**import** numpy **as** np

**import** pandas **as** pd

**import** matplotlib.pyplot **as** plt

**import** statsmodels.api **as** sm

**from** numpy.polynomial.polynomial **import** polyfit

**from** sklearn.linear\_model **import** LinearRegression

## ****1 - Business Problem****

**\_Deliverytime -> Predict delivery time using sorting time**

## ****2 - Data collection and description****

In [2]:

df **=** pd**.**read\_csv("C:/Users/Sohail/OneDrive/Data Science/Assignments/Simple Linear Regression/Raw Data/delivery\_time.csv")

### ****Scatter Plot****

In [3]:

x **=** df['Sorting Time']

y **=** df['Delivery Time']

In [4]:

b, m **=** polyfit(x, y, 1)

plt**.**scatter(x, y)

plt**.**plot(x, y, '.')

plt**.**plot(x, b **+** m **\*** x, '-')

plt**.**title('Scatter plot Delivery Time')

plt**.**xlabel('Sorting Time')

plt**.**ylabel('Delivery Time')

plt**.**show()

As displayed in the scatter plot, the data does contains some outliers, but there is potive correlation between delivery time and sorting Time

### ****Correlation Analysis****

In [5]:

corr **=** np**.**corrcoef(x, y)

Corr  
array([[1. , 0.82599726], [0.82599726, 1. ]])

The correlation between delivery time and sorting Time is high (83%)

## ****3 - Regression Model****

### ****1 - No transformation****

In [6]:

model **=** sm**.**OLS(y, x)**.**fit()

predictions **=** model**.**predict(x)

In [7]:

model**.**summary()

Out[7]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.955 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.953 |
| **Method:** | Least Squares | **F-statistic:** | 424.5 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 6.12e-15 |
| **Time:** | 20:16:10 | **Log-Likelihood:** | -57.349 |
| **No. Observations:** | 21 | **AIC:** | 116.7 |
| **Df Residuals:** | 20 | **BIC:** | 117.7 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 2.5652 | 0.125 | 20.603 | 0.000 | 2.306 | 2.825 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1.504 | **Durbin-Watson:** | 1.305 |
| **Prob(Omnibus):** | 0.471 | **Jarque-Bera (JB):** | 0.508 |
| **Skew:** | -0.348 | **Prob(JB):** | 0.776 |
| **Kurtosis:** | 3.310 | **Cond. No.** | 1.00 |

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

### ****2 - Log Transformation of X****

In [8]:

x\_log **=** np**.**log(df['Sorting Time'])

In [9]:

model **=** sm**.**OLS(y, x\_log)**.**fit()

predictions **=** model**.**predict(x\_log)

In [10]:

model**.**summary()

Out[10]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.975 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.974 |
| **Method:** | Least Squares | **F-statistic:** | 791.0 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 1.48e-17 |
| **Time:** | 20:17:59 | **Log-Likelihood:** | -51.035 |
| **No. Observations:** | 21 | **AIC:** | 104.1 |
| **Df Residuals:** | 20 | **BIC:** | 105.1 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 9.6706 | 0.344 | 28.124 | 0.000 | 8.953 | 10.388 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 3.656 | **Durbin-Watson:** | 1.453 |
| **Prob(Omnibus):** | 0.161 | **Jarque-Bera (JB):** | 2.164 |
| **Skew:** | 0.772 | **Prob(JB):** | 0.339 |
| **Kurtosis:** | 3.298 | **Cond. No.** | 1.00 |

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

### ****3 - Log Transformation of Y****

In [11]:

y\_log **=** np**.**log(df['Delivery Time'])

In [12]:

model **=** sm**.**OLS(y\_log, x)**.**fit()

predictions **=** model**.**predict(x)

In [13]:

model**.**summary()

Out[13]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.917 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.912 |
| **Method:** | Least Squares | **F-statistic:** | 219.7 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 3.00e-12 |
| **Time:** | 20:18:03 | **Log-Likelihood:** | -25.284 |
| **No. Observations:** | 21 | **AIC:** | 52.57 |
| **Df Residuals:** | 20 | **BIC:** | 53.61 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 0.4008 | 0.027 | 14.821 | 0.000 | 0.344 | 0.457 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 2.572 | **Durbin-Watson:** | 1.446 |
| **Prob(Omnibus):** | 0.276 | **Jarque-Bera (JB):** | 1.346 |
| **Skew:** | -0.275 | **Prob(JB):** | 0.510 |
| **Kurtosis:** | 1.889 | **Cond. No.** | 1.00 |

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

### ****4 - Log Transformation of X & Y****

In [14]:

model **=** sm**.**OLS(y\_log, x\_log)**.**fit()

predictions **=** model**.**predict(x\_log)

In [15]:

model**.**summary()

