

# Practice of AI

C2: Machine learning & Data analyze

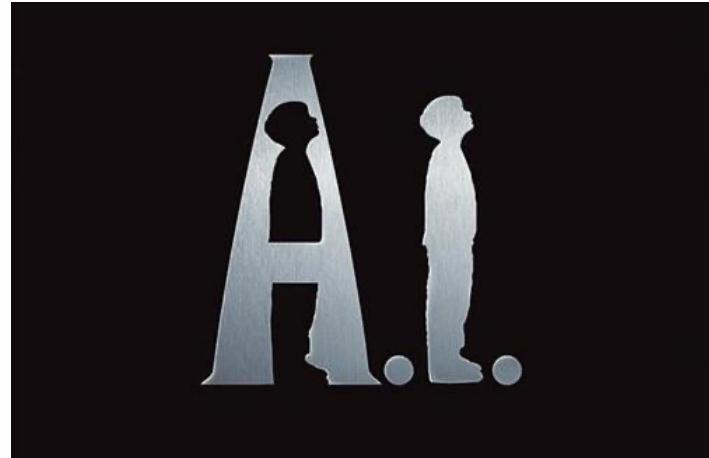
Jim Xie

2020/10/6



# Goal

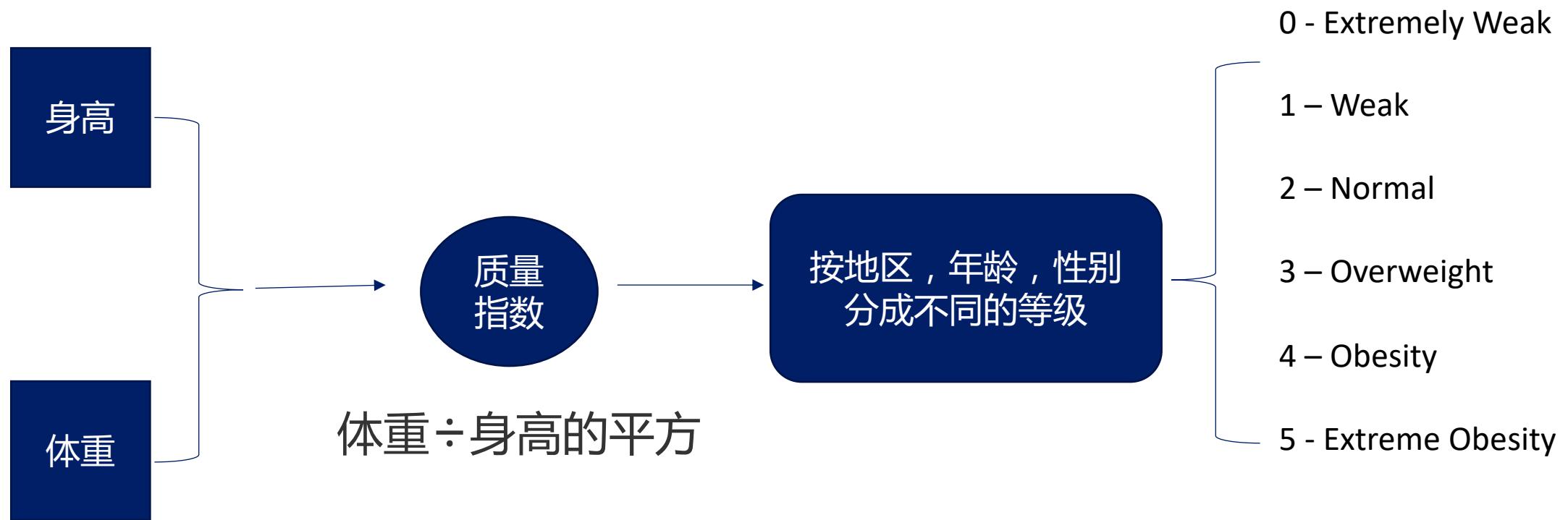
---



了解ML流程，进行数据探索分析和特征选取

例子：训练一个BMI分级模型

# BMI等级



# 试验样本

	Gender	Height	Weight	Level
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
5	Male	189	104	3



- 性别
- 身高
- 体重
- 健康等级

{ 0 - Extremely Weak  
1 – Weak  
2 – Normal  
3 – Overweight  
4 – Obesity  
5 - Extreme Obesity

[http://10.206.67.123:8888/edit/dataset/500\\_Person\\_Gender\\_Height\\_Weight\\_Index.csv](http://10.206.67.123:8888/edit/dataset/500_Person_Gender_Height_Weight_Index.csv)

# BMI

---

Body Mass Index 分类识别  
( 简化版 )

## Demo

<http://10.206.67.123:8888/notebooks/JimXie/src/C2/BMI-draft.ipynb>

# 感受

---

1. 这么简单？感觉结果稀里糊涂就出来了？
2. 靠谱吗？敢在product环境使用吗？
3. 样本里只有身高，体重数据，如果无效怎么办？

[识别无效的例子](#)

