6月总结

Created By	
Stakeholders	
Status	
Type	Technical Spec
Created	@June 28, 2022 10:35 AM
Last Edited Time	@June 28, 2022 2:30 PM
Last Edited By	

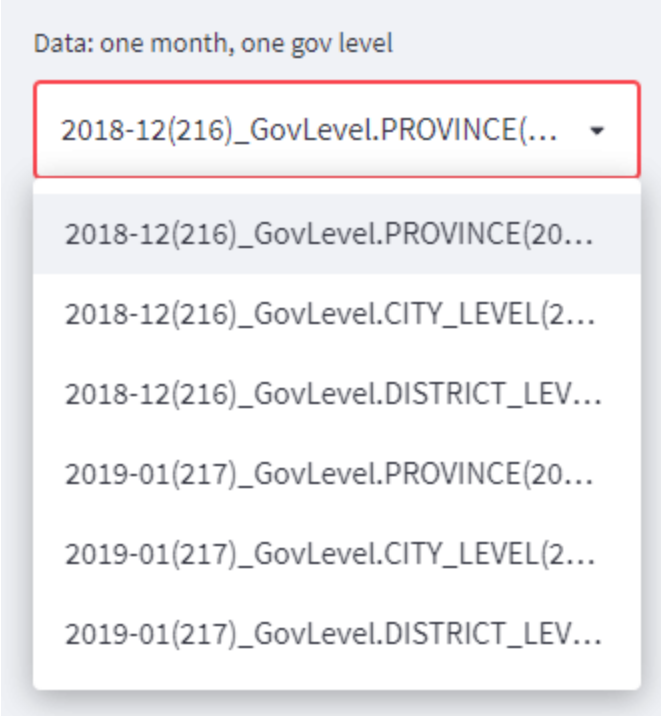
目标

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。

- 关联性定义：
- 定义1、2、3.a项作为6月交付的目标

关联性挖掘

1. 相关性上：
    a. 输入：一个时间点，一个样本集合（省、市、区，或按人口）



   b. 输出：一张表，按每对index的相关强度排序，

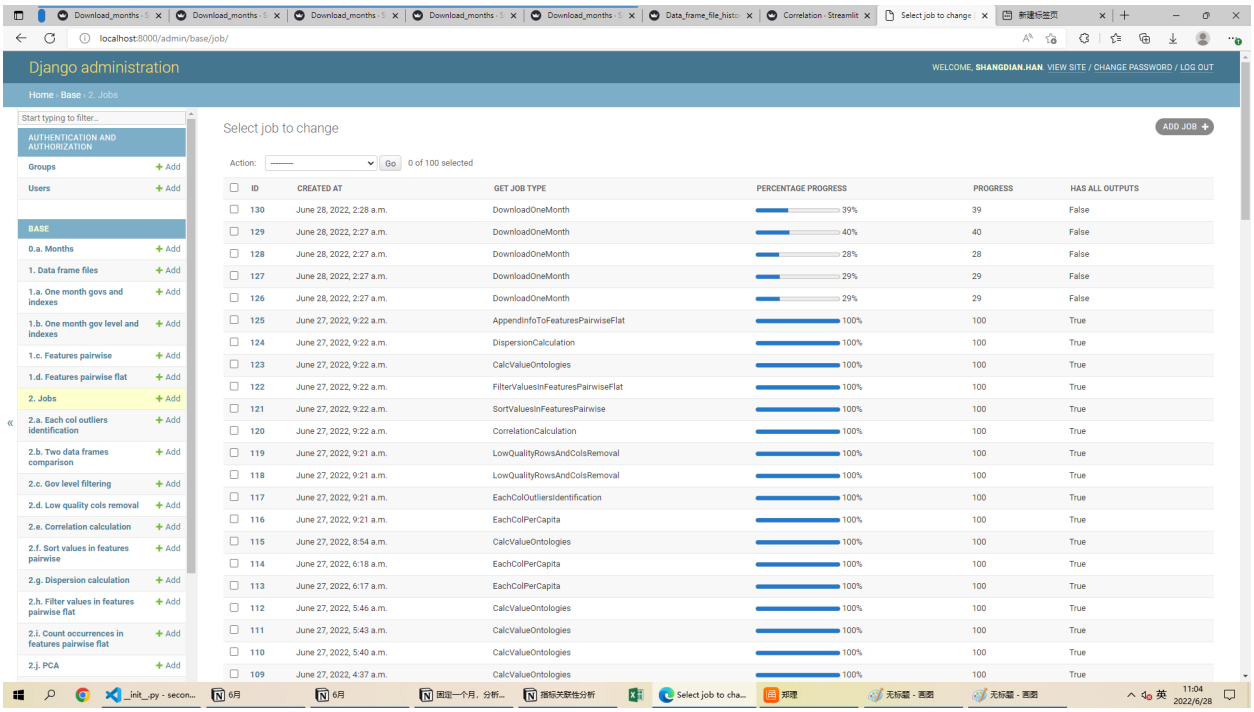
      `index_id_1, index_id_2, index_name_1, index_name_2, 相关强度,`
      `dispersion_1, dispersion_2, 其他关于index的信息（type, sub_type, …）`

