吐血的给你写一份数据使用说明，我下载下来已经要吐血了。。。。。

我不知道python咋回归，但是应该还是要整理成面板数据的，但是这里有一个问题，我下午要问老师的，你要是有空可以先想想。就是非平衡面板数据和平衡面板数据，我看了很多关于上市公司的研究最后都是把非平衡面板数据处理成平衡面板数据了，我不知道是不是一定要这样做。他的做法应该是把不是每个截面上都有数据的样本删掉了，然后缺失部分指标的样本也删了，最后使得每个截面上的个体是一样的。就像有的公司后来才上市，不是在2011-2018年整个跨度都有数据的，那这种样本删掉，然后对于部分指标缺失这种情况我有点疑惑，刚才大概看了下数据，发现政府补贴和绿色专利都会存在某个截面上为0的情况，那这种会不会造成影响还是应该把这种数据删掉，就感觉，哦哦对这个可以通过控制年份效应去实现。

我尝试用vlookup通过股票代码去匹配了，简直吐血，然后我数据透视表用的不是很好，实在是弄不成面板。。。。。

压缩包里有一份制造业上市公司股票代码excel（8），初步样本是2458个，但后面处理后不知道还会剩多少个样本，我觉得按之前的文献来看，最后剩下个800-1300个公司都是合理的。然后这里面除了**绿色专利（4）**其他的数据都是我从国泰安数据库下载的，在下载之前已经筛选了字段，只挑了对这个研究直接或间接有用的字段，然后每个压缩包里的txt文件是对各字段的含义、赋值、单位的说明。（压缩包4是单独下载的，不是从数据库，所以里面包含了全体上市公司，要筛选一下，筛选出A股制造业上市公司，根据文件9）

~~压缩包（1）董监高个人特征文件：年龄、性别、受教育程度（学历）~~

~~这个麻烦的是，这个要先把它变成截面数据，把一个时间一个公司的所有高管的年龄啊什么平均下，你看那个变量定义表就行。这个表下载下来的时候已经筛选过了，这里面都是高管。但是压缩包2无法筛选，就是还包含高管以外的人。所以最好压缩包2按照1里面的人来匹配。~~

~~压缩包（2）董监高任职情况表：任期~~

~~压缩包（3）公司基本情况文件：成立日期~~

**压缩包（4）绿色专利数据：绿色专利授权量（获得量）这份数据里面的公司不仅仅是制造业，所以要根据文件（9）去筛选一下。**

压缩包（5）政府补助：看每年合计数据那一行就ok

~~压缩包（6）利润表：营业收入、净利润。这份文件因为涉及到营业收入增长率的计算，所以时间跨度是2010-2018而不是2011-2018，研究设计中有计算公式，营业收入增长率=（本年营业收入-上年营业收入）/上年营业收入，然后注意下财务报表都是年末12.31公布的。~~

~~压缩包（7）资产负债表：总资产~~

~~压缩包（8）中国上市公司股权性质文件：股权性质及编码。与研究设计中的不同，要转换下~~