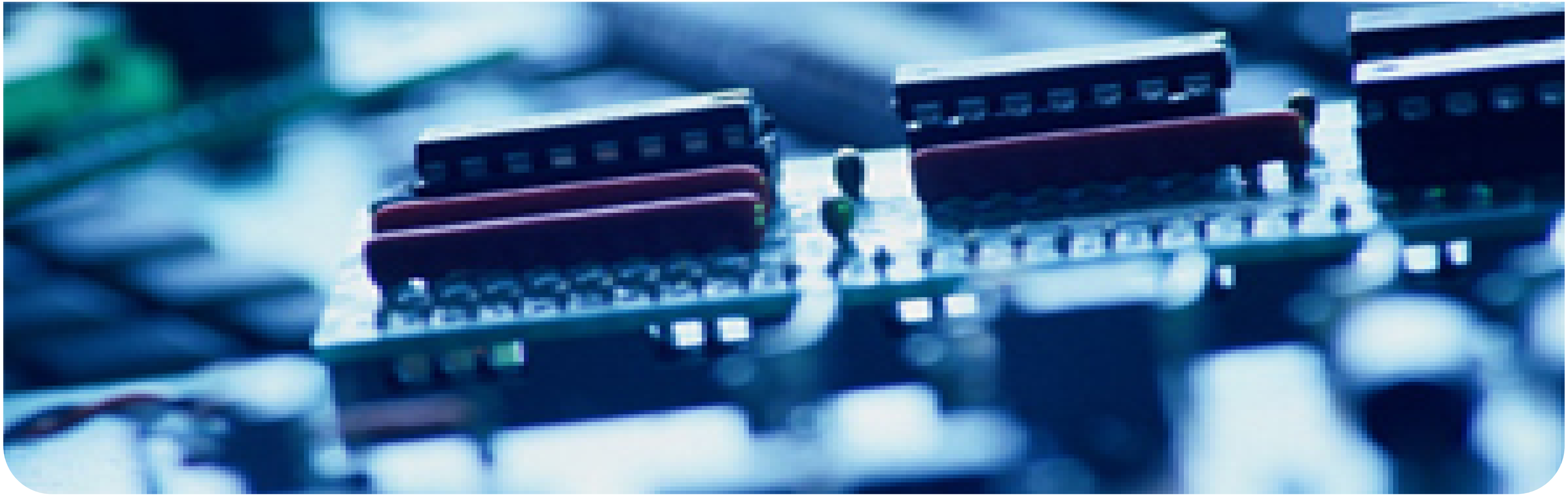
数据挖掘和大数据分析

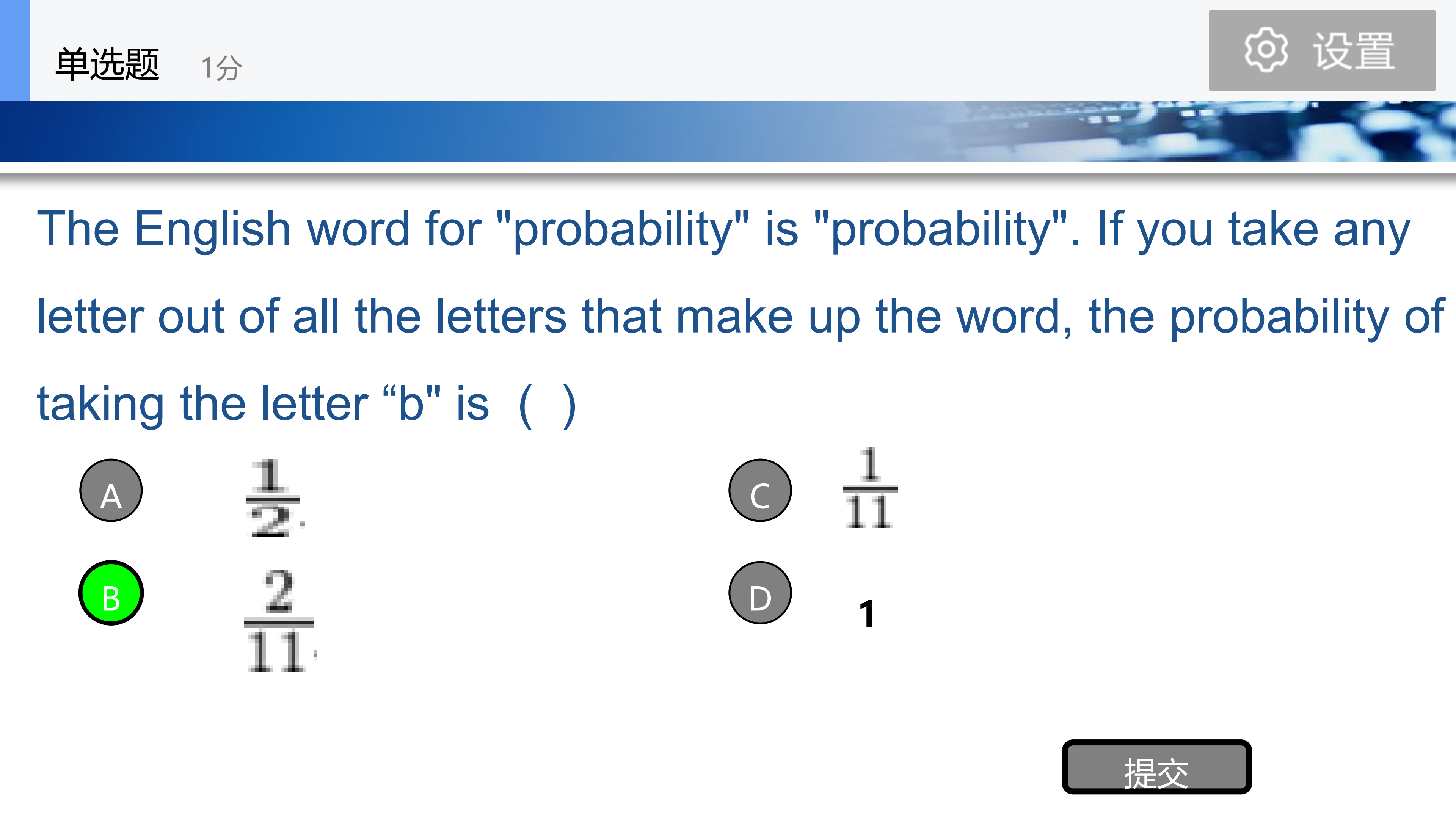


Outline

① Probability Review

② Regression ★

③ DM Lab3



单选题 1分

⚙ 设置

The English word for "probability" is "probability". If you take any letter out of all the letters that make up the word, the probability of taking the letter “b" is （ ）

A $\frac{1}{2}$

[✓ B $\frac{2}{11}$]

C $\frac{1}{11}$

D 1

提交

填空题 0分

设置

此题未设置答案，请点击右侧设置按钮

Take one of the 52 playing cards and find out the probability of the following events:

