B

丰田二手车价格影响因素的探究

2018110760 李晨茜

一、研究背景

二手车是指在公安交通管理机关登记注册，在达到国家规定的报废标准之前或在经济实用寿命期内服役，并仍可继续使用的机动车辆。二手车比较适合预算数额比较低或者刚拿到驾照的新手，可以帮助他们在尽可能优惠的价格下买到心仪的车。

然而，国内二手车市场鱼龙混杂，不诚信经营、定价不合理的情况也不在少数，本报告通过分析丰田二手车的价格及相关因素，希望能缓解或解决该问题。

二、研究问题

本案例希望研究影响丰田二手车价格的因素，最终模型可以用于：

（1）二手车的定价

（2）帮助消费者更好地选择二手车

三、数据预处理

首先对数据中涉及的变量进行整理，如下表所示：

表格 1 变量说明

变量名			变量说明	变量类型	备注
因变量		Price	价格	连续型变量	
自变量	基本信息	Age	已使用时长	连续型变量	单位：月
		KM	已行驶公里数	连续型变量	
		Quarterly_Tax	公路税	分类变量	小于 85.5 的为低税组；大于 85.5 的为高税组
		Weight	重量	连续型变量	
	功能性信息	Tow_Bar	是否有牵引杆	分类变量	yes=1,no=0
		Automatic	是否自动挡	分类变量	yes=1,no=0
		CC	气缸容量	分类变量	小于 1601 的为低气缸容量组；大于 1601 的为高气缸容量组。
		Doors	车门数	分类变量	2-5
		Fuel_Type	燃料类型	分类变量	Petrol,Diesel,CNG
		HP	马力	分类变量	小于 110.5 的为低马力组，高于 110.5 的为高马力组
		Mfr_Guarantee	是否在保修期	分类变量	yes=1,no=0
		ABS	是否有防抱死制动系统	分类变量	yes=1,no=0
		Airbag_2	是否有安全气囊	分类变量	yes=1,no=0
		Airco	是否有空调	分类变量	yes=1,no=0
		Boardcomputer	是否有车载电脑	分类变量	yes=1,no=0
		CD_Player	是否有车载 CD	分类变量	yes=1,no=0
		Central_Lock	是否有中央锁	分类变量	yes=1,no=0
		Powered_Windows	是否是电动窗户	分类变量	yes=1,no=0
		Radio	是否有收音机	分类变量	yes=1,no=0
		Sport_Model	是否有运动模式	分类变量	yes=1,no=0
		Backseat_Divider	是否有后座隔档	分类变量	yes=1,no=0
		Met_Color	是否金属颜色	分类变量	yes=1,no=0

经检查，发现该数据中没有缺失值。 那异常值呢？

为避免后续在回归模型中出现虚拟变量陷阱和多重共线性，将 4 车门作为 Doors 变量的基准组；将 CNG 作为 Fuel_Type 变量的基准组。并把 Quarterly_Tax 小于 85.5 的作为低税组，高于 85.5 的作为高税组；把 CC 小于 1601 的作为低气缸容量组，高于 1601 的作为高气缸容量组；把 HP 小于 110.5 的作为低马力组，高于 110.5 的作为高马力组。（85、1600 和 110 分别是 Quarterly_Tax、CC 和 HP 的 3/4 分位点）

四、数据可视化

1.因变量（价格）可视化

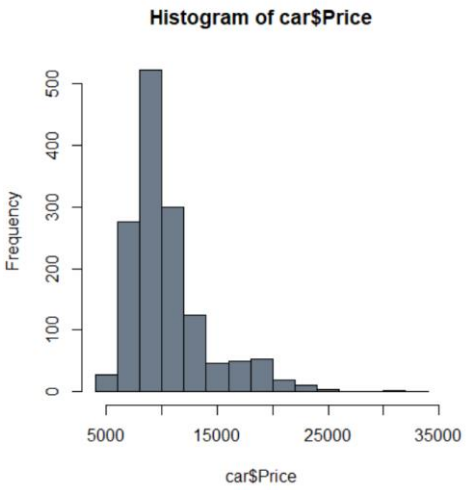

图 2 二手车价格条形图

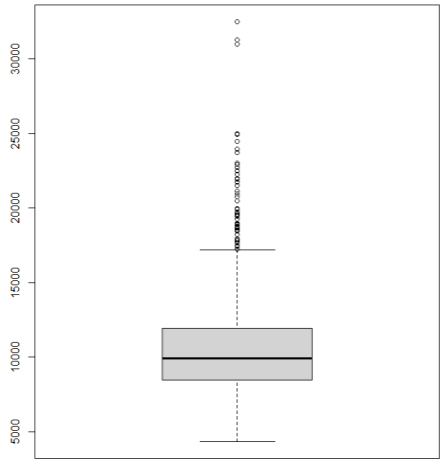

图 1 二手车价格箱线图

通过绘制直方图、箱线图以及对价格数据的初步整理可以发现，丰田二手车价格大部分集中在 8450-11950 元，最低价格为 4350，最高价格为 32500。呈现出右偏的趋势。

2.价格影响因素（自变量）可视化

（1）二手车使用时长

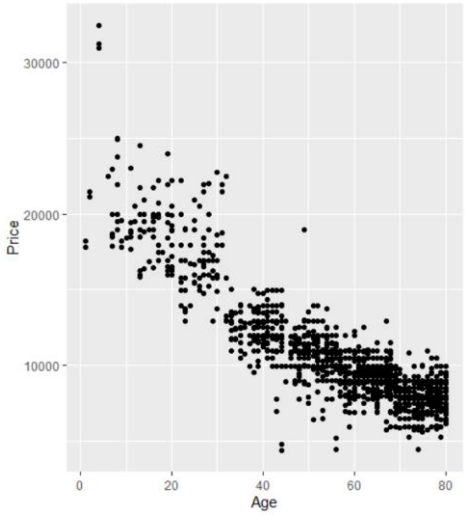

图 3 二手车使用时长和价格散点图

可以通过散点图发现价格和使用时长之间呈负相关的线性关系，说明控制其他因素不变，随着丰田二手车已使用时长的增加，二手车价格呈下降趋势。经过对已使用时长和价格的方差分析，发现其 p 值<0.001，说明已使用时长和价格之间的线性关系在 95%的置信水平下十分显著。

（2）已行驶公里数

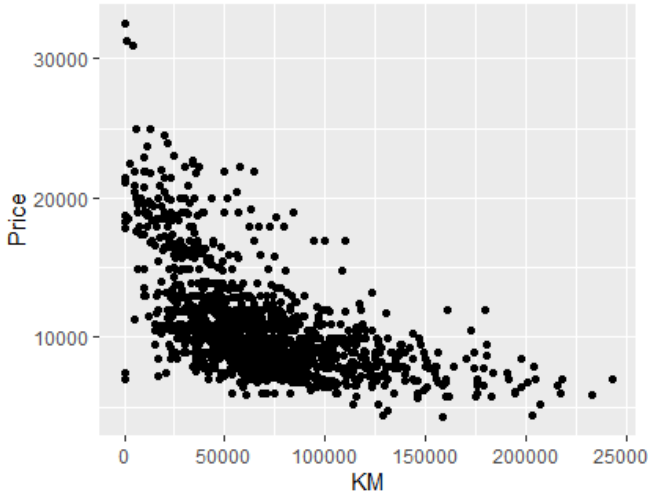

图 4 二手车已行驶公里数和价格散点图

可以通过散点图发现已行驶公里数和丰田二手车的价格也大致呈负相关关系，即控制其他因素不变的情况下，随着已行驶公里数增长，丰田二手车价格大致呈下降趋势。通过单因素方差分析，可以发现已行驶公里数的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（3）是否在保修期

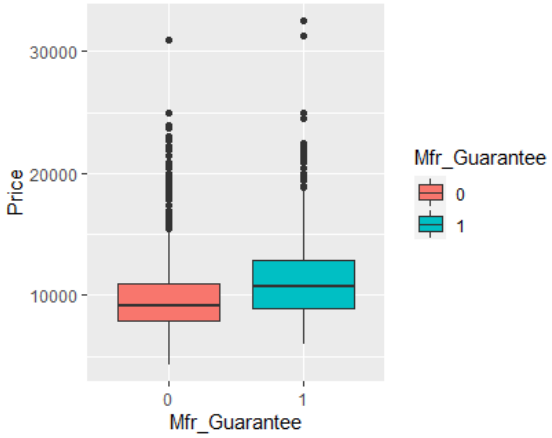

图 5 保修期和价格分类箱线图

可以从箱线图看出，其他因素保持不变的前提下，在保修期的二手车价格平均比不在保修期的高一些。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（4）税

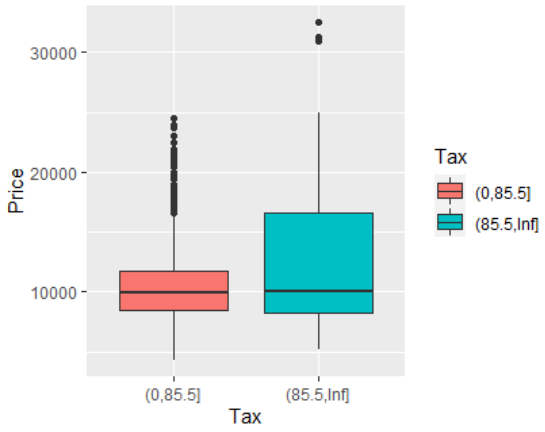

图 6 税和价格分类箱线图

如前文所述，将 Quarterly_Tax 在 85.5 以下的分到低税组，高于 85.5 的分到高税组。从箱线图可以看出两个组的二手车价格均值相差不大，但高税组包含更多价格较高的二手车。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（5）重量

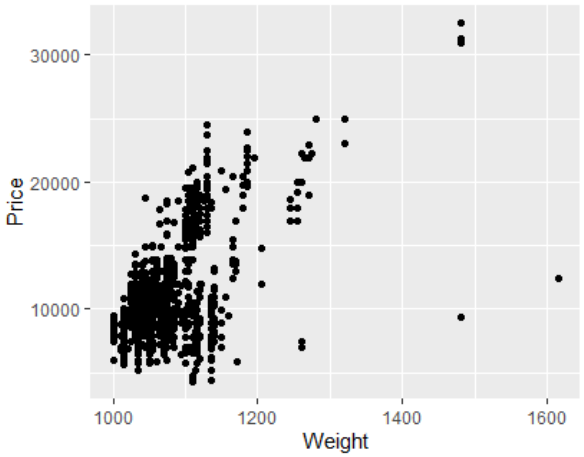

图 7 二手车重量和价格散点图

从散点图可以发现价格较高的二手车多出现在重量较高的位置。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（6）是否有牵引杆

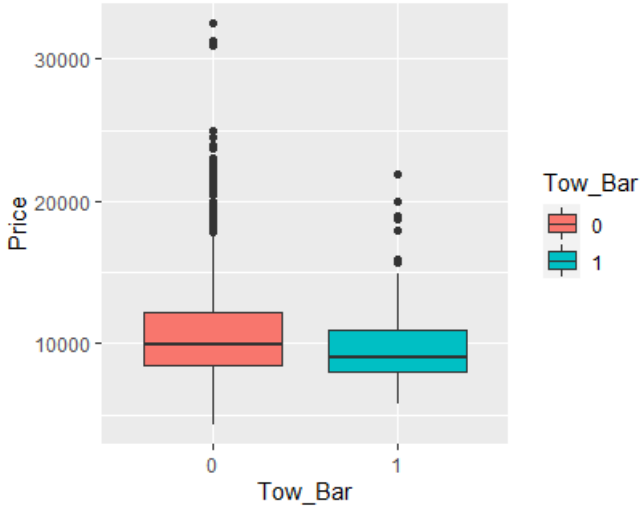

图 8 牵引杆和价格分类箱线图

从箱线图可以看出有牵引杆的二手车比没有的价格要偏低一些。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（7）是否是自动挡

