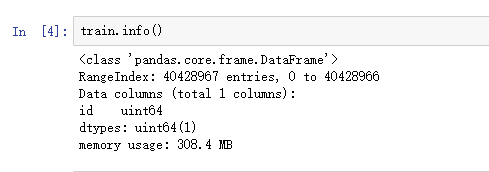
# 各列特征分析

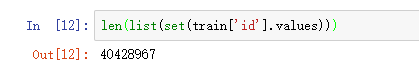
## id

### 说明: ad identifier （广告 ID）



Int64类型 , 总40428967条记录。

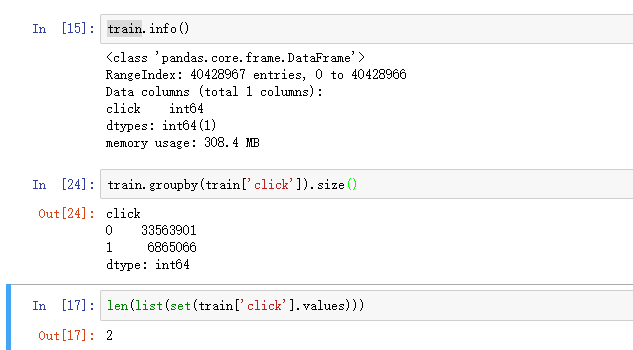
Id中去重之后，记录数不变，id中没有重复记录



### 处理方式: drop

## click

### 说明: 是否被点击，其中 0 为不被点击，1 为不被点击，此列为目标变量



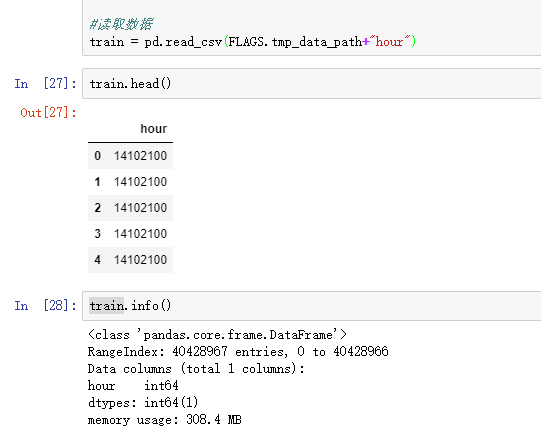
样本中0的样本数量是1样本数量的5倍

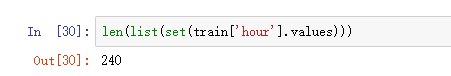
### 处理方式：

测试集默认 0 ,进行特征工程

## hour

### 说明: 广告被点击的时间戳





### 处理方式

截取 日

