DSC 680 – White Paper

Project 2 – Milestone 2

Clayton Evans

***Business Problem***

The business problem for this project will be to compare the various variables to the median home price to see if there are other predictable factors other than the houses being along the Charles River to determine value.

***Background/History***

The background for the data is that it comes from the U.S. Census Bureau and is focused on the housing in the city of Boston, Massachusetts.

***Data Explanation***

The data is explained as below:

1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town.
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10 million)
6. RM - average number of rooms per dwelling
7. AGE - proportion of owner-occupied units built prior to 1940
8. DIS - weighted distances to five Boston employment centres
9. RAD - index of accessibility to radial highways
10. TAX - full-value property-tax rate per $10,000
11. PTRATIO - pupil-teacher ratio by town
12. B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
13. LSTAT - % lower status of the population
14. MEDV - Median value of owner-occupied homes in $1000's

***Methods***

Methods used will be various data cleansing, building test and train groupings, regression methods and then various graphing elements to show the correlations.

***Analysis***

The analysis will be a linear regression model as it is a very strong model to show the relationship between variables. Plus, from these models it is possible to build confidence interval and can help display which variables are of value and which are not.

***Conclusion***

The key factor that impacts the price of a house based on this data model is the number of rooms. The more rooms a house tends to have, the higher the price of the house. The key here is that all rooms are in that count not just bedrooms, etc.

***Assumptions***

It would be likely that the number of rooms would have an impact on the price of the house. That said, it is also assumed that if there were more elements to the data set such as size of the house, it too could have been of value.

***Limitations***

The limitations in this case really may only come from the amount of data over a set number of years. The data can still be of value, but may have been able to paint a much more broader picture with more years of data to see how trends change.

***Challenges***

The key challenge will be to ensure that one can show the true relevance of each column in the data set. It will be important to get a feel for what is truly an impact on the housing price or just data noise.

***Future Uses/Additional Applications***

Being that this examination uses a multivariate approach to analysis, it will have a myriad of possible other uses where there are a good number of variables and one key predictor variable that one is attempting to predict.

***Recommendations***

Ideas such as parks, size of house and location (crime, close to shopping, etc) will need to be considered as well to give the most accurate picture so for those data elements that are missing int would be crucial in a future run of the predictive model to have other variables.

***Implementation Plan***

Though there is a strong correlation for the price of homes based on number of rooms, there would be no set implementation as there could be other factors to consider. The key though is that we would know of one very solid predictor on the price of housing in Boston.

***Ethical Assessment***

This data has a column that can have race involved in the analysis thus the implications that racism could be a factor is something to consider. One will have to see how relevant the column is at it may need to be one in various model runs to consider removing.