【报告】赛题 A 的数据分析报告问题一

——2010YYDS，彭扬

1.分析目标确认

查看对应的数据分布，通过对比分析等方法得到各维度信息量与目标值（二手车交易金额）的变化关系。

2. 查看数据

2.1 数据类型

附件 1 共计 36 维特征，原始数据给定的是 txt 文档（不包含变量名），我将变量名称添加到了数据中，并转换为了 EXCEL 的 xlsx 格式（train_data_raw.xlsx）。此外，我将展销时间、注册日期、上 牌日期等数据格式从 object 转换为了 Datetime，如下表。

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	float64
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 4	float64
车辆颜色	int64	匿名特征 5	int64
车辆所在城市 id	int64	匿名特征 6	int64
国标码	float64	匿名特征 7	object
过户次数	int64	匿名特征 8	float64
载客人数	int64	匿名特征 9	float64
注册日期	datetime64[ns]	匿名特征 10	float64
上牌日期	datetime64[ns]	匿名特征 11	object
国别	float64	匿名特征 12	object
厂商类型	float64	匿名特征 13	float64
年款	float64	匿名特征 14	int64
排量	float64	匿名特征 15	object
变速箱	float64	交易价格	float64

2.2 缺失值占比

原始数据（包括附件 1 与附件 2）的缺失值占如下表所示，通过对比分析得出一下结论：

1. 训练数据与验证数据中匿名特征 4、7、15 缺失值占比均超过了 35%，**建议删除该维度**。
2. 其余缺失特征需要进一步探索其与数据之间的关系来考虑数据清洗的方法（见 2.3）。
3. 训练数据中不包含缺失值的特征共计 22 维，每一维 30000 行数据，详情可见：**trian_data_notnull.xlsx**。验证数据中不包含缺失值的特征也共计 22 维，但是**二者的特征种类并不一致**，详情可见 **val_data_notnull.xlsx**。

训练数据				验证数据			
特征名称	缺失值占比	特征名称	缺失值占比	特征名称	缺失值占比	特征名称	缺失值占比
车辆 id	0.00%	燃油类型	0.00%	车辆 id	0.00%	燃油类型	0.00%
展销时间	0.00%	新车价	0.00%	展销时间	0.00%	新车价	0.00%
品牌 id	0.00%	匿名特征 1	5.27%	品牌 id	0.00%	匿名特征 1	6.80%
车系 id	0.00%	匿名特征 2	0.00%	车系 id	0.00%	匿名特征 2	0.00%
车型 id	0.00%	匿名特征 3	0.00%	车型 id	0.00%	匿名特征 3	0.00%
里程	0.00%	匿名特征 4	40.36%	里程	0.00%	匿名特征 4	37.26%
车辆颜色	0.00%	匿名特征 5	0.00%	车辆颜色	0.00%	匿名特征 5	0.00%
车辆所在城市 id	0.00%	匿名特征 6	0.00%	车辆所在城市 id	0.00%	匿名特征 6	0.00%
国标码	0.03%	匿名特征 7	60.15%	国标码	0.00%	匿名特征 7	66.30%
过户次数	0.00%	匿名特征 8	12.58%	过户次数	0.00%	匿名特征 8	8.32%
载客人数	0.00%	匿名特征 9	12.48%	载客人数	0.00%	匿名特征 9	8.26%
注册日期	0.00%	匿名特征 10	20.80%	注册日期	0.00%	匿名特征 10	24.62%
上牌日期	0.00%	匿名特征 11	1.54%	上牌日期	0.00%	匿名特征 11	1.46%
国别	12.52%	匿名特征 12	0.00%	国别	7.92%	匿名特征 12	0.02%
厂商类型	12.14%	匿名特征 13	5.40%	厂商类型	7.50%	匿名特征 13	5.20%
年款	1.04%	匿名特征 14	0.00%	年款	2.12%	匿名特征 14	0.00%
排量	0.00%	匿名特征 15	91.93%	排量	0.00%	匿名特征 15	94.38%
变速箱	0.00%	交易价格	0.00%	变速箱	0.00%		

2.3 数据清洗

去除缺失值占比较大的特征维度之后，余下的包含缺失值的特征表如下所示（四舍五入）。

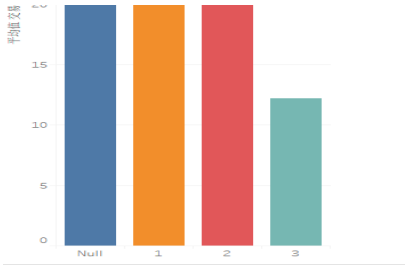
从表中我们可以发现，训练集与测试集之间的缺失值占比类型大致相同，主要是：“**国别、厂商类型、年款、匿名特征 1，8，9，10，11，13**”。其余的如“国标码，变速箱”等由于缺失值占比较小，且验证集中没有缺失，所以**建议：将训练数据中的“国标码”、“变速箱”特征下的缺失值进行“行删除”，删除“匿名特征 12”维度列。**

