2020年3月22日 21:18

1.加载数据

2.简略观察数据

- 使用train_data.head().append(train_data.tail())
- 使用train_data.shape查看数据的形状
- 使用describe()查看数据的一些信息(个数、平均数、方差、最值等)
- 使用info()查看数据中每一列的non-null的个数以及类型

```
SaleID            150000 non-null int64
name              150000 non-null int64
regDate           150000 non-null int64
model             149999 non-null float64
brand             150000 non-null int64
bodyType          145494 non-null float64
fuelType          141320 non-null float64
gearbox           144019 non-null float64
power             150000 non-null int64
```

4.判断数据缺失和异常

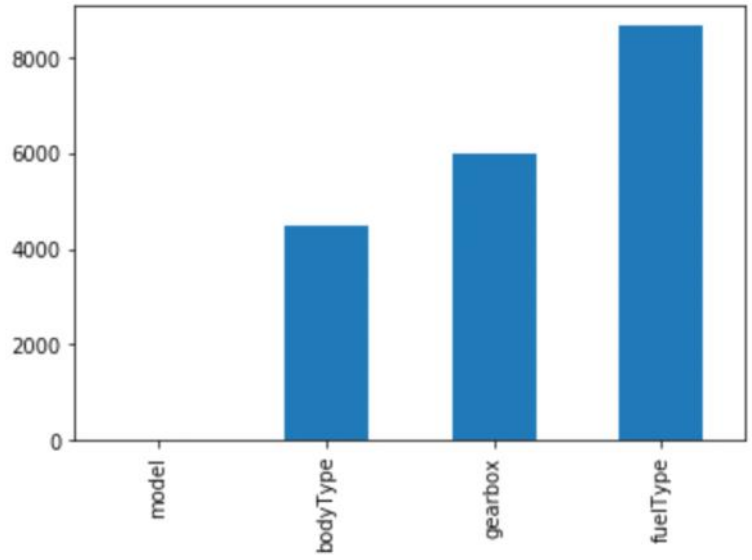
- 使用train_data.isnull().sum()查看每一列存在的null情况

```
SaleID                0
name                  0
regDate               0
model                 1
brand                 0
bodyType           4506
fuelType           8680
gearbox            5981
power                 0
kilometer             0
notRepairedDamage     0
```

- 绘制缺失图

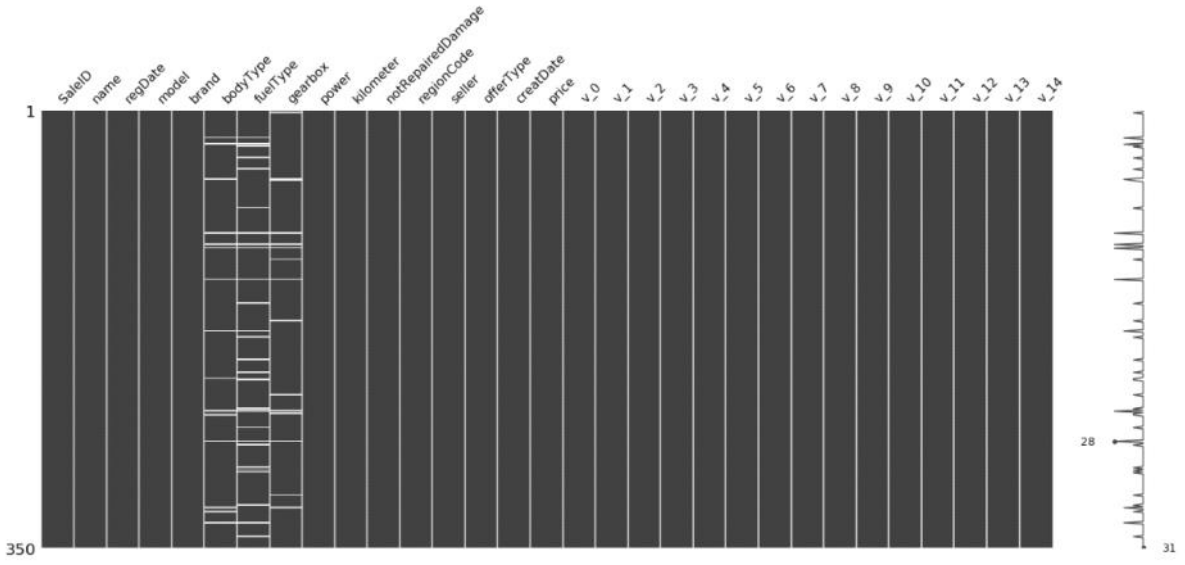
```
#缺失值绘图
missing = train_data.isnull().sum()
missing = missing[missing > 0]
missing.sort_values(inplace=True)
missing.plot.bar()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1785d0a8448>
```



