数据集格式说明

**一、二手车平台数据集**

（一）基本信息

1. 数据源：

第一车网（https://www.iautos.cn/）、二手车之家（https://www.che168.com/）、车三百（https://www.che300.com/）。

1. 数据规模：
2. 第一车网：

最终清洗后数据共1540条，共14个字段。

1. 二手车之家：

最终清洗后数据共1178条，共14个字段。

1. 车三百：

最终清洗后数据共2567条，共17个字段。

1. 处理后：

整合后数据共7622条，共23个字段。

1. 数据类型：

本数据集均为结构化数据。

（二）字段说明

各字段名及含义为：

1. **车型标题：**字符型，包含品牌、车系、具体车型和年款等信息，有助于确定车辆的具体型号。
2. **上牌时间：**日期型，格式为“XXXX年XX月”，可用于分析车辆使用时长对保值率的影响。
3. **里程：**数值型，以万公里为单位，是车辆使用程度的重要指标。
4. **售价：**数值型，精确到具体金额，单位为万元，直接反映二手车辆价格。
5. **车辆所属地：**字符型，具体到市级行政区，可能与地区经济、市场需求等因素对保值率的影响有关。
6. **新车含税价：**数值型，用于计算车辆的折旧情况，对分析保值率至关重要。
7. **排量、油耗、变速箱、车型级别、车厢结构、能源类型：**这些数据类型多样，能从不同角度反映车辆的特性，可能与保值率存在潜在联系。

此外，数据整合后，定义并加入了以下新字段：

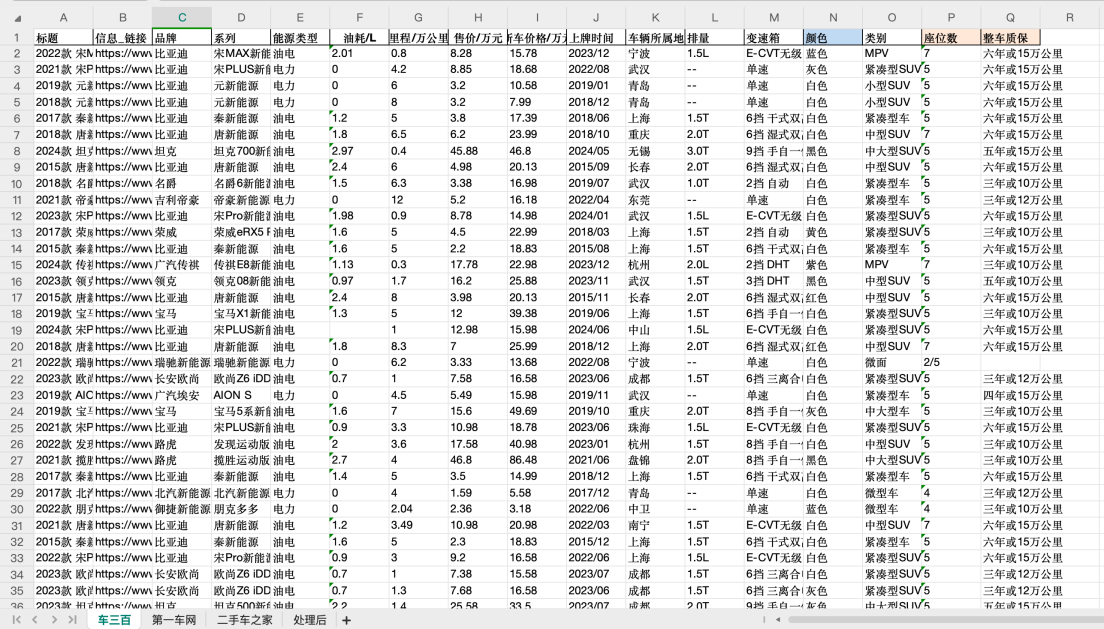
1. **折旧比：**根据valueLoss和newCarPrice计算折旧比，公式为折旧比 = valueLoss / newCarPrice。在计算过程中，对于newCarPrice为0的情况，将对应的折旧比设为NaN，以避免除零错误。
2. **使用时长：**根据regDate和当前日期（假设为2024 - 12 - 09）计算使用时长，单位为天数，通过(current\_date - regDate).dt.days计算。
3. **年平均折旧额：**根据valueLoss和使用时长计算每年平均折旧额，公式为每年平均折旧额 = valueLoss / (使用时长 / 365)。对于使用时长为0的情况，将对应的每年平均折旧额设为NaN，避免除零错误。
4. **使用强度：**根据mileAge和使用时长计算里程使用强度，公式为里程使用强度 = mileAge / 使用时长。同样，对于使用时长为0的情况，将对应的里程使用强度设为NaN。
5. **保值率变化率：**根据resaleValueRate和使用时长计算保值率变化率，这里简单使用差分计算，即保值率变化率 = resaleValueRate.diff() / 使用时长.diff()。对于第一个值为NaN的情况，使用bfill方法进行后向填充。

