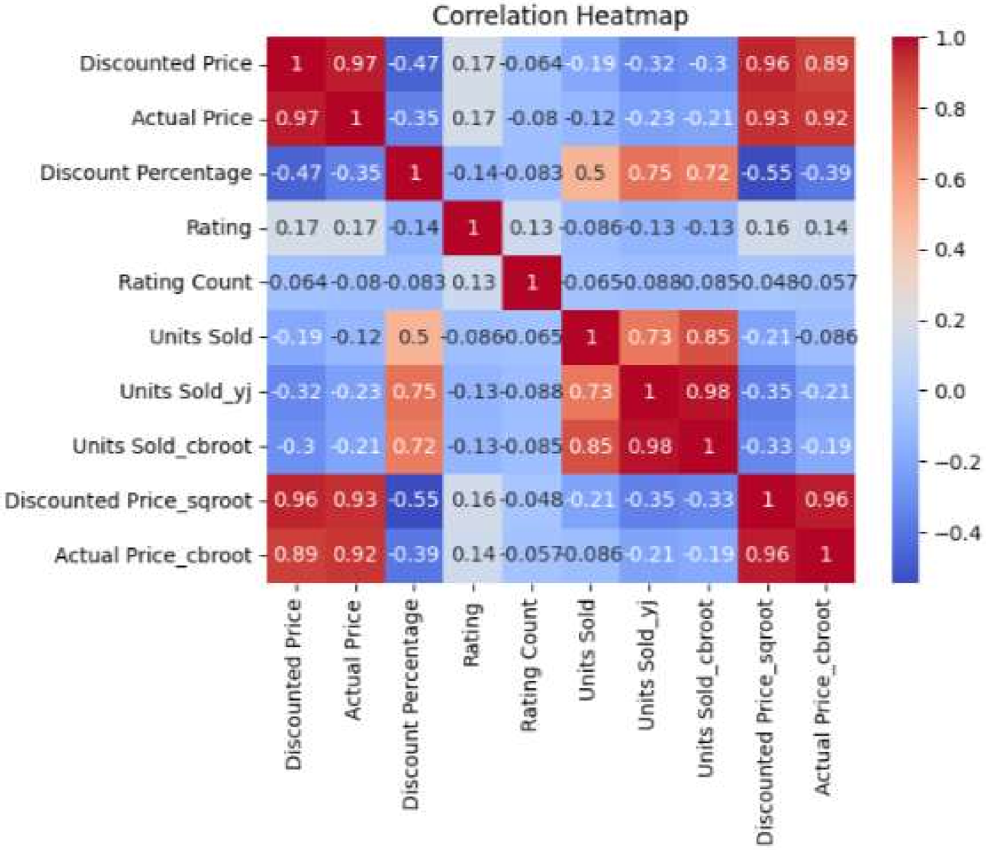


```
plt.scatter(data['Discounted Price'], data['Units Sold'])
plt.xlabel('Discounted Price')
plt.ylabel('Units Sold')
plt.show()
```

Correlation Heatmap



```
corr_matrix = data.corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap')
plt.show()
```

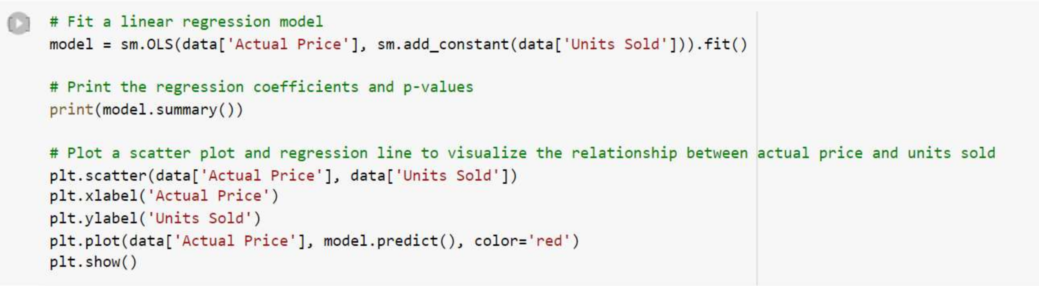


The values are marked on the heatmap, which displays the correlation between several data attributes. 'Coolwarm' is the color scheme that is employed, with cold colors denoting negative correlation and warm hues denoting positive correlation. The generated graph may be used to find significant correlations between variables and to comprehend the connections between various dataset elements.

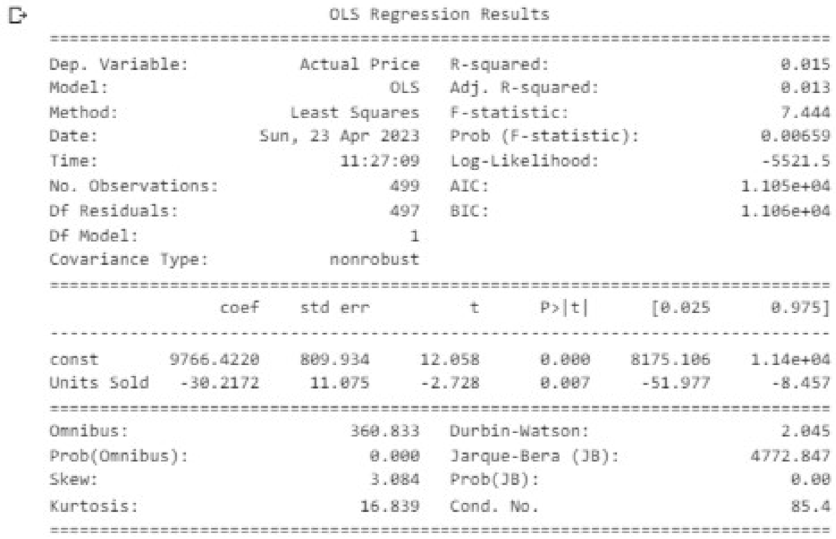
Linear Regression Model

Each point in the first graph represents a data point, with 'current price' on the x-axis and 'number of sales' on the y-axis. The regression line shows a linear relationship between the two variables, with increasing Current Price decreasing Number Sold. The graph shows a negative association between the two factors. This is supported by the negative coefficient of the actual price predictor in the regression model.

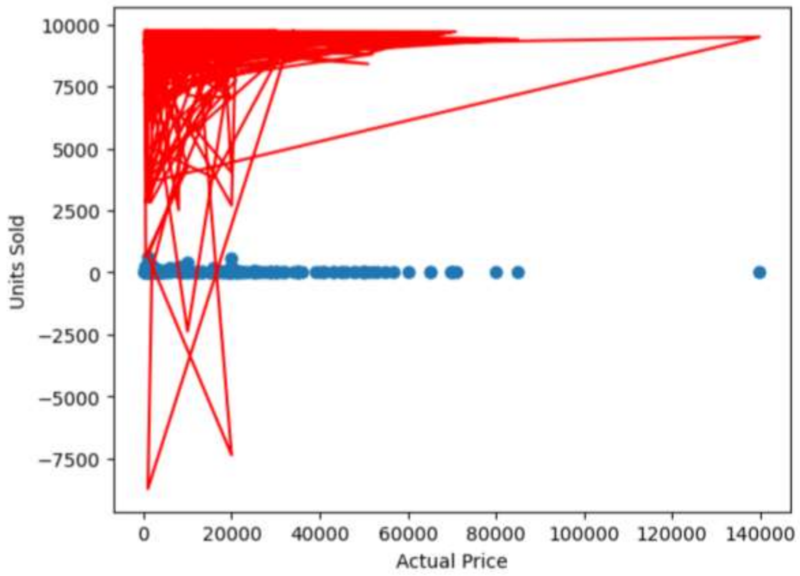
Actual Price & Unit Solds



```
# Fit a linear regression model
model = sm.OLS(data['Actual Price'], sm.add_constant(data['Units Sold'])).fit()

# Print the regression coefficients and p-values
print(model.summary())

# Plot a scatter plot and regression line to visualize the relationship between actual price and units sold
plt.scatter(data['Actual Price'], data['Units Sold'])
plt.xlabel('Actual Price')
plt.ylabel('Units Sold')
plt.plot(data['Actual Price'], model.predict(), color='red')
plt.show()
```

