



# 6月总结报告

Created By	
Stakeholders	
Status	
Type	Technical Spec
Created	@June 29, 2022 10:02 AM
Last Edited Time	@June 29, 2022 12:34 PM
Last Edited By	

## 1. 项目的目标

### 1.1. 目标综述

#### 6月目标

#### 7月目标

### 1.2. 计划简述

## 2. 阶段性结果

### 2.1. 目前进度

### 2.2. 流程

### 2.3. 典型结果

### 2.4. 存在的问题

### 2.5. 下一步

## 3. 支撑内容

### 3.1. 功能函数

### 3.2. 参数管理工具

任务管理

文件管理

结果可视化

数据的历史轨迹可视化

批量执行任务

### 3.3. 代码细节

## 4. 【附录】 总体计划

目标：

特征之间的关联性：

关联性定义：

交付内容：

注意事项：

---

# 1. 项目的目标

## 1.1. 目标综述

开发出一套**通用的工具**，可以针对一组样本集合和一组特征集合，提取出在该样本集合下相关联的特征组/对，并且输出每组/对关联性的强弱。

- 筛选条件偏向苛刻，宁可结果数量少，也要保证输出的关联性是真实的。

## 6月目标

- 相关性上的关联性
- 排序名次上的关联性
- 等长时间序列上的关联性

## 7月目标

- 不等长时间序列上的关联性
- 随机变量分布上的关联性
- 分箱同时出现频次上的关联性

## 1.2. 计划简述

- 使用经典方法挖掘关联性。

- 手动查看结果（列表、画图），提出对于我们的数据的特殊洞察，再特殊处理部分数据。
- 开发一个具备前端、后端、数据库的数据处理与挖掘框架，以更快、更清晰准确地执行任务。
- 继承已经写好的类，以快速地开发更高级的挖掘方法。

## 2. 阶段性结果

### 2.1. 目前进度

- 数据来源：DAAS
- 指标：daas\_index\_list中的5116个指标
- 样本：“区县”+“省辖县”（区县级）
- 关联性挖掘方法
  - Pearson相关性
  - Spearman相关性 - 排序名次
  - 等长的时间序列的相似性
    - 同时
    - 错开X个月

### 2.2. 流程

1. 固定一个时间点（时间序列则固定一个地区）
2. 筛选出“区县级”的样本（时间序列则选择一个时间段）
3. 对指标进行分类，按照指标类型处理数据
4. 去除离群值
5. 去除质量过低的列（指标）
6. 去除质量过低的行（地区）
7. 计算关联性
8. 筛选关联性强的指标对
9. 用去除质量过低的行（地区）的结果计算离散度

10. 合并：关联性结果、离散度、指标类型、指标其他信息

11. 最后在Excel中按照关联性、离散度、type、sub\_type等信息筛选

## 2.3. 典型结果

index_id1	index_id2	value	index_name1	index_name2
91787	40	0.705906467	机构专利申请量 (年度)	高新技术企业数量
91785	40	0.701502325	专利申请总量(年度)	高新技术企业数量
91791	40	0.692721321	专利授权总量(年度)	高新技术企业数量
91793	40	0.691710227	机构专利授权量 (年度)	高新技术企业数量
91788	40	0.595312821	ICT专利授权量 (年度)	高新技术企业数量
91794	40	0.58818517	ICT专利申请量 (年度)	高新技术企业数量
91785	57188	0.532236703	专利申请总量(年度)	电气机械和器材 制造业企业数量
40	57188	0.5227335	高新技术企业数量	电气机械和器材 制造业企业数量
91785	57378	0.519863098	专利申请总量(年度)	印刷和记录媒介 复制业企业数量
149710	107512	0.509286179	每万人房地产业 个体工商户数量 (大类)	每万人中介机构 相关兴趣点数量
91785	57190	0.508759881	专利申请总量(年度)	计算机、通信和 其他电子设备制 造业企业数量
91791	57188	0.508459229	专利授权总量(年度)	电气机械和器材 制造业企业数量
91787	57188	0.508136387	机构专利申请量 (年度)	电气机械和器材 制造业企业数量

index_id1	index_id2	value	index_name1	index_name2
70577	62467	0.505779712	近2年信息传输、软件和信息 技术服务业企业 数量增长（大 类）	近2年信息技术 产业企业数量增 长（大类）
141252	60864	0.503633327	本年度新增注册 专业技术服务业 企业数量	软硬件技术产业 企业数量
91787	57394	0.501225415	机构专利申请量 (年度)	专用设备制造业 企业数量
91785	57394	0.500651556	专利申请总量(年 度)	专用设备制造业 企业数量

