Machine Learning
Assignment 1 & 2 Report

Name: Mai Khaled

ID: 40-2019

First of all, we read the csv file using pandas libraries with the help of this method **pd.read_csv().** The csv files contained some NaN values so we replaced them with zeros using **replace(np.nan, 0**). We need to preprocess data before passing it to any machine learning model. Normalization is useful when the data has varying scales or when the features have different ranges. To normalize the features, we got each feature subtracted it form the mean and then we divided it by the standard deviation. The features in the dataset started with the 3rd column.

The features used were the bedrooms and bathrooms located in the 3rd and 4th columns in the dataset. We also used the 2nd column because it contains the prices of the houses.

Then, we took the data in the dataset and we split it using **np.split()** to 60% training set, 20% cross validation and 20% test set.

Moreover, we computed the cost function on the training features and training prices, and theta by using the equation below.

h=np.dot(train_x,theta)

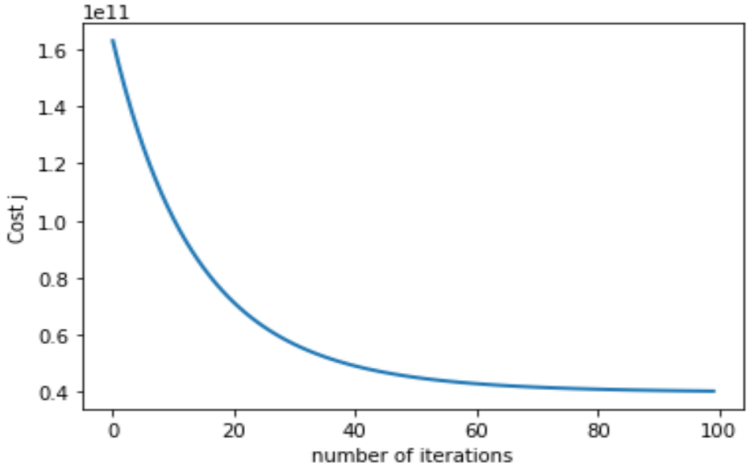
j = (1/(2 * train_m)) * np.sum((np.square(h - train_y)))

And then we computed the gradient descent to get the value of the theta that will lead to the minimum cost. The gradient descent takes training features, training prices, theta, alpha, and number of iterations. The equation below shows how it was computed.

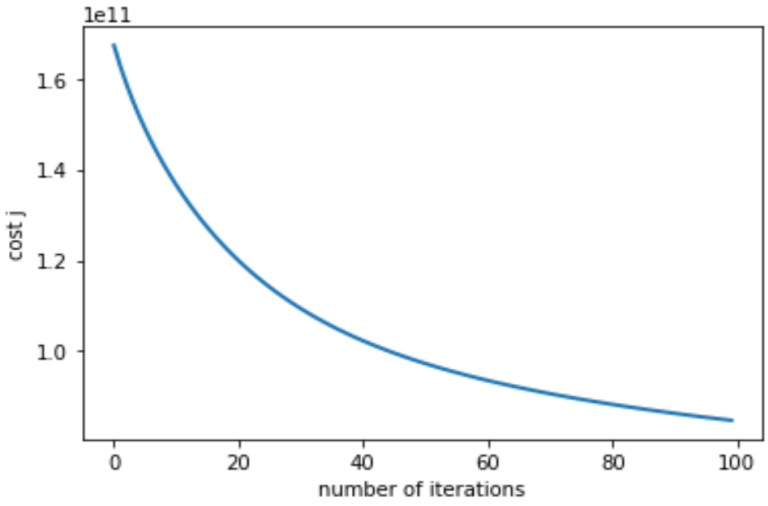
h=np.dot(train_x,theta)

theta =theta-((alpha/train_m)*(np.dot(train_x.T,h- train_y)))

