

DATA MINING

FINAL PROJECT

Mustafa KURÇEREN



NOT: …………………..….

**ANKARA MEDİPOL ÜNİVERSİTESİ**

**BİLGİSAYAR PROGRAMCILIĞI BÖLÜMÜ**

VERİ MADENCİLİĞİ DERSİ ARA SINAV ÖDEVİ

**ADI SOYADI:……………………………………………… NO:……………………………**

On the “datamining.csv” data set which is given,there will be done changes about the necessary cleanings and data types.(every

process will be told);

1. Summary analysis of data set will be done and this summary screen image will be shown on the project,
2. Correlation will be calculated by Kendall,Spearman and Pearson techniques,

* Screen matrix image of correlation (result of correlation) will be written exactly.
* On the datas which have meaningful relationships on the matrix, results will be evaluated and will be posted in the

comment section.

For the every dual data type which is meaningful on the correlation;

1. Linear Regression slope-intercept values will be determined,
   * Graphic will be done as Plotting
   * f(x) function will be found

1. Linear Regression method will be applied,
   * train-test data will be calculated separately
   * Plotting Graphic will be created
   * Linear R2 value will be calculated

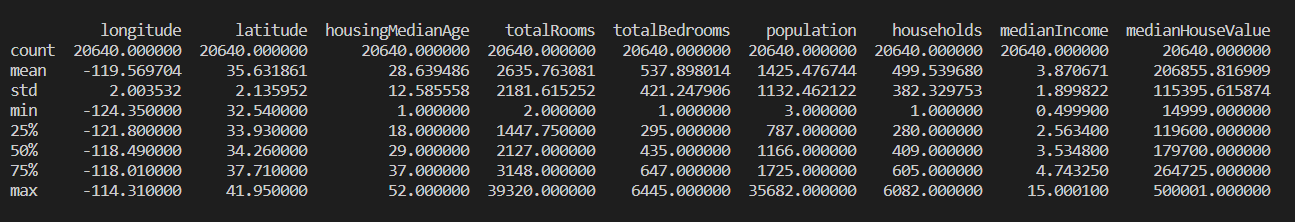
* All results will be evaluated and written in the comment section

*Başarılar dilerim.*

On the “datamining.csv” data set which is given,there will be done changes about the necessary cleanings and data types.(every

process will be told);

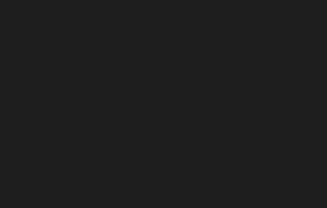
1. Summary analysis of data set will be done and this summary screen image will be shown on the project,

 ANALYZE USING DESCRIPTION

**metin, levha, plaka içeren bir resim

Açıklama otomatik olarak oluşturuldu** ANALYZE USING STATISTICS

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu****metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

1. Correlation will be calculated by Kendall,Spearman and Pearson techniques,

* Screen matrix image of correlation (result of correlation) will be written exactly.
* On the datas which have meaningful relationships on the matrix, results will be evaluated and will be posted in the

comment section.

Making Correlation with Only Corr()

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Making Correlation with Kendall

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Making Correlation with Pearson

**masa içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Making Correlation with Spearman

metin, bilgisayar içeren bir resim

Açıklama otomatik olarak oluşturuldu

For the every dual data type which is meaningful on the correlation;

3-) Linear Regression slope-intercept values will be determined

* + Graphic will be done as Plotting
  + f(x) function will be found

Linear Regression Slope-Itercept Values

**metin içeren bir resim

Açıklama otomatik olarak oluşturuldu**

Text version:

longitude latitude slope intercept value: [ -0.98577886 -82.23742528]

totalRooms totalBedrooms slope intercept value: [ 0.17955303 64.6387697 ]

totalRooms population slope intercept value: [ 0.44492845 252.75075521]

totalRooms households slope intercept value: [ 0.16096511 75.27377369]

totalBedrooms population slope intercept value: [ 2.36044072 155.80036918]

totalBedrooms households slope intercept value: [ 0.88930475 21.18442068]

population households slope intercept value: [ 0.30628668 62.93513925]

medianIncome medianHouseValue slope intercept value: [41793.8492019 45085.57670327]

f(x) Function Equations

longitude-latitude Slope Intercept: f(x)=(-0.98577886 x) + ( -82.23742528)

totalRooms-totalBedrooms Slope Intercept: f(x)=( 0.17955303 x) + ( 64.6387697)

totalRooms-population Slope Intercept: f(x)=( 0.44492845 x) + ( 252.75075521)

totalRooms-households Slope Intercept: f(x)=( 0.16096511 x) + ( 75.27377369)

totalBedrooms population slope Intercept: f(x)=( 2.36044072 x) + (155.80036918)

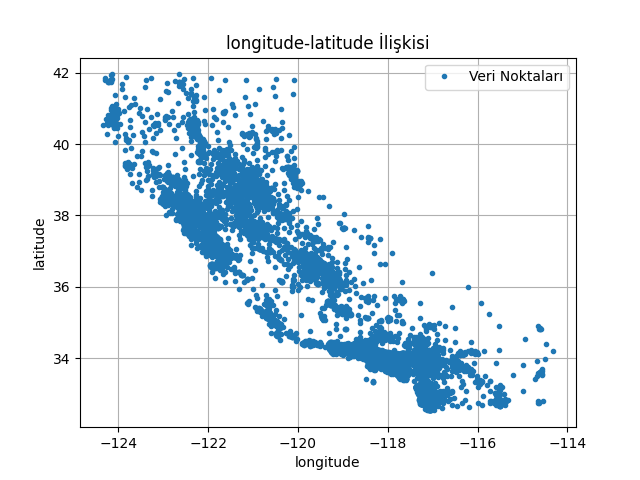
totalBedrooms households slope Intercept f(x)=(2.36044072 x ) +(155.80036918)

population households Slope Intercept : f(X)= (0.30628668 x ) + (62.93513925)

