

# 20214064-李恺阳-第1次作业

---

## 数据可视化第1次作业

### 版本状态:

<input type="checkbox"/> 进行	<input checked="" type="checkbox"/> 提交
-----------------------------	--

---

### 学生信息:

• 姓名: 李恺阳		
• 专业: 数据科学与大数据技术		
• 学号: 20214064		

---

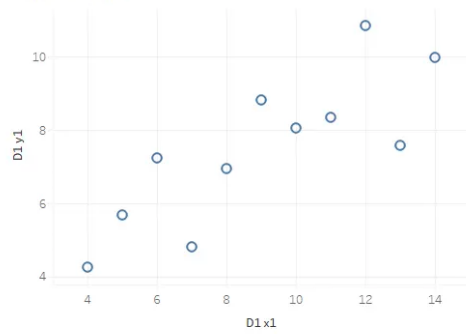
### 任务描述:

- 任务描述:
  - 作业任务1:
  - 设计或开发工作:
    - 用可视化软件或工具（包括但不限于Tableau, Excel, Weka, Weave）对课程第1讲教学内容中所涉及的Anscombe's quartet数据集进行分析，将四组数据进行可视化；
    - 提示:
      - 每组数据可以单独用1个散点图，最后生成4个散点图；
      - 请注意图表的设置，配色要美观；
      - 尝试使用1个散点图，对每组数据使用不同的颜色进行编码；
  - 文字阐述工作:
    - 根据你自己的知识来说明这四组数据的分布特征；
    - 提示:
      - 可以结合均值、方差、相关系数等统计特征进行分析；引用必要的文献；给出必要的公式；
    - 结合这个任务，用适当文字阐述数据可视化的意义和价值；
- 
- - 作业任务2:
  - 用脚本语言或编程语言，计算四组数据的最小二乘法回归线方程；
    - 提示:

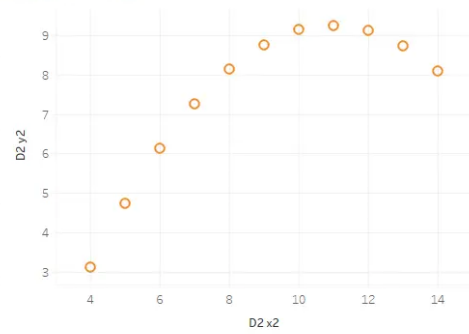


- 将四组数据在一个图表中可视化

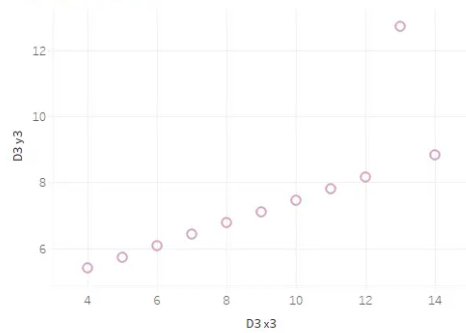
数据组D1散点图



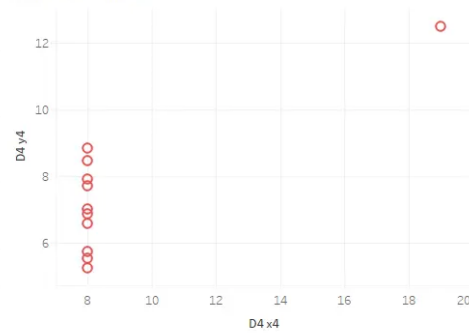
数据组D2散点图



数据组D3散点图



数据组D4散点图



- 分析说明这四组数据的分布特征

- 分别计算四组数据的均值、方差、相关系数

- 计算公式

- $$\mu = \frac{\sum_{x=1}^n y_x}{n}$$

- 方差: 
$$s^2 = \frac{\sum_{i=1}^n (y_i - \mu)^2}{n}$$

- 相关系数: 
$$\gamma = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{x=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

- 使用python计算各组数据的相关指标

```
1 import numpy as np
2 import pandas as pd
3 # 读取Excel文件并去掉前三行
4 df = pd.read_excel("Anscombe's quartet.xlsx", skiprows=[0, 1])
5 data = np.array(df)
6 # 遍历每组数据, 计算均值、方差和相关系数
7 for i in range(4):
8     mean_x = data[:, i*2].mean()
9     print(f"\nStatistics for Group {i + 1}:\n")
10    print(f"Mean_X: {mean_x:.2f}")
11    mean_y = data[:, i*2+1].mean()
12    print(f"Mean_Y: {mean_y:.2f}")
13    var_x = np.var(data[:, i*2], ddof=0)
14    print(f"Variance_X: {var_x:.2f}")
15    var_y = np.var(data[:, i*2+1])
16    print(f"Variance_Y: {var_y:.2f}")
17    corr = np.corrcoef(data[:, i*2], data[:, i*2+1])[0, 1]
18    print(f"Correlation: {corr:.2f}")
19
```