## 2.4. 存在的问题

- 某些type的指标（如POI）结果不理想
- 参数选择依赖经验，不一定是最佳的值
- 样本的选择可以更细致（如东部地区、人口大于100万）

## 2.5. 下一步

- ☐ 进一步针对部分指标进行特殊处理
- ☐ 调参
- ☐ 多线程
- ☐ 开发更高级的关联性挖掘（7月目标）

# 3. 支撑内容

## 3.1. 功能函数

- 去除离群值

```
def remove_outliers_series_with_median_deviation(
    series: pd.Series, outlier_sd_threshold: float
) -> pd.Series:
    dropped_na = series.dropna()
    deviation = np.abs(dropped_na - np.median(dropped_na))
```

```

median_dev = np.median(deviation)
scaled_dev = (deviation / median_dev) if median_dev else None

result = (
    dropped_na[scaled_dev < outlier_sd_threshold]
    if scaled_dev is not None
    else pd.Series(index=series.index, dtype=np.float64)
)

return result

```

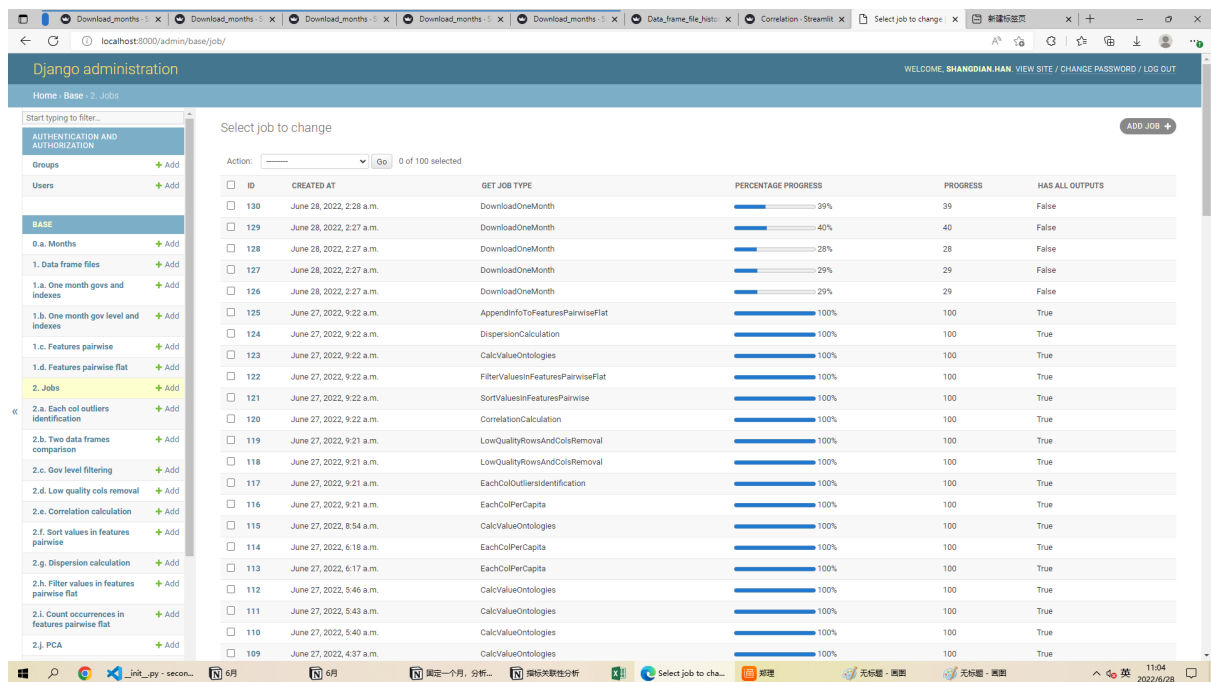
- 相关性
  - pearson (相关性)
  - kendall
  - spearman (时间序列)
- 离散度
  - Standard deviation
  - Interquartile range
  - Coefficient of variation
  - 【目前使用】 Quartile coefficient of dispersion