2. 排序名次上：
   a. 输入：输入：一个时间点，一个样本集合（省、市、区，或按人口）
   b. 输出：输出：一张表，按每对index的相关强度排序，
      index_id_1, index_id_2, index_name_1, index_name_2, 相关强度,
      dispersion_1, dispersion_2, 其他关于index的信息（type, sub_type, …）
3. 时间序列上：
   a. 输入：连续几个时间点，一个样本（如全国）
   b. 输出：输出：一张表，按每对index的相关强度排序，
      index_id_1, index_id_2, index_name_1, index_name_2, 相关强度,
      dispersion_1, dispersion_2, 其他关于index的信息（type, sub_type, …）

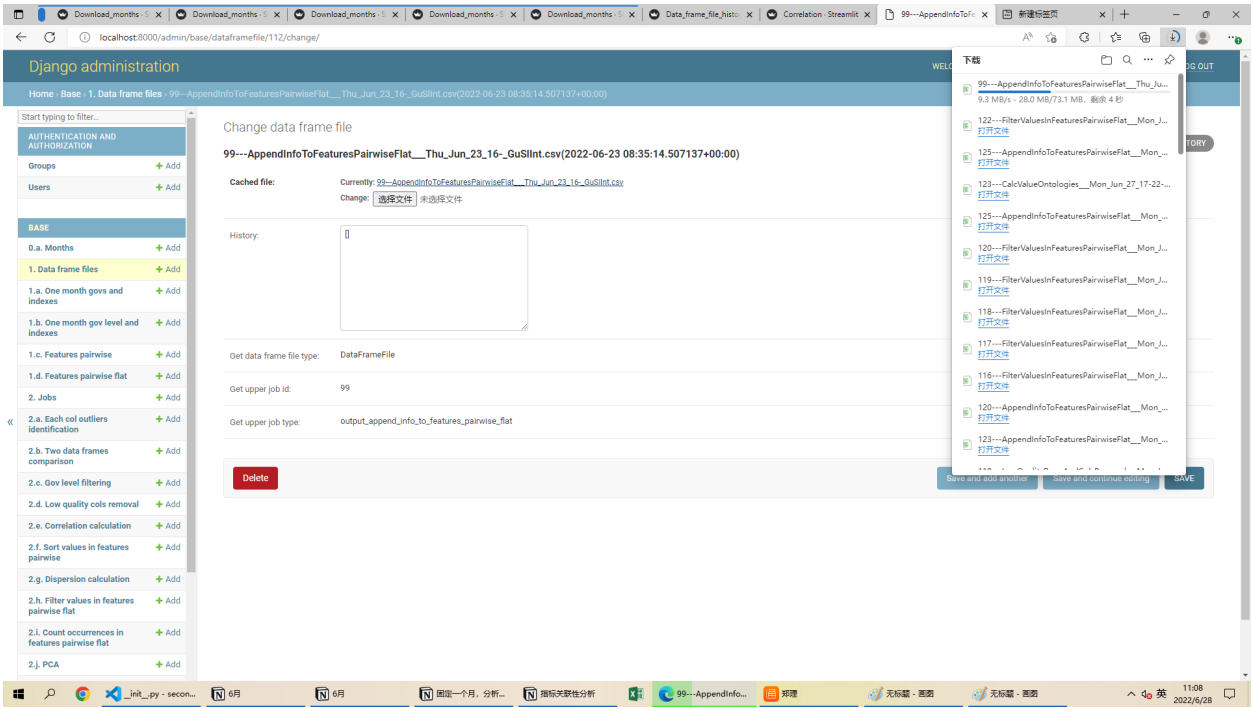
工具

一个具备前端、后端、数据库的数据处理框架，可以开放端口给内部人员使用以下功能：

任务管理

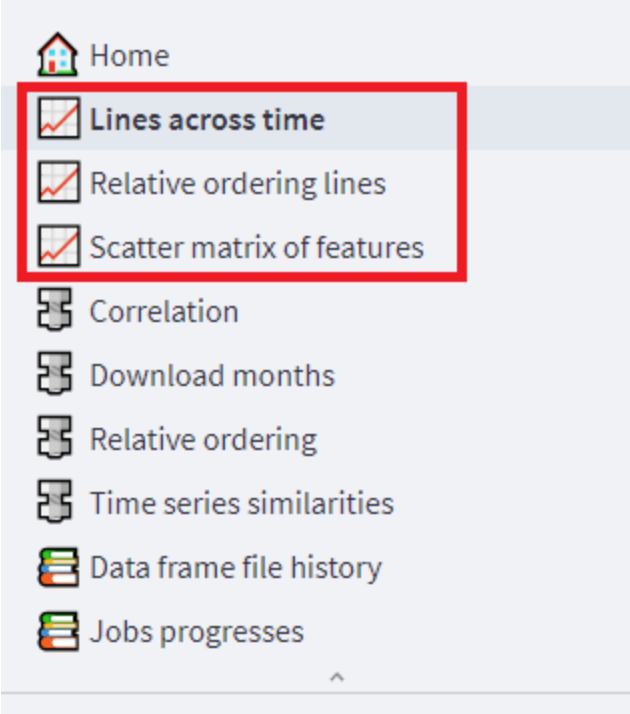

Django administration
WELCOME, SHANGDIAN.HAN. VIEW SITE / CHANGE PASSWORD / LOG OUT
Home › Base › 2. Jobs
Select job to change
ADD JOB
Action: Go 0 of 100 selected
ID | CREATED AT | GET JOB TYPE | PERCENTAGE PROGRESS | PROGRESS | HAS ALL OUTPUTS
130 June 28, 2022, 2:28 a.m. DownloadOneMonth 39% 39 False
129 June 28, 2022, 2:27 a.m. DownloadOneMonth 40% 40 False
128 June 28, 2022, 2:27 a.m. DownloadOneMonth 28% 28 False
127 June 28, 2022, 2:27 a.m. DownloadOneMonth 29% 29 False
126 June 28, 2022, 2:27 a.m. DownloadOneMonth 29% 29 False
125 June 27, 2022, 9:22 a.m. AppendInfoToFeaturesPairwiseFlat 100% 100 True
124 June 27, 2022, 9:22 a.m. DispersionCalculation 100% 100 True
123 June 27, 2022, 9:22 a.m. CalcValueOntologies 100% 100 True
122 June 27, 2022, 9:22 a.m. FilterValuesInFeaturesPairwiseFlat 100% 100 True
121 June 27, 2022, 9:22 a.m. SortValuesInFeaturesPairwise 100% 100 True
120 June 27, 2022, 9:22 a.m. CorrelationCalculation 100% 100 True
119 June 27, 2022, 9:21 a.m. LowQualityRowsAndColsRemoval 100% 100 True
118 June 27, 2022, 9:21 a.m. LowQualityRowsAndColsRemoval 100% 100 True
117 June 27, 2022, 9:21 a.m. EachColOutliersIdentification 100% 100 True
116 June 27, 2022, 9:21 a.m. EachColPerCapita 100% 100 True
115 June 27, 2022, 8:54 a.m. CalcValueOntologies 100% 100 True
114 June 27, 2022, 6:18 a.m. EachColPerCapita 100% 100 True
113 June 27, 2022, 6:17 a.m. EachColPerCapita 100% 100 True
112 June 27, 2022, 5:46 a.m. CalcValueOntologies 100% 100 True
111 June 27, 2022, 5:43 a.m. CalcValueOntologies 100% 100 True
110 June 27, 2022, 5:40 a.m. CalcValueOntologies 100% 100 True
109 June 27, 2022, 4:37 a.m. CalcValueOntologies 100% 100 True
AUTHENTICATION AND AUTHORIZATION
Groups
Users
BASE
0.a. Months
1. Data frame files
1.a. One month govs and indexes
1.b. One month gov level and indexes
1.c. Features pairwise
1.d. Features pairwise flat
2. Jobs
2.a. Each col outliers identification
2.b. Two data frames comparison
2.c. Gov level filtering
2.d. Low quality cols removal
2.e. Correlation calculation
2.f. Sort values in features pairwise
2.g. Dispersion calculation
2.h. Filter values in features pairwise flat
2.i. Count occurrences in features pairwise flat
2.j. PCA

