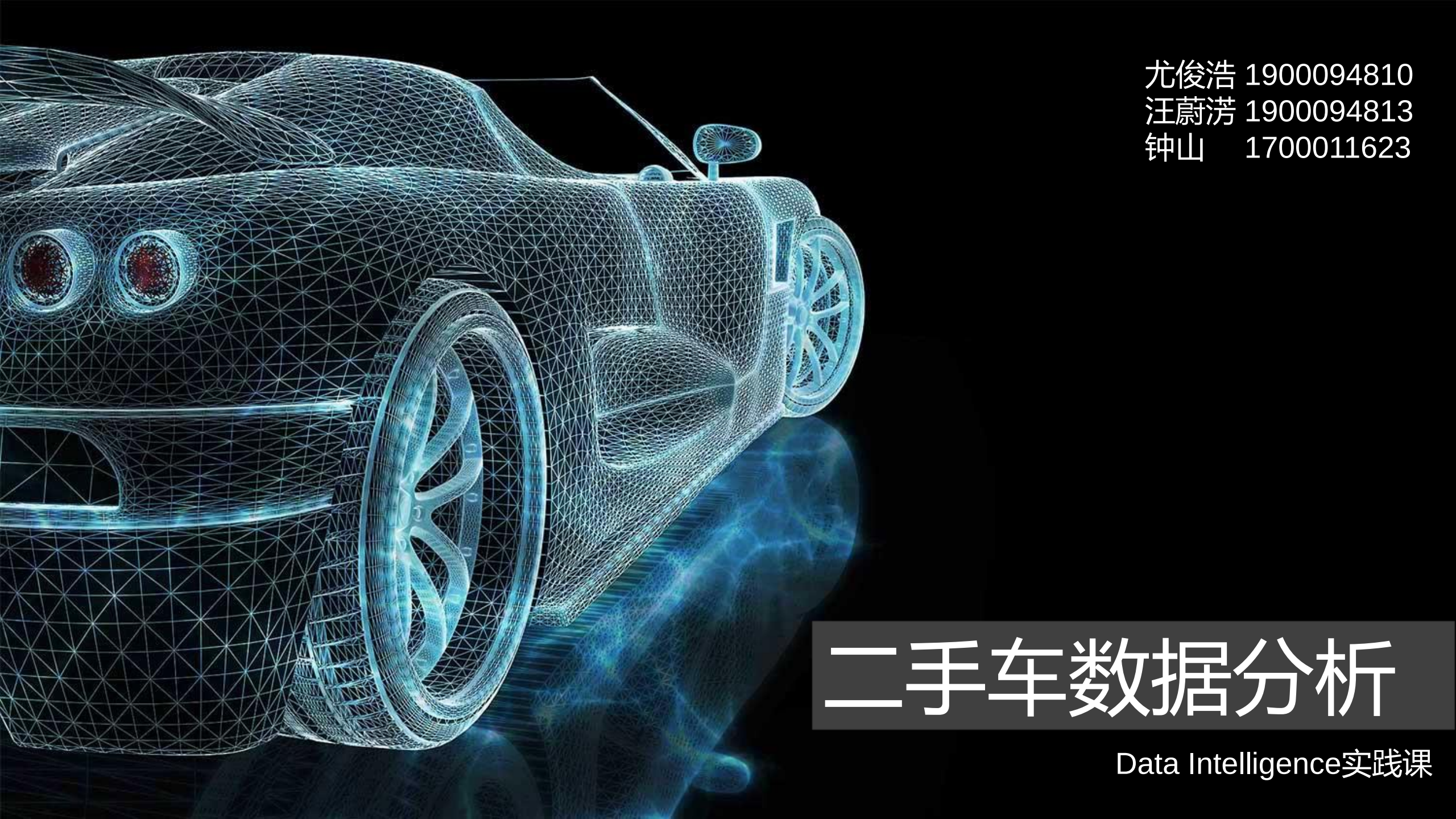
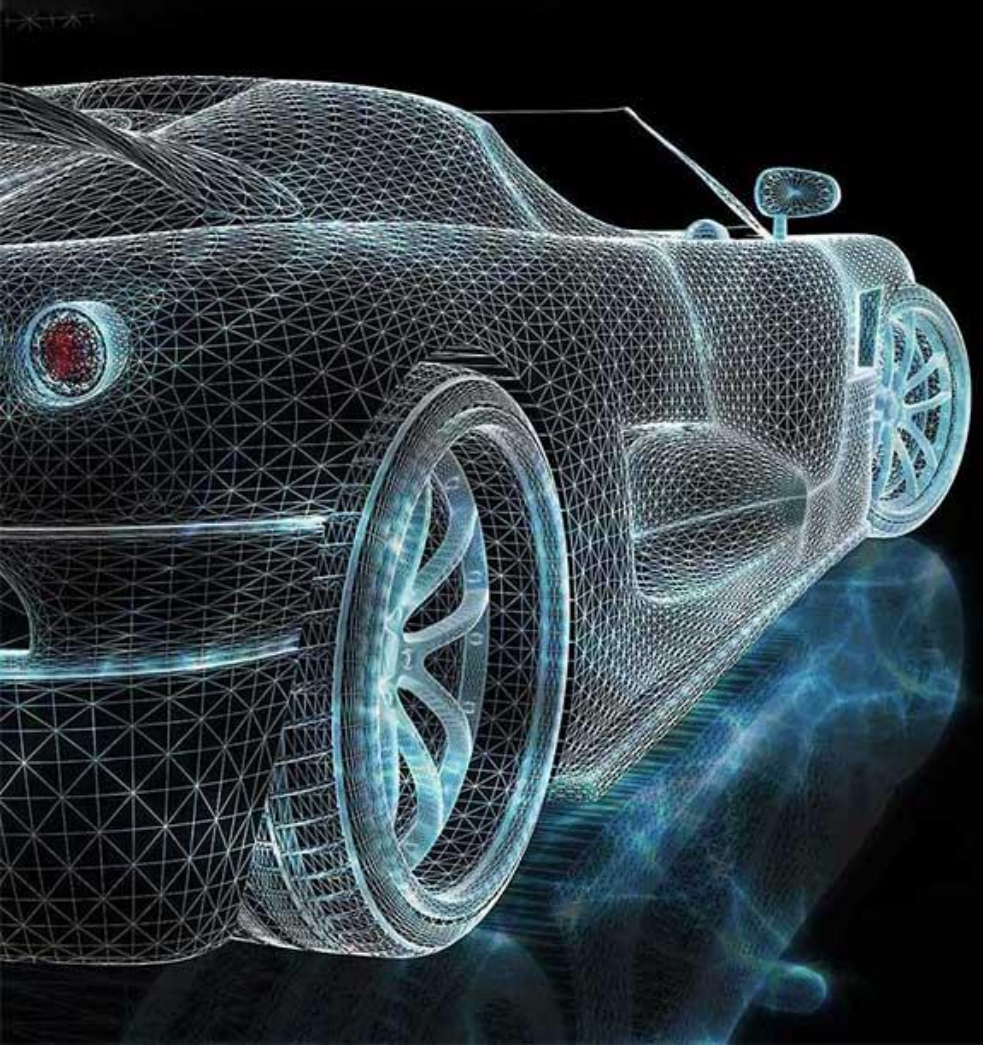


尤俊浩 1900094810
汪蔚涝 1900094813
钟山　1700011623

二手车数据分析

Data Intelligence实践课

大纲



01 项目背景

02 数据预处理

03 数据分析

04 回归模型的构建

1 项目背景

简 单 介 绍 此 项 目

目标

对从Kaggle得到Ebay-Kleinanzeigen（德国网络交易平台）的二手车数据集进行分析

1. 分析各变量的相关性

尝试找出有趣的关系

2. 分析价格与其他变量的相关性

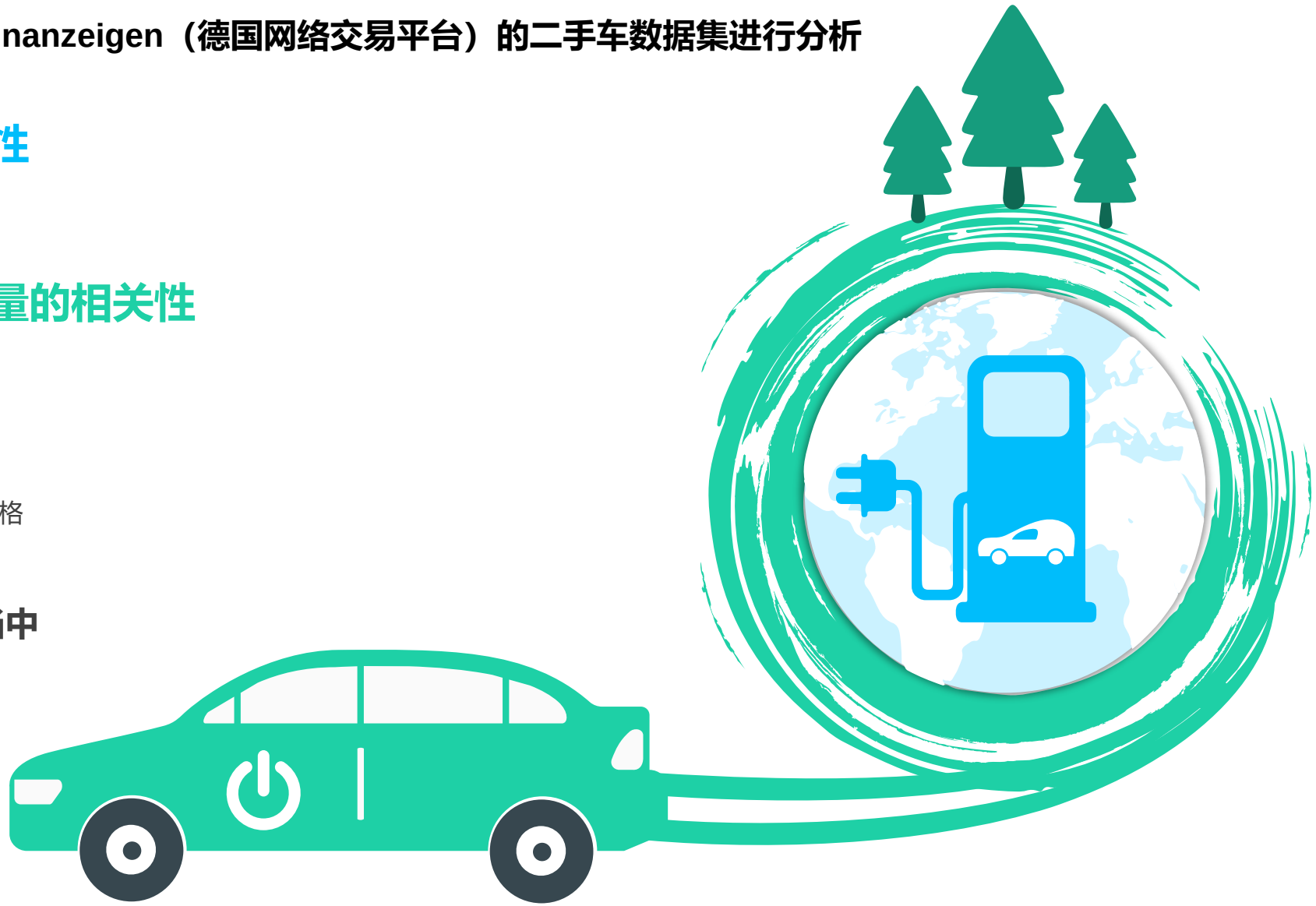
探究影响二手车价格的因素

3. 构建回归模型

通过机器学习预测二手车的价格

从不同的维度挖掘数据集当中

有用的信息



2 数据预处理

数据概览和数据清洗

数据概况

- 来自于Ebay 上的二手车转卖专区
- 数据于2016年爬取
- 德国的网络交易平台

371528
Total Data

303820
After Cleaning

20
Variables

面临的问题

-语言不统一
-数据杂乱无章
-有些变量没有意义
-很多重复数据
-有很多空值

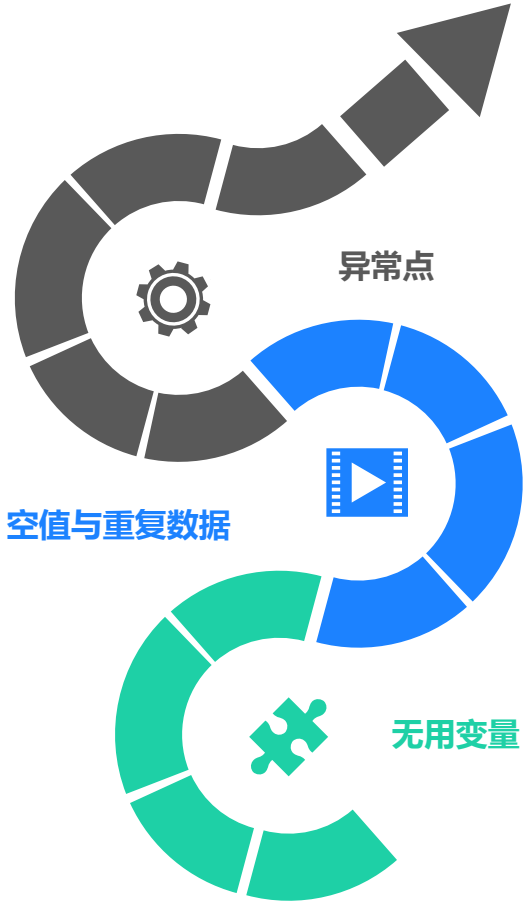
待处理的部分

重复数据量

```
In [57]: df.duplicated(keep='first').sum()
Out[57]: 6842
```

空值分布

name	0
price	0
vehicleType	10791
yearOfRegistration	0
gearbox	5313
powerPS	0
model	11383
kilometer	0
monthOfRegistration	0
fuelType	15376
brand	0
notRepairedDamage	42856
car_age	0
horse_power	0



不合理的值

	price	yearOfRegistration	powerPS	kilometer	monthOfRegistration
count	3.715280e+05	371528.000000	371528.000000	371528.000000	371528.000000
mean	1.729514e+04	2004.577997	115.549477	125618.688228	5.734445
std	3.587954e+06	92.866598	192.139578	40112.337051	3.712412
min	0.000000e+00	1000.000000	0.000000	5000.000000	0.000000
25%	1.150000e+03	1999.000000	70.000000	125000.000000	3.000000
50%	2.950000e+03	2003.000000	105.000000	150000.000000	6.000000
75%	7.200000e+03	2008.000000	150.000000	150000.000000	9.000000
max	2.147484e+09	9999.000000	20000.000000	150000.000000	12.000000

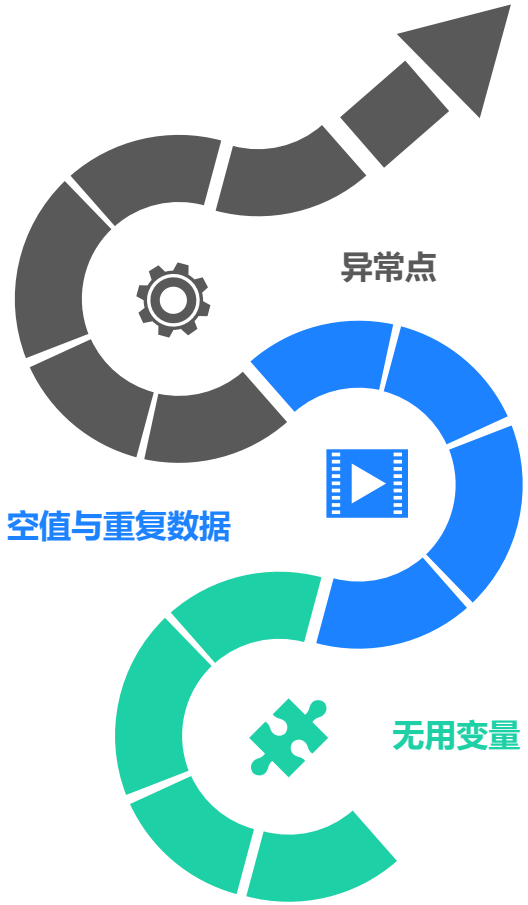
对分析无用的变量

“nrOfPictures” 是无效的
“seller” 与 “offerType” 值占比悬殊
“postalCode”、“dateCreated”、“dateCrawled”等变量对分析无用

处理方式

填补空值

采用Back Fill来填充空值



仅保留正常辖域的变量

保留车价处于100$到200000$的数据
保留注册年份处于1960年至2016年间的数据
保留引擎马力处于10PS至1000PS的数据

```
Too new: 14680
Too old: 289
Too cheap: 13320
Too expensive:  232
Too few km:  0
Too many km:  0
Too few PS:  41040
Too many PS:  835
```

去除无用变量

把上述所提到的变量都去除，其中包括了：“dateCrawled”、“offerType”、“postalCode”、“dateCreated”、“lastSeen”、“nrOfPictures”、“abtest”、“seller”

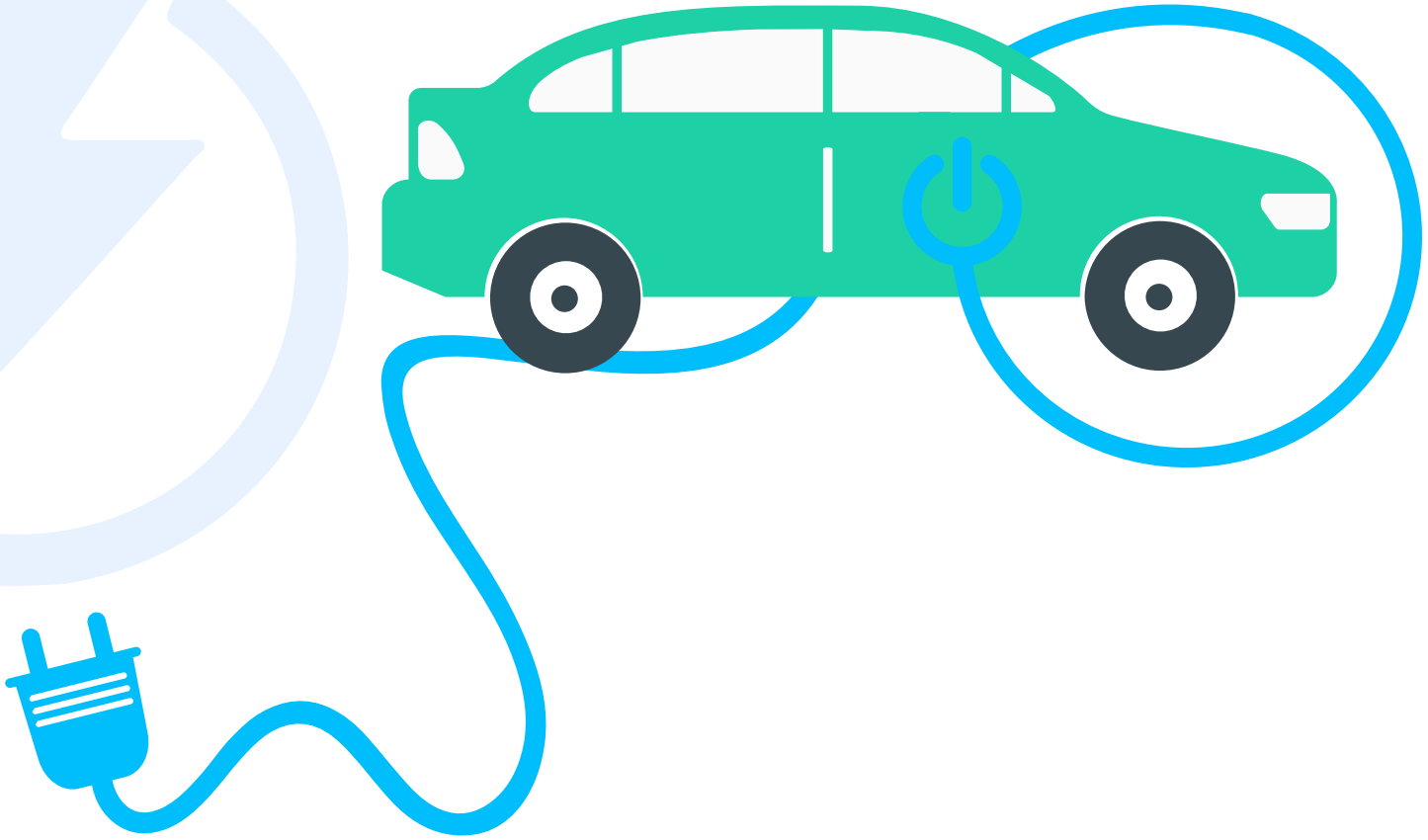
进一步的处理

1 **翻译数据**
把德文的变量名以及值都翻译成英文

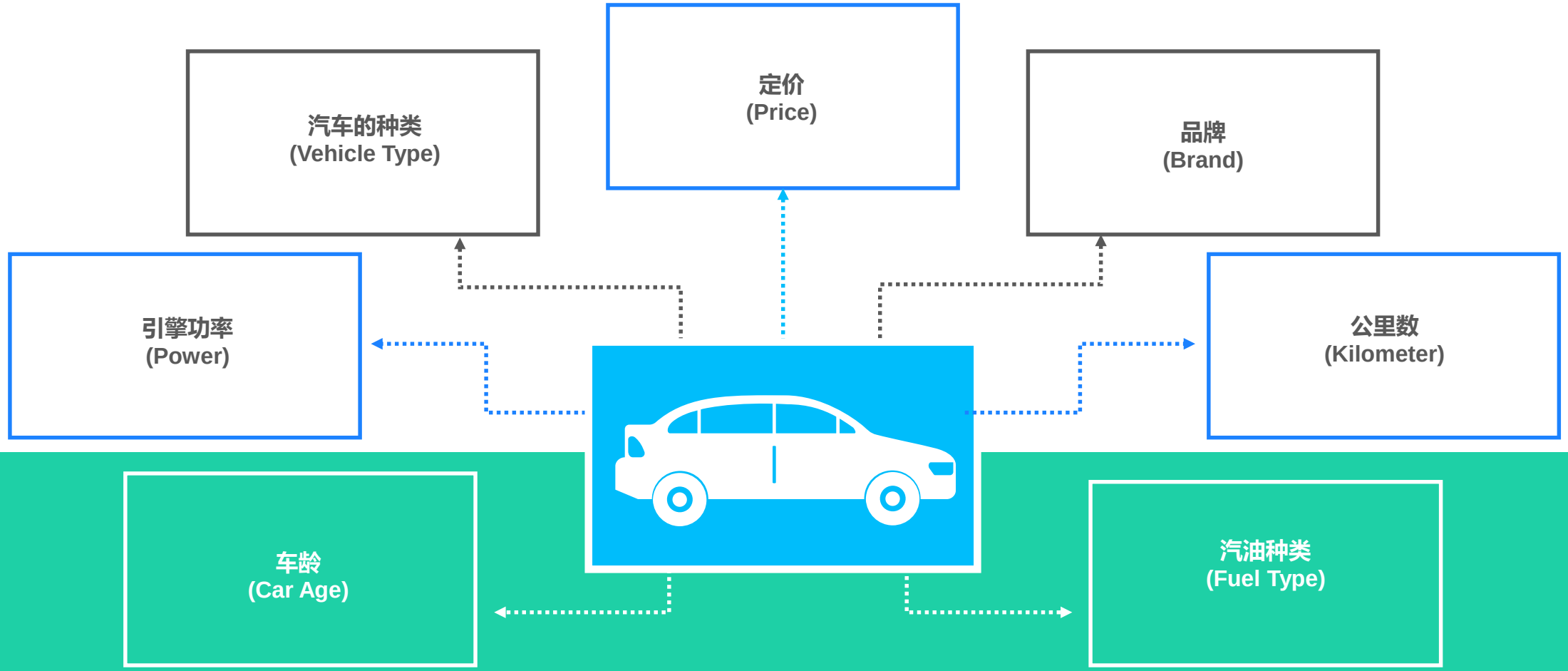
2 **创建中文词典**
方便数据可视化时通过中文来表示

3 **创建新的变量**
用车龄“car_age”来表示从注册年份到2016年的间隔

4 **单位的统一**
把公制马力PS转换成常用的马力HP



几个重要的变量


汽车的种类
(Vehicle Type)
定价
(Price)
品牌
(Brand)
引擎功率
(Power)
公里数
(Kilometer)
车龄
(Car Age)
汽油种类
(Fuel Type)

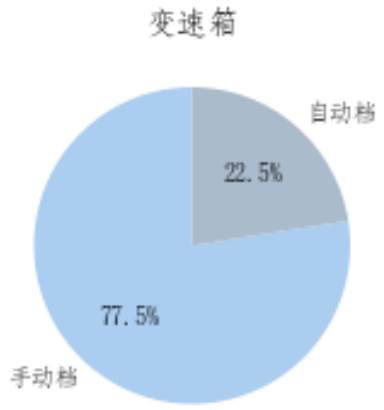
3 数据分析

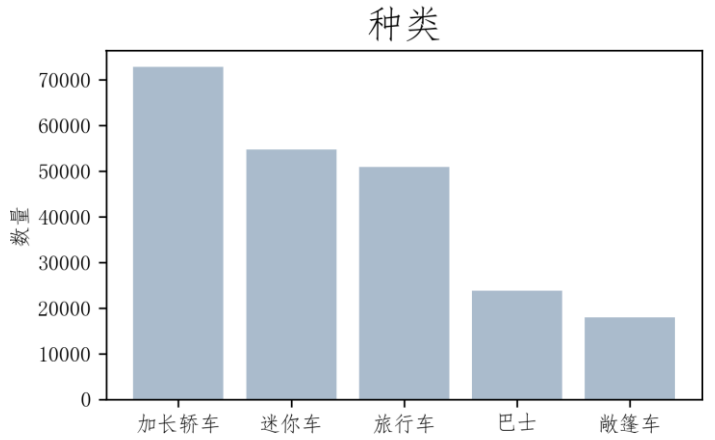
将数据可视化与进行相关性分析

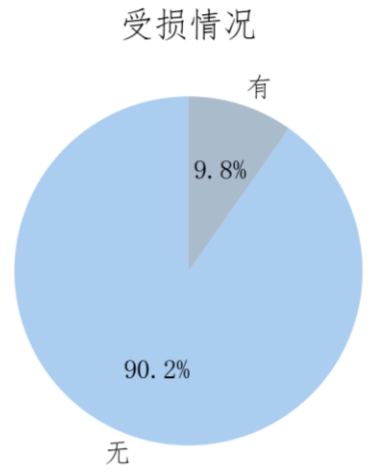
简单的视图

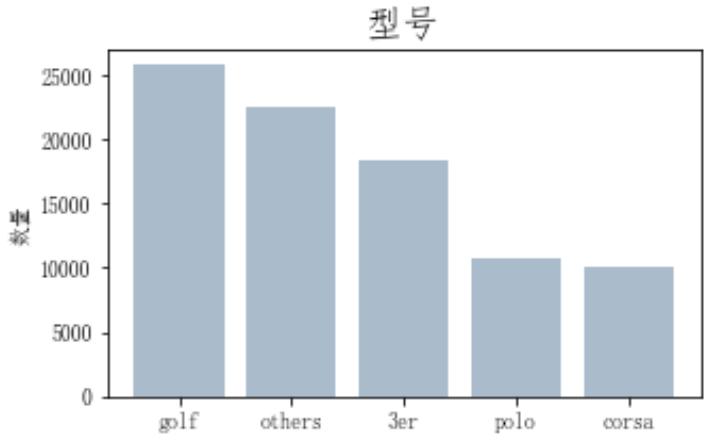
Brief View on the Variables

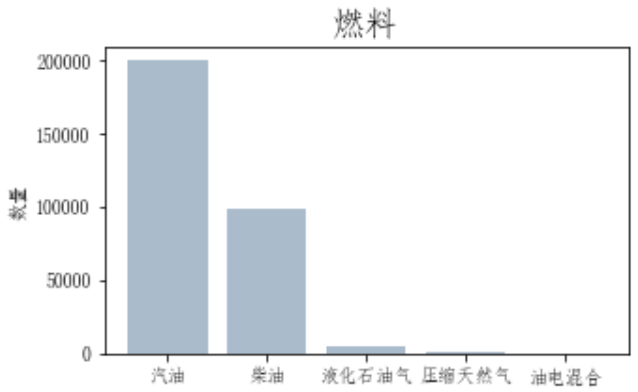
- ✓ **变速箱:**
  分为手动档和自动档
- ✓ **受损情况:**
  分为受损和无受损.
- ✓ **燃料:**
  分为为汽油、柴油、液化石油气、压缩天然气和油电混合
- ✓ **汽车种类:**
  主要有加长轿车、迷你车、旅行车、巴士和敞篷车
- ✓ **型号:**
  主要有golf, 3er, polo, corsa 等等

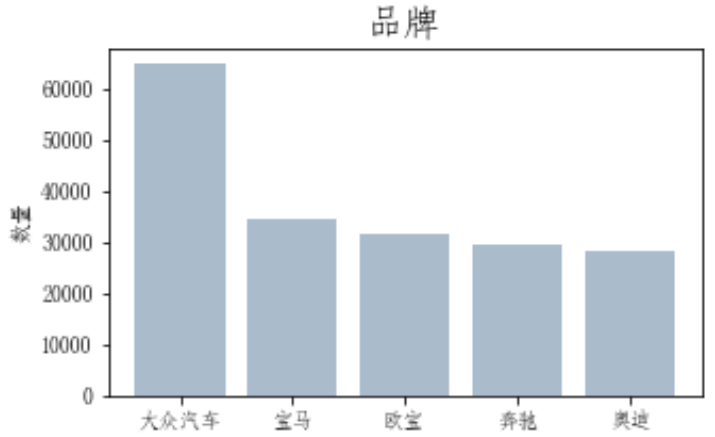




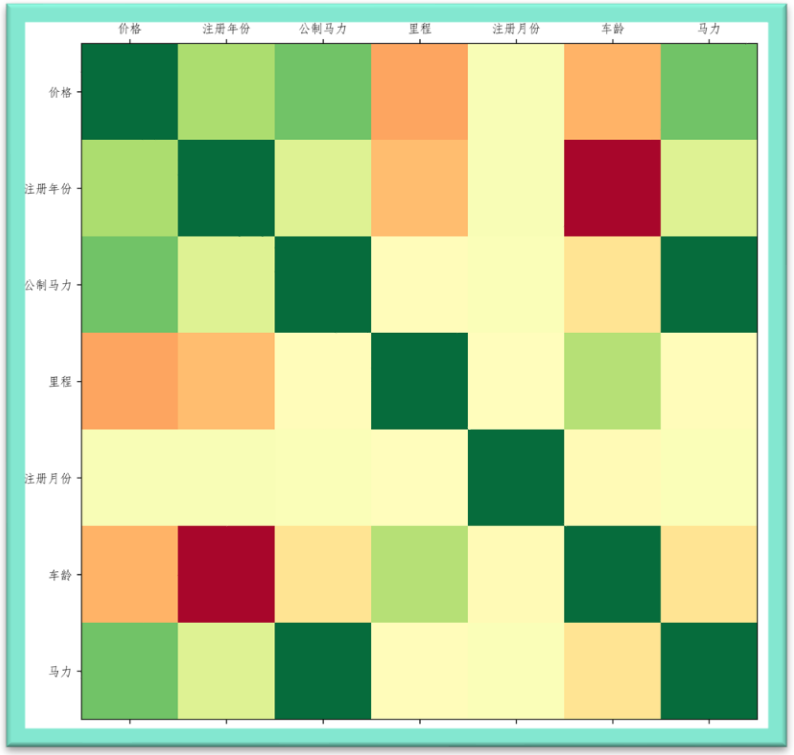


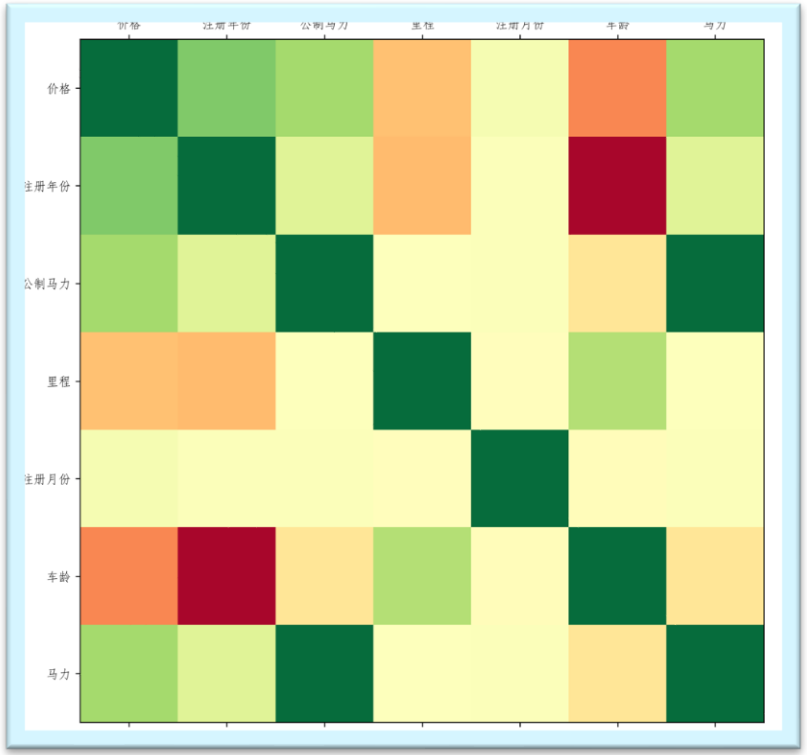


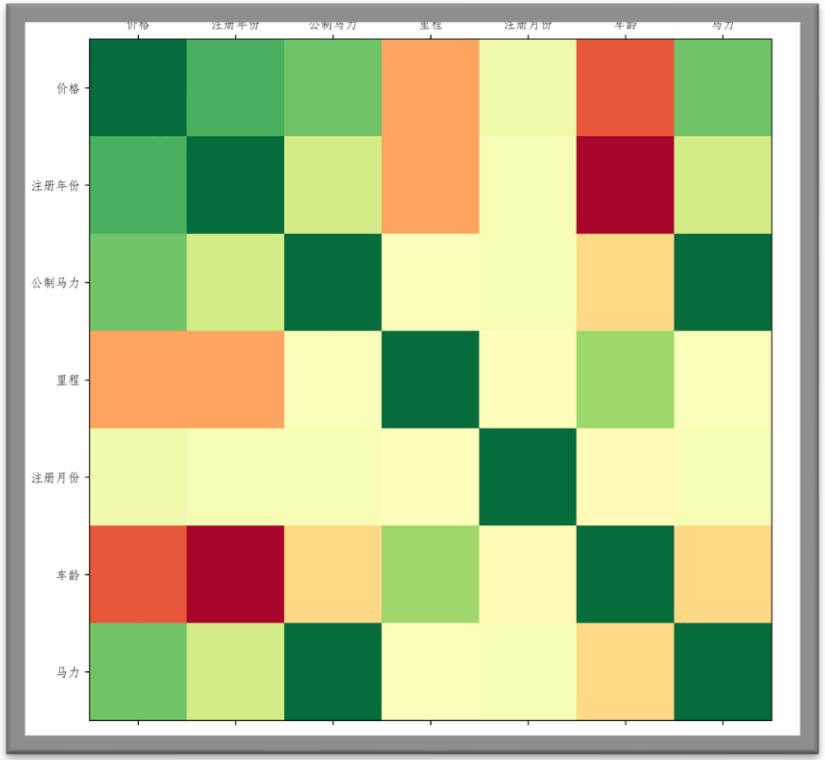




相关性分析 (Correlation Graph)







Pearson

衡量线行相关变量之间关系程度

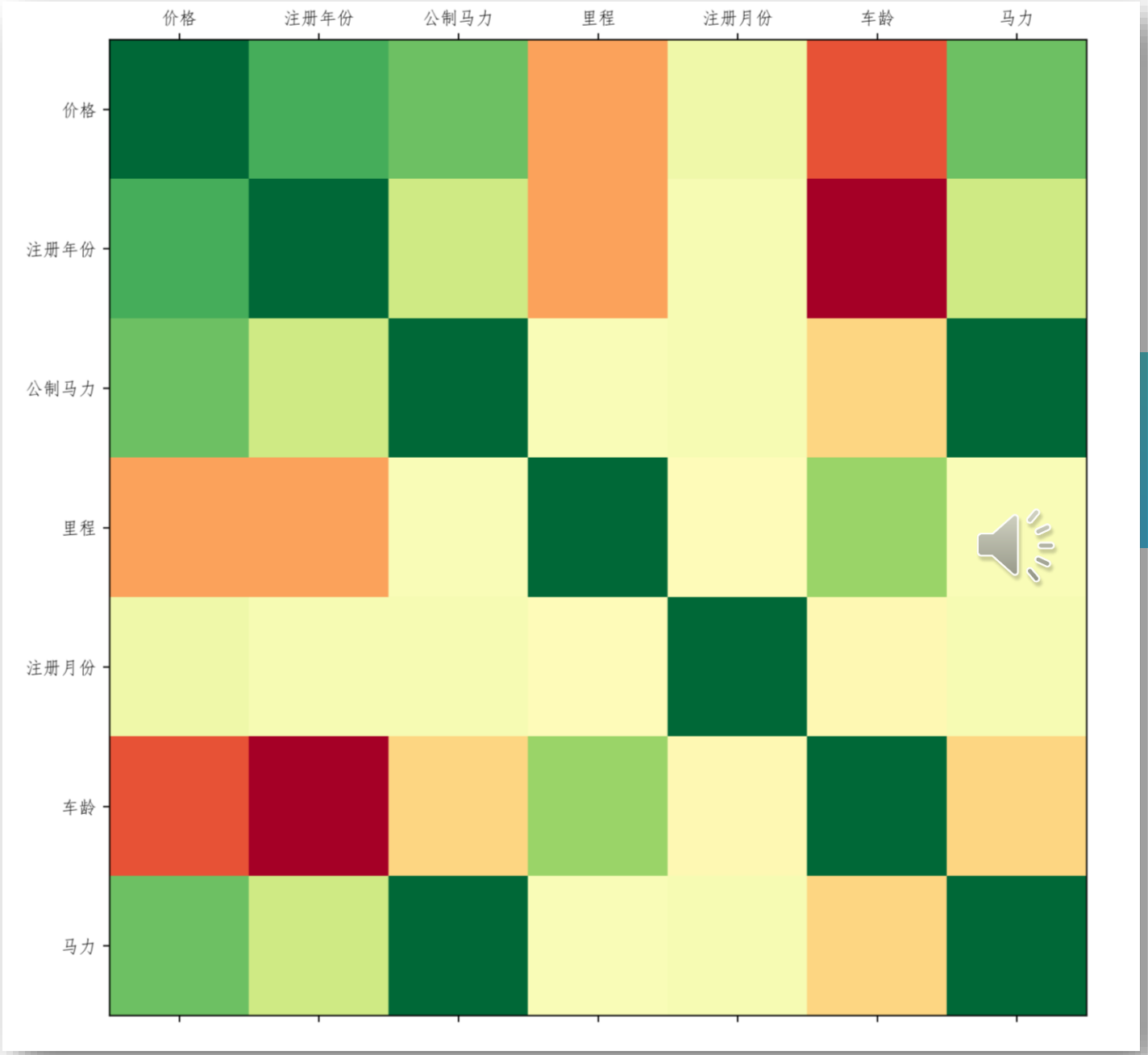
Kendall

适用于分类变量均为有序分类的情况

Spearman

适用范围最广，但统计效能要低一些

结果相差不大，用Spearman来进一步说明

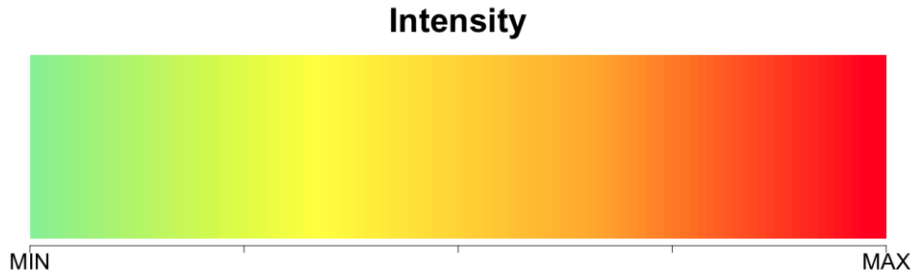


各变量
相关性 分析

Spearman Correlation

显然车龄和注册年份是**强负相关**的；**车龄与里程数**是**正相关**的。
马力与车龄也是**负相关**的，这可能是因为造车的工艺不断进步所致。

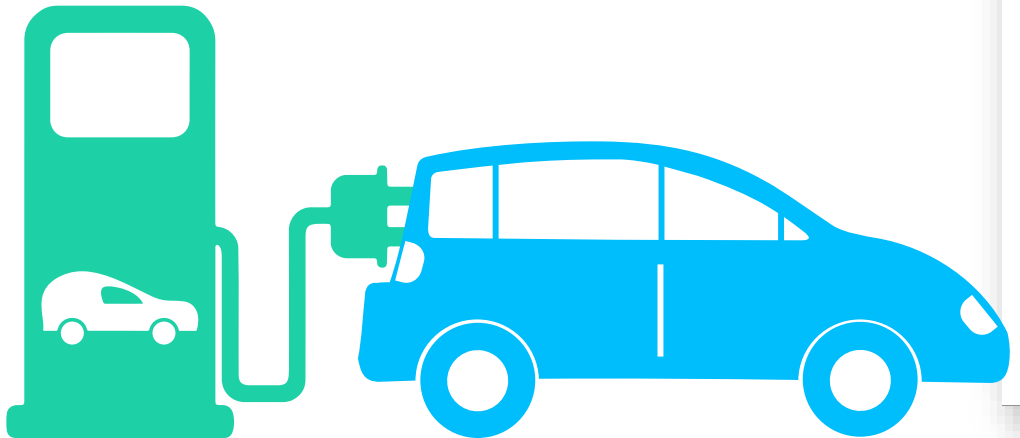
我们重点想要观察的是**价格**与其他变量的相关性
车的价格与其引擎功率是**正相关**的，而**与里程、车龄**都是**负相关**的。



