Capstone Project - IBM Data Science Professional Certificate

Presented By: Gaurav Duhoon

Influence of high leverage points

Problem Statement:

I am going to discuss an important issue of the influence of what are called high leverage points. And these are points that can be considered influential observation.

Python Tools:

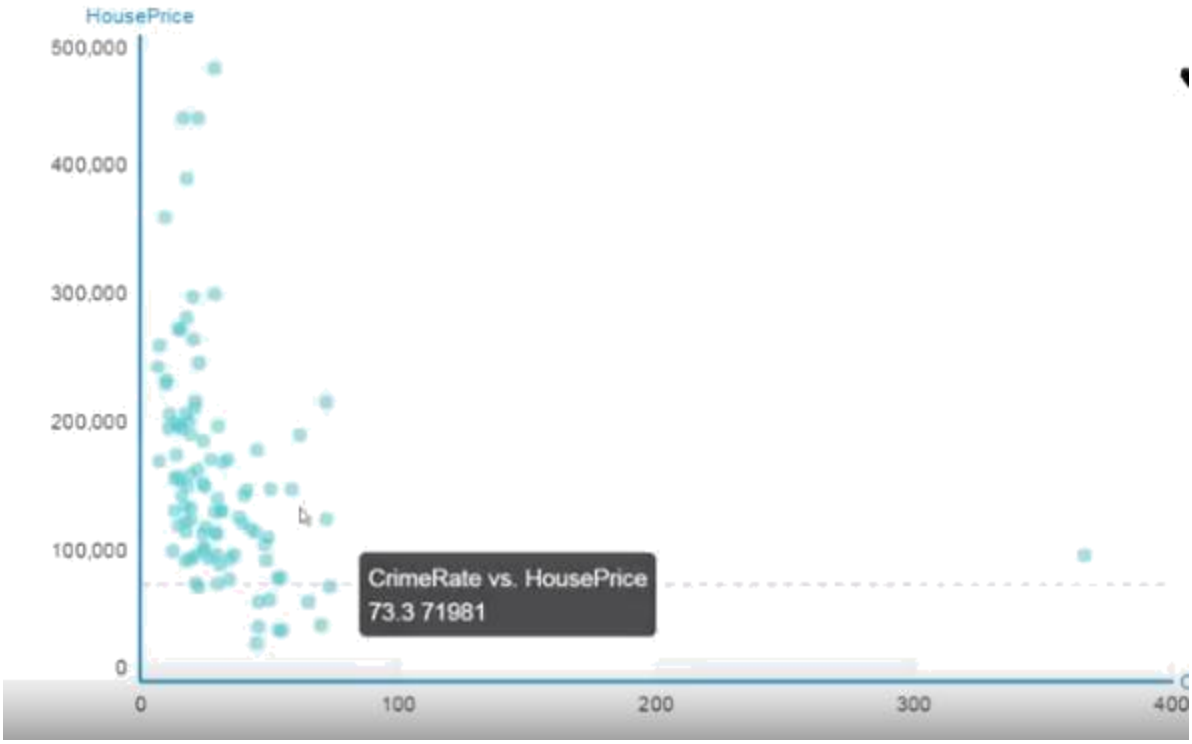
(i)SFrame

SFrame is really scalable data structure for dealing with big tables of data. And that data search with SFrame is part of package called GraphLab.

Load some house value vs. crime rate data

So, the dataset, where in particular, our dataset consists of the average house price in a whole collection of towns in the greater Philadelphia region. And we also have information about crime rates in each one of these towns. As well as how far that town is from center city and center city is the downtown region of Philadelphia. Dataset is from Philadelphia, PA and includes average house sales price in a number of neighborhoods. The attributes of each neighborhood we have include the crime rate ('CrimeRate'), miles from Center City ('MilesPhila'), town name ('Name'), and county name ('County').

By analyzing this data, we have made scatter plot of what's the relationship between average house sales prices, and crime rates.



Okay, so here we are, we're going do just command to show a scatter plot of, on the x axis we have crime rate. And each one of these little blue circles or cyan, light blue circles is a different town in our dataset. And we have a total of 98 different towns. And on the y axis what we have is the average house value in that town. Okay. And so, from this you can see that there's some relationship between our crime rate and our house sales price.