截取 整点时间

截取 周几

截取 小时数 从21日00点开始

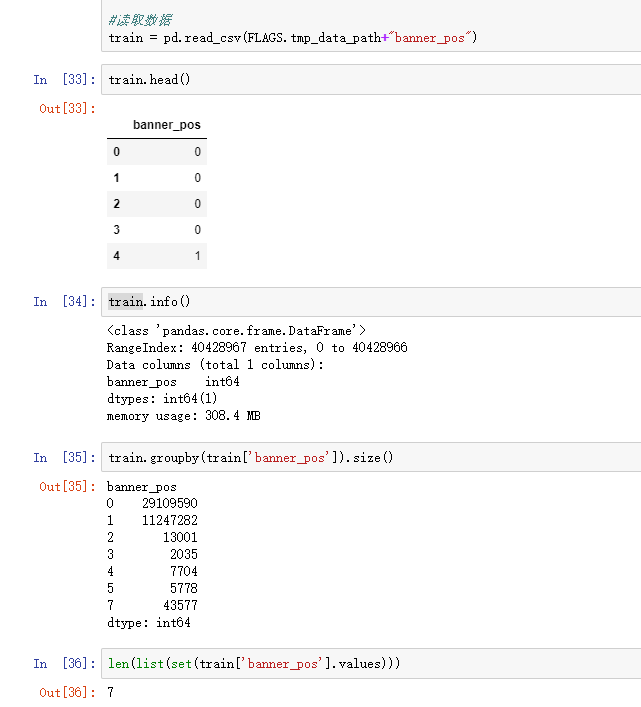
截取 上一小时几点

截取 下一小时几点

截取 是否工作日

## banner\_pos

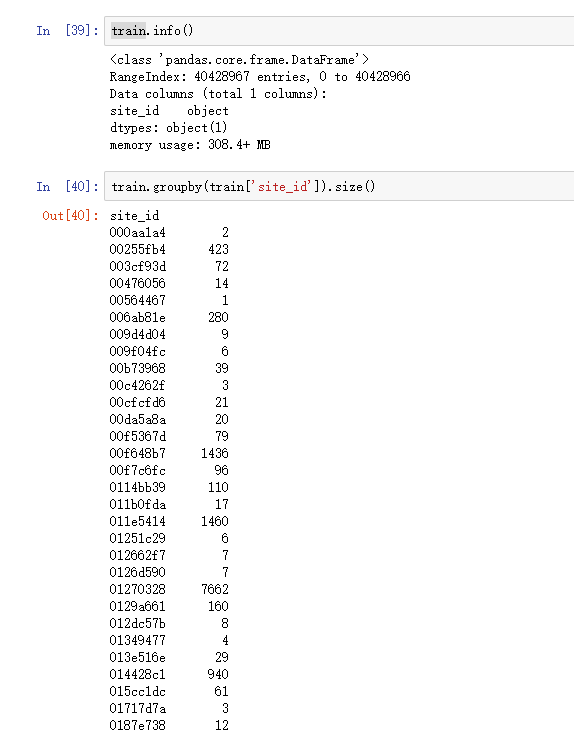
说明：广告所处位置

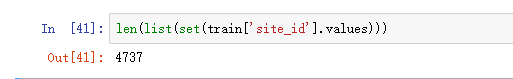


感觉无需调整，可以观察下，不同位置的投放量和对应的点击率

## site\_id

### 说明：站点 ID

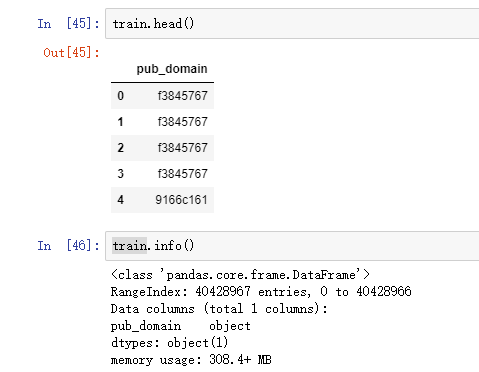


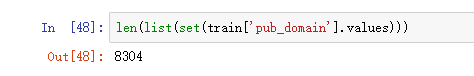


### 处理方式:有4000多个site\_id,和app\_id 一起处理

## site\_domain

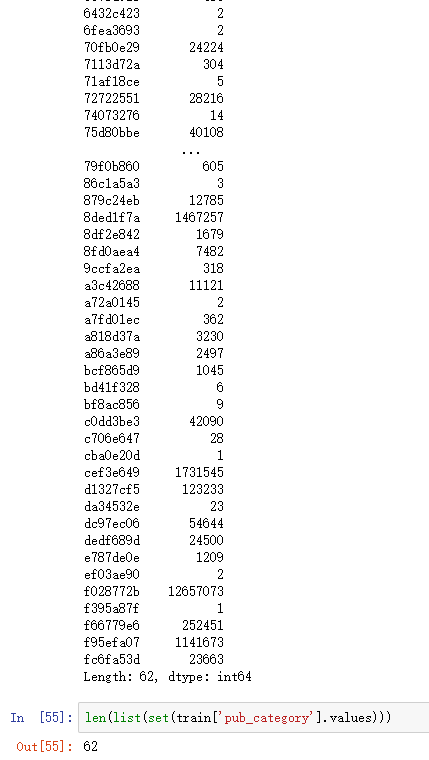
说明: 站点领域





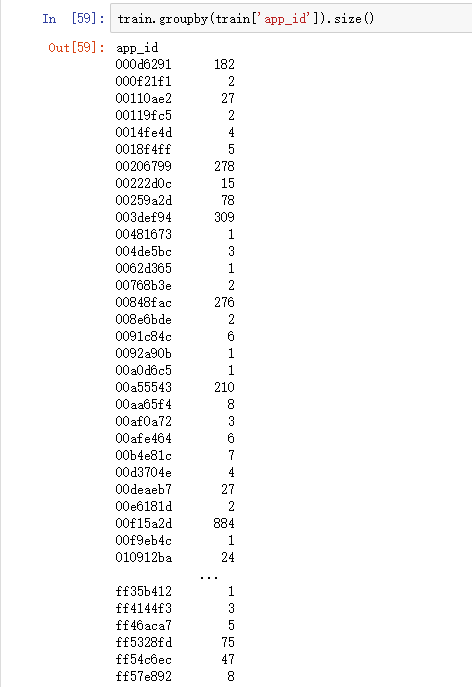
## pub\_category

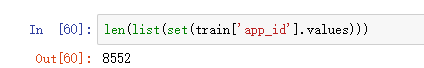
说明：站点类别



## app\_id

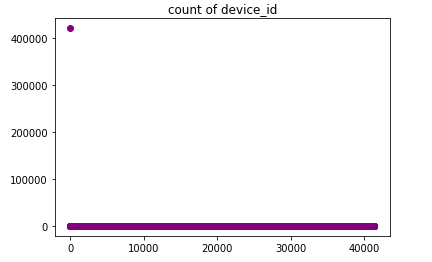
说明：移动端id

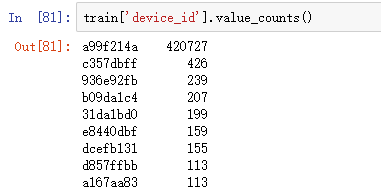




## device\_id

说明:设备id,肯定能用户登入的手机型号或者电脑型号





### 处理方式

将设备相关的信息全部串联，作为用户id，用来分析用户历史行为

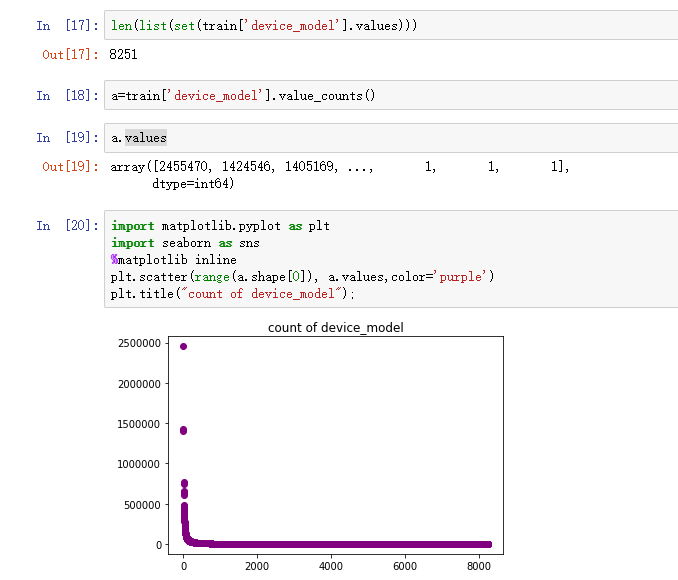
## device\_ip

说明:设备的ip地址



## device\_model

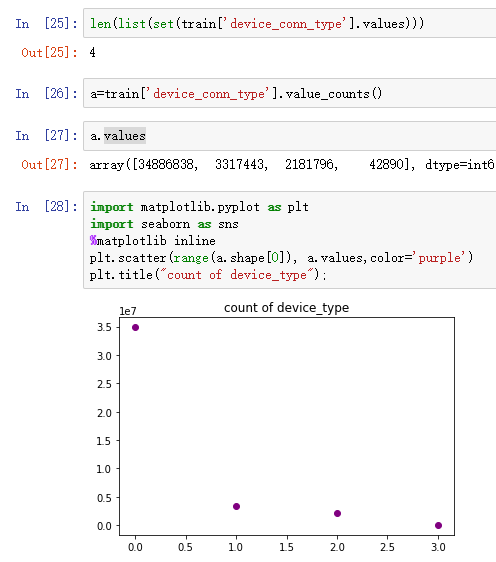