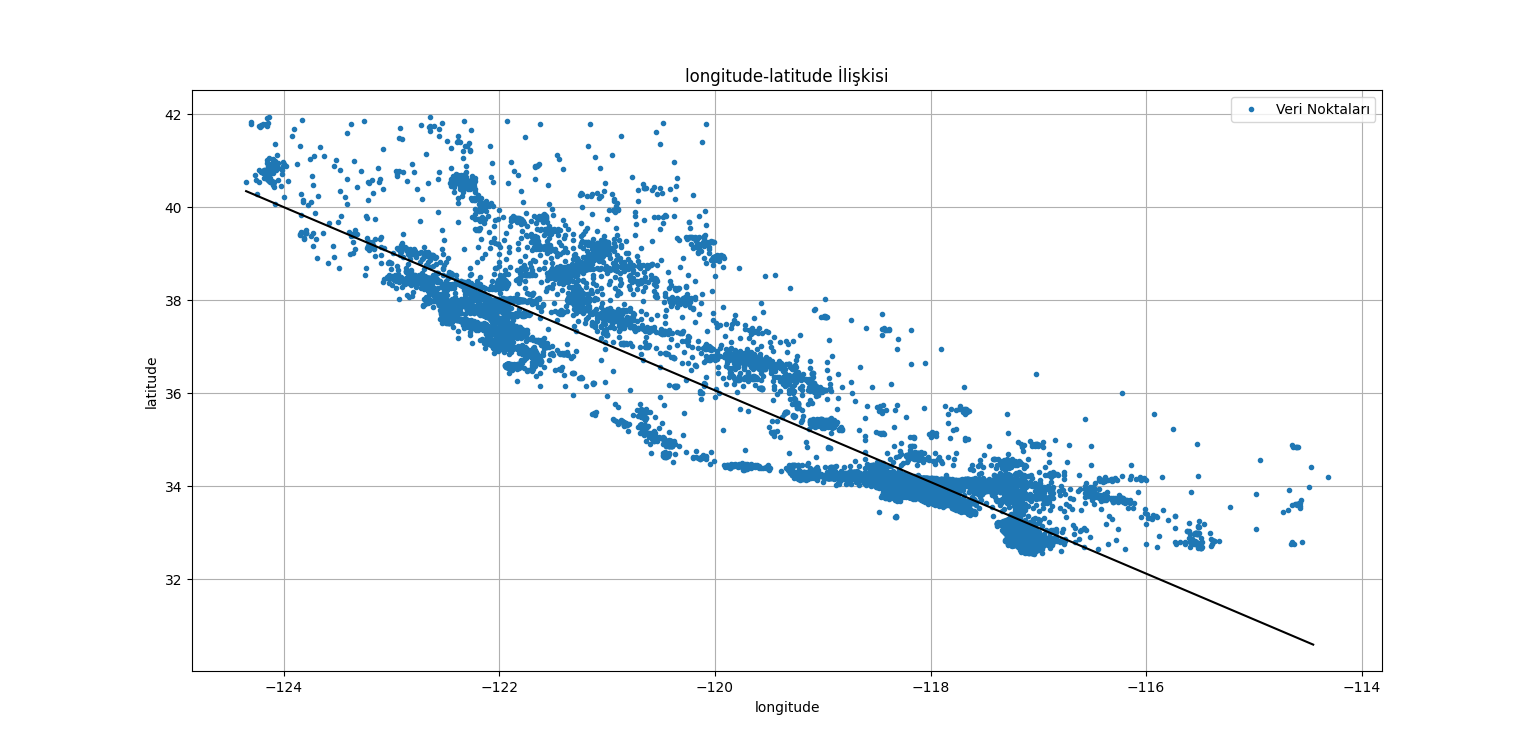
medianIncome-medianHouseValue Slope Interce: f(x)=( 41793.8492019 x) + ( 45085.57670327)

