1. 赛题理解
   1. 背景、目标、数据、评估指标、使用什么模型解决什么类型的问题
2. 数据探索
   1. 数据预览

data.info()、data.describe()

* 1. 变量类别（连续型、类别型）
  2. 变量分析
     1. **单变量分析**
        1. 连续型：中心分布趋势和range。

|  |  |
| --- | --- |
| 中心趋势 | 离散化程度 |
| 均值 | Range |
| 中位数 | 分位数 |
| 众数 | IQR |
| 最大值 | 方差var |
| 最小值 | 标准差std |
|  | 偏度 |
|  | 峰度 |

可视化：直方图、箱形图

for i in range(变量个数): # 对每一个变量做单变量分析

plt.subplot(几行, 几列, i + 1) # 子图

sns.boxplot(train.iloc[:, i], orient=’v’, width=0.5)

plt.ylabel(train.columns.tolist()[i])

plt.show()

* + - 1. 类别型：频次、频率。

可视化：柱形图

* + 1. **双变量分析**
       1. 连续型 vs. 连续型

相关性分析：0不相关、强相关

可视化：散点图

* + - 1. 类别型 vs. 类别型

将类别型变量分组，计算两两之间的频次和占比

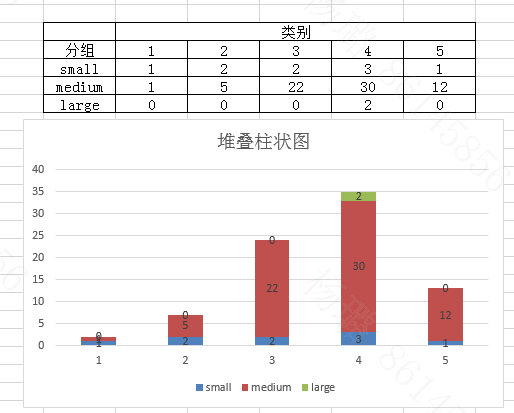
卡方检验：对多个样本率（构成比）的比较，也即对频次分布做拟合检验，找出与目标变量最相关的k个特征

from sklearn.feature\_selection import SelectKBest, chi2

chi2(X, y) # 计算X与y的拟合度

SelectKBest(chi2, k=10).fit\_transform(X, y) # 选择拟合度最高的前10个特征

可视化：堆叠柱状图



* + - 1. 类别型 vs. 连续型

可视化：小提琴图。结合了箱线图和密度图的特征。

seaborn.violinplot()

* 1. 缺失值填充
     1. 一般连续型变量：平均数、中值，类别型变量：众数
     2. 预测模型

缺失值记录作为testSet，完整的作为trainSet，将缺失的变量作为target，进行分类/回归预测

* 1. 异常值
     1. 检测
        1. 箱线图：不在-1.5IQR~1.5IQR之间的样本点

封顶方法：5%~95%范围外的样本点

与平均值mean距离3\*std的样本点

* + 1. 处理
       1. 删除
       2. 转换（如取log）
       3. 填充（类似缺失值）
       4. 区分（存在大量的异常值，则将异常值作为另外一个组。两组分别建立模型，然后将两组的输出合并）
  1. 变量转换

使用直方图、核密度估计等对单变量进行分析时，变量的分布不均，为使其类似于正态分布，需要对取值区间进行转换

* + 1. 按比例缩放、标准化（min-max scale、z-score）
    2. 对数变换，使非线性关系转换成线性关系（log）
    3. 分布倾斜

左倾斜——对变量取平方、立方、指数；

右倾斜——对变量取平方根、立方根、对数

* + 1. 把变量分组（pd.cut）
  1. 变量生成
     1. 派生变量：从现有变量创建新变量
     2. 哑变量
  2. 总结