(1) Draw out a red heart [   ]

(2) Draw out a red old K [   ]

(3) Draw a plum J [   ]

(4) Draw a card that is not Q [   ]

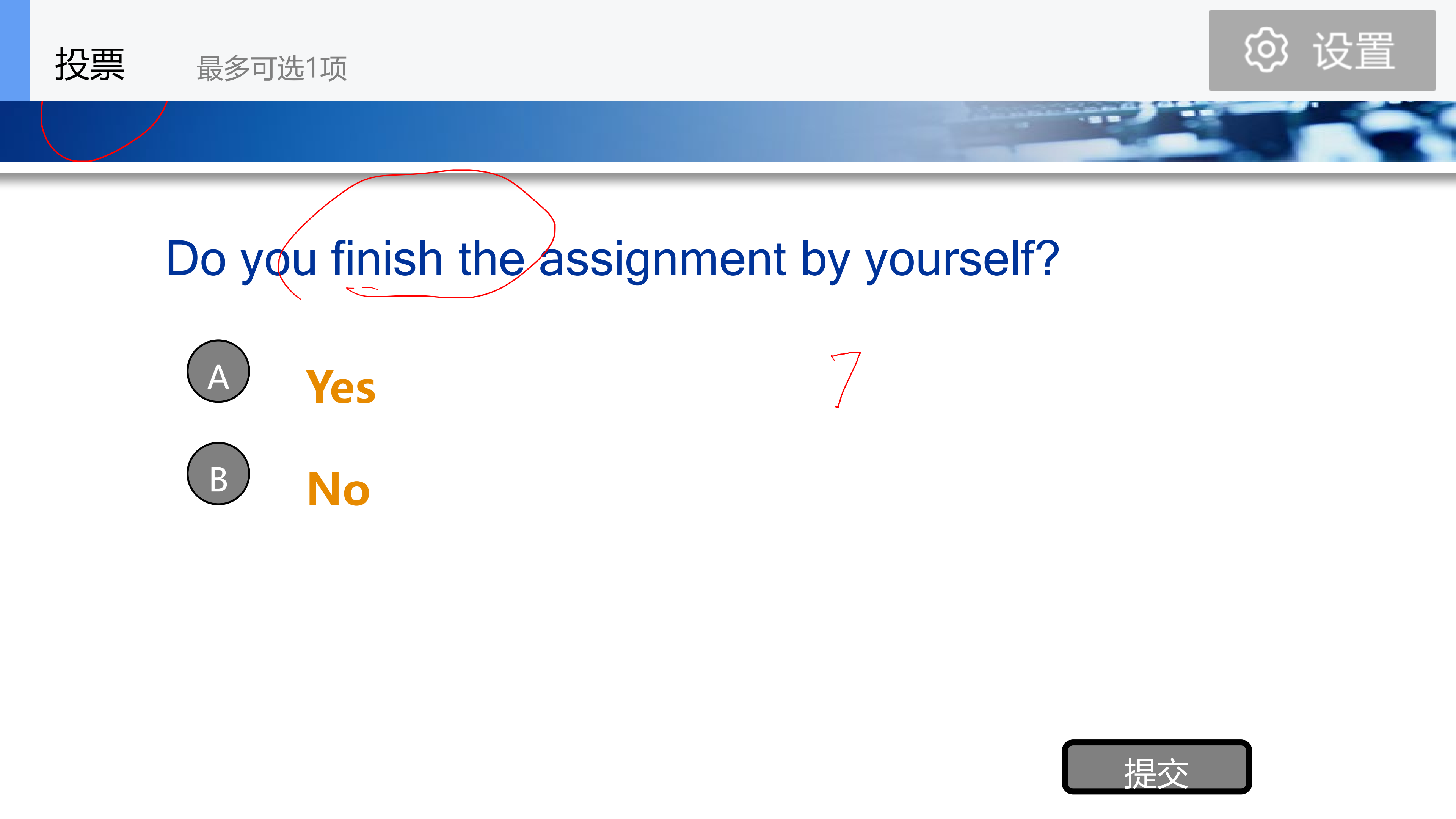
正常使用填空题需3.0以上版本雨课堂

作答

Assignment

- ✓ Two point distribution
- ✓ Binomial distribution
- ✓ Geometric distribution
- ✓ Poisson distribution
- ✓ Uniform distribution
- ✓ Exponential distribution
- ✓ Normal distribution

①Formula ②Coding ③Figure



Do you finish the assignment by yourself?