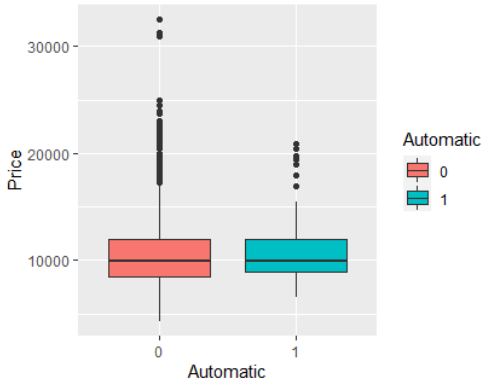

图 9 自动挡和价格分类箱线图

从箱线图发现是否是自动挡对二手车价格影响不大。但价格极高的几个二手车都出现在非自动挡组，考虑可能是因为这些车是豪华轿车、跑车、专业越野车等。因为在普通轿车中，由于自动变速箱成本较高，故自动挡价格比同款手动挡要高；然而在豪华轿车、跑车、专业越野车中，由于跑车在行驶过程中保持高速是需要不停的换挡来实现的，而自动挡是靠自己的程序来换挡，这些程序在执行的过程中难免会出现迟缓，这对跑车提速、变速带来了诸多限制。所以，跑车一般都是采用手动挡来满足消费者的需求。

通过单因素方差分析，可以发现该变量的 p 值=0.21>0.05，说明它和二手车价格的关系在 95%的置信水平下不显著。

（8）气缸容量

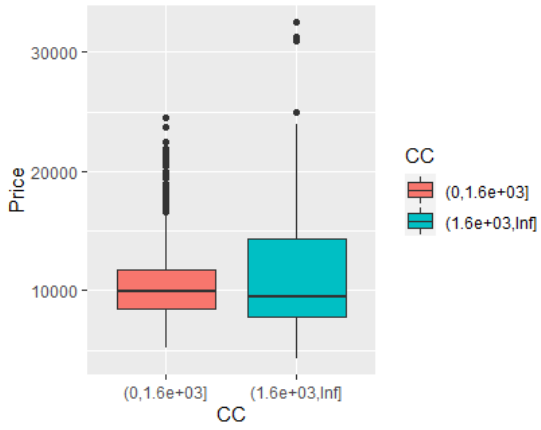

图 10 气缸容量和价格分类箱线图

从箱线图可以发现，气缸容量较大的汽车反而具有相对偏低的平均价格。但价格极高的几个二手车数据也出现在气缸容量大的一组，符合前文的假设——这几辆车可能是豪华轿车、跑车、专业越野车类型——因为豪华轿车、跑车、专业越野车为了取得更大动力，普遍采用 6 缸、8 缸、12 缸等的 V 型或 W 型多缸数发动机，即气缸容量相对更大。

通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在95%的置信水平下比较显著。

（9）车门数

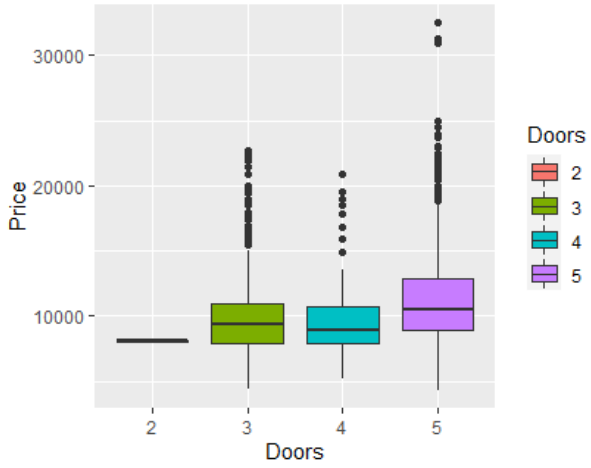

图 11 车门数和价格分类箱线图

从箱线图发现，3 门和 5 门的二手车的平均价格高于 2 门和 4 门的。经查阅资料得知，后档风玻璃与后背箱为一体，开后背箱同时后档风玻璃跟着也一块起，这就是第三/五门，这种车型也被称为掀背型车，通常常见于高级轿车的配置中。因此它们的价格偏高也就符合常理了。同时发现价格大于 30000 的几个数据出现在五门掀背车，符合前文的猜想——这几款车属于高级轿车/跑车/高级越野车。

通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在95%的置信水平下比较显著。

（10）燃料种类

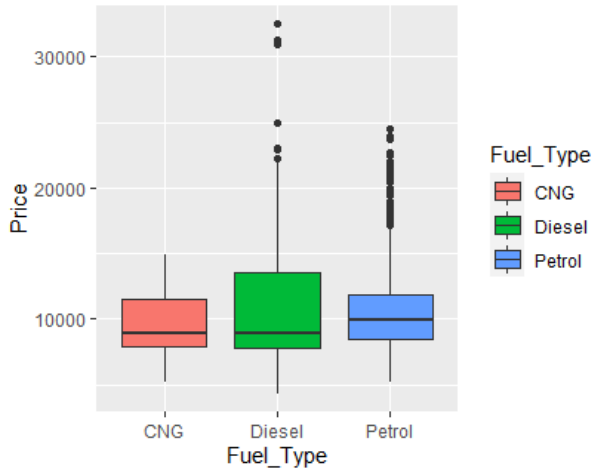

图 12 燃料种类和价格分类箱线图

从该箱线图可以看出柴油和汽油车相对以天然气为燃料的车要更贵一些。通过单因素方差分析，可以发现该变量的 p 值=0.0446<0.05，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（11）马力

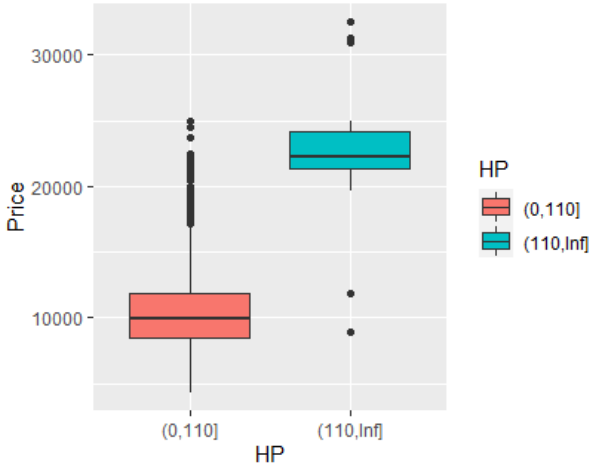

图 13 马力和价格分类箱线图

如前文所述，将马力大于 110 的车分为高马力组，而马力小于 110 的车分为低马力组。从箱线图可以发现，高马力组车的价格明显高于低马力组。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（12）是否有防抱死制动系统

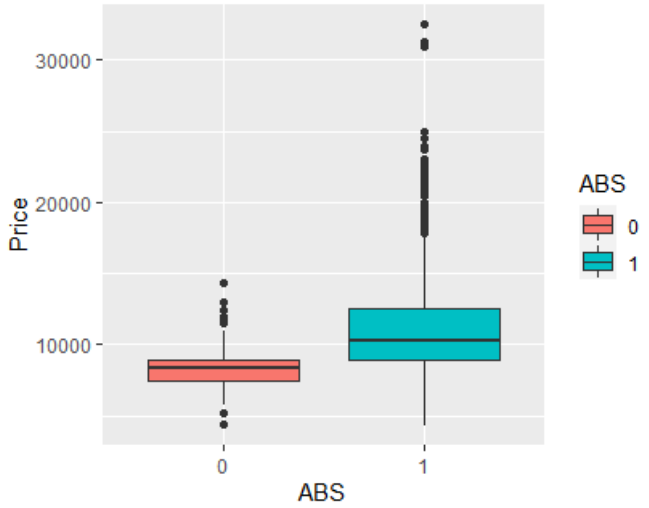

图 14 ABS 和价格分类箱线图

从箱线图可以看出，具有 ABS 的二手车价格相比没有 ABS 的车平均价格更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（13）是否有安全气囊

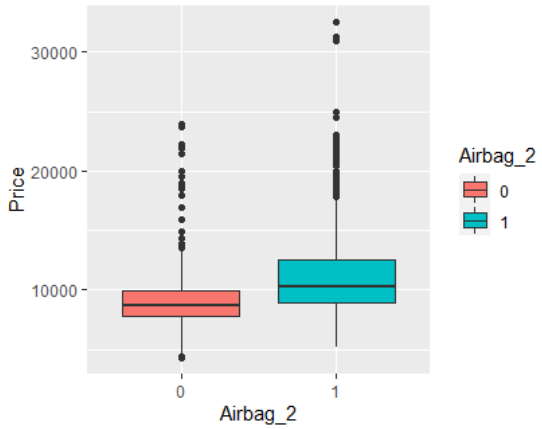

图 15 安全气囊和价格分类箱线图

从箱线图可以发现，具有安全气囊的二手车平均价格明显更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（14）是否有车载空调

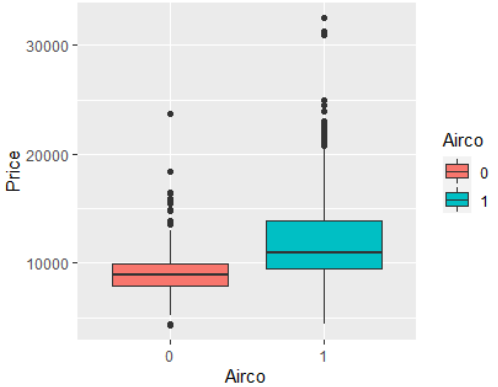

图 16 车载空调和价格分类箱线图

从箱线图可以发现，具有车载空调的二手车平均价格明显更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（15）是否有车载电脑

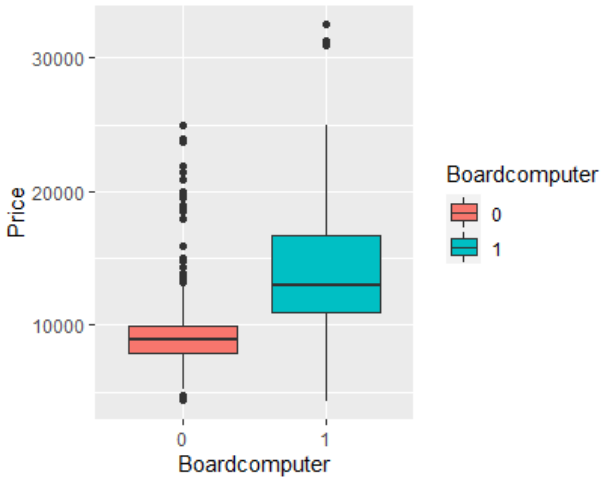

图 17 车载电脑和价格分类箱线图

从箱线图可以发现，具有电脑面板的二手车平均价格明显更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（16）是否有车载 CD 播放器

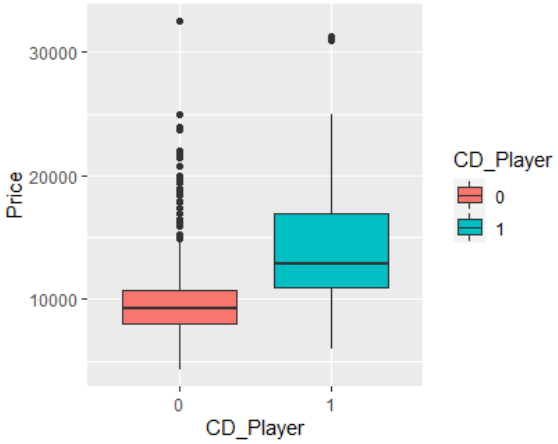

图 18 车载 CD 播放器和价格分类箱线图

从箱线图可以发现，具有车载 CD 播放器的二手车平均价格更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（17）是否有中央锁

