【报告】赛题 A 的数据分析报告问题二

——2010YYDS，彭扬

1.分析目标确认

联合查看对比分析附件 1 与附件 4，通过归因分析等方法得到各维度信息量与目标值（成交周期）的变化关系。

2. 查看数据　　data merge为最后处理过的数据文件，可用于建模

2.1 数据类型

将附件 1 与附件 4 组合排列，并给定标题行，转换为 EXCEL 的 xlsx 格式（相比于问题一，新增了 5 列特征）。

特征名称	数据类型	特征名称	数据类型
车辆 id	int64	燃油类型	int64
展销时间	datetime64[ns]	新车价	float64
品牌 id	int64	匿名特征 1	int64
车系 id	int64	匿名特征 2	int64
车型 id	int64	匿名特征 3	int64
里程	float64	匿名特征 5	int64
车辆颜色	int64	匿名特征 6	int64
车辆所在城市 id	int64	匿名特征 8	int64
国标码	int64	匿名特征 9	int64
过户次数	int64	匿名特征 10	int64
载客人数	int64	匿名特征 14	int64
注册日期	datetime64[ns]	交易价格	float64
上牌日期	datetime64[ns]	***上架时间***	***object***
国别	int64	***上架价格***	***float64***
厂商类型	int64	***价格调整***	***object***
年款	int64	***下架时间***	***object***
排量	float64	***成交时间***	***object***
变速箱	int64		

2.2 特征重构

将附件 4 的内容与附件 1 拼接。并对“价格调整”栏进行维度重构，提取出 5 维新的特征，如下：“是否调价”“调价次数”“调价频率（次/天）”“最终调价时间”“最终调价”。如下图 1：

是否调价	调价次数	调价频率	最终调价时间	最终调价	成交时间	成交时间New	成交周期
False	0	0.0	2021-06-25	7.38	2021-07-23	2021-07-23	28
False	0	0.0	2021-06-29	4.38	2021-06-30	2021-06-30	1
False	0	0.0	2021-06-30	5.9	2021-07-19	2021-07-19	19

图 1 计算所得新增特征维度表示意图

此外，考虑到二手车交易收到价格的影响较大，所以对比新车价等因素，最终对“初始降价”（二手车第一次定价与新车价的差值）、“最终降价”（二手车交易价格与新车价的差值）、“最终降价比”等。

2.3 缺失值占比

附件 4 中的缺失值占比如下：

特征名称	数据类型
车辆 id	0.00%
上架时间	0.00%
上架价格	0.00%
价格调整	0.00%
下架时间	0.00%
成交时间	20.00%

题目中提到“附件 4“门店交易训练数据”包括 6 个字段，如下表所示，其中所 有 carid 等相关信息包含在附件 1“估价训练数据”中。”,下架时间（成交车辆下架时间和成交时间相同）。

序号	Features	Description
1	carid	车辆 id
2	pushDate	上架时间
3	pushPrice	上架价格
4	updatePriceTimeJson	{价格调整时间：调整后价格}
5	pullDate	下架时间(成交车辆下架时间和成交时间相同)
6	withdrawDate	成交时间

图 2 题 A 中问题二给定的信息

因此，**建议可以根据拼接的数表来对车辆的成交时间进行缺失值的填补，并以此计算出“成交周期”**

2.4 重复值检验

经过查验，该数据不含有重复值（**最终所得清洗数据，可见：data_merge.xlsx**）

3.数据分析

3.1 时间分布（给定成交速度）

对时间分布进行描述性统计，可得下表：

特征名称	展销时间	注册日期	上牌日期	上架时间	成交时间 New	成交周期
count	9993	9993	9993	10000	10000	10000
mean	41:11.7	29:54.1	14:11.6	14:21.1	2020/12/14 1:01	22.1992
min	2020/1/1 0:00	2004/9/1 0:00	2005/5/18 0:00	2020/1/1 0:00	2020/1/7 0:00	0
25%	2020/7/28 0:00	2012/11/1 0:00	2013/2/21 0:00	2020/7/23 0:00	2020/8/18 0:00	6
50%	2020/11/21 0:00	2015/7/1 0:00	2015/11/23 0:00	2020/11/14 0:00	2020/12/6 12:00	13
75%	2021/4/13 0:00	2017/9/1 0:00	2017/12/14 0:00	2021/4/9 0:00	2021/4/23 0:00	28
max	2021/7/31 0:00	2021/4/1 0:00	2021/6/15 0:00	2021/7/31 0:00	2021/12/6 0:00	277
mode						1

对成交周期总体分布进行查看，如下图：

将成交周期，对数据按照“周”进行分箱操作，可得“成交速度”字段：

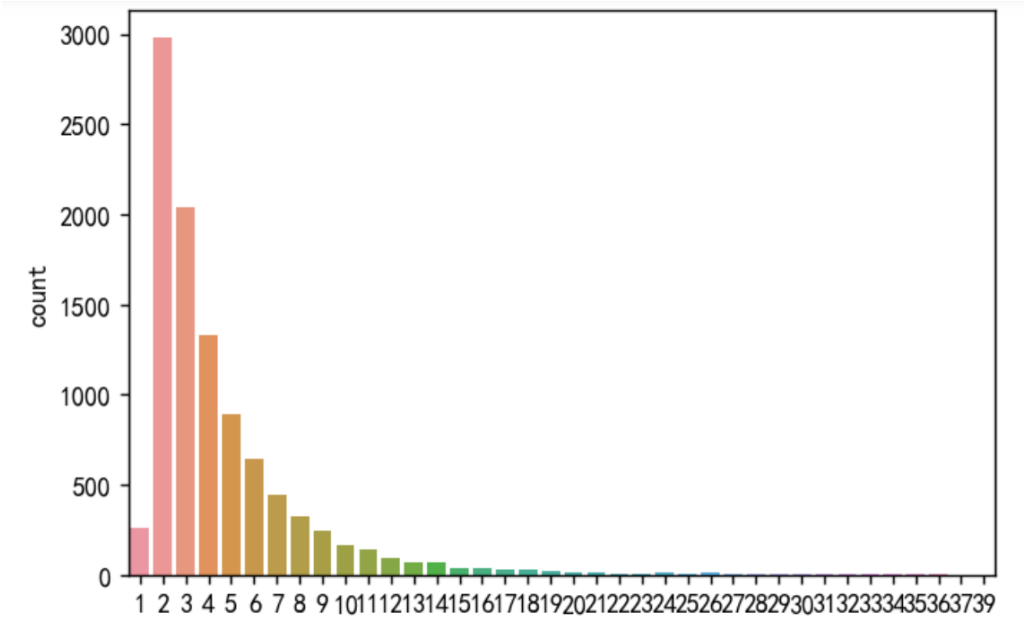


图 3 成交速度的交易额分布量

可以发现，一周内（1<x<=7）内车辆的成交量最大。

- 1：1 天
- 2：1 周
- 3：2 周
- 4：1 月
- 5：1 季度
- 6：1 年

3.2 相关性分析

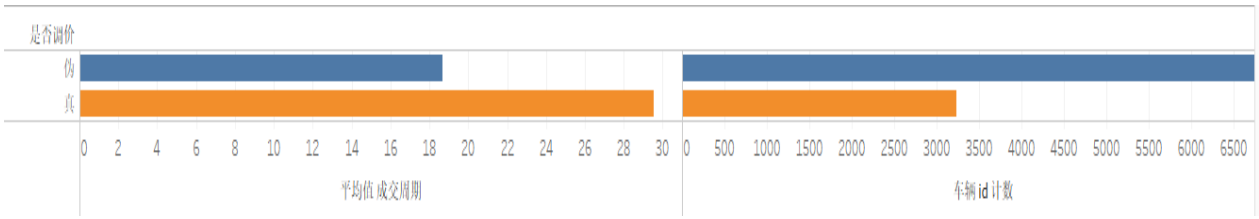
对所有特征进行相关性分析可以得到下表，对相关性系数较大的特征进行进一步的分析。

特征名称	相关系数	特征名称	相关系数
成交周期	**1**	最终降价	0.059323911
成交速度	**0.980349538**	新车价	0.052457721
调价次数	**0.300546697**	车系 id	0.051894532
是否调价	**0.280551791**	里程	0.042839727
初始降价比	**0.199944821**	国别	0.037832715
调价频率	**0.19917991**	匿名特征 5	0.037463127
最终降价比	**0.18322691**	变速箱	0.027262854
上架价格	**0.163675884**	匿名特征 3	0.020543588
交易价格	**0.153693636**	载客人数	0.019474376
最终调价	**0.153693636**	品牌 id	0.018615293
过户次数	**0.144847351**	匿名特征 14	0.017409408
车辆所在城市 id	**0.124689677**	车辆颜色	0.016838842
上牌日期	**0.101191014**	车辆 id	0.016780908
注册日期	0.096513808	匿名特征 6	0.014172403
国标码	0.092519186	排量	0.013975654
车辆级别	0.091841812	匿名特征 2	0.013762206
年款	0.085424535	匿名特征 1	0.012901308
展销时间	0.07758872	车型 id	0.010081574
初始降价	0.073243575	匿名特征 10	0.009689279
厂商类型	0.071484836	燃油类型	0.006866573
匿名特征 8	0.066543931	匿名特征 9	0.001363688

3.3 调价次数、是否调价

（1）是否调价

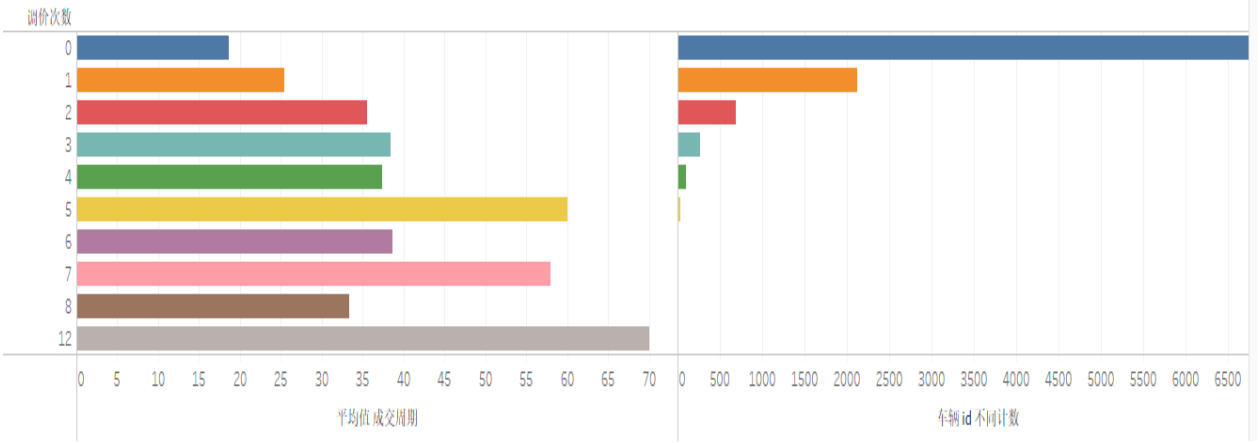
从下图看，未进行调价的车辆交易周期明显比经过调价的车辆交易周期小。

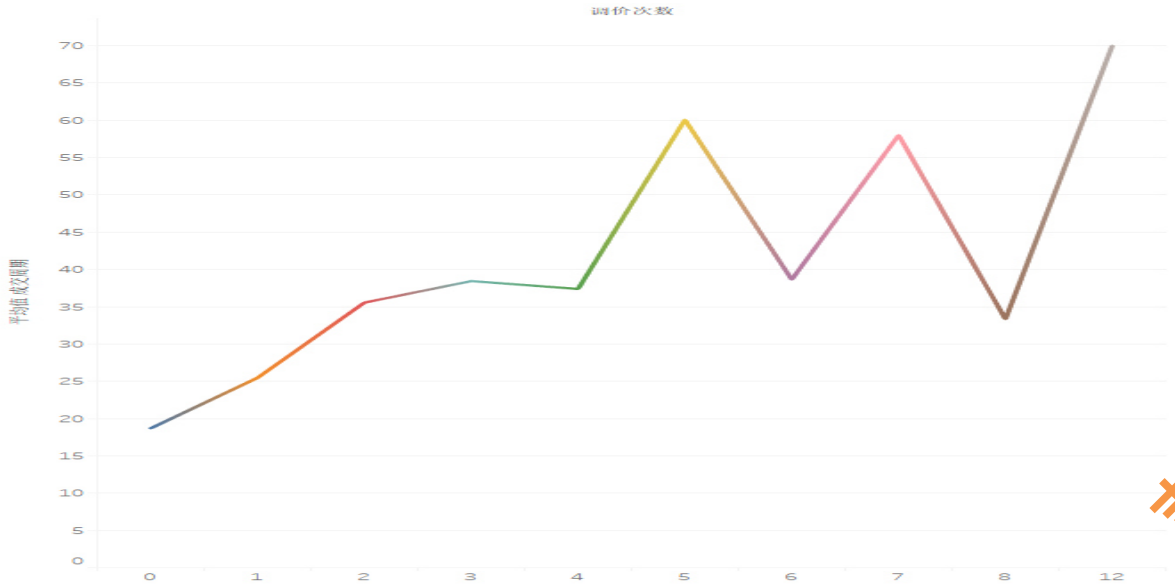


（2）调价次数

如下所示，交易车辆中未调价的车辆交易额最大，且平均交易周期最小；调价次数最多的车辆，其交易周期最大；随着调价次数的增大，车辆的平均交易周期增大。

这说明，车辆的初始定价对车辆的成交周期的影响较大，降价次数对成交周期有影响。