$$QCD = \frac{Q3 - Q1}{Q1 + Q3}$$

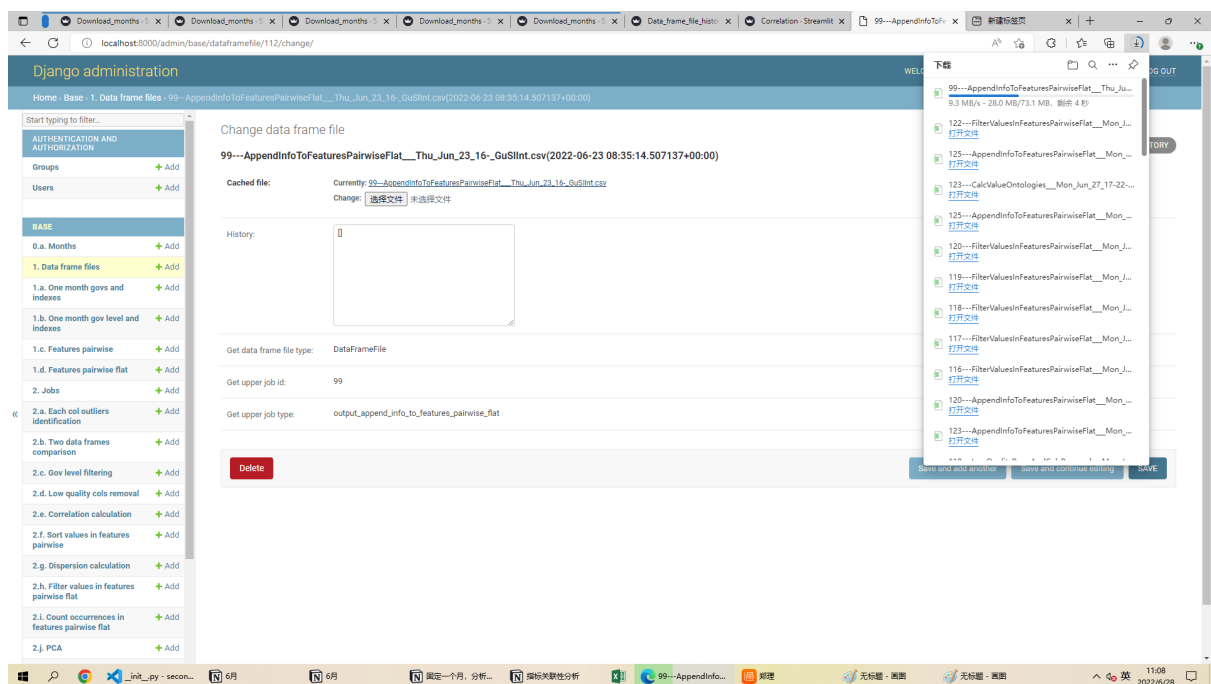
## 3.2. 参数管理工具

- 数据处理与挖掘的任务管理：查看进度、查看参数、添加、修改、删除
- 文件管理：最终输出、中间结果（缓存、方便内部开发）
- 结果可视化：相关性、排序名次、时间序列
- 数据的历史轨迹（原始数据、方法、参数、中间结果）可视化
- 批量执行任务：多线程、读取缓存