文件管理

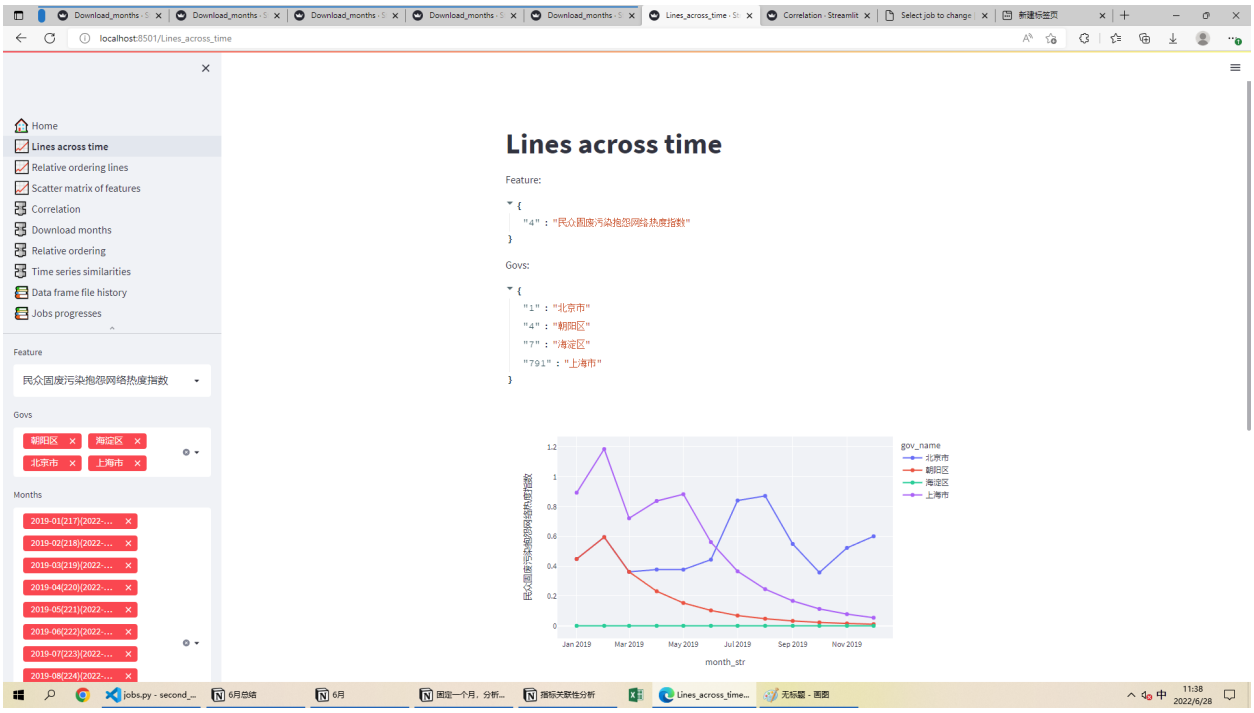

Django administration
Home › Base › 1. Data frame files › 99---AppendInfoToFeaturesPairwiseFlat___Thu_Jun_23_16-_GuSIInt.csv(2022-06-23 08:35:14.507137+00:00)
Change data frame file
99---AppendInfoToFeaturesPairwiseFlat___Thu_Jun_23_16-_GuSIInt.csv(2022-06-23 08:35:14.507137+00:00)
Cached file: Currently: 99---AppendInfoToFeaturesPairwiseFlat___Thu_Jun_23_16-_GuSIInt.csv
Change: 选择文件 未选择文件
History:
Get data frame file type: DataFrameFile
Get upper job id: 99
Get upper job type: output_append_info_to_features_pairwise_flat
Delete
Save and add another
Save and continue editing
SAVE
下载
99---AppendInfoToFeaturesPairwiseFlat___Thu_Ju...
9.3 MB/s - 28.0 MB/73.1 MB，剩余 4 秒
122---FilterValuesInFeaturesPairwiseFlat___Mon_J...
125---AppendInfoToFeaturesPairwiseFlat___Mon_...
123---CalcValueOntologies___Mon_Jun_27_17-22-...
120---FilterValuesInFeaturesPairwiseFlat___Mon_J...
119---FilterValuesInFeaturesPairwiseFlat___Mon_J...
118---FilterValuesInFeaturesPairwiseFlat___Mon_J...
117---FilterValuesInFeaturesPairwiseFlat___Mon_J...
116---FilterValuesInFeaturesPairwiseFlat___Mon_J...
120---AppendInfoToFeaturesPairwiseFlat___Mon_...
123---AppendInfoToFeaturesPairwiseFlat___Mon_...
打开文件