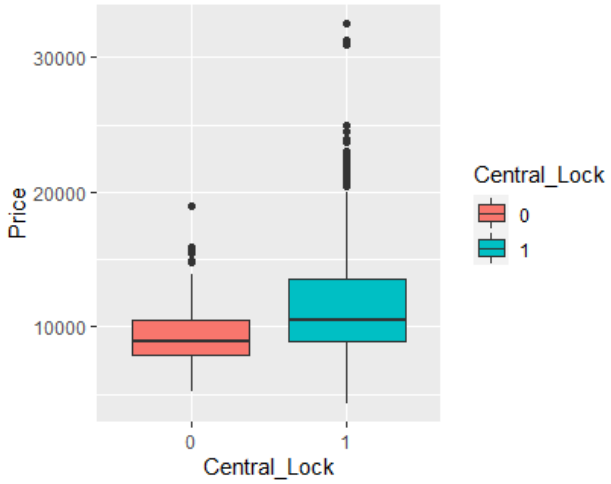

图 19 中央锁和价格分类箱线图

从箱线图可以发现，具有中央锁的二手车平均价格更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（18）是否有运动模式

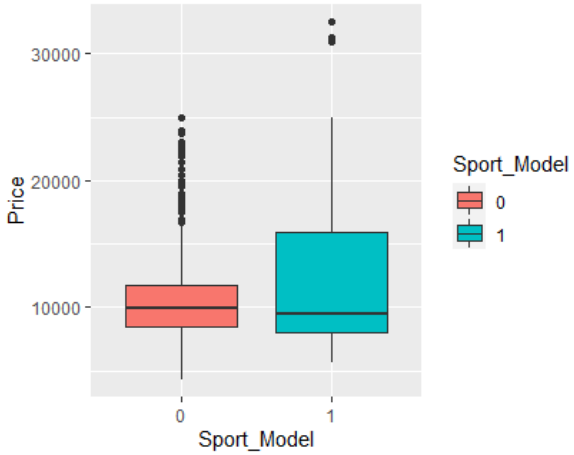

图 20 运动模式和价格分类箱线图

从箱线图可以发现，具有运动模式的二手车平均价格略微低于没有运动模式的。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下比较显著。

（19）是否有后座分隔

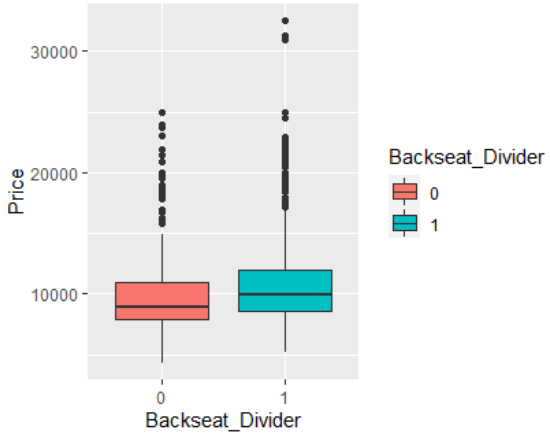

图 21 后座分隔和价格分类箱线图

从箱线图可以发现，具有后座分隔的二手车平均价格更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（20）是否是金属颜色

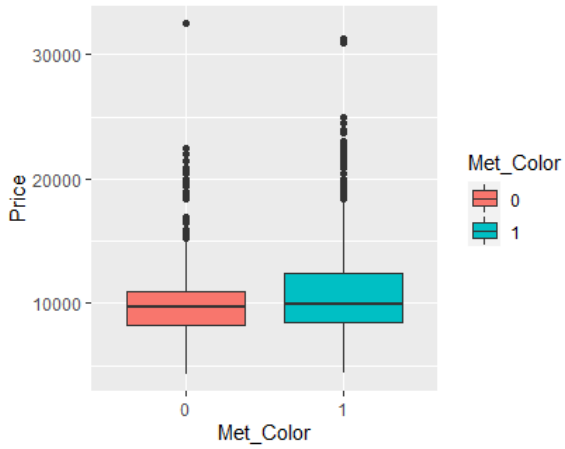

图 22 金属颜色和价格分类箱线图

从箱线图可以发现，是金属颜色的二手车和不是金属颜色的二手车平均价格差别不大。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（22）是否有电动车窗

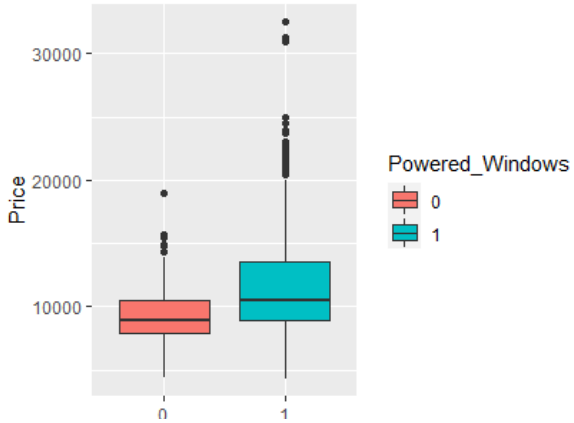

图 23 电动车窗和价格分类箱线图

从箱线图可以发现，具有电动车窗的二手车平均价格明显更高。通过单因素方差分析，可以发现该变量的 p 值<0.001，说明它和二手车价格的关系在 95%的置信水平下十分显著。

（23）是否有车载收音机

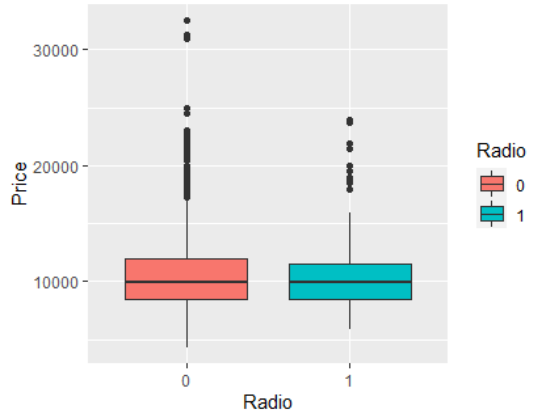

图 24 车载收音机和价格分类箱线图

从箱线图可以发现，有收音机的二手车和没有收音机的二手车平均价格差别不大。通过单因素方差分析，可以发现该变量的 p 值=0.113>0.05，说明它和二手车价格的关系在 95%的置信水平下不显著。