## 任务管理

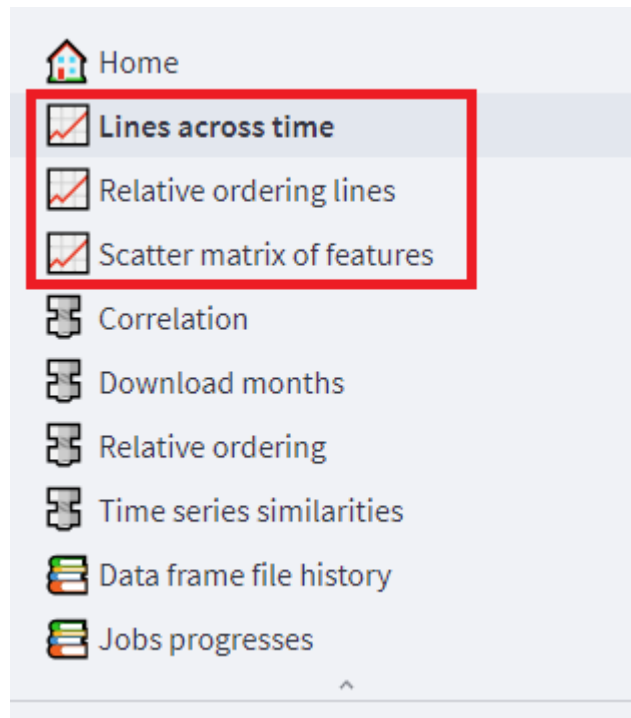


## 文件管理

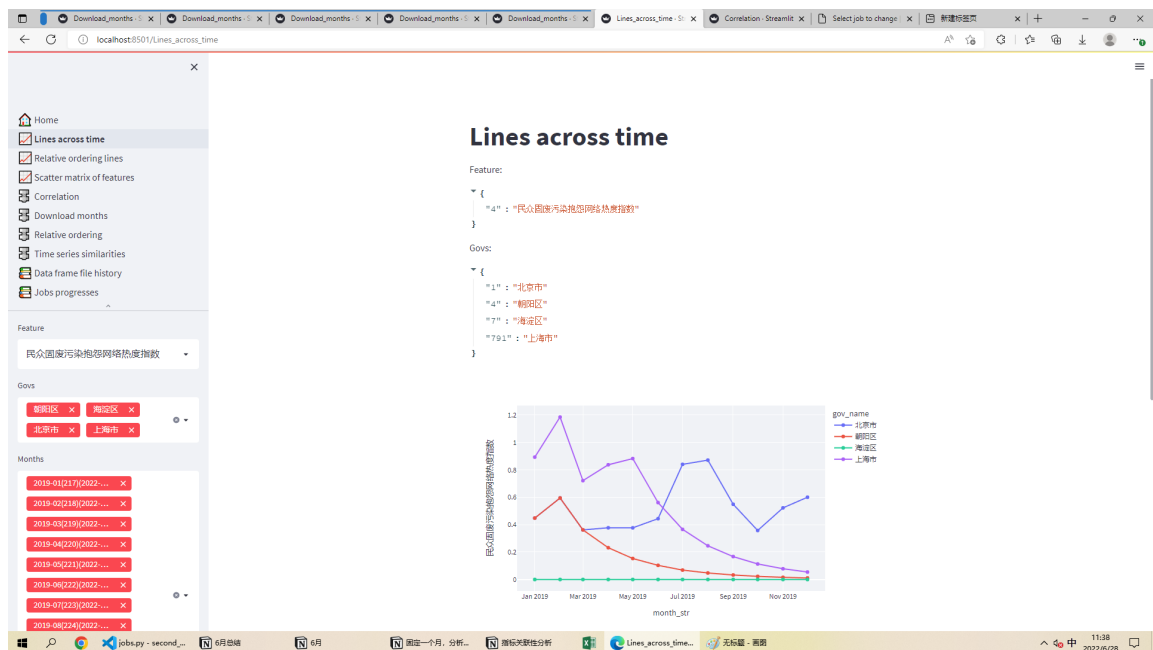