（三）数据格式与存储结构

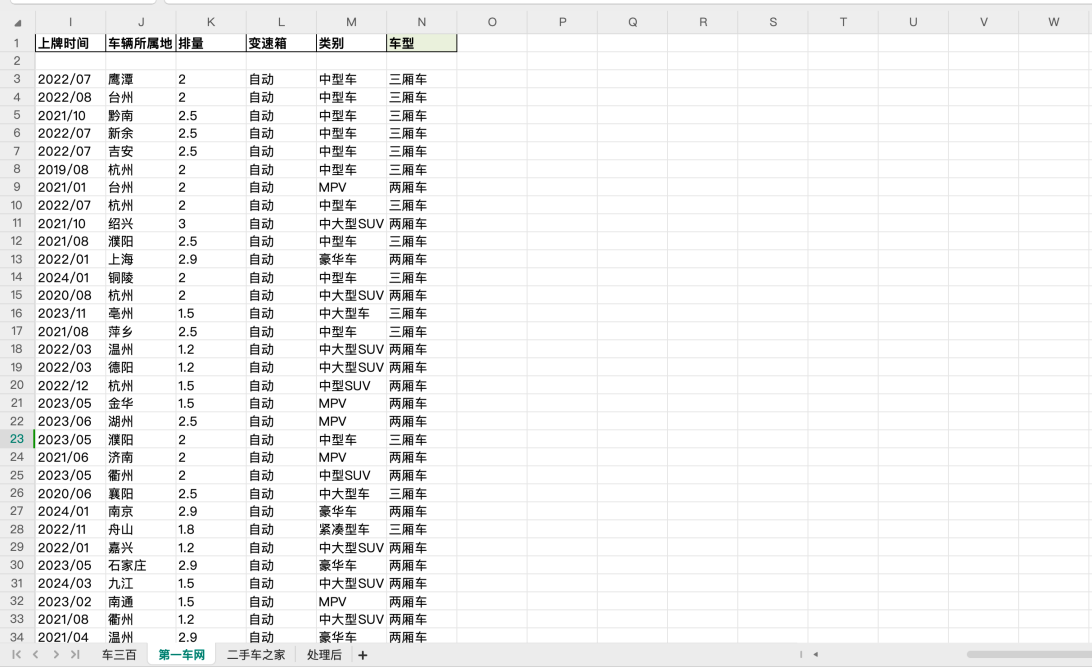
1. 文件格式：本数据集存储于Excel文件中，格式为“.xlsx”；
2. 文件结构：文件中包含字段行，字段行后为具体数据行。

