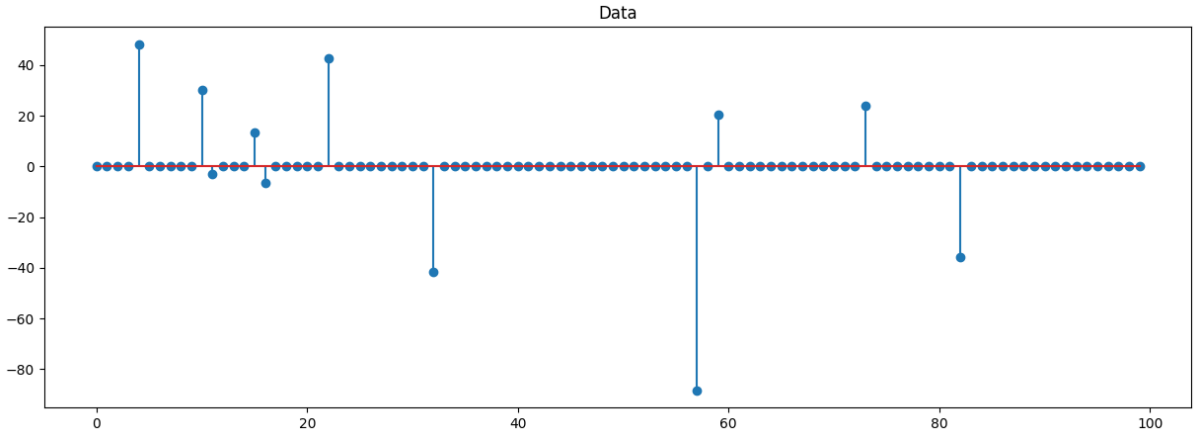
EN3150 Assignment 01
Learning from data and related challenges and linear models for regression
K.D. Wijeratne – 200722T

1. Data Preprocessing

2.



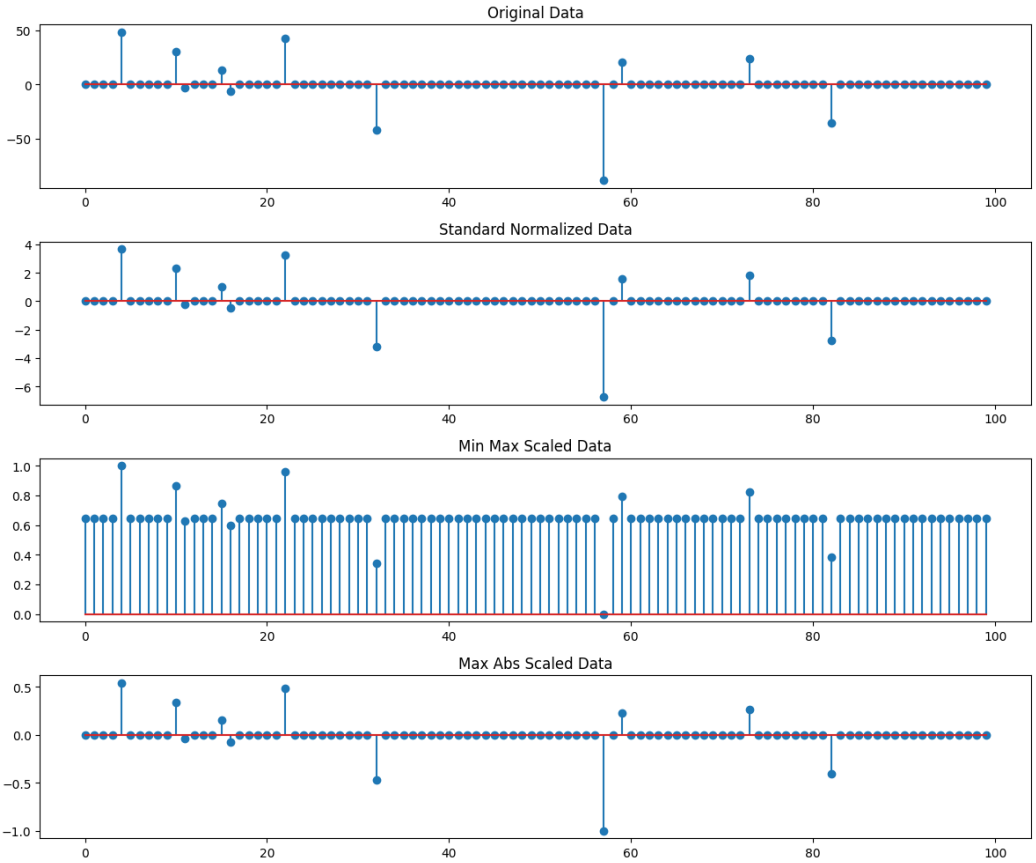
3.

```python
#3. MaxAbsScaler
from sklearn.preprocessing import MaxAbsScaler
scaler = MaxAbsScaler().fit(signal)
scaled_data_max_abs = scaler.transform(signal)
```

```python
#3. MinMaxScaler
def min_max_scale(data):
    min_val = np.min(data)
    max_val = np.max(data)
    scaled_data = (data - min_val) / (max_val - min_val)
    return scaled_data

scaled_data_min_max = min_max_scale(signal)
print("min of scaled data", np.min(scaled_data_min_max), "max of scaled data", np.max(scaled_data_min_max))
```

```python
#3. StandardNormalizer
def StandardNormalizer(data):
    mean = np.mean(data)
    std = np.std(data)
    scaled_data = (data - mean) / std
    return scaled_data

normalized_data_standard = StandardNormalizer(signal)
print("mean of scaled data", np.mean(normalized_data_standard), "std of scaled data", np.std(normalized_data_standard))
```

