Department of Computer Science and Engineering (Data Science)

Subject: Machine Learning – I (DJ19DSC402)

AY: 2022-23

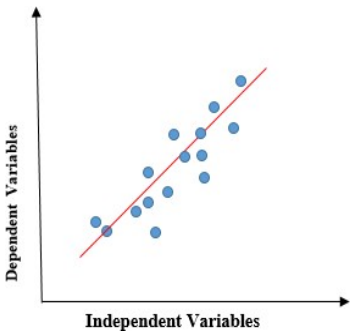
Experiment 1

(Regression)

Aim: Implement Linear Regression on the given Dataset and apply Regularization to overcome overfitting in the model.

Theory:

- **Linear Regression**: Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. *If there is a single input variable (x), such linear regression is called* ***simple linear regression****. And if there is more than one input variable, such linear regression is called* ***multiple linear regression*****.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best. *To calculate best-fit line linear regression uses a traditional slope-intercept form.*

$$y = mx+b \implies y = a_0+a_1x$$

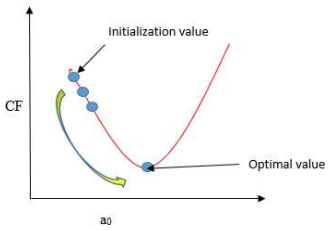
y= Dependent Variable; x= Independent Variable; a0= intercept; a1 = Linear regression coefficient.

- **Cost function:** The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**. In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values. *By simple linear equation y=mx+b we can calculate MSE as: Let's y = actual values, y_i = predicted values*

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Model parameters ***xi, b (a_0, a_1)*** can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

- **Gradient descent:** Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE). A regression model uses gradient descent to update the coefficients of the line (a0, a1 => xi, b) by reducing the cost function by a random selection of coefficient values and then iteratively update the values to reach the minimum cost function.



To update a_0 and a_1, we take gradients from the cost function. To find these gradients, we take partial derivatives for a_0 and a_1.

$$J = \frac{1}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)^2$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(a_0 + a_1 \cdot x_i - y_i) \cdot x_i$$

$$\frac{\partial J}{\partial a_0} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$\frac{\partial J}{\partial a_1} = \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

$$a_0 = a_0 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i)$$

$$a_1 = a_1 - \alpha \cdot \frac{2}{n}\sum_{i=1}^{n}(pred_i - y_i) \cdot x_i$$

- **Regularization:** When linear regression is underfitting there is no other way (given you can't add more data) then to increase complexity of the model making it polynomial regression (cubic, quadratic, etc…) or using other complex model to capture data that linear regression cannot capture due to its simplicity. When linear regression is overfitting, number of columns(independent variables) approach number of observations there are two ways to mitigate it

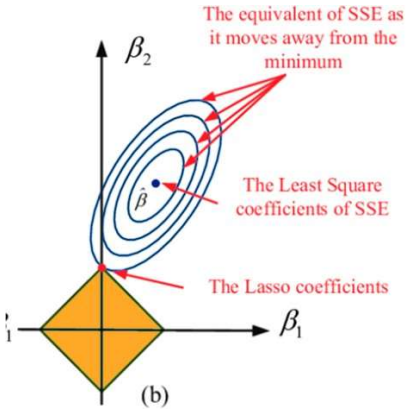
  1. Add more observations
  2. Regularization

  Since adding more observations is time consuming and often not provided we will use regularization technique to mitigate overfitting. There are multiple regularization techniques, all

share the same concept of **adding constraints on weights** of independent variables(except theta_0) however they differ in way of constraining. We will go through three most popular regularization techniques: Ridge regression (L2) and Lasso regression (L1)

- **Lasso Regression**

The word "LASSO" denotes Least Absolute Shrinkage and Selection Operator. Lasso regression follows the regularization technique to create prediction. It is given more priority over the other regression methods because it gives an accurate prediction. Lasso regression model uses shrinkage technique. In this technique, the data values are shrunk towards a central point similar to the concept of mean. The lasso regression algorithm suggests a simple, sparse models (i.e. models with fewer parameters), which is well-suited for models or data showing high levels of multicollinearity or when we would like to automate certain parts of model selection, like variable selection or parameter elimination using feature engineering. Lasso Regression algorithm utilises L1 regularization technique It is taken into consideration when there are more number of features because it automatically performs feature selection.