Out[15]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.972 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.970 |
| **Method:** | Least Squares | **F-statistic:** | 688.7 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 5.72e-17 |
| **Time:** | 20:18:09 | **Log-Likelihood:** | -13.899 |
| **No. Observations:** | 21 | **AIC:** | 29.80 |
| **Df Residuals:** | 20 | **BIC:** | 30.84 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 1.5396 | 0.059 | 26.244 | 0.000 | 1.417 | 1.662 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1.636 | **Durbin-Watson:** | 1.727 |
| **Prob(Omnibus):** | 0.441 | **Jarque-Bera (JB):** | 1.137 |
| **Skew:** | 0.304 | **Prob(JB):** | 0.566 |
| **Kurtosis:** | 2.035 | **Cond. No.** | 1.00 |

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

### ****5 - Sq Root Transformation of X****

In [16]:

x\_sqrt **=** np**.**sqrt(df['Sorting Time'])

In [17]:

model **=** sm**.**OLS(y, x\_sqrt)**.**fit()

predictions **=** model**.**predict(x\_sqrt)

In [18]:

model**.**summary()

Out[18]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.975 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.973 |
| **Method:** | Least Squares | **F-statistic:** | 772.0 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 1.88e-17 |
| **Time:** | 20:18:14 | **Log-Likelihood:** | -51.284 |
| **No. Observations:** | 21 | **AIC:** | 104.6 |
| **Df Residuals:** | 20 | **BIC:** | 105.6 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 6.9466 | 0.250 | 27.785 | 0.000 | 6.425 | 7.468 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 6.818 | **Durbin-Watson:** | 1.334 |
| **Prob(Omnibus):** | 0.033 | **Jarque-Bera (JB):** | 4.599 |
| **Skew:** | 1.090 | **Prob(JB):** | 0.100 |
| **Kurtosis:** | 3.708 | **Cond. No.** | 1.00 |

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

### ****6 - Square Root Transformation of Y****

In [19]:

y\_sqrt **=** np**.**sqrt(df['Delivery Time'])

In [20]:

model **=** sm**.**OLS(y\_sqrt, x)**.**fit()

predictions **=** model**.**predict(x)

In [21]:

model**.**summary()

Out[21]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.930 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.927 |
| **Method:** | Least Squares | **F-statistic:** | 266.0 |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 5.09e-13 |
| **Time:** | 20:18:19 | **Log-Likelihood:** | -31.484 |
| **No. Observations:** | 21 | **AIC:** | 64.97 |
| **Df Residuals:** | 20 | **BIC:** | 66.01 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 0.5926 | 0.036 | 16.309 | 0.000 | 0.517 | 0.668 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 1.452 | **Durbin-Watson:** | 1.434 |
| **Prob(Omnibus):** | 0.484 | **Jarque-Bera (JB):** | 1.105 |
| **Skew:** | -0.328 | **Prob(JB):** | 0.575 |
| **Kurtosis:** | 2.087 | **Cond. No.** | 1.00 |

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

### ****7 - Square Root Transformation of X & Y****

In [22]:

model **=** sm**.**OLS(y\_sqrt, x\_sqrt)**.**fit()

predictions **=** model**.**predict(x\_sqrt)

In [23]:

model**.**summary()

Out[23]:

OLS Regression Results

|  |  |  |  |
| --- | --- | --- | --- |
| **Dep. Variable:** | Delivery Time | **R-squared (uncentered):** | 0.987 |
| **Model:** | OLS | **Adj. R-squared (uncentered):** | 0.987 |
| **Method:** | Least Squares | **F-statistic:** | 1542. |
| **Date:** | Sun, 08 Mar 2020 | **Prob (F-statistic):** | 2.10e-20 |
| **Time:** | 20:18:22 | **Log-Likelihood:** | -13.658 |
| **No. Observations:** | 21 | **AIC:** | 29.32 |
| **Df Residuals:** | 20 | **BIC:** | 30.36 |
| **Df Model:** | 1 |  |  |
| **Covariance Type:** | nonrobust |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **coef** | **std err** | **t** | **P>|t|** | **[0.025** | **0.975]** |
| **Sorting Time** | 1.6364 | 0.042 | 39.267 | 0.000 | 1.549 | 1.723 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Omnibus:** | 0.176 | **Durbin-Watson:** | 1.461 |
| **Prob(Omnibus):** | 0.916 | **Jarque-Bera (JB):** | 0.231 |
| **Skew:** | -0.179 | **Prob(JB):** | 0.891 |
| **Kurtosis:** | 2.632 | **Cond. No.** | 1.00 |

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

## ****4 - Output Interpretation****

We will use Model 7 as it has the best R square value

1 - p-value < 0.01  
Thus the model is accepted

2 - coefficient == 1.64  
Thus if the value of Sorting Time is increased by 1, the predicted value of Delivery Time will increase by 1.64

3 - Adj. R-sqared == 0.987  
Thus the model explains 98.7% of the variance in dependent variable

In [ ]: