Keshav Shankar
ECE 1395
Assignment 2

1) $(X_1, X_2, Y) = [[1,1,8], [2,2,6], [3,3,4], [4,4,7]]$

$X_0 = [[1], [1], [1], [1]]$

theta 1 $= [0, 1, 0.5]$ theta 2 $= [10, -1, -1]$

theta 1 $J(\theta_0, \theta_1, \theta_2) = \frac{1}{2(4)} \sum_{1}^{4} (h_\theta(x^{(i)}) - y^{(i)})^2$

$$= \frac{1}{8}\Big[((0)(1) + (1)(1) + (0.5)(1) - 8)^2 + ((0)(1) + (1)(2) + (0.5)(2) - 6)^2 + ((0)(1) + (1)(3) + (0.5)(3) - 4)^2 + ((0)(1) + (1)(4) + (0.5)(4) - 7)^2\Big]$$

$$= \boxed{8.4375}$$

theta 2 $J(\theta_0, \theta_1, \theta_2) = \frac{1}{2(4)} \sum_{1}^{4} (h_\theta(x^{(i)}) - y^{(i)})^2$

$$= \frac{1}{8}\Big[((10)(1) + (-1)(1) + (-1)(1) - 8)^2 + ((10)(1) + (-1)(2) + (-1)(2) - 6)^2 + ((10)(1) + (-1)(3) + (-1)(3) - 4)^2 + ((10)(1) + (-1)(4) + (-1)(4) - 7)^2\Big]$$

$$= \boxed{0}$$

```
Cost of theta 1: 8.4375
Cost of theta 2: 0.0
```

My manual calculation matches that of the program, so I can verify it.

2)

```
Estimate of theta:
[[0.15008399]
 [0.53227665]
 [0.50840193]]
Cost after 15 iterations: 8.31
```

3)

```
Estimate of theta:
[[10.]
 [-1.]
 [-1.]]
```

- Yes, there is a significant difference.
- This is because gradient descent is iterative while the normal equation computes in one iteration.
- There are various things we can do to ensure a same result.
  - feature scaling: normalize the data so gradient descent isn't as sensitive. (regularization)
  - learning rate: adjust gradient descent learning rate so its not too large/small to cause divergence or slow convergence.

4c)

```
X has 179 points and 2 features per point.
Y has 179 points and 1 labels per point.
```

4e)

```
Estimate of theta:
[[6.3621373 ]
 [7.07005981]]
```

4g)

```
The error in prediction using gradient descent is: 17.56 %
The error in prediction using normal is: 11.26 %
```

↑ Also image for 4h

4h) The error in prediction using gradient descent is larger than in using the normal equation. This is most likely since gradient descent is for larger data sets with more features. This explains gradient descent's convergence issues as well.

4i) Using an α of 0.003 was best since the line most looks like a convex, and it also converges nearing 0 cost. An α of 0.001 does not curve enough, and α's of 0.03 & 3 converge too quickly.

5a)

```
Engine Size: Mean = 1611.11 , Std = 383.53
Car Weight: Mean = 1292.28 , Std = 238.74
CO2 Emission: Mean = 102.03 , Std = 7.35
Size of X: 3
Size of y: 1
```

note: X includes added bias feature of 1's.

5b)

```
Estimate of theta:
[[5.34080907e-05]
 [3.37662671e-01]
 [3.14894136e-01]]
```

5c)

```
Prediction for co2 emission: 106.587
```

note, this is un-normalized (converted back after calc.)