**Plotting Graphic**

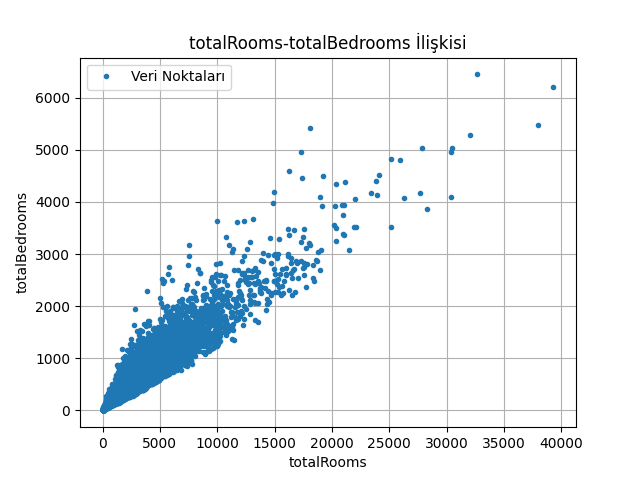
a)“longitude” and “latitude” Graphic

****

a.1)“longitude” and “latitude” Graphic and Line Formed with

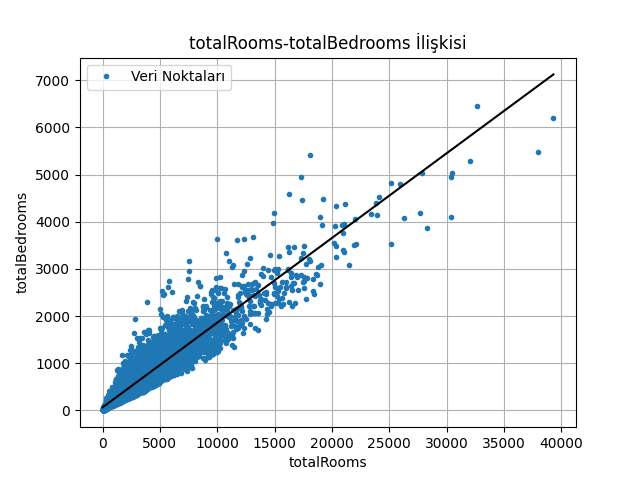
**** Slope-Itercept

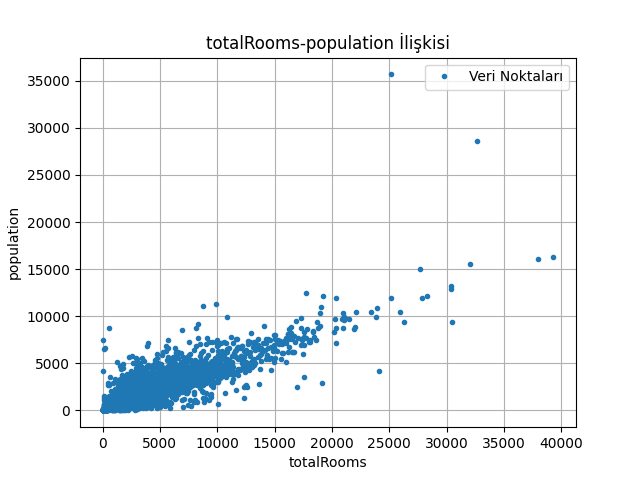
b)“totalRooms” and “totalBedrooms” Graphic



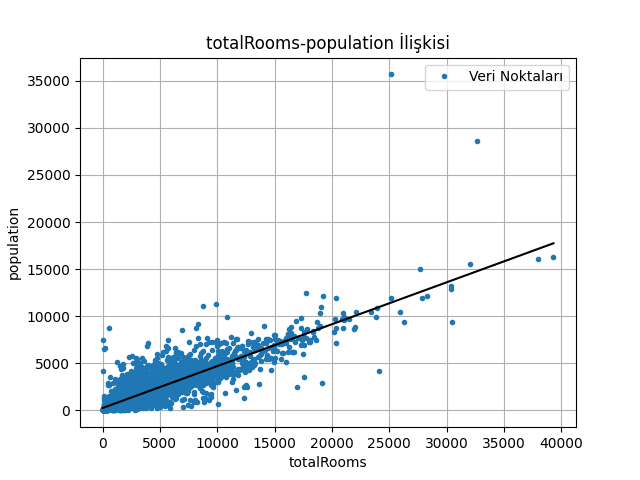
b.1)“totalRooms” and “totalBedrooms” Graphic and Line Formed with