## ■ 计算结果

### □ 数据组D1

```
Statistics for Group 1:

Mean_X: 9.00
Mean_Y: 7.50
Variance_X: 10.00
Variance_Y: 3.75
Correlation: 0.82
```

### □ 数据组D2

```
Statistics for Group 2:

Mean_X: 9.00
Mean_Y: 7.50
Variance_X: 10.00
Variance_Y: 3.75
Correlation: 0.82
```

### □ 数据组D3

```
Statistics for Group 3:

Mean_X: 9.00
Mean_Y: 7.50
Variance_X: 10.00
Variance_Y: 3.75
Correlation: 0.82
```

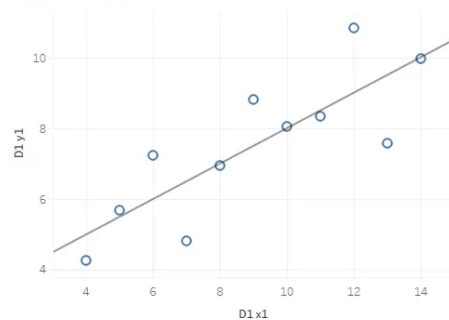
### □ 数据组D4

```
Statistics for Group 4:

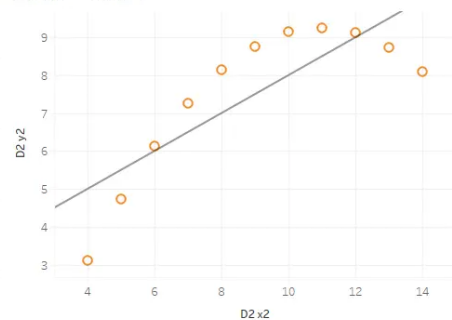
Mean_X: 9.00
Mean_Y: 7.50
Variance_X: 10.00
Variance_Y: 3.75
Correlation: 0.82
```

## ■ 对四组数据均进行线性拟合

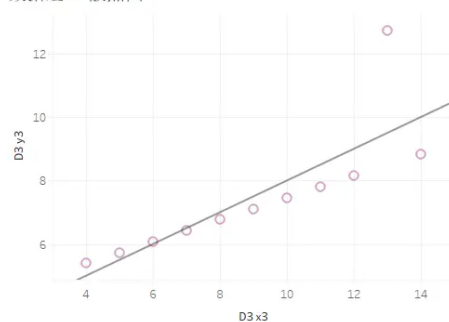
数据组D1散点图



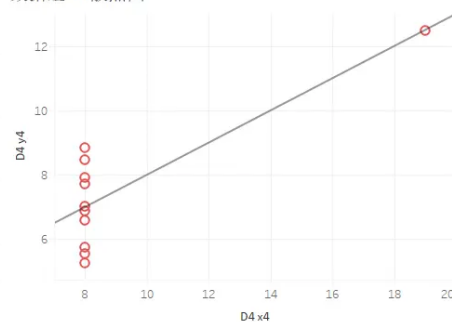
数据组D2散点图



数据组D3散点图



数据组D4散点图



- 由上图可知，拟合结果极为相似
- 结果分析
  - 从各项指标来看，四组数据的均值、方差、相关指数均相等，线性拟合极为相似，仅从数据上来看，四组数据反映出的情况极为相似。但从实际绘图上来看，第一三组都接近线性分布，第三组更加精确，而第二组数据更为接近二次分布，第四组数据与线性分布相差甚远。从上述分析可知，仅看数据的均值、方差、相关指数等指标会导致实验结果并不可靠，而数据可视化的操作让实验结果的可靠性大大提升。因此，数据可视化对于数据分析挖掘方面是极为重要的。

