# import neccessery libraries

In [26]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sms
from numpy.polynomial.polynomial import polyfit
from sklearn.linear_model import LinearRegression

import warnings
warnings.filterwarnings('ignore')
```

# Delivery time > Predict delivery time using sorting time

# import data

In [3]:
```
import os
```

In [4]:
```
os.getcwd()
```

Out[4]:
```
'C:\\Users\\Akarsh\\assignment 4'
```

In [5]:
```
os.chdir('C:\\Users\\Akarsh\\Desktop\\assignments')
```

In [6]:
```
os.getcwd()
```

Out[6]:
```
'C:\\Users\\Akarsh\\Desktop\\assignments'
```

In [8]:
```
delivery_data = pd.read_csv('delivery_time.csv')
delivery_data
```

Out[8]:

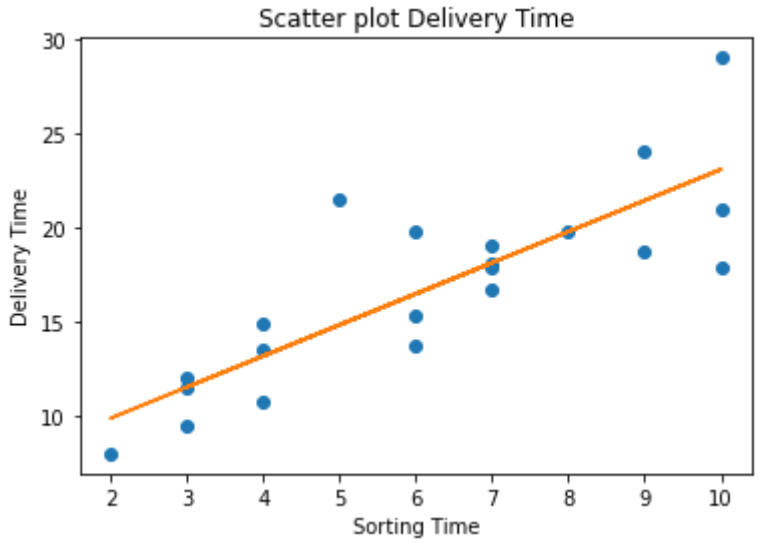
| | Delivery Time | Sorting Time |
|---|---|---|
| **0** | 21.00 | 10 |
| **1** | 13.50 | 4 |
| **2** | 19.75 | 6 |
| **3** | 24.00 | 9 |
| **4** | 29.00 | 10 |
| **5** | 15.35 | 6 |
| **6** | 19.00 | 7 |
| **7** | 9.50 | 3 |
| **8** | 17.90 | 10 |
| **9** | 18.75 | 9 |

| | Delivery Time | Sorting Time |
|---|---|---|
| **10** | 19.83 | 8 |
| **11** | 10.75 | 4 |
| **12** | 16.68 | 7 |
| **13** | 11.50 | 3 |
| **14** | 12.03 | 3 |
| **15** | 14.88 | 4 |
| **16** | 13.75 | 6 |
| **17** | 18.11 | 7 |
| **18** | 8.00 | 2 |
| **19** | 17.83 | 7 |
| **20** | 21.50 | 5 |

## display the scatter plot

In [14]:
```
x = delivery_data['Sorting Time']
y = delivery_data['Delivery Time']
b, m = polyfit(x, y, 1)
```

In [13]:
```
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()
```




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

# Correlation Analysis

In [15]:
```
corr = np.corrcoef(x, y)
```

In [16]:
```
corr
```

Out[16]:
```
array([[1.        , 0.82599726],
       [0.82599726, 1.        ]])
```

# correlation between delivery time and sorting Time is 83% high

# Regression Model

In [18]:
```
model = sms.OLS(y, x).fit()
predictions = model.predict(x)
```

In [19]:
```
model.summary()
```

Out[19]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 6.12e-15 |
| **Time:** | 13:32:48 | **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 |

Notes:

[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.

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

# Log Transformation of X

In [39]:
```
x_log = np.log(delivery_data['Sorting Time'])
```

In [40]:
```
model = sms.OLS(y, x_log).fit()
predictions = model.predict(x_log)
model.summary()
```

Out[40]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 1.48e-17 |
| **Time:** | 13:46:21 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# Log Transformation of y

In [41]:
```
y_log = np.log(delivery_data['Delivery Time'])
```

In [42]:
```
model = sms.OLS(y_log, x).fit()
predictions = model.predict(y_log)
model.summary()
```

Out[42]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 3.00e-12 |
| **Time:** | 13:46:22 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# Log Transformation of X & Y

In [43]:
```
model = sms.OLS(y_log, x_log).fit()
predictions = model.predict(x_log)
model.summary()
```

Out[43]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 5.72e-17 |
| **Time:** | 13:46:23 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# square root tranformation of x

In [44]:
```
x_sqrt = np.sqrt(delivery_data['Sorting Time'])
```

In [45]:
```
model = sms.OLS(y, x_sqrt).fit()
predictions = model.predict(x_sqrt)
model.summary()
```

Out[45]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 1.88e-17 |
| **Time:** | 13:46:27 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# square root tranformation of y

In [35]:
```
y_sqrt = np.sqrt(delivery_data['Delivery Time'])
```

In [46]:
```
model = sms.OLS(y_sqrt, x).fit()
predictions = model.predict(y_sqrt)
model.summary()
```

Out[46]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 5.09e-13 |
| **Time:** | 13:47:07 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# Square Root Transformation of X & Y

In [50]:
```
model = sms.OLS(y_sqrt, x_sqrt).fit()
predictions = model.predict(x_sqrt)
model.summary()
```

Out[50]:

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:** | Sat, 12 Feb 2022 | **Prob (F-statistic):** | 2.10e-20 |
| **Time:** | 13:48:50 | **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 |

Notes:
[1] $R^2$ is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.

# We will use square root transformation of X & Y 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 [ ]: