实验八：pandas数据处理与分析

一、实验目的

1. 掌握对缺失值数据处理的方法。

2. 掌握主键合并的几种方法。

3. 掌握数据标准化的方法。

4. 掌握时间信息提取的方法及其时间数据的算术运算。

5. 掌握分组聚合的处理方法与步骤。

二、实验内容与要求

1. 用户用电量数据呈现一定的周期性关系，missing\_data.csv表中存放了用户A、用户B和用户C的用电量数据，其中存在缺失值，请使用Scipy库中interpolate模块中的拉格朗日插值法对缺失值进行插补，并检测插值操作是否成功。

2. 线路线损数据、线路用电量趋势下降数据和线路告警数据是识别用户窃漏电与否的3个重要特征。请读取ele\_loss.csv与alarm.csv表，查看两表的形状，以线路编号ID和时间date两个键值作为主键进行内连接，查看合并后的数据。

3. 算法的种类非常多，当涉及到空间距离计算、梯度下降等，就必须进行标准化处理。读取model.csv，定义标准差标准化函数，使用函数分别对线路线损特征、线路用电量趋势下降特征、线路告警特征进行标准化，查看标准化后的数据。

4. 用户信息更新表Training\_Userupdate和登录信息表Training\_LogInfo中均存在大量的时间数据，提取时间数据内存在的信息，一方面可以加深对数据的理解，另一方面能够探索这部分信息和目标的关联程度。同时用户登录时间、借款成交时间、用户信息更新时间这些时间的时间差信息也能够反映出P2P网络贷款不同用户的行为信息。（1）使用to\_datetime函数转换用户信息更新表Training\_Userupdate和登录信息表Training\_LogInfo的时间字符串。（2）使用year、month、week等方法提取用户信息更新表Training\_Userupdate和登录信息表Training\_LogInfo中的时间信息。（3）计算用户信息更新表Training\_Userupdate和登录信息表Training\_LogInfo中两时间的差，以分钟计算。

5. 分析用户信息更新表Training\_Userupdate和登录信息表Training\_LogInfo时，还可以结合用户编号进行分组聚合，然后进行组内分析。通过组内分析得出每组组内的最早和最晚信息更新时间、最早和最晚登录时间、信息更新的次数、登录的次数等信息。

三、实验作业提交

1. 实验完成后，检查是否达到实验目的。

2. 请将实验源程序打包为一个压缩文件，压缩文件请以“实验8源程序+学号+姓名”命名，如：实验8源程序001夏芸.rar。

3. 请撰写实验报告，实验报告文档以“实验8报告+学号+姓名”命名。

4. 将实验报告与源程序分别提交到ftp://172.16.94.8/作业/夏芸/Python数据分析/实验8中的实验报告与源程序文件夹。