# 简化流程

---



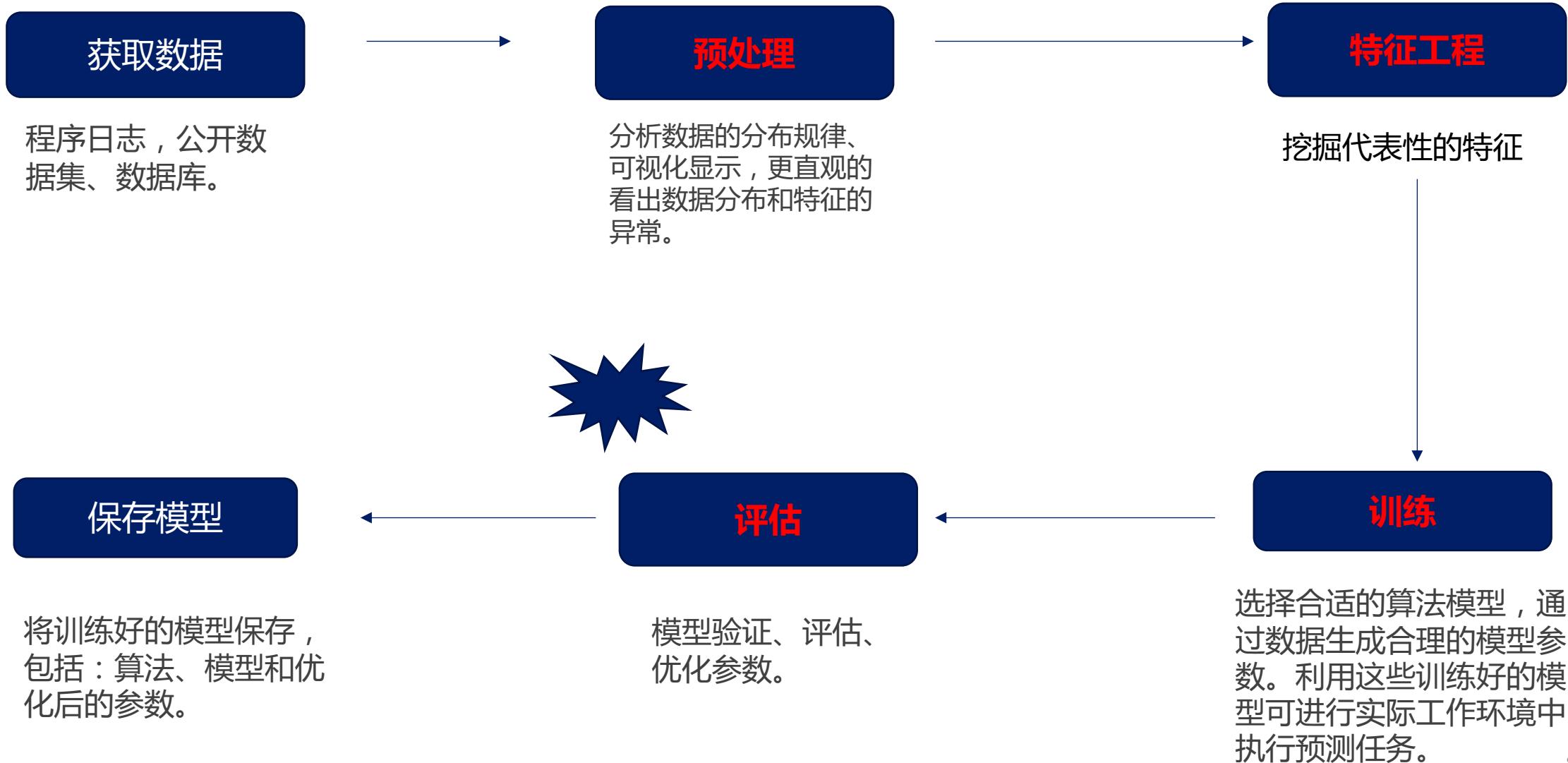
获取数据

数据预处理

模型训练

模型评估

# 细化流程



# 思考题

---

## 脑筋急转弯：

- 现在要在机场装个摄像头，对恐怖分子进行识别；
- 如何设计识别算法，达到90%+的精准度？

精准度        识别正确的样本数量 ÷ 总样本数量

# 基本概念

---

- 样本，特征，数据集分别是什么？

# 结构数据的特征和样本

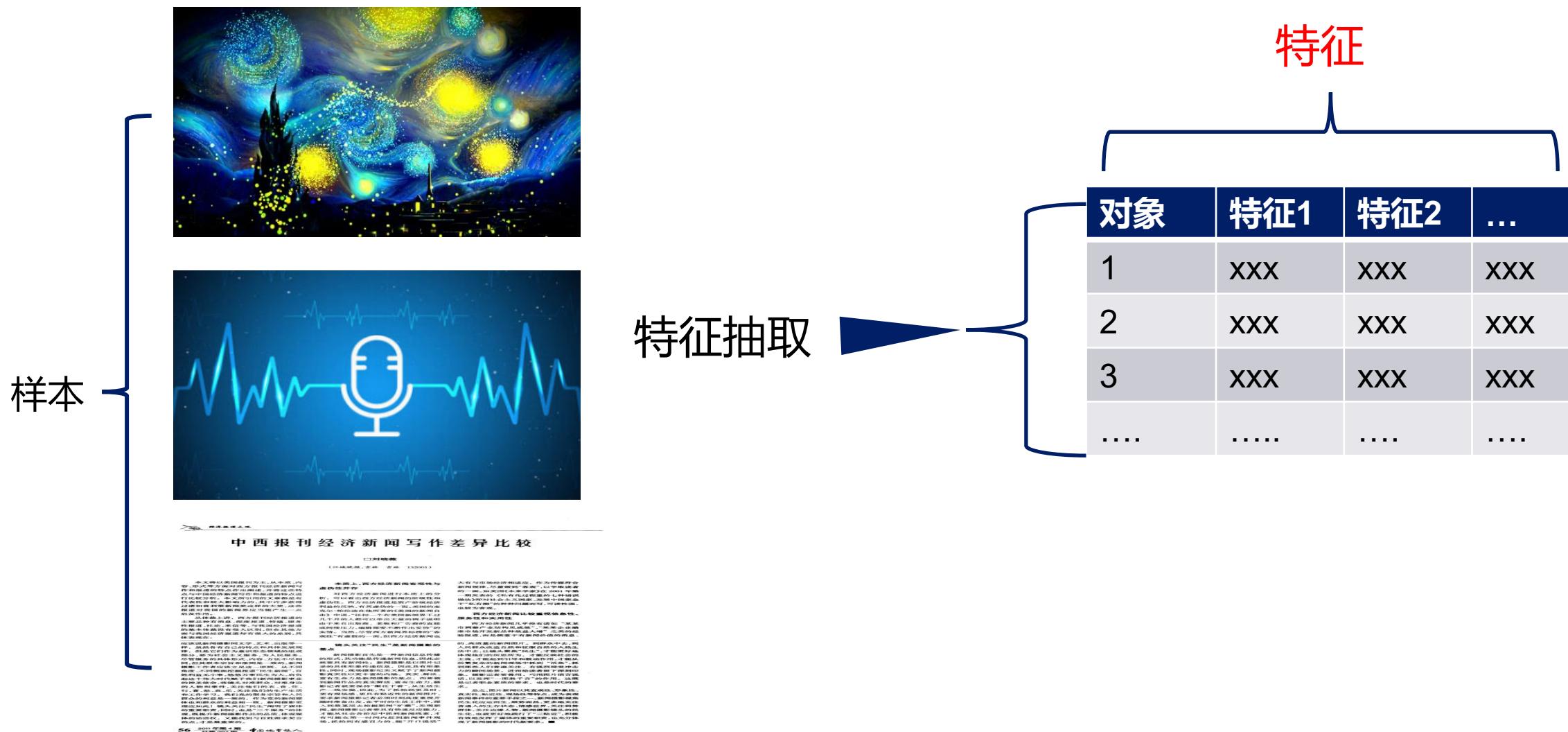
**特征** (输出特征，输入特征)

	Gender	Height	Weight	Level
0	Male	174	96	4
1	Male	189	87	2
2	Female	185	110	4
3	Female	195	104	3
4	Male	149	61	3
5	Male	189	104	3

**样本**  
(训练样本  
测试样本)

标注标签

# 非结构数据的特征和样本



# 数据集划分

## 数据集: 样本的集合

类型	类比	作用	经验比例 ( 小样本 )	经验比例 ( 百万级大样本 )
训练集	学生课本	训练模型	60% 或 70%	98%
验证集	作业	寻找最佳参数	20% 或 0%	1%
测试集	考试	评估模型性能 ( 不会改变学习算法或参数 )	20% 或 30%	1%

1. 训练集上训练 → 2. 验证集上评估参数 → 3. 测试集上测试结果

# 数据预处理

---

- 样本扫描
- 数据清洗
- 数据变换
- 数据可视化

# 样本扫描

---

拿到样本后，先进行整体扫描，对数据产生一个宏观的印象。

- 读取样本
- 样本大小
- 数据类型
- 缺失值或空值
- 有哪些值
- 统计信息
- .....