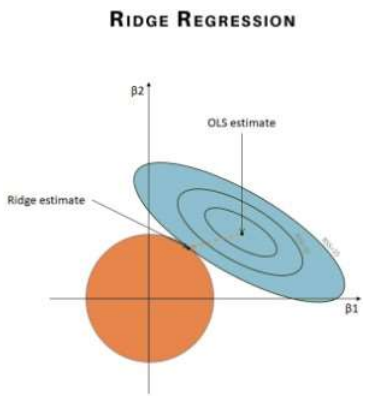
Residual Sum of Squares + λ * (Sum of the absolute value of the coefficients)

The equation looks like:

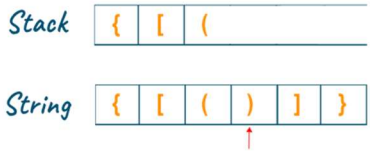
$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

- **Ridge Regression**

Ridge Regression is another type of regression algorithm in data science and is usually considered when there is a high correlation between the independent variables or model parameters. As the value of correlation increases the least square estimates evaluates unbiased values. But if the collinearity in the dataset is very high, there can be some bias value. Therefore, we create a bias matrix in the equation of Ridge Regression algorithm. It is a useful regression method in which the model is less susceptible to overfitting and hence the model works well even if the dataset is very small.



The cost function for ridge regression algorithm is:



Where λ is the penalty variable. λ given here is denoted by an alpha parameter in the ridge function. Hence, by changing the values of alpha, we are controlling the penalty term. Greater the values of alpha, the higher is the penalty and therefore the magnitude of the coefficients is reduced.We can conclude that it shrinks the parameters. Therefore, it is used to prevent multicollinearity, it also reduces the model complexity by shrinking the coefficient.

Lab Assignments to complete in this session

Use the given dataset and perform the following tasks:

Dataset 1: Simulate a sine curve between 60° and 300° with some random noise.

Dataset 2: food_truck_data.csv

Dataset 3: housing.csv

1. Perform Linear Regression on Dataset 1 and Dataset 2 by computing cost function and gradient descent from scratch.

2. Use sklearn to perform linear regression on Dataset 2, show the scatter plot for best fit line using matplotlib and show the results using MSE.

3. To perform regularization on linear model build using Linear Regression on Dataset2.

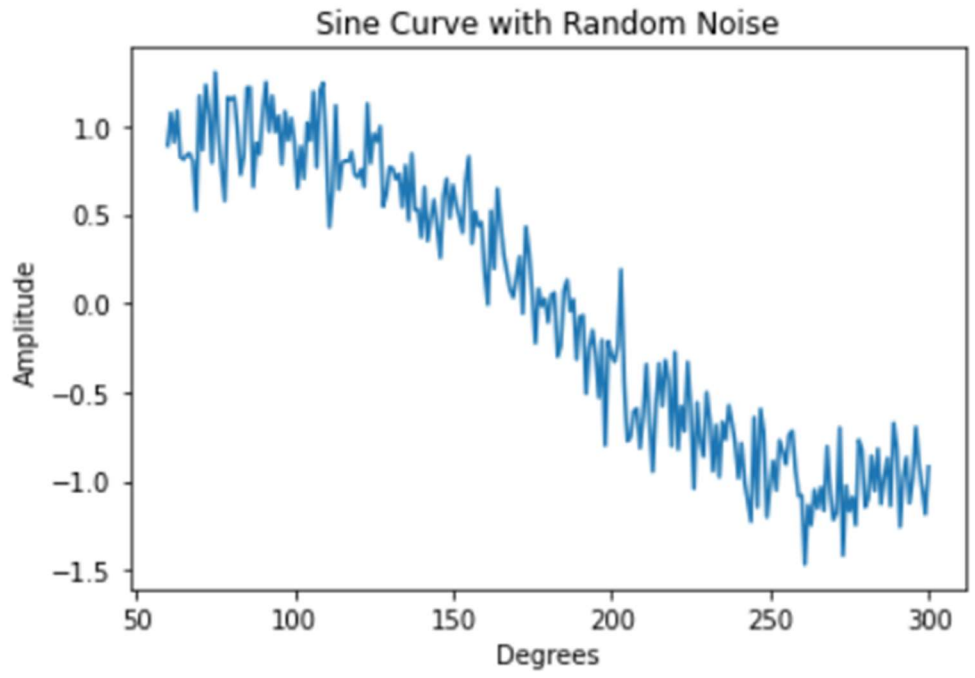
1.

