**实 验 报 告**

|  |  |
| --- | --- |
| **课程名称：** | **Python数据分析** |
| **学生姓名：** | **刘博严** |
| **学生学号：** | **202030390308** |
| **学生专业：** | **电子商务** |
| **开课学期：** | **2022-2023学年第一学期** |

**电子商务系**

**2022年11月**

**目 录**

[实验一 数据抓取 3](#_Toc25935990)

实验二 数据清洗 12

实验三 地区实证…………………...………………………………………………………………..22

[实验四 企业实证 3](#_Toc27392798)4

# 实验一 数据抓取

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **地 点：** | B10 楼 | 208房； | **实验台号：** |  |
| **实验日期与时间：** | 2022/11/7 | | **评 分：** |  |
| **预习检查纪录：** |  | | **实验教师：** |  |
| **批改意见** |  | | | |

【实验目的及要求】

掌握Python 爬虫的使用。使用Python的Request、lxml、BeautifulSoup、selenium等工具包，从知识产权官方媒体的抓取相关数据

【实验内容】

第一部分，基础部分

1. 国家知识产权局（2008及以后）专利统计数据的收集
2. 从佰腾网完成各省不同类型专利数据的收集
3. 主要省份知识产权官方微博数据的抓取
4. 主要省机关报数据的抓取

国家知识产权局专利统计数据的抓取

1. 百度搜索：https://www.cnipa.gov.cn/，打开以上链接



点击“数据”，找到“国家知识产权局统计年报”，输入年份，点击查询



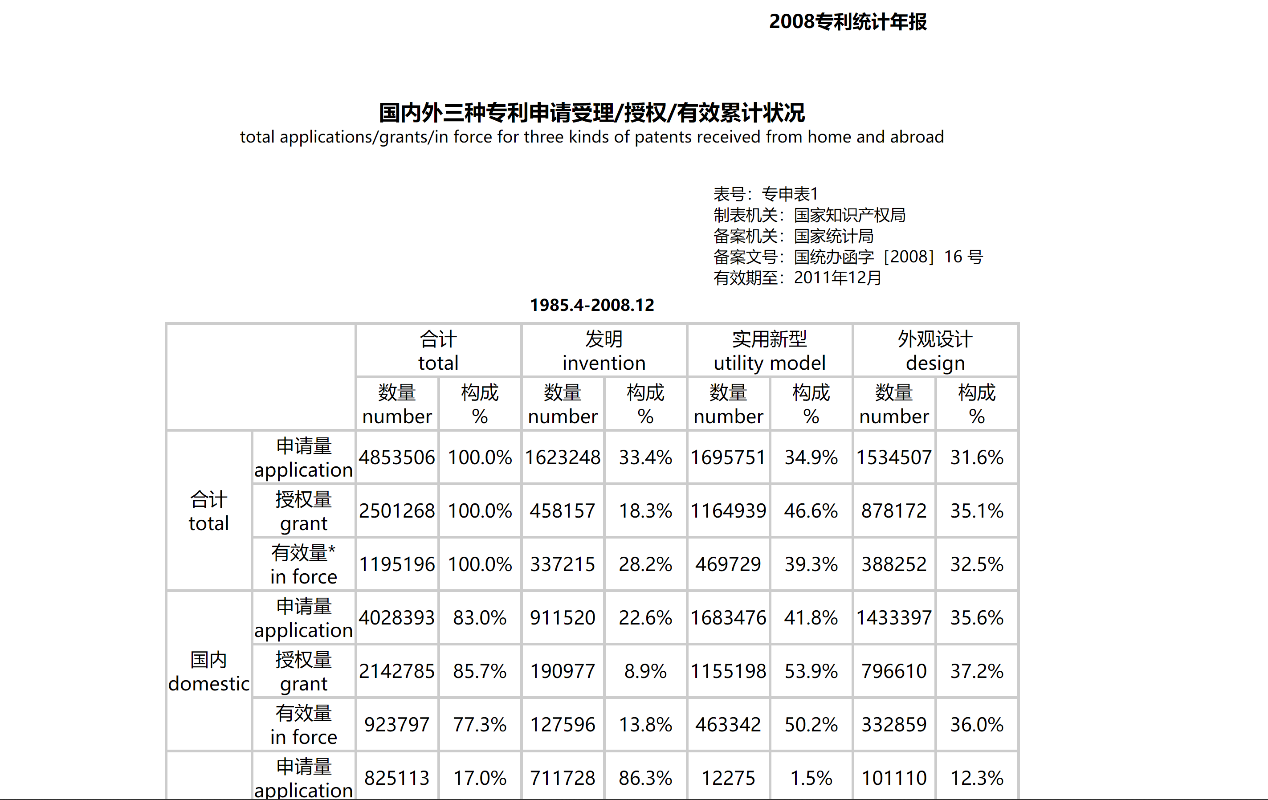
1. 获取各年专利统计年报的子页面专利申请状况、专利申请授权状况、专利有效状况、专利行政执法状况的url。（提示由于url类似，可以考虑直接生成）



3、获取专利申请状况（专利申请授权状况、专利有效状况、专利行政执法状况做相同处理）子页面的所有url，



4、抓取对应页面的表格数据，并且输出到excel中



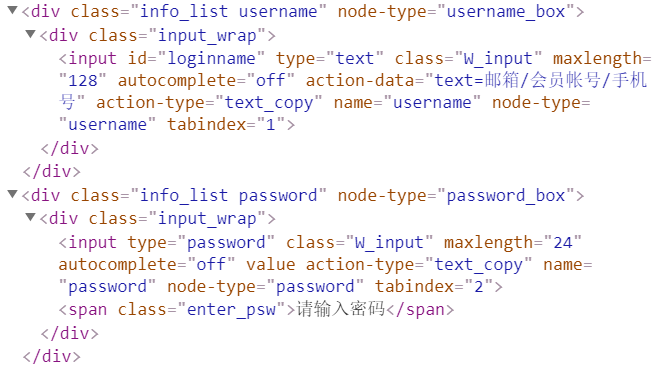
各省知识产权官方微博数据的抓取（以北京知识产权官方微博为例）

一、使用自己的微博账号模拟登录微博（使用selenium库）

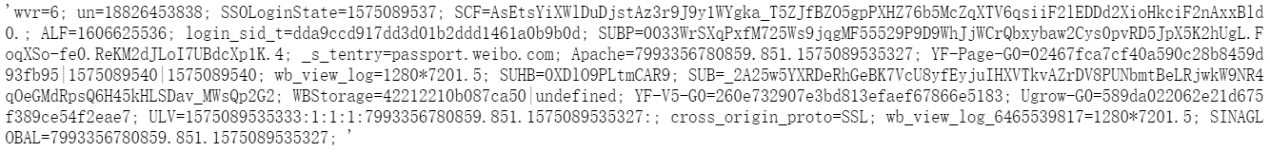
1. 查看谷歌浏览器的版本，在（<http://npm.taobao.org/mirrors/chromedriver/>）中下载相应的chromedriver。

2.在谷歌浏览器中的开发者工具(F12)中找到微博首页（https://weibo.com/）登录框的网页源代码。

（提示：可以使用user = browser.find\_element\_by\_id和pwd = browser.find\_element\_by\_name来锁定账号密码两个输入框，使用user.send\_keys和pwd.send\_keys来传入账号和密码）

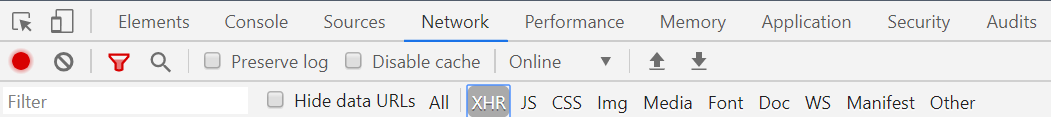


3.成功登录之后，使用browser.get\_cookies()并且将其整合成一个字符串。

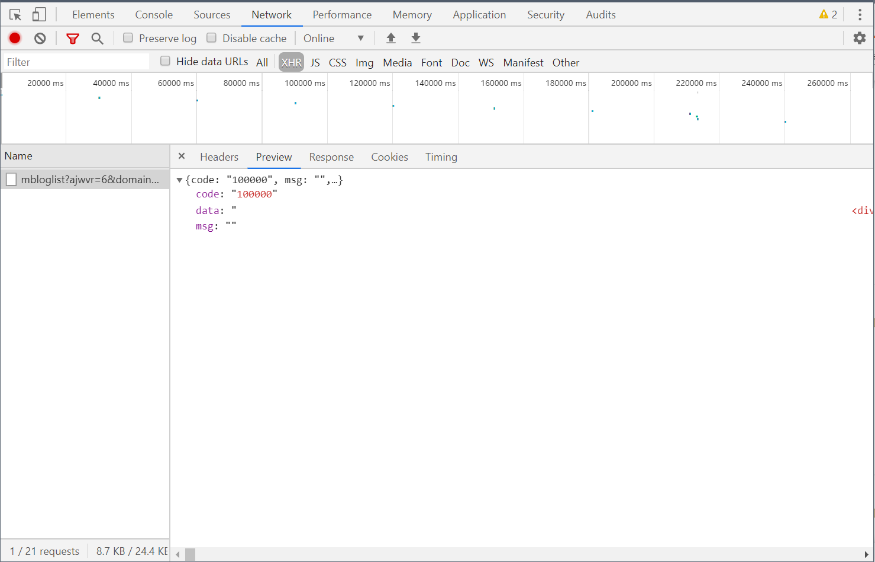


二、数据抓取

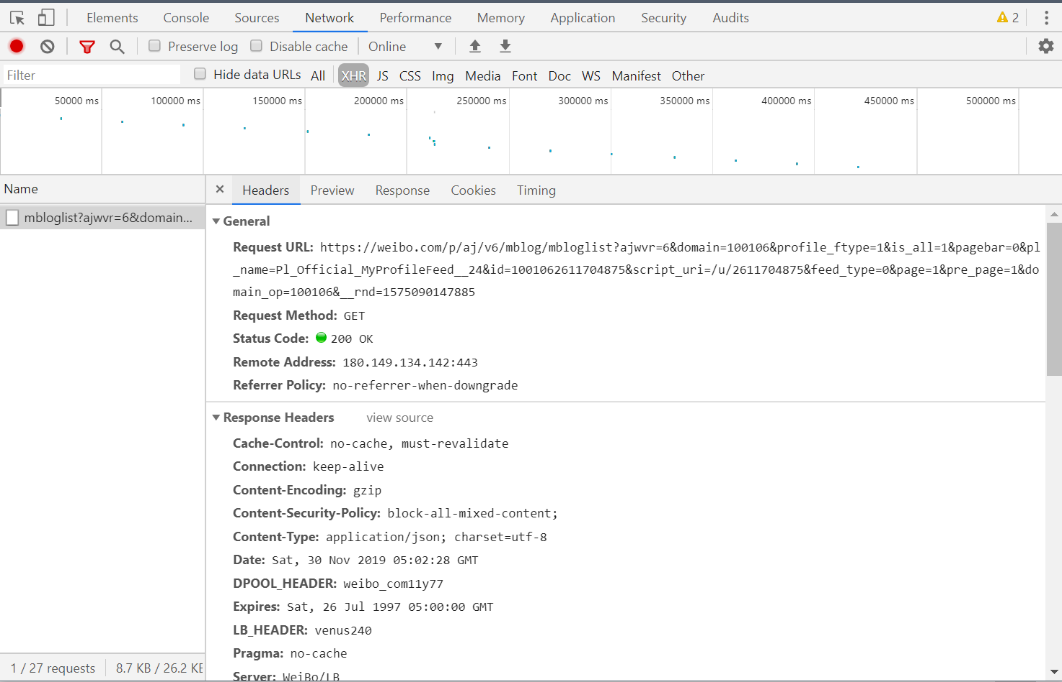
4.浏览北京知识产权官方微博（<https://weibo.com/u/2611704875?profile_ftype=1&is_all=1>），打开开发者工具（F12），点击Network选项并且选择XHR。



