Universidade de São Paulo
Instituto de Física de São Carlos
Mathematical-Computational Modeling

Linear Least Squares

Éverton Luís Mendes da Silva (10728171)

1 Introduction

In this project, three different data sets from "Scikit-Learn"(python library) will be covered. In each of them, a study on the "Linear Least Squares"method will be carried out.

2 Datasets

The data sets have samples with individual features about our reality. The first one is the "Boston house prices datasets", the second is "Breast cancer wisconsin (diagnostic)", and for last "California housing dataset".

7.2.1. Boston house prices dataset

Data Set Characteristics:

Number of Instances:	506
Number of Attributes:	13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.
Attribute Information (in order):	• 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 to 1940 • DIS weighted distances to five Boston employment centres • 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 • MEDV Median value of owner-occupied homes in $1000's

Figure 1

7.2.7. Breast cancer wisconsin (diagnostic) dataset

Data Set Characteristics:

Number of Instances:	569
Number of Attributes:	30 numeric, predictive attributes and the class
Attribute Information:	• radius (mean of distances from center to points on the perimeter) • texture (standard deviation of gray-scale values) • perimeter • area • smoothness (local variation in radius lengths) • compactness (perimeter^2 / area - 1.0) • concavity (severity of concave portions of the contour) • concave points (number of concave portions of the contour) • symmetry • fractal dimension ("coastline approximation" - 1)

Figure 2

7.3.7. California Housing dataset

Data Set Characteristics:

Number of Instances:	20640
Number of Attributes:	8 numeric, predictive attributes and the target
Attribute Information:	• MedInc median income in block • HouseAge median house age in block • AveRooms average number of rooms • AveBedrms average number of bedrooms • Population block population • AveOccup average house occupancy • Latitude house block latitude • Longitude house block longitude

Figure 3

3 Linear Least Squares(LLS)

The intuitive method used is based on systems of linear equations. These parameters can be described as vector expressions.

$$\vec{p} = U\vec{y} \tag{1}$$

$$U = (A^T A)^{-1} A^T \tag{2}$$

Where 'p' is a column matrix of the coefficients of a polynomial and 'y' is a column matrix of the data(same thing to 'A').

The polynomial equations below were used to minimize data.

- linear

$$y = a_0 + a_1 x \tag{3}$$

- quadratic

$$y = a_0 + a_1 x + a_2 x^2 \tag{4}$$

- cubic

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \tag{5}$$

- Plane Surface

$$f(x, y) = a_0 + a_1 x + a_2 y \tag{6}$$

- Quadratic Surface

$$f(x, y) = a_0 + a_1 x + a_2 y + a_3 xy + a_4 x^2 + a_5 y^2 \tag{7}$$

4 Boston

4.1 Line or Curves

The first image is a two-dimensional Plot, which points are the Data and the Line a LLS('linear').

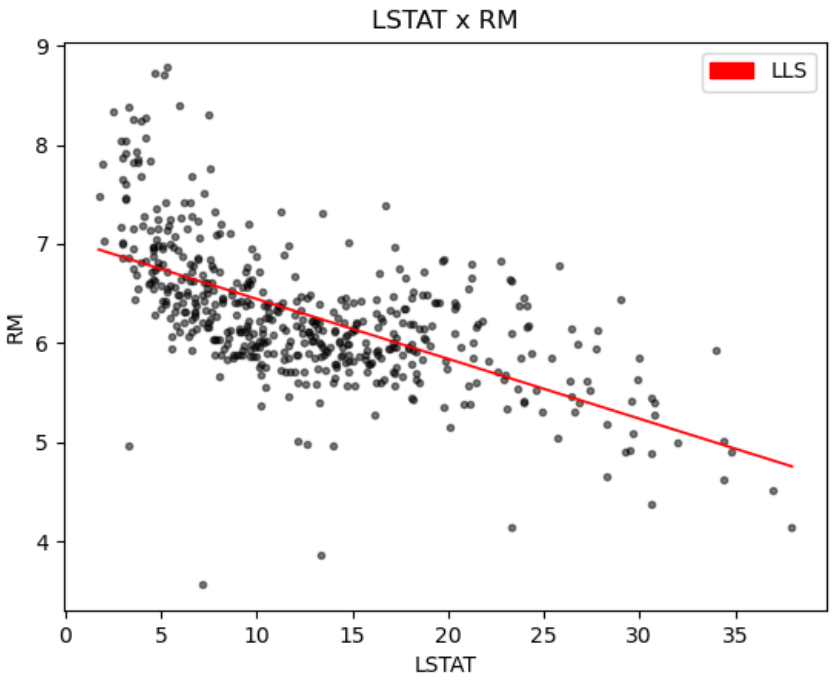

Figure 4 a0= 7.048 a1= -0.060

The next two images are with LLS ('quadratic') and LLS ('cubic'), respectively.