Slope-Itercept

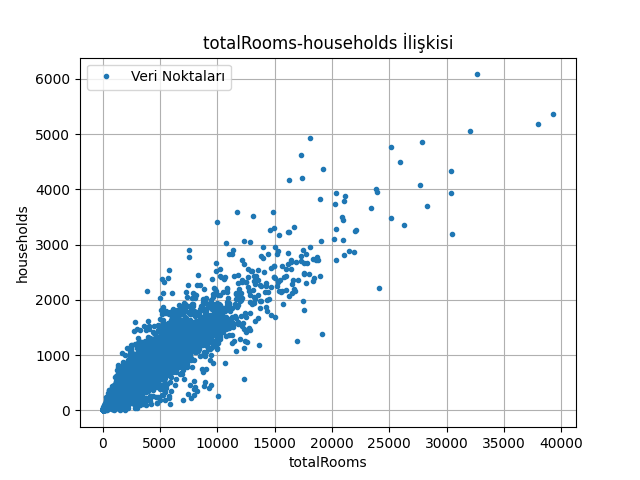


c-)”totalRooms” and “population” Graphic

c.1)“totalRooms” and “population” Graphic and

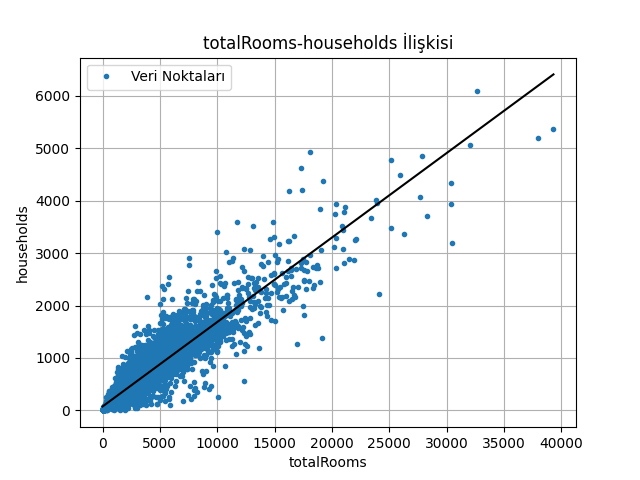
Line Formed with Slope-Itercept

d)“totalRooms” and “households” Graphic

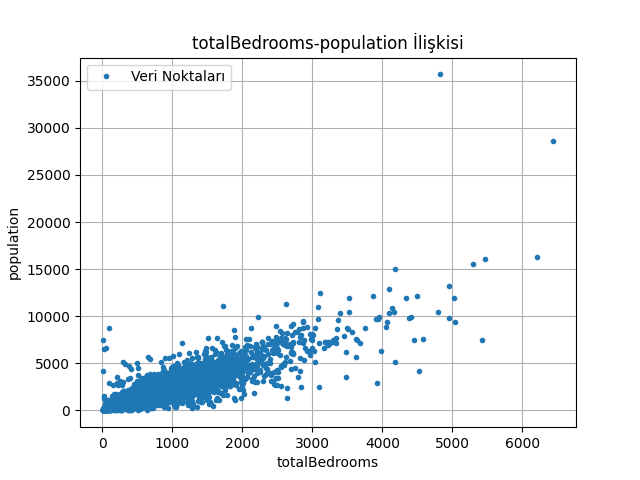


d.1)“totalRooms” and “households” Graphic and Line Formed

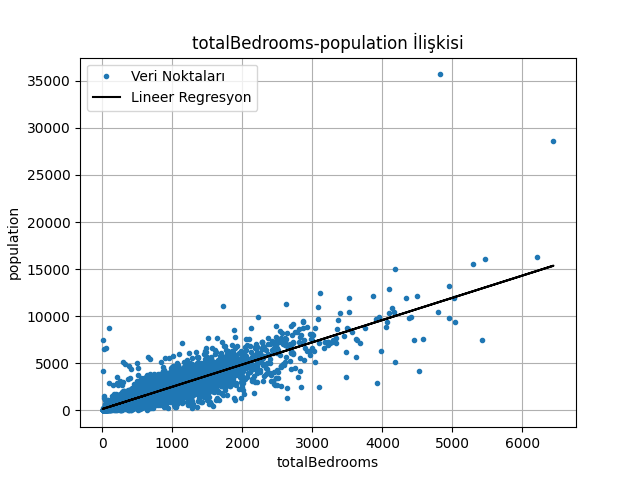
with Slope-Itercept



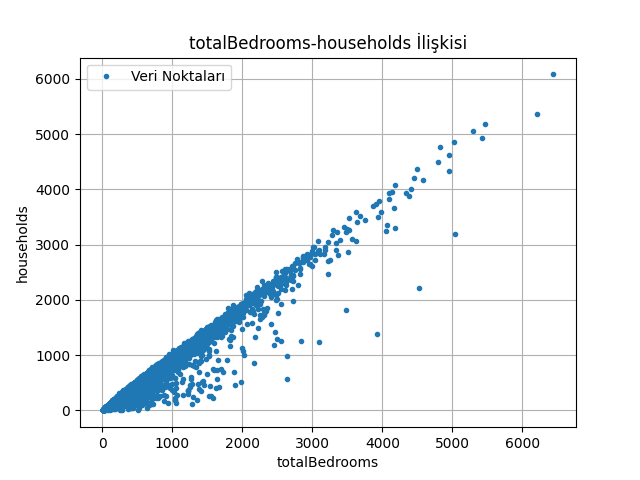
e)“totalBedrooms” and “population” Graphic



e.1)“totalBedrooms” and “population” Graphic and Line

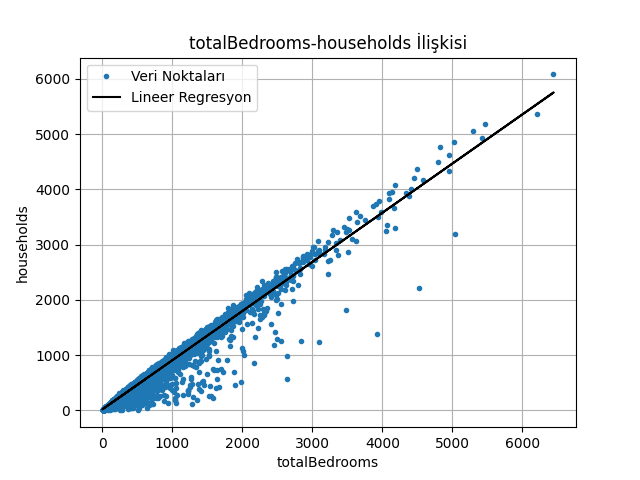
Formed with Slope-Itercept

f)“totalBedrooms” and “households” Graphic



f.1)“totalBedrooms” and “households” Graphic and Line Formed

with Slope-Itercept



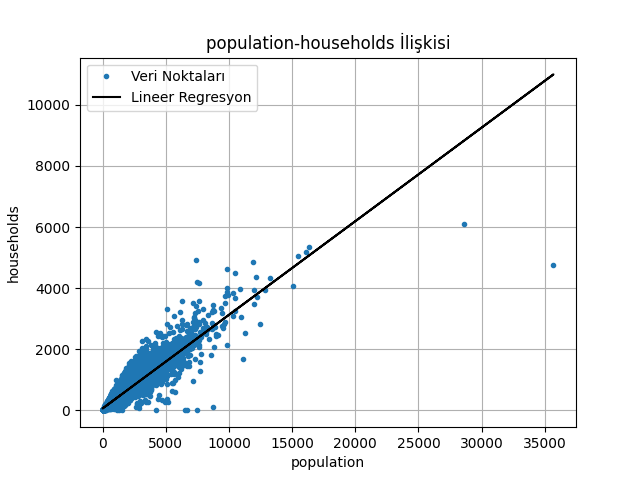
g)“ population” and “households” Graphic