热图 Heatmap

进一步的
具体化
数值化

车价与其变速箱种类相关
马力与变速箱种类、品牌和车的种类相关



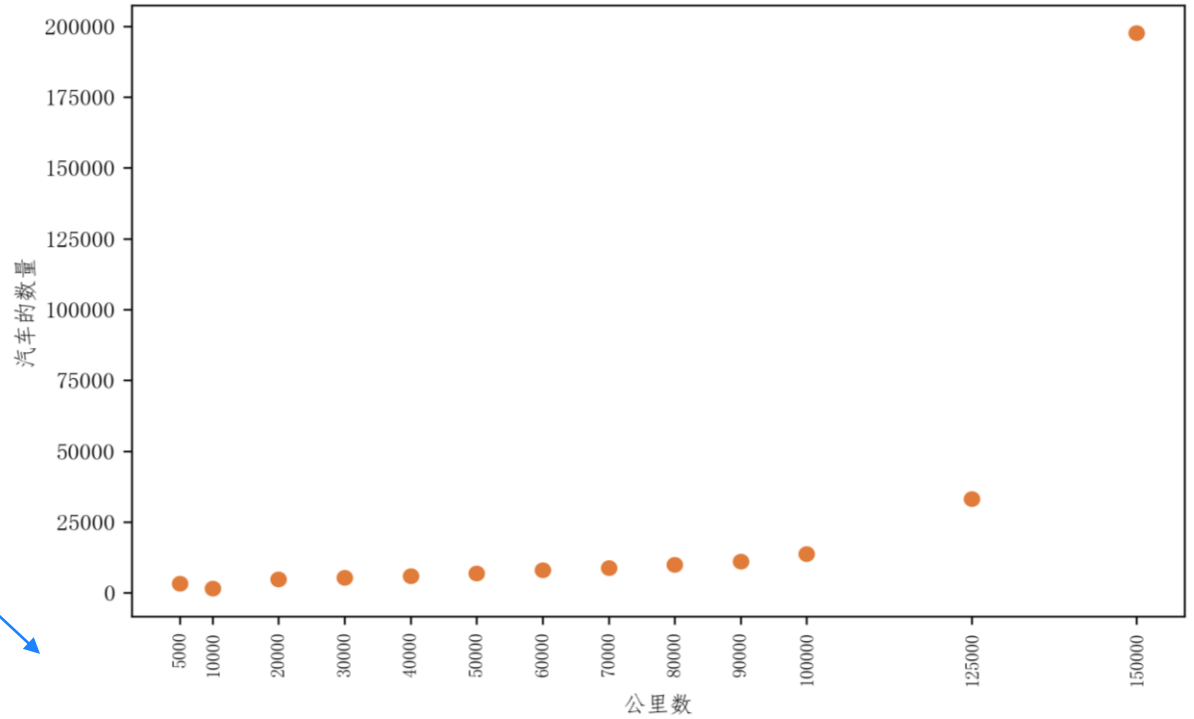
公里数 与 汽车数量 售卖价格 之间的关系

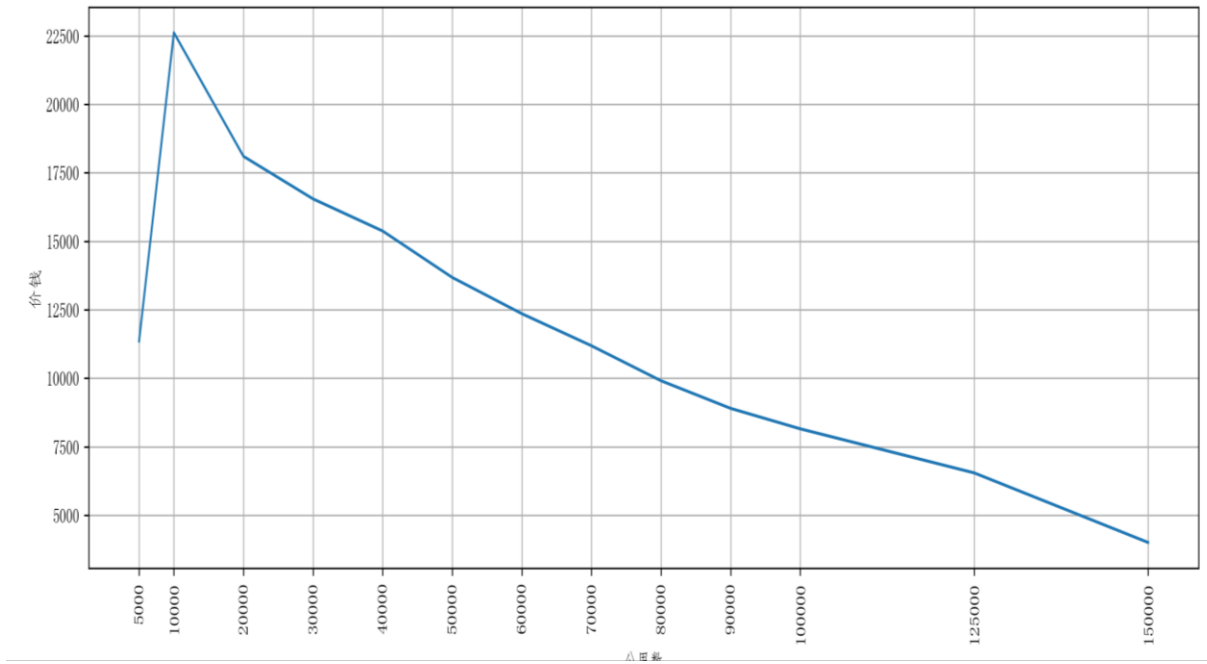
公里数是指汽车已经行驶过多少公里

公里数越低代表车子损耗程度越低

数据显示二手车公里数的分布介于5000至150000km

最低的公里数是5000km，数量少于5000辆
最高的则是150000km，数量在20000辆左右





公里数

越低

车价在10000km时达到峰值
买家会**相信**
二手车**保存**得比较**完善**

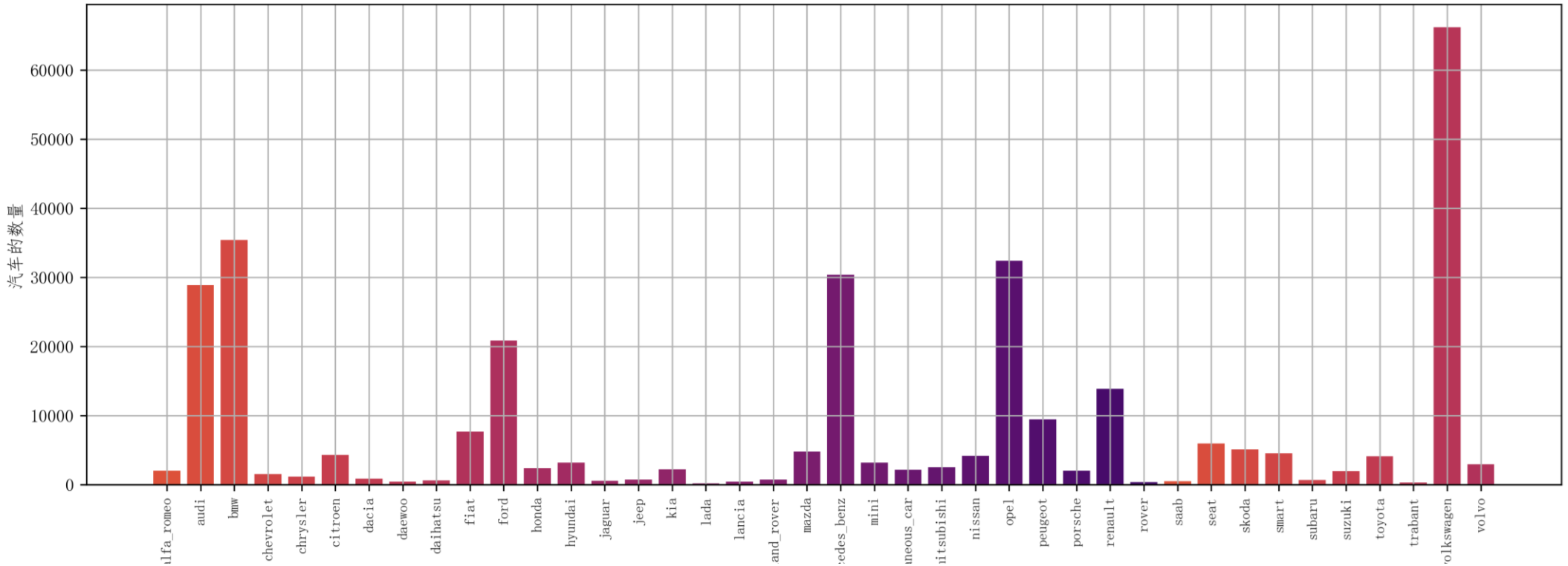
越高

二手车的价格越低
买家会**质疑**
二手车的**性能**和**零件状况**

图 表