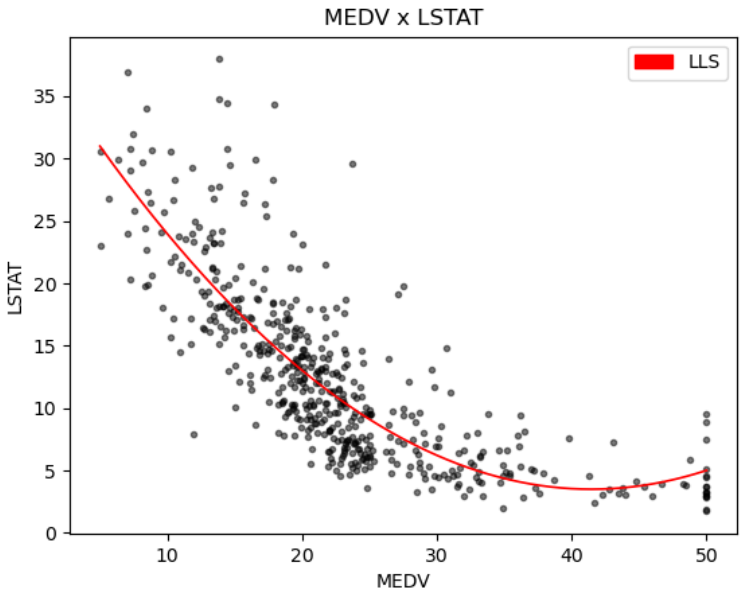

Figure 5 a0= 3.904e+01, a1= -1.715e+00, a2= 2.068e-02

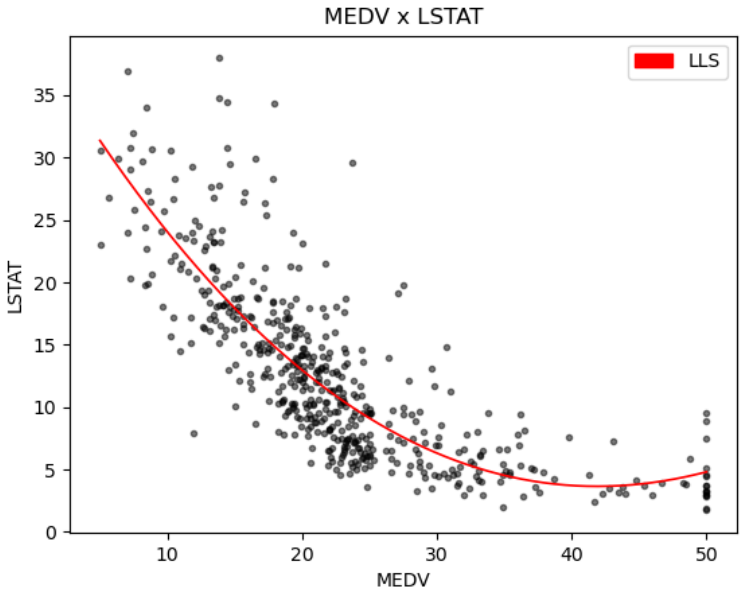

Figure 6 a0= 3.992e+01, a2= -1.840e+00 a3= 2.586e-02 a4=-6.189e-05

4.2 Surfaces

The two next images are a three-dimensional Plot, which points are in DataBoston and the Surface a LLS ('Plane Surface').

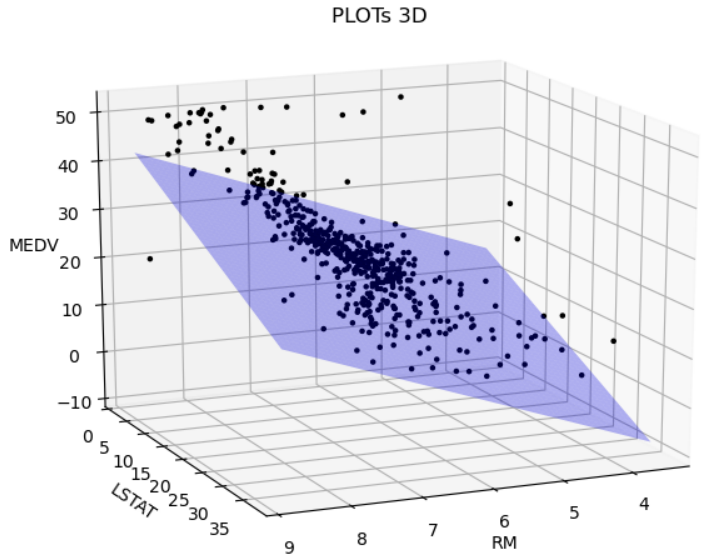

Figure 7 a0= -1.358, a2= 5.094, a3= -0.642

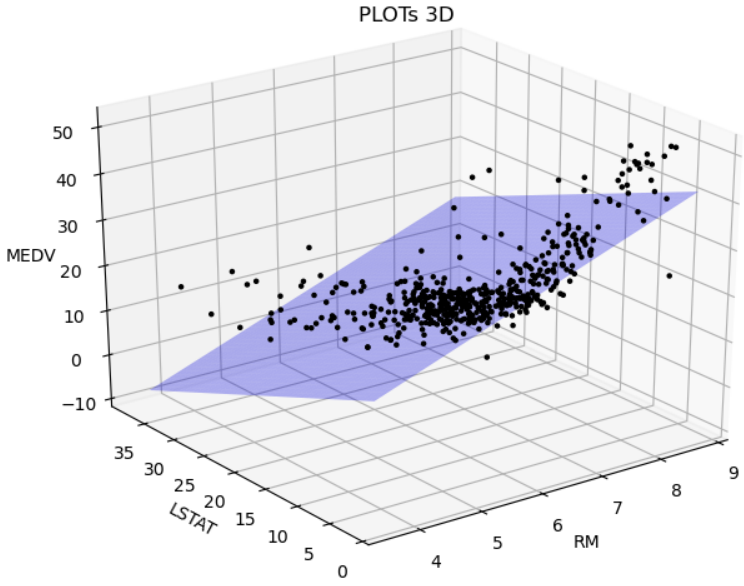

Figure 8 a0= -1.358, a1= 5.094, a2= -0.642

The two next images are a three-dimensional Plot, which points are the Data and the Surface a LLS ('Quadratic Surface').

