










固定一个月，分析指标间的线性相关性

 Created By	
 Stakeholders	
 Status	
 Type	Technical Spec
 Created	@June 9, 2022 11:16 AM
 Last Edited Time	@June 9, 2022 6:01 PM
 Last Edited By	

[目标](#)

[解决方案](#)

[定义](#)

[步骤](#)

[问题](#)

目标

- 相关性上的关联性：相关性上：线性相关、指数相关等。容易解读。
- 固定一个月，输入样本集合，挖掘出指标间的线性相关性。

解决方案

定义

- 样本分类 `GovLevel`
 - 全部 `ALL = -1`
 - 国（仅一个样本） `NATION = 0`
 - 省：省 `PROVINCE = 1`
 - 市：市 & 直辖市 `CITY_LEVEL = 2`
 - 区：区县 & 省辖县 `DISTRICT_LEVEL = 3`
- 指标 `prod_indexes_info.json`
 - 仅使用“已经上线”的指标
- 月 `month_int_to_str`
 - “2001-01”的id是1
 - “2020-12”的id是240

步骤

1. 下载并缓存一个月的全部数据

```
def save_one_month(month: int) -> None:
```

2. 读取一个月的全部数据

```
def read_one_month_values(month: int,  
                           govs_ids: list = mda.GOV_S_IDS,  
                           prod_indexes_ids: list = mda.PROD_INDEXES_IDS  
                           ) -> pd.DataFrame:
```

	gov_id	month	42367	42368
0	0.0	252.0	84673.8069	607218.0
1	1.0	252.0	3795.7635	26443.0
2	2.0	252.0	157.9043	1101.0
3	3.0	252.0	121.5172	849.0
4	4.0	252.0	265.7773	1851.0
...
3195	3213.0	252.0	11.3032	84.0
3196	3214.0	252.0	0.9243	14.0
3197	3215.0	252.0	0.7727	13.0
3198	3216.0	252.0	0.8182	13.0
3199	3227.0	252.0	0.0000	8.0

3. 计算线性相关性

- `gov_ids_or_gov_level` 可以选择：省、市、区，或自定义gov_id数组；默认为全部样本。
- `prod_indexes_ids` 可以自定义index_id数组；默认为全部“已经上线”的指标。

```
class OneMonthCorrAnalyzer():
    def __init__(self, month: int,
                  gov_ids_or_gov_level=mda.GovLevel.ALL,
                  prod_indexes_ids=mda.PROD_INDEXES_IDS):
        # ...
        analyzer = OneMonthAnalyzer(252)
```

```
analyzer.corr
```

	42367	42368	42385	42386
42367	NaN	0.99998	0.978592	0.982164
42368	NaN	NaN	0.978596	0.982290
42385	NaN	NaN	NaN	0.999618
...

```
analyzer.sorted_corr
```

139303	139113	1.0
140134	139021	1.0
139221	139035	1.0
		...
57585	64668	-1.0

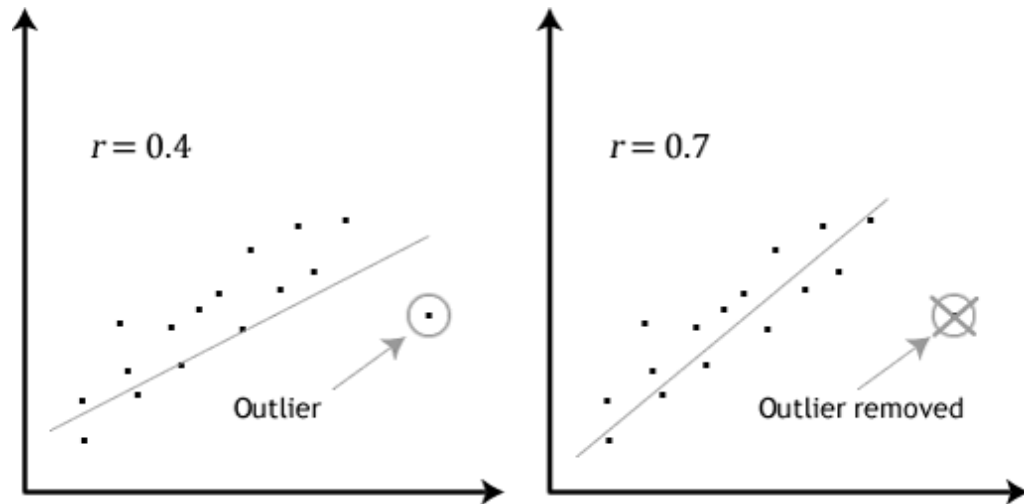
4. 筛选出“好的”相关的指标组

- 筛选的动机：关联分析中，极可能存在找到大量的关联性强的特征组，但是其中的绝大多数对于应用场景来讲是没有价值的，所以需要过滤步骤。过滤步骤的解决思路后面补。
- 条件：
 - 绝对值大于某个数值，如0.8
 - 在两个指标上都有效的样本数量
 - 在两个指标上都有效的样本数量 / 总样本数量
 - 两个指标name的最大公共子序列的长度
 - 两个指标name的最大公共子序列的长度 / 两个指标name的长度之和
 - 两个指标name的最大公共子串的长度
 - 两个指标name的最大公共子串的长度 / 两个指标name的长度之和

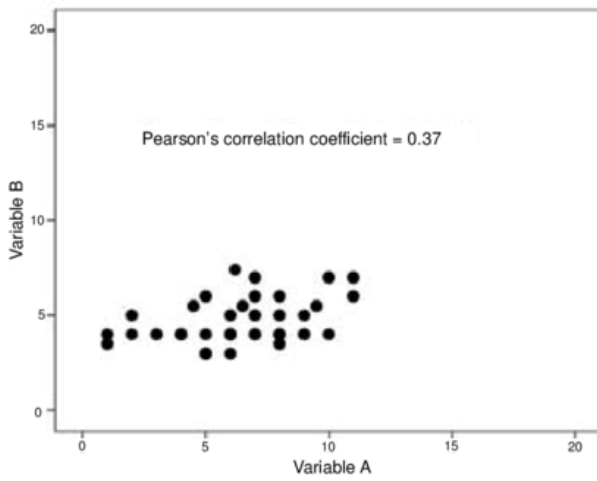
问题

- 使用不同的相关性计算方法？
 - pearson [目前]
 - kendall
 - spearman
- 再细分gov类型？
- 多个指标组合起来的特征？

- 跨越多个月？
- 怎么处理离群值？



a)



b)

