数据科学导论实验 5：ESI 学科分析与排名预测

姓名：杨云天

学号：**10245501405**

一、 基于多学科排名的全球高校聚类分析

1. 构建排名特征张量

从 22 个 ESI 学科中提取每所高校的排名（依据引用数 Cites）。其中，入榜高校使用其实际排名；未入榜高校引入惩罚机制，填入该学科的最大排名*2（即该学科总机构数*2）。

没有惩罚机制时，未入榜学科被视为"缺失值"，聚类算法可能产生偏差，综合性大学与专业院校难以公平比较。添加惩罚机制后，所有高校在所有学科都有排名，聚类基于完整数据矩阵，能更好识别“全面发展型”vs“特色突出型”高校。

- 构建特征张量 df_features：9990 个机数 × 22 个学科（下图为部分）

	Agricultural Sciences	Biology & Biochemistry	Chemistry	Clinical Medicine	Computer Science	Economics & Business
CHINESE ACADEMY OF SCIENCES	1	3	1	188	1	60
CHINESE ACADEMY OF AGRICULTURAL SCIENCES	2	175	357	4360	826	1086
UNITED STATES DEPARTMENT OF AGRICULTURE (USDA)	3	179	756	1476	738	1086
CHINA AGRICULTURAL UNIVERSITY	4	215	398	3470	273	1086
INRAE	5	54	348	696	591	320

2. 使用 k-means 算法聚类，k = 5

2.1. 数据标准化

- 使用库与方法：scikit-learn 的 StandardScaler 类，实现 Z-score 标准化
- 目的与原理：
  - 将各特征转换为均值为 0、标准差为 1 的标准正态分布
  - 计算公式：

$$z = \frac{x - \mu}{\sigma}$$

- 必要性：避免某些数值较大的特征主导聚类过程

2.2. K-means 算法

- 利用 scikit-learn 的 KMeans 类，基于划分的无监督学习算法
- 参数：
    - n_clusters=5：聚类数 k = 5
    - random_state=42：设置随机种子，确保结果可重现
    - n_init=20：使用不同质心种子运行算法的次数，避免局部最优

3. 聚类结果分析

3.1. 簇规模统计

```
各簇高校数量:
  簇 0: 5100 所高校
  簇 1: 417 所高校
  簇 2: 792 所高校
  簇 3: 3198 所高校
  簇 4: 483 所高校
```

3.2. 华东师范大学定位

```
EAST CHINA NORMAL UNIVERSITY:
所属簇: 1
```

4. PCA 降维实现聚类可视化

4.1. PCA（主成分分析）原理

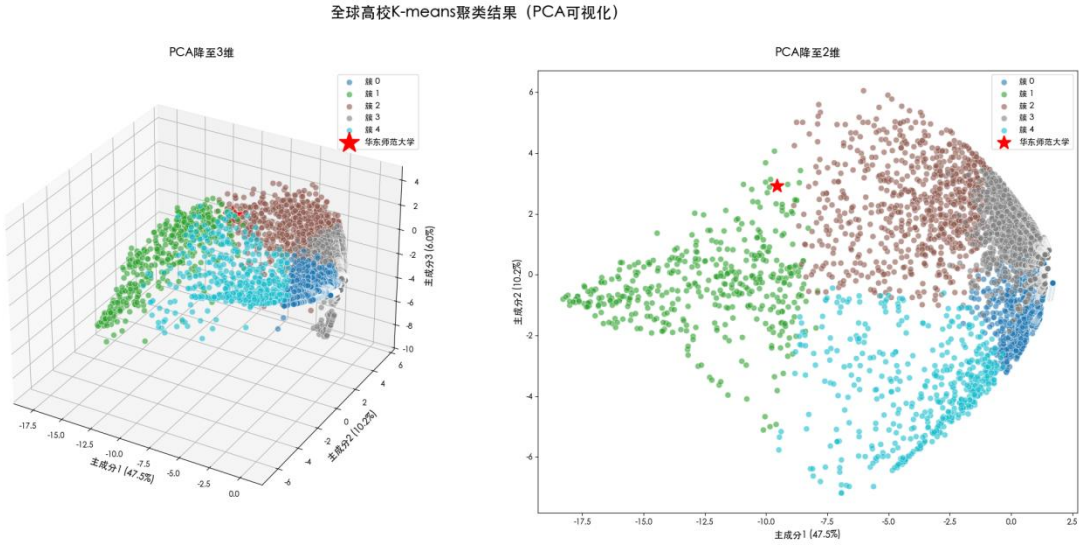
PCA（主成分分析）是一种无监督的线性降维方法。其核心思想是通过线性变换，将高维数据映射到低维空间，同时尽可能保留数据中的大部分变异信息。

4.2. PCA 降维的作用

- 数据简化：高维数据往往包含冗余信息，PCA 降维可以减少数据的维度，使数据更易于处理和可视化，就像图中把高校相关的高维特征数据降维到 3 维和 2 维后，能在 3D 空间和 2D 平面中直观展示聚类结果。
- 去噪：在降维过程中，那些方差较小的成分会被舍弃，从而在一定程度上去除数据中的噪声，保留主要信息。
- 特征提取：从原始的众多特征中提取出最具代表性的主成分，这些主成分能够概括原始数据的大部分特征，有助于后续的数据分析，让分析过程更高效且更聚焦于关键信息。

4.3. PCA 降维效果

4.3.1. 可视化



4.3.2. 坐标含义：

- 左图（PCA 降至 3 维）：

  - 横轴：主成分 1：47.5%表示该主成分能够解释原始数据总方差的 47.5%，它是从原始高维数据中提取出的、包含信息最多的一个维度。

  - 纵轴：主成分 2，10.2%表示其解释原始数据总方差的比例为 10.2%，是仅次于主成分 1、能解释较多剩余信息的维度。

  - 竖轴：主成分 3，6.0%表示解释原始数据总方差的比例为 6.0%，是三个主成分中解释方差比例相对较小的，但仍能补充部分信息。

- 右图（PCA 降至 2 维）同理

4.3.3. PCA 降维可视化结果分析

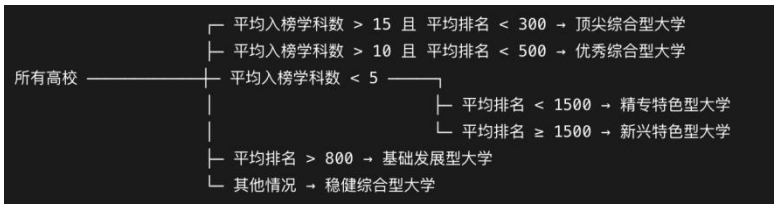
从图中可以看到，华东师范大学在 3 维空间和 2 维空间中，都处于绿色簇（簇 1）的分布区域内。这表明，基于 K - means 聚类以及 PCA 降维后的特征，华东师范大学与绿色簇内的其他高校在相关特征上具有较高的相似性，所以被归为同一类簇。

5. 各簇特征分析

5.1. 总体情况统计

簇	高校数	平均入榜学科数	入榜学科平均排名	最佳排名
0	5100	1.5	3032	1
1	417	19.3	373	1
2	792	9.7	985	4
3	3198	2.1	1368	3
4	483	8.8	771	5

5.2. 高校类型决策树



5.3. 各簇类型推断

```
簇 0: 新兴特色型大学
簇 1: 优秀综合型大学
簇 2: 基础发展型大学
簇 3: 精专特色型大学
簇 4: 稳健综合型大学
```

6. 与华东师范大学相似的高校

6.1. 思路

计算与华东师范大学欧氏距离最相近的 5 所高校

欧氏距离公式：

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

6.2. 结果

```
与华东师范大学最相似的 5 所高校（簇 1 - 优秀综合型大学）:
1. UNIVERSITY OF BASQUE COUNTRY
相似度距离: 2.596
2. UNIVERSITY OF ELECTRONIC SCIENCE & TECHNOLOGY OF CHINA
相似度距离: 3.077
3. SOUTHWEST UNIVERSITY - CHINA
相似度距离: 3.635
4. UNIVERSITY OF MILANO-BICOCCA
相似度距离: 4.250
5. SOUTHEAST UNIVERSITY - CHINA
相似度距离: 4.348
```

二、 华东师范大学学科画像探索性分析

1. 获取华东师范大学各个学科排名

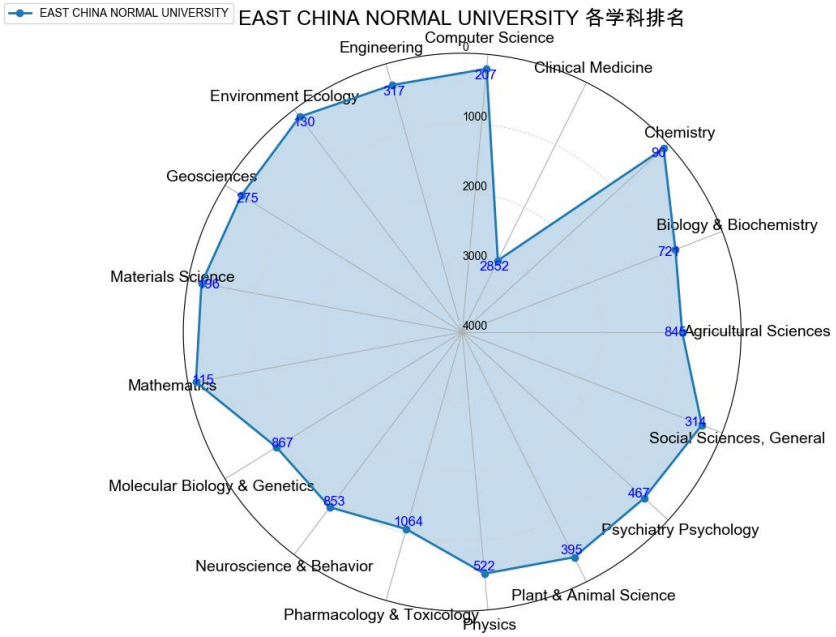
subject_names: 保存学科名称
subject_ranks: 保存学科排名
rank_df: dataframe（右图）

	Subject	Rank
0	Agricultural Sciences	845
1	Biology & Biochemistry	721
2	Chemistry	90
3	Clinical Medicine	2852
4	Computer Science	207
5	Economics & Business	/
6	Engineering	317
7	Environment Ecology	130
8	Geosciences	275
9	Immunology	/
10	Materials Science	196

	Subject	Rank
11	Mathematics	115
12	Microbiology	/
13	Molecular Biology & Genetics	867
14	Multidisciplinary	/
15	Neuroscience & Behavior	853
16	Pharmacology & Toxicology	1064
17	Physics	522
18	Plant & Animal Science	395
19	Psychiatry Psychology	467
20	Social Sciences, General	314
21	Space Science	/

2. 学科分布雷达图

雷达图可以将多个指标整合在一个二维平面的雷达形状图表中，每个指标对应雷达的一条轴，数据点在各轴上的位置连接成多边形，方便直观地比较不同对象在多个指标上的表现，能清晰展现出对象的优势指标、劣势指标以及整体的均衡性等情况。



- 未上榜学科：

Economics & Business, Immunology, Microbiology, Multidisciplinary, Space Science

- 优势学科：

Chemstry (90), Mathematics (115), Environment Ecology (130)

- 相对劣势学科：

Clinical Medicine (2852), Pharmacology & Toxicology (1064),
Molecular Biology & Genetics (867)