结果可视化

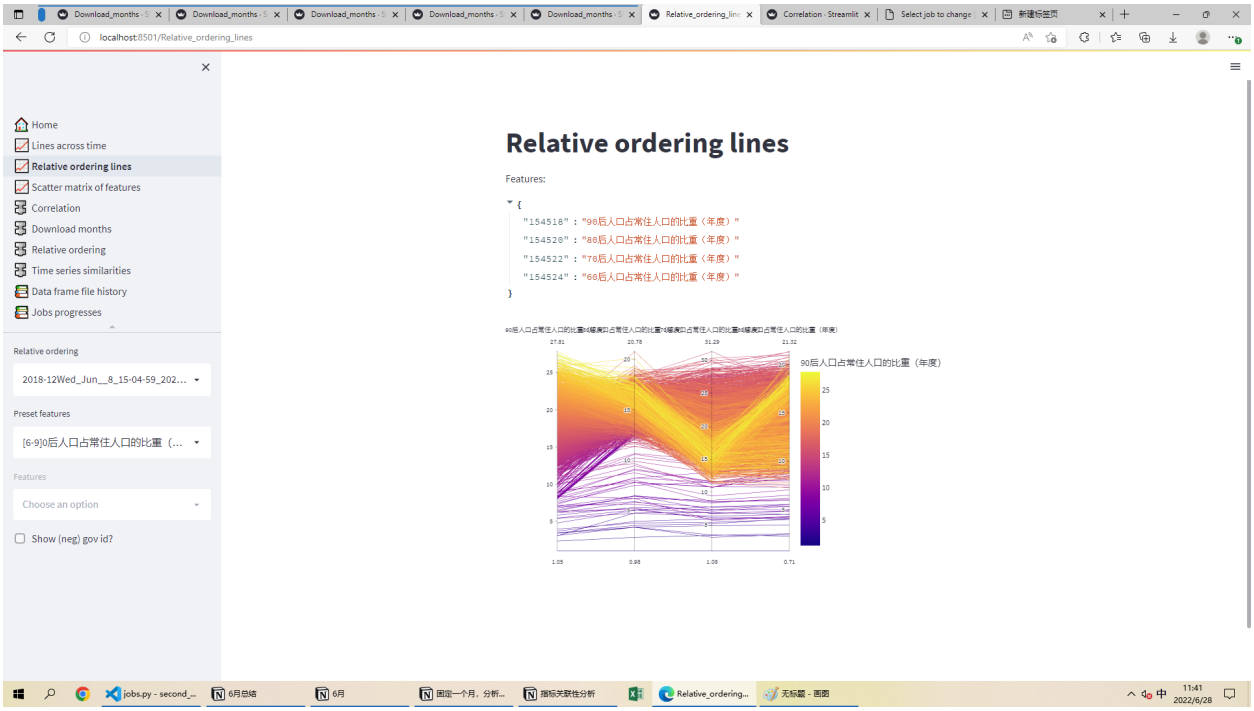
图表类型



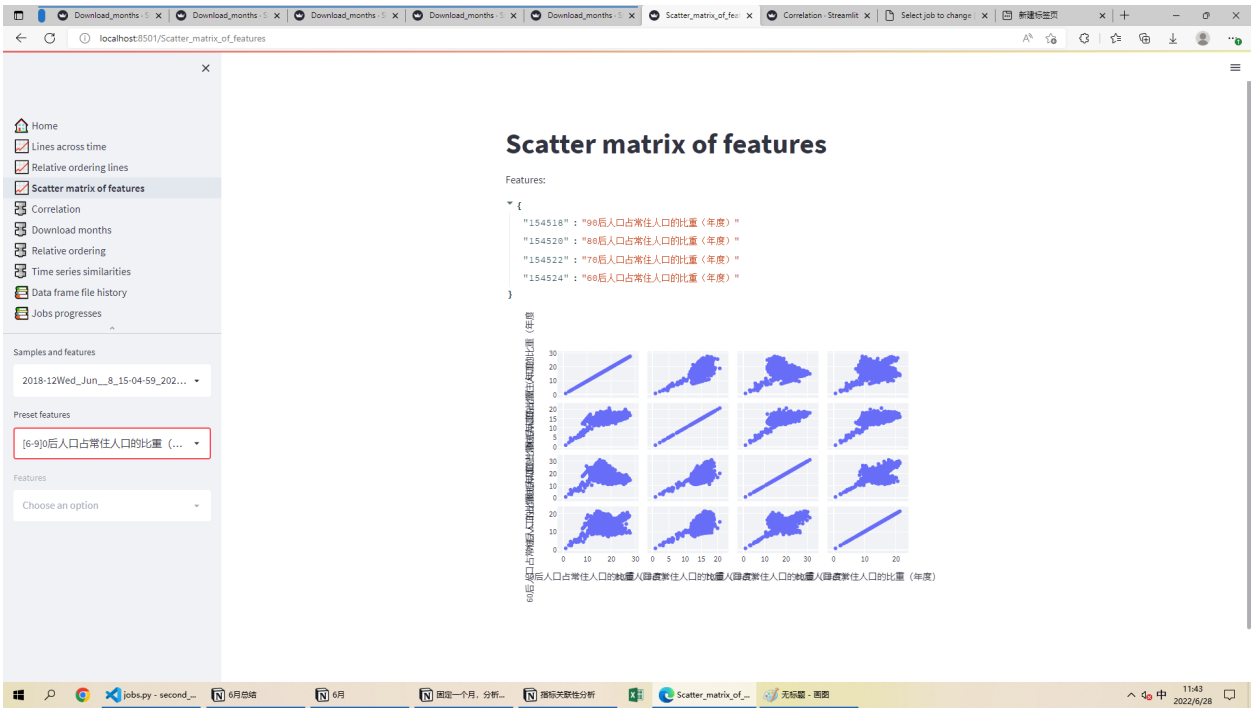
时间序列


Lines across time
Feature:
{
"4" : "民众固废污染抱怨网络热度指数"
}
Govs:
{
"1" : "北京市"
"4" : "朝阳区"
"7" : "海淀区"
"791" : "上海市"
}
Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Time series similarities
Data frame file history
Jobs progresses
Feature
民众固废污染抱怨网络热度指数
Govs
朝阳区
海淀区
北京市
上海市
Months
2019-01(217)(2022-...
2019-02(218)(2022-...
2019-03(219)(2022-...
2019-04(220)(2022-...
2019-05(221)(2022-...
2019-06(222)(2022-...
2019-07(223)(2022-...
2019-08(224)(2022-...
gov_name
北京市
朝阳区
海淀区
上海市
Jan 2019
Mar 2019
May 2019
Jul 2019
Sep 2019
Nov 2019
month_str