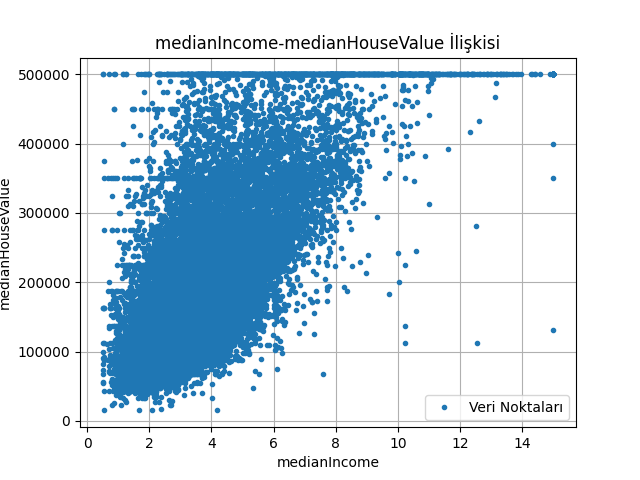
çizelge içeren bir resim

Açıklama otomatik olarak oluşturuldu

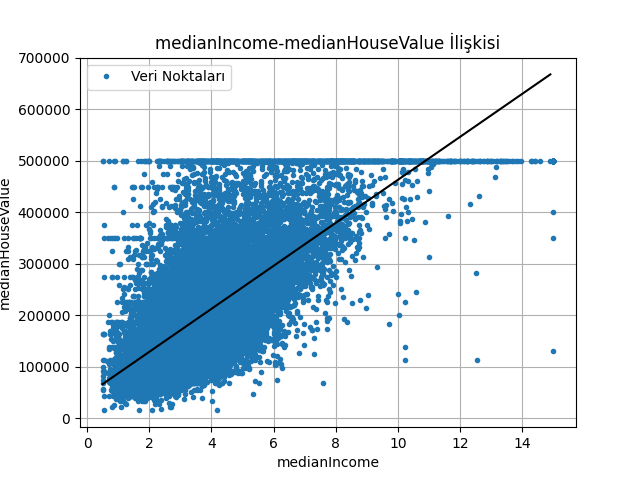
g.1)“ population” and “households” Graphic and Line

Formed with Slope-Itercept



h)“medianIncome” and “medianHouseValue” Graphic

h.1)“medianIncome” and “medianHouseValue” Graphic and Line

Formed with Slope-Itercept 

1. Linear Regression method will be applied,
   1. train-test data will be calculated separately
   2. Plotting Graphic will be created
   3. Linear R2 value will be calculated

* All results will be evaluated and written in the comment section

a)“longitude” and “latitude”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

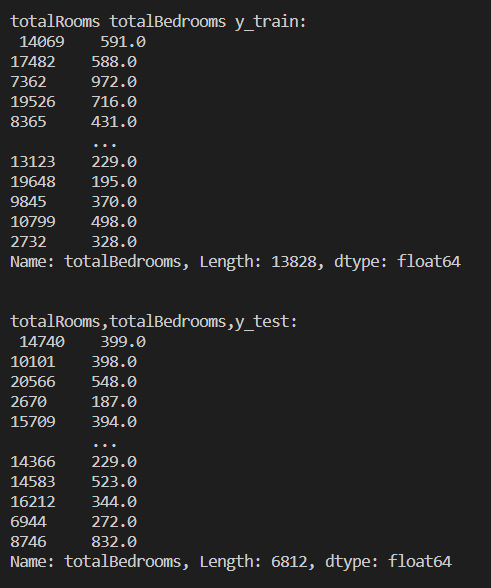
Açıklama otomatik olarak oluşturuldu

çizelge içeren bir resim

Açıklama otomatik olarak oluşturulduPlotting Graphic

b)“totalRooms” and “totalBedrooms”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

metin içeren bir resim

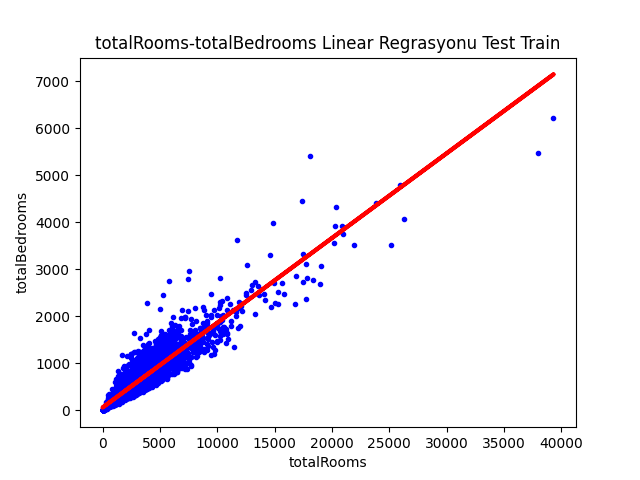
Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic



c)“totalRooms” and “population”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

metin içeren bir resim

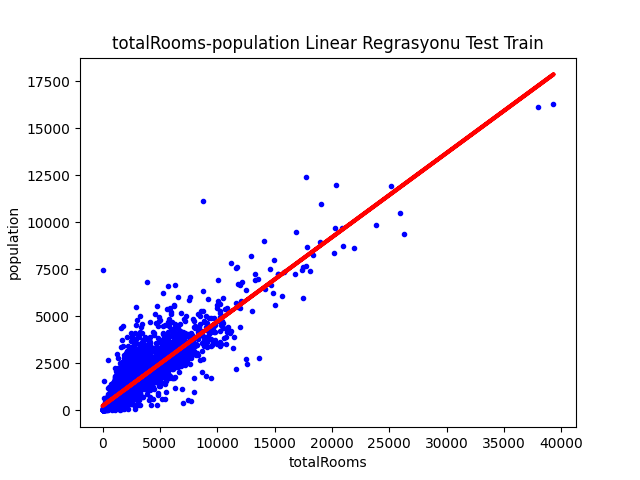
Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic



d)“totalRooms” and “households”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