余下含有缺失值的特征，按照缺失占比大小进行排序，逐个进行分析

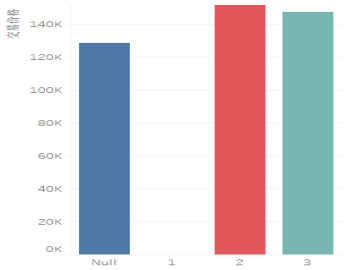
训练数据		验证数据	
特征名称	缺失值占比	特征名称	缺失值占比
国标码	0.03%		
国别	**12.52%**	**国别**	**7.92%**
厂商类型	**12.14%**	**厂商类型**	**7.50%**
年款	**1.04%**	**年款**	**2.12%**
变速箱	0.00%	匿名特征 12	0.02%
匿名特征 1	**5.27%**	**匿名特征 1**	**6.80%**
匿名特征 8	**12.58%**	**匿名特征 8**	**8.32%**
匿名特征 9	**12.48%**	**匿名特征 9**	**8.26%**
匿名特征 10	**20.80%**	**匿名特征 10**	**24.62%**
匿名特征 11	**1.54%**	**匿名特征 11**	**1.46%**
匿名特征 13	**5.40%**	**匿名特征 13**	**5.20%**

（1）匿名特征 10

匿名特征 10 属于分类变量，包含 Null 缺失值在内一共是 4 个类别，各类之间与 price（交易价格）之间的关系如下图所示：



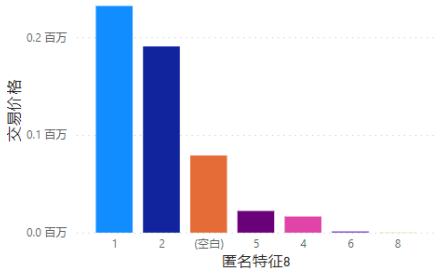
每类中的平均交易价格



总计的交易价格

因此，对于“匿名特征 10”，**建议将 Null 缺失值重构为类别 4.**

（2）匿名特征 8



同理，建议

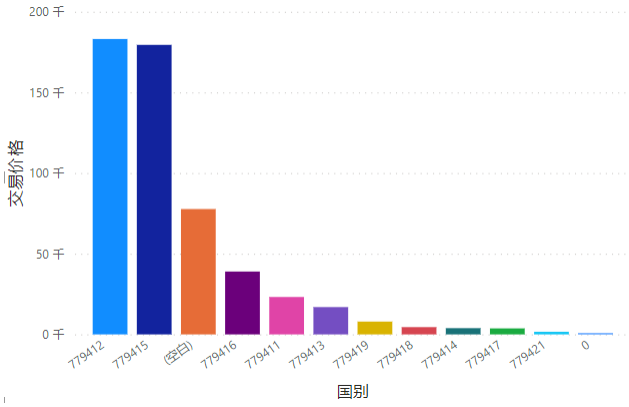
① 将 NULL 缺失值重构为类别 3；

② 或者聚类

③ 或者舍弃该列

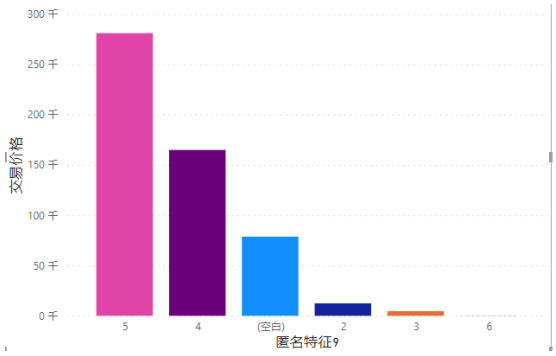
本人采用的是第一种方法。（建模的时候可以考虑再舍弃）

（3）国别



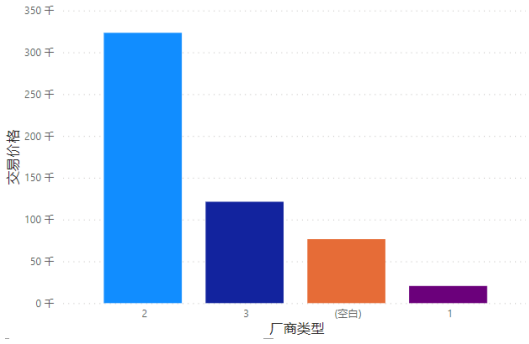
将国别进行排序，共得到 12 类数字。同理分析，**建议将 NULL 缺失值重构为类别 773420，将类别 0 重命名为 773410**

（4）匿名特征 9



同理分析，**建议将 NULL 缺失值重构为类别 1**

（5）厂商内型



同理，**建议将 NULL 缺失值重构为类别 4**

（6）匿名特征 13

“匿名特征 13”包含 NULL 值在内一共有 168 维度类别，由于类别数量较大，所以我们无法进行简单的填补，此外通过观察匿名特征 13 的数据分布，我猜测特征是与时间有关的类别，类似最终你的交易日期什么的。除了某一个月交易额激增，其余无显著差异。

综合分析，**我建议可以采取一下几种方法的一种：**

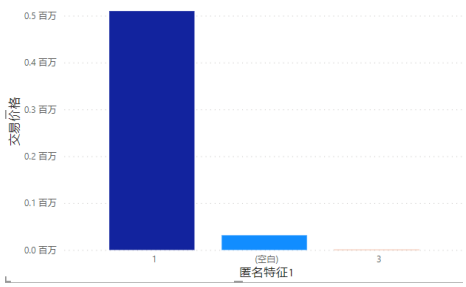
① **删除缺失值，最好删除该列（就是建模的时候不要考虑该维度特征）**

② **对缺失值进行填补（但是训练集填补后，验证集也要填补，可能误差较大）**

③ **深度学习算法可以用缺失值进行训练或预测吗？（@叶子汗）**

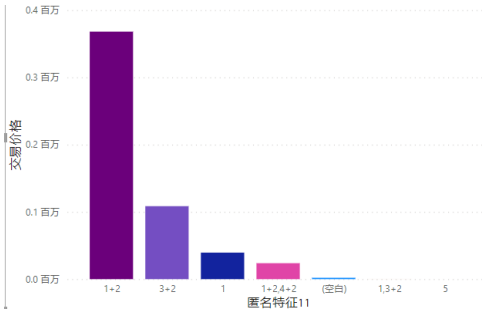
本人使用的是第一种方法。

（7）匿名特征 1



“匿名特征”1 为分类变量，通过图表分析，**建议将 NULL 缺失值重构为类别 2**

（8）匿名特征 11

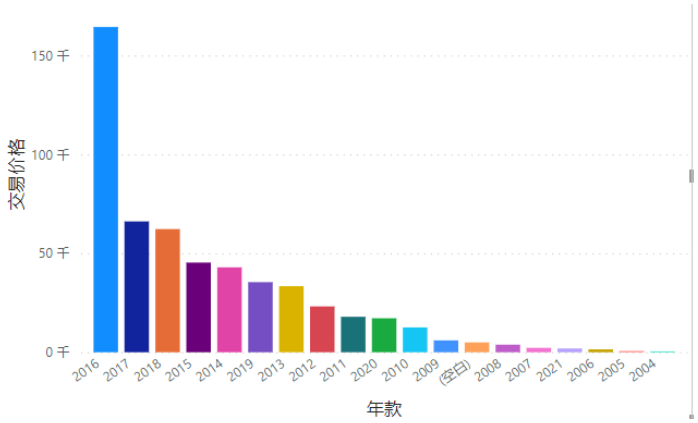



```
In  [222]:  1  data_1["匿名特征11"].unique()
Out[222]:  array(['1', '1+2', nan, '3+2', '1+2,4+2', '1,3+2', '5'], dtype=object)

In  [221]:  1  data_2["匿名特征11"].unique()
Out[221]:  array(['1+2', nan, '1+2,4+2', '1', '3+2'], dtype=object)
```