常用函数示例：[常用pandas函数](#)

# 数据清洗

	Gender	Height	Weight
0	Male	174	96
1	Male		87
2	Female	185	110
3	Female	1197	104
4	Male	149	61

替换成编码  
( 0,1 )

删除 or 插值

```
import pandas as pd
df1 = pd.read_csv('./data/500_Person_Gender_Height_Weight_Index.csv')
df1['Height'].fillna(df1.Height.mean())
#df1['Height'].fillna(df1.Height.mode())
#df1['Height'].fillna(df1.Height.median())
df1.dropna()
df1.dropna(axis=1)
df1.replace("Male", 0,inplace=True)
df1.replace("Female", 1,inplace=True)
```

插值方法	样本中的值	pandas函数
平均数	169.00	mean()
中位数	170.00	median()
众数	168.00	mode()
回归数	通过模型预测得到	

# 可视化

---

## 可视化的作用和常用方法

# 可视化常用方法

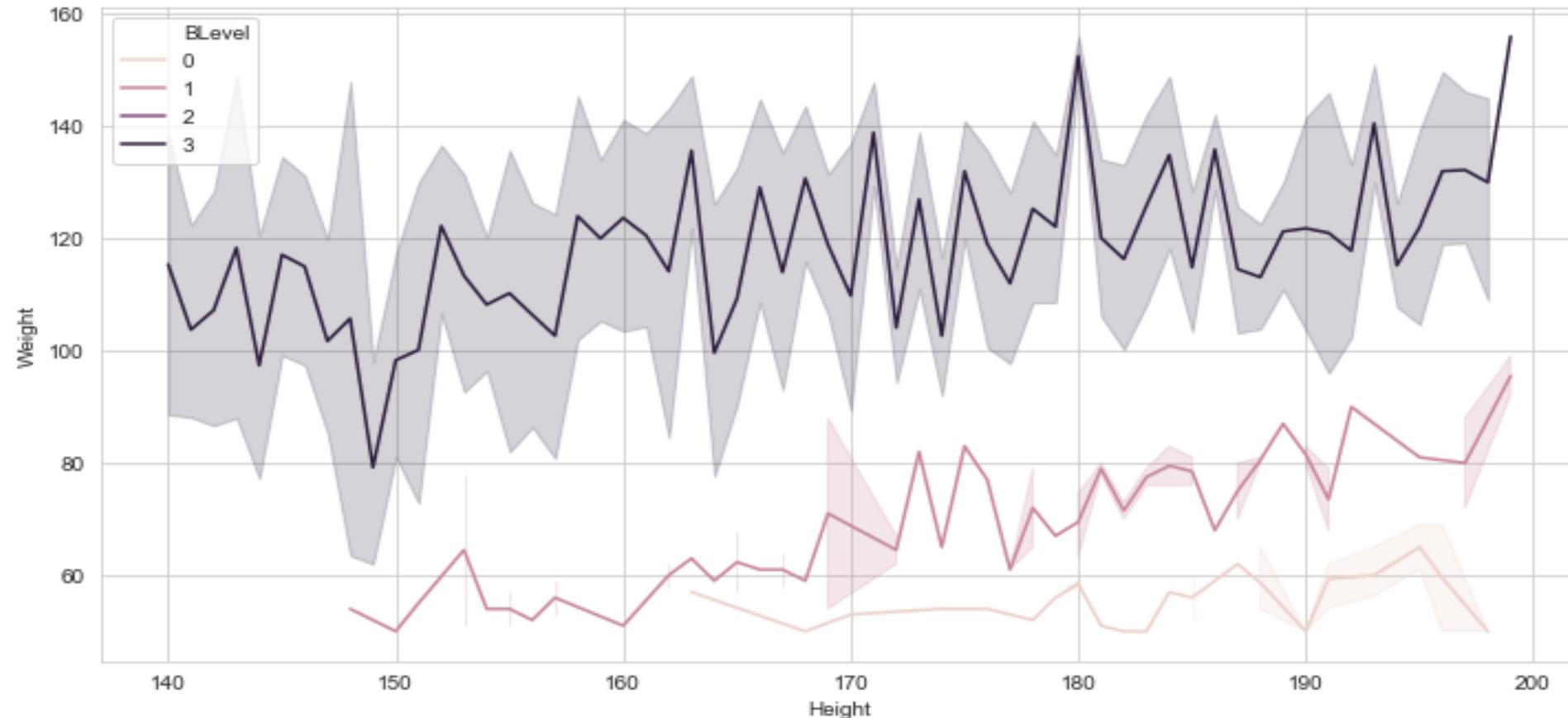
---

可视化 = 作图 + 数据转换

- 趋势图
  - 散点图
  - 直方图
  - 热力图
  - .....
- 
- 归一化
  - 标准化
  - .....

# 趋势图 : Good case

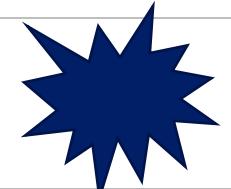
观察走势：试验样本中，身高和体重是否有关联？



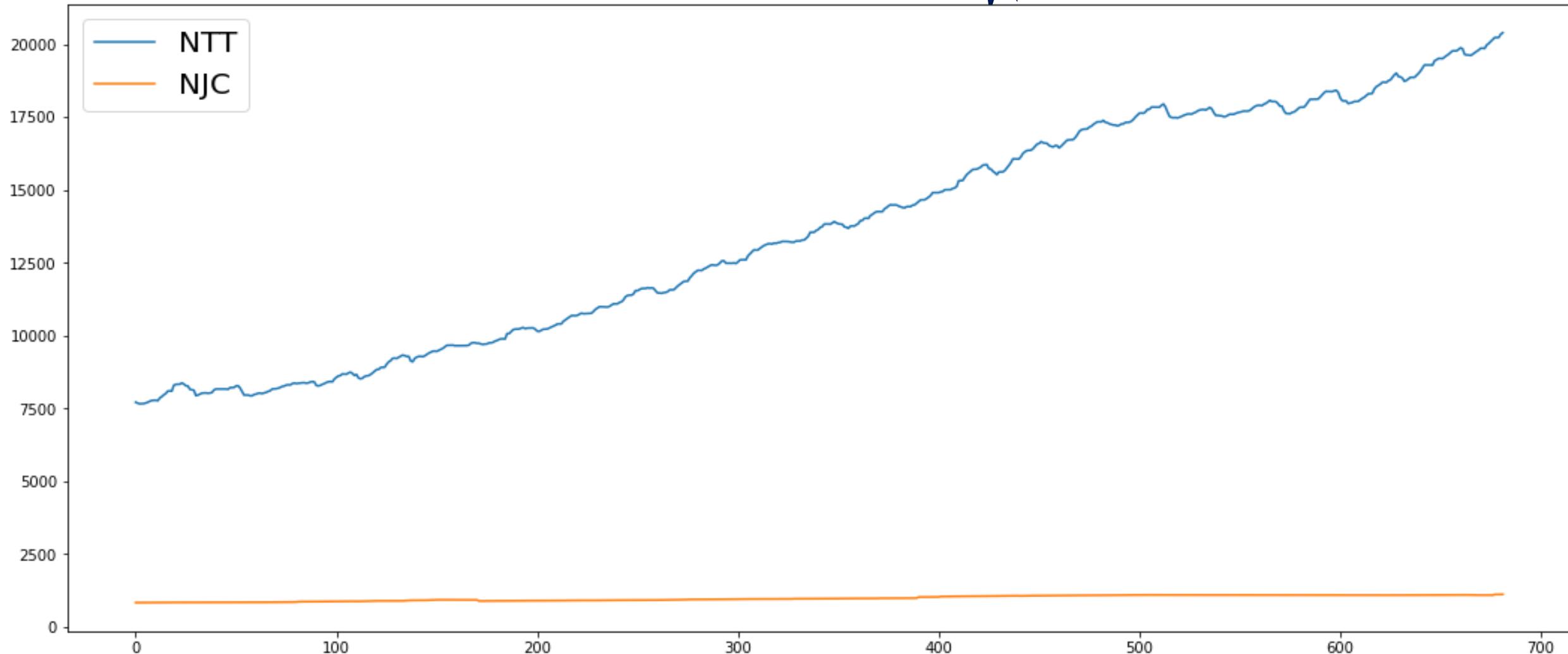
```
plt.figure(figsize=(12,6))
sns.lineplot(x="Height",y="Weight",hue='Level',ci=95,data=df1)
```

