```python
from google.colab import files
uploaded = files.upload()
```

⇅ [Choose Files] chicago_taxi_fares.csv
  • **chicago_taxi_fares.csv**(text/csv) - 113 bytes, last modified: 6/26/2025 - 100% done
  Saving chicago_taxi_fares.csv to chicago_taxi_fares.csv

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

# Load CSV data
data = pd.read_csv("chicago_taxi_fares.csv")

# Assigning Dependent and Independent variables
x = data.iloc[:, :-1].values         # trip_distance as input (X)
y = data.iloc[:, 1].values           # fare_amount as output (y)

# Splitting dataset into training and testing sets (80% train, 20% test)
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=42)

# Training the Linear Regression model
model = LinearRegression()
model.fit(x_train, y_train)

# Predicting fares for test set
y_pred = model.predict(x_test)

# Printing model equation
print(f"Model: fare = {model.coef_[0]:.2f} * distance + {model.intercept_:.2f}")

# --- Plot for Training Set ---
plt.scatter(x_train, y_train, color='blue')
plt.plot(x_train, model.predict(x_train), color='red')
plt.title('FARE VS DISTANCE (Training Set)')
plt.xlabel('Trip Distance (miles)')
plt.ylabel('Fare Amount ($)')
plt.grid(True)
plt.show()

# --- Plot for Testing Set ---
plt.scatter(x_test, y_test, color='blue')
plt.plot(x_train, model.predict(x_train), color='red')  # Use train data for the line
plt.title('FARE VS DISTANCE (Testing Set)')
plt.xlabel('Trip Distance (miles)')
plt.ylabel('Fare Amount ($)')
plt.grid(True)
plt.show()
```

```
Model: fare = 2.73 * distance + 1.77
```

FARE VS DISTANCE (Training Set)

FARE VS DISTANCE (Testing Set)