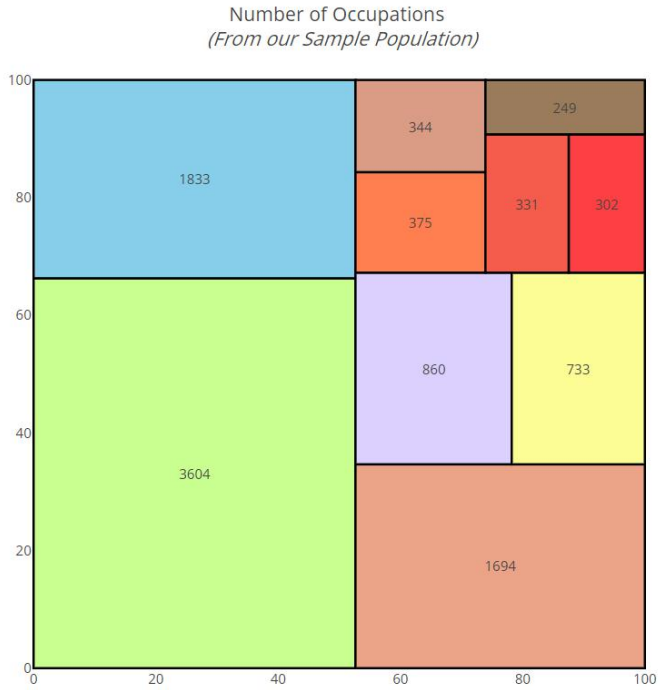
- 数据导入后，首先检查重复观测和变量缺失值，该数据集当中没有上述情况。
- 应当将 duration（电话时长）删掉，虽然该变量与存款与否高度相关，但 duration 是在电话挂断后记录的，而此时已经知道客户是否存款，则对结果预测没有意义。我们的目的是在每次电话之前预测客户是否存款。
- pdays 中超过 70%取值为-1，意义不明确。
- 定量变量的标准化对逻辑回归是重要的，可以较快达到最优；而对决策树等其他算法来说不必需。

- 关于响应变量严重偏倚问题的解决：数据偏倚往往使分类结果偏向较多观测的类，此时可以砍掉多数类的样本（欠采样），或对少数类样本进行 Bootstrap 抽样（即有放回抽样），从而形成 1：1 样本，但前者会导致信息丢失，从而欠拟合，后者会导致样本简单复制，从而过拟合。2002 年 Chawla 提出了 **SMOTE 采样**，即合成少数过采样技术。该技术是目前处理非平衡数据的常用手段，收到学术界和工业界的一致认同。该算法在模拟生成新样本的过程中采用了 KNN 技术：即对于每个少数类样本的 K 个近邻，随机挑选其中 N 个样本进行随机线性插值，从而构造出新的少数类样本，将其与原数据合并，形成新的训练集。

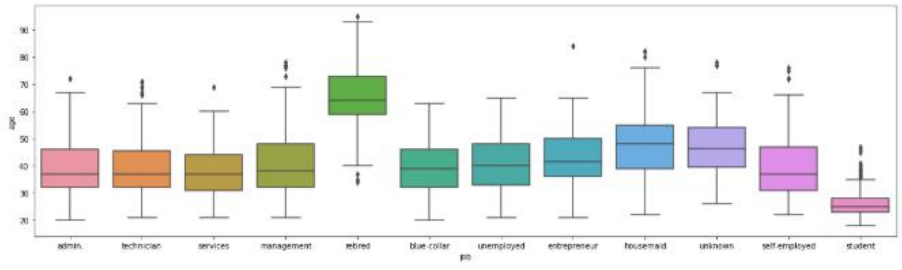
1. **定性自变量**：对每个定性变量分类计数作柱状图或饼图。二分的变为 0/1 型，多分的变为 dummy variables。