# 趋势图 : Bad case

观察走势: 产品销售增长是否有规律?



数据差别大，看不清楚



# 数据变换：归一化

---

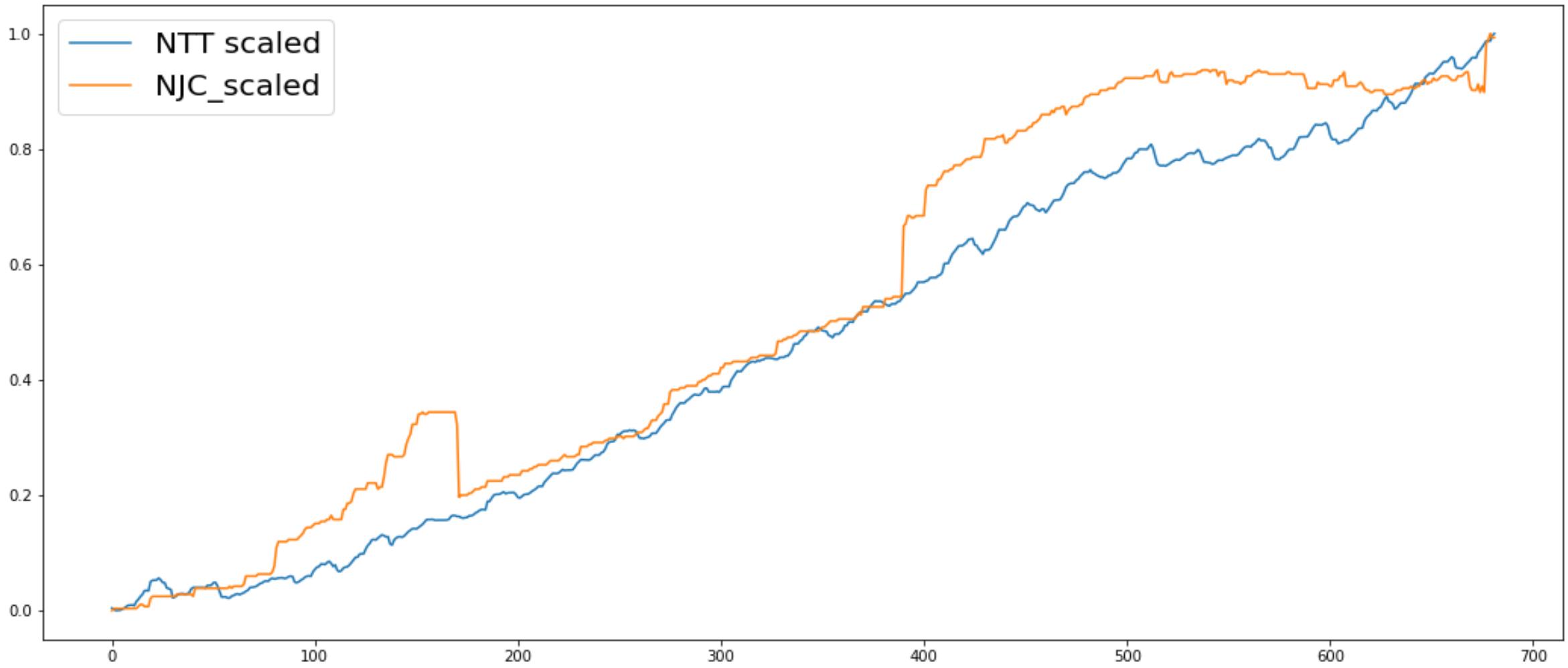
产生原因： 单位不同导致 (如白细胞个数和身高数据 )

解决方法： 将样本中的数据压缩至0 ~ 1之间。 (去量纲)

最值归一化 :  $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$  ( x为当前值 )

Method	Data					
Raw data	[ 529    578    466    437    318 ]					
MaxMinScal	[ 0.81    1.    0.569    0.45769    0. ]					