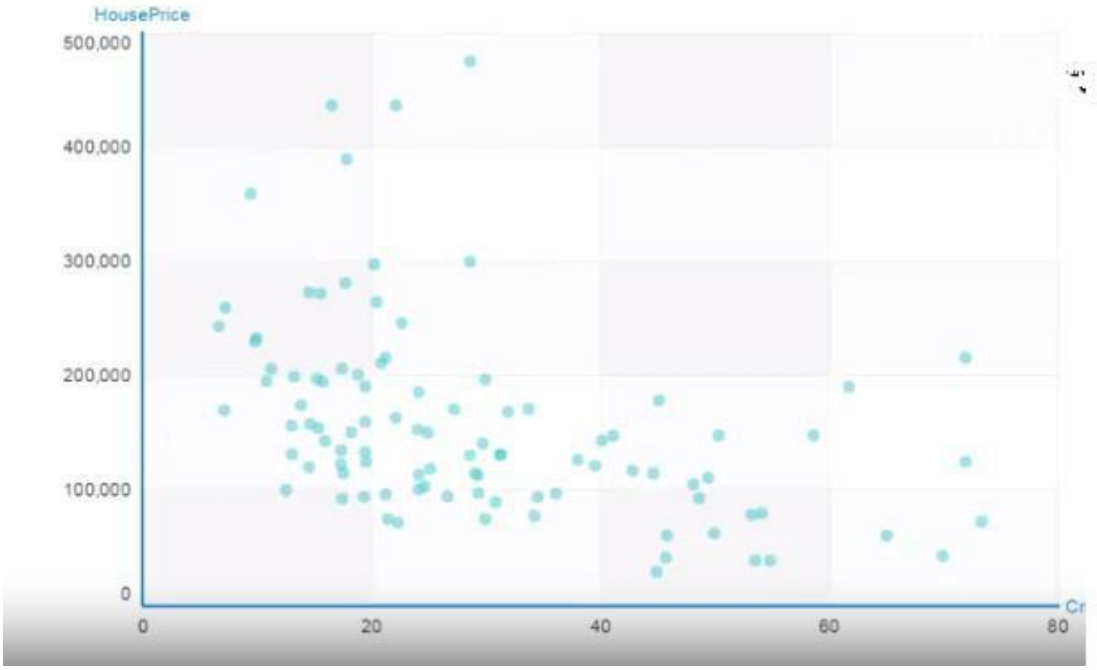
Exploring the data:

The house price in a town is correlated with the crime rate of that town. Low crime towns tend to be associated with higher house prices and vice versa.

Remove Center City and redo the analysis:

Center City is the one observation with an extremely high crime rate, yet house prices are not very low. This point does not follow the trend of the rest of the data

very well. A question is how much including Center City is influencing our fit on the other datapoints. Let's remove this datapoint and see what happens.



High leverage points:

Center City is said to be a "high leverage" point because it is at an extreme x value where there are not other observations. As a result, recalling the closed-form solution for simple regression, this point has the *potential* to dramatically change the least squares line since the center of x mass is heavily influenced by this one point and the least squares line will try to fit close to that outlying (in x) point. If a high leverage point follows the trend of the other data, this might not have much effect. On the other hand, if this point somehow differs, it can be strongly influential in the resulting fit.

Influential observations:

An influential observation is one where the removal of the point significantly changes the fit. As discussed above, high leverage points are good candidates for

being influential observations, but need not be. Other observations that are *not* leverage points can also be influential observations (e.g., strongly outlying in y even if x is a typical value).

Remove high-value outlier neighborhoods and redo analysis:

Based on the discussion above, a question is whether the outlying high-value towns are strongly influencing the fit. Let's remove them and see what happens.

Do the coefficients change much?

```
crime_model_noCC.get('coefficients')
```

]:

name	index	value
(intercept)	None	225204.604303
CrimeRate	None	-2287.69717443

[2 rows x 3 columns]

```
crime_model_nohighend.get('coefficients')
```

]:

name	index	value
(intercept)	None	199073.589615
CrimeRate	None	-1837.71280989

[2 rows x 3 columns]

Conclusion:

We see that removing the outlying high-value neighborhoods has *some* effect on the fit, but not nearly as much as our high-leverage Center City data points.