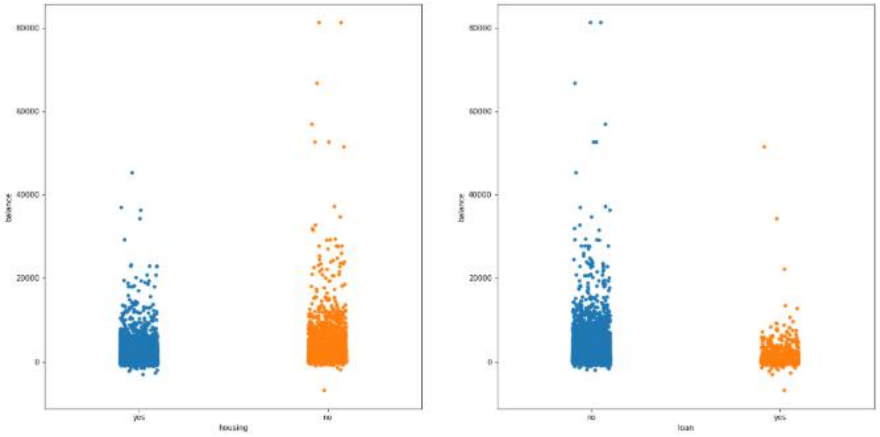


2. **定量自变量**：每个变量作直方图或箱线图，发现存在离群值的变量，可能是噪声数据，对其进一步分析：计算极端值比例，若较少，用均值等替换；若异常值或缺失值比例太大，将变量直接删去。