4.



5. Nonzero elements before and after the normalization has changed in standard normalization and in min max scaling but has remained the same in max abs scaling.

```
Number of non zero elements in the signal is 11
Number of non zero elements in the signal after standard normalization is 100
Number of non zero elements in the signal after min max scaling is 99
Number of non zero elements in the signal after max abs scaling is 11
```

6. Max Abs Scaler: Max Abs Scaler scales the data by dividing each data point by the maximum absolute value of the data set. It relatively maintains the relationship between data points and scales them to the range [-1,1]. This scaling method does not handle the outliers well as it is based on the maximum absolute value.

   Min Max Scaler: Scales the data to a specific range generally [0,1] and the scaling is done by subtracting the minimum value in the data set from each data point and then dividing by the difference of maximum value and minimum value of the dataset. This method also does not handle outliers well as it is dependent on maximum and minimum value.

   Standard Normalization: This method scales data to have a mean 0 and a standard deviation of 1. This is calculated by subtracting the mean of the dataset from each point and then dividing by the standard deviation. This does not scale the data set to a specific range like above methods and does not handle outliers well if there's extreme values in the data set.

7. Max Abs Scaler:
    a. Distribution: Does not change the original shape and preserves the relative relationship between datapoints.
    b. Structure: Maintains the original patterns in data
    c. Scale: Scales to a range of [-1,1]

Min Max Scaler

a. Distribution: The relative relationship between data points is consistent but the data points are shifted to a different range.
b. Structure: Maintains the original patterns in data
c. Scale: Scales the data into a range generally [0,1] based on minimum and maximum values.

Standard Normalization

a. Distribution: Centres the data around a mean of 0.b
b. Structure: Maintains the original patterns in data
c. Scale: Scaled by the standard deviation. But does not have specific range like previous methods.

Model: Standard Normalization. Assuming that the errors in linear regression model are normally distributed, we can use such data to compare the relative importance of different features and improve the interpretation of regression coefficients.