1.SINE CURVE BETWEEN 60 TO 300 WITH NOISE

```
import numpy as np
import matplotlib.pyplot as plt

# parameters
start_degrees = 60
end_degrees = 300
step_degrees = 1
amplitude = 1
noise_amplitude = 0.2
```

```
[7] #sine wave
degrees = np.arange(start_degrees, end_degrees + step_degrees, step_degrees)
radians = np.deg2rad(degrees)
sine_wave = amplitude * np.sin(radians)

# noise
noise = noise_amplitude * np.random.randn(len(sine_wave))
sine_wave_with_noise = sine_wave + noise
```

```
plt.plot(degrees, sine_wave_with_noise)
plt.xlabel('Degrees')
plt.ylabel('Amplitude')
plt.title('Sine Curve with Random Noise')
plt.show()
```



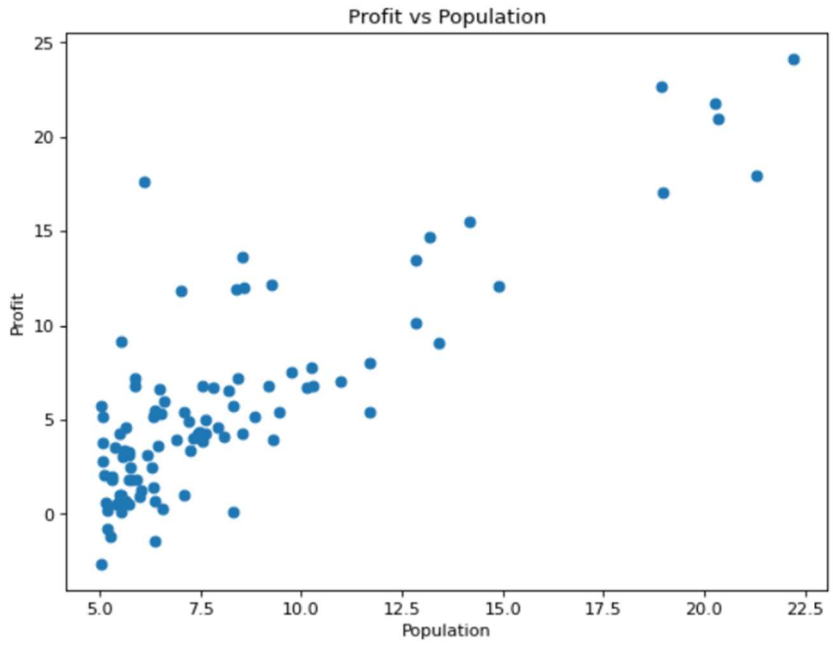
FOOD TRUCK DATA SET

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from matplotlib.pyplot import figure
```