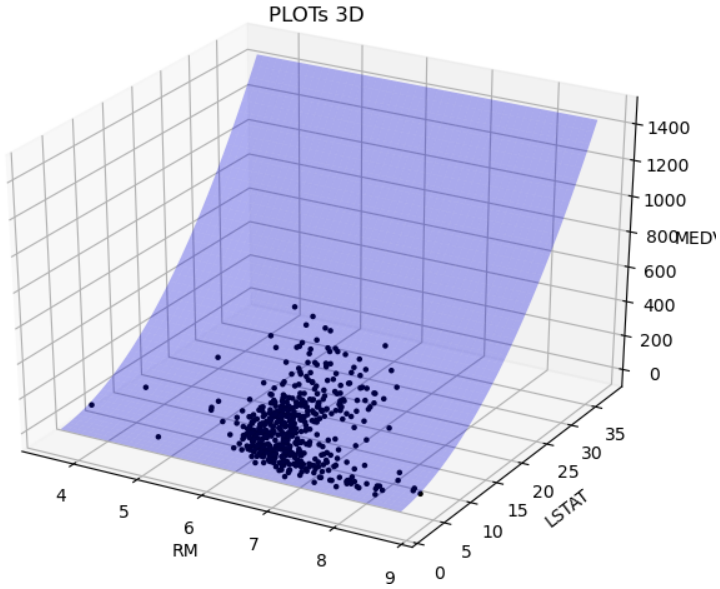

Figure 9 a0= 2.17487139e-12, a2= 3.73244491e-12, a3=1.000e+00, a4= 1.243e-14, a5= -5.480e-13, a6=3.995e-15

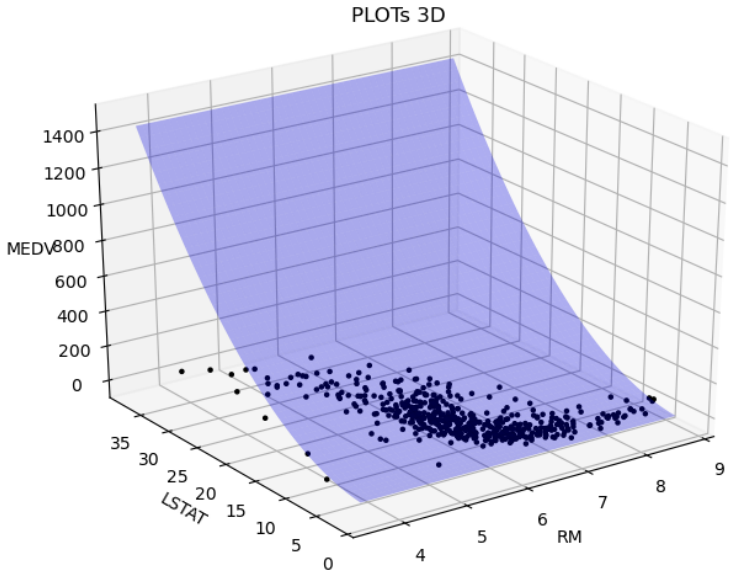

Figure 10 a0= 2.17487139e-12, a2= 3.73244491e-12, a3=1.000e+00, a4= 1.243e-14, a5= -5.480e-13, a6=3.995e-15

5 Cancer

5.1 Line or Curves

The first image is a two-dimensional Plot, which points are the Data and the Line a LLS('linear').

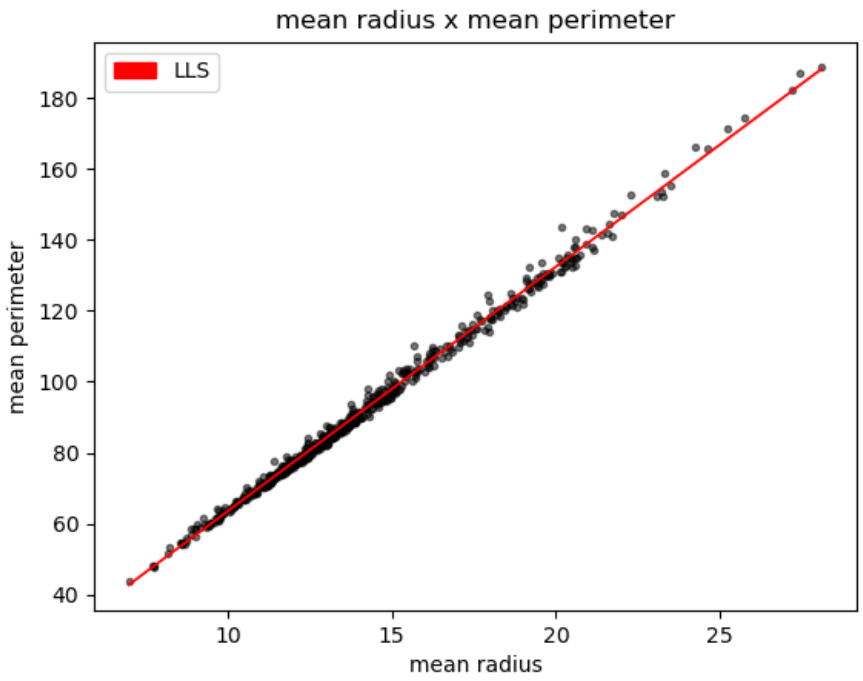

Figure 11 a0=-5.232, a1= 6.880

The next two images are with LLS ('quadratic') and LLS ('cubic'), respectively.