# 趋势图：归一化后

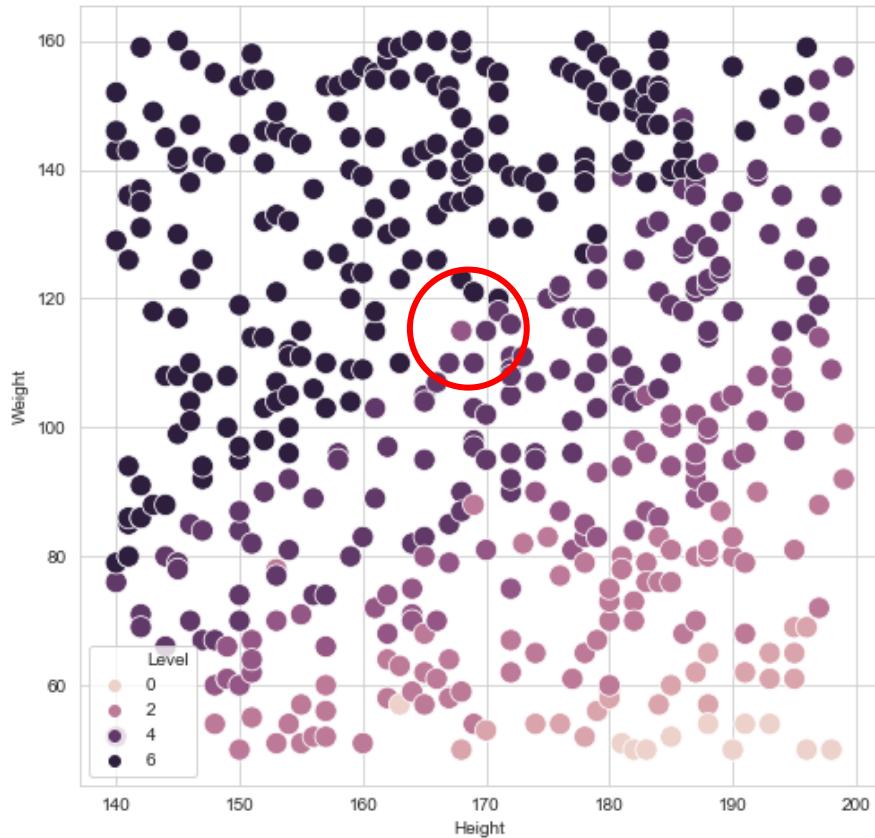


优点：特征呈现出来了，相对关系不变

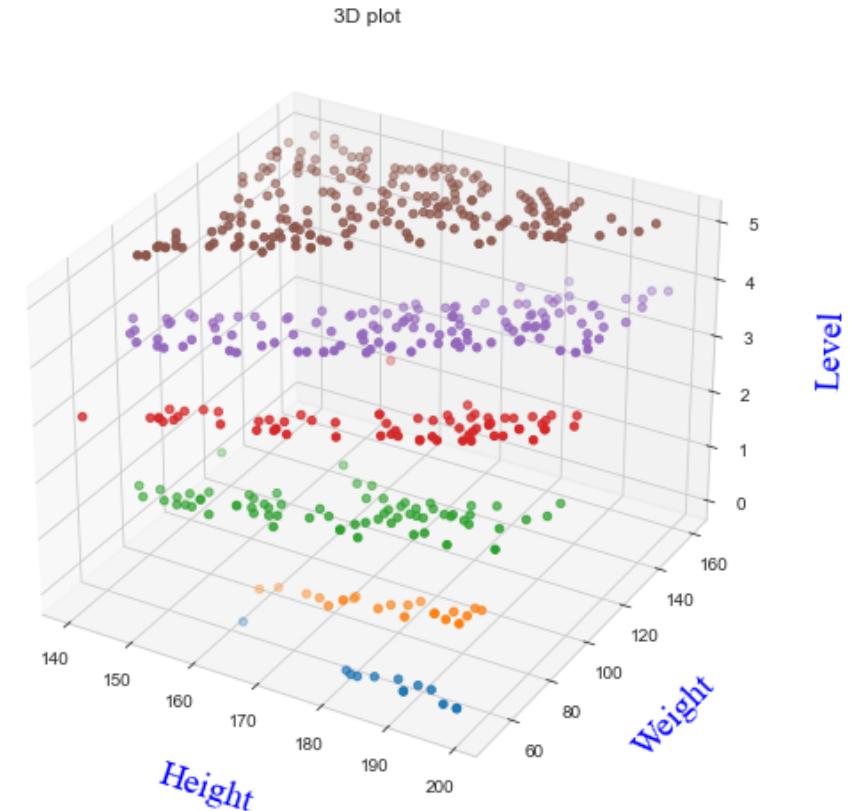
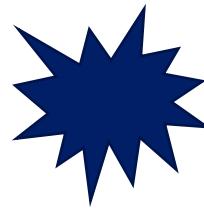
缺点：需要预先知道最大/最小值，异常值不敏感

# 散点图: Bad case

观察数据的分布：肥胖程度，身高，体重的分布



- 异常数据不明显
- 类别多，辨识度不够



```
ax = sns.scatterplot(x='Height', y="Weight", s=150, hue="Level", data=df3)
```

# 数据变换：标准化

变换方法： $x' = \frac{x - \mu}{\sigma}$  <http://10.206.67.123:8888/notebooks/JimXie/excise/excise-1.ipynb>

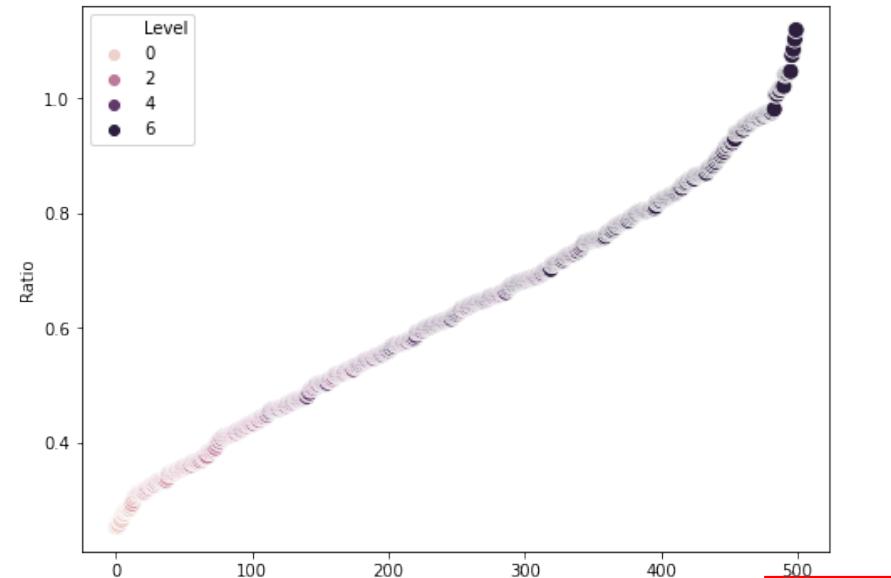
Method	Data				
Raw data	[ 529                578                466                437                318 ]				
Z-score	[ 0.71550817    1.26850345    0.00451425    -0.32276867    -1.6657572 ]				

- 作用1：去量纲
- 作用2：异常值检测
- 作用3：更容易显现特征

# 散点图: Good case

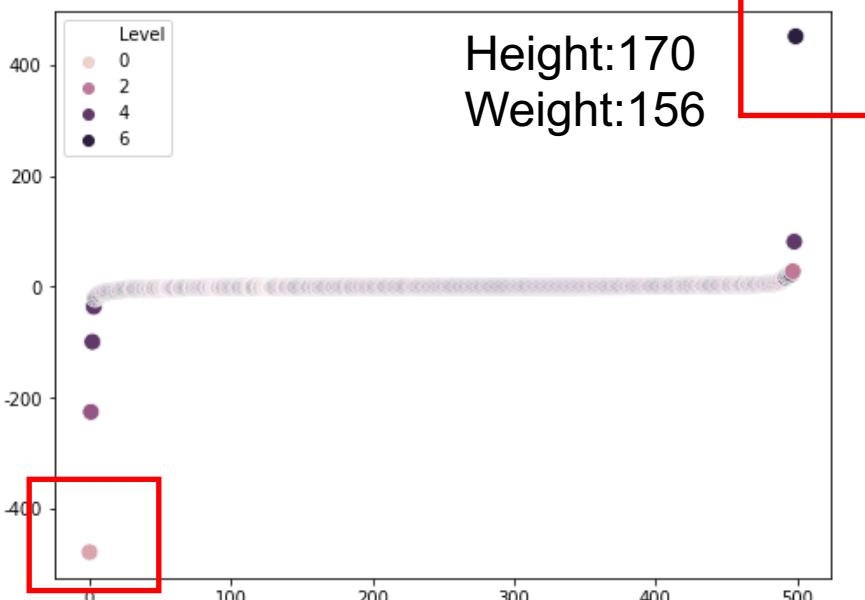
	Gender	Height	Weight	Ratio
490	Male	140	143	1.02
491	Male	143	149	1.04
492	Female	140	146	1.04
493	Male	140	146	1.04
494	Female	151	158	1.05
495	Male	148	155	1.05

