**datawhale-天池二手车价格预测-Task2 数据分析**

**一、缺失值查看，可视化及处理**

## 1) 查看每列的存在nan情况

Train\_data.isnull().sum()

Test\_data.isnull().sum()

# nan可视化

missing = Train\_data.isnull().sum()

missing = missing[missing > 0]

missing.sort\_values(inplace=True)

missing.plot.bar()

注意：这里的缺失值包括nan，'-'，等很多情况。比如通过查看数据信息我们可以知道只有 notRepairedDamage这一列类型为object ，需要我们去查看这一列有哪些类型的数据

Train\_data['notRepairedDamage'].value\_counts()

Train\_data['notRepairedDamage'].replace('-', np.nan, inplace=True)

Train\_data['notRepairedDamage'].value\_counts()

Test\_data['notRepairedDamage'].value\_counts()

Test\_data['notRepairedDamage'].replace('-', np.nan, inplace=True)

以下两个类别特征严重倾斜，一般不会对预测有什么帮助，故这边先删掉，当然你也可以继续挖掘，但是一般意义不大

Train\_data["seller"].value\_counts()

Train\_data["offerType"].value\_counts()

del Train\_data["seller"]

del Train\_data["offerType"]

del Test\_data["seller"]

del Test\_data["offerType"]

二、了解预测值的分布

johnson SU distribution学习： <https://blogs.sas.com/content/iml/2020/01/27/johnson-su-distribution.html>

<https://zhuanlan.zhihu.com/p/26869997>

## 1) 总体分布概况（无界约翰逊分布等）

#了解我们最关心的价格的分布情况

import scipy.stats as st

y = Train\_data['price']

plt.figure(1); plt.title('Johnson SU')

sns.distplot(y, kde=False, fit=st.johnsonsu)

plt.figure(2); plt.title('Normal')

sns.distplot(y, kde=False, fit=st.norm)

plt.figure(3); plt.title('Log Normal')

sns.distplot(y, kde=False, fit=st.lognorm)

#结果可以看出price的价格是左偏，不符合正态分布的

#查看price的偏度和峰度

sns.distplot(Train\_data['price']);

print("Skewness: %f" % Train\_data['price'].skew())

print("Kurtosis: %f" % Train\_data['price'].kurt())

思考：如何判断数据的转换程度？如何判断最佳拟合？

#查看每一列的数据分布情况

sns.distplot(Train\_data.skew(),color='blue',axlabel ='Skewness')

sns.distplot(Train\_data.kurt(),color='orange',axlabel ='Kurtness')

数据分布的指标：偏度和峰度

偏度：

Definition:是描述数据分布形态的统计量，其描述的是某总体取值分布的**对称性**，简单来说就是数据的不对称程度。。

偏度是三阶中心距计算出来的。

（1）Skewness = 0 ，分布形态与正态分布偏度相同。

（2）Skewness > 0 ，正偏差数值较大，为正偏或右偏。长尾巴拖在右边，数据右端有较多的极端值。

（3）Skewness < 0 ，负偏差数值较大，为负偏或左偏。长尾巴拖在左边，数据左端有较多的极端值。

（4）数值的绝对值越大，表明数据分布越不对称，偏斜程度大。

计算公式：

Skewness=E[((x-E(x))/(\sqrt{D(x)}))^3]

| Skewness| 越大，分布形态偏移程度越大。

峰度：

是描述某变量所有取值分布形态陡缓程度的统计量，简单来说就是数据分布顶的**尖锐程度**。

峰度是四阶标准矩计算出来的。

（1）Kurtosis=0 与正态分布的陡缓程度相同。

（2）Kurtosis>0 比正态分布的高峰更加陡峭——尖顶峰

（3）Kurtosis<0 比正态分布的高峰来得平台——平顶峰

计算公式：

Kurtosis=E[ ( (x-E(x))/ (\sqrt(D(x))) )^4 ]-3