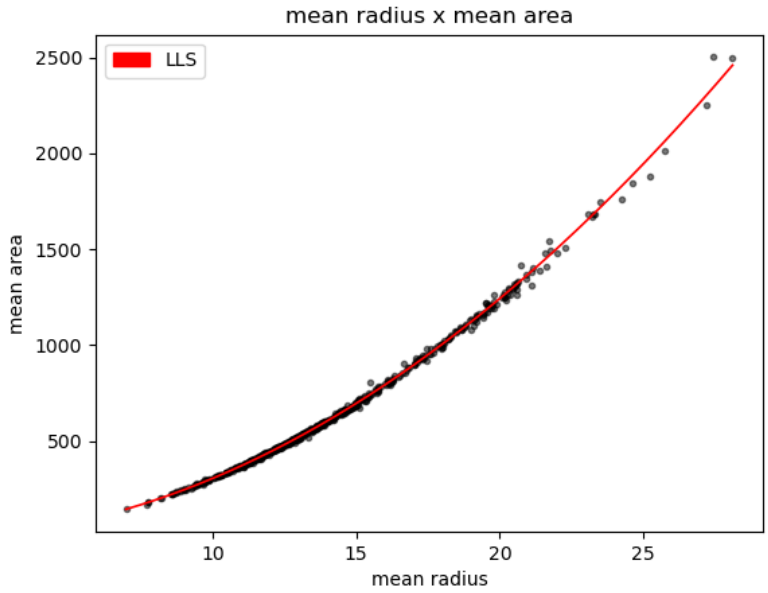

Figure 12 a0=-10.516, a1=0.436, a2=3.109

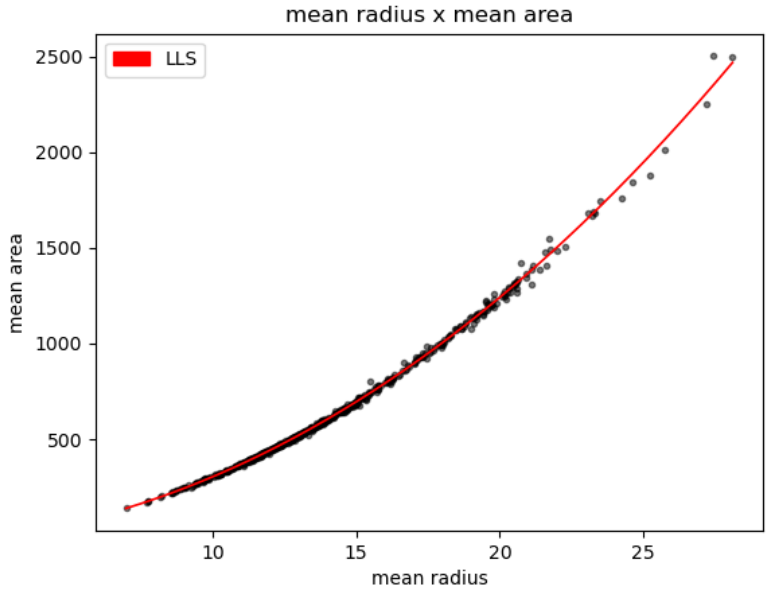

Figure 13 a0=-4.373e+01, a1=7.0653e+00, a2=2.690e+00, a3=8.390e-03

5.2 Surfaces

The two next images are a three-dimensional Plot, which points are in DataBoston and the Surface a LLS ('Plane Surface').

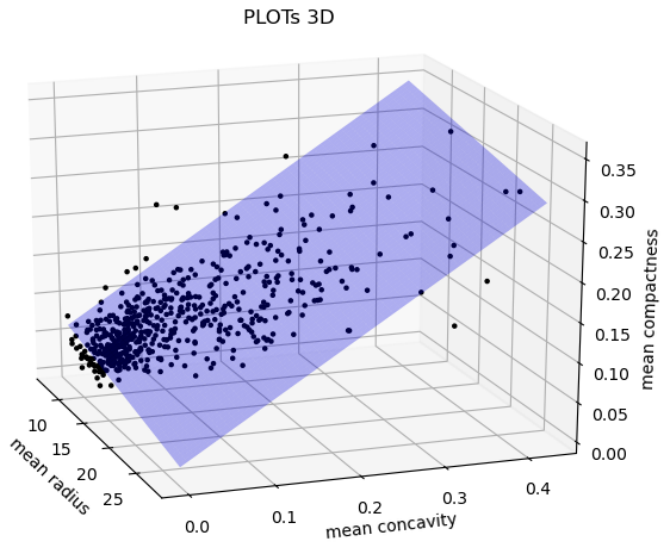

Figure 14 a0= 0.081, a1=-0.002, a2=0.660

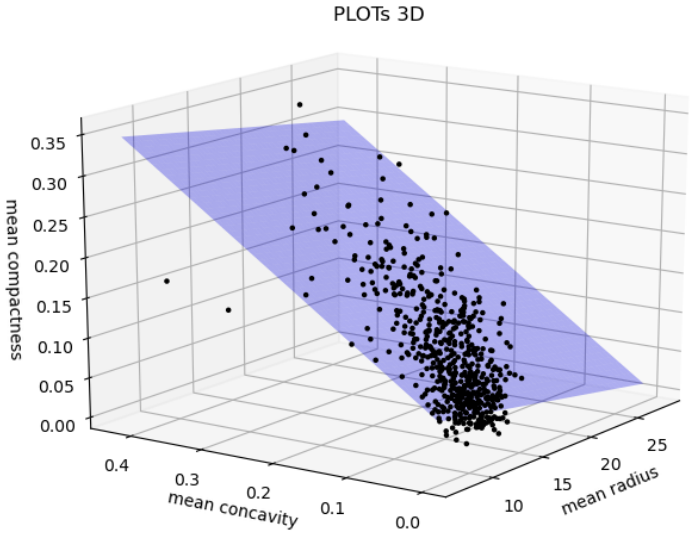

Figure 15 a0= 0.081, a1=-0.002, a2=0.660

The two next images are a three-dimensional Plot, which points are in DataBoston and the Surface a LLS ('Quadratic Surface').