```
<Figure size 640x480 with 0 Axes>
<Figure size 640x480 with 0 Axes>
```

```
[40] df=pd.read_csv("/content/foodtruck.txt")
```

```
figure(figsize=(8, 6), dpi=80)
plt.scatter(df['Population'], df['Profit'])
plt.xlabel('Population')
plt.ylabel('Profit')
plt.title('Profit vs Population')
plt.show()
```



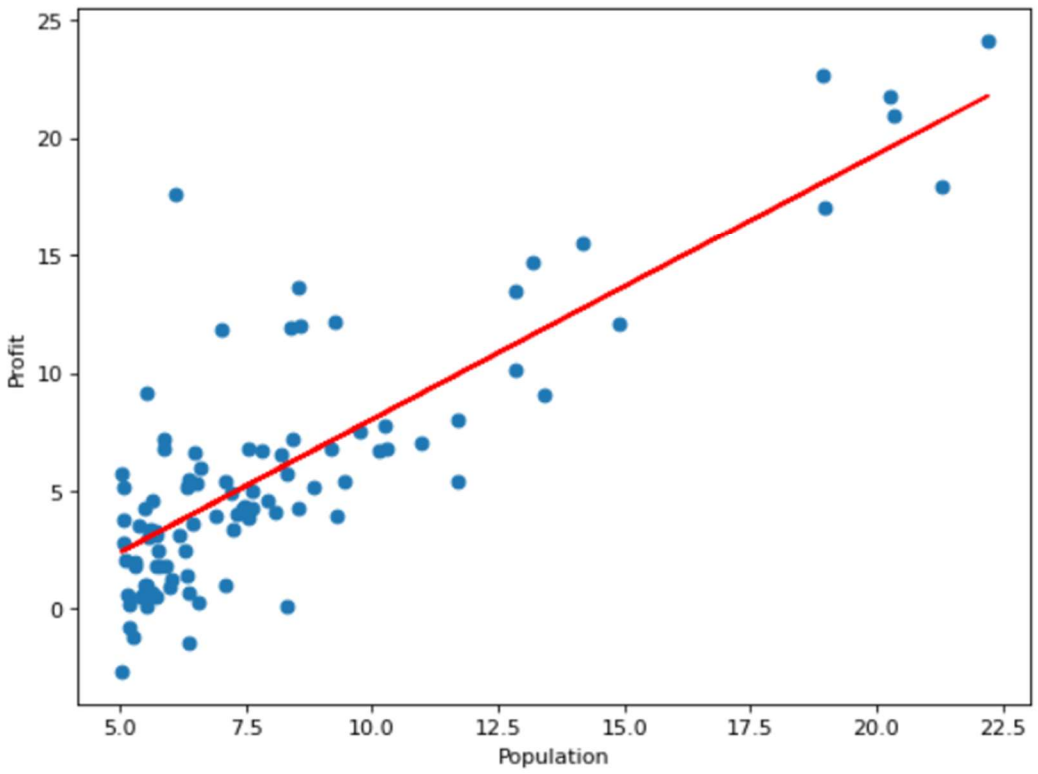
```
[32] def cost_function(X, y, theta):
         m = len(y)
         J = np.sum((X.dot(theta) - y) ** 2) / (2 * m)
         return J
```

```
[33] X = np.c_[np.ones(df.shape[0]), df['Population'].values]
     y = df['Profit'].values.reshape(-1, 1)
     theta = np.zeros((X.shape[1], 1))
```

```
     def gradient_descent(X, y, theta, alpha, num_iters):
         m = len(y)
         J_history = []

         for i in range(num_iters):
             theta = theta - (alpha / m) * X.T.dot(X.dot(theta) - y)
             J_history.append(cost_function(X, y, theta))

         return theta, J_history
```

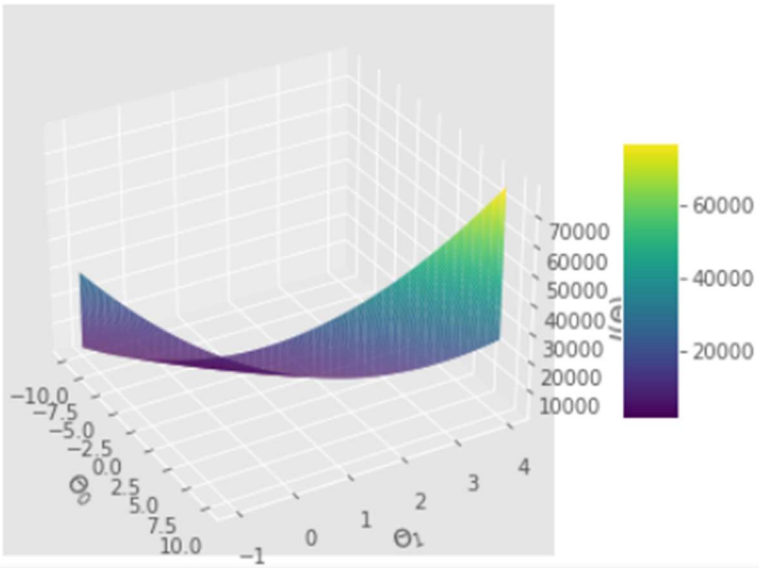
```
[35] alpha = 0.01
     num_iters = 1000
     theta, J_history = gradient_descent(X, y, theta, alpha, num_iters)
```