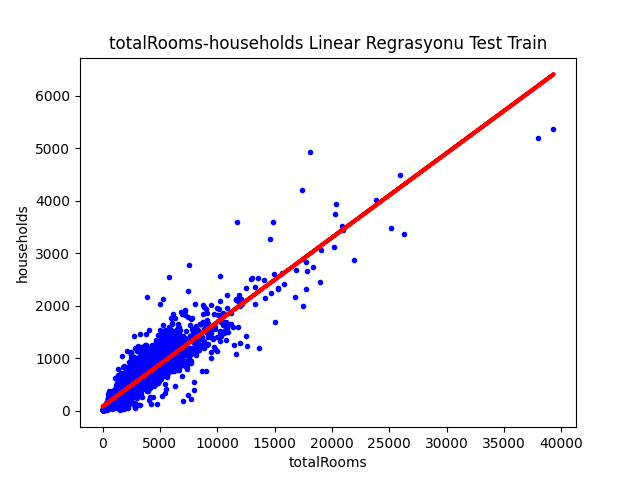
metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic

e)“totalBedrooms” and “population”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

metin içeren bir resim

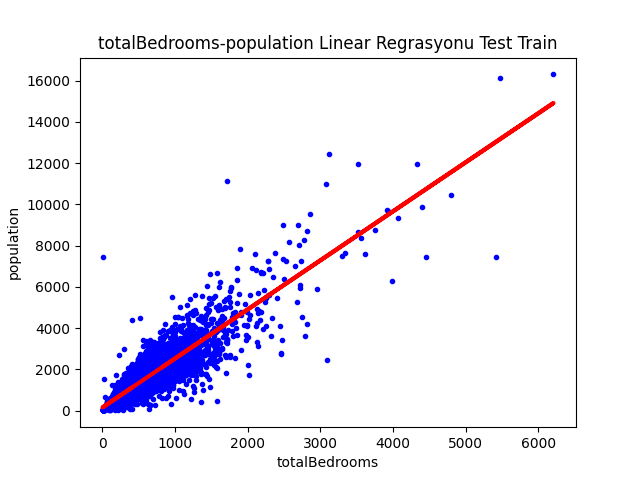
Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic



f)“totalBedrooms” and “households”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldumetin içeren bir resim

Açıklama otomatik olarak oluşturuldu

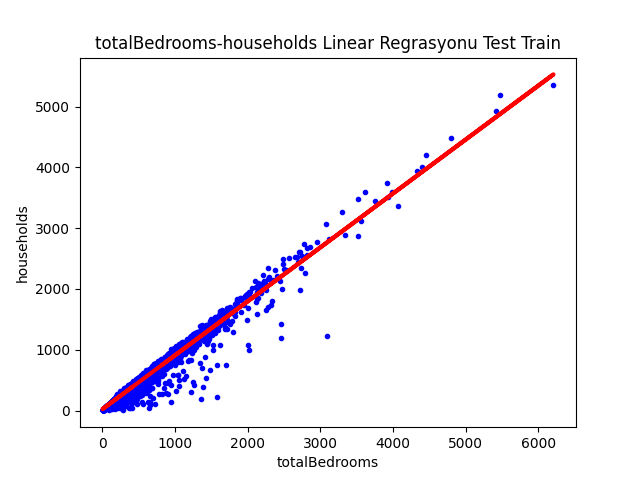
Linear R2 Value



Predict Value



Plotting Graphic



g)“ population” and “households”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value



Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic



e)“medianIncome” and “medianHouseValue”

Train-Test Data Terminal Image

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Linear R2 Value

metin içeren bir resim

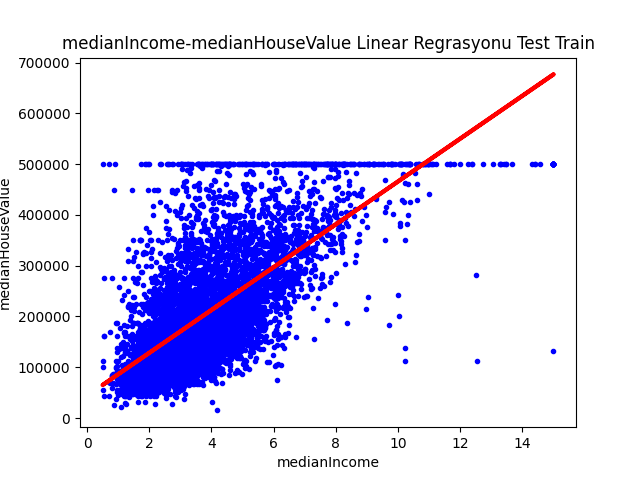
Açıklama otomatik olarak oluşturuldu

Predict Value

metin içeren bir resim

Açıklama otomatik olarak oluşturuldu

Plotting Graphic



Comments

**Cleaning Process on the Data Set, Actions and Comments**

1.question: On the “datamining.csv” data set which is given,there will be done changes about the necessary cleanings and data types (every process will be told);