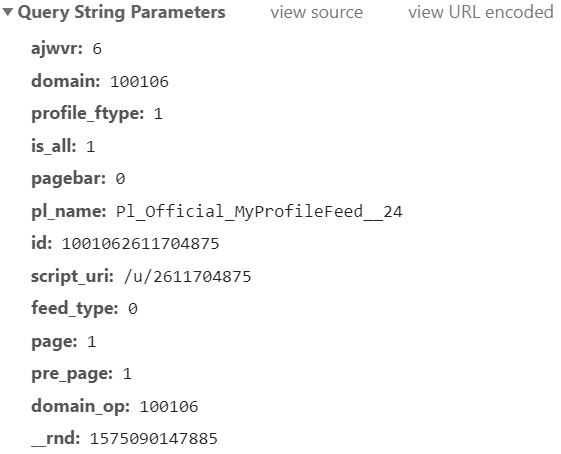
下拉网页，直到刷到新的微博内容，观察XHR选项卡的Preview中的内容。



通过观察可以发现每条微博内容都存在了‘data’中，再观察headers选项卡



所以我们需要使用request的get方法，并且模仿request url的格式来获得所有的微博内容。



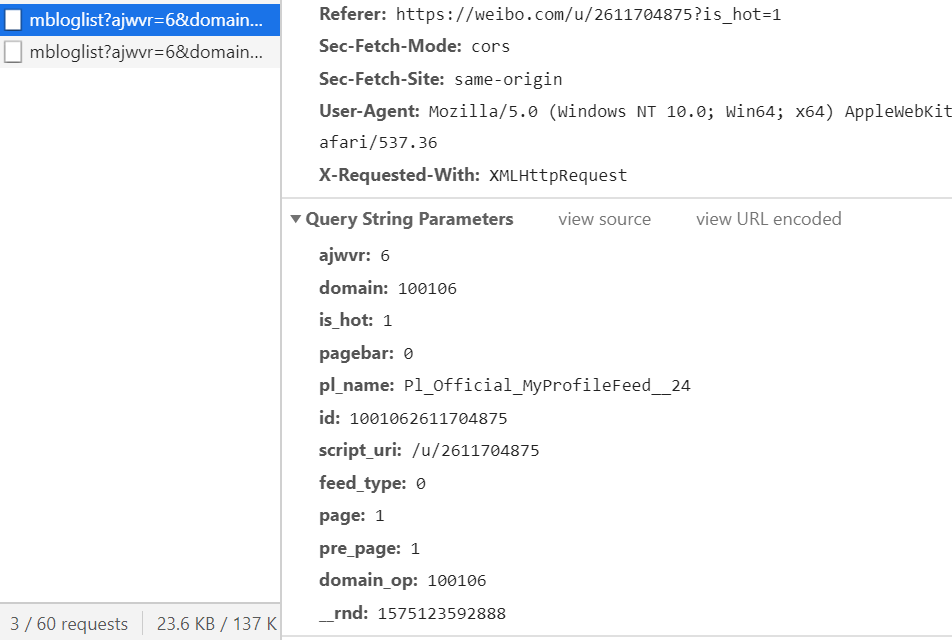
提示：只有上面参数中的（pagebar、id、script\_uri、page、pre\_page、\_\_rnd）才是变动的

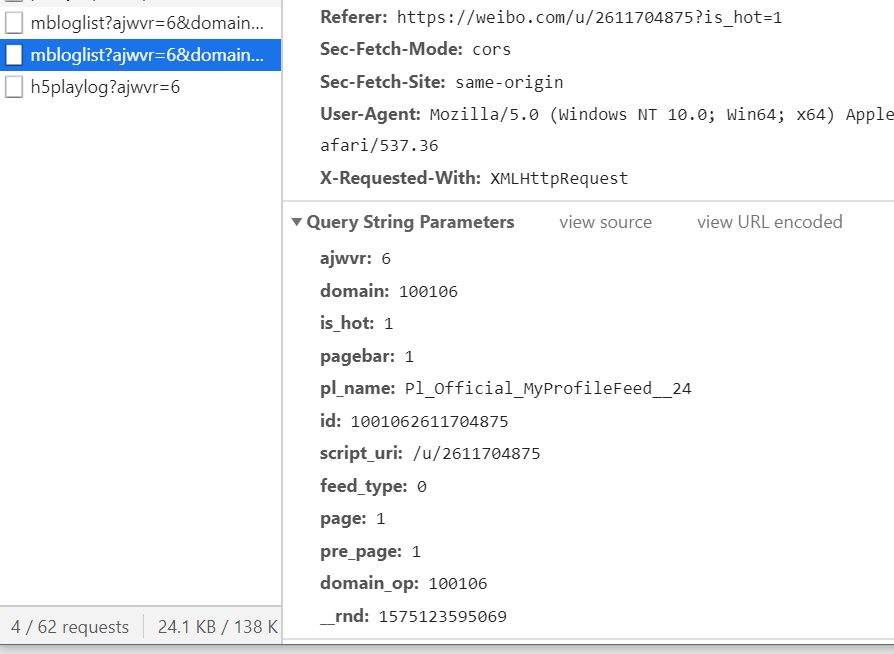
\_\_rnd为毫秒级时间戳（提示：可以用time库来获取）

id和script\_url可以在该微博网页源代码的head标签下找到：（id为下面截图的page\_id, script\_url为’/u/’加上id的后十位数



将微博页面下拉到最底下，然后发现XHR中出现了两个新的Ajax请求链接，我们可以发现每页都会进行两次Ajax请求：



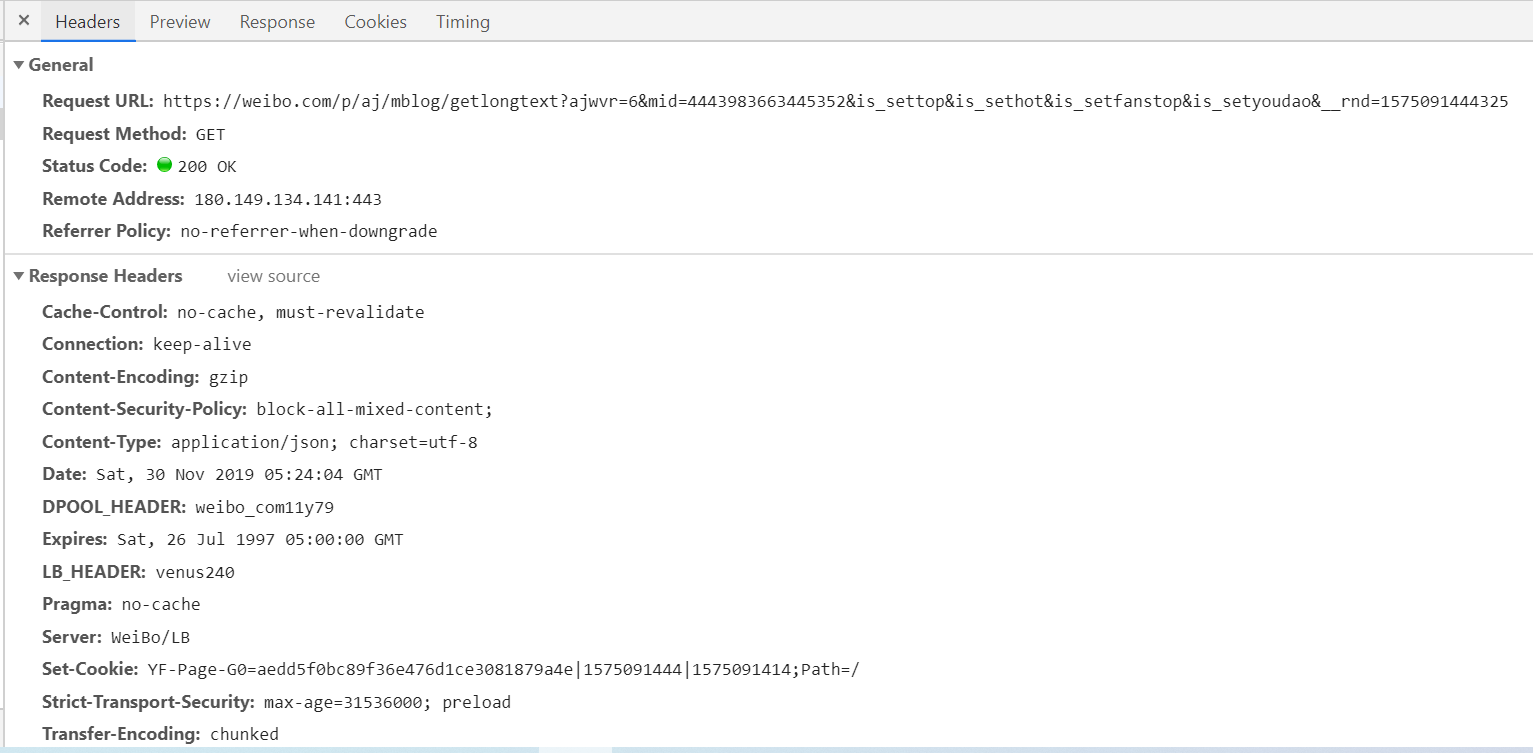


观察两条Ajax请求连接参数的变化，我们可以发现Pagebar从0变成了1，page和pre\_page都没有变化。但是刚打开网页没有下拉页面时，网页已经有了一次Ajax请求了（因为已经有微博展示再网页中），所以尝试将Pagebar设为0,page设为1，pre\_page设为0，进行一次Ajax请求，就会发现能返回刚打开微博页面的微博内容，因此微博在每一页都会请求三次Ajax，第i页对应的pagebar, page,pre\_page参数分别为(0,i,i-1),(0,i,i),(1,i,i)。只要我们通过模仿请求这三次Ajax，就能获取每页的所有微博内容。