```
figure(figsize=(8, 6), dpi=80)
plt.scatter(df['Population'], df['Profit'])
plt.plot(df['Population'], X.dot(theta), color='r')
plt.xlabel('Population')
plt.ylabel('Profit')
plt.show()
```

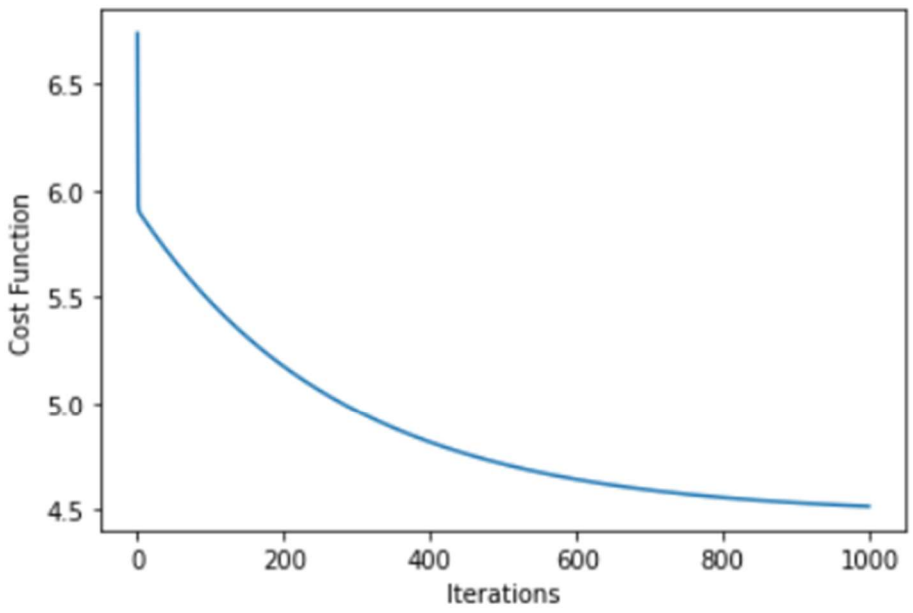


```
fig = plt.figure(figsize = (12, 5))
ax = fig.gca(projection ='3d')

surf = ax.plot_surface(theta_0, theta_1, cost_values, cmap ='viridis')
fig.colorbar(surf, shrink=0.5, aspect=5)

plt.xlabel("$\Theta_0$")
plt.ylabel("$\Theta_1$")
ax.set_zlabel("$J(\Theta)$")

ax.view_init(30, 330)

plt.show()
```

```
<ipython-input-18-c8d4d21a53c8>:2: MatplotlibDeprecationWarning: Calling
  ax = fig.gca(projection ='3d')
```



```
plt.plot(range(num_iters), J_history)
plt.xlabel('Iterations')
plt.ylabel('Cost Function')
plt.show()
```



```
[ ]  def predict(x, theta):
         y_pred = np.dot(theta.transpose(), x)
         return y_pred
```

```
y_pred_1 = predict(np.array([1, 4]), theta)*10000
y_pred_1
```

```
31700.53956989978
```

```
y_pred_2 = predict(np.array([1, 8.3]), theta)*10000
y_pred_2
```

```
66436.15389505132
```

SUING SKLEARN

```
[50] import numpy as np
     import matplotlib.pyplot as plt
     import pandas as pd
     import seaborn as sns
     import pandas as pd
     from sklearn.linear_model import LinearRegression
```