## 结果可视化

- 图表类型



## • 时间序列

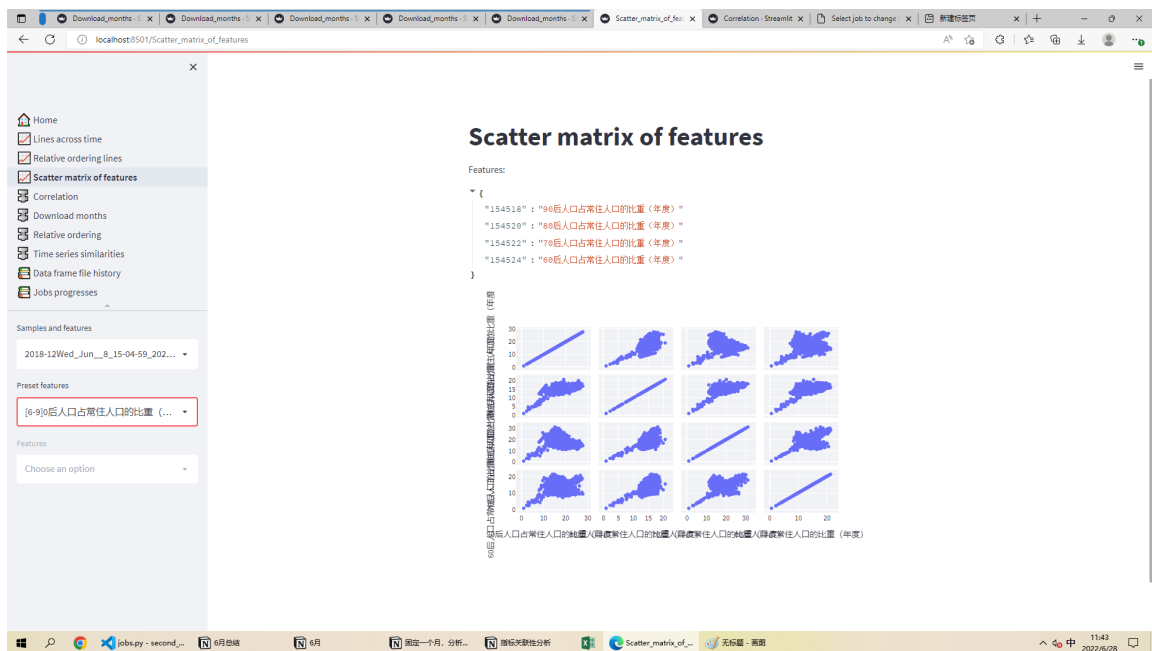