5、当微博内容需要点击“展开全文”才会显示全部内容时，我们需要建立一个新方法



进行开发者工具之后，在网页中点击“展开全文”，可以发现请求了一个新的链接：



其中的参数mid也能在网页源码中找到。模拟该请求，创建一个方法。

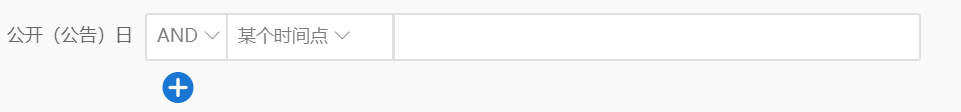
6、使用beautifulsoup来解析第4步获取到的数据，当“发现”内容中有“展开全文”时再调用第5步的方法。

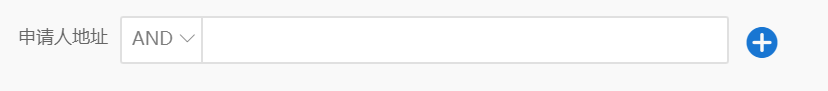
7、将获取到的所有微博内容存在一个dataframe中，并输出到excel中



从佰腾网完成各省不同类型专利数据的收集

1、打开链接<https://www.baiten.cn/gjs.html>，找到以下两个检索条件





分别输入2018，广东，然后观察检索结果



我们需要抓取的数据为左边各种专利的种类数量

2、打开开发者工具（F12），刷新一下网页，找到Network中XHR的名为filter的请求，



可以发现我们需要的数据就存在其返回的Json中

3、模仿filter请求，获得全国各省1985-2019年所有的专利种类数量数据。存在一个dataframe并且输出到excel中



省机关报数据的抓取（以北京日报为例）

1. 抓取一天的北京日报内容

<http://bjrb.bjd.com.cn/html/2019-11/30/node_1.htm>，打开以上链接

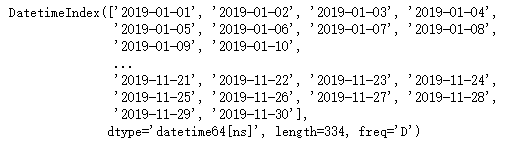
1. 获取各个版面的链接，提示，此处可以用正则表达式

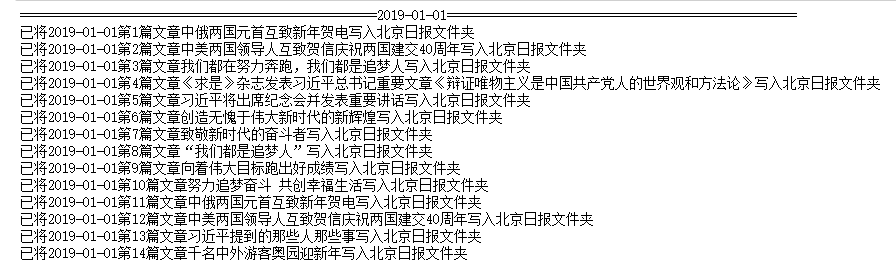


1. 获取各个版面内的文章的链接，也可以用正则表达式来匹配



1. 获取每篇文章的内容并写入文件夹中即可，温馨提示：文件夹中的文件名称不能含有\/:\*?<>|等符号，否则会报非法参数的错误。
2. 然后将时间改成2019年，此处可以用时间序列





# 实验二 数据清洗

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **地 点：** | B10 楼 | 208房； | **实验台号：** |  |
| **实验日期与时间：** | 2022/11/14 | | **评 分：** |  |
| **预习检查纪录：** |  | | **实验教师：** |  |
| **批改意见** |  | | | |

【实验目的及要求】

掌握Python数据预处理技术。使用Python 的re、pandas库对抓取和下载的数据进行清洗和规整等操作

【实验内容】

第一部分，基础部分

1.知识产权年鉴数据清洗，包括各地区知识产权执法、确权等数据

2.机关报数据清洗，包括各地区机关报数据中相关性指标

3.知识产权微博数据清洗，包括各地区的关注度指标

**知识产权年鉴数据清洗（即实验一的国家知识产权局数据）**

一、知识产权年鉴中各地区每年（这里以2018年为例）专利执法数（统计各个地区的立案、结案数） 数据为：h1.xls（各地区管理专利工作的部门专利执法统计表）

1.读取数据，只需要前四列，数据从第‘2’行开始有效，存在data中。



2.给列重命名，这里为['地区', '年份', '立案', '结案']，。



3.筛选出年份这一列为2018的数据。



4.将地区这一列数据改为对应的中文，并将地区设置为index。



5.计算结案与立案的比值，变量名为结案/立案（或结案率）。



6.将数据存入excel中。

二、统计各地区每年查处执法假冒量

1.读入数据，数据为h5.xls（各地区查处假冒专利执法统计表（一））,



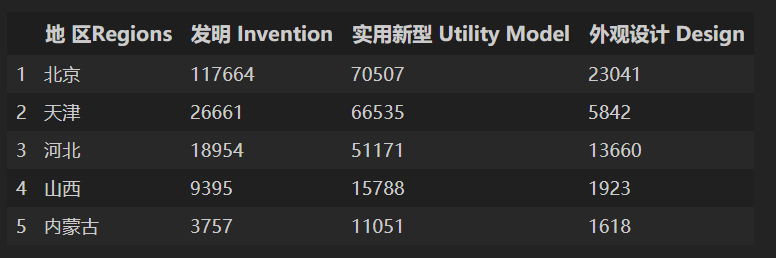
2.重命名列名，重设索引值，将数据保存到excel。



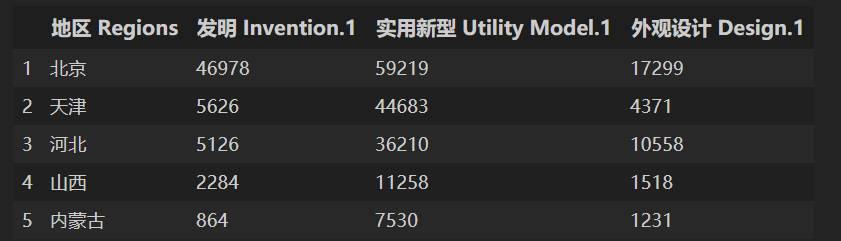
三、地区专利情况： 各地区3种专利的申请量、有效量，发明专利授权量/申请量，发明专利/(发明+实用新型)

1.读入数据，用a4（国内三种专利申请量），b3（国内三种专利申请授权量），c2（国内三种专利有效量）并选择需要的行和列，分别存入df1、df2、df3

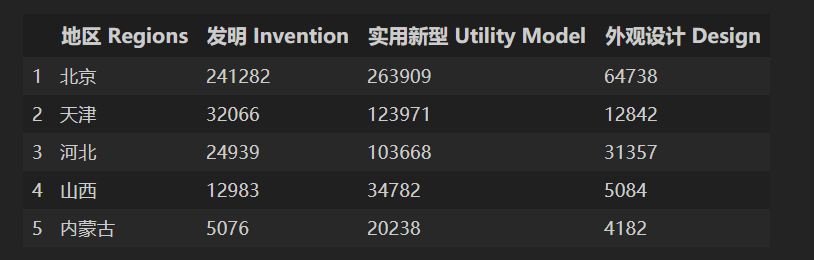
申请量：



申请授权量：



有效量：



2.重新设置df1、df2、df3的列名，







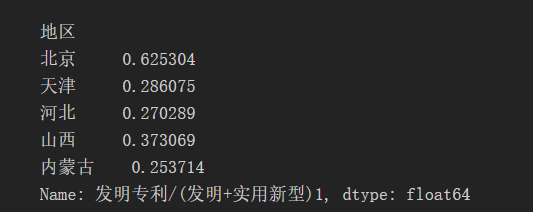
3.将df1、df2、df3合并存入df，



4.计算发明专利授权量/申请量，并填入df中。



5.对申请量、授权量和有销量分别计算 发明专利/(发明+实用新型)，也填入df中，将df存入excel。



6.将一、二、三中的三个数据表整合

7.其余年份的数据做类似处理。

8.以(年份,地区)作为二重行索引，把各年份数据整合，获得知识产权年鉴清洗后的数据



**机关报数据清洗**

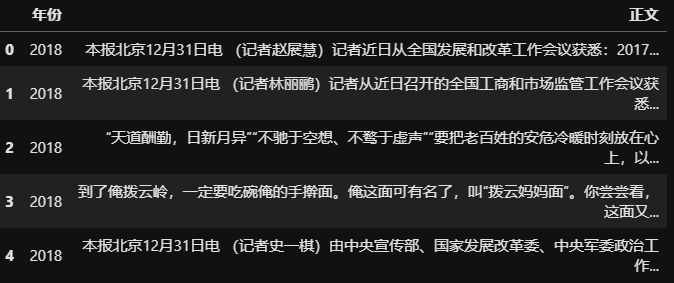
本例子使用人民日报2018-2019年的数据，引入的库：os, re, pandas

*\*\*\*（由于平台本身原因，上传数据比较麻烦，建议将日报数据整合到一个表格里面（csv也可以），****已整理好，可以直接用 人民日报.xlsx****）*

1、读取数据：人民日报.xlsx

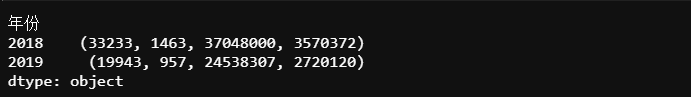


2、新增年份列，为标题前四个字符。然后除去标题列



3、计算每年的知识产权文章数量以及每年知识产权文章字数和总文章字数

使用正则表达式去掉文章中多余的空白以及，。？；：‘’、！（）《》等符号， 遍历所有文章的内容，若文章内容中存在知识产权， 专利， 商标， 版权， 著作权五个关键词中的一个或者多个，即记为知识产权类文章。计算每年知识产权类文章的数量，字数， 以及该年份所有文章的数量，字数



3、通过第2步中的数据计算出知识产权平均字数（每年知识产权类文章总字数除以知识产权类文章的数量），每年的知识产权类文章数量与全部新闻数量的比例， 每年的知识产权类文章的字数与全部新闻总字数的比例，然后将其放入DataFrame



**微博数据清洗**

1.用pd.read\_excel()读取上次实验获取的数据,以广东知识产权微博.xlsx为例，这里展示了前5条。



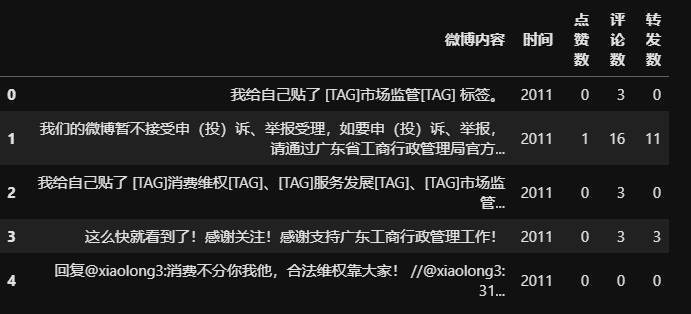
2．重新构建data1选取需要的列['微博内容', '时间', '点赞数', '评论数', '转发数']