（四）示例数据

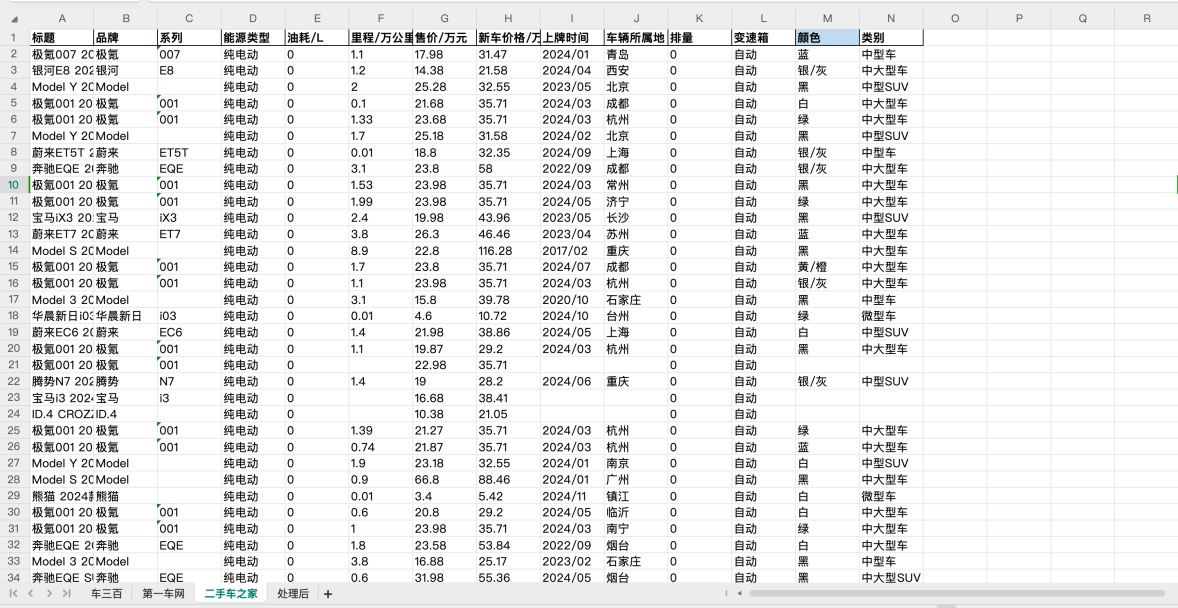
1. 车三百：



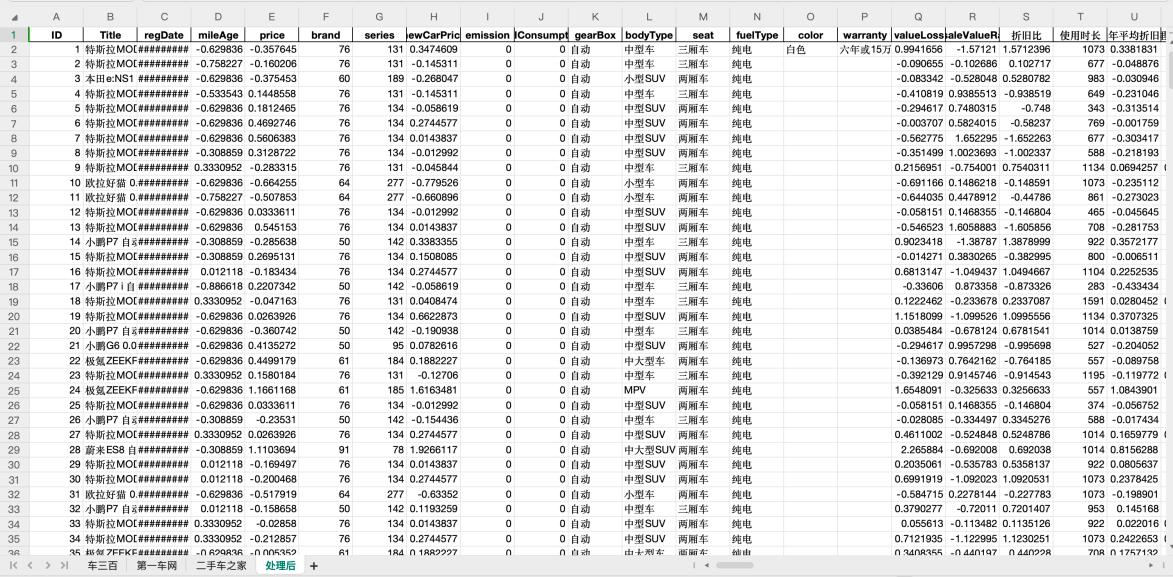
2. 第一车网：



1. 二手车之家：



1. 处理后数据：



（五）数据集使用注意事项

数据清洗：

1. Sheet“车三百”& Sheet“第一车网”& Sheet“二手车之家”：分别是车三百、第一车网及二手车之家等三个二手车平台的交易数据，且已进行过数据清洗；
2. Sheet“处理后”：上述三个平台数据整合且经过标准化处理及编码后数据。

（六）数据集用途

1. 用于后续模型训练及相关分析与预测；

2. 适用于新能源二手车市场。

**二、文本数据集**

（一）数据集基本信息

1. 数据源：知乎（https://www.zhihu.com/）、豆瓣（https://www.douban.com）及新华社（https://www.xinhuanet.com/）。
2. 数据规模：
3. 原始数据：共1024条数据，共7个字段；
4. 处理后数据：（储存在Sheet处理结果中）：共972条数据，共5个字段。
5. 数据类型

