**CEBD 1100 Introduction to Python**

Rogerio Lessa Da Paixao

Activity week 8

Big Data Project.

Data analysis Boston Dataset

Description of the dataset (5 first rows)

The Boston dataset consisted of 14 attribute variables and 506 observations listed below:

* MED - Median value of owner-occupied homes (in $ 1000)
* CRIM - Per capita Crime rate by town
* ZN - Proportion of residential land zoned for lots over 25.000 sq.ft
* INDUS - Proportion of non-Retail Business Acres per Town
* CHAS - Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
* NOX - Nitric oxides concentration (parts per 10 million)
* RM - Average number of rooms per dwelling
* AGE - Proportion of Owner –occupied units built prior 1940
* DIS - Weighed distances to five Boston employment centers
* RAD - Index of accessibility to radial highways
* TAX - Full-value property-tax rate per $ 10.000
* PTRATIO – Pupil-teacher ratio by town
* B - 1000 (Bk – 0.63)^2 where Bk is the proportion of blacks by town
* LSTAT - % lower status of the population

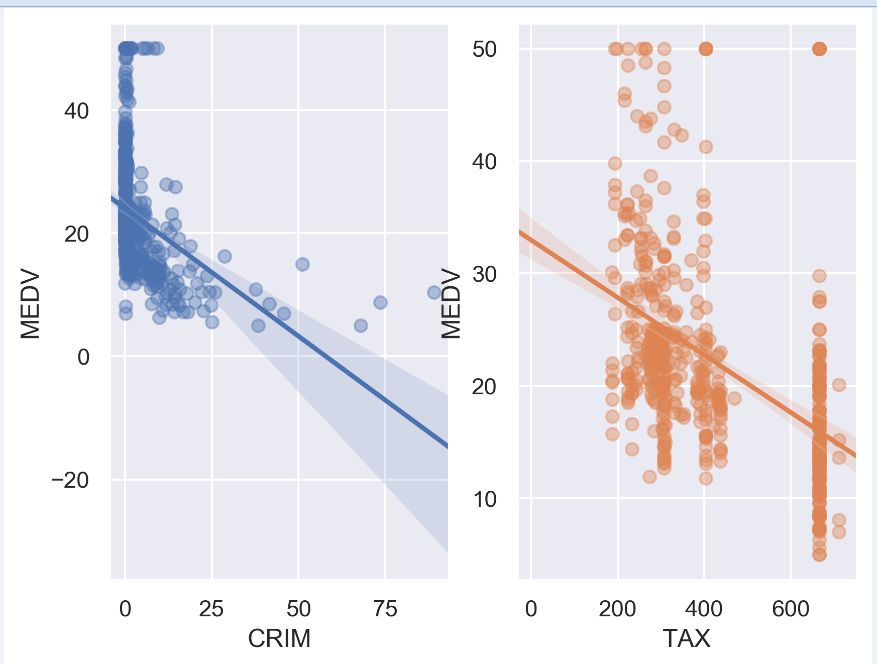
**Table 1. Attributes**

|  |
| --- |
| MEDV CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT |
| 0 24.0 0.00632 18.0 2.31 0.0 0.538 6.575 65.2 4.0900 1.0 296.0 15.3 396.90 4.98 |
| 1 21.6 0.02731 0.0 7.07 0.0 0.469 6.421 78.9 4.9671 2.0 242.0 17.8 396.90 9.14 |
| 2 34.7 0.02729 0.0 7.07 0.0 0.469 7.185 61.1 4.9671 2.0 242.0 17.8 392.83 4.03 |
| 3 33.4 0.03237 0.0 2.18 0.0 0.458 6.998 45.8 6.0622 3.0 222.0 18.7 394.63 2.94 |
| 4 36.2 0.06905 0.0 2.18 0.0 0.458 7.147 54.2 6.0622 3.0 222.0 18.7 396.90 5.33 |

The relationship between the independent variable MEDV (Median Value of Owner-Occupied Homes in $1000) and the independent variables CRIM (Per Capita Crime Rate by Town) and TAX (Full-value Property-tax Rate per $1000). According to the Fig.1 the median value of homes is higher when the crime rates are lower and the median value of the homes tend to decrease as the property tax increase.

The correlation among the variables can be observed in Table 1.

**Fig. 1 Linear regressions MEDV, CRIM and TAX**



**Table 2. Correlations**

|  |
| --- |
| **CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT target** |
| CRIM 1.000000 -0.200469 0.406583 -0.055892 0.420972 -0.219247 0.352734 -0.379670 0.625505 0.582764 0.289946 -0.385064 0.455621 -0.388305 |
| ZN -0.200469 1.000000 -0.533828 -0.042697 -0.516604 0.311991 -0.569537 0.664408 -0.311948 -0.314563 -0.391679 0.175520 -0.412995 0.360445 |
| INDUS 0.406583 -0.533828 1.000000 0.062938 0.763651 -0.391676 0.644779 -0.708027 0.595129 0.720760 0.383248 -0.356977 0.603800 -0.483725 |
| CHAS -0.055892 -0.042697 0.062938 1.000000 0.091203 0.091251 0.086518 -0.099176 -0.007368 -0.035587 -0.121515 0.048788 -0.053929 0.175260 |
| NOX 0.420972 -0.516604 0.763651 0.091203 1.000000 -0.302188 0.731470 -0.769230 0.611441 0.668023 0.188933 -0.380051 0.590879 -0.427321 |
| RM -0.219247 0.311991 -0.391676 0.091251 -0.302188 1.000000 -0.240265 0.205246 -0.209847 -0.292048 -0.355501 0.128069 -0.613808 0.695360 |
| AGE 0.352734 -0.569537 0.644779 0.086518 0.731470 -0.240265 1.000000 -0.747881 0.456022 0.506456 0.261515 -0.273534 0.602339 -0.376955 |
| DIS -0.379670 0.664408 -0.708027 -0.099176 -0.769230 0.205246 -0.747881 1.000000 -0.494588 -0.534432 -0.232471 0.291512 -0.496996 0.249929 |
| RAD 0.625505 -0.311948 0.595129 -0.007368 0.611441 -0.209847 0.456022 -0.494588 1.000000 0.910228 0.464741 -0.444413 0.488676 -0.381626 |
| TAX 0.582764 -0.314563 0.720760 -0.035587 0.668023 -0.292048 0.506456 -0.534432 0.910228 1.000000 0.460853 -0.441808 0.543993 -0.468536 |
| PTRATIO 0.289946 -0.391679 0.383248 -0.121515 0.188933 -0.355501 0.261515 -0.232471 0.464741 0.460853 1.000000 -0.177383 0.374044 -0.507787 |
| B -0.385064 0.175520 -0.356977 0.048788 -0.380051 0.128069 -0.273534 0.291512 -0.444413 -0.441808 -0.177383 1.000000 -0.366087 0.333461 |
| LSTAT 0.455621 -0.412995 0.603800 -0.053929 0.590879 -0.613808 0.602339 -0.496996 0.488676 0.543993 0.374044 -0.366087 1.000000 -0.737663 |
| target -0.388305 0.360445 -0.483725 0.175260 -0.427321 0.695360 -0.376955 0.249929 -0.381626 -0.468536 -0.507787 0.333461 -0.737663 1.000000 |