说明: 设备模式



## device\_type

## device\_conn\_type

说明:设备联网方式

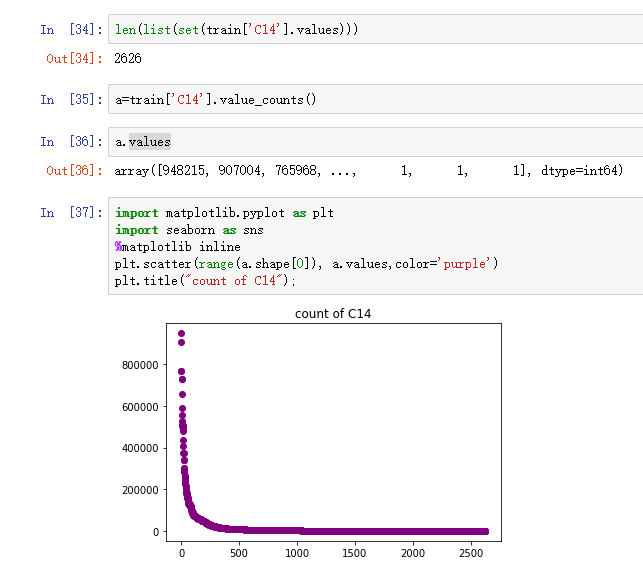


## C1

观察这些特征的关联度，考虑是否做特征筛选

## C14

说明:应该是某一字段脱敏后的id或者索引值



## C15

## C16

## C17

## C18

## C19

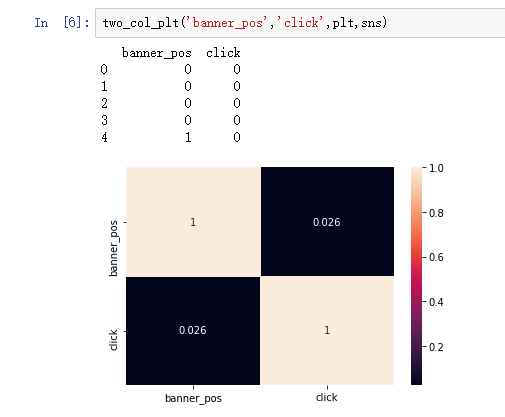
## C20

## C21

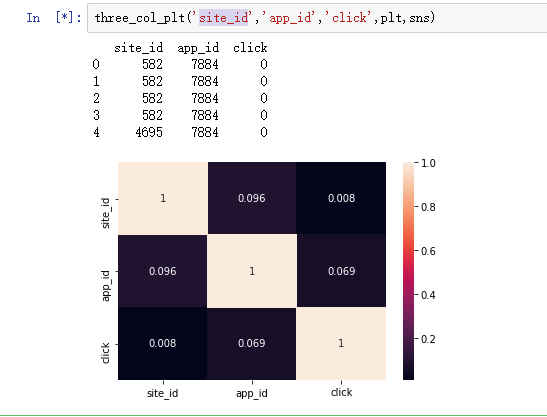
# 特征之间的联合分布

## banner\_pos + click

说明：广告所处位置应该和是否点击存在很强的关联关系



## site\_id or app\_id +click



## device\_model + device\_conn\_type