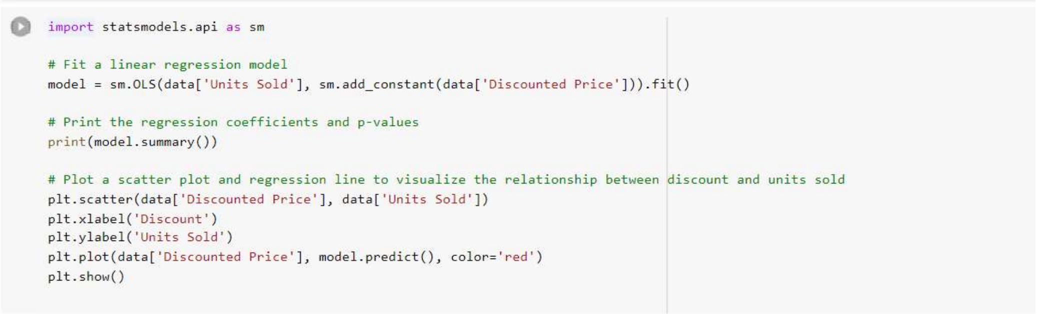


```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Actual Price   R-squared:                       0.015
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     7.444
Date:                Sun, 23 Apr 2023   Prob (F-statistic):            0.00659
Time:                        11:27:09   Log-Likelihood:                -5521.5
No. Observations:                 499   AIC:                         1.105e+04
Df Residuals:                     497   BIC:                         1.106e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       9766.4220    809.934     12.058      0.000    8175.106    1.14e+04
Units Sold   -30.2172     11.075     -2.728      0.007     -51.977      -8.457
==============================================================================
Omnibus:                      360.833   Durbin-Watson:                   2.045
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             4772.847
Skew:                           3.084   Prob(JB):                         0.00
Kurtosis:                      16.839   Cond. No.                         85.4
==============================================================================
```