I used pandas library for necessary cleaning processes then I had my cvs file read with read.cvs.Firstly before taking any process,I learned how many columns and rows I have with rowcount=df.shape[0] and columcount= df.shape[1]. So I can learn how much I need my data for processing.After these processes,firstly I started to cleaning process with dropna(inplace=True) command I deleted rows which are empty then I saw any rows

have changed and that means in all my data has any empty data.After that I wanted to clean the same rows

with drop\_dublicates(inplace=True) command then I checked my row count again and I saw there is any

decrease in my row count and that means there is any data which repeating itself.So I passed my last process.I checked data types with my Df.info() command and I saw that they were all float64.I didn’t need any change in here because my numerical datas are in the data type it should be so I can start my data analysis.After these processes I understood that I can start my analysis without any data loss and it will affect truth of my analysis in

a good way.

**Data Set Summary Analysis Comment**

2.question: Summary analysis of data set will be done and this summary screen image will be shown on the project;

Here I saw the rows’ average with mean command and this way I can access a few result.Firstly I can talk

about latitude and longitude.When I searched the average value and max min value on the google maps,I

learned that my data is about the houses in ABD California City.And when I looked at to max and min value I

saw that location of houses is mostly on the east west line.Lastly when I saw the standard deviation is 2 I

learned that these data addresses are in neighborhoods very close to eachother.

HousingMedianAge:I learned that buildings are 28 years old on average with analysis about HousingMedianAge column.When I see the standard deviation is 12 it means there is a big difference between buildings’ age and buildings are middle-aged but age distribution is wide.

TotalRooms:Here I can see the average is 2635 and it shows that there are regions with too many rooms and regions with lesser rooms and there is not much dispersion about that.There are some reasons about the

low diversity and one of the reasons is standard deviation is 2181 and the other reason is it is too close to average. There is a big difference between min and max values but there is not much value which far above or

far below the average.

TotalBedrooms:For the total bedrooms column,average number of bedrooms is 537 and standard deviation is 421 so it shows us there is not much number diversity.Here I can make this comment,min value of bedroom is 1 and there is no place without a bedroom.

Population:There are very crowded neighborhoods but 1424 people accommodate on average.Max

value is 35682 so we can say that there are rarely very big neighborhoods and it has an average population.

Households:There is 499 people in building on average and standard deviation is 382.Generally,buildings are built to accommodate 500 people.And there are also buildings where a minimum of 1 person lives.There

are max 6082 people in a building and these buildings are few.

Median Income: Average income of people is 3.870 dollars in a building.And this represents a moderate income level.There are also people who are rich and have high income level in that neighborhood but they are

in the minority because an average value is close to min value.

MedianHouseValue:It is about the value of buildings.You can buy a house for 206.855 dollar from California.House prices are too close to eachother because standard deviation is very close to average.Even for 14.000 dollar you can find a house but it is few.The most expensive house in here is 500.000 dollar.

**Correlation Analysis**

Correlation will be calculated by Kendall,Spearman and Pearson techniques;

* Screen matrix image of correlation (result of correlation) will be written exactly,
* On the datas which have meaningful relationships on the matrix, results will be evaluated and will be posted in the comment section.

**Strong Correlation:**

Value that we find for strong correlation depends on proximity distance to -1 and 1.Duals which are

above the 0.6 or -0.6 correlation values,we can say a strong correlation.

**“**longitude” X “latitude” = -0.924664 ,this is a really strong correlation and it shows there is an inverse proportion between the longitude and latitude.And that means if longitude increases also latitude increases.

“totalRooms” X “totalBedrooms”= 0.929893,this is a really strong correlation and it shows there is correct proportion between the total number of rooms and total number of bedrooms.And that means if total number

of rooms increases also total number of bedrooms increases.

“totalRooms” X “population” =0.857126, this is a really strong correlation and it shows there is a correct proportion between the total number of rooms and total population number of neighborhood around the building.And that means if total number of rooms increases also total population number of neighborhood

around the building increases.

“totalRooms” X “households”=0.918484 ,this is a really strong correlation and it shows there is correct

proportion between the total number of rooms and population of households.And that means if total number

of rooms increases also population of households increases.

“totalBedrooms” X “population”= 0.878026, this is really strong correlation and it shows there is correct proportion between the total number of bedrooms and population of neighborhood.And that means if total number of bedrooms increases also population of neighborhood increases.

“totalBedrooms” X “households”= 0.979829, it is almost too close to 1 so it is very strong correlation and

it shows there is a correct proportion between the total number of bedrooms and population of households.

And that means if total number of bedrooms increases also population of households increases

“population” X “households”= 0.907222, this is a really strong correlation and it shows there is a correct proportion between the population of neighborhood and population of households.And that means if

population of households increases also population of neighborhood increases.

“medianIncome” X “medianHouseValue”= 0.688075, this is a kinda strong correlation and it shows there

is a correct proportion between the income level of people who live in building and sale value of building.And

that means if income level of people who live in building decreases also sale value of building decreases.

**Weak Correlation :**

. Value that we find for weak correlation depends on proximity distance to -1 and 1.Duals which are below the

-0.6 or 0.6 correlation values ,we can say a weak correlation.

(Notice: **“**longitude”X**“**longitude”=1.0000 , these results are excluded and self-made correlations already result

in “1”.)