2.Linear regression on real world data

2.

```
# Split data into training and testing sets
X_train , X_test , y_train , y_test = train_test_split ( df [['TV' , 'radio' , 'newspaper']] , df ['sales'] , test_size =0.2 , random_state =42)
```

3,4.

```
For training data set:
RSS:  432.8207076930262
RSE:  1.6551046999139907
MSE2:  2.705129423081414
R2:  0.8957008271017818

           Coefficients  Standard Error  t-statistic   p-value
TV             0.044730        0.001111    40.263322  0.000000
radio          0.189195        0.002778    68.116584  0.000000
newspaper      0.002761        0.003059     0.902667  0.368077

For testing data set:
RSS_test:  126.96389415904413
RSE_test:  1.8278826848155574
MSE_test:  3.1740973539761033
R2_test:  0.899438024100912

           Coefficients  Standard Error  t-statistic   p-value
TV             0.044730        0.002517    17.768430  0.000000
radio          0.189195        0.006846    27.635915  0.000000
newspaper      0.002761        0.007031     0.392702  0.696734
```

5. Yes there is a relationship between advertising budgets and sales. By observing the p values we can conclude that there is a significant relationship between budgets allocated for TV and radio but there is no such significance in budget allocated for newspaper.

6. Radio budget. As the coefficient of radio is the highest, we can say that it has a higher contribution towards sales.

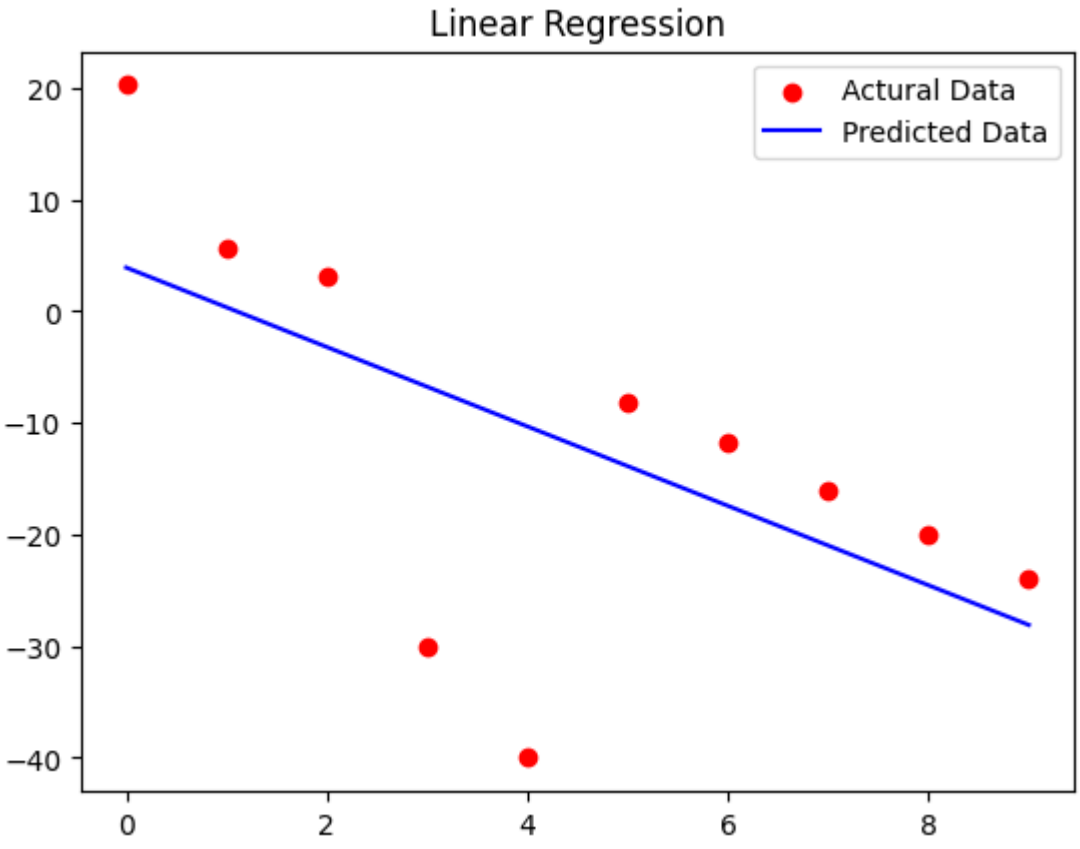
7.

```
Case 1:  [5851.09335992]
Case 2:  [2239.45494077]
Case 3:  [9462.73177906]
```

According to the above predictions we can say that allocating 50 000 USD to radio yields more sales than the other two ways. This result is obtained due to the high coefficient of regression in radio.