Ⓐ Yes

Ⓑ No

7

DATA ANALYTICS:

DATA MINING AND BIG DATA

—— Statistic 2

Regression

Variables

(dependent and independent)

Function Clear Relationship

Regression

? Function No Clear Relationship

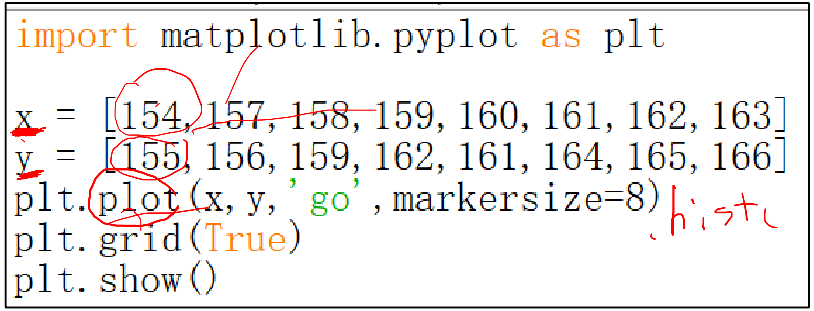
$C = \pi \times d$

$\mathbf{Y} = \mathbf{a} + \mathbf{b}\mathbf{X} + \text{£}$

(x_1, Y_1) (x_2, Y_2) (x_3, Y_3) (x_i, Y_i)

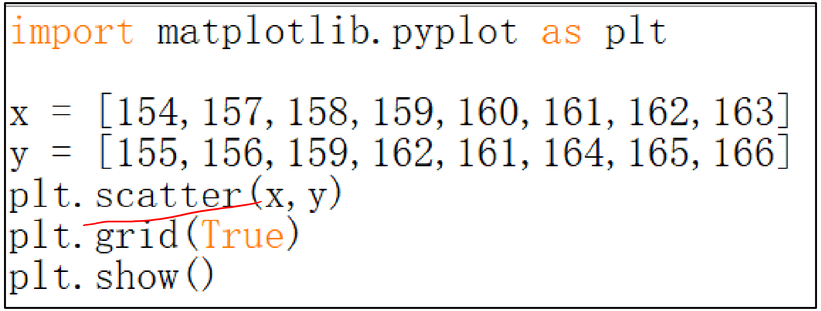
$Y_i = a + bx_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

Regression



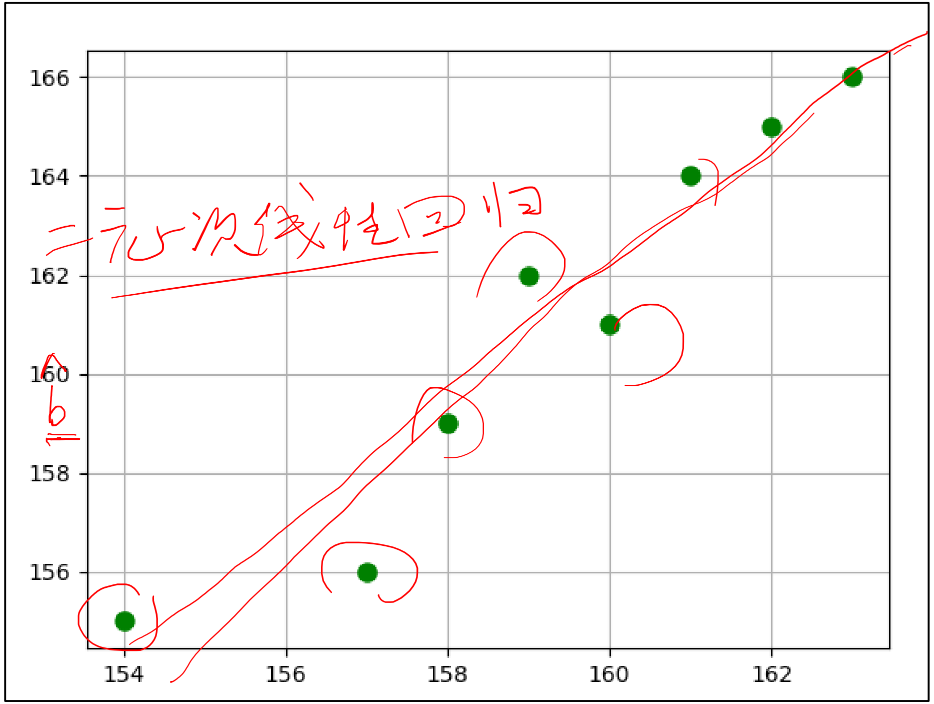
```
import matplotlib.pyplot as plt

x = [154, 157, 158, 159, 160, 161, 162, 163]
y = [155, 156, 159, 162, 161, 164, 165, 166]
plt.plot(x, y, 'go', markersize=8)
plt.grid(True)
plt.show()
```

hist(



```
import matplotlib.pyplot as plt

x = [154, 157, 158, 159, 160, 161, 162, 163]
y = [155, 156, 159, 162, 161, 164, 165, 166]
plt.scatter(x, y)
plt.grid(True)
plt.show()
```

$\hat{a}$



投票 最多可选1项

设置

Please coding the task in 5 minutes.

