数据可视化

基本步骤

1.新建 datafram
2.从 txt 文件中读取，json 格式的数据。
3.删除新建时的空行(首行)
4.用.info 方法查看数据格式
5.用.head 方法查看数据内容
6.定义问题：质量问题和清洁度问题
7.备份和清洗数据
7.5 迭代定义问题和清洗
8.可视化分析
9.5.迭代定义问题和清洗

关键步骤的详细内容：

1.从 txt 文件中读取数据

数据无法从一次性直接转换成字典类型，因为数据是一条一条写进文件，以换行符分隔，最后用遍历的方式一条一条的读取出来（for line in open():），然后用 json.loads()将字符串加载为字典类型。并入 dataframe 的语句为：dataframe.append(dict,ignore_index=True)。忽略索引，在 dict 中就可以不需要考虑索引，在 append 的时候自动创建。

2.定义问题：质量问题和清洁度问题

质量问题：指的是缺失值、异常值、无效值等数据完整性、有效性、准确性、一致性的内容层面问题
清洁度问题：指的是：
（1）记录/变量重复:行列内容重复或者可以从另外的行列推导计算出
（2）变量/值混用：以值作为列名或者以变量作为一个值，具体值在另外一列出现
（3）单个表中存在多种观测类型
（4）一种观测类型存在多个表当中

3.备份和清洗数据

个人认为备份数据的目的，在于快速地推到重来，不用来浪费时间再去整理一次。
清洗数据主要有以下几个步骤：
① 确认数据质量问题，例如：
1.修改部分列的数据类型
2.删除错误提取的内容
3.填充空值为 None
② 确认数据整洁度问题：

1.删除无法利用的列

2.拆分提取的分数列为分子和分母

③ 逐个处理数据问题：

优先级为：数据整洁度问题 > 质量问题

数据质量问题：完整性 > 有效性 > 准确性 > 一致性

数据整洁度问题： 重复行 > 需要计算的列 > 重复列 > 记录值是变量或者列标题为数值(需要归位)

第一步将无效的或者不需要的数据给剔除，将需要的数据补齐；第二步进其它洁度问题的处理；第三步是数据质量问题中的有效性、准确性和一致性。在进行变量或数值归位之后，也许需要计算新变量。计算新变量之后，很可能需要删掉某些列。这些都都可以延后处理，等到迭代再进行，因为很可能后面还需要用到此处看似多余的列。

关于数据类型的转换问题 ,可以统一在最后进行转换 ,若影响数据处理 ,则在数据处理之前进行转换。

数据质量问题的含义：

完整性：对应数据集的数据量是否相同或者近似

有效性：极端值、异常值、缺失值，看看是否符合列数据的规范

准确性：符合规范，但是无法反映真实情况，例如在数据取值范围内，但是却是错误的值。

一致性：有效且符合规范，格式是否一致。(几乎)

数据整洁度问题，按字面意思理解即可，不理解参考这里

④迭代前三个步骤：数据清洗过程中很可能会有新的认识和发现，需要对数据问题进行补充和处理。

4.可视化分析

连续数据的分析：

① 描述统计：df_status.describe()，集中趋势、离散趋势的度量

② 箱线图：dataframe.boxplot() Series.plot.box()，集中趋势、离散趋势的直观理解

③ 散点图：

可以直观感受，连续变量随着时间的分布、两个变量之间的相关程度。要注意的是 X_Y 轴的数据要适当地进行归一化 (比例尺相同，相关关系才准确)

相关语句：

```
a.plt.scatter(x,y,c=color,s=scale*100,alpha=0.9,edgecolors='none',cmap=cm)
b.dataframe.plot(y='label',ylim=[0,2],style='.',alpha=0.5,figsize=(12,6))
```

分类数据分析：

① 条形图：展示各个分类数量分布，要是数量太多就横着放，会更容易观

```
status.plot(kind='bar')
df_img.p1.value_counts().head(10).plot(kind='barh')
```

② 气泡图：在散点图的基础上加颜色，大小。可以进行 4 个维度的分析。其中 X,Y，大小必须是连续型数据，颜色可以是连续也可以是分类数据。

总结：数据清洗和可视化分析是可以系统化的，从数据问题分类着手进行数据问题的界定和处理。从变量类型和变量个数着手，进行数据可视化，考虑变量各种组合就可以探索不同的变量间的关系。