Baltimore Buildings Dataset

Demari Green, Adrian Holley, Errol Stoute

Assisted by Radhouane Chouchane

Synopsis

Our Project consisted of finding the data of buildings in baltimore city. Our code would run in order to search for a target location/property. The project initially just shows a list of locations throughout the city which we narrow down into certain property specifics such as area, perimeter, or listing dates etc in order to help someone find a target location. We were going to then find a price based on size of the location and whether it was for sale or not. Then we were going to add a part that would allow us to see if the location was in a suitable condition and needed to be renovated and how much would the cost of renovation would be which would then give us a value for the building after completed renovation.

Setbacks

We went about cleaning up the dataset by removing columns that were empty or strings because they did not help us achieve the programs purpose. We kept to using ints and floats because that would help us go through the machine learning process. We needed to then be able to find out which buildings were able to be purchased and or needed to be renovated to a suitable condition.

Abstract/Goal

The main objective of our project was to make a program that uses information about a building that accurately returns the value of our target. The target feature within our dataset we chose to pick was AREA_ so this would be the main feature or column of values we would try and predict.. To do that we would need to load the data in and read the csv. Then we would need to split data in test sets and training sets by using scikit-learns. Then to do a linear regression on the training set so that we can then make the predictions on test set which would then return the predicted values to the test. Then from that, we'd compare the predicted values we got to the actual values and should be accurate or close to that. So overall, this would be the machine learning that predicts values within datasets accurately or close to. We also wanted to point out any outliers within are target which we did see.

Imports Used For The Project

Matplotlib - to create plots and graphs to gives a visual representation of what the machine is able to generate.

Pandas - a library that is used for datasets such as converting a csv file to a list.

Sklearn - used for machine learning to create models such a classification, regression, etc.

Numpy - working with arrays, matrices, and other mathematical functions.

Logistic Regression- predicts relationships between binary variables

Linear Discriminant Analysis - dimensionality reduction technique used to help classify