A Yes

B No

提交

Regression

$$f(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu^{mle} = \arg\max p(\overset{y}{x_1}, \overset{y}{x_2}, \ldots \overset{y}{x_N} \mid \overset{a+bx_i}{\mu}, \sigma^2)$$

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right]$$

$$= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right].$$

$$Q(a,b) = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

偏导

$$\left.\begin{aligned} \frac{\partial Q}{\partial a} &= -2\sum_{i=1}^{n}(y_i - a - bx_i) = 0 \\ \frac{\partial Q}{\partial b} &= -2\sum_{i=1}^{n}(y_i - a - bx_i)x_i = 0 \end{aligned}\right\}$$

ln ab = ln a + ln b

argmax argmin

估计值

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

平均值

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Regression

Mum H x/cm	154	157	158	159	160	161	162	163
Daughter Hy/cm	155	156	159	162	161	164	165	166

关联性？

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\left(\bar{x}\right)^2\right)\left(\sum_{i=1}^{n}y_i^2-n\left(\bar{y}\right)^2\right)}}$$

S_{xy}

$\sqrt{S_{xx}\,S_{yy}}$

平均值

Regression

i	xi	yi	(xi)^2	(yi)^2	xi*yi
1	154	155	23716	24025	23870
2	157	156	24649	24336	24492
3	158	159	24964	25281	25122
4	159	162	25281	26244	25758
5	160	161	25600	25921	25760
6	161	164	25921	26896	26404
7	162	165	26244	27225	26730
8	163	166	26569	27556	27058
∑	1274	1288	202944	207484	205194

Regression

$$\bar{x} = \frac{\sum x_i}{n} = 159.25 \qquad \bar{y} = \frac{\sum y_i}{n} = 161$$

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\left(\bar{y}\right)^2\right)}}$$

$x \cdot y$

$r \leq 1$

$$= \frac{205194 - 8 \times 159.25 \times 161}{\sqrt{202944 - 8 \times 159.25^2}\sqrt{207484 - 8 \times 161^2}}$$

$$= \frac{80}{\sqrt{59.5 \times 116}} \approx 0.963$$

Regression

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\left(\bar{x}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i - 8\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - 8\left(\bar{x}\right)^2} \approx 1.345$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} \approx -53.191$$

回归:

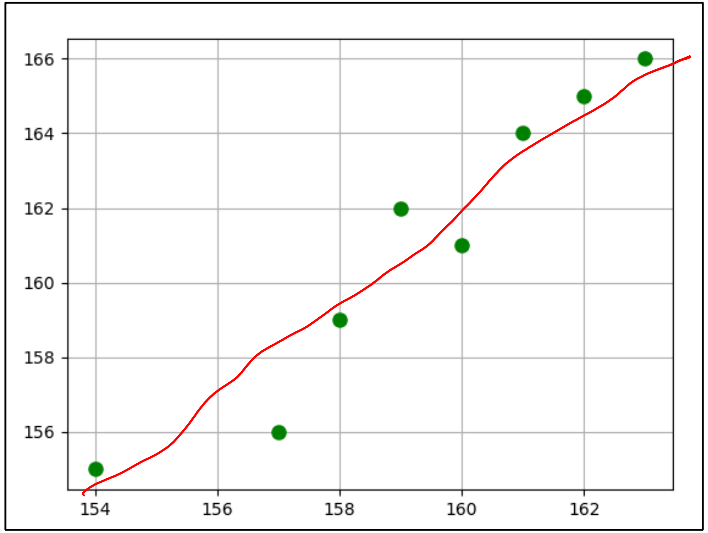
回归 二乘法 — 最小二乘

$$\hat{y} = -53.191 + 1.345x$$

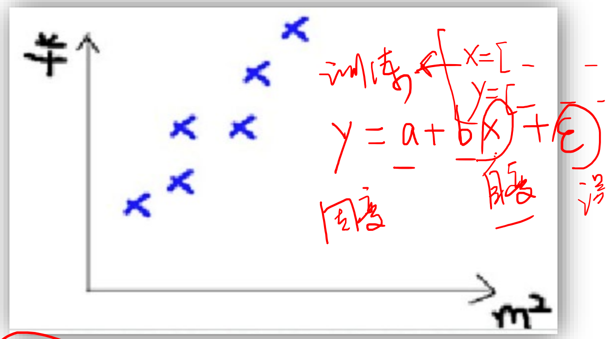
$\varepsilon \sim N(0, \sigma^2)$

正态分布

作业



Regression

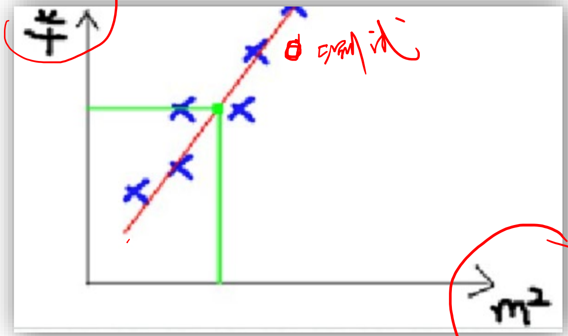


训练 x=[- -]
y=[- - -]