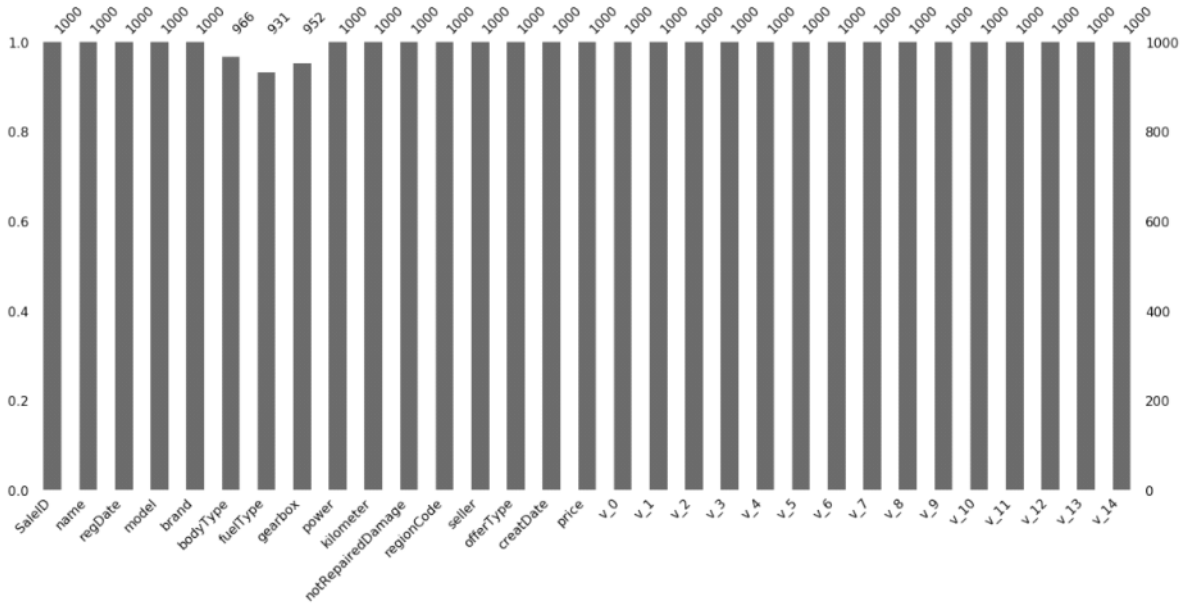

```
msno.matrix(train_data.sample(350))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1785f071388>
```



```
msno.bar(train_data.sample(1000))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x1785f174988>
```



- 异常值检测

  使用info()查看每一类中的数据类型，对非数值型进行处理，将一些没有意义的换成nan.

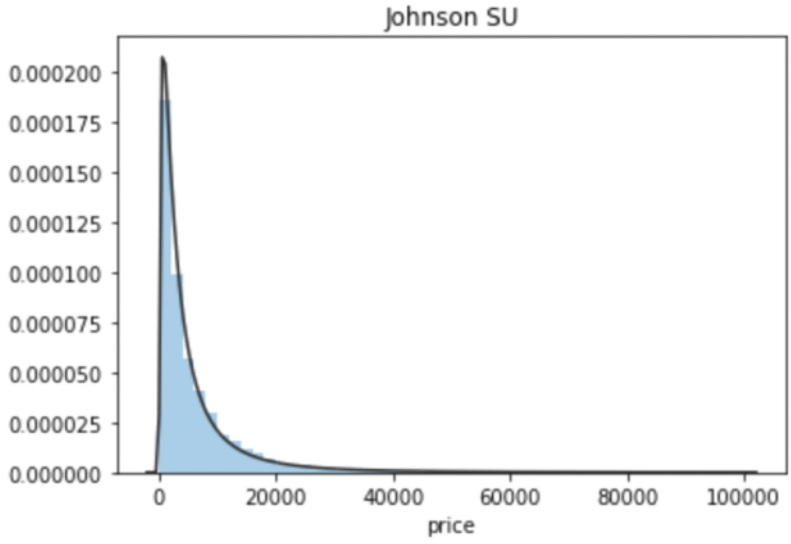
- 对没有意义的数据进行删除

5.了解数据的分布

- Johnson su分布

```
y = train_data['price']
plt.figure(1)
plt.title('Johnson SU')
sns.distplot(y,kde=False,fit=st.johnsonsu)
```

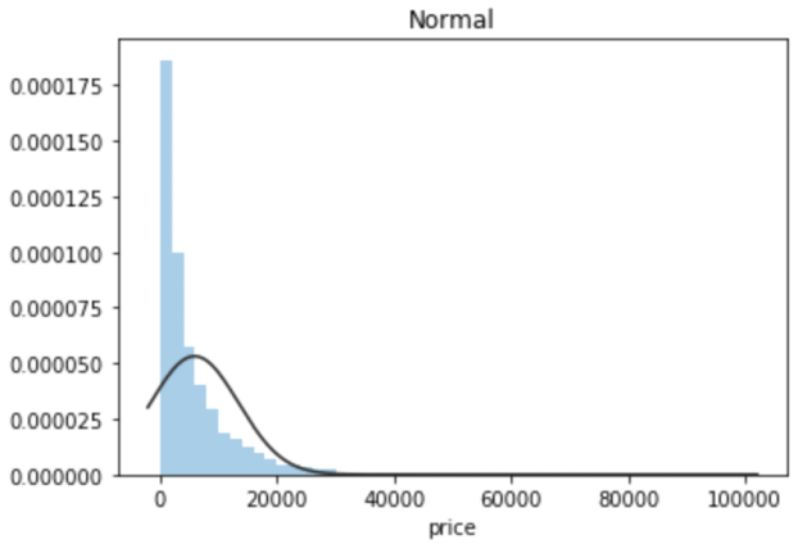
```
<matplotlib.axes._subplots.AxesSubplot at 0x1786569bf08>
```



- 正态分布

```
plt.figure(2)
plt.title('Normal')
sns.distplot(y,kde=False,fit=st.norm)
```

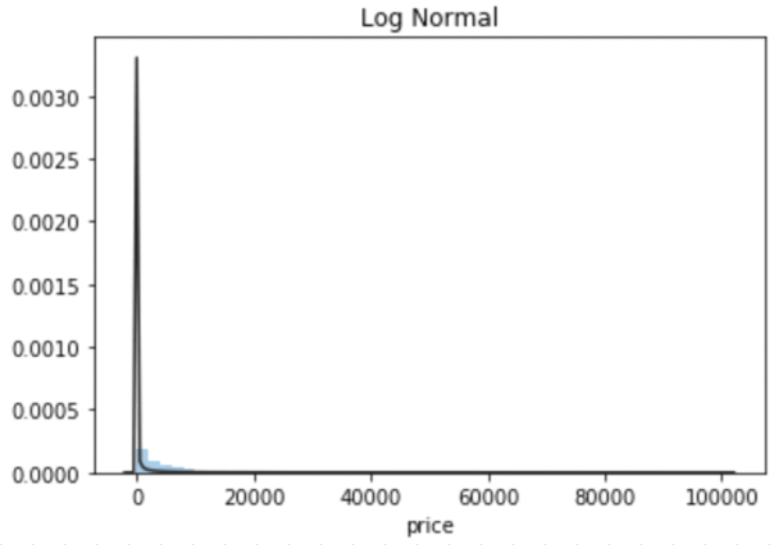
```
<matplotlib.axes._subplots.AxesSubplot at 0x178656c73c8>
```



★ Log normal分布

```
plt.figure(3)
plt.title('Log Normal')
sns.distplot(y,kde=False,fit=st.lognorm)
```

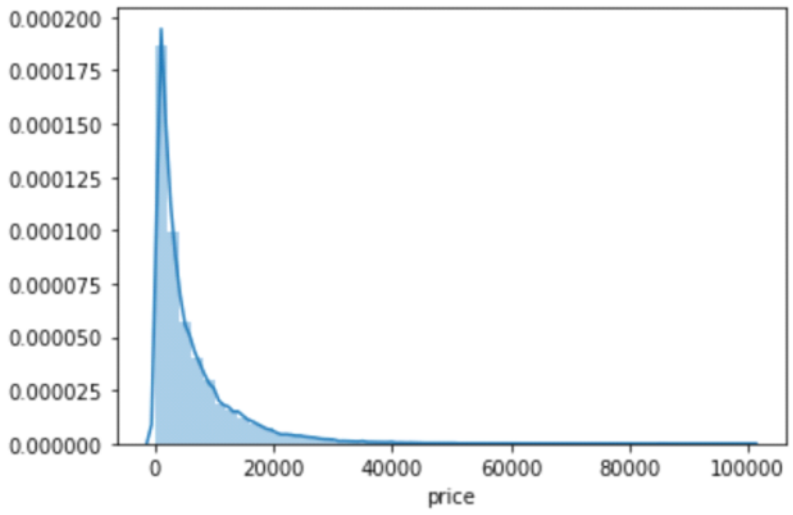
```
<matplotlib.axes._subplots.AxesSubplot at 0x1786586c188>
```



★ 查看偏度和峰度(skewness and kurtosis)

```
#计算skewness偏度、kurtosis峰度
sns.distplot(train_data['price'])
print('skewness:%f' % train_data['price'].skew())
print('skewness:%f' % train_data['price'].kurt())
```

```
skewness:3.346487
skewness:18.995183
```



6.对特征进行分类

- 手动对数据类型进行分类标记
- 对特征进行nunique分布

```
# 特征nunique分布
for cat_fea in categorical_features:
    print(cat_fea + "的特征分布如下：")
    print("{}特征有个{}不同的值".format(cat_fea, Train_data[cat_fea].nunique()))
    print(Train_data[cat_fea].value_counts())
```

```
name的特征分布如下：
name特征有99662个不同的值
708         282
387         282
55          280
1541        263
203         233
           ...
5074          1
7123          1
11221         1
13270         1
174485        1
Name: name, Length: 99662, dtype: int64
```

7.数字特征分析