直观的看一看变量之间的相关性

品牌与数量



20% 大众汽车(Volkswagen)
高达65000辆
远远超出其他

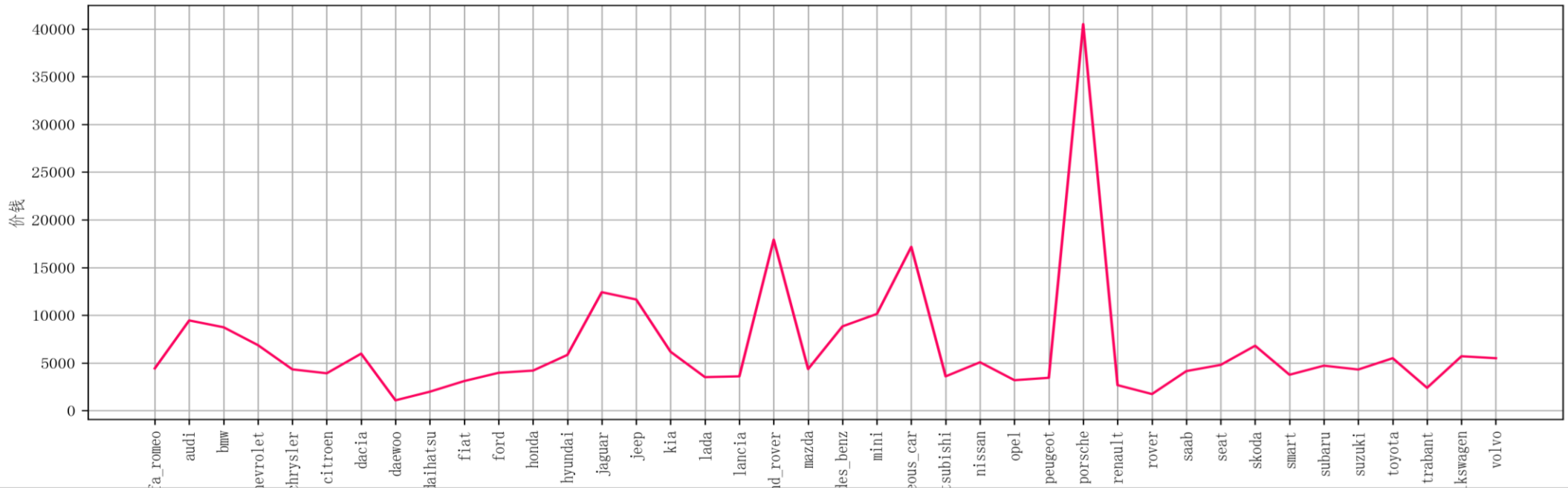
10% 宝马(BMW)
高达35000辆

9% 欧宝(Opel)
大约31000辆
与宝马差得不大

其次为 奔驰(Mercedes Benz)， 奥迪(Audi)， 福特(Ford)， 雷诺(Renault)

其他的均不超过1万辆

品牌与售卖价格



保时捷 (Porsche)

价格是最高的，远远超出其他
平均值为40000欧元
大约30万人民币

其他品牌(Sonstige Auto)

第三高，十分接近路虎的平均价格
平均值为16000欧元
大约12万人民币

路虎 (Land Rover)

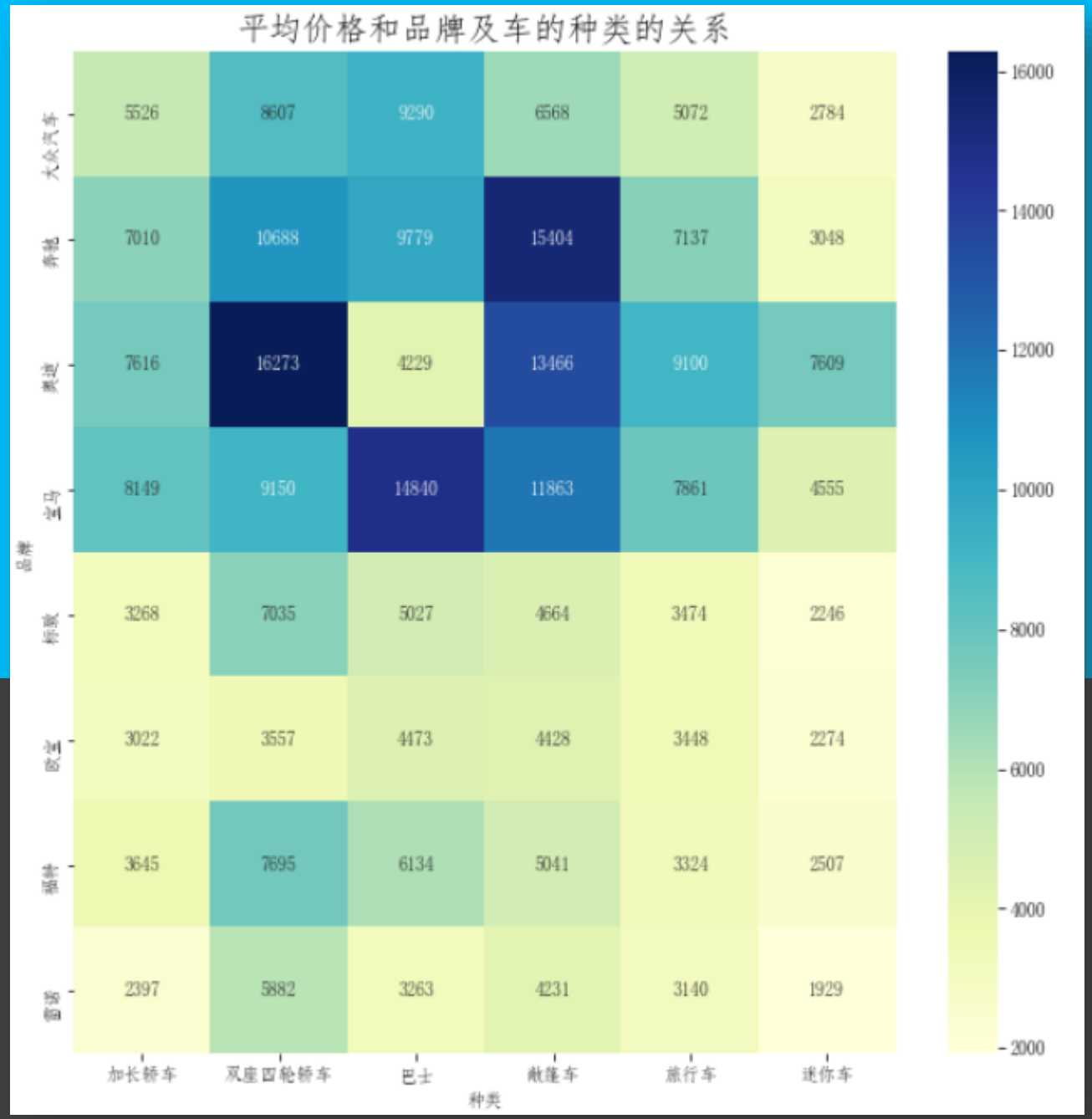
第二高
平均值为17000欧元
大约13万人民币

大宇汽车 (Daewoo)

最低
平均值为2000欧元
大约1万人民币

其他的都介于3000-10000欧元之间

常见品牌和汽车种类与价格之间的关系



最高 奥迪的双座四轮轿车

紧接着是
奔驰的敞篷车
宝马的巴士
奥迪的敞篷车

最低 雷诺的迷你车

紧接着是
福特、欧宝的迷你车
雷诺的加长轿车

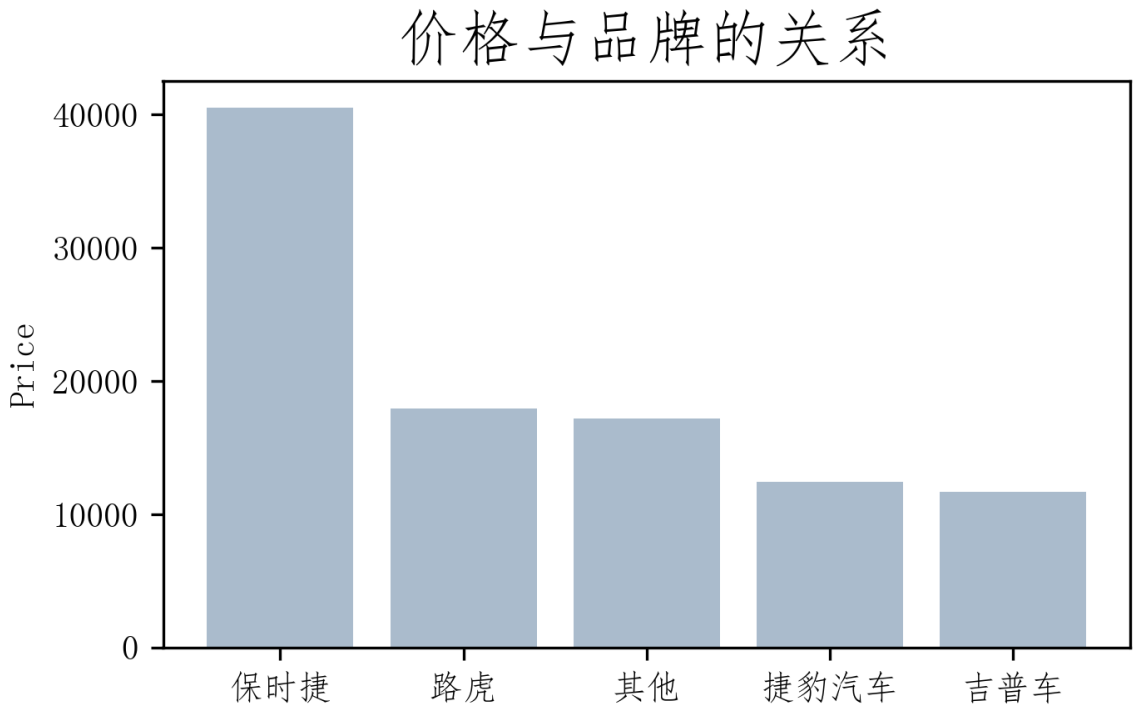
品牌作为汽车的耐久和性能的保证，而品牌自身的定位也成为二手车定价最重要的基准。

汽 车 品 牌

C a r B r a n d

小结 汽车品牌与数量、价格的关系

二手车售卖的价格都与品牌自身的价位有关；
二手车的数量也在一定程度上反映出了当地各品牌所占据的市场份额

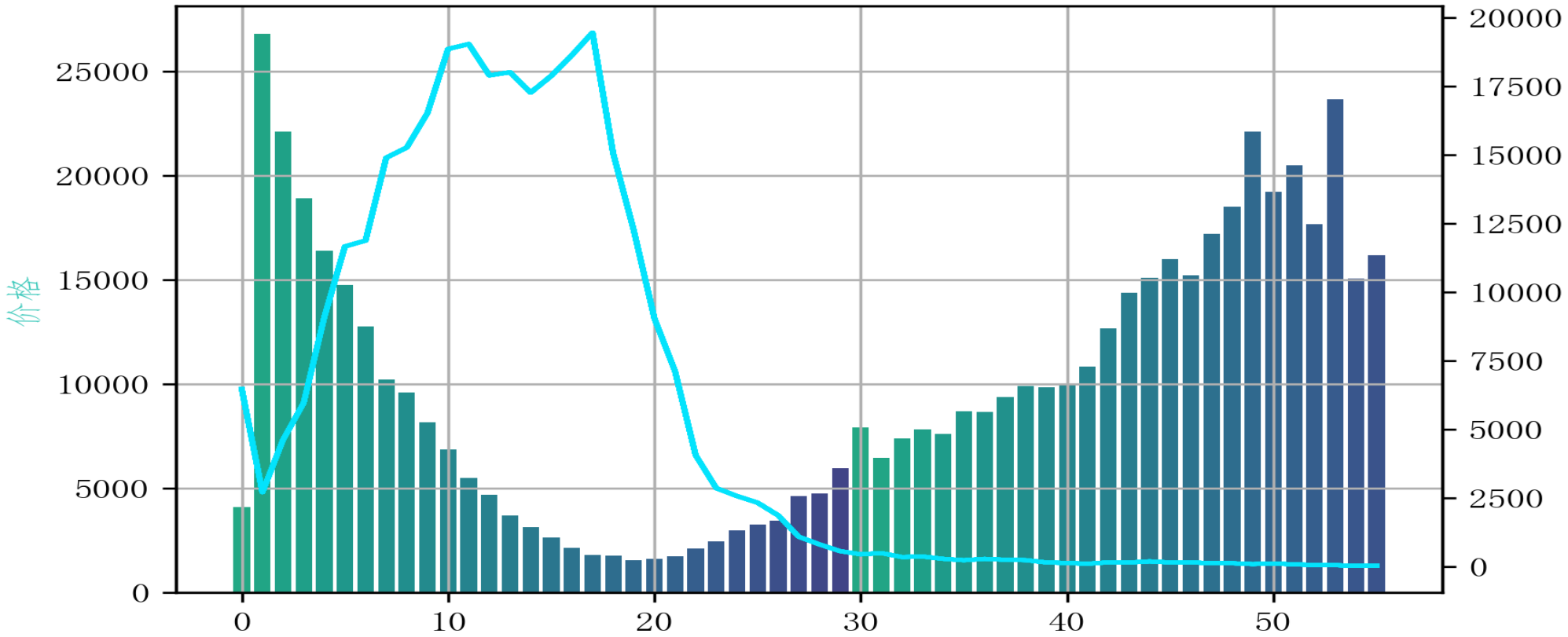


车龄与数量和价格的关系

大多数的二手车的车龄是介于10至20年之间
由此可见，**车主大多选择在购入汽车的10年后决定转卖汽车**
越新的车，价格越高，随后逐渐跌价，车龄到了大约30年就开始增加，证明**古董车**是具有一定收藏价值的

柱状图是价格

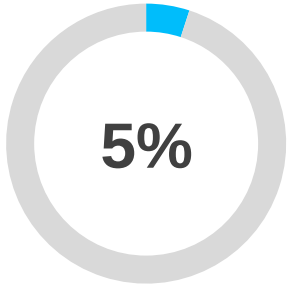
线型图是数量



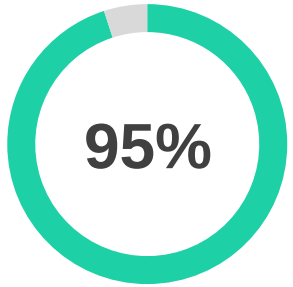
关于车龄的总结

数量

高车龄和低车龄



中间的车龄



高车龄和低车龄
中间的车龄

高车龄指大于40年，低车龄指小于5年
车龄介于10年至40年间

价格

以人民币来呈现

> 11万

高车龄和低车龄的车平均售价均大于11万
峰值更是达到了20万
可见新车或有一定车龄的古董车价值最高
新车车况保留比较好
古董车则是有一定的收藏意义

虽然中间车龄的车子占多数，但是价格完全不占优势
可见在市场供给越多的车子出价普遍较低

二手车的数量与价格有呈反比的趋势

4 回归模型

对价格进行预测

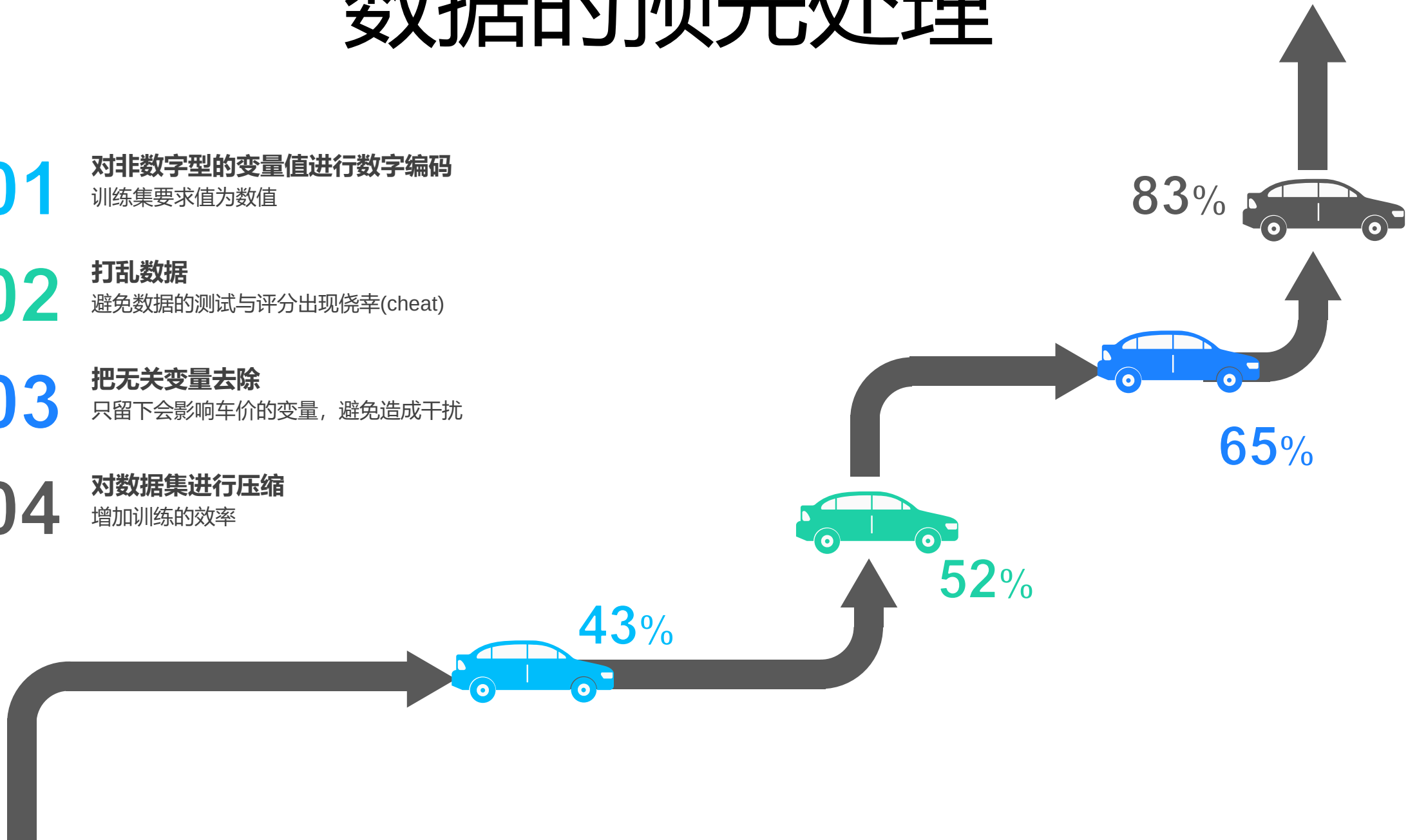
数据的预先处理

01 **对非数字型的变量值进行数字编码**
训练集要求值为数值

02 **打乱数据**
避免数据的测试与评分出现侥幸(cheat)

03 **把无关变量去除**
只留下会影响车价的变量，避免造成干扰

04 **对数据集进行压缩**
增加训练的效率



回归模型的构建

1.构造训练集与测试集

定义因变量y为车价，而自变量X为除了车价以外其余变量的集合。
分离数据，将200行数据加入训练集，而其余数据都纳入训练集。

2. 模型的选用

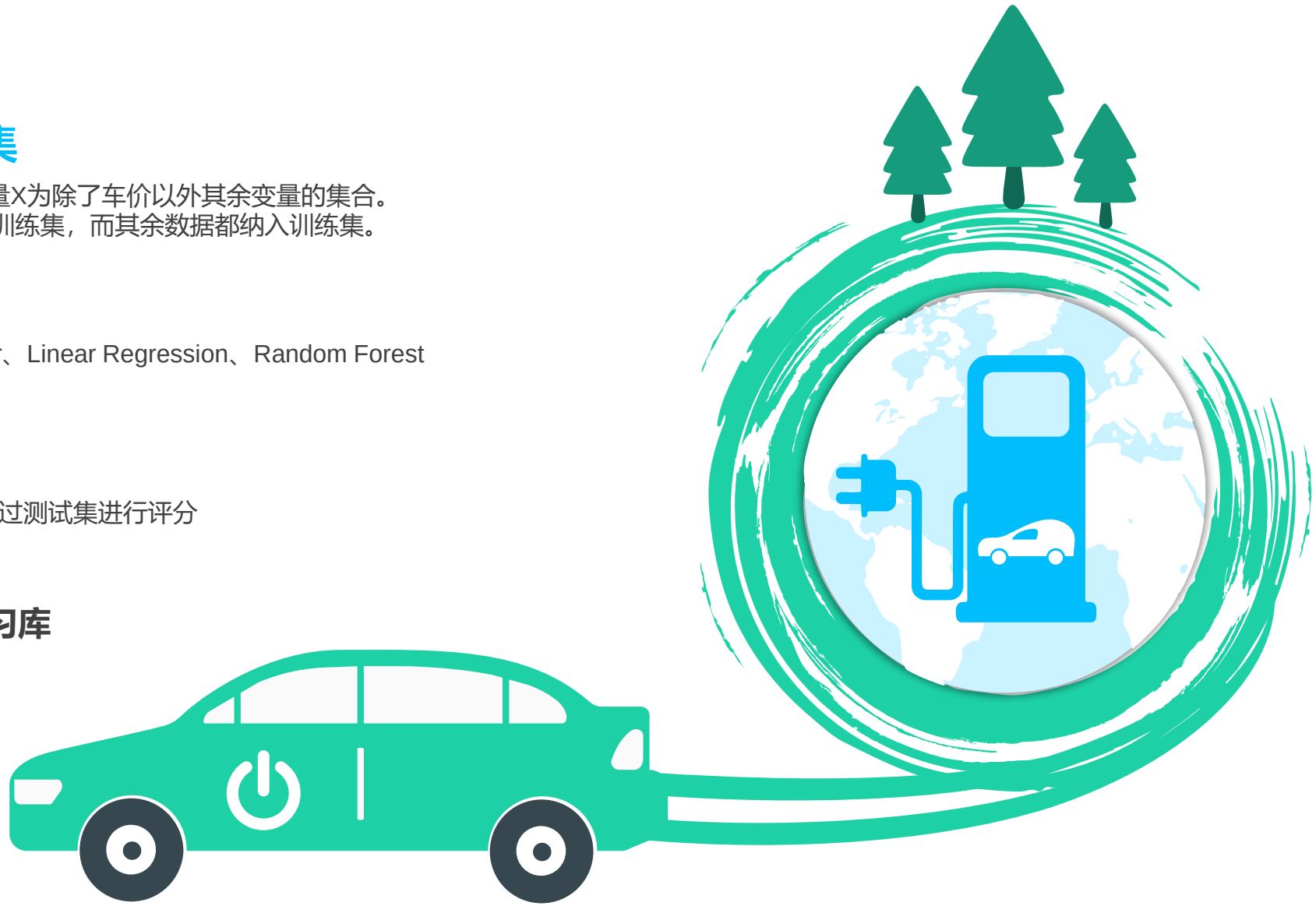
Linear SVR、SGD Regressor、Linear Regression、Random Forest Regressor

3. 模型的训练

把训练集喂给模型，过后再通过测试集进行评分

利用Python自带的机器学习库

构建回归模型！



评分结果

总的来说

评分结果看起来还可以接受，但是对于某些测试数据却出现了奇怪的预测结果。

比如对于第四个测试数据，三个模型都给出了卖家应该付钱来让人拿走他的车的结论。而对于部分数据来说，预测的结果还是相当准确的。

```
LinearSVR(C=1.0, dual=True, epsilon=0.0, fit_intercept=True,
          intercept_scaling=1.0, loss='epsilon_insensitive', max_iter=1000,
          random_state=None, tol=0.0001, verbose=0)

SGDRegressor(alpha=0.0001, average=False, early_stopping=False, epsilon=0.1,
             eta0=0.01, fit_intercept=True, l1_ratio=0.15,
             learning_rate='invscaling', loss='squared_loss', max_iter=1000,
             n_iter_no_change=5, penalty='l2', power_t=0.25, random_state=None,
             shuffle=True, tol=0.001, validation_fraction=0.1, verbose=0,
             warm_start=False)

LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

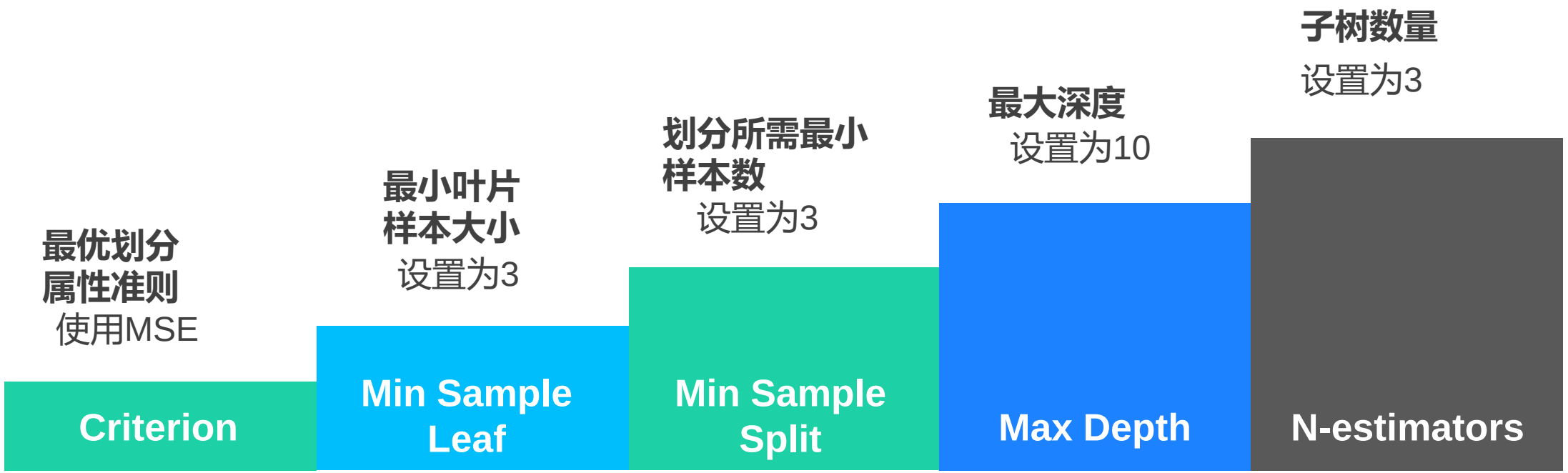
Linear SVR	SGD Regressor	Linear Regression
0.590	0.601	0.590
预测值 vs 真实值	预测值 vs 真实值	预测值 vs 真实值
5808.12 / 7500	8604.47 / 7500	9444.65 / 7500
11230.71 / 14100	13497.77 / 14100	13836.65 / 14100
7448.49 / 7499	9446.27 / 7499	11593.40 / 7499
-540.27 / 650	-1392.46 / 650	-1731.35 / 650

Random forest regressor

评分最高的模型

评分达到了0.87，
而且也没有出现奇怪的预测了！

预测值	真实值
7012.62	7500
14820.82	1400
6677.04	7499
971.06	650



最后

结论与反思

总结反思

结论

引擎功率越强，二手车价值越高
车龄越高，二手车价值普遍上越低
二手车越稀有（市场供给越少），价值越高

拥有一定程度车龄的车可能会成为具有收藏价值的古董车

燃料种类与变速箱种类也是二手车定价的主要指标

通过Python可以非常方便的建立回归模型

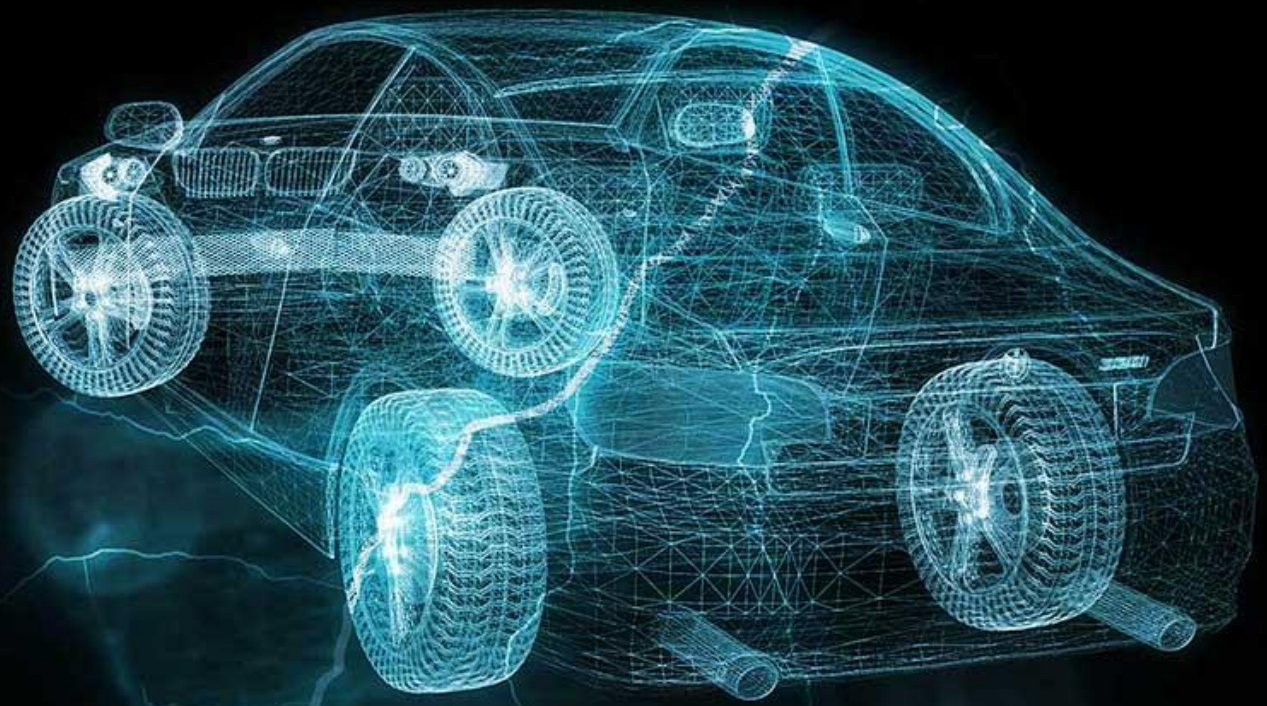
不足之处

在数据的处理上还有许多进步的空间

理论上可以构建模型由车款推导出如变速箱等缺失的变量

没能挖掘出更多给我们带来惊喜的结论

回归模型预测出了不合常理的价格



谢谢

Thank You!