To give an example **“**medianIncome” X “population”= 0.004834, this is a value which is too far to 1 and that means there is no relationship between the income level of people who live in building and population number

of neighborhood around the building.

**Linear Regression Analysis Comment**

For the every dual data type which is meaningful on the correlation;

1. Linear Regression slope-intercept values will be determined,

* Graphic will be done as Plotting
* f(x) function will be found

I took duals’ slope\_intercept values which are related to eachother and greater than 0.6 from my correlation methods.And I created f(x) function with these values.That f(x) function provided me get a linear regression.I created binary graphs,which I found,as a plotting and I also created graphic by adding f(x) function graphic inside binary graphs.And graphic that I found was covering eachother perfectly except one dual (“medianlncome” and “medianHauseValue”).When I looked at the graphic I thought that they were

incompatible and already corr value wasn’t too close to 1.We’ll see better everything on the results of Test

Train graph.

1. Linear Regression method will be applied;

* 1. train-test data will be calculated separately
  2. Plotting Graphic will be created
  3. Linear R2 value will be calculated
* All results will be evaluated and written in the comment section

Here,I took part of the data of the binaries for train model and I tested it with some of them and I took images of graphs.Then I got results and I created plotting graph.The graphs I got was almost an exact match.

After the graph of dual ,which I suspected (“medianIncome and medianHouseValue”) ,I calculated Linear R2

value and results are:

longitude-latitude R2 value: 0.8530734992629649 that value is close to 1 and accuracy is good enough

totalRooms-totalBedrooms R2 value: 0.8565628206396894 that value is close to 1 and accuracy is good enough

totalRooms-population R2 value: 0.7512234851734285 that value is a bit dubious but it is still good enough

totalRooms-households R2 value: 0.8396946924819624 that value is close to 1 and accuracy is good enough

totalBedrooms-population R2 value: 0.7845703640993615 that value is almost close to 1 and accuracy is good enough

totalBedrooms-households R2 value: 0.9643243067169097 that value is too close to 1 and it is almost exact

accuracy

population-households R2 value: 0.8323149304375397 that value is close to 1 and accuracy is enough

medianIncome-medianHouseValue R2 value: 0.45998761856422554 that value is too far to 1 and doing nothing

with it could lead us to very misleading conclusions and I suspect the graph of that so it is not enough for us.

**The Observations I Have Made**

When I used only corr() command correlation that I got has same result with pearson correlation and I

realized that when I didn’t specify any correlation it worked with pearson.There are images of both in the

project.

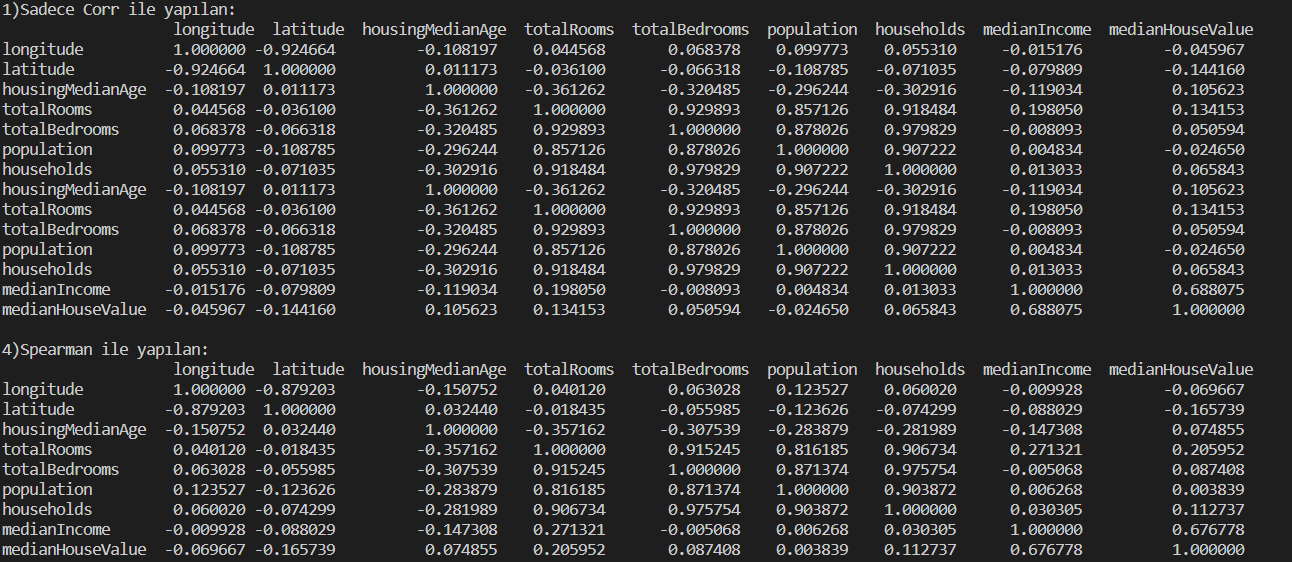
If I enter no information when I integrated my datas for the plt.plot() graph,color=”blue”,linestyle=”solid”.

When I wanted a dot graph I need to do linestyle=none.

**Some Of Mistakes I Got**

**A long time,my correlations didn’t give some correlations’ output while I was printing so I had to**

**open and close the programme.When I printed more I encountered with the same mistake again.**



**Here I printed all correlations but it gave to me only that binary and it passed 2 and 3 number itself.**

**While I was printing y\_test and y\_train values I got the same mistake which I got in the correlation**

**and I don’t know the reason but it can be that my computer is not enough.**