Assignment #2 Regression Report

The dataset I selected for the sake of this assignment was the Boston Housing Prices dataset which is detailed here (https://archive.ics.uci.edu/ml/machine-learning-databases/housing/).

Our null hypothesis is that there is no relationship between a neighborhood being on the Charles River and the median value of owner-occupied homes in that area. We will use three control variables in our analysis which are detailed below. We will test at an alpha level of 0.05. We believe these controls are the most important features when valuing homes and have clear relationships with our target variables. Obviously, the average size and distance will be important as well as the age of the home. In a deeper analysis we would consider looking at the interactions as well.

The independent variables for this analysis were the following: 'RM' which is the average number of rooms per dwelling, 'DIS' which is the weighted distances to five Boston employment centers, 'AGE' which is the proportion of owner-occupied units built prior to 1940, and 'CHAS' which is a dummy variable for whether the neighborhood is on the Charles River (1 if the tract bounds river and 0 otherwise). The dependent variable is 'MEDV' which is the median value of owner-occupied homes in $1000's.

As evidenced in the tables below, we find that the p-value associated with the t-statistic of our test variable 'CHAS' is less than 0.001 and thus at our alpha level we can reject the null hypothesis that there is no relationship between a neighborhood being on the Charles River and the prices of owner-occupied homes in that neighborhood.

	coef	std err	[0.025	0.975]
RM	6.073439	0.228404	5.625767	6.521112
DIS	−1.098095	0.188232	−1.467030	−0.729159
AGE	−0.170218	0.012365	−0.194453	−0.145984
CHAS	5.449137	1.142720	3.209406	7.688868

	coef	std err	t	P>|t|	[0.025	0.975]
RM	6.0734	0.228	26.620	0.000	5.625	6.522
DIS	-1.0981	0.188	-5.840	0.000	-1.468	-0.729
AGE	-0.1702	0.012	-13.782	0.000	-0.194	-0.146
CHAS	5.4491	1.141	4.774	0.000	3.206	7.692