3．首先对微博内容这一列进行处理，去掉其中一些不需要的字符串（如：'...'，

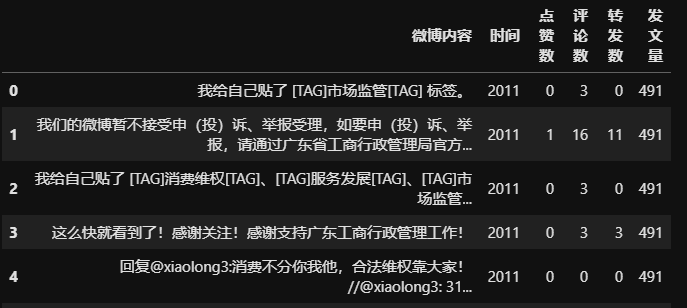
'\u200b'，'O网页链接'，'\u3000'），这里可以直接用向量化字符串函数。

4.对时间进行处理，只保留年份，用transform（）函数



5.计算这个微博每一年的发文量，在data1中新增发文量这一列，用对应年份的发文量进行填充。





6.分析该省微博对知识产权的重视程度，用知识产权平均字数（一年的知识产权类字数/这一年全部文章字数，知识产权类字数是指包含知识产权、专利、商标、版权和著作权5个关键字其中一个或多个的微博的总字数）衡量，在data1中新增知识产权平均字数这一列，用对应年份的知识产权平均字数进行填充。



7.最后，使用to\_excel（）函数将数据存下备用。

# 实验三 地区实证

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **地 点：** | B10 楼 | 208房； | **实验台号：** |  |
| **实验日期与时间：** | 2022/11/21 | | **评 分：** |  |
| **预习检查纪录：** |  | | **实验教师：** |  |
| **批改意见** |  | | | |

【实验内容】

1） 清洗并整合中国统计年鉴数据，与实验二的数据进行整合

2） 构建省级知识产权保护指数

3）知识产权保护对地区创新产出的影响实证

一、清洗中国统计年鉴数据并与实验二的数据进行整合

（1）中国统计年鉴数据清洗与整合（GDP数据）

1、读取GDP中“2009-2013地区生产总值.xls”并清洗，得到2009-2013年的gdp数据df1。



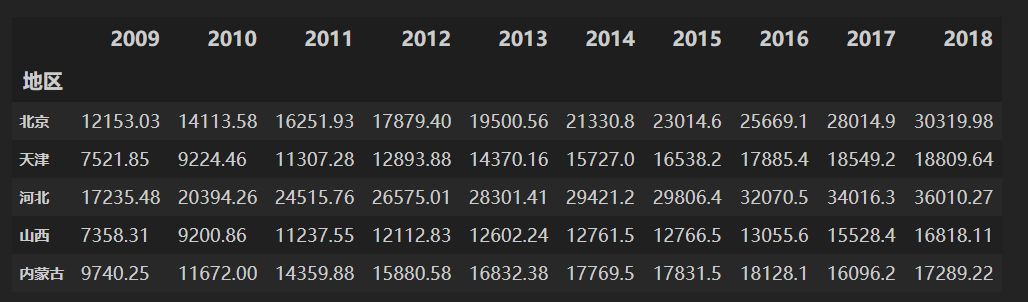
2、分别读取2014-2017年的地区生产总值excel文件，并清洗，得到2014到2017年的数据df2。（提示：2014年到2017年的数据存储格式是一样的，可以使用循环）



3、读取“2018各地区生产总值.xls”，提取2018年的数据df3。



4、将上述三个df合并得到综合2009-2018年的gdp数据,并存入“gdp.xlsx”（单位：亿元）



（2）实验二和实验三数据清洗及整合

1.读取“知识产权年鉴数据.xlsx”并存到df1



2.从知识产权年鉴数据中，选取需要的列['Unnamed: 0','Unnamed: 1','结案/立案','专利\_申请量','专利\_授权量','职务专利授权量百分比']，将列名重命名为['年份','地区','结案率','专利\_申请量','专利\_授权量','职务专利授权量百分比']

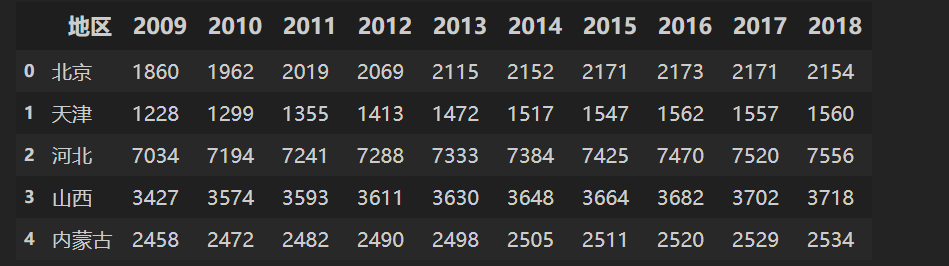


3. 将年份这一列的数据按照对应的年份（利用fillna里面的method参数）填充好



4.将‘结案率’中的无穷大‘np.inf’替换为0，将大于1的结案率换为1

5.读取“各省人口数（2009-2018）.xlsx”，得到各省人口数据df2，（单位：万人） 单位万人，但乘的是10万。



6.使用melt函数将数据变成如下格式：



7.将以上两个dataframe整合得到df。（注意删掉2008年的数据，因为缺少人口数据）



8. 将整理好的GDP数据（单位是亿元），各省进出口总额（单位：亿美元），全社会固定资产投入（单位：亿元），外商直接投资（FDI）（单位：万美元）做同样处理，整合到df。（本指南用的美元兑人民币汇率为6.5286）



二、自变量的构建

（1）知识产权保护强度指数IPRs

***IPRs =Fs\*IPRc***

其中IPRs代表省份S修正后的知识产权保护水平，Fs代表省份s的执法力度，IPRc代表国家层面的知识产权保护水平,为4.86。

执法力度Fs有三个衡量指标：1）知识产权行政保护水平，用专利侵权案件的结案率（结案/立案）衡量；2） 产权社会保护水平，用国内专利申请中有专业代理机构的比例（职务专利授权量百分比）来度量；3）民众知识产权保护意识，用人均专利申请量（省专利申请量除以省总人口数）表示。采用主成分分析，根据三个指标的解释百分比加权计算出各省知识产权保护指数。

1)计算人均专利申请量，用每年各省专利申请量除以省人口数。刚刚把结案率中>1的变成1.为什么又出现1.476190?



2）选出['结案率','职务专利授权量百分比','人均专利申请量']三列数据，利用PCA计算解释百分比（主要用到sklearn.decomposition里的PCA，和sklearn.preprocessing里的StandardScaler用于标准化）这个数据不存在，就算存在在前面也要求了清洗结案率。就算是拿结案/得出结案率。这个表格的也是错的。3的结案率是0，不是1.0625。不知道是为什么。



得到如下解释百分比即权重



对['结案率','职务专利授权量百分比','人均专利申请量']进行加权求和然后乘以4.86，得到省级知识产权保护指数。



（2）人均GDP



（3）经济开放度（进出口总额/GDP）



（4）GDP的哑变量：将数据按GDP的中位数分为两组，大于中位数的赋值为1，否则为0

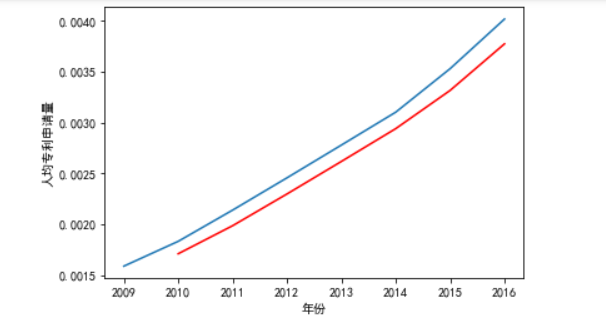


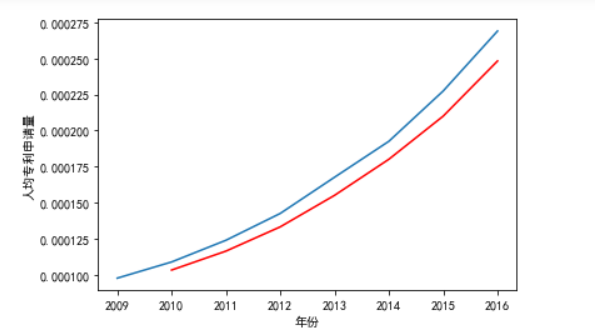
1. 知识产权保护对地区创新产出的影响实证

***1***描述性统计



1. 对每个省的人均专利申请量和人均专利授权量做时间序列预测， 不知道使用什么库，不会做。





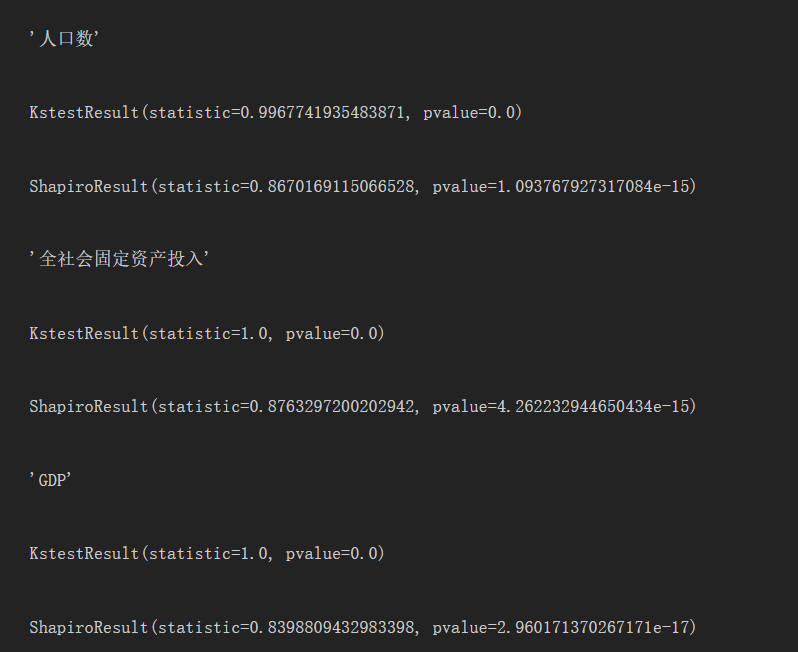
1. 正态性检验

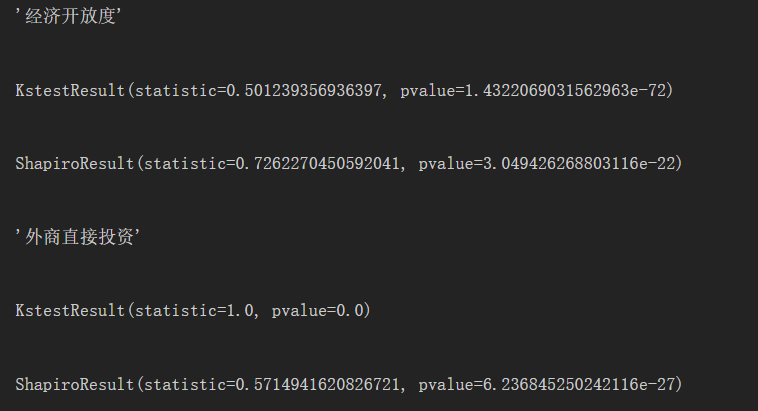
检验变量为人均专利授权量、人均专利申请量、知识产权保护指数、人口数、全社会固定资产投入、gdp、经济开放度和外商直接投资