```
df=pd.read_csv("/content/foodtruck.txt")
df.head()
```

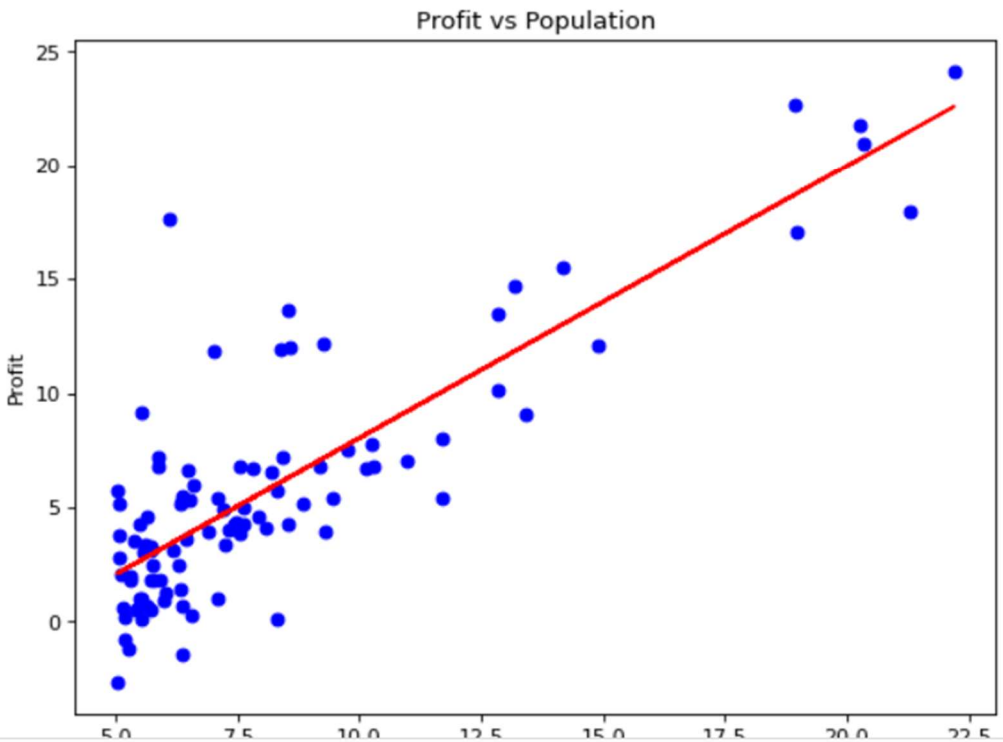
	Population	Profit
0	6.1101	17.5920
1	5.5277	9.1302
2	8.5186	13.6620
3	7.0032	11.8540
4	5.8598	6.8233

```
[53] X = df.iloc[:, :-1].values
     y = df.iloc[:, -1].values

     regressor = LinearRegression()
     regressor.fit(X, y)

     print('Coefficients:', regressor.coef_)
     print('Intercept:', regressor.intercept_)
```

0s

```
Coefficients: [1.19303364]
Intercept: -3.8957808783118484
```

```
figure(figsize=(8, 6), dpi=80)
plt.scatter(X, y, color='blue')

plt.plot(X, regressor.predict(X), color='red')

plt.xlabel('Population')
plt.ylabel('Profit')
plt.title('Profit vs Population')
plt.show()
```



HOUSING Dataset

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
from mpl_toolkits.mplot3d import Axes3D
from matplotlib.pyplot import figure
```

```
[11] df=pd.read_csv("/content/Housing.csv")
     df.head()
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

```
figure(figsize=(8, 6), dpi=80)
sns.scatterplot(df['area'],df['price']).set_title('PRICE V/S AREA')
plt.show()
```

cost function