身高 , 体重比  
( 原始 )



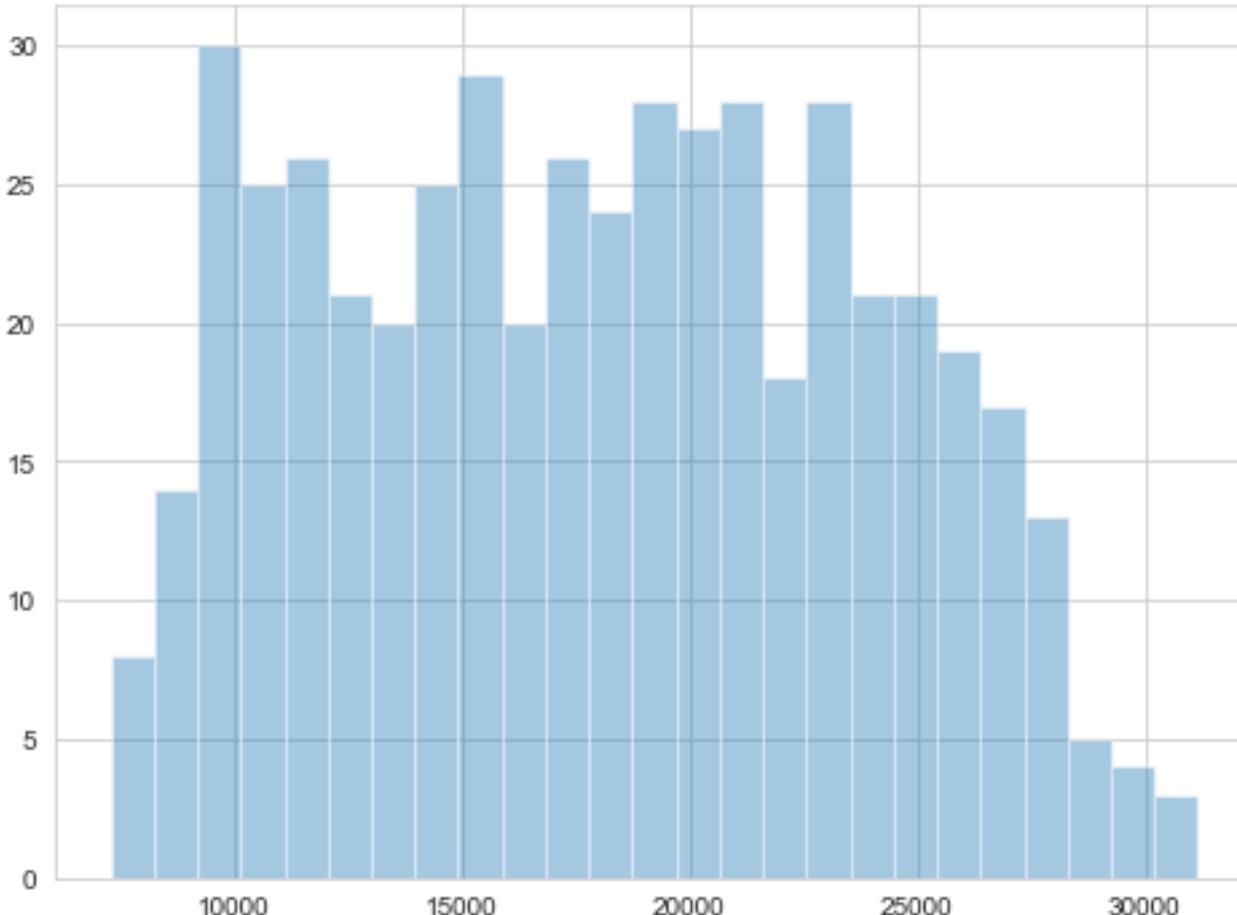
	Gender	Height_scaled	Weight_scaled	Height	Weight	Ratio
490	Female	-0.12	-1.45	168	59	12.23
491	Male	-0.12	-1.73	168	50	14.57
492	Male	0.06	1.08	171	141	16.76
493	Male	0.06	1.27	171	147	19.63
494	Male	0.06	1.27	171	147	19.63
495	Female	0.06	1.42	171	152	22.03
496	Female	0.06	1.51	171	155	23.46
497	Female	-0.06	-1.61	169	54	27.86
498	Male	0.00	0.28	170	115	81.27
499	Female	0.00	1.54	170	156	451.50

身高 , 体重比  
( 标准化后 )