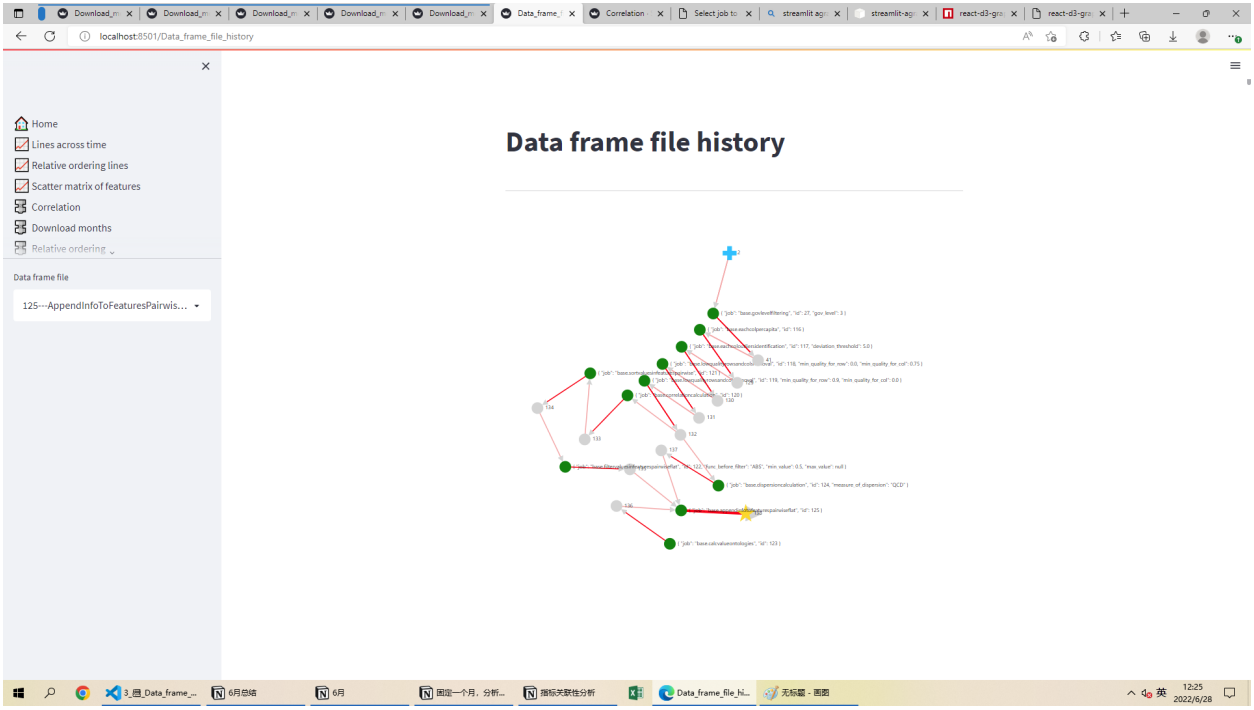
排序名次

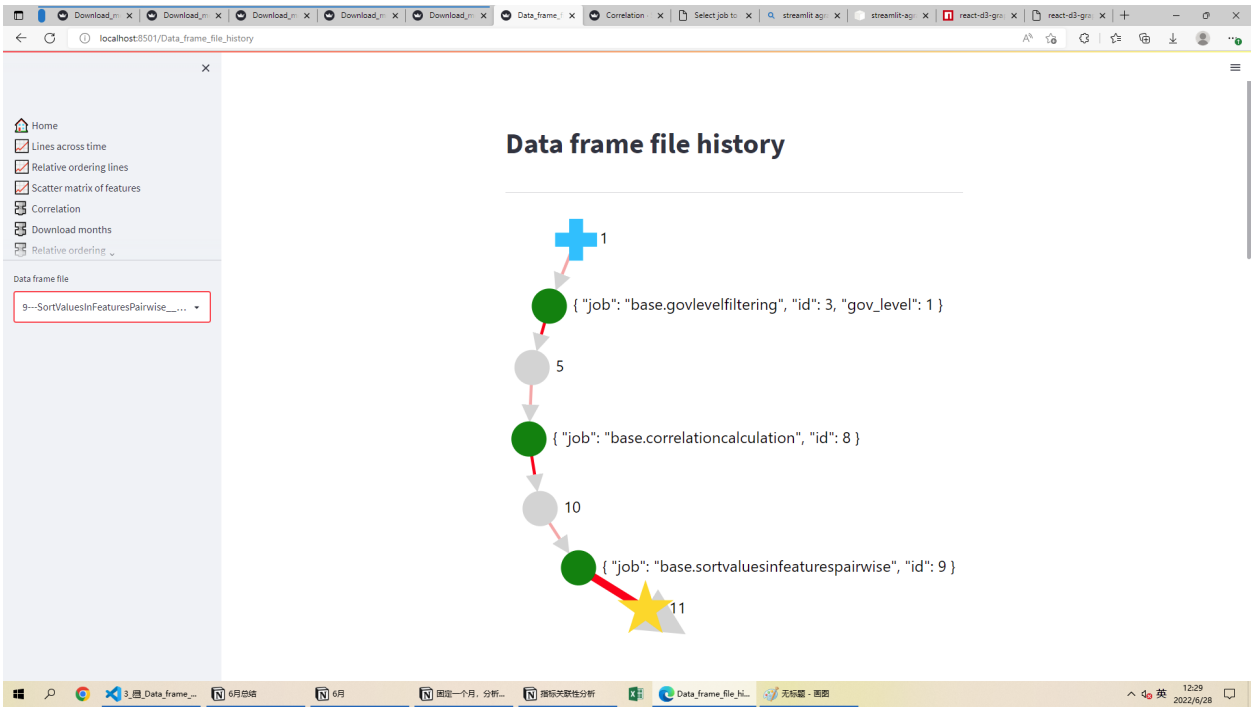

Relative ordering lines
Features:
{
"154518" : "90后人口占常住人口的比重（年度）"
"154520" : "80后人口占常住人口的比重（年度）"
"154522" : "70后人口占常住人口的比重（年度）"
"154524" : "60后人口占常住人口的比重（年度）"
}
Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Time series similarities
Data frame file history
Jobs progresses
Relative ordering
2018-12Wed_Jun__8_15-04-59_202...
Preset features
[6-9]0后人口占常住人口的比重（...