$y = \underline{a} + \underline{b}x + \varepsilon$

固定 服务 误差

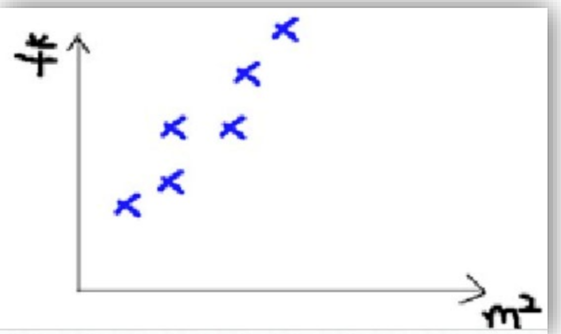
集中 分数

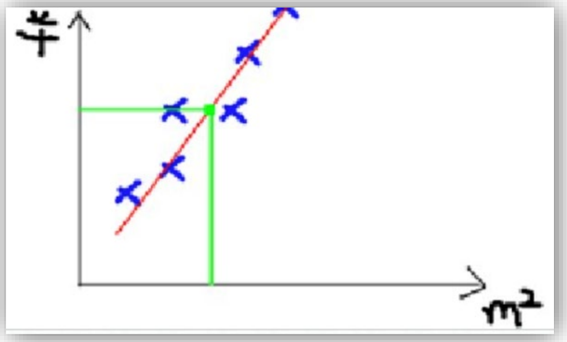


d 测试

```
import pandas as pd
from io import StringIO
from sklearn import linear_model
import matplotlib.pyplot as plt
```

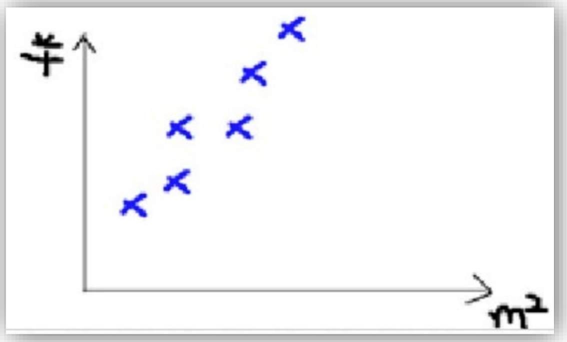
Regression

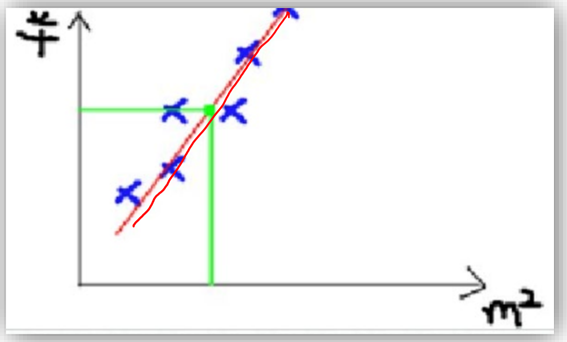




```
csv_data = 'square_feet,price\n150,6450\n200,7450\n250,8450\n300,9450\n350,11450\n400,15450\n600,18450\n'
df = pd.read_csv(StringIO(csv_data))
print(df)
regr = linear_model.LinearRegression()
regr.fit(df['square_feet'].values.reshape(-1, 1), df['price'])
a, b = regr.coef_, regr.intercept_
```

Regression





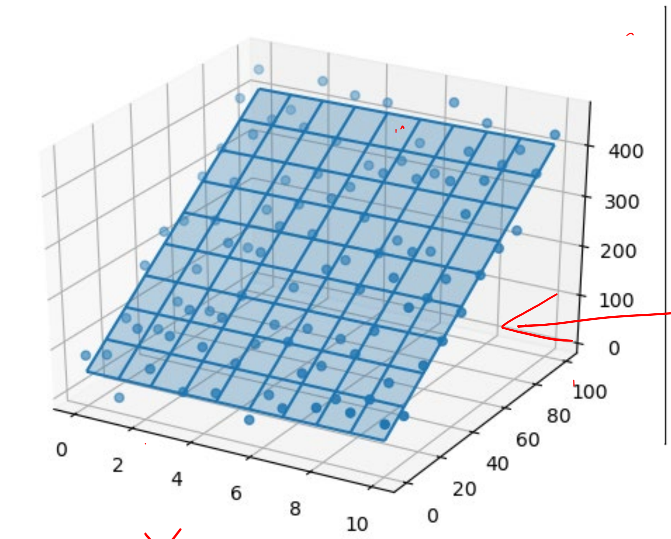
```
area = 238.5
print(a * area + b)
print(regr.predict([[238.5]]))
```

测试

area = 1, 2, 3

[1], [2], [3]