**提示：**

1. 先导入相关库from scipy import stats，可以使用stats .sktest和stats . shapiro函数
2. 检验的零假设为变量服从正态分布。p值越小，变量越有可能不服从正态分布



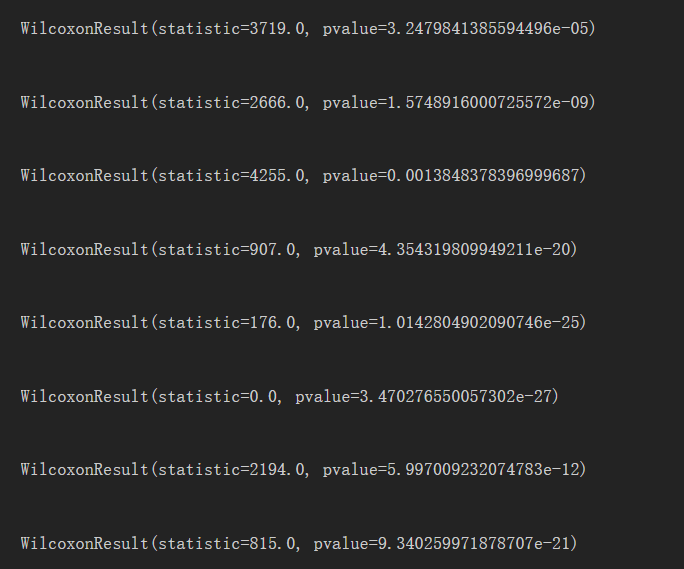




1. 非参数检验：wilcoxons检验

根据gdp哑变量分组，检验两组数据在人均专利授权量、人均专利申请量、知识产权保护指数、人口数、全社会固定资产投入、GDP、经济开放度和外商直接投资有无显著差异。

**注：**如果p值较小（比如小于或等于给定的显著性水平，譬如0.05）则可以拒绝零假设。使用stats.wilcoxon函数。

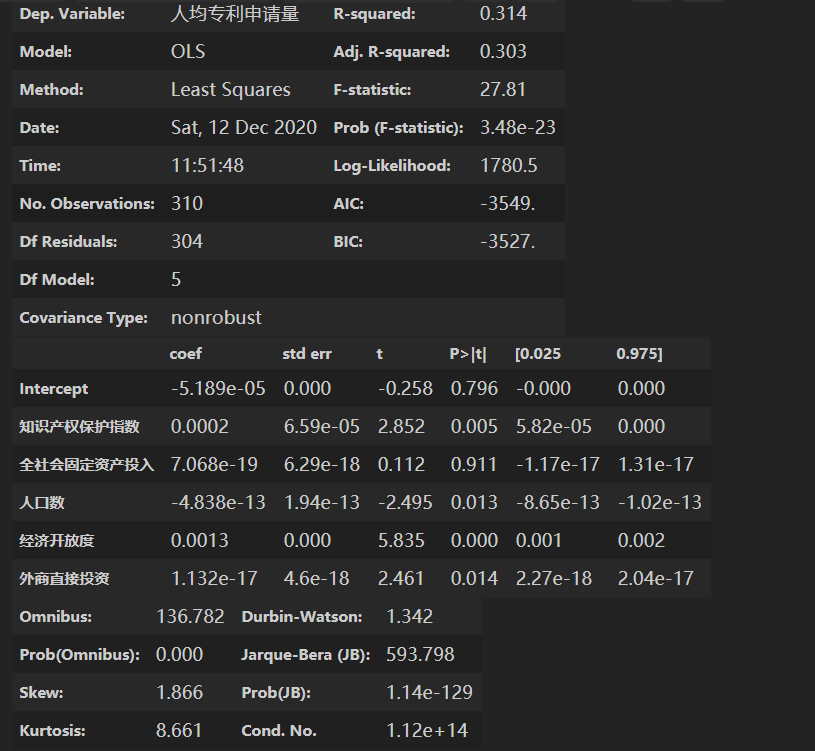


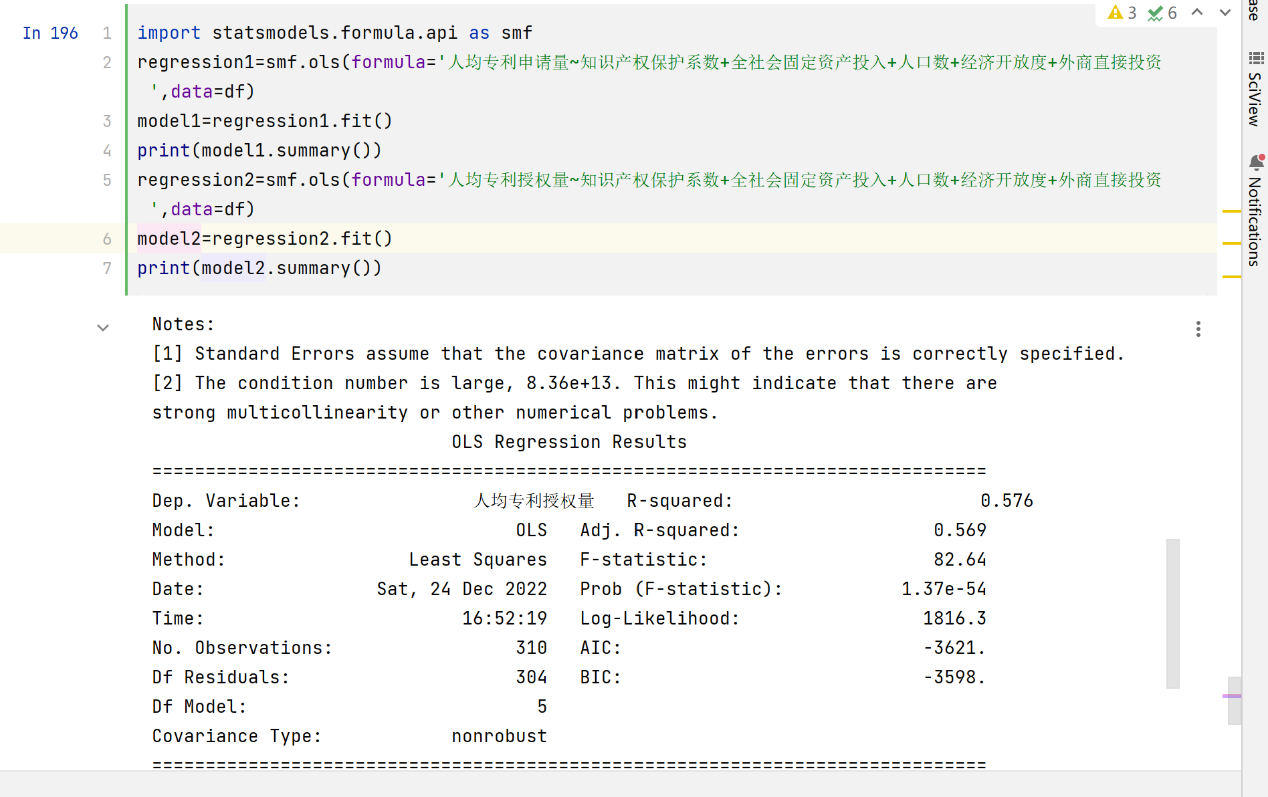
1. 相关性分析：检验变量两两之间的相关性



1. 回归分析

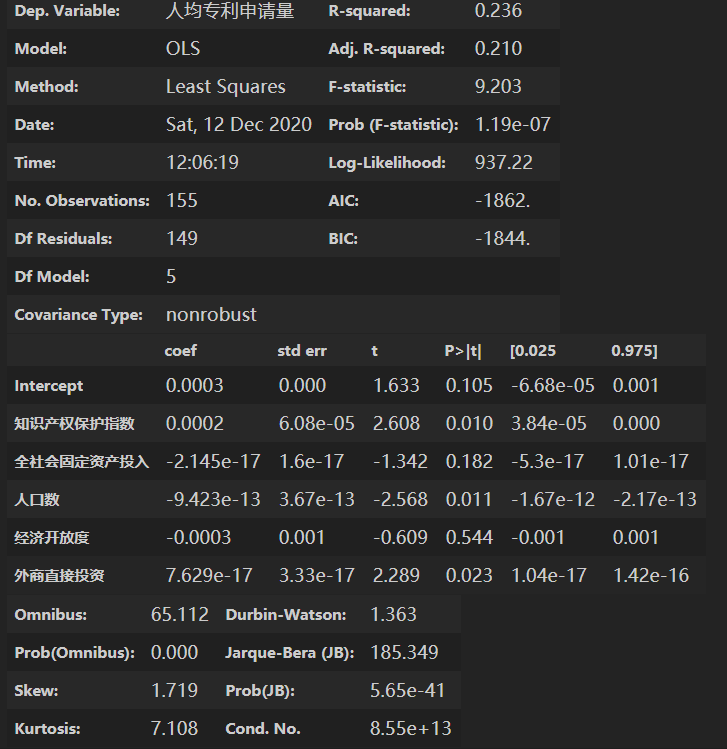
分别以'人均专利申请量','人均专利授权量'为因变量，以'知识产权保护指数 + 全社会固定资产投入 + 人口数 + 经济开放度 + 外商直接投资'为自变量，回归分析（提示：可以使用statsmodels.formula.api 函数，共两个回归结果，指南只截了一个图，**对回归结果进行分析**）

