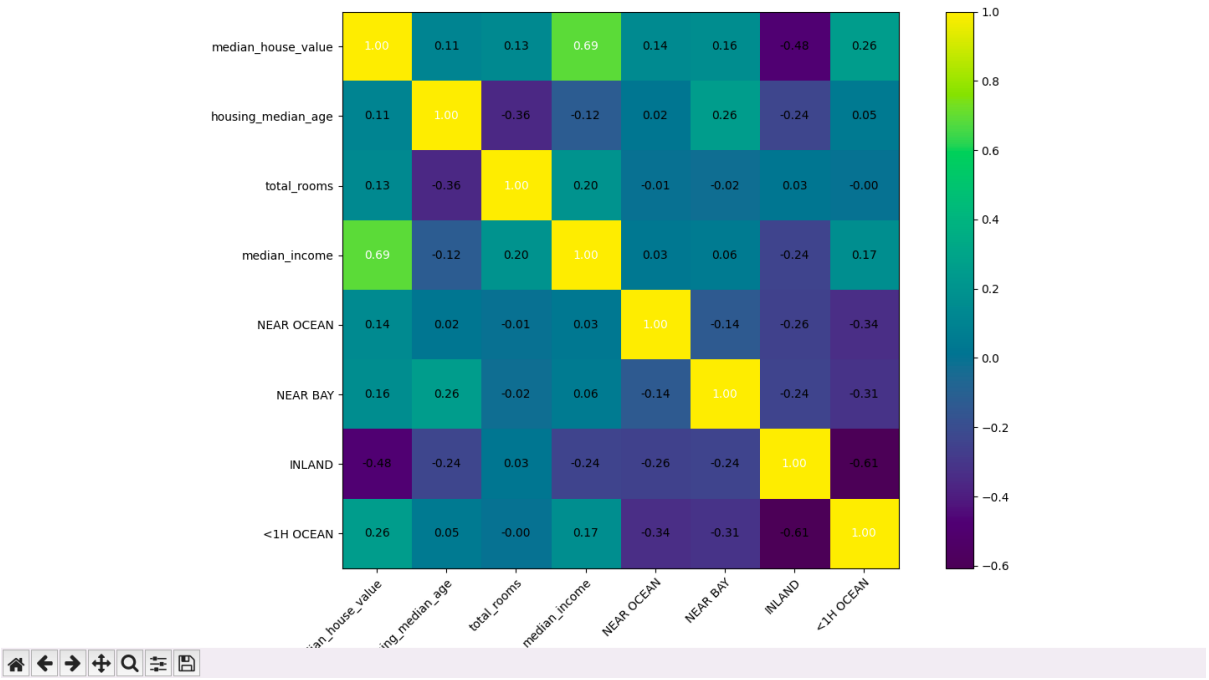
The very first step for the assignment was actually analyzing the data, exploring the features that were provided to us as input. This involves taking a look at our output value which is the median house value and seeing how it changes as our input features change. The target function which is the third element and unknown to us would give the ideal house value estimator. As for the last two elements, I used the data after carefully analyzing it, splitting it into testing and training, and then finally being able to get the hypothesis by predicting values for the house value.

A good way to check for the accuracy of the model would be by visuallizing our results on a graph, where we could have the difference between predicted values and actual values and check whether our model gives us results where the difference is minimal. This will help us in identifying how off our predicted values our from the actual values in the testing data.

For the feature selection, I tried for columns which specifically were correlated to our output column which was the median house value. Some of the features such as longitude and latitude made no sense at all so I did not consider them. For others, I looked carefully at the pearson plot and tried to get correlations which were greater than atleast 0.1. I know this is a small correlation but compared to some of the other features which had a very very low correlation with the median house value, it seemed a lot. I also tried to be careful about not choosing features which were highly correlated to eachother and not to the median housing value. Plotting a few scatterplots with the output variable helped to better see the correlation as well.



I think my model considered the average income and Inland to be the most influential features since these to had the highest correlation with the median value of the house. After 9-10 iterations of the training set, the graph begins to flatten out, so I think that's probably a good time to stop the training.

I played around with the learning rate for a bit just to get to the point where my cost and epoch graph would look reasonable and i would be able to get a good minimum. I found this number to be 0.15, the graph looks good and does not seem like it will diverge. I started off with a high learning rate and checked for low ones as well and found 0.15 to be the most optimal one.

I did thorough preprocessing on the dataset, I started off by changing the categorical data into numerical data in order for it to be fed to our model. For the nan values which were only present in one of the columns, i filled them out using the mean method. I made sure I had separated variables for my features and for the output variable. I also made sure there were no outliers in my dataset, removing them gave me better accuracy with the expected value closer to the actual value. I removed the outliers using the percentiles process mentioned in the slides. Finally, I split the data into testing and training by first randomizing the dataset and then assigning 80% of that to the training set and the remaining 20% to the testing.