五、数据分析（线性回归模型）

分析结果可以以表格形式放出来的，虽然放在描述性分析文字中也行，但不够直观。

以价格为因变量，去除掉上一步单因素方差分析结果不显著的 Automatic 和 Radio 变量，其余 20 个变量作为自变量，进行线性回归，得到结果如下表所示：

表格 2 线性回归模型结果

回归系数也是很重要的一个结果，是用于模型解读的，为啥没放上来

变量	p 值	备注
截距项	0.700	
已使用时长	<0.001	
已行驶公里数	<0.001	
高税组	<0.001	基准值：低税组
重量	<0.001	
牵引杆	0.094	
大气缸容量	0.655	基准值：小气缸容量
三门车	0.496	基准值：两门车
四门车	0.801	
五门车	0.569	
燃料类型-柴油	0.748	基准值：CNG
燃料类型-汽油	<0.001	
高马力	<0.001	
保修	<0.001	
防抱死制动系统	0.033	
安全气囊	0.088	
车载空调	<0.001	
电脑面板	0.230	

车载 CD 播放器	0.003
中央锁	0.99
电动车窗	<0.001
运动模式	<0.001
后座分隔	0.855
金属颜色	0.450

可以发现有很多变量在 95%的置信水平下不显著，且该回归方程的调整R^2=0.8846，说明该回归模型能够解释 0.8846 的因变量变异的程度。对该模型进行模型诊断：

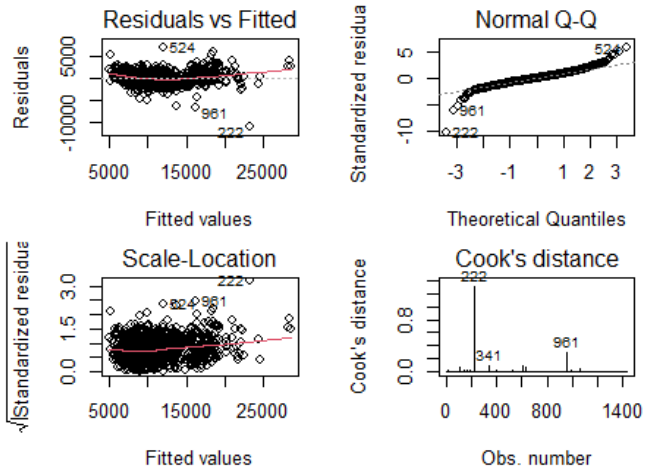

图 25 模型诊断图

表格 3 模型方差膨胀因子结果

变量	GVIF	DF	GVIF^(1/(2*DF))
AGE	4.03	1	2.01
KM	2.07	1	1.44
QUARTERLY_TAX	5.8	1	2.41
WEIGHT	3.25	1	1.8
TOW_BAR	1.1	1	1.05
CC	31.02	1	5.57
DOORS	1.71	3	1.09
FUEL_TYPE	37.4	2	2.47
HP	2.4	1	1.55
MFR_GUARANTEE	1.18	1	1.09
ABS	2.11	1	1.45
AIRBAG_2	3.04	1	1.74
AIRCO	1.7	1	1.3
BOARDCOMPUTER	2.5	1	1.58
CD_PLAYER	1.48	1	1.22
CENTRAL_LOCK	4.48	1	2.12
POWERED_WINDOWS	4.54	1	2.13
SPORT_MODEL	1.33	1	1.15
BACKSEAT_DIVIDER	2.38	1	1.54

MET_COLOR	1.12	1	1.06

发现 QQ 图中两侧的尾巴偏离直线，说明误差不服从正态分布而是服从 t 分布。并且 cook 图中可见明显的强影响点。并且通过计算方差膨胀因子，发现变量之间存在严重的多重共线性。

考虑到模型诊断出的问题和前文对二手车价格做可视化分析时发现其存在右偏，考虑对其做对数变换以稳定方差，改善分布不对称的现象。然而在实际操作中发现，将价格做对数变换后得到的线性模型的拟合优度反而降低了，并且也没有有效解决 QQ 图两侧尾部偏离直线的问题和多重共线性的，因此在本文中不使用因变量的对数形式。 但线性回归是有假设条件的，正态性是其中一个比较重要的条件。

将模型使用 AIC 的方法进行变量的选择和剔除，得到结果如下表所示：

表格 4 AIC 后的回归模型结果

变量	回归系数	标准误	p 值	备注
截距项	-826.1	1292	0.595	
已使用时长	-116.6	2.905	<0.001	
已行驶公里数	-0.0178	0.001	<0.001	
高税组	1357	219.5	<0.001	基准值：低税组
重量	16.44	1.106	<0.001	
牵引杆	-136.2	74.9	0.069	
两门车	-630	879.2	0.789	基准值：两门车
三门车	-235.4	119.4	<0.001	
五门车	-531.2	121.6	0.015	
燃料类型-柴油	-44.56	321.2	0.889	基准值：CNG
燃料类型-汽油	1911	355.9	<0.001	
高马力	3497	322.5	<0.001	
保修	327.2	69.44	<0.001	
防抱死制动系统	-235.7	117.8	0.045	
安全气囊	-256.8	103.5	0.013	
车载空调	301.5	83.65	<0.001	
车载 CD 播放器	255.7	92.95	0.006	
电动车窗	458.9	80.39	<0.001	
运动模式	475.8	78.95	<0.001	

该模型说明：控制其他因素不变，已使用时间（月）增加 1 月，二手车价格平均下降 116.6 欧元；控制其他因素不变，已行驶公里数增加 1 公里，二手车价格平均下降 0.0166 欧元；控制其他因素不变，高税组比低税组平均贵 1357 欧元；控制其他因素不变，二手车重量增加 1，二手车价格平均上升 16.44 欧元；控制其他因素不变，有牵引杆的二手车比没有牵引杆平均便宜 136.2 欧元；控制其他因素不变，两门车比四门车平均贵 235.4 欧元，三门车比四门车平均便宜 394.7 欧元，五门车比四门车平均便宜 295.9 欧元；控制其他因素不变，燃料为柴油的二手车比燃料为天然气的二手车平均便宜 44.56 欧元，燃料为汽油的二手车比燃料为天然气的二手车平均贵 1911 欧元；控制其他因素不变，高马力的二手车比低马力的二手车平均贵 3497 欧元；控制其他因素不变，有保修的二手车比没有保修的平均贵 327.2 欧元；控制其他因素不变，有 ABS 的二手车比没有 ABS 的平均便宜 235.7 欧元；控制其他因素不变，有安全气囊的二手车比没有安全气囊的平均便宜 256.8 欧元；控制其他因素不变，有空调的二手车比没有的平均贵 301.5 欧元；控制其他因素不变，有 CD 播放器的二手车比没有的平均贵 255.7 欧元；控制其他因素不变，有电动车窗的二手车比没有的平均贵