3. 各学科实力矩阵分析

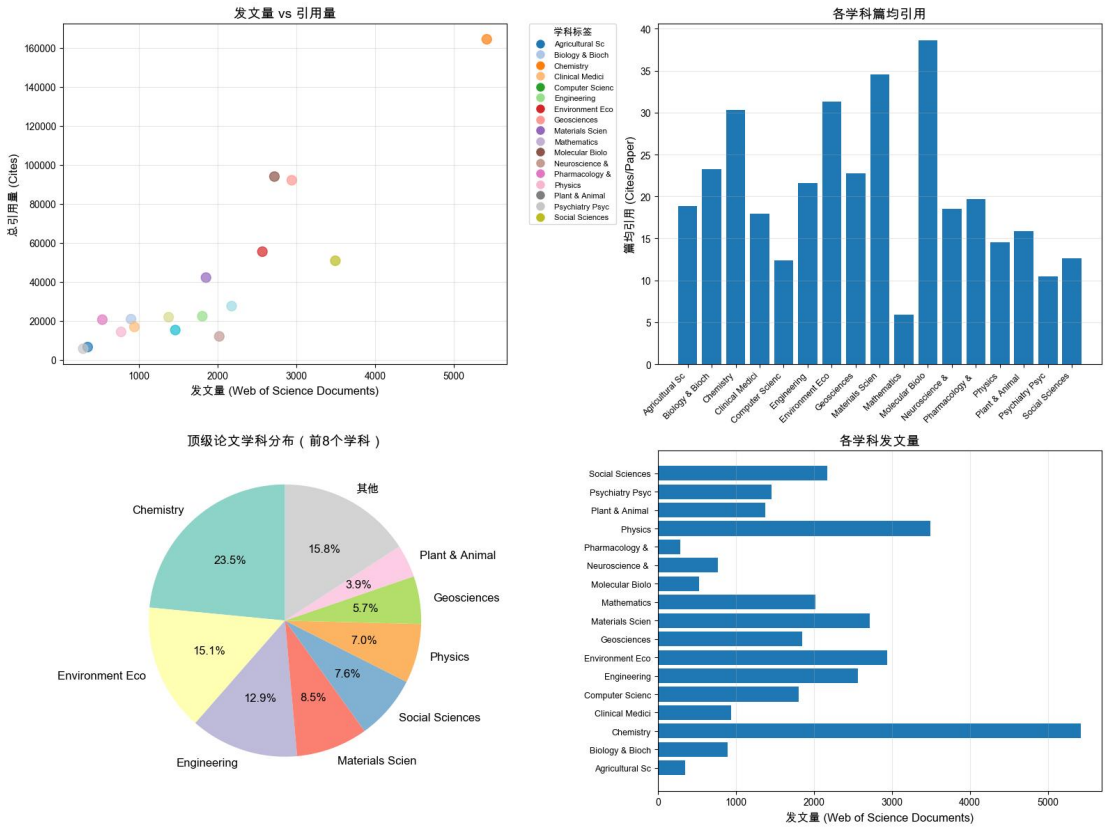
分析 ESI 学科排名中的四项指标：

- 发文量（Web of Science Documents）：指在 Web of Science 数据库中，各学科发表的论文数量，反映学科的研究产出规模。

- 总引用量（Cites）：是各学科发表的论文被引用的总次数，体现学科研究成果的影响力和受关注程度。

- 篇均引用（Cites/Paper）：总引用量除以发文量得到的平均值，用于衡量单篇论文的平均影响力，可反映学科研究质量的平均水平。

- 顶级论文学科分布：展示不同学科在顶级论文中所占的比例，能体现各学科在高水平研究成果方面的贡献情况。



4. 制定学科打分机制

4.1. 评分权重参数与打分公式

weight_docs = 0.3

weight_cites = 0.3

weight_top_papers = 0.4（看重高质量论文）

综合得分 = 发文量归一化 * 0.3 + 引用量归一化 * 0.3 + 顶级论文数归一化 * 0.4

4.2. 各学科评分（仅上榜学科）

	Subject	Docs_Score	Cites_Score	Top_Papers_Score	Composite_Score
0	Chemistry	1.000	1.000	1.000	1.000
1	Environment Ecology	0.543	0.560	0.643	0.588
2	Materials Science	0.502	0.572	0.363	0.467
3	Engineering	0.474	0.337	0.548	0.462
4	Physics	0.645	0.309	0.299	0.406
5	Social Sciences, General	0.401	0.167	0.325	0.301
6	Geosciences	0.341	0.256	0.242	0.276
7	Computer Science	0.333	0.136	0.159	0.204
8	Mathematics	0.373	0.073	0.140	0.190
9	Plant & Animal Science	0.254	0.133	0.166	0.182
10	Biology & Biochemistry	0.165	0.127	0.115	0.134
11	Psychiatry Psychology	0.269	0.093	0.045	0.126
12	Clinical Medicine	0.173	0.103	0.076	0.113
13	Neuroscience & Behavior	0.142	0.087	0.045	0.087
14	Molecular Biology & Genetics	0.098	0.125	0.038	0.082
15	Agricultural Sciences	0.064	0.040	0.025	0.041
16	Pharmacology & Toxicology	0.053	0.035	0.032	0.039

三、 学科排名预测模型

1. 数据说明

学科总数：22 个
数据特征：
- Web of Science Documents
- Cites
- Top Papers
- 机构在该学科的全球排名

- 忽略 Cites/Paper 原因：篇均引用可以通过 Cites 与 Web of Science Documents 计算获得，相关性极高，无需选取。

2. 数据集划分

训练集：60%
验证集：20%
测试集：20%
注: 每个学科的数据在划分前进行随机打乱（random_state=42）

3. 数据预处理

3.1. 数据归一化
- 排名：反向归一化到 [0,1]，排名越靠前（数字越小），归一化值越接近 1
- 特征：正向归一化到 [0,1]
- 使用 sklearn 的 MinMaxScaler
- 归一化原因：量纲差异会导致数值大的特征在计算中占据主导地位，例如引用量（万级）会掩盖顶级论文数（百级）。对于回归模型，量纲差异会影响距离计算或系数估计，导致结果偏向数值范围大的特征，无法客观反映各特征的真实重要性。