Regression



多元回归方程

Y X

一元一次

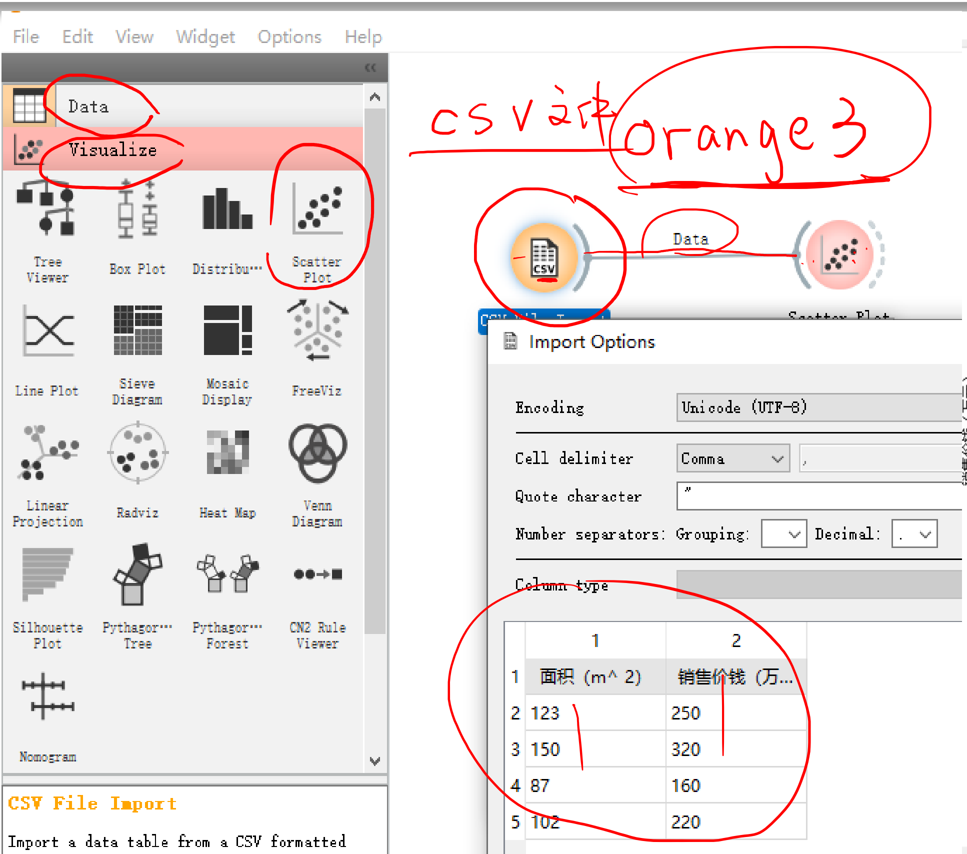
$Z = aX + bY + \varepsilon$

二元一次

```
plt.scatter(df['square_feet'], df['price'], color='blue')
plt.plot(df['square_feet'],
regr.predict(df['square_feet'].values.reshape(-1,1)),
color='red', linewidth=4)
plt.show()
```

X

Regression



CSV File Import

Import a data table from a CSV formatted

Import Options

Encoding: Unicode (UTF-8)

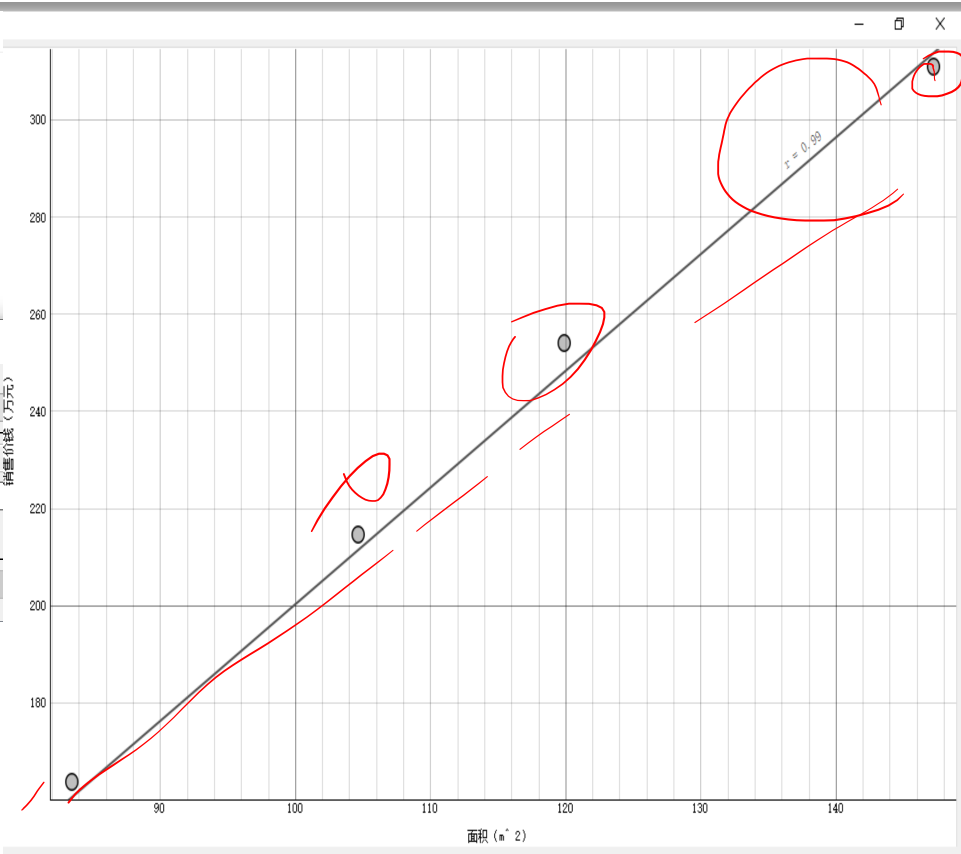
Cell delimiter: Comma ,

Quote character: "

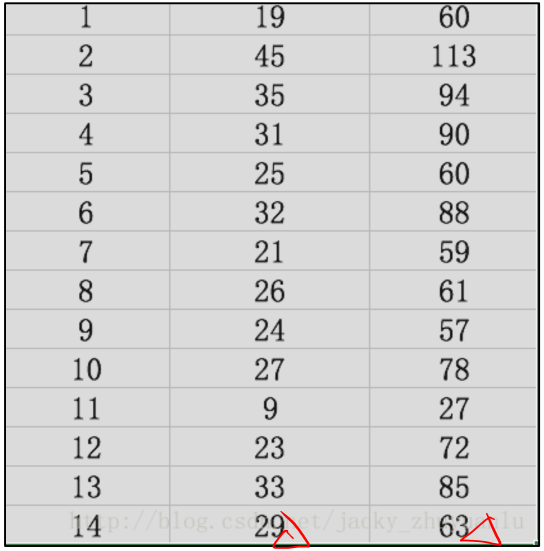
Number separators: Grouping: Decimal: .