Features
Choose an option
Show (neg) gov id?
90后人口占常住人口的比重（年度）

相关性

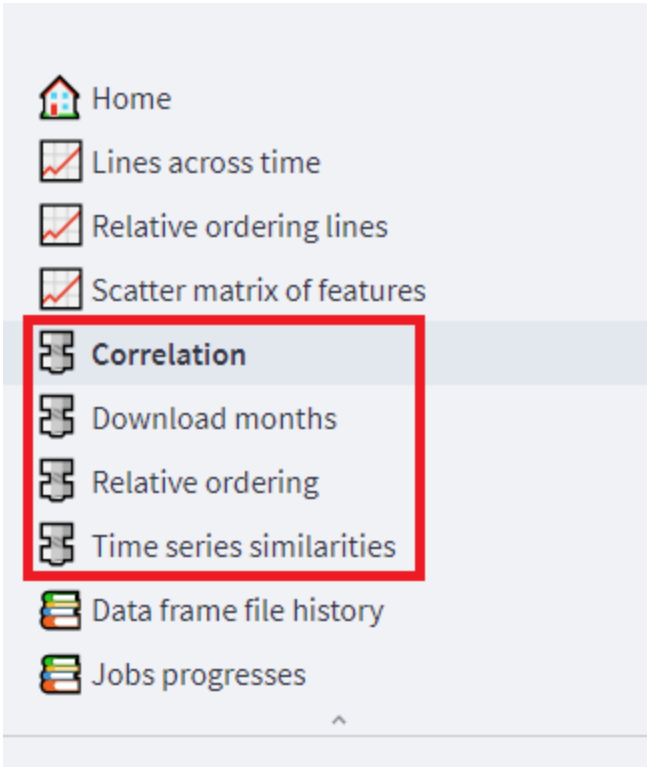

Scatter matrix of features
Features:
{
"154518" : "90后人口占常住人口的比重（年度）"
"154520" : "80后人口占常住人口的比重（年度）"
"154522" : "70后人口占常住人口的比重（年度）"
"154524" : "60后人口占常住人口的比重（年度）"
}
Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Time series similarities
Data frame file history
Jobs progresses
Samples and features
Preset features
Features
Choose an option