## • 排序名次

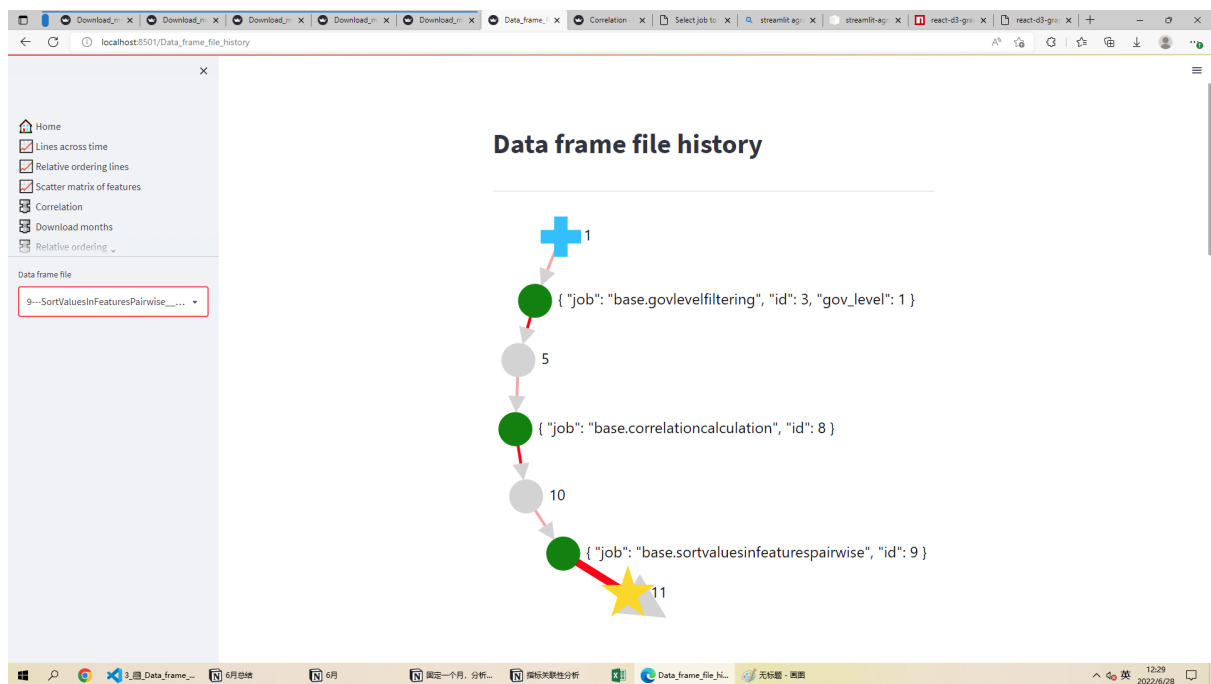
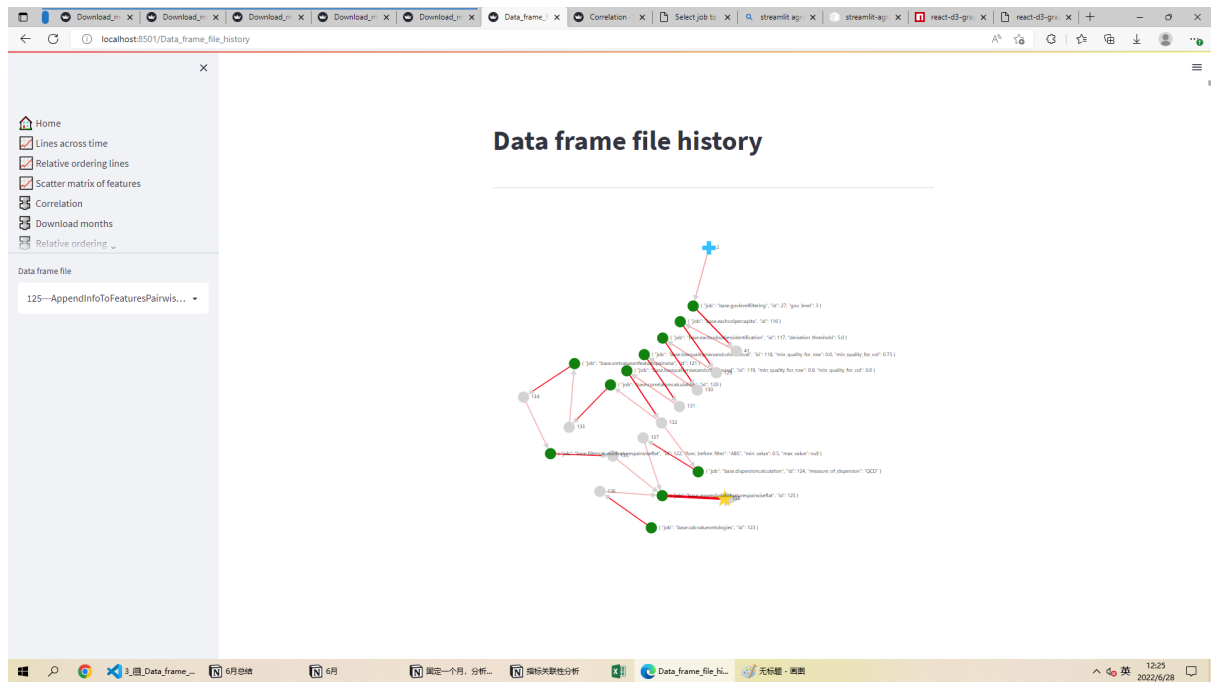