本数据集已将非结构化文本数据转换为结构化数据。

（二）字段说明

各字段名及其含义为：

1. 原始数据：

（1）问题标题：字符串类型，代表该条回答所针对的问题；

（2）问题内容：由于此字段所示内容与问题标题完全相同，故此字段下数据为空；

（3）答主昵称：字符串类型，代表该条回答的答主，其中包含屏蔽字符\*；

（4）回答时间：字符串类型，代表该条回答的发布时间；

（5）赞同数：整数型数值数据，代表该条回答所收到的赞同数量；

（6）评论数：字符串类型，（包含了多余字符，需进一步清洗），代表该条回答下的评论数；

（7）回答内容：字符串类型，代表该条回答的内容；同时还包含了答主、发布时间、赞同数量、评论数量等额外数据及段落、换行等复杂格式，需进一步清洗。

1. 处理后数据：
2. 答主：清洗后的答主昵称信息；
3. 内容：清洗后的回答内容，已不含换行符、段落符等；
4. 发布时间：回答内容的发布时间；
5. 赞同数：回答所收到的赞同数量，数值型数据；
6. 评论数：回答所收到的评论数量，数值型数据。

（三）数据格式与存储结构

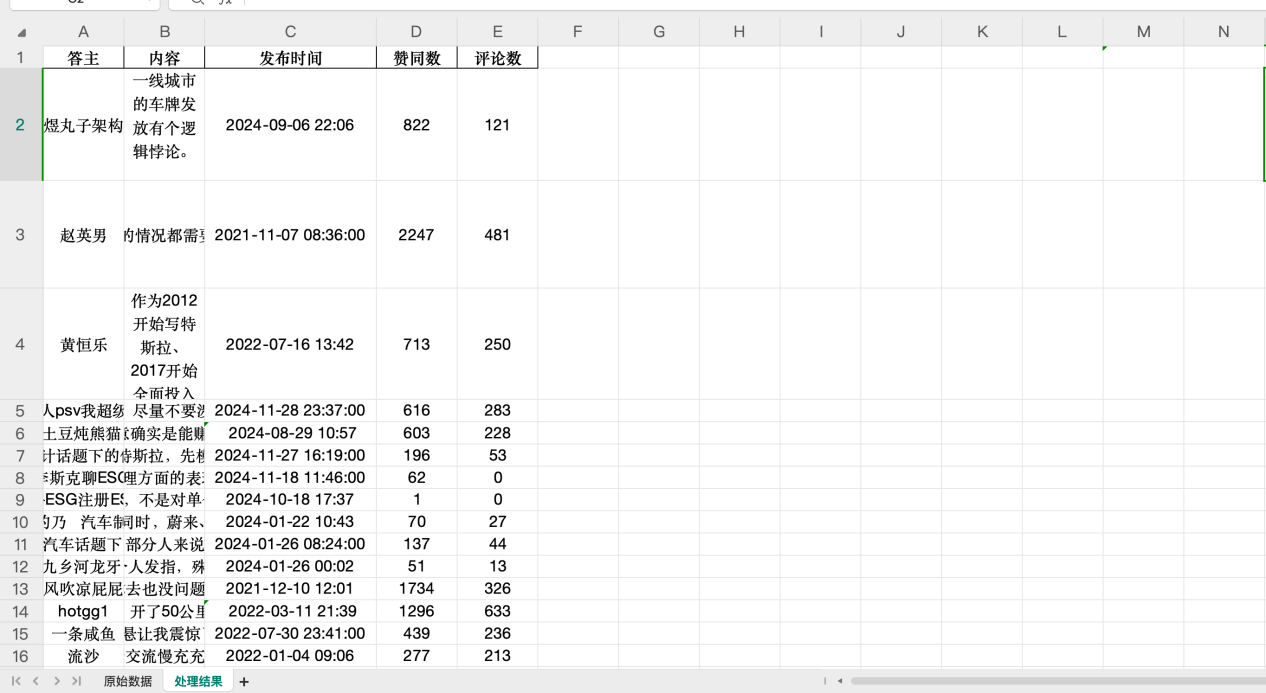
1. 文件格式：本数据集存储于Excel文件中，格式为“.xlsx”；
2. 文件结构：文件中包含字段行，字段行后为具体数据行。

（四）示例数据

1. 原始数据：



2. 处理后数据：



（五）数据集使用注意事项

1. Sheet原始数据：为从回答社区平台爬取的原始数据；

2. Sheet处理结果：将原始数据预处理后的数据，后续基于此数据展开分析。

（六）数据集用途

1. 用于后续大众情感及主题分析；

2. 适用于新能源二手车市场。