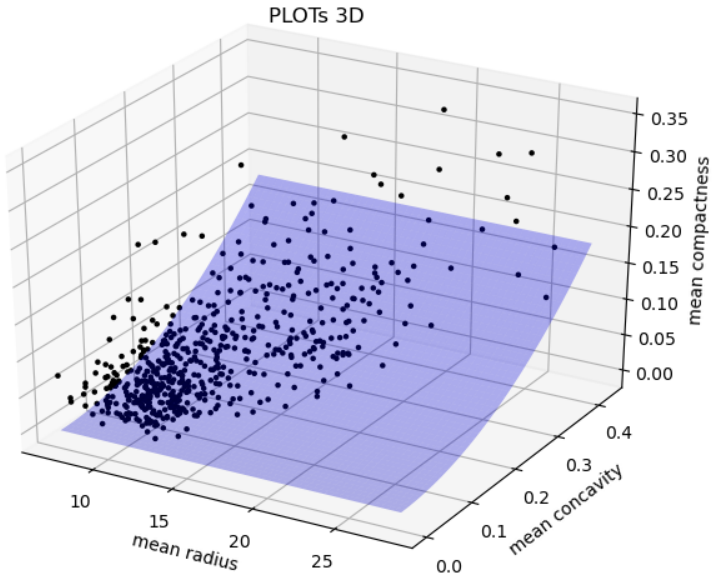

Figure 16 a0=-1.852e-14, a1=2.1776e-15, a2=1.000e+00, a3=-1.996e-15, a4=-6.041e-17, a5=9.631e-15

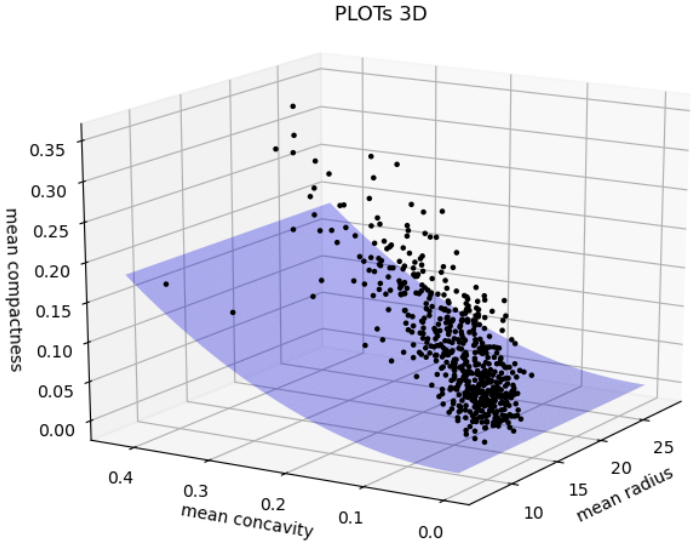

Figure 17 a0=-1.852e-14, a1=2.1776e-15, a2=1.000e+00, a3=-1.996e-15, a4=-6.041e-17, a5=9.631e-15

6 California

6.1 Line or Curves

The first image is a two-dimensional Plot, which points are the Data and the Line a LLS('linear').

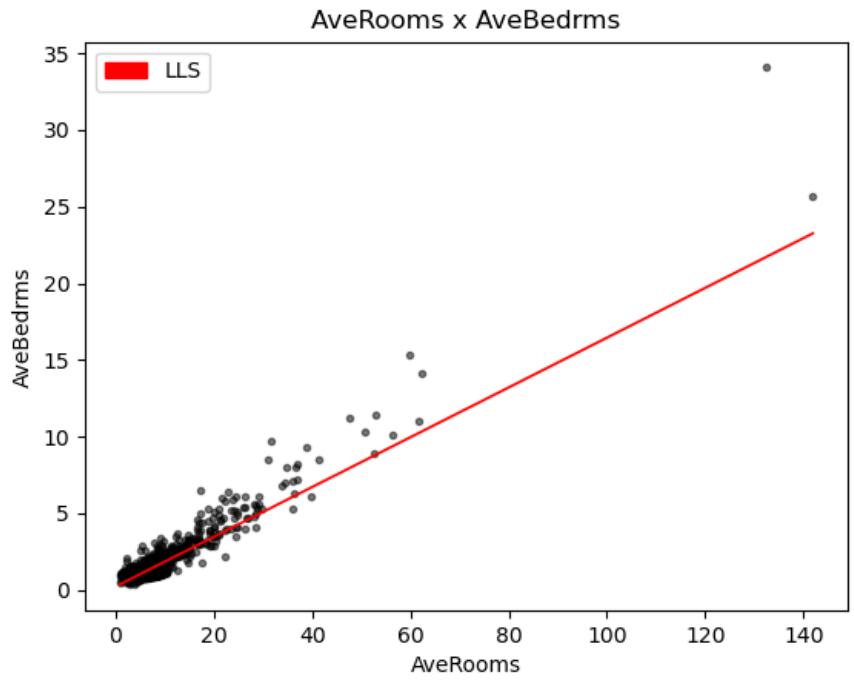

Figure 18 a0= 0.215, a1=0.162

The next two images are with LLS ('quadratic') and LLS ('cubic'), respectively.

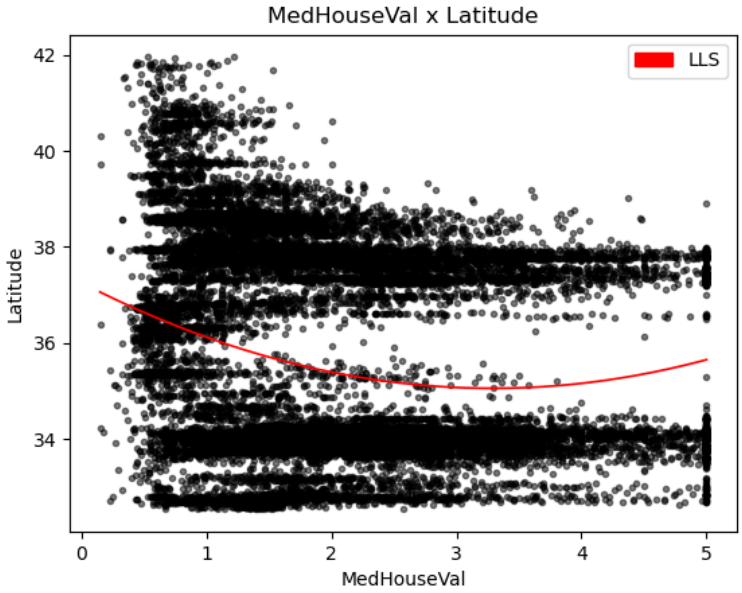

Figure 19 a0=37.258, a1=-1.337, a2=0.203

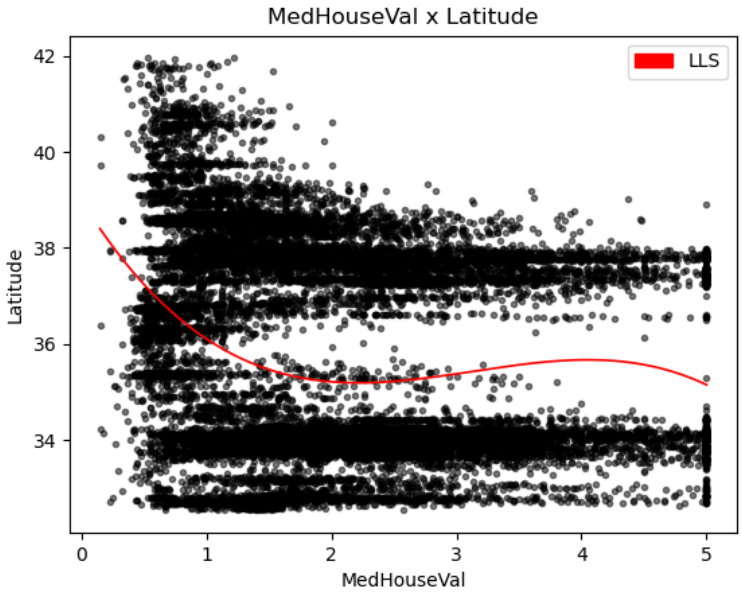

Figure 20 a0=38.998, a1=-4.197, a2=1.464, a3=-0.155

6.2 Surfaces

The two next images are a three-dimensional Plot, which points are in DataBoston and the Surface a LLS ('Plane Surface').