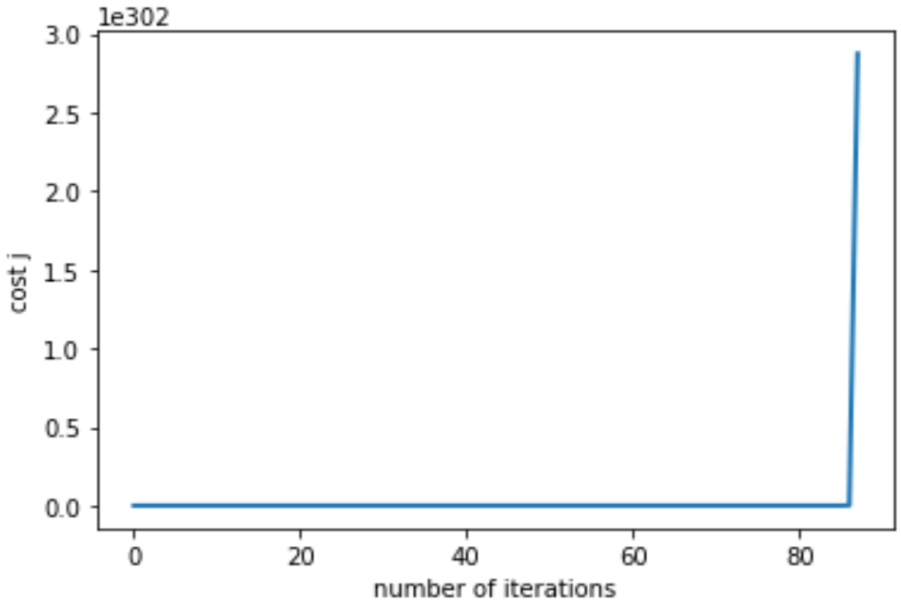
1. Trying out **alpha=0.03** and **number of iterations=100**, the graph showed that by increasing the number of iterations the cost decreased which showed that they are inversely proportional. And this was our target to get the minimum cost function for different values of theta.



2. By changing the degree for our hypothesis (h) to **degree 2**, we recomputed the cost function and the gradient descent. It shows that the difference between the cost and number of iterations is not that large as in **degree 1**. **Alpha=0.01** and **number of iterations= 100**

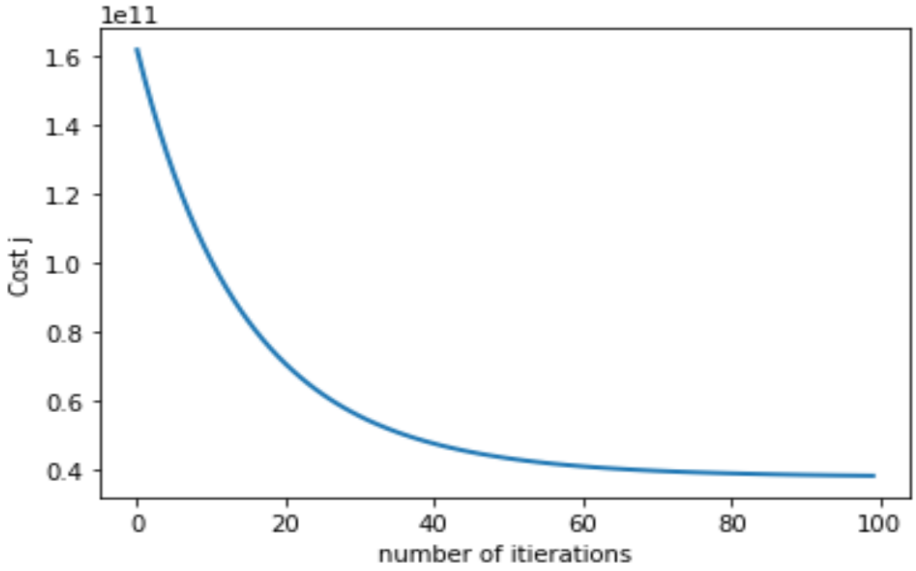


3. After that, we change the degree of our hypothesis (h) to **degree 3**. And passing it to the cost function and the gradient descent. The graph below shows that the power of 3 in the hypothesis will lead to very large values when multiplying it with the large values in the features which will lead to an overflow in the values.
   The graph also shows the cost will increase when we have a large increase in the number of iterations. **Alpha=0.01** and **number of iterations= 100**