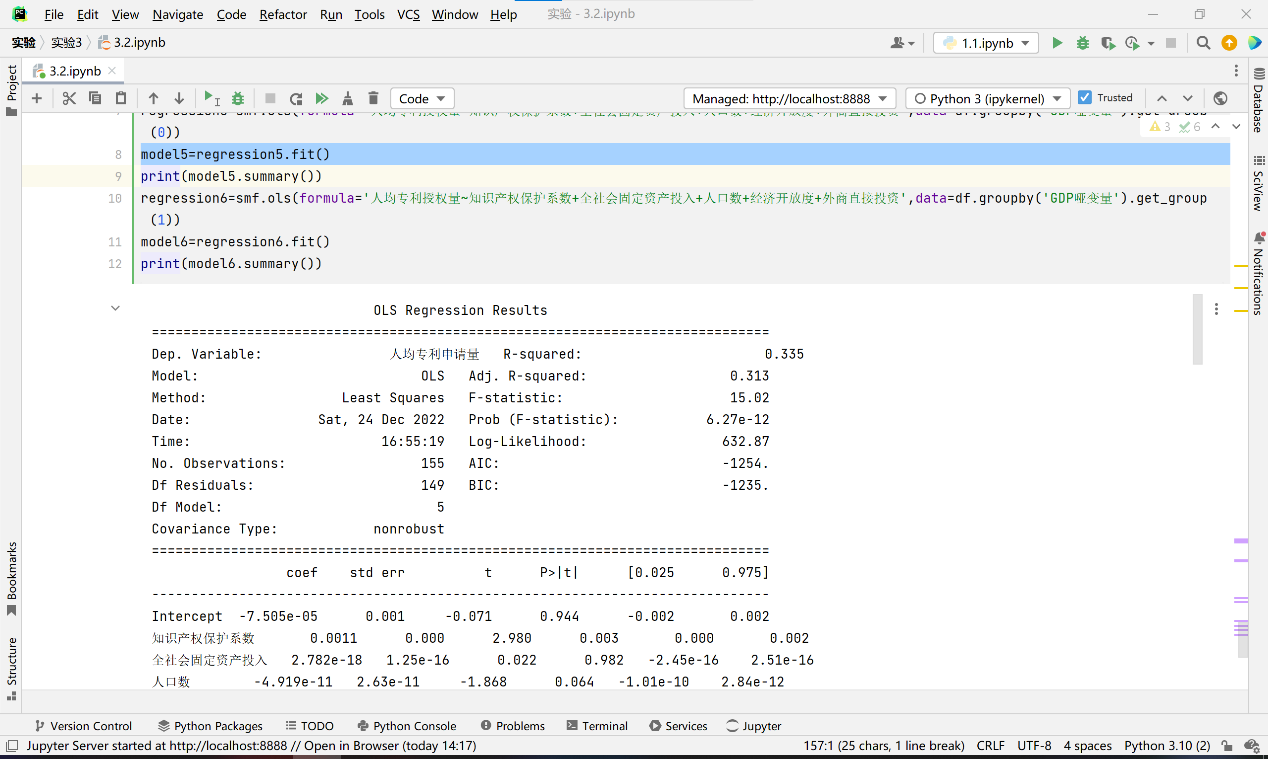


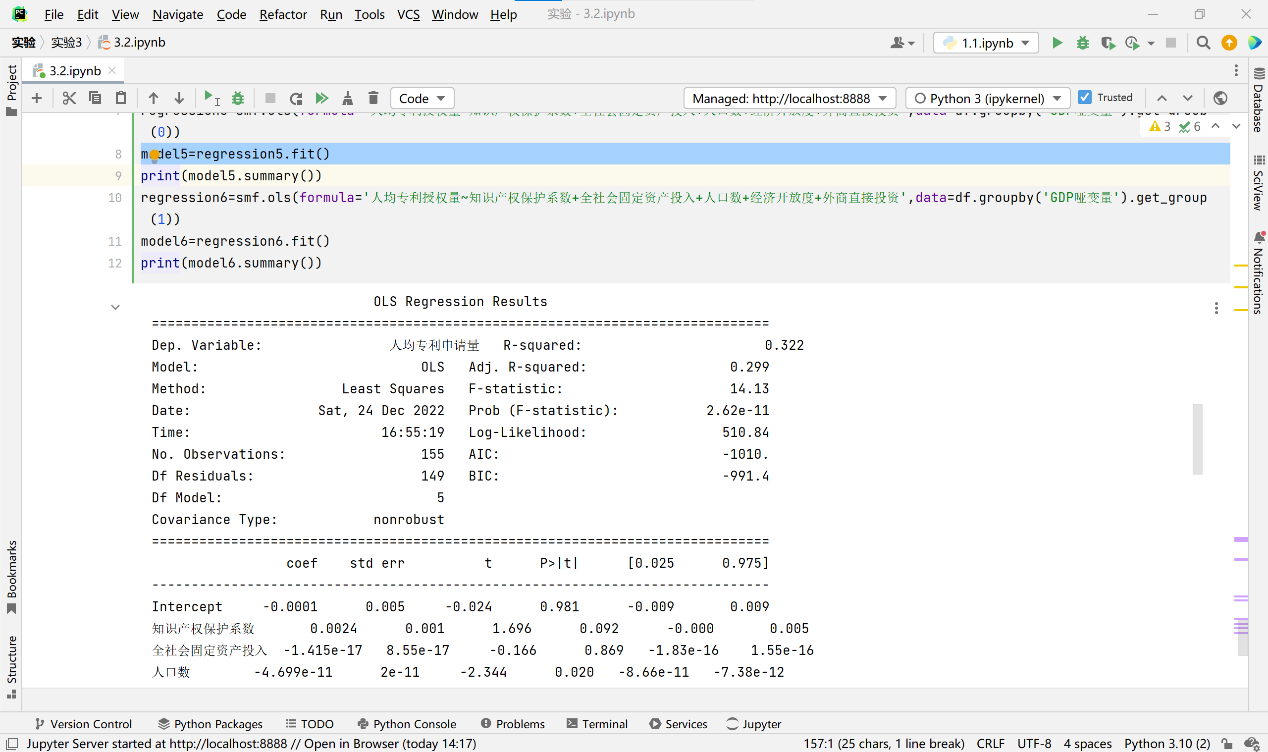


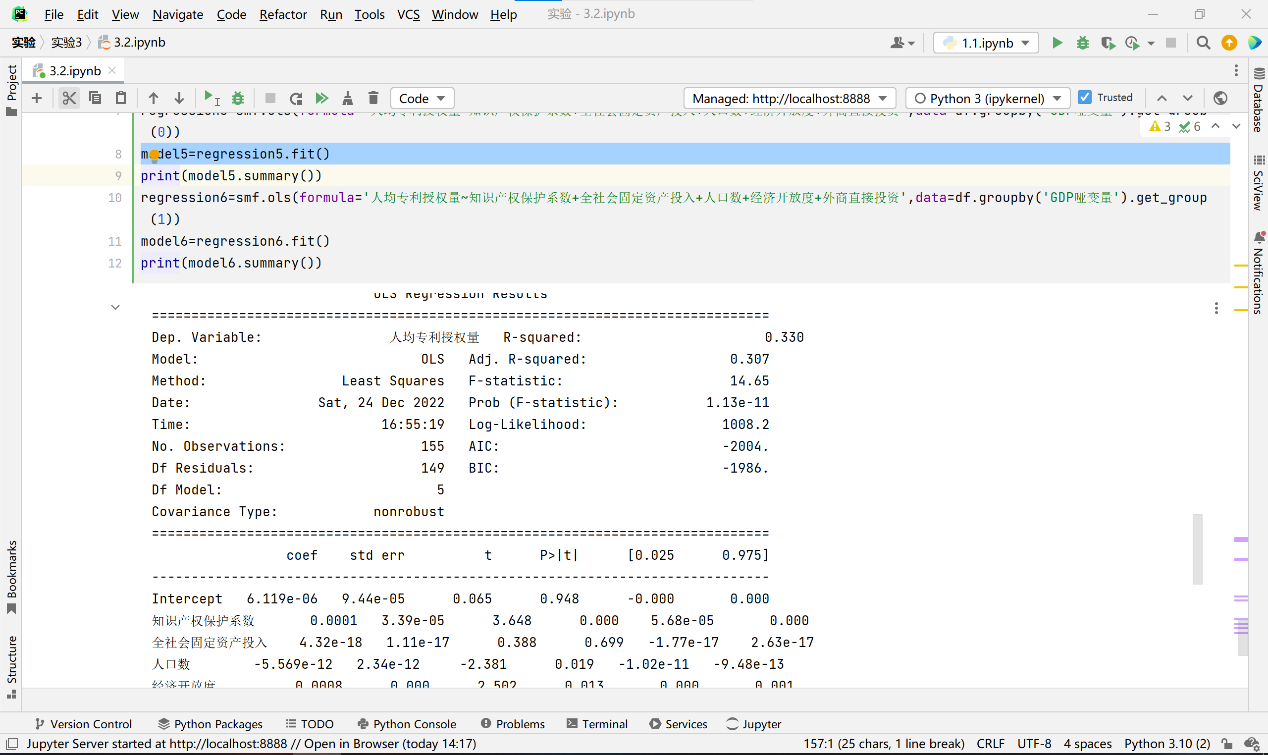
1. 分组回归

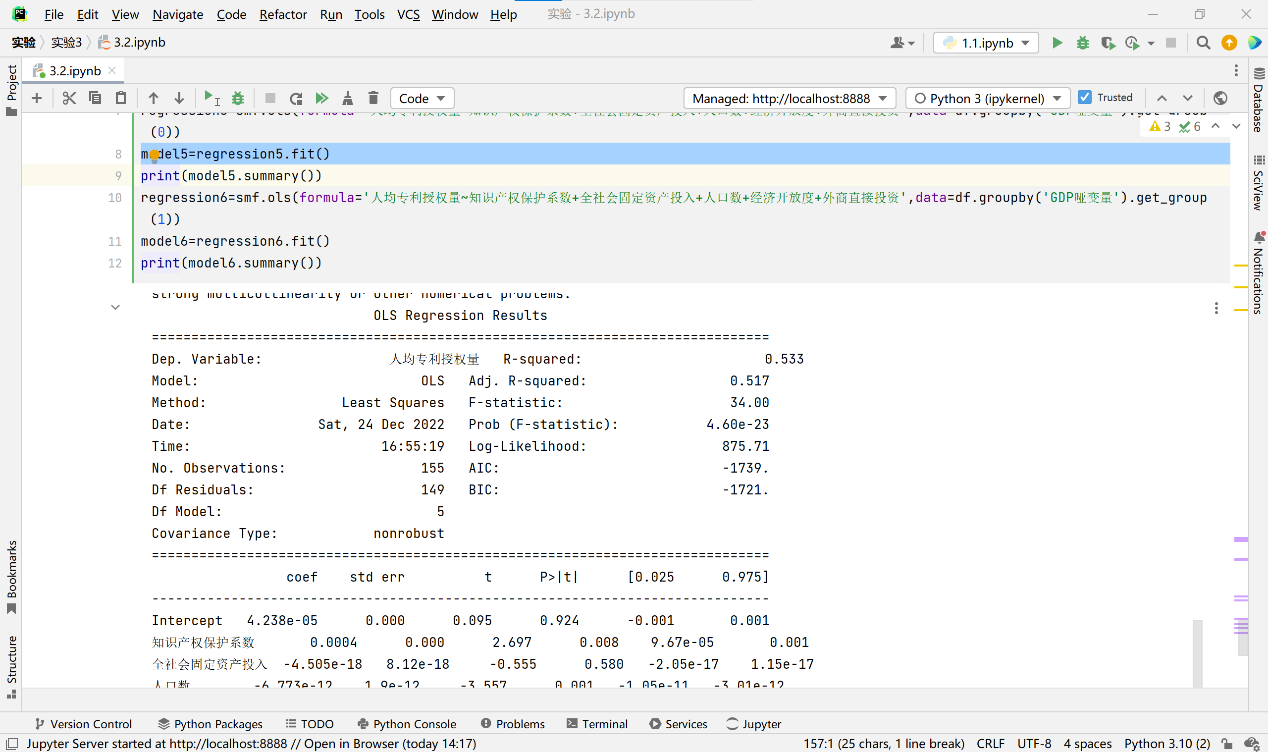
根据GDP\_哑变量分组进行同上的回归分析。（提示：共四个回归结果）











# 实验四 企业实证

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **地 点：** | B10 楼 | 303房； | **实验台号：** |  |
| **实验日期与时间：** | 2022/11/24 | | **评 分：** |  |
| **预习检查纪录：** |  | | **实验教师：** |  |
| **批改意见** |  | | | |

