任务1.1

TSLA美股数据预处理（2012.4.17-2022.4.17共10年数据）

数据相关信息及实验要求：

1. 数据为csv格式，每一列分别为日期（Date），收盘价（Price），开盘价（Open），当日最高价（High），当日最低价（Low）。

2. 股市不是每天开盘的（周末和法定节假日休市）。非交易日的各项数据为NaN，请使用编程的方式删除该数据，并在屏幕上打印出删除行数的统计信息。

3. 源数据应当每个交易日存在一条记录，请使用编程的方式检查并删除日期重复的数据，并在屏幕上打印出删除行数的统计信息。

4. 在中国，股价的波动范围为通常为±10%（即涨停和跌停），但在美股中没有这项规定。请使用编程的方式计算在10年间共有多少次股价单日涨跌幅度超过±10%。具体的计算公式为：(今日收盘 - 昨日收盘) / 昨日收盘 × 100%。请在屏幕上分别打印出总计天数的统计信息。

5. 股价在一天之内往往有高低波动，请分别计算每日股价最高与最低之间的差值，存在Diff列中，与经过前述数据清洗后的结果合并后保存为新的csv文件。

注：数据源来自证券公司，为本门课程实验添加了适量脏数据，模拟实际应用中可能出现的情况，以供同学们练习。

任务1.2

鸢尾花数据预处理（样本包括约150组正常+10组异常数据）

1. 鸢尾花有四个特征，分别为：萼片长度（Sepal Length）、萼片宽度（Sepal Width）、花瓣长度（Petal Length）、花瓣宽度（Petal Width），单位为厘米。

2. 鸢尾花分为山鸢尾（Iris Setosa）、杂色鸢尾（Iris Versicolour）和为吉尼亚鸢尾（Iris Virginica）三种，在数据集最后一列（Class）分别以0、1、2三种标签表示。

3. 请将数据有缺失的行删除，请将出现明显错误数据的行删除（负数等）。

5. 求出每种鸢尾花萼片长度的平均值、中位数和标准差。

6. 归一化：创建一种标准化形式的鸢尾花特征数据，其值正好介于0和1之间，这样最小值为0，最大值为1。

7. 将数据清洗及归一化后的结果保存为新的csv文件。