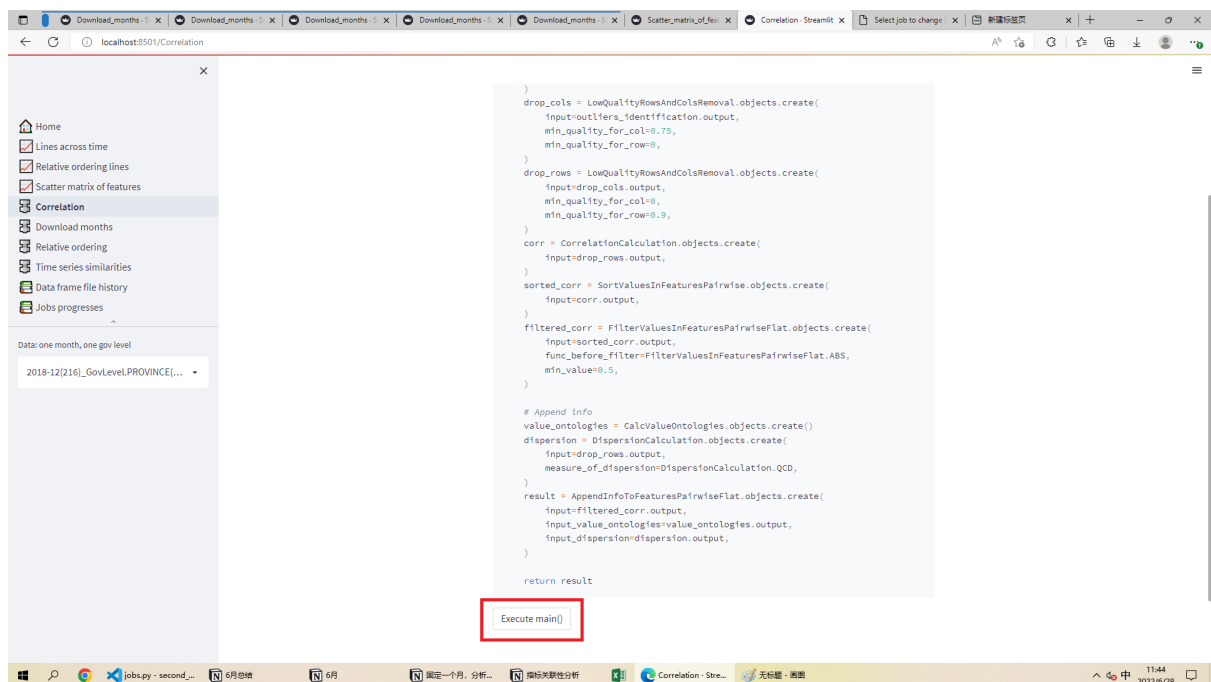
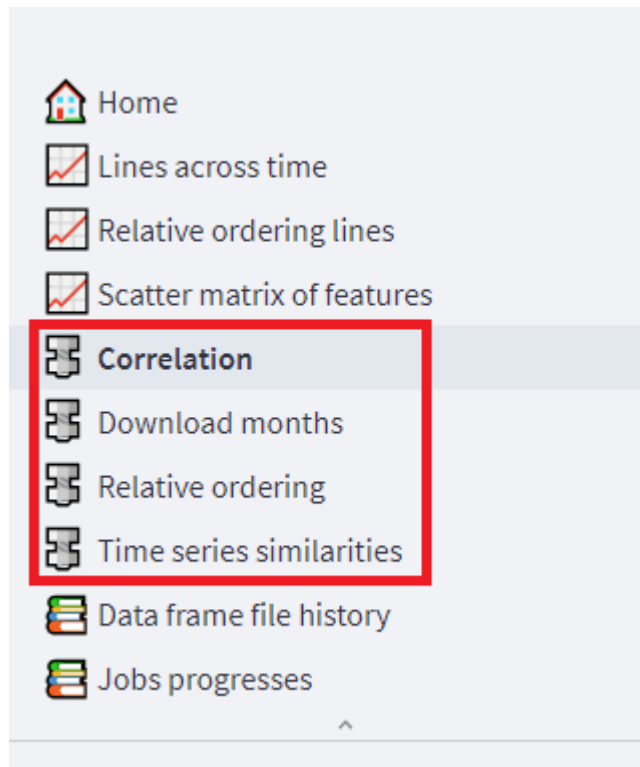
## • 相关性