```
def cost_function(X, y, theta):
  m = len(y)
  y_pred = X.dot(theta)
  error = (y_pred - y) ** 2
  return 1/ (2*m) * np.sum(error)

m = df.price.size
X = np.append(np.ones((m, 1)), df.price.values.reshape(m, 1), axis=1)
y = df.area.values.reshape (m, 1)
theta = np.zeros((2, 1))

cost_function (X,y, theta)
```

```
15614473.13669725
```

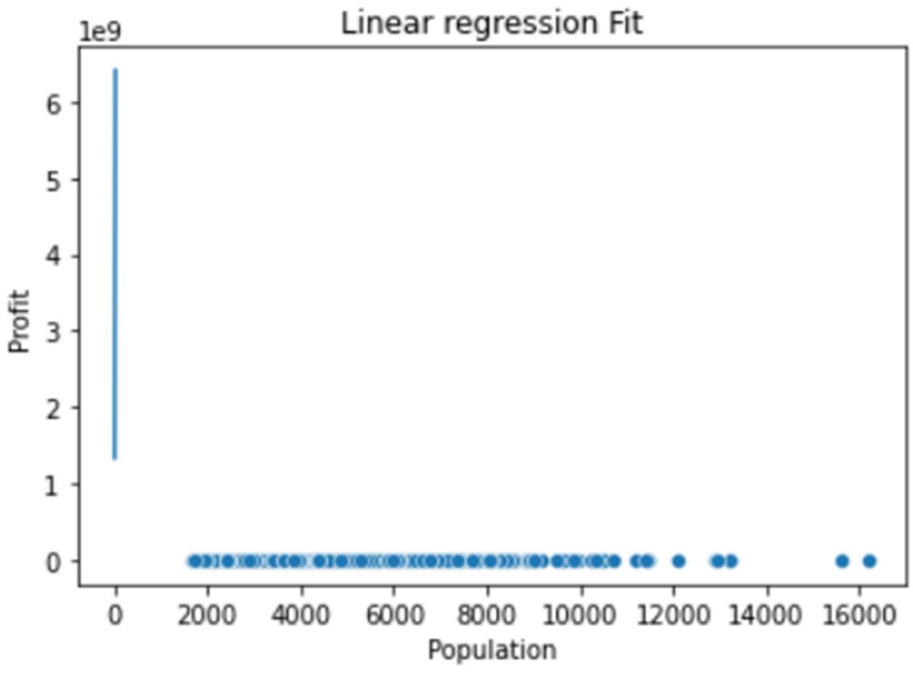
GRADIENT

```
def gradient_descent(X, y, theta, alpha, iterations):
  m = len(y)
  costs =[]
  for i in range(iterations):
    y_pred = X.dot(theta)
    error = np.dot(X.transpose(), (y_pred - y))
    theta -= alpha * 1/m *error
    costs.append(cost_function(X, y, theta))
  return theta, costs
theta, costs = gradient_descent(X, y, theta, alpha =0.01, iterations=1)

print("h(x) = {} + {}x1".format(str(round(theta[0, 0], 2)), str(round(theta[1, 0], 2))))
```

```
h(x) = 51.51 + 267229200.5x1
```

```
[24] theta.shape
     theta

     array([[5.15054128e+01],
            [2.67229201e+08]])
```

```
theta = np.squeeze(theta)
sns.scatterplot(df['area'],df['price'])

x_value = [x for x in range(5, 25)]
y_value = [(x * theta[1] + theta[0]) for x in x_value]
sns.lineplot(x_value, y_value)

plt.xlabel("Population")
plt.ylabel("Profit")
plt.title("Linear regression Fit");
```



3.

REGULARIZATION

```
import pandas as pd
from sklearn.linear_model import Ridge
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
```

```
data = pd.read_csv('/content/foodtruck.txt')
X_train, X_test, y_train, y_test = train_test_split(data.iloc[:, :-1], data.iloc[:, -1], test_size=0.2, random_state=42)

reg = Ridge(alpha=0.5)

reg.fit(X_train, y_train)

y_pred = reg.predict(X_test)

mse = mean_squared_error(y_test, y_pred)

print("Mean Squared Error:", mse)
```

```
Mean Squared Error: 15.700929974169245
```