## 任务二：用脚本语言或编程语言，计算四组数据的最小二乘法回归线方程

- 使用python编写最小二乘法回归线方程
  - 最小二乘法思想：
    - 最小二乘法将最小化误差平方之和来作为目标，从而找到最优模型，这个模型可以拟合观察数据
    - 最小二乘法步骤：
      - 残差： $e_i = y_i - (\omega \times x_i + b)$
      - $Min \sum_i^n e_i^2$
      - 得到回归线方程  $y = \omega \times x + b$
  - Python 是一种解释型、面向对象、动态类型的编程语言
  - 代码：

```

1  import numpy as np
2  import pandas as pd
3  import matplotlib.pyplot as plt
4  df = pd.read_excel("Anscombe's quartet.xlsx", skiprows=[0, 1])
5  data = np.array(df)
6  plt.figure(figsize=(8,8), dpi=80)
7  for j in range(4):
8      data_x = data[:,j*2]
9      data_y = data[:,j*2+1]
10     m = len(data_y)
11     x_bar = np.mean(data_x)
12     sum_yx = 0
13     sum_x2 = 0
14     sum_delta = 0
15     for i in range(m):
16         x = data_x[i]
17         y = data_y[i]
18         sum_yx += y * (x - x_bar)
19         sum_x2 += x ** 2
20     # 根据公式计算w
21     w = sum_yx / (sum_x2 - m * (x_bar ** 2))
22     for i in range(m):
23         x = data_x[i]

```

```

24     y = data_y[i]
25     sum_delta += (y - w * x)
26     b = sum_delta / m
27     print(f"Group {j + 1}: y = {w:.2f} * x + {b:.2f}")
28     pred_y = w * data_x + b
29     ax1 = plt.subplot(2,2,j+1)
30     ax1.scatter(data_x, data_y)
31     ax1.plot(data_x, pred_y, c='orangered', label='line')
32     ax1.set_title(f"\nStatistics for Group {j + 1}:\n")
33 plt.show()
34

```

○ 运行结果:

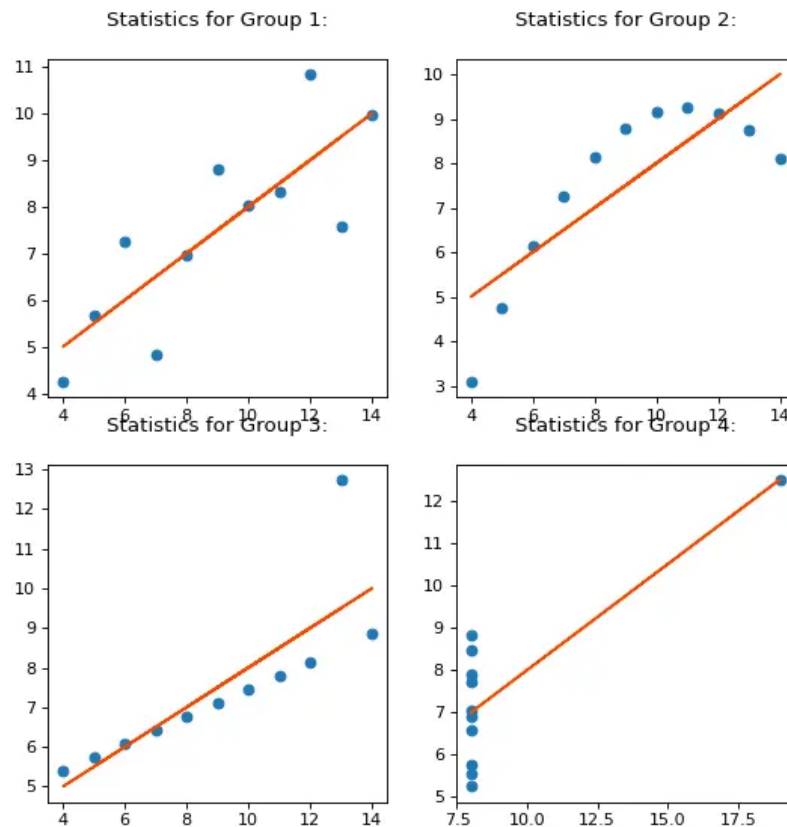
```

Group 1: y = 0.50 * x +3.00
Group 2: y = 0.50 * x +3.00
Group 3: y = 0.50 * x +3.00
Group 4: y = 0.50 * x +3.00

```

■ 四组数据的回归线方程:

■ 回归线与散点图



○ 结果分析:

■ 从上述结果来看，四组数据所得到的回归线方程是一样的，但是四组数据的真实分布趋势确实完全不同的。因此，数据可视化是数据分析过程中保障实验可靠性的重要操作之一。