Column type

	1	2
1	面积（m^ 2)	销售价钱（万...
2	123	250
3	150	320
4	87	160
5	102	220



DM Lab3



1	19	60
2	45	113
3	35	94
4	31	90
5	25	60
6	32	88
7	21	59
8	26	61
9	24	57
10	27	78
11	9	27
12	23	72
13	33	85
14	29	63

Please Open “test1.csv” File

回归

```
import numpy
from pandas import read_csv
from matplotlib import pyplot as plt
from sklearn.linear_model import LinearRegression
```

pip install pandas
pip install matplotlib
pip install sklearn

① 关联度？ ② 一元一次回归直线

60 ?

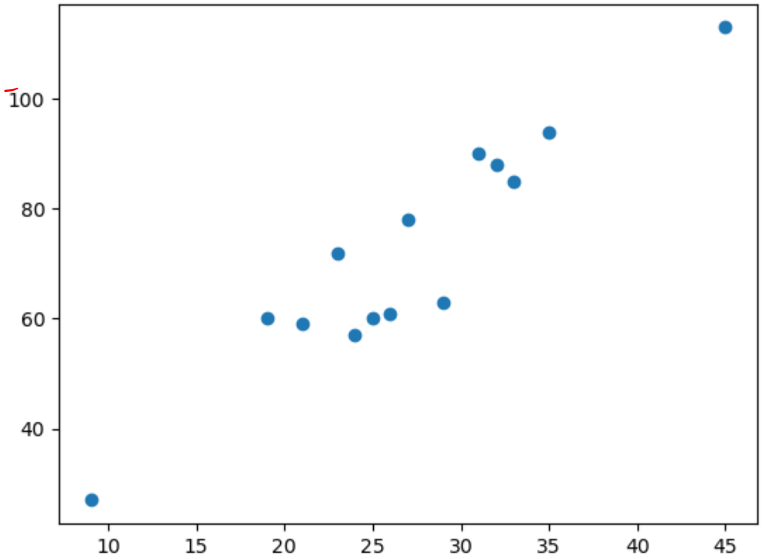
DM Lab3

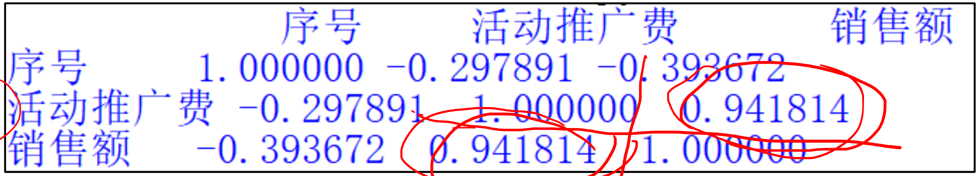
*.py 相对路径

```
data = read_csv('test1.csv')

plt.scatter(data.活动推广费,data.销售额)

plt.show()

u = data.corr()

print(u)
```

0.942

94.2%





```
             序号        活动推广费        销售额
序号        1.000000  -0.297891  -0.393672
活动推广费  -0.297891   1.000000   0.941814
销售额      -0.393672   0.941814   1.000000
```

活动推广费 = 60，销售额=？

投票 最多可选1项

设置

Please coding the task in 15 minutes.

A Yes

B No

提交

DM Lab3

lrModel = LinearRegression() ③

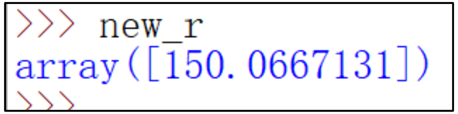
lrModel.fit(x,y)   75   销售额 = ?

lrModel.predict([[60]])

alpha = lrModel.intercept_ ① 截距

beta = lrModel.coef_ ②

new_r = alpha + beta*numpy.array([60])



```
>>> new_r
array([150.0667131])
>>>
```