3.Linear regression impact on outliers

2.



4.

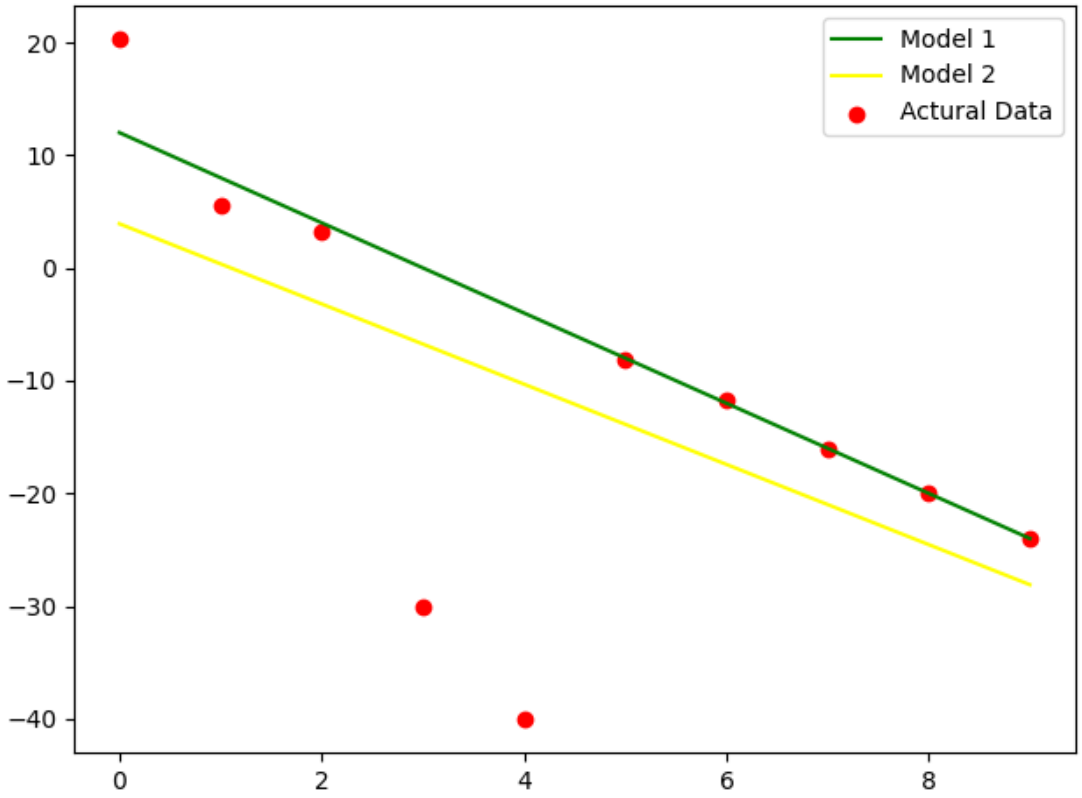
```
def loss_function(N,b, y, y_pred):
    y = np.array(y)
    y_pred = np.array(y_pred)
    return (1/N)*np.sum(((y-y_pred)**2)/((y-y_pred)**2 + b**2))

y1_pred = -4*x + 12
y2_pred = -3.55*x + 3.91

print ('Loss function for Model 1' , loss_function(10,1,y,y1_pred))
print ('Loss function for Model 2' , loss_function(10,1,y,y_pred))
```

```
Loss function for Model 1 0.435416262490386
Loss function for Model 2 0.9732472128655365
```

5.The Loss function (which corresponds to the error between actual values and predicted values) is less in model 1. Therefore, that model is much suitable for the above dataset.

6.The robust estimator reduces the weight of the outliers in the loss function by choosing model parameters which reduce the loss function thereby decreasing the adverse effects of the outliers.

7.



8. As Beta increases it places more emphasis on reducing the influence of outliers on the model parameters, making it more resistant to outliers. If this Beta value if too high, it could result in a model too focused in on reducing the impact of outliers that it underfits the data as it is resistant to capture the patterns of the majority data.

GitHub Link of the codes: https://github.com/KasuniWijeratne/EN3150---Pattern-Recognition