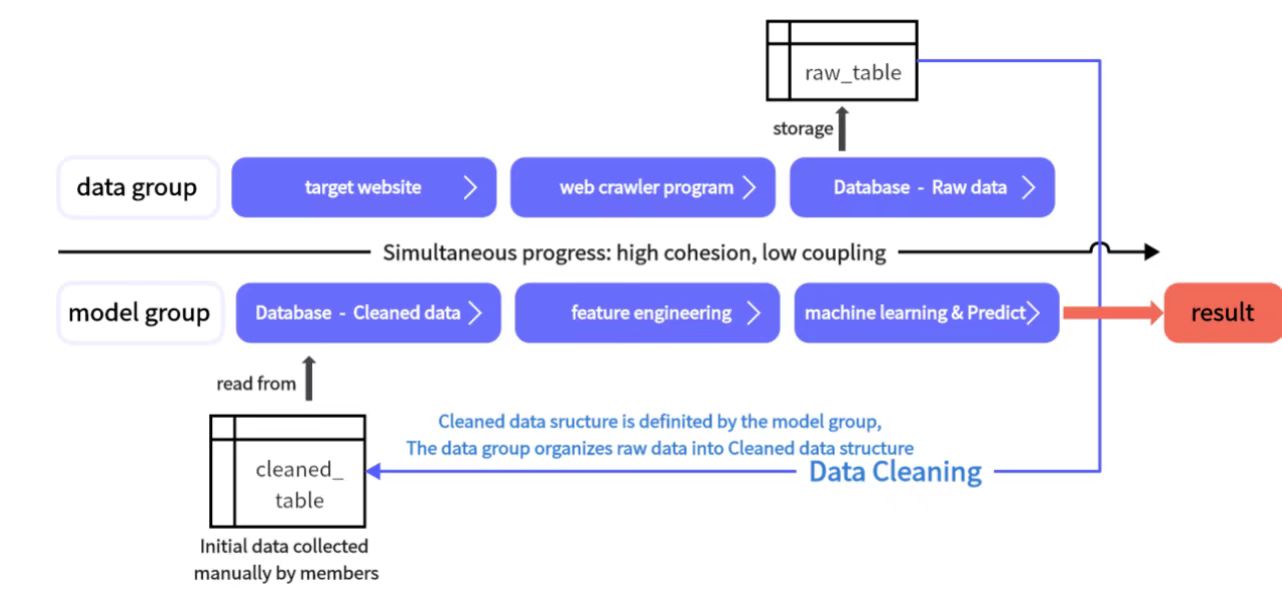
**中期报告-zny**

**2 Project plan and division of labor**



**3 Project progress and preliminary results**

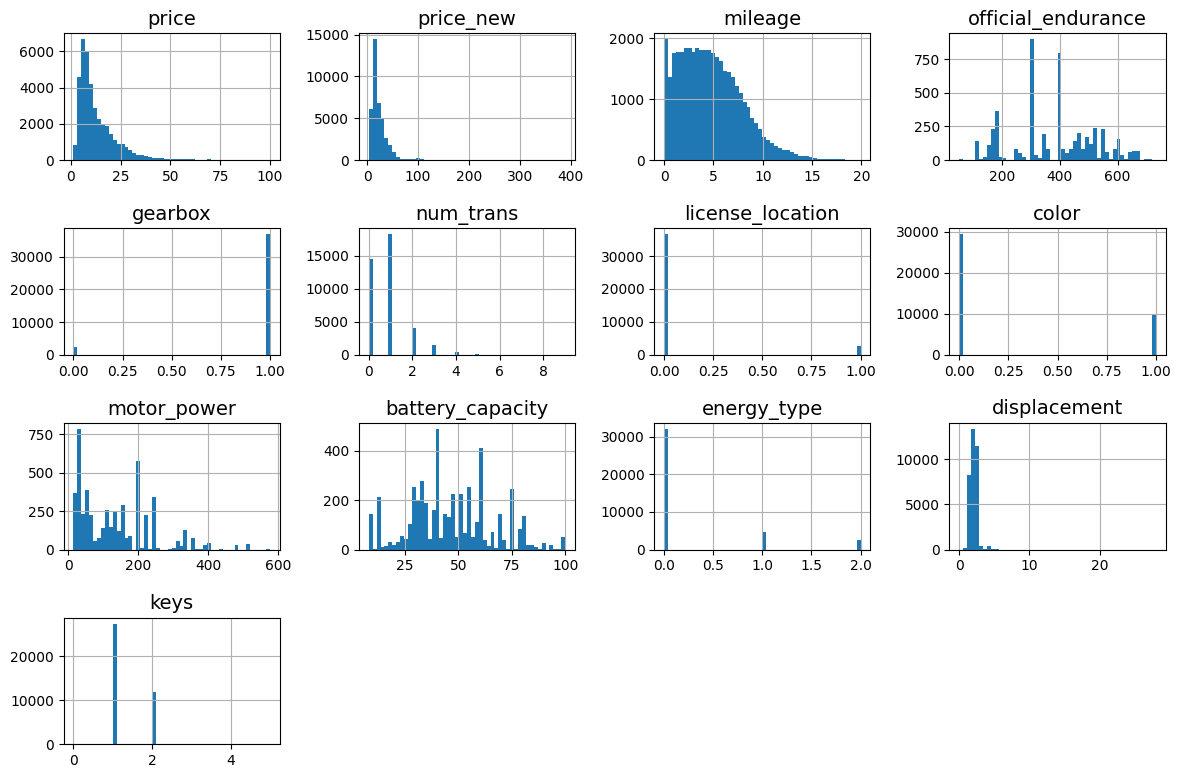
**Part2 Data Preprocessing**

The crawler program obtained a total of 39,216 pieces of data, including 17 features.

For better statistical analysis, we performed data preprocessing, including cleaning and preliminary data correlation analysis.

First, the purpose of cleaning is to convert each feature value to a numeric type (such as int64 or float64) for subsequent feature engineering operations. The specific conversion is as follows:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 列名 | 说明 | 清洗前数据类型 | 清洗后数据类型 | 备注 |
| 1 | title | 该车在网站上的标题 | object | object |  |
| 2 | price | 该车在网站上的标价 | float64 | float64 |  |
| 3 | price\_new | 该车初次购买时的含税价格 | float64 | float64 |  |
| 4 | date\_regi | 该车首次上车牌的日期 | object | date |  |
| 5 | mileage | 该车里程表所显示行驶的总里程，单位为万公里 | object | float64 |  |
| 6 | official\_endurance | 该车单次充电所能行驶的预估距离，仅限电动车 | object | float64 |  |
| 7 | standard | 该车所属的中国汽车排放标准，主要有新能源、国五、国六、国六b | Object | object |  |
| 8 | gearbox | 该车的变速箱类型，分为自动和手动，仅限燃油车 | Object | int64 | 手动0，自动1 |
| 9 | num\_trans | 该车过户的总次数 | Object | float64 |  |
| 10 | license\_location | 车牌所在地 | Object | float64 | 0：其他城市，1：一线城市（北上广深） |
| 11 | color | 该车的车身颜色 | Object | float64 | 0：黑色、白色、深灰色、银灰色，1：其他颜色 |
| 12 | motor\_power | 该车电动机的总功率，仅限电动车 | Object | float64 |  |
| 13 | battery\_capacity | 该车的电池容量，仅限电动车 | Object | float64 |  |
| 14 | battery\_type | 该车的电池类型，仅限电动车 | Object | object |  |
| 15 | energy\_type | 区分油车、电车和混合 | Object | float64 | 0：油车，1：电车，2：混合 |
| 16 | displacement | 该车的气缸的总容积，单位分为L (涡轮增压)和T (自然吸气)，T\*1.4变成L，排放越大性能越好 | Object | float64 |  |
| 17 | keys | 该车配备的钥匙数量 | Object | float64 |  |
| 18 | brand | 从title分离得到，该车品牌（清洗后数据才有的特征） | Object | object |  |

After the format is unified, draw the boxplot of each feature (each column) to view the data distribution:

Taking the price as an example, the prices of vehicles sold on the Guazi used car platform are concentrated in the range of 100,000 ± 50,000 yuan. Taking the location of the license plate as an example, more vehicle license plates sold are located in non-first-tier cities (0 value).

Further correlation analysis, through the Dataframe.corr() method, get the correlation measurement value of other features and price:

price 1.000000

price\_new 0.891951

motor\_power 0.888454

battery\_capacity 0.717950

displacement 0.660271

official\_endurance 0.520394

gearbox 0.212435

energy\_type 0.190409

keys 0.088873

num\_trans 0.079684

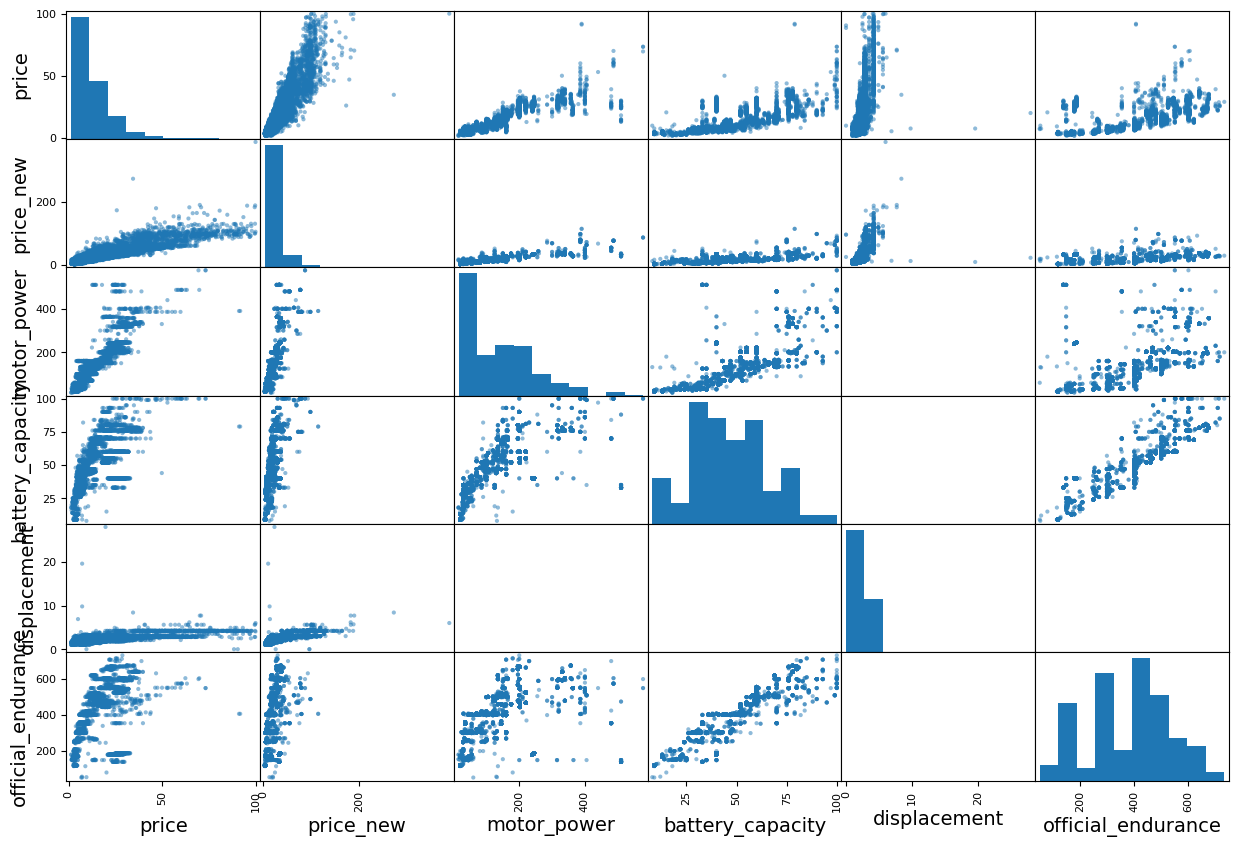
license\_location 0.074309

color -0.028627

mileage -0.084578

Name: price, dtype: float64

Plot the regression plot for the top 5 features with the highest scores:



It can be seen that price\_new and displacement have the greatest impact rate (slope) on price. It can be used as an important indicator for subsequent feature engineering.