【实验目的及要求】

1. 企业基本面数据与知识产权数据可视化
   1. 从下载的不同格式数据中， 解释出包含但不限于企业知识产权数据、ROE/ROA数据、资产负债表率、总营业收入、无形资产、资产总额等数据，
   2. 以上数据与之前的数据进行整合，并选择合适的图例进行可视化
2. 回归实证

以企业的专利产出作为因变量，地区知识产权指数等其他数据作为自变量，对以上数据进行统计检验。实现截面数据回归

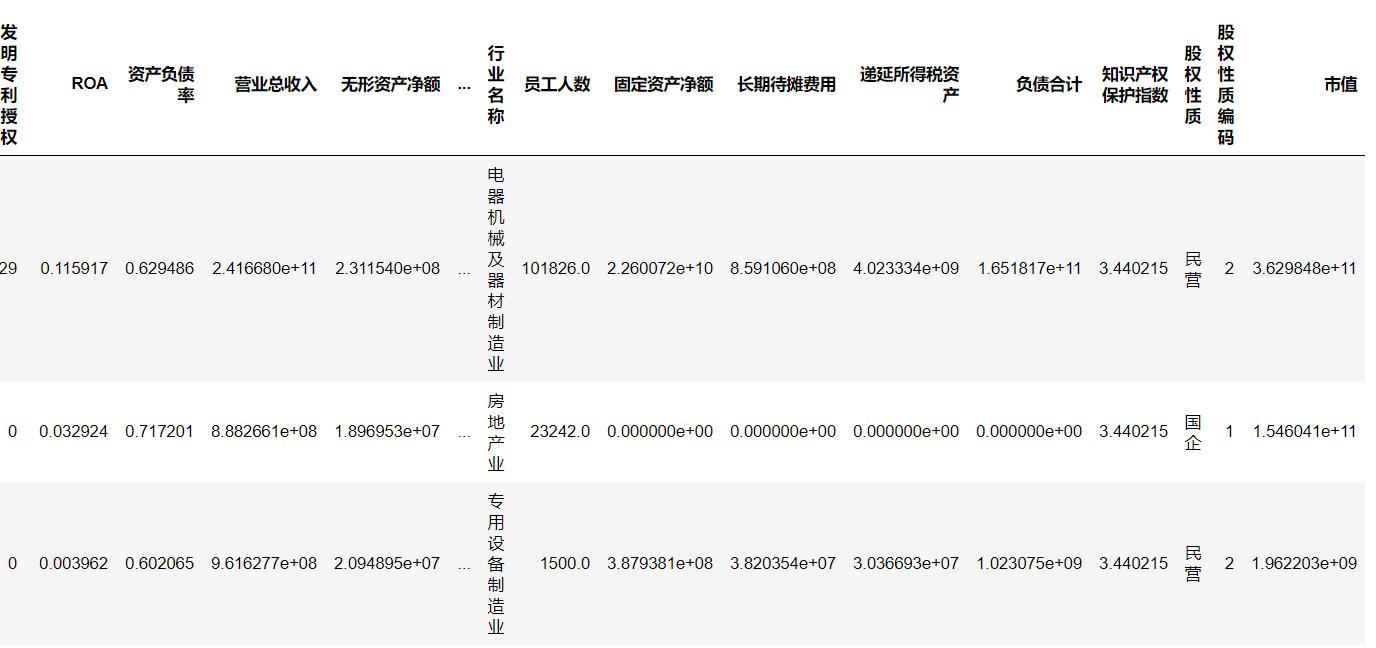
1. 企业创新100指数：

从专利密集型行业分类后，分别选择不同行业排名靠前的公司（共100家）作为企业创新100指数成分股，企业创新的绩效基准和股票价格指数

一、企业基本面数据与知识产权数据可视化

**1、公司市值**

使用国泰企业数据.xlsx，将公司市值数据整合：



**2、公司成立日期**

读取文件公司成立日期.xlsx，整合到上面的df，并删除没有2017年数据的公司所有记录，





**3、计算专利产出、Age、公司规模、托宾Q值、国有企业哑变量**

**专利产出：**用专利授权量除以资产总计，

Age：公司成立年限+1取对数（成立年限：成立日到截至日期，***同一个公司每年的成立年限是不一样的***）

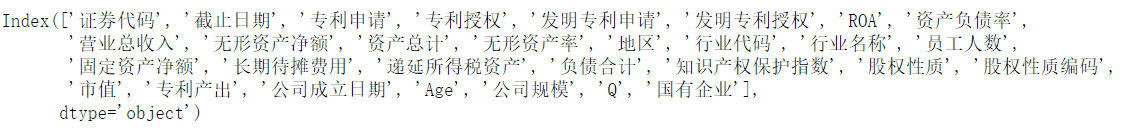
公司规模：为年末公司总资产取对数

托宾Q值：（股票市值+净债务）/有形资产现行价值，其中有形资产现行价值=资产总额-待摊费用-无形资产-递延税款借项）

**国有企业哑变量：1表示是，0表示否。根据股权性质编码，为1的是国有企业（注意有的企业有多个属性）**



4、整合后的数据包含如下列名：



二、回归实证

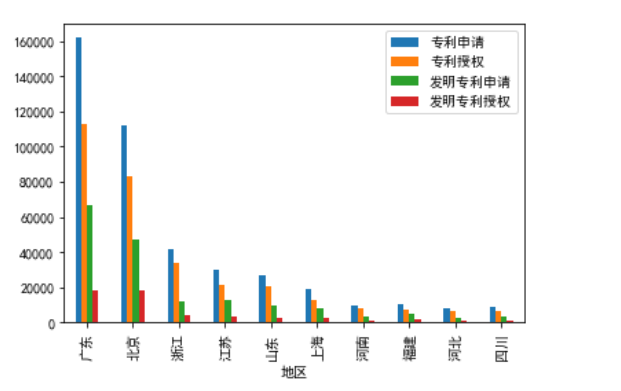
*用到的列为：['证券代码', '截止日期', '专利申请', '专利授权', '发明专利申请', '发明专利授权','专利产出', 'ROA', '资产负债率','无形资产率', '地区', '行业代码', '知识产权保护指数','市值','Age', '公司规模', 'Q', '国有企业']*

1、描述性统计：

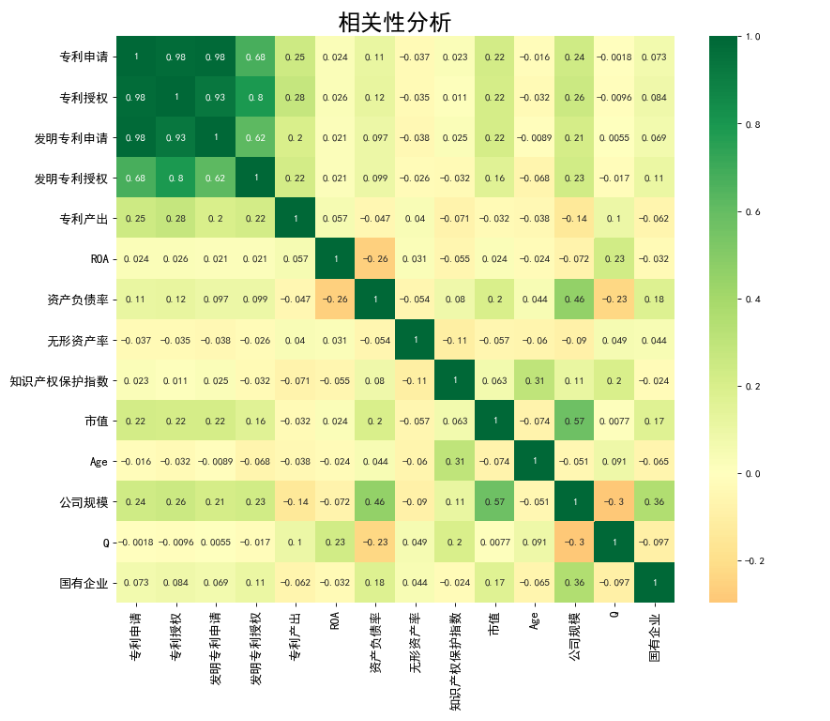


2、地区专利分布分析

*可以选择排名为前十的数据*

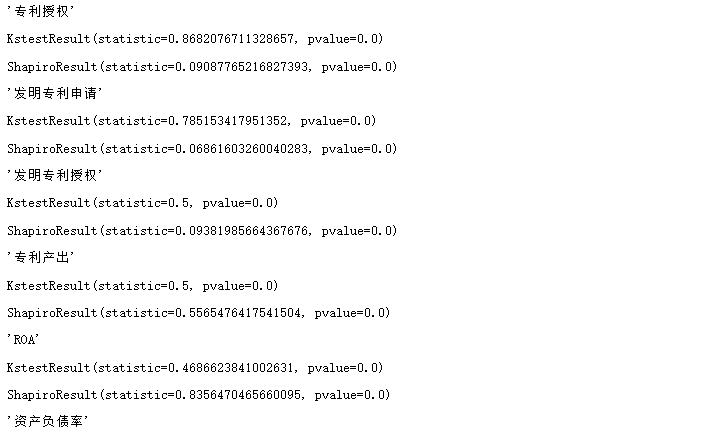


3、变量的相关性分析



4、正态检验

['专利申请', '专利授权', '发明专利申请', '发明专利授权', '专利产出', 'ROA', '资产负债率', '无形资产率', '知识产权保护指数', '市值', 'Age', '公司规模', 'Q']