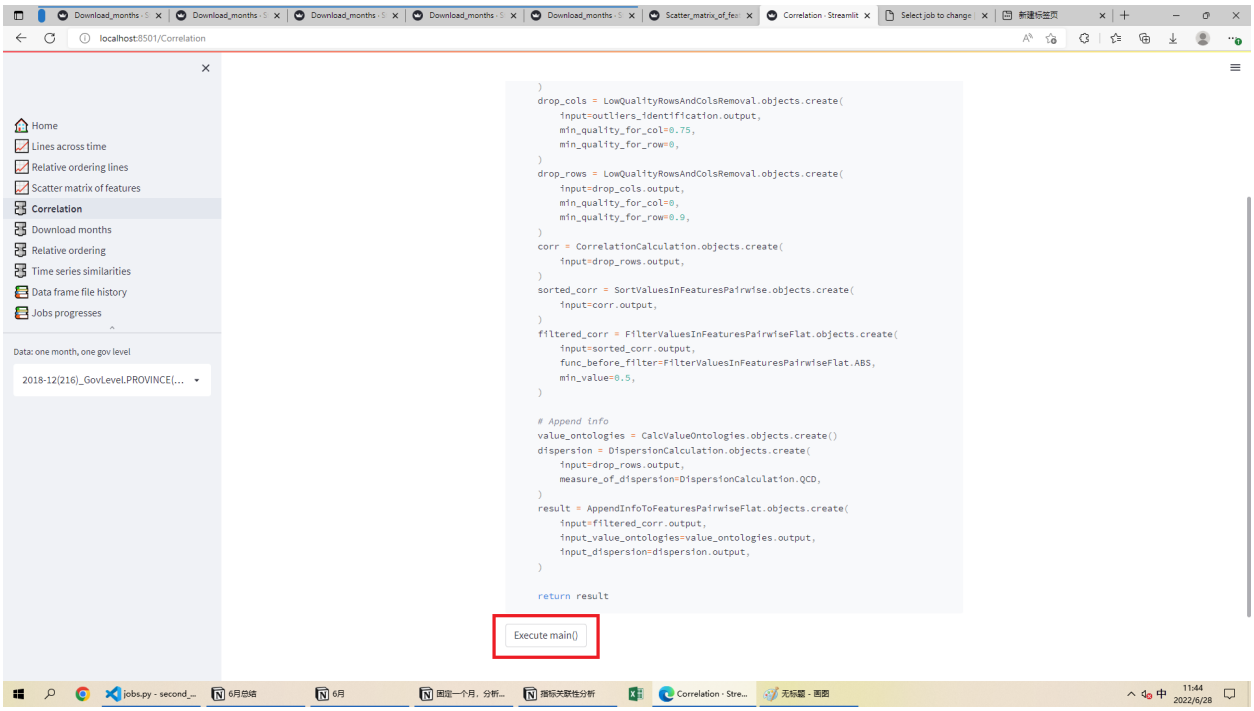
数据处理历史可视化


Data frame file history
Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Data frame file


localhost:8501/Data_frame_file_history
Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Data frame file
9---SortValuesInFeaturesPairwise_...
Data frame file history
1
{ "job": "base.govlevelfiltering", "id": 3, "gov_level": 1 }
5
{ "job": "base.correlationcalculation", "id": 8 }
10
{ "job": "base.sortvaluesinfeaturespairwise", "id": 9 }
11

批量执行任务


Home
Lines across time
Relative ordering lines
Scatter matrix of features
Correlation
Download months
Relative ordering
Time series similarities
Data frame file history
Jobs progresses



代码细节

关联性挖掘

- 相关性
  1. 固定一个时间点
  2. 筛选出“区县级”的样本
  3. 按照指标类型，处理数据
     a. 数量 → 除以人口
     b. 比例 → 不变
     c. 增量 → 除以人口
     d. 增幅 → 不变
     e. 分数 → 不变
     f. 人均数量 → 不变
     g. 价格 → 不变

h. 全国占比 → 除以人口

4. 去除离群值（`deviation_threshold=5`）

```
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
```

5. 去除质量过低的列（`min_quality_for_col=0.75`）
6. 去除质量过低的行（`min_quality_for_row=0.9`）
7. 计算相关性
8. 筛选相关性强的指标对（`func=ABS, min_value=0.5`）

$$|x| > 0.5$$

9. 用这个任务去除质量过低的列（min_quality_for_col=0.75）的结果计算离散度（`measure_of_dispersion=QCD`）

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

10. 合并：相关性结果、离散度、指标类型（按照指标类型，处理数据）、指标其他信息
11. 最后在Excel中筛选：
    a. $\mathrm{correlation} > 0.6$
    b. 两个指标都：$\mathrm{dispersion} > 0.4$

c. 两个指标的type不同

d. 两个指标的sub_type不同

e. 两个指标的类型（按照指标类型，处理数据）一致

- 排序名次

  方法基本和相关性 相同，除了：

  - 计算关联性前需要将`NaN`补成指标上的平均值。
  - 关联性的计算是：

    1. 在两个指标上分别进行样本排序，若两个样本的值相同则拥有相同的序号。
    2. 两个排序的距离：$\sum_{样本}((序号_1 - 序号_2)^2)$

- 时间序列

  方法基本和排序名次 相同，除了：

  - 关联性的计算是：

    1. 将每个指标标准化
    2. 两个指标的距离：$\sum_{时间}((标准化值_1 - 标准化值_2)^2)$

缓存

- 从DAAS上读取有点慢，所以我们可以缓存每一版（一个时间点）的所有数据。
- 小任务的结果和参数都会被储存，以方便内部开发。

编写新的任务类型

- 很容易，继承`Job`类：
  - 输入：
    - `DataFrameFile | tuple[DataFrameFile] | None`
    - 参数
  - 输出：
    - `DataFrameFile | tuple[DataFrameFile]`
- 任务的管理、缓存等都会自动被处理好。

下一步

- [ ] 多线程
- [ ] 调参
- [ ] 岔开时间的时间序列相似性
- [ ] 更高级的关联性挖掘

问题

-