考虑到训练数据与测试数据之间分类缺失的类型无法确定，所以**建议删除该列特征**

（9）年款



“年款”特征表示的是该车是哪年出的，通过分析，**建议将缺失值 NULL 重构为类别 2003.**

到此，训练数据与验证数据中的缺失值均被处理。

2.4 重复值占比

通过对缺失值清洗后的数据进行验证，发现，给定数据中不含有重复值，可以放心使用。

数据清洗后的文件名称为：train_data_new.xlsx;
val_data_new.xlsx。

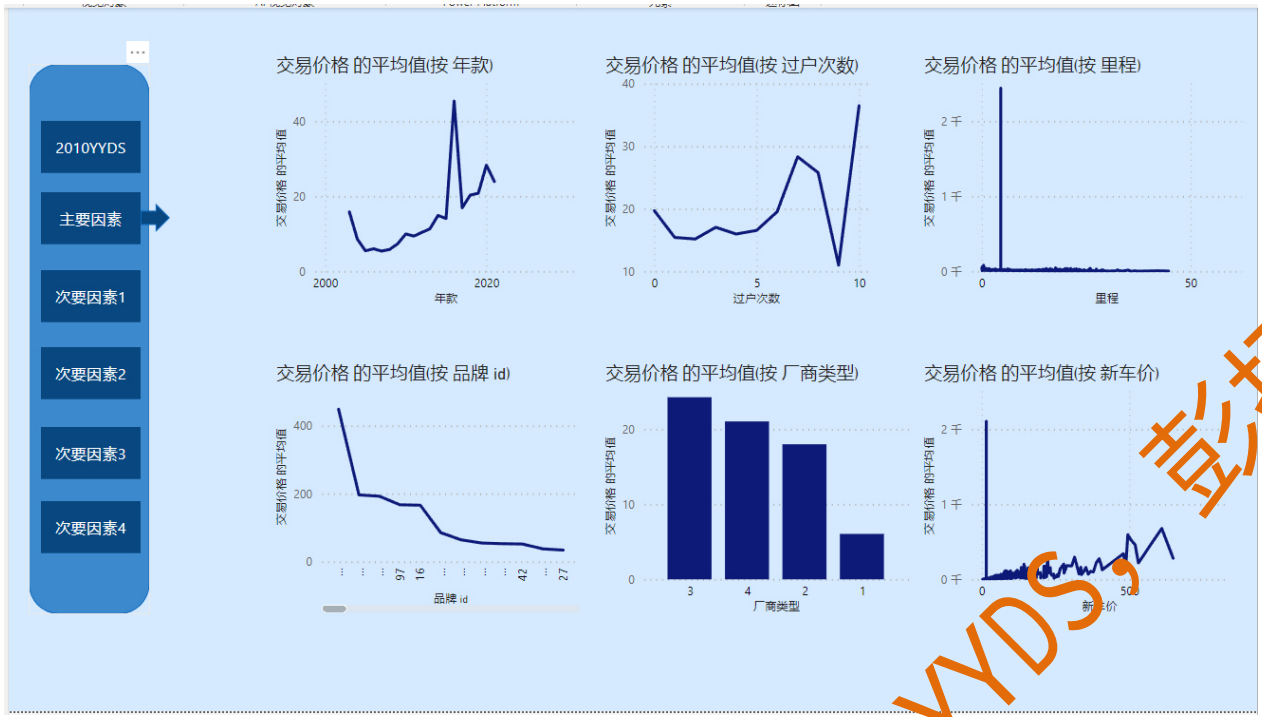
3.数据分析

数据清洗完成之后，数据维度变成了 30 维。

具体的数据分析结果，可见下方的可视化面板链接报表（做得比较丑，见谅）：

https://app.powerbi.cn/view?r=eyJrIjoiYzgzYTc1NzMtNmZhNC00MWFhLTg3N2MtM2NjZmJmMTY1ZjFhIiwidCI6IjJhM2M2YjRjLTYxODgtNGU4Yi1hODQzLTM1ZjU2YjdmZmY1NyJ9

形如（共计 6 页，可以点击，同一页面的变量数据会随之变化）：



特征工程

1．对特征进行互信息检验，发现所有特征与标签之间的互信息均大于 0.

2．使用随机森林算法对特征进行嵌入选择，可得对应的特征排序如下所示，**本人选择排名为 1 的 14 维特征，最后所得数据为**：**train_data_new_RFE.xlsx**。各位也可以依照此排名进行特征选择来构建对应的模型。

特征名称	特征排序	特征名称	特征排序
车辆 id	1	匿名特征 2	3
匿名特征 3	1	国别	4
新车价	1	变速箱	5
排量	1	匿名特征 6	6
年款	1	载客人数	7
匿名特征 5	1	匿名特征 8	8
注册日期	1	燃油类型	9
上牌日期	1	车辆所在城市 id	10
车系 id	1	匿名特征 9	11
展销时间	1	匿名特征 10	12
车辆颜色	1	匿名特征 14	13
里程	1	厂商类型	14
车型 id	1	匿名特征 1	15
品牌 id	1	国标码	16
过户次数	2		