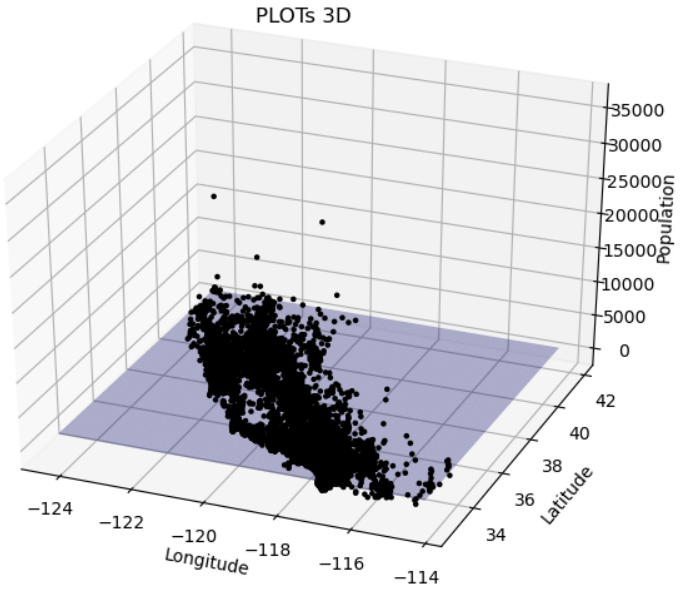

Figure 21 a0=3.198e+03, a1= -3.181e+00, a2=-6.043e+01

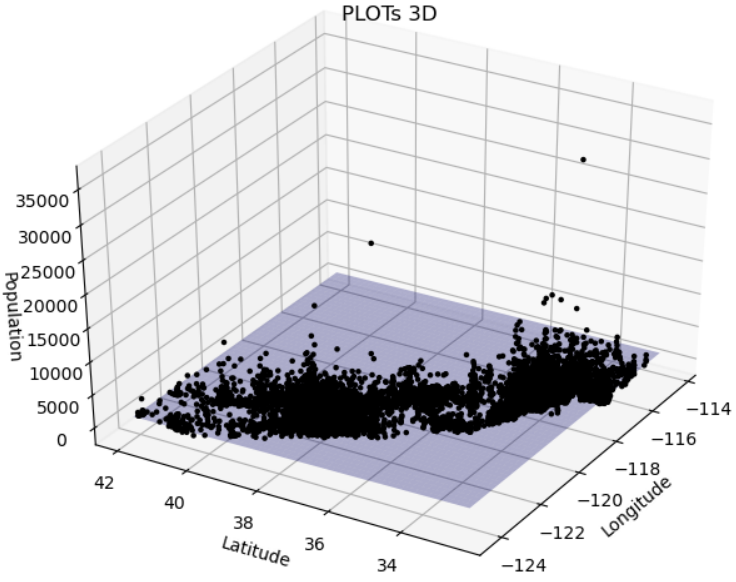

Figure 22 a0=3.198e+03, a1= -3.181e+00, a2=-6.043e+01

The two next images are a three-dimensional Plot, which points are in DataBoston and the Surface a LLS ('Quadratic Surface').

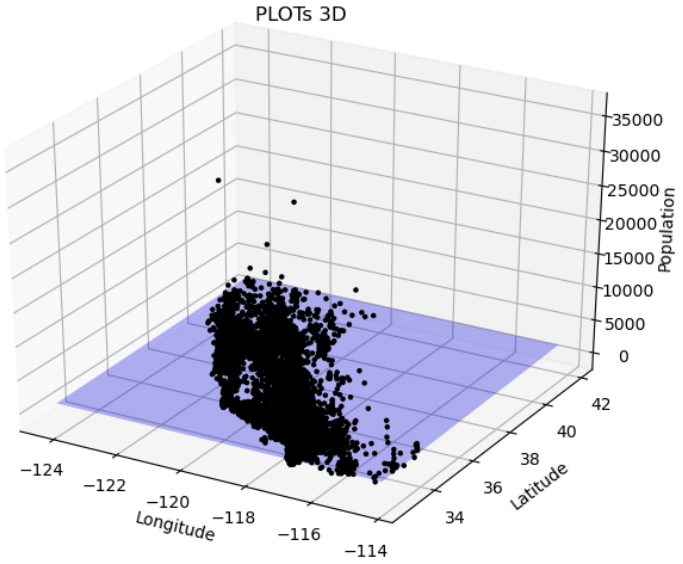

Figure 23 a1=3.161e-05, a2=6.808e-07, a3=1.000e+00, a4= 5.526e-09, a5=3.653e-09, a6=2.067e-09

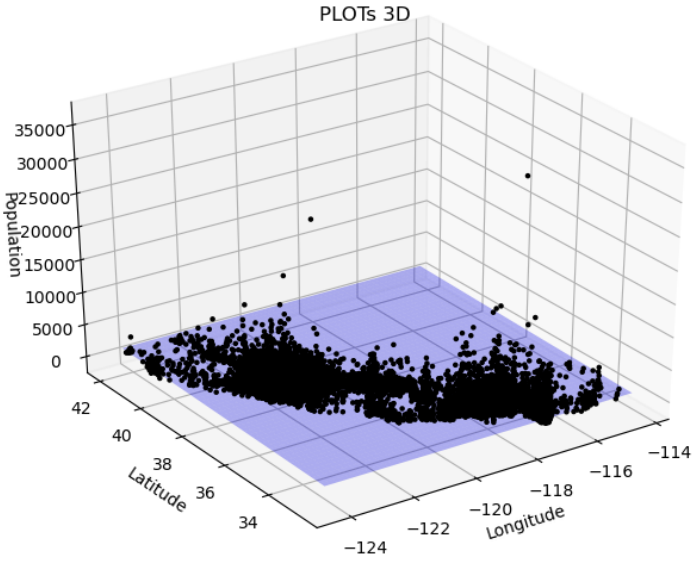

Figure 24 a1=3.161e-05, a2=6.808e-07, a3=1.000e+00, a4= 5.526e-09, a5=3.653e-09, a6=2.067e-09

Lastly, the change of the opacity of the points allow us to see the clustering of them.

