1 淘宝内衣购买分析 ¶

目的：

- 非规则数据如何处理
- 掌握时间处理
- 如何选择维度
- seaborn与pandas操作

2 数据整理

2.1 实际工作中，我们的数据来自哪里

- 更多的根据自己产品采集数据
- 开源数据集
- 第三方数据

2.2 准备工作：

- 下载数据集
- 查看数据集
- 设定整理目标

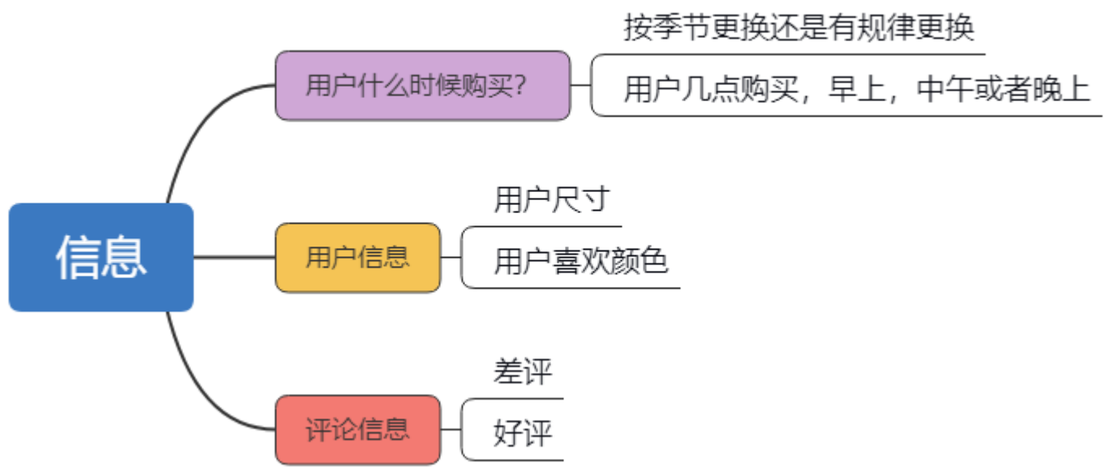
数据内容：

- 文本文件
- 主要内容：2017-04-20 13:06:04,颜色分类:肤色薄款;尺码:38/85C,不错给婆婆买的，准备再买两件

从数据得到信息：

- 时间，颜色分类，尺寸，评论
- 数据不规范需要提取
- 目标： 提取时间，类别，尺寸，评论

通过这些我们可以获取什么信息？



2.3 将数据整理成DataFrame对象

知识点：

- 文件操作
- 正则
- 列表

实现思路：

- 逐行读取文件
- 使用正则切分数据
- 将数据添加到列表中
- 创建DataFrame对象

实现如下：

In [4]:
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import re
path = r'F:\database\pandas_dir\cup_all.txt'
f = open(path, encoding='utf-8')
result = []
for line in f:
    t = re.sub(r'(颜色分类:)|(尺码:)', '', line.strip())
    t = re.split(r'[,;]', t, maxsplit=3)
    result.append(t)

df = pd.DataFrame(result, columns=['date', 'colortype','bsize', 'comment'])
```

In [ ]:
```
1
```

3 时间分析

需求：

- 什么时候备货
- 是么时候在线

知识点：

- 对时间进行处理：按照月，日，小时拆分
- 知识点：pandas时间处理，period

3.1 时间处理

- 将date转成DatetimeIndex
- 使用DatetimeIndex将其转换成月，日，小时

In [ ]:
```
#将时间列转DatetimeIndex
dindex = pd.to_datetime(df.date.values)
#设置Period为Day
df['day'] = dindex.to_period('D')
#设置Period为Month
df['month'] = dindex.to_period('M')
#设置为小时
df['hour'] = dindex.strftime('%H')
df.head()
```

3.2 时间统计

- 按照月进行统计

In [ ]:
```
#sns设置，字体1.2倍
sns.set(font_scale=1.2)
#支持中文
sns.set_style({"font.sans-serif":['simhei','Droid Sans Fallback']})
#画布大小
plt.figure(figsize=(10,4))
#时间排序
morder = sorted(df.month.unique())
#使用countplot进行统计，并按时间排序
ax = sns.countplot(df.month, order=morder)
#设置x轴标签旋转60度
_ = ax.set_xticklabels(ax.get_xticklabels(), rotation=60)
ax.set_title('评价数量/月')
ax.set_ylabel('数量')
ax.set_xlabel('月')
```

- 按照小时排序

In [ ]:
```
#通过评论信息，查看用户在线时间
ax = sns.countplot(df.hour)
#设置x轴标签旋转60度
_ = ax.set_xticklabels(ax.get_xticklabels(), rotation=60)
ax.set_title('评价数量/小时')
ax.set_ylabel('数量')
ax.set_xlabel('小时')
```

3.3 总结

- 通过月评价数量：在3月开始备货，到了45月是换机季节，多准备货源
- 通过小时评价量：用户在8点开始，就开始大量上线，一直到晚上11点，客流下降

4 用户属性分析

目的:

- 备货准备：尺寸，颜色，类型

知识点:

- pandas中的str方法与正则表达式

4.1 尺寸

- 查看数据:
- 根据ABCD简单获取尺寸

In [ ]:
```
df.head()
```

- 解决方式：使用正则去获取ABCD

In [ ]:
```
df['barsize'] = bsize.str.extract(r'([ABCD])')
_ = sns.countplot(df.barsize)
ax.set_title('大小')
ax.set_ylabel('数量')
ax.set_xlabel('size')
```

4.2 类型

类型：颜色，材质等，因为信息混到一起，我们不在做拆分

- 统计各个类型数量

In [ ]:
```
type_count = df.colortype.value_counts()
type_count
```

- 结果：几百个类型，没办法可视化?
- 过滤销量小于1000的值

In [ ]:
```
type_count =type_count[type_count>1000]
type_count
```

In [266]:
```
tcount = type_count.reset_index()
plt.figure(figsize=(20,4))
ax = sns.barplot(x='index', y='colortype', data=tcount)
_ = ax.set_xticklabels(ax.get_xticklabels(), rotation=90)
ax.set_title('类型与销量')
ax.set_ylabel('数量')
ax.set_xlabel('类型')
```

4.3 总结

- 根据大小，备货尽量选择AB，然后准备C，稍微准备点D
- 根据类型，我们可以选择大家喜欢的选择颜色，进行备货

扩展:

- 对类型与颜色再次提取，提取出更多颜色
- 对品论信息进行分类，但是品论没有对应的商品，所以无法确认商品好坏

Contents