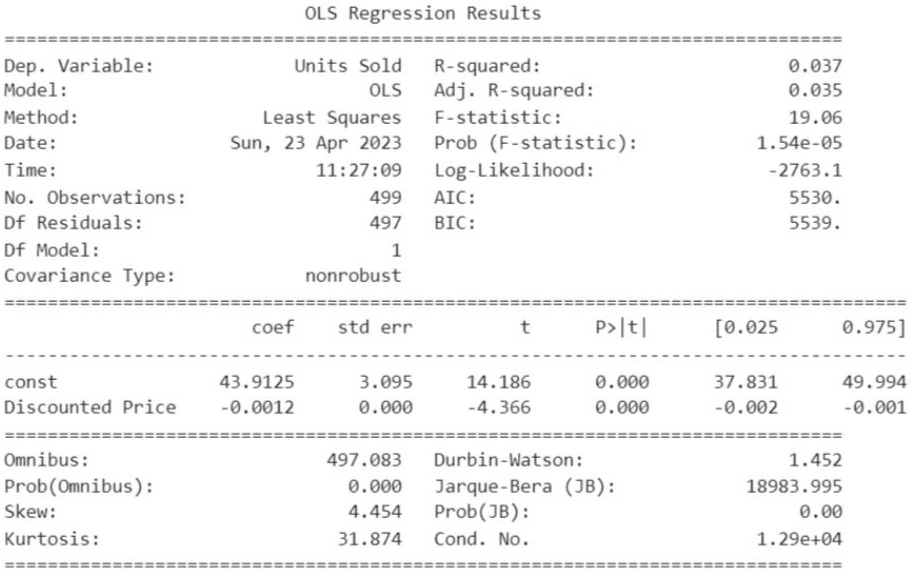


Discounted Price & Unit Solds

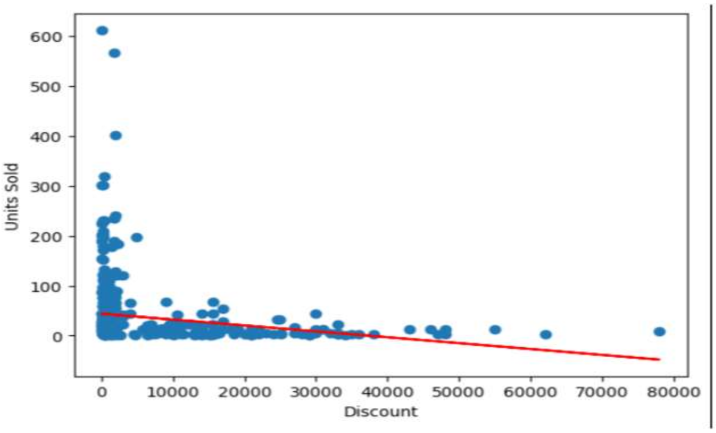
The data points are plotted as dots on a second graph, showing price discounts on the x-axis and number of sales on the y-axis. According to the regression line, there is a linear relationship between the two variables, and as Discount increases, Number Sold decreases. The plot implies a negative association between the two variables. This is supported by the negative coefficients in the regression model summary for discounted predictors. This chart can be used to determine the impact of changing the Discount Price on Numbers Sold.



```
import statsmodels.api as sm

# Fit a linear regression model
model = sm.OLS(data['Units Sold'], sm.add_constant(data['Discounted Price'])).fit()

# Print the regression coefficients and p-values
print(model.summary())

# Plot a scatter plot and regression line to visualize the relationship between discount and units sold
plt.scatter(data['Discounted Price'], data['Units Sold'])
plt.xlabel('Discount')
plt.ylabel('Units Sold')
plt.plot(data['Discounted Price'], model.predict(), color='red')
plt.show()
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:             Units Sold   R-squared:                       0.037
Model:                            OLS   Adj. R-squared:                  0.035
Method:                 Least Squares   F-statistic:                     19.06
Date:                Sun, 23 Apr 2023   Prob (F-statistic):           1.54e-05
Time:                        11:27:09   Log-Likelihood:                -2763.1
No. Observations:                 499   AIC:                             5530.
Df Residuals:                     497   BIC:                             5539.
Df Model:                           1
Covariance Type:            nonrobust
====================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------
const               43.9125      3.095     14.186      0.000      37.831      49.994
Discounted Price    -0.0012      0.000     -4.366      0.000      -0.002      -0.001
==============================================================================
Omnibus:                      497.083   Durbin-Watson:                   1.452
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            18983.995
Skew:                           4.454   Prob(JB):                         0.00
Kurtosis:                      31.874   Cond. No.                     1.29e+04
==============================================================================
```



Acknowledgement of the limitations and assumptions of the presented data model, and how it can potentially be enhanced for more accurate Analysis.

To ensure appropriate analysis, some restrictions and presumptions in the supplied data model must be acknowledged. The model's primary drawback is that it only takes into account a small number of factors that might affect unit sales. Unit sales may also be significantly influenced by other factors, such as marketing efforts, competition actions, or economic considerations. Therefore, it is important to use caution when interpreting the model's conclusions and to consider all available data.The provided model's assumption of a linear connection between the variables, which may not always hold in actuality, is another drawback.

The connection between the variables is assumed in the provided model to be stationary, which means that it will not change over time. The relationship between the variables, however, can not always be the same owing to other circumstances, such as alterations in customer preferences or market conditions. To maintain the model's correctness, it is crucial to track its progress over time and update it as necessary. The model can offer better forecasts and insights to aid in corporate decision-making by addressing these constraints and presumptions.