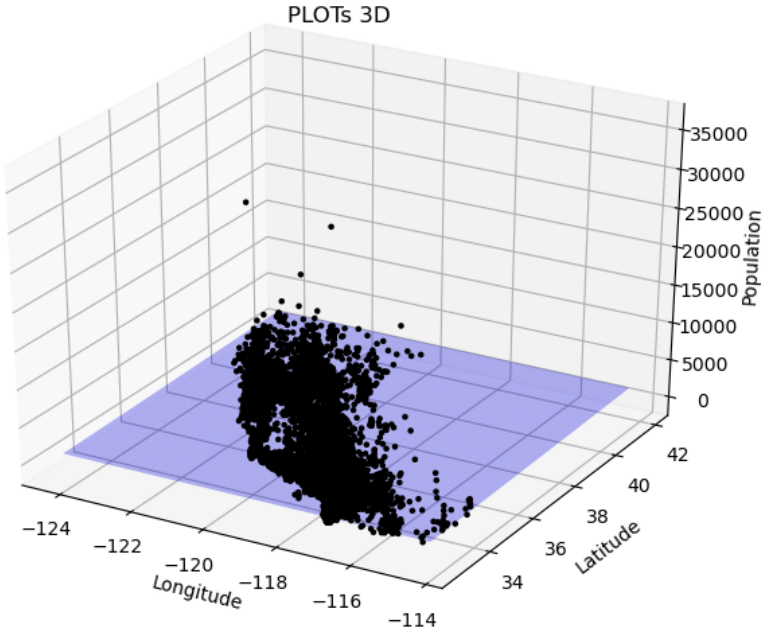

Figure 25: alpha=1

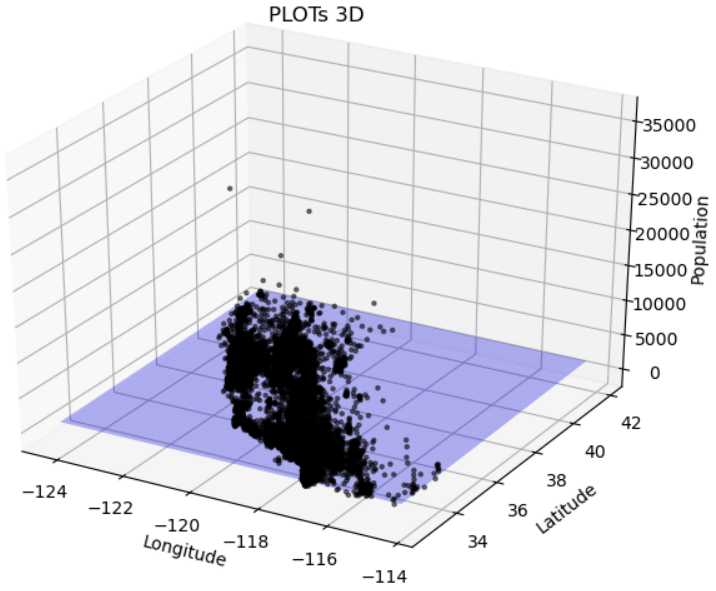

Figure 26: alpha=0.5

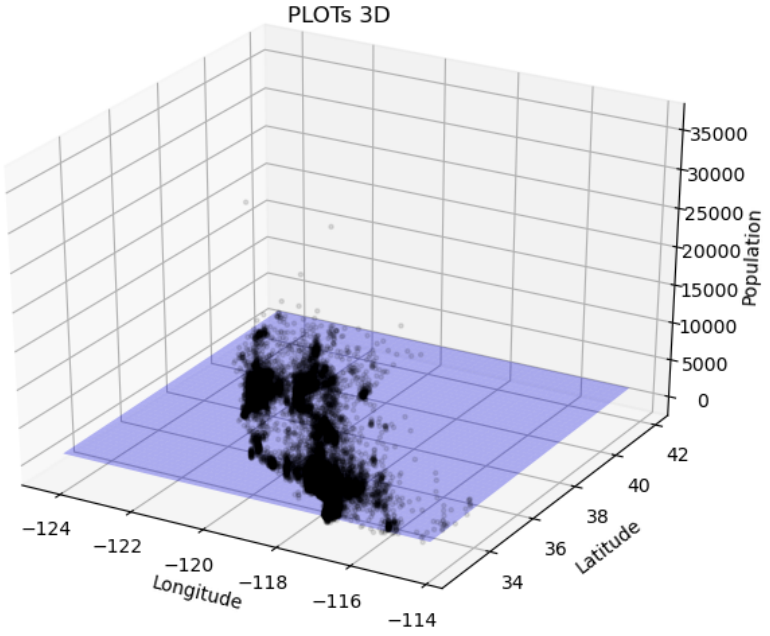

Figure 27: alpha=0.1

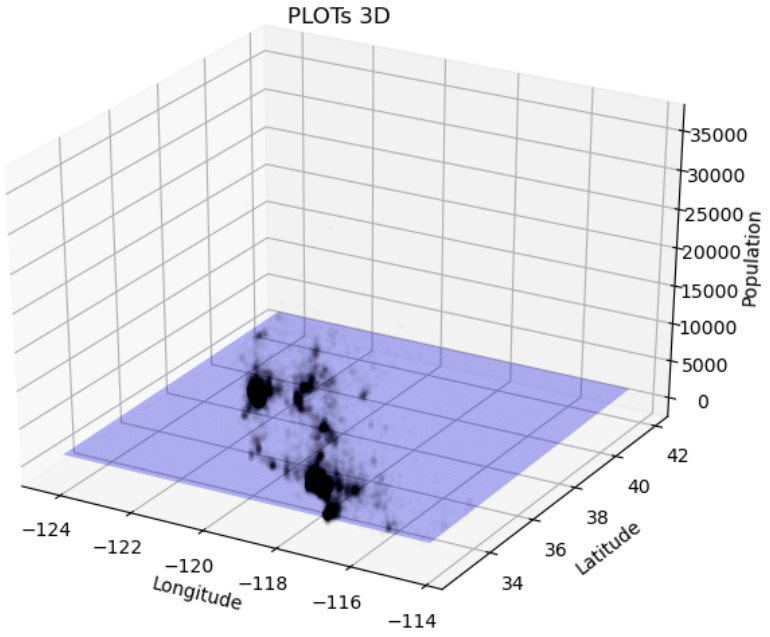

Figure 28: alpha = 0.01

7 Referências

[1] Figures 1, 2, 3 of the Data sets come from:
https://scikit-learn.org/stable/index.html

[2] da Silva, Éverton Luís Mendes. Project Codes in 'Project1' in Repository 'Mathematical-Computational Modeling' :
https://github.com/evertonmendes/Mathematical-Computational-Modeling