458.9 欧元；控制其他因素不变，有运动模式的二手车比没有的平均贵 255.7 欧元；

且该模型的调整 R^2=0.8848，说明该模型能解释 0.8848 的因变量的变异情况。该模型的 p 值<0.001，说明该模型是显著的，可以有效解释自变量和因变量之间的关系。

对该模型进行回归诊断：

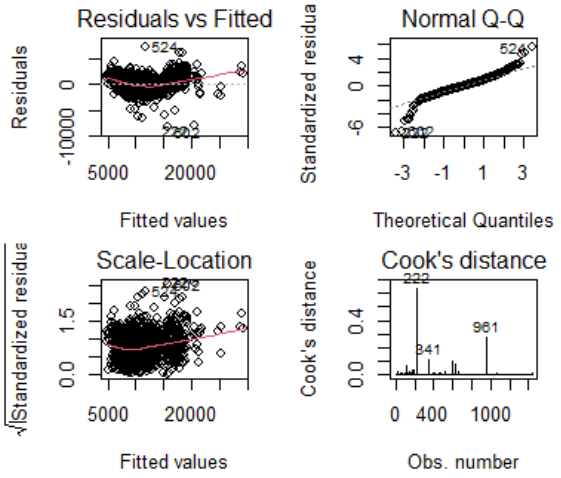

图 26 模型诊断图

发现 QQ 图两端尾部偏离直线的情况有所改善，cook 距离图中也不再存在强影响点。

表格 5 方差膨胀因子结果

变量名	GVIF	DF	GVIF^(1/(2*DF))
AGE	2.76	1	1.66
KM	2.04	1	1.43
QUARTERLY_TAX	4.77	1	2.18
WEIGHT	3.21	1	1.79
TOW_BAR	1.07	1	1.03
DOORS	1.37	3	1.05
FUEL_TYPE	5.43	2	1.53
HP	1.35	1	1.16
MFR_GUARANTEE	1.11	1	1.05
ABS	2	1	1.41
AIRBAG_2	2.03	1	1.43
AIRCO	1.66	1	1.29
CD_PLAYER	1.4	1	1.18
POWERED_WINDOWS	1.51	1	1.23
SPORT_MODEL	1.24	1	1.11