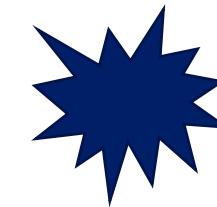


# 直方图 : Bad case

频次统计 : 组合特征 体重\*身高 在不同区间统计



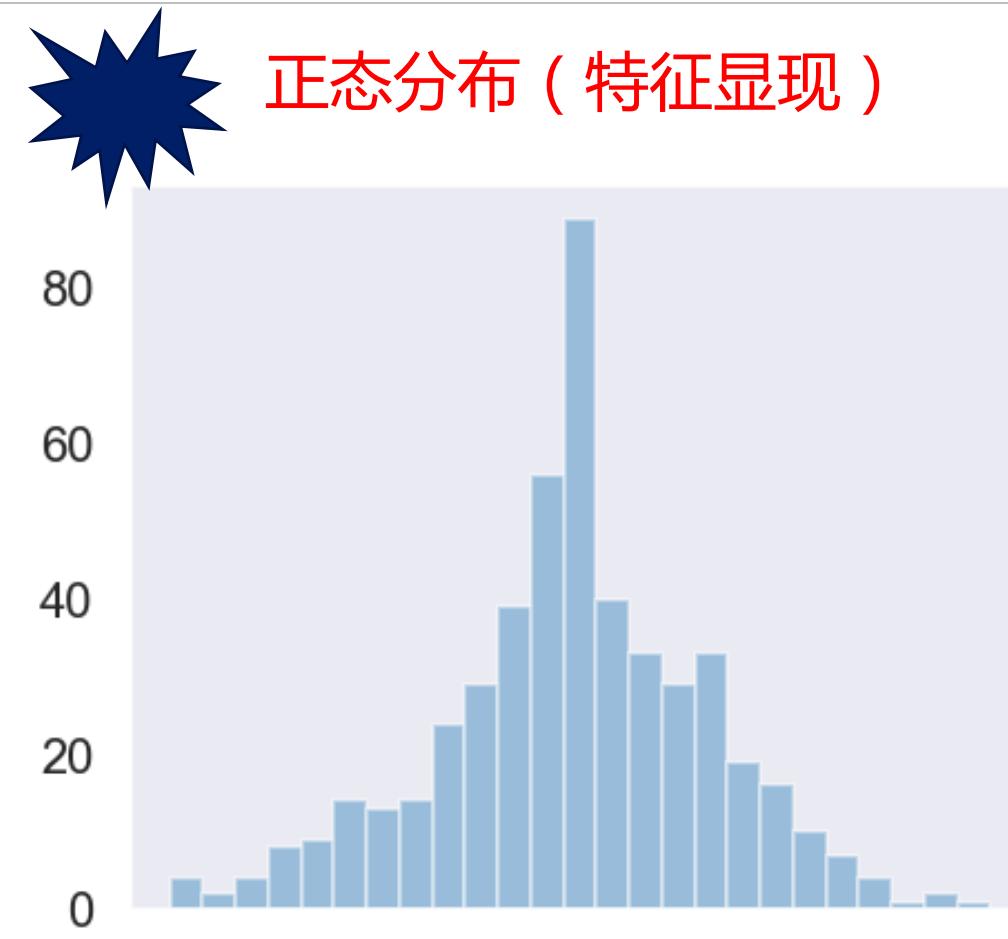
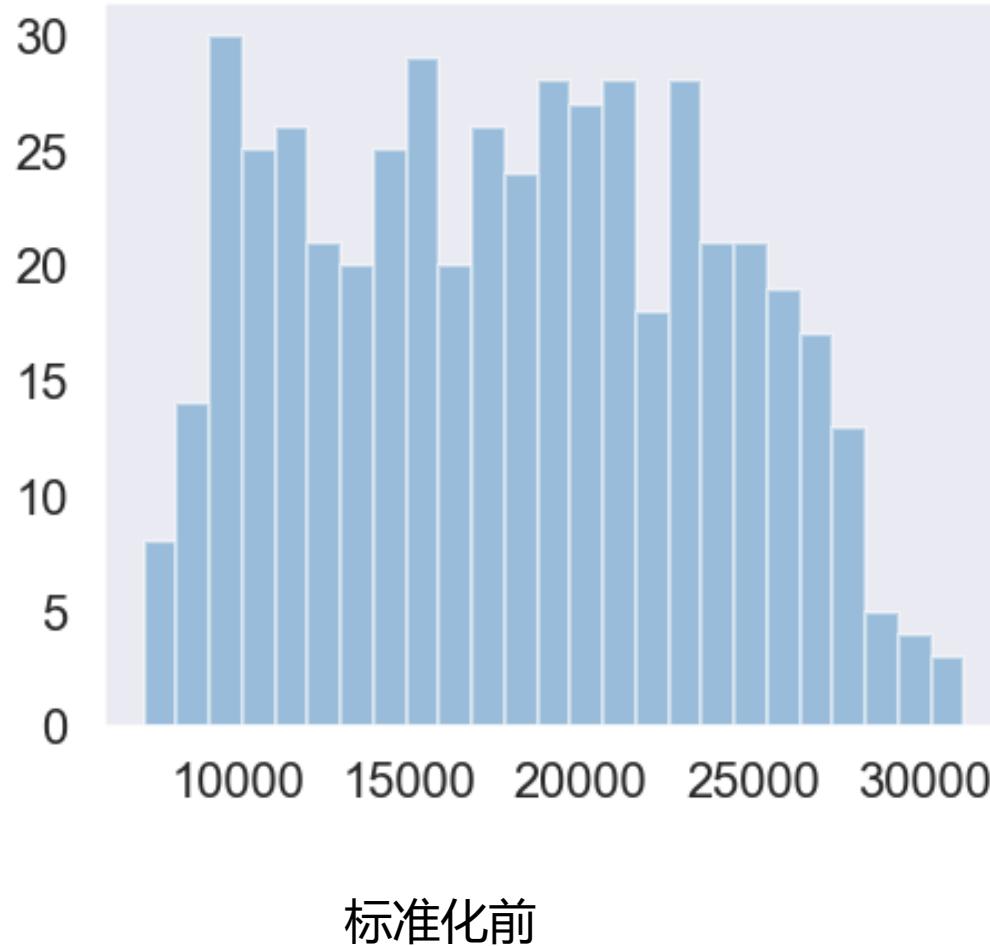
```
sns.distplot(df3['Height']*df3['Weight'], bins=25)
```



没发现什么规律

# 直方图 : Good case

频次统计 : 体重\*身高 在不同区间统计

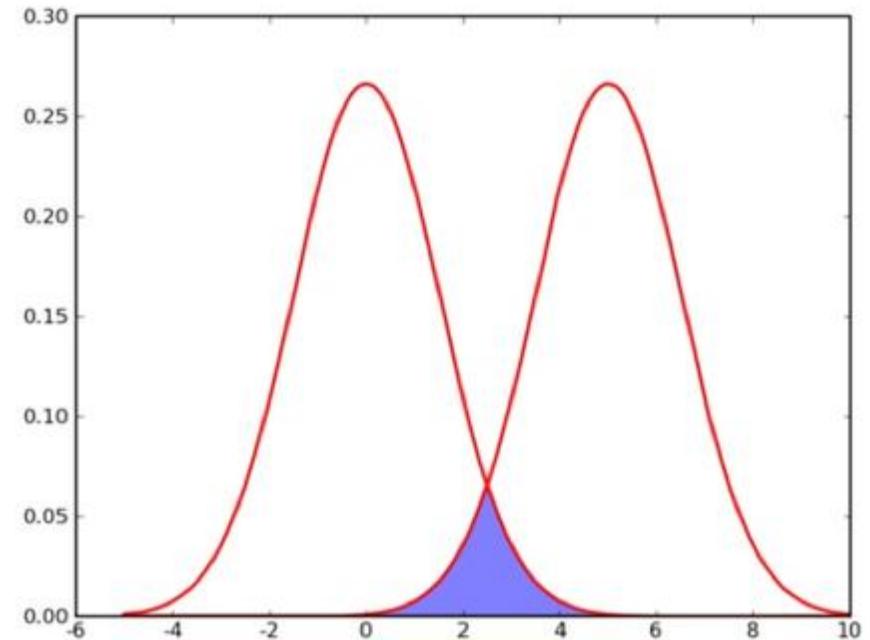


正态分布 ( 特征显现 )