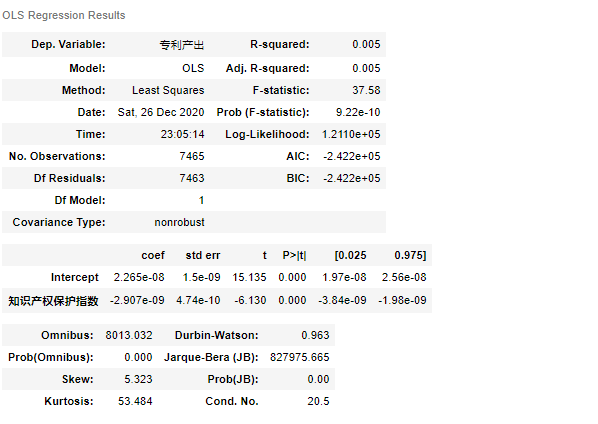


5、因变量为专利产出，与知识产权保护指数进行回归分析，分析相关性，

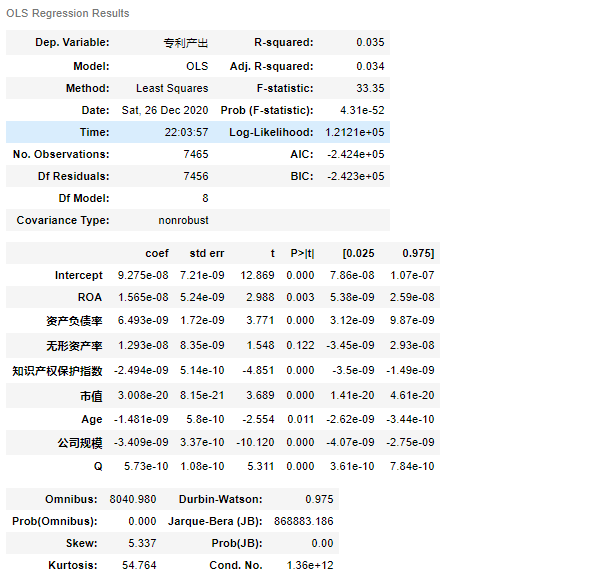
按照国有变量进行分许，再分析相关性。

分析企业性质对专利产出的影响。

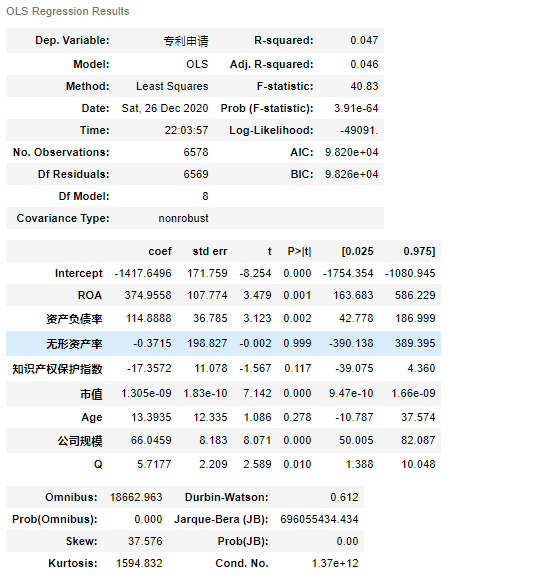
国有企业的专利产出较私营企业较差。



6、以'专利产出'作为因变量，用'ROA', '资产负债率', '无形资产率', '知识产权保护指数', '市值', 'Age', '公司规模', 'Q'，作为因变量进行多元回归分析。



7.根据国有企业分组分别进行同6的回归分析（应该有2个图分别进行分析）分析企业性质对专利产出的影响。

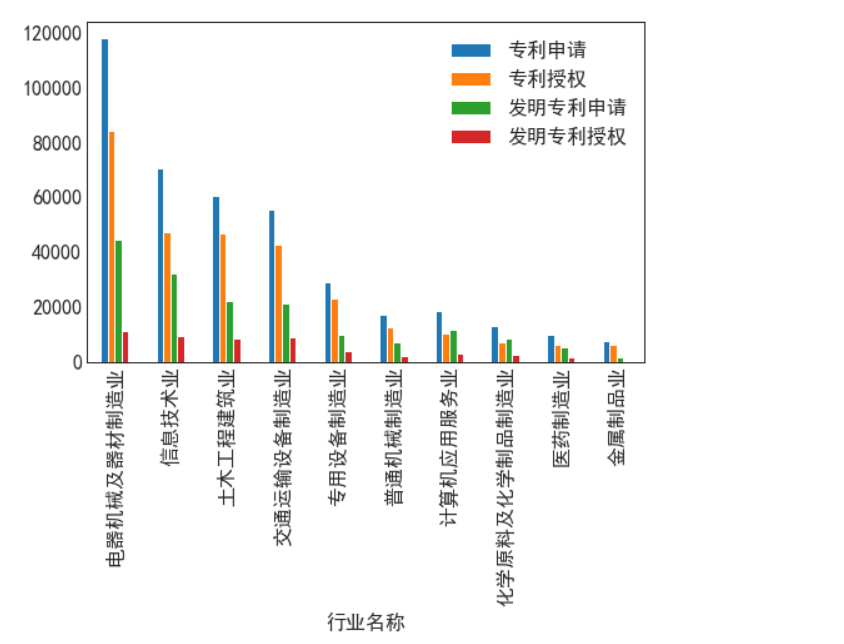


三、企业创新100指数

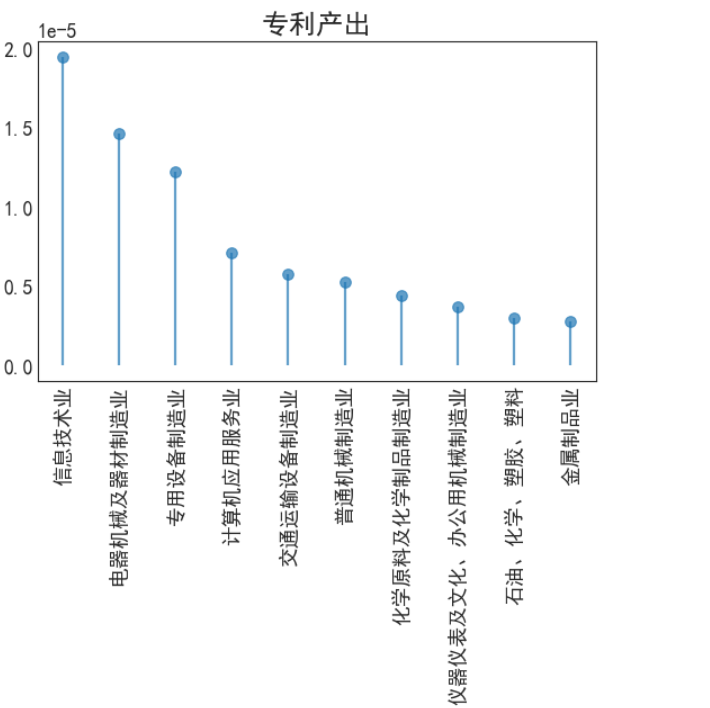
从专利密集型行业，分别选择不同行业排名靠前的公司（共100家）作为企业创新100指数成分股，构建企业创新的绩效基准和股票价格指数

***以下策略仅供参考（可以考虑更多的变量进行股票的选择和权重的确定）***

***1、总共63个行业，找出专利数据排名前10个行业数据，并画图***

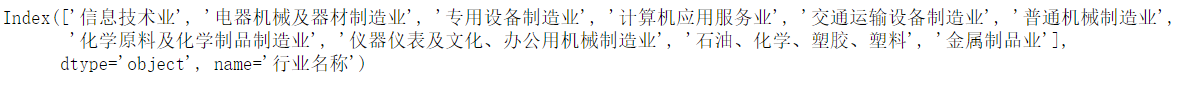


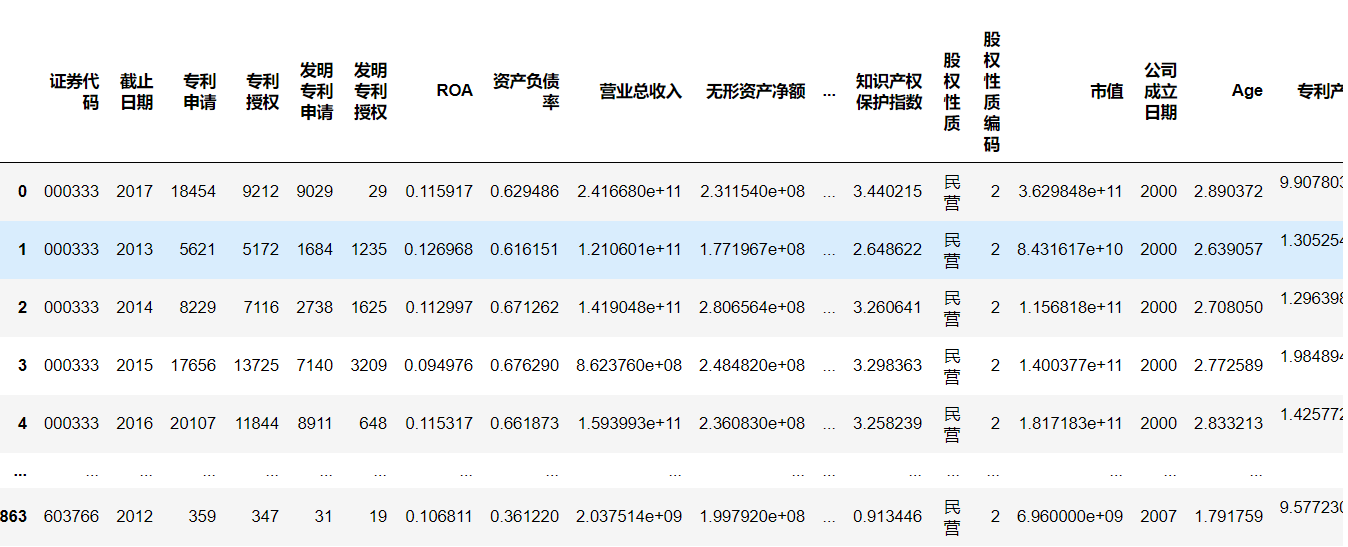
***2、找出专利产出前10名的行业，绘图***



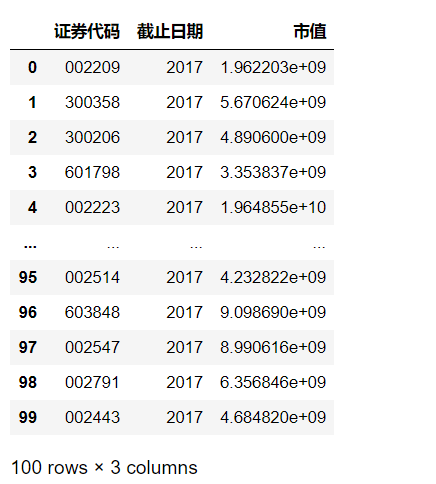
3、***编写自己的选股策略***

***1）找到原来数据中行业名称为上面的10个行业的数据***





***2）找出每个行业专利产出前10的数据***



***3）按照每只股票2017年该股票市值占100只股票市值总和的比重构建权重，得到创新100指数***



批改意见：