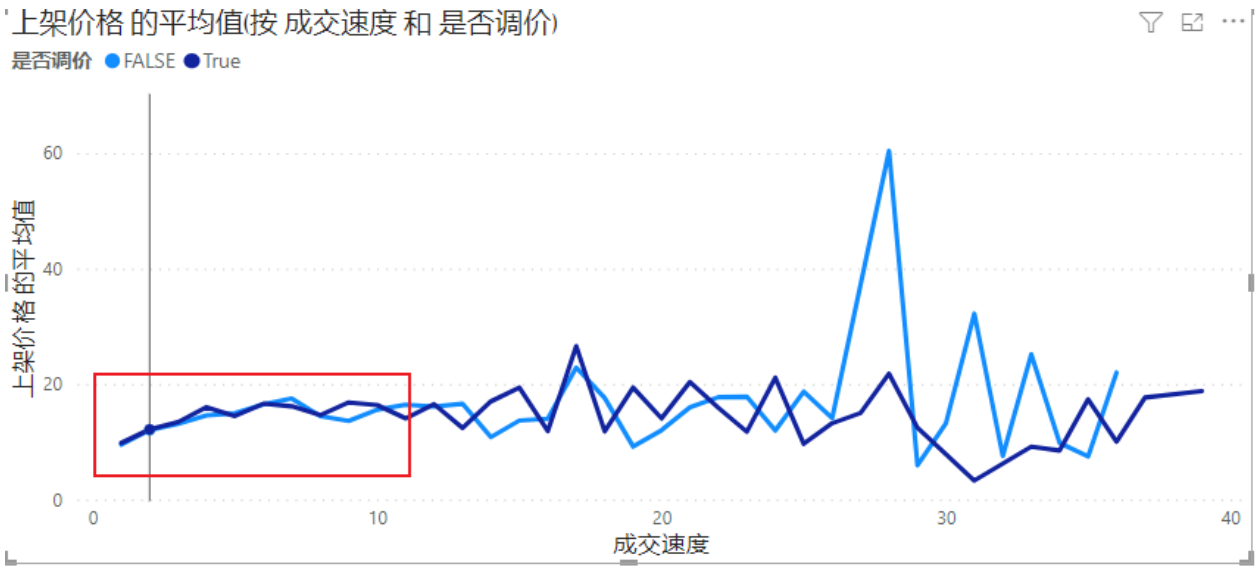
3.4 初始降价比、最终降价比

（1）初始降价比（与新车价相比）

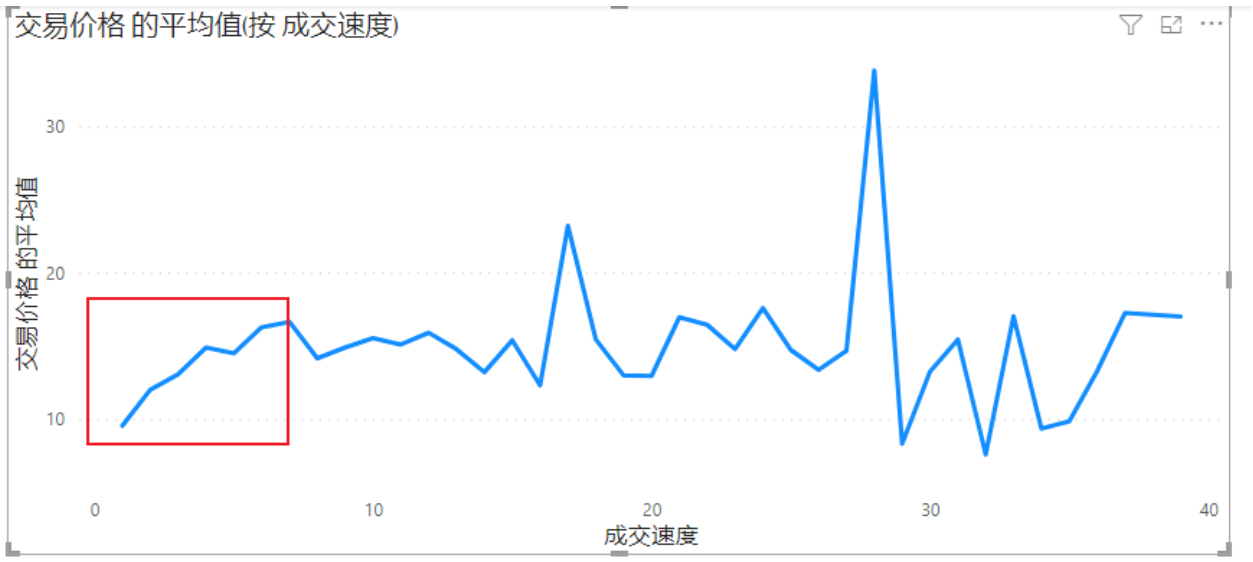
（2）最终降价比（与新车价相比）

可以发现，降价比对成交速度有一定影响，降价比较大时，二手车的成交速度周期较小。。

3.5 上架价格、交易价格



从上图中可以发现，在前两个月内，上架价格越高，其成交速度越大。



从图中可以发现，在前 1 个月内完成交易的二手车辆，其平均交易价格较低，且随着成交速度的增长，其平均交易价格也在增长

3.6 过户次数、上牌日期、城市位置