# 正太分布的应用

---

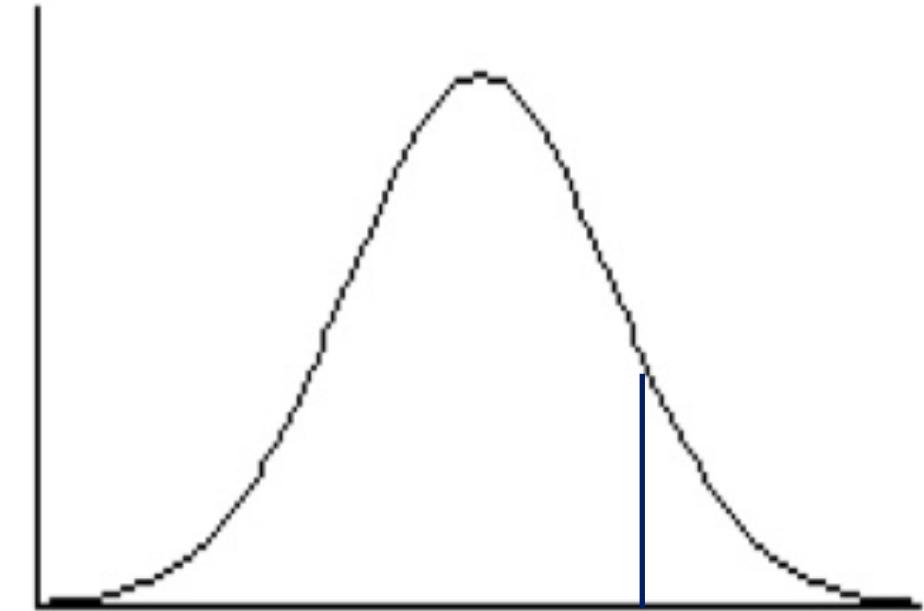
- 样本质量检验
- 发现新特征
- 模型本身
- .....



# 正太分布的应用：出试卷

---

- 举例：
- 趋势出题目，难易程度怎么才算合理？
  - 太容易，起不到筛选效果
  - 太难，有人可能做过类似的，不公平



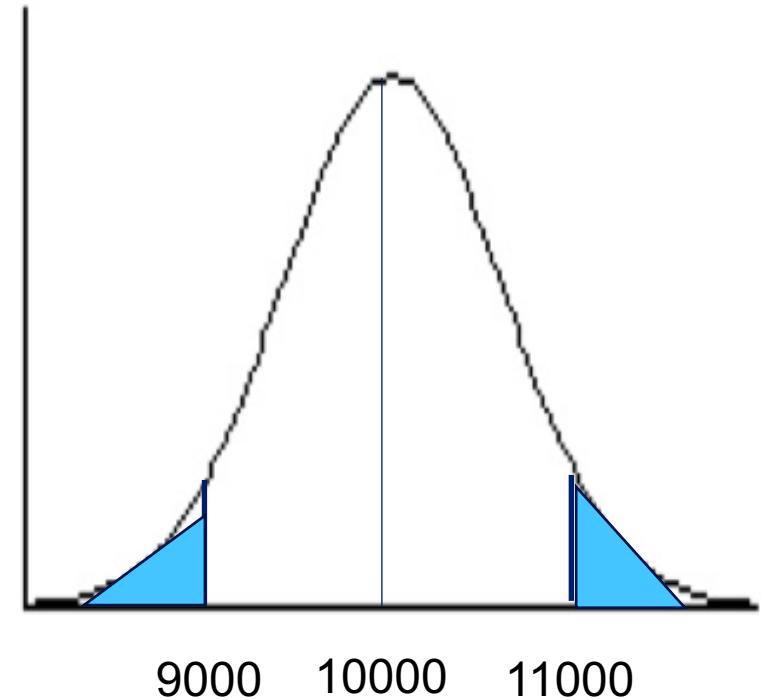
题库组题->试做->平均分->正态分布

# 正太分布的应用：有效性检验

---

举例：

- 客户平均每天1万封垃圾邮件；
- 新feature上线后，平均每天9000封垃圾邮件；
- 新feature是否有效？

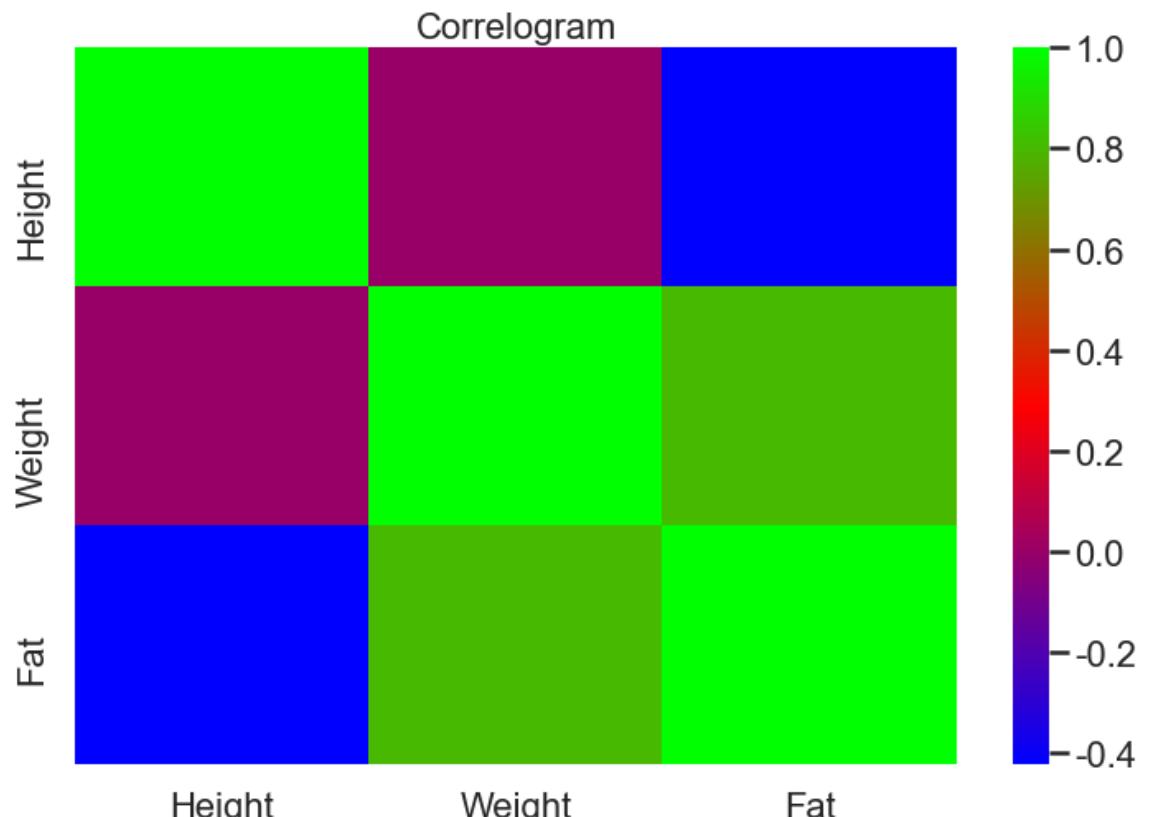


# 热力图

观察多维数据的分布区间

样本的关联系数矩阵

	Height	Weight	Level
Height	1.00	0.00	-0.42
Weight	0.00	1.00	0.80
Level	-0.42	0.80	1.00



```
sns.heatmap(corr,cmap='brg', annot=False)
```

# 小结

---



- 样本、特征、数据集
- 数据清洗
- 可视化和常见统计图
- 数据转换
- 通过标准化发现异常值
- 正态分布的应用

# Thanks

2020-8-15