对该模型计算方差膨胀因子，发现该模型中基本不存在严重的多重共线性问题。

六、附录代码

```
library(mice)
library(ggplot2)
library(corrplot)
library(car)

#读取数据&预处理
```

```r
car<-read.csv('D://学习//大三上//探索性数据分析
//hw1//ToyotaCorolla_part.csv',header=TRUE)
head(car)
md.pattern(car)
car$Quarterly_Tax<-cut(car$Quarterly_Tax,c(0,85.5,Inf))
car$CC<-cut(car$CC,c(0,1601,Inf))
car$HP<-cut(car$HP,c(0,110.5,Inf))
car$Doors<-
factor(as.character(car$Doors),levels=c("4","2","3","5"))
car$Fuel_Type<-factor(as.character(car$Fuel_Type),levels =
c("CNG","Diesel","Petrol"))


#可视化
#因变量价格
boxplot(car$Price)
summary(car$Price) #均值在 10731，有一些数值极大的离群值
par(mfrow=c(1,1))
hist(x=log(car$Price),col='lightsteelblue4')
#价格+使用时长
ggplot(car)+geom_point(aes(x=Age,y=Price)) #线性负相关
cor(car$Price,car$Age,use='complete.obs')
summary(aov(car$Price~car$Age))
#已使用公里数+价格
ggplot(car,aes(x=KM,y=Price))+geom_point()
cor(car$Price,car$KM,use = 'complete.obs')
#是否在保修期+价格
car$Mfr_Guarantee<-as.character(car$Mfr_Guarantee)
ggplot(car)+geom_boxplot(aes(x=Mfr_Guarantee,y=Price,fill=Mfr_Guara
ntee)) #发现在保修期的二手车价格更高
#Quarterly_Tax + Price
ggplot(car)+geom_boxplot(aes(x=Tax,y=Price,fill=Tax))
#Weight+Price
ggplot(car,aes(x=Weight,y=Price,))+geom_point()
car$Tow_Bar<-as.character(car$Tow_Bar)
ggplot(car)+geom_boxplot(aes(x=Tow_Bar,y=Price,fill=Tow_Bar))
car$Automatic<-as.character(car$Automatic)
ggplot(car)+geom_boxplot(aes(x=Automatic,y=Price,fill=Automatic))
car$CC<-as.character(car$CC)
ggplot(car)+geom_boxplot(aes(x=CC,y=Price,fill=CC))
car$Doors<-as.character(car$Doors)
ggplot(car)+geom_boxplot(aes(x=Doors,y=Price,fill=Doors))
ggplot(car)+geom_boxplot(aes(x=Fuel_Type,y=Price,fill=Fuel_Type))
ggplot(car)+geom_boxplot(aes(x=HP,y=Price,fill=HP))
```

```
car$ABS<-as.character(car$ABS)
ggplot(car)+geom_boxplot(aes(x=ABS,y=Price,fill=ABS))
car$Airbag_2<-as.character(car$Airbag_2)
ggplot(car)+geom_boxplot(aes(x=Airbag_2,y=Price,fill=Airbag_2))
car$Airco<-as.character(car$Airco)
ggplot(car)+geom_boxplot(aes(x=Airco,y=Price,fill=Airco))
car$Boardcomputer<-as.character(car$Boardcomputer)
ggplot(car)+geom_boxplot(aes(x=Boardcomputer,y=Price,fill=Boardcomp
uter))
car$CD_Player<-as.character(car$CD_Player)
ggplot(car)+geom_boxplot(aes(x=CD_Player,y=Price,fill=CD_Player))
car$Central_Lock<-as.character(car$Central_Lock)
ggplot(car)+geom_boxplot(aes(x=Central_Lock,y=Price,fill=Central_Lo
ck))
car$Powered_Windows<-as.character(car$Powered_Windows)
ggplot(car)+geom_boxplot(aes(x=Powered_Windows,y=Price,fill=Powered
_Windows))
car$Radio<-as.character(car$Radio)
ggplot(car)+geom_boxplot(aes(x=Radio,y=Price,fill=Radio))
car$Sport_Model<-as.character(car$Sport_Model)
ggplot(car)+geom_boxplot(aes(x=Sport_Model,y=Price,fill=Sport_Model
))
car$Backseat_Divider<-as.character(car$Backseat_Divider)
ggplot(car)+geom_boxplot(aes(x=Backseat_Divider,y=Price,fill=Backse
at_Divider))
car$Met_Color<-as.character(car$Met_Color)
ggplot(car)+geom_boxplot(aes(x=Met_Color,y=Price,fill=Met_Color))

summary(aov(car$Price~car$KM))
summary(aov(car$Price~car$Mfr_Guarantee))
summary(aov(car$Price~car$Quarterly_Tax))
summary(aov(car$Price~car$Weight))
summary(aov(car$Price~car$Tow_Bar))
summary(aov(car$Price~car$Automatic))#
summary(aov(car$Price~car$CC))
summary(aov(car$Price~car$Doors))
summary(aov(car$Price~car$Fuel_Type))
summary(aov(car$Price~car$HP))
summary(aov(car$Price~car$ABS))
summary(aov(car$Price~car$Airbag_2))
summary(aov(car$Price~car$Airco))
summary(aov(car$Price~car$Boardcomputer))
summary(aov(car$Price~car$CD_Player))
summary(aov(car$Price~car$Central_Lock))
```

```
summary(aov(car$Price~car$Sport_Model))
summary(aov(car$Price~car$Backseat_Divider))
summary(aov(car$Price~car$Met_Color))
summary(aov(car$Price~car$Powered_Windows))
summary(aov(car$Price~car$Radio))#


#线性回归
#未取价格对数
lm_car1<-
lm(Price~Automatic+Radio+Age+KM+Quarterly_Tax+Weight+Tow_Bar+CC+Doo
rs+Fuel_Type+HP+Mfr_Guarantee+ABS+Airbag_2+Airco+Boardcomputer+CD_P
layer+Central_Lock+Powered_Windows+Sport_Model+Backseat_Divider+Met
_Color,data=car)
summary(lm_car1) #AR=0.8794
par(mfrow=c(2,2),mai=c(0.75,0.8,0.25,0.2))
plot(lm_car1,which = c(1,2,3,4)) #发现 QQ 图的尾巴偏离直线，因变量并不服从
正态分布
                                #cook 距离可见明显的强影响点
round(vif(lm_car1),2)
#取价格对数后
lm_car2<-
lm(log(Price)~Age+KM+Quarterly_Tax+Weight+Tow_Bar+CC+Doors+Fuel_Typ
e+HP+Mfr_Guarantee+ABS+Airbag_2+Airco+Boardcomputer+CD_Player+Centr
al_Lock+Powered_Windows+Sport_Model+Backseat_Divider+Met_Color,data
=car)
summary(lm_car2)#p=0.863
round(vif(lm_car2),2)
par(mfrow=c(2,2),mai=c(0.75,0.8,0.25,0.2))
plot(lm_car2,which = c(1,2,3,4)) #残差图可见明显的喇叭状，说明存在异方
差；QQ 图说明因变量仍然不服从真该分布

#AIC(用的是 log(price))
lm_car3<-step(lm_car2)
summary(lm_car3)#p=0.8632
round(vif(lm_car3),2)
drop1(lm_car3) #发现 drop 掉 doors 之后 AIC 增加的最少（增加 0）
#去掉 doors 之后重新回归
lm_car4<-
lm(log(Price)~Age+KM+Quarterly_Tax+Weight+Tow_Bar+Automatic+Fuel_Ty
pe+HP+Mfr_Guarantee+Airbag_2+Airco+CD_Player+Powered_Windows+Sport_
Model+Backseat_Divider,data=car)
summary(lm_car4)
drop1(lm_car4)
```

```
#去掉 sport_model，airbag_2 之后重新回归
lm_car5<-
lm(log(Price)~Age+KM+Quarterly_Tax+Weight+Tow_Bar+Automatic+Fuel_Ty
pe+HP+Mfr_Guarantee+Airco+CD_Player+Powered_Windows+Backseat_Divide
r,data=car)
summary(lm_car5)

par(mfrow=c(2,2),mai=c(0.75,0.8,0.25,0.2))
plot(lm_car2,which = c(1,2,3,4))

#AIC(not log(price))
lm_car6<-step(lm_car1)
summary(lm_car6) #ar=0.8848
round(vif(lm_car6),2)
par(mfrow=c(2,2),mai=c(0.75,0.8,0.25,0.2))
plot(lm_car6,which = c(1,2,3,4))

lm_car7<-
lm(Price~Airco+CD_Player+Mfr_Guarantee+Powered_Windows+Sport_Model+
HP+Quarterly_Tax+Weight+KM+Age,data=car)
round(vif(lm_car7),2)
lm_car8<-
lm(Price~Age+KM+Quarterly_Tax+Weight+Tow_Bar+Automatic+Fuel_Type+HP
+Mfr_Guarantee+Airco+CD_Player+Powered_Windows+Sport_Model+Backseat
_Divider,data=car)
summary(lm_car7) #AR=0.8766
par(mfrow=c(2,2),mai=c(0.75,0.8,0.25,0.2))
plot(lm_car7,which = c(1,2,3,4))
```