关于此次任务的结果汇报

代码思路分析

代码方面主要是写了一个模块(data\_Refreshing\_module)，将各类方法整合后以直接引用的形式处理各个省市的数据。

我们需要收集和处理的数据是在各个省市各地区一年度的各个维度的数据，每一个维度的数据经过手动收集后集成多个格式相同的excel表格，再用python进行数据的清洗工作。

**下面分析data\_refreshing\_module模块的各个方法。**

def combine\_two\_dataframe(df1, df2):

首先合并两个excel表的内容。这里有两个点需要注意，一个是合并的是维度数据而不是时间跨度的数据，因此需要横向合并数据的列。第二点是excel表会重复表头，也就是各个地区的列，因此在合并完成后需要将重复的列删除其一。

def batch\_process\_csv(dir\_str):

该方法将会用来批量合并各个维度的excel表格，避免代码的重复和累赘，这里将会用到combine\_two\_dataframe方法。这里用到os库，定位到放置所有维度csv文件的的文件夹下，利用循环一一读取所有csv文件并合并成一个表格后返回。

def city\_code\_assign(df):

由于数据的组织形式，该方法并没有用到。方法用于给每一个地区附上它们的编码，结果放置于新的最后一列。由于处理数据的形式，每一个城市或区县只出现一次，因此人工赋值效率更高。

def rank\_city\_code(df):

方法用于按城市的编码给所有城市排序。当人工赋值完成之后，使用该方法给所有地区或城市进行按照编码的排序。

def missing\_data\_filling(df):

方法用于按照要求为所有excel表中的缺失值进行补全。

**下面是对特定省市数据清洗代码的分析。**

首先用到的sys库导入data\_refreshing\_module作为临时库使用。定位到所有维度csv文件所在的文件夹的位置之后用batch\_process\_csv方法返回合并所有数据的dataframe并输出csv文件，命名为raw result。

将该csv文件进行人工赋值各地区或城市的编码后，重新读入成dataframe。使用rank\_city\_code和missing\_data\_filling方法进行排序和缺失值处理。最后的输出是清洗完成后的数据，命名成finished result.

数据收集

北京数据问题汇总:

1. 北京-资源中”人均水资源”按常住人口年平均数计算。2006-2010年数据根据全国第六次人口普查进行了修订。各区的资源数据也没有被挖掘到.
2. 北京各区的土地面积通过常住人口密度和常住人口数量不精确计算。
3. 北京的总就业人口应该是不包括个体户或法人单位等。
4. 北京的健康维度无法收集到各区养老机构数据
5. 北京的教育维度由于各区高校,职校,幼儿园等数据统计不一致,最终只选取普通中学和小学.
6. 北京农业无法搜索到灌溉面积,上海地区也是同样无法查找到。
7. 北京人口维度中的老年人口和未成年人人口数量是没有的，因为北京地区年鉴中对此数据的收集标准不同，例如北京市年鉴统计年龄段以65和17为界线。
8. 北京生活维度中农村数据不具备。