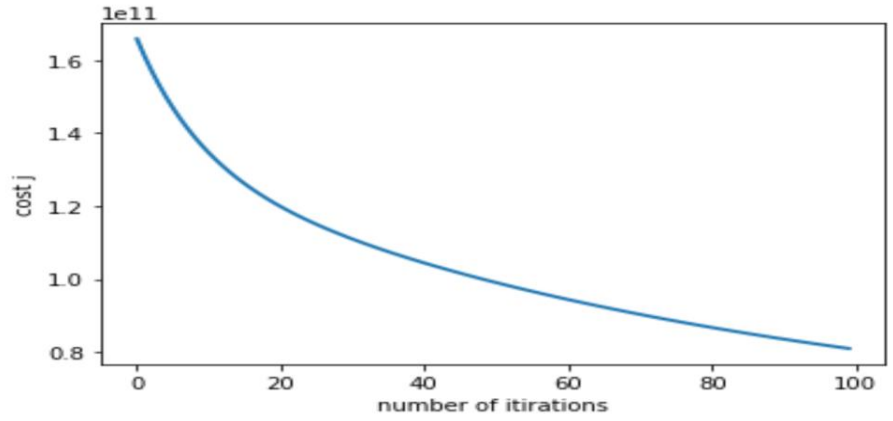


Now by moving to the **cross validation set**, we also applied to it the cost function and the gradient descent function.

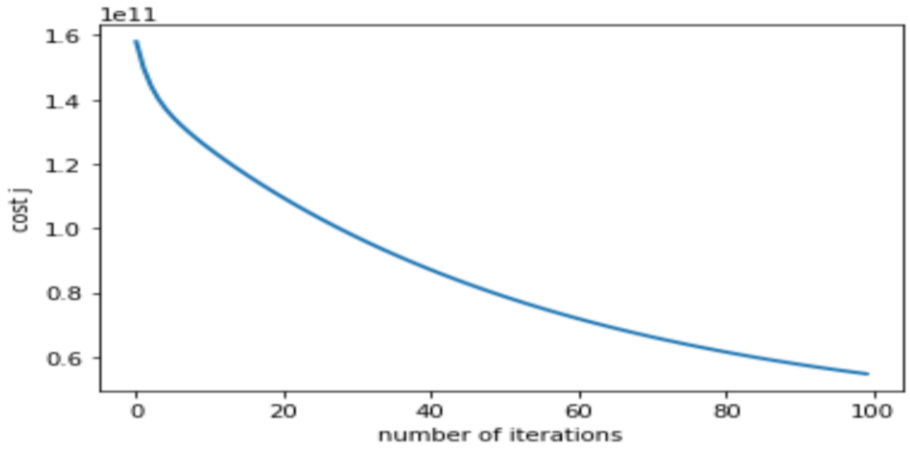
1. **Alpha= 0.03 and number of iterations =100** with hypothesis **degree of 1**
   Large decrease in the cost function without decreasing so much in the number of iterations.



2. **Alpha= 0.01 and number of iterations =100** with hypothesis **degree of 2**
   The cost does not decrease much when increasing the number of iterations.

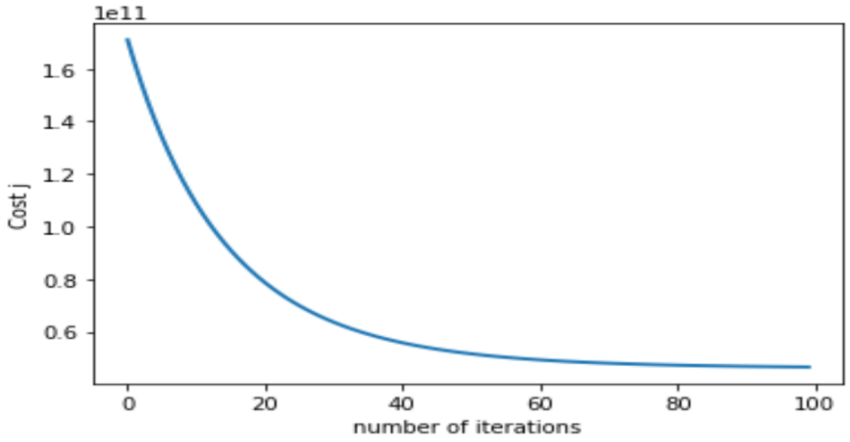


3. **Alpha= 0.01 and number of iterations =100** with hypothesis **degree of 3.**
   This graph just shows slight difference then the one with degree of 2.