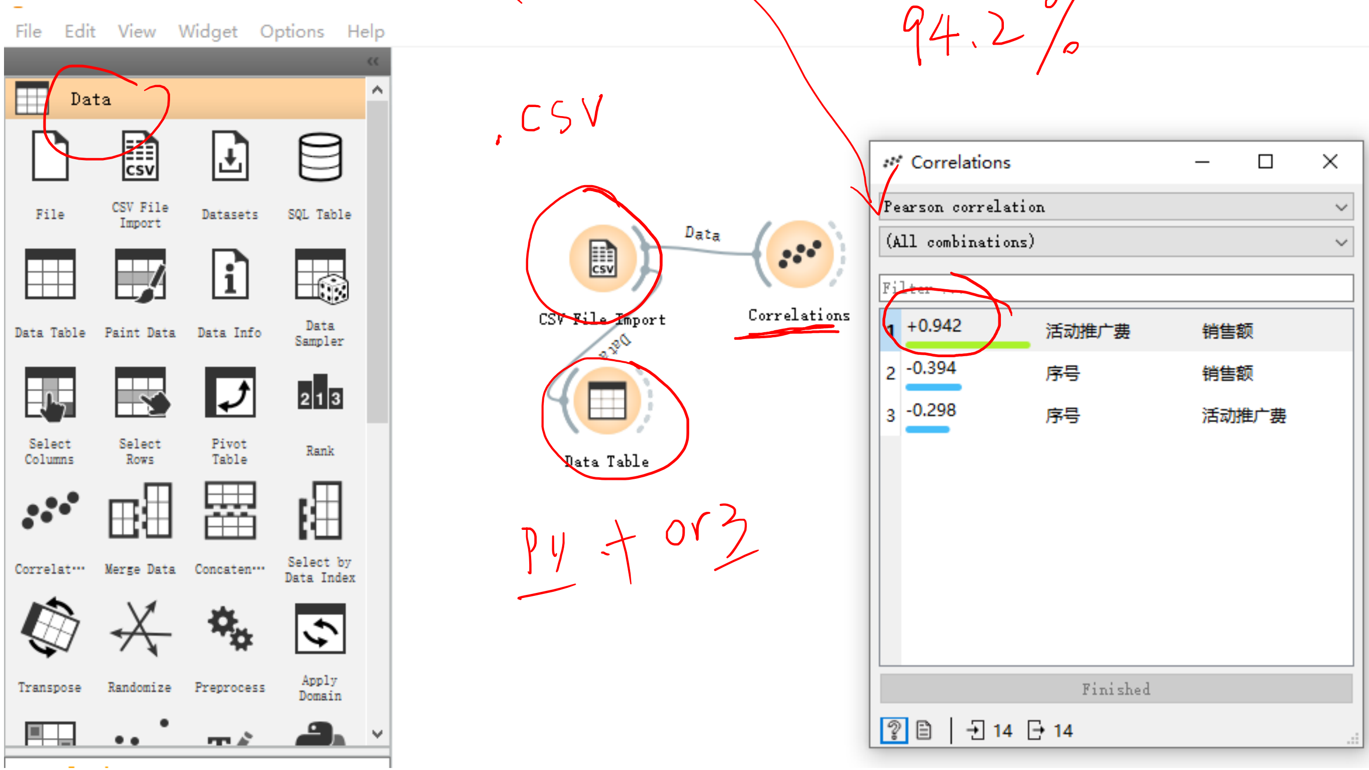
活动推广费 = 60，销售额=150

DM Lab3

orange3

94.2%

.csv



Pearson correlation

(All combinations)

Filter...

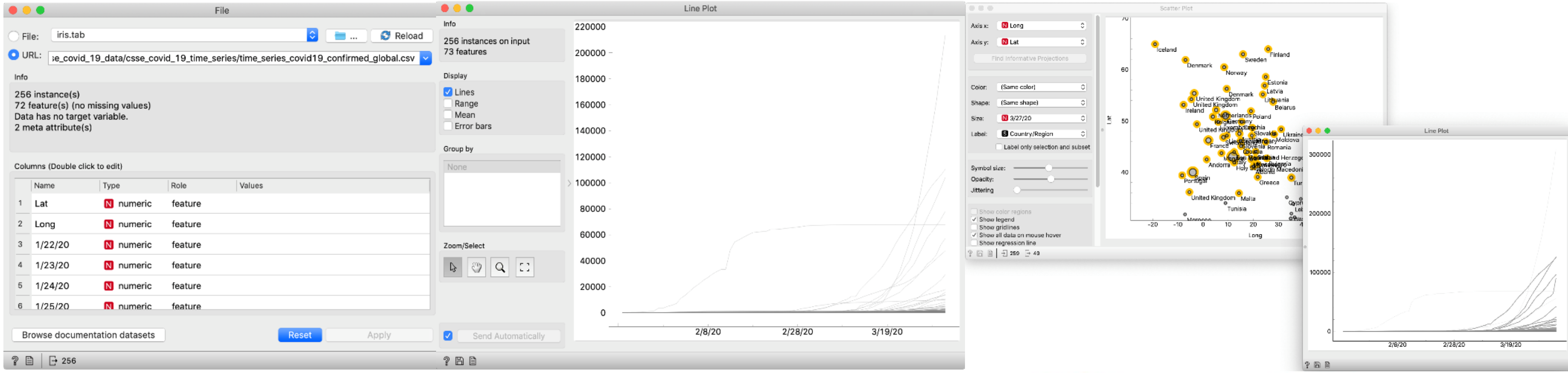
1	+0.942	活动推广费	销售额
2	-0.394	序号	销售额
3	-0.298	序号	活动推广费

Finished

14 14

Py + or3

Data Mining COVID-19 Epidemics: Part 1



File

○ File: iris.tab

● URL: ;e_covid_19_data/csse_covid_19_time_series/time_series_covid19_confirmed_global.csv

Info

256 instance(s)
72 feature(s) (no missing values)
Data has no target variable.
2 meta attribute(s)

Columns (Double click to edit)

	Name	Type	Role	Values
1	Lat	N numeric	feature	
2	Long	N numeric	feature	
3	1/22/20	N numeric	feature	
4	1/23/20	N numeric	feature	
5	1/24/20	N numeric	feature	
6	1/25/20	N numeric	feature	

Line Plot

Info

256 instances on input
73 features

Display

☑ Lines
□ Range
□ Mean
□ Error bars

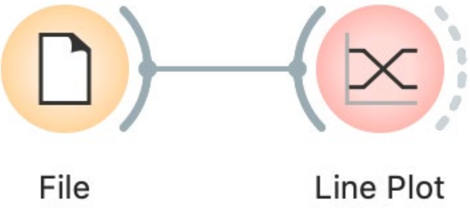
Group by

None

☑ Send Automatically

Scatter Plot

Axis x: N Long
Axis y: N Lat
Color: (Same color)
Shape: (Same shape)
Size: N 3/27/20
Label: S Country/Region
□ Label only selection and subset
□ Show color regions
☑ Show legend
□ Show gridlines
☑ Show all data on mouse hover
□ Show regression line



https://orange.biolab.si/blog/2020/2020-04-02-covid-19-basic/

贵在坚持！