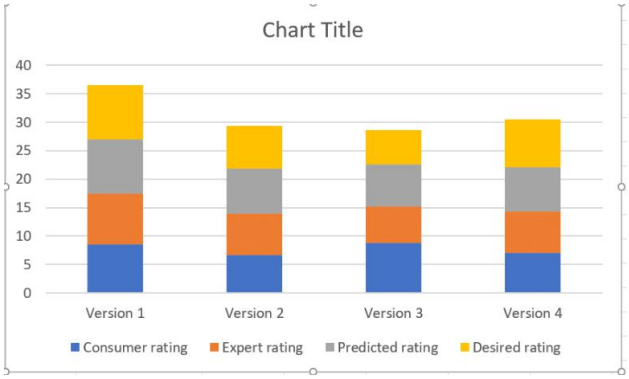
3. **自变量之间：**

①可以根据不同职业种类或受教育程度作余额的箱线图或 strip plot（得出某些职业余额更高、不同受教育水平余额差别不大的结论）；

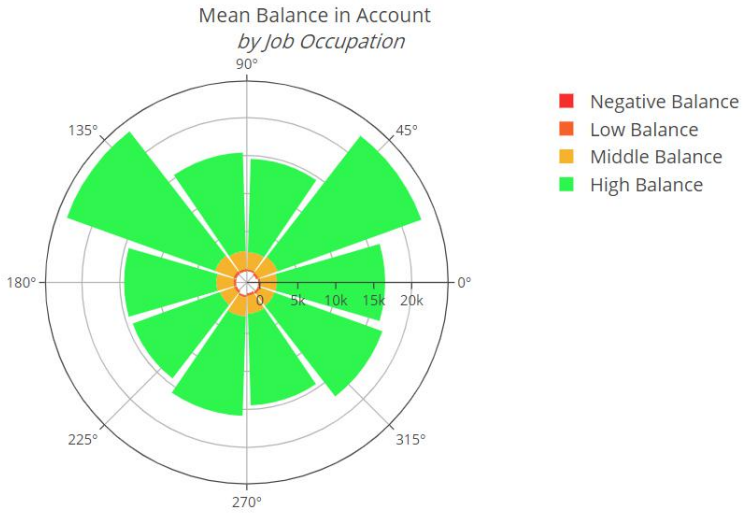


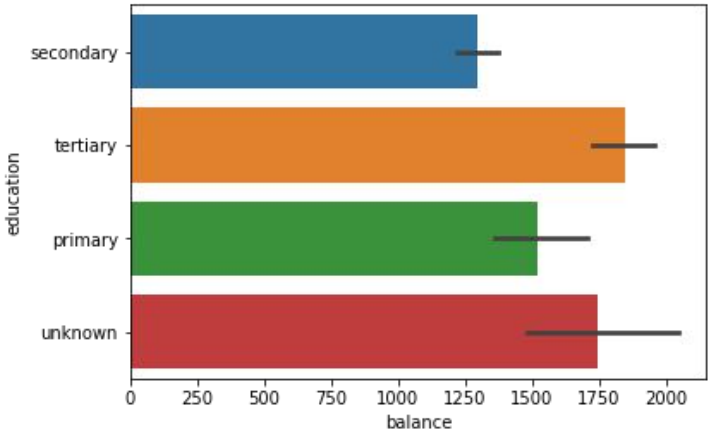


②可以作分层柱状图（stacked column chart）：比如将余额分高中低三类，然后看每种职业的人群中三类余额所占比例；

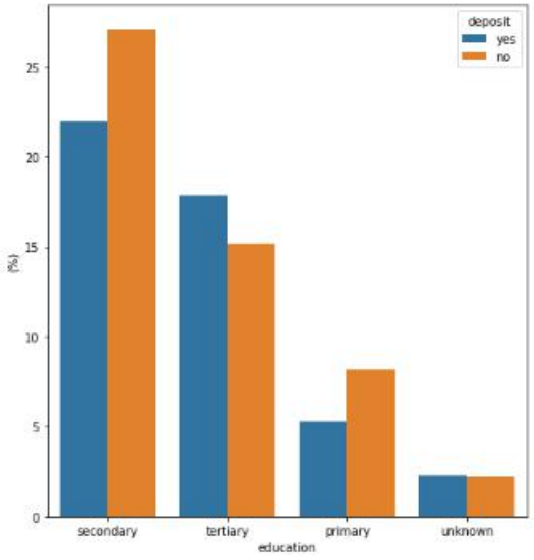


③可以在某定性变量不同种类下，分别作某定量变量的直方图、分布曲线，或比较数字特征；如绘制不同婚姻状况下余额的分布，发现已婚人士余额更多，但差别不太显著；

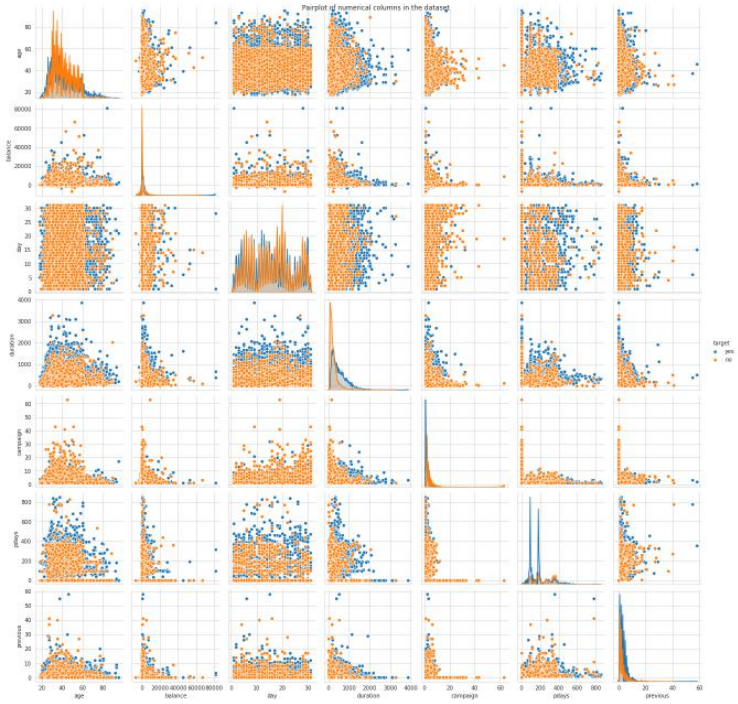


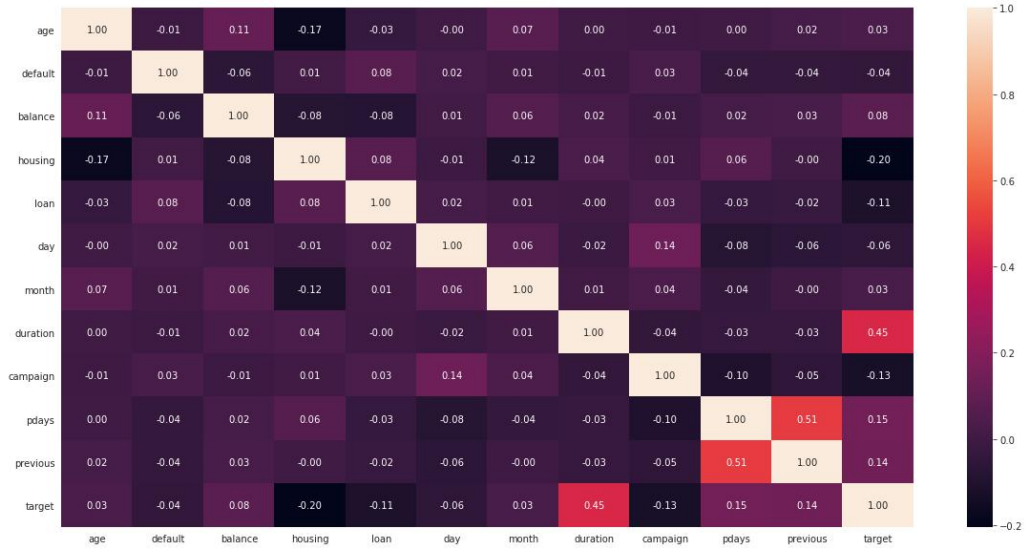


④根据某定性变量不同类别，绘制另一定性变量的计数；如根据不同年龄层，对营销的成功和失败次数计数，发现 20 岁以下和 60 岁以上人群成功率较高；观察不同月份与营销活动次数的相关性，发现营销活动主要集中在 1、4、8 月。

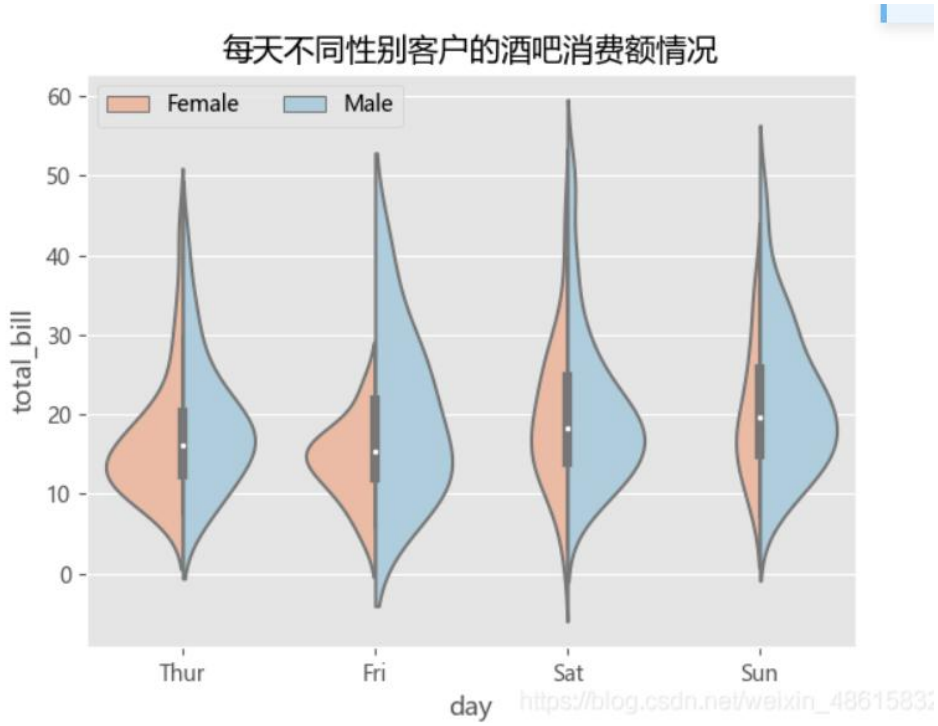


⑤散点图矩阵和相关系数矩阵（热力图）





⑥根据两个定性变量将样本划分为若干类，而后逐一观察另一定量变量的分布，或绘制小提琴图（其中核密度图的宽度代表样本量）。



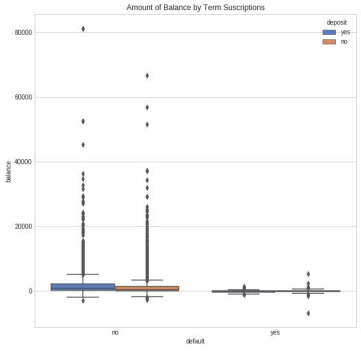
4. **响应变量：**

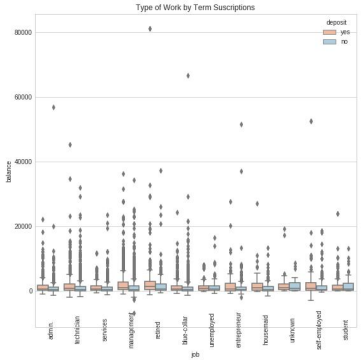
①绘制饼图，若两类比例接近、数据比较均匀，可以用准确率作为模型的评判指标。

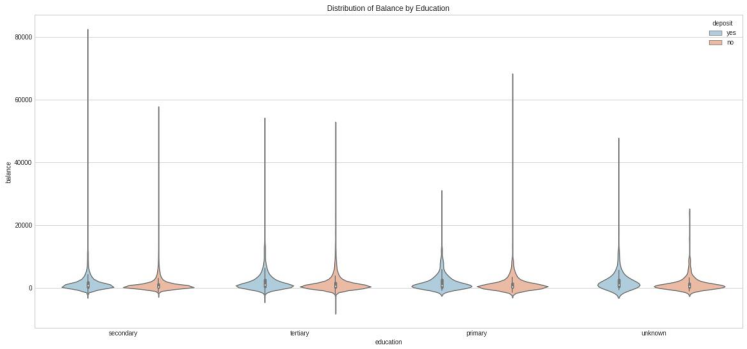
②绘制响应变量随每个定性变量分类计数的柱状图，对每个群体存款的倾向进行大致判断（如：蓝领更倾向不存款，已婚客户更倾向存款……）

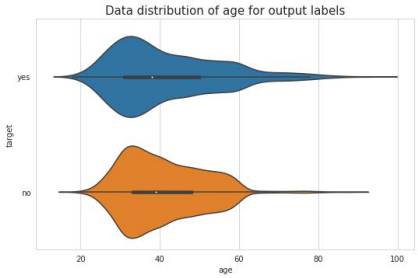
③计算响应变量取不同值时，对应定量变量的数字特征（如均值、标准差、最小最大值、四分位数），或绘制箱线图、分布图、strip plot，并对存款与不存款两群体该定量变量的特征做粗略描述（如：存款的人账户余额和年龄较高、此次营销活动中联系次数较低）

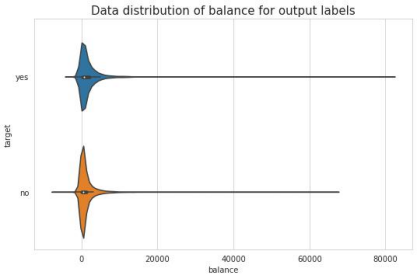
④根据某定性变量不同水平以及不同响应变量水平，绘制某定量变量的箱线图或分布图。

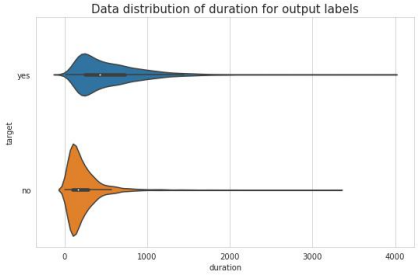

Amount of Balance by Term Suscriptions
deposit
yes
no
balance
80000
60000
40000
20000
0
no
yes
default

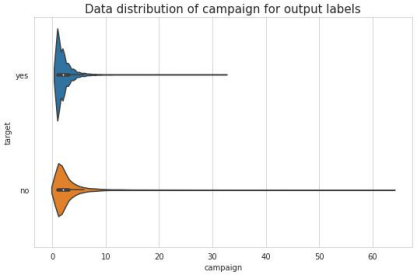

Type of Work by Term Suscriptions
deposit
yes
no
balance
80000
60000
40000
20000
0
admin.
technician
services
management
retired
blue-collar
unemployed
entrepreneur
housemaid
unknown
self-employed
student
job

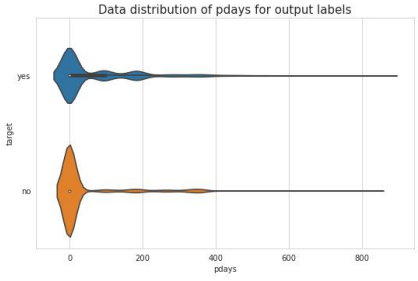

Distribution of Balance by Education
deposit
yes
no
balance
80000
60000
40000
20000
0
secondary
tertiary
primary
unknown
education

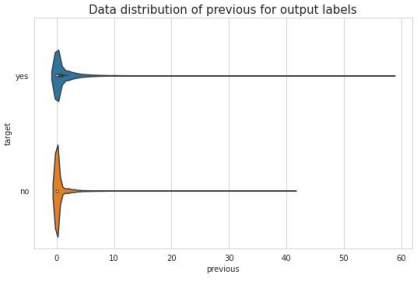

Data distribution of age for output labels
target
yes
no
20
40
60
80
100
age


Data distribution of balance for output labels
target
yes
no
0
20000
40000
60000
80000
balance


Data distribution of duration for output labels
target
yes
no
0
1000
2000
3000
4000
duration


Data distribution of campaign for output labels
target
yes
no
0
10
20
30
40
50
60
campaign


Data distribution of pdays for output labels
target
yes
no
0
200
400
600
800
pdays


Data distribution of previous for output labels
target
yes
no
0
10
20
30
40
50
60
previous