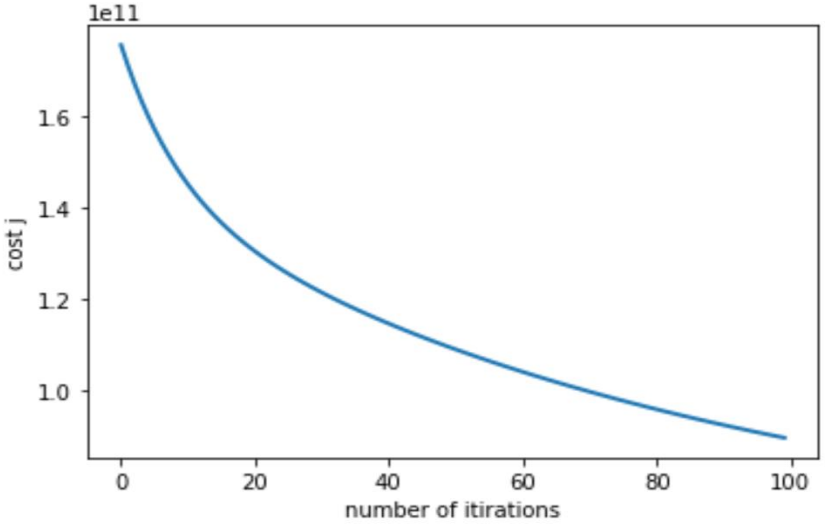


Finally, by moving to the **test set** we also applied to it the cost function and the gradient descent function.

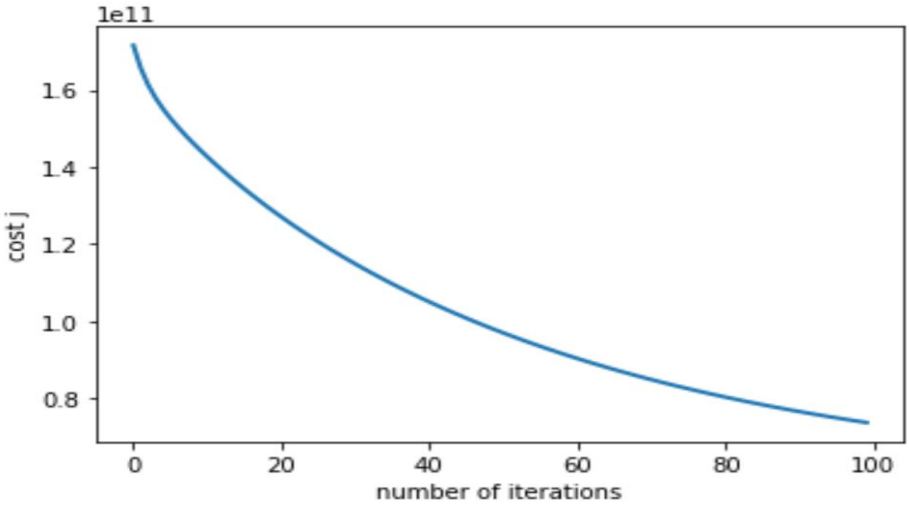
1. **Alpha= 0.03 and number of iterations =100** with hypothesis **degree of 1**
   Large decrease in the cost function without decreasing so much in the number of iterations.



2. **Alpha= 0.01 and number of iterations =100** with hypothesis **degree of 2**
   The cost does not decrease much when increasing the number of iterations.



3. **Alpha= 0.01 and number of iterations =100** with hypothesis **degree of 3.**
   This graph just shows slight difference then the one with degree of 2.



In conclusion, the best hypothesis is the one with degree 1 because the graph shows that it's the one with the minimum cost in relation with the other degrees of the hypothesis. And this hypothesis contains the two features bedrooms and bathrooms.