## 数据的历史轨迹可视化



## 批量执行任务



### 3.3. 代码细节

- 按照指标类型，处理数据
  1. 数量 → 除以人口
  2. 比例 → 不变

3. 增量 → 除以人口
  4. 增幅 → 不变
  5. 分数 → 不变
  6. 人均数量 → 不变
  7. 价格 → 不变
  8. 全国占比 → 除以人口
- 去除离群值 ( `deviation_threshold=5` )
  - 去除质量过低的列 ( `min_quality_for_col=0.75` )
  - 去除质量过低的行 ( `min_quality_for_row=0.9` )
  - 筛选相关性强的指标对 ( `func=ABS, min_value=0.5` )

$$|x| > 0.5$$

- 最后在Excel中筛选：
  - `correlation > 0.6`
  - 两个指标都：`dispersion > 0.4`
  - 两个指标的type不同
  - 两个指标的sub\_type不同
  - 两个指标的类型 ( 按照指标类型，处理数据 ) 一致

## 4. 【附录】 总体计划

### 目标：

开发出一套**通用的工具**，可以针对一组样本集合和一组特征集合，提取出在该样本集合下相关联的特征组/对，并且输出每组/对关联性的强弱。

### 特征之间的关联性：


#### 关联性定义：

样本集合下的特征关联性，即在样本集合中，两组或多组特征的值在以下几个方面相似，且均存在正相关、负相关，以及对应关联性强弱的结果：

1. **排序名次上**：基于不同的特征的值进行样本排序，样本在排序结果的名次上一致性高的特征组/对。容易解读。
2. **相关性上**：线性相关、指数相关等。容易解读。
3. **时间序列上**：在同一个样本的不同特征之间，给定时间窗口下（这个时间窗口不一定要一样，甚至不一定要一样长），不同特征的取值上是否存在相似性。考虑周期性。
  - a. 在等长窗口上，可以直接用相似度来量化，方法有很多。容易解读。
  - b. 不等长窗口上用动态规划来做。不容易解读，到时候再看怎么描述。
4. **分布上**：将样本集合在一项特征上的取值看做一个变量的分布，看不同变量分布之间的散度（评估方法有很多）。散度用于量化两个随机变量之间的独立性和相关性。不容易解读，到时候再看怎么描述好。


#### 独立性检验

假设检验（Test of Hypothesis）又称为显著性检验（Test of Statistical Significance）。在抽样研究中，由于样本所来自的总体其参数是未知的，只能根据样本统计量对其所来自总

 <http://www.cnblogs.com/zhangchaoyang/articles/2642032.html>

#### 相关性检验--Spearman秩相关系数和皮尔森相关系数

本文给出两种相关系数，系数越大说明越相关。你可能会参考另一篇博客独立性检验。皮尔森相关系数 皮尔森相关系数（Pearson correlation coefficient）也叫皮尔森积差相关系数

 <http://www.cnblogs.com/zhangchaoyang/articles/2631907.html>

5. **出现频次上**：将不同特征的值进行分箱操作后，得到样本集合在特征上的分类，分析不同特征组的分类在样本集合上同时出现的频次，同时出现次数多的即相似。（典型的关联规则挖掘，算法复杂度高）。容易解读。

## 交付内容：

1. 一个工具（程序）：
  - 特征间关联性挖掘：输入为一组样本在一组特征上的数据，输出为若干组关联特征组和每组相应的关联性强弱量化（值域在-1~1之间）。
  - 时间序列关联性挖掘：输入为一个样本在一组连续版本上的一组特征数据，输出为若干组关联特征和每组关联特征对应的时间窗口，以及关联性强弱量化（值域在-1~1之间）。

- 每种关联性定义下的输出结果都要保留。

### 注意事项：

1. 以上定义1、2、3.a项作为6月交付的目标；
2. 以上定义3.b、4、5作为7月的交付目标；
3. 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。
4. 针对不同类型关联性的解读文字，能够直接用到产品中，后面补。