3.2. 反归一化:
将预测的归一化值转换回实际排名。

4. 模型说明

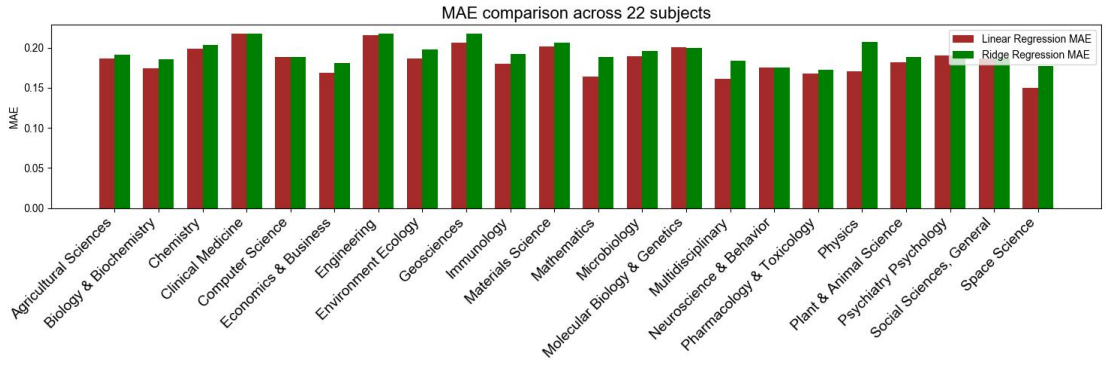
4.1. 线性回归 Linear Regression
- 参数数量: 4 个（3 个系数 + 1 个截距）
- 特点：模型结构简单，计算速度快。
- 适用场景：变量与结果近似存在线性关系，特征之间无强相关性时效果较好。

4.2. 岭回归 Ridge Regression
- 特点：在线性回归的基础上引入 L2 正则化项，有效限制模型系数大小，提升泛化能力。
- 适用场景：特征数量较多、变量间存在相关性，或者数据容易过拟合时。

- 参数搜索范围: 10^-3 到 10^3 （20 个候选值）

5. 评价指标

5.1. MAE：平均绝对误差
- 含义：预测排名与实际排名的平均差距
- 越小越好



5.2. 两种模型 MAE 分析

从图中看，部分学科两种模型 MAE 持平，多数学科线性回归略优于岭回归，原因可能如下：

- 数据特性：这些学科的特征数据不存在严重的多重共线性问题。线性回归的核心假设是特征间线性无关，当数据满足这一条件时，线性回归能直接通过最小二乘法高效且准确地拟合数据，无需额外的正则化干预。而岭回归的正则化（L2 惩罚项）在此种无严重多重共线性的场景下，反而可能因对系数的“过度收缩”，导致模型对数据的拟合程度略微下降，进而使 MAE 不如线性回归。

- 模型特征：线性回归是更简洁的模型，在数据本身适配线性模型且无复杂干扰的情况下，越简洁的模型往往越能直接捕捉数据的核心规律，避免因引入额外参数而产生的“模型冗余”，从而在预测误差（MAE）上体现出优势。

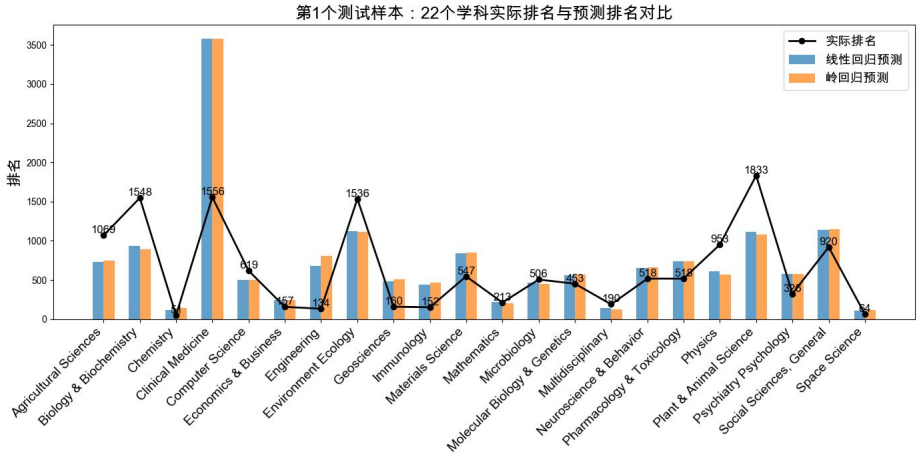
6. 预测结果

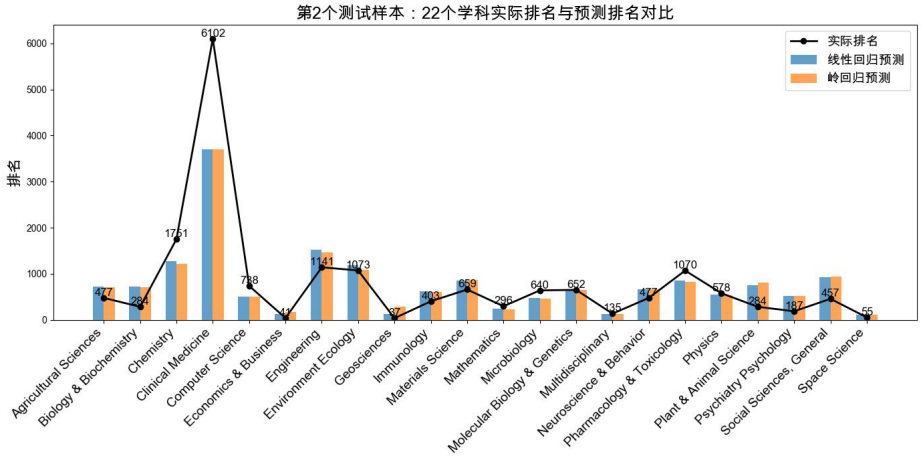
6.1. 输出
predictions_lr：列表，包含 22 个元素
- 每个元素是一个学科的线性回归预测结果
- 形状： （测试集样本数, 1）
- 数值：归一化值

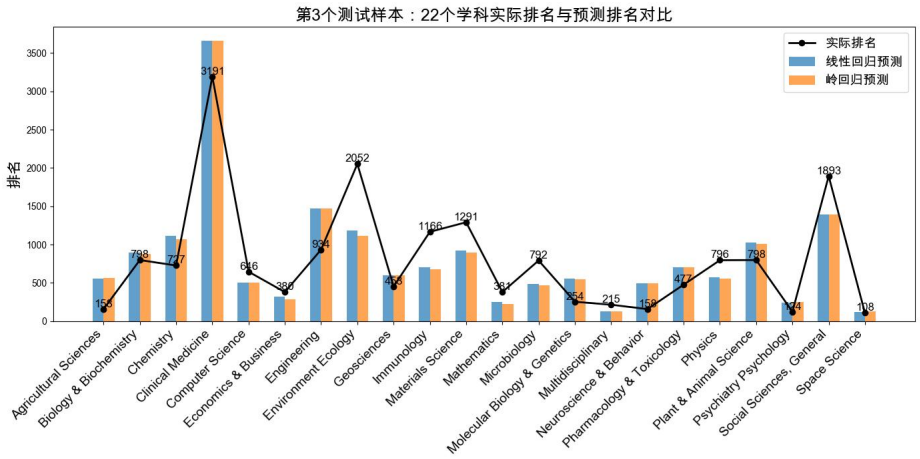
predictions_ridge：列表，包含 22 个元素
- 每个元素是一个学科的岭回归预测结果
- 形状： （测试集样本数, 1）
- 数值：归一化值

6.2. 输出数据处理
- 反归一化，得到排名（精确到整数）
- 若排名出现负数，则改为 1

6.3. 三个测试样本可视化







6.4. 分析

部分学科预测较准确：像 Computer Science、Economics & Business 等学科，实际排名和两种回归方法的预测排名较为接近，说明在这些学科上，线性回归和岭回归的预测效果较好。

部分学科预测偏差大：像 Clinical Medicine、Environment Ecology 等学科，线性回归和岭回归预测排名相对真实排名高许多，与实际值仍有一定差距。