



淘宝母婴数据分析

小组成员：刘金龙 刘晨阳 张晋哲 梁茜

汇报人：梁茜



CONTENTS

01 选题与意义

Part one

02 基本工作概述

Part tow

03 分析过程与 成果展示

Part three

04 结论与收获

Part four

The background of the slide is a light beige, textured surface. In the four corners, there are clusters of stylized green leaves. These leaves are layered and have white veins, giving them a paper-cut or origami-like appearance. The leaves in the top-left and bottom-left corners are a lighter green, while those in the top-right and bottom-right corners are a darker green. The bottom-right cluster also features some small yellow dots and a small yellow smiley face on one of the leaves.

01

选题与意义

THE PART ONE

01 选题与意义



选题意义

熟悉数据科学分析方法与流程

锻炼分工合作、代码编写、语言组织等能力

通过实践加强对于理论知识的理解



数据集来源

选取天池阿里公开数据集 Baby Goods Info Data, 包含:
(sample)sam_tianchi_mum_baby.csv 包含了消费者在
淘宝或天猫提供的900多个孩子的生日和性别。

(sample)sam_tianchi_mum_baby_trade_history.csv
该表包含淘宝会员的历史交易信息。



数据分析工具

使用python语言进行分析

01 选题与意义



数据集介绍

选取天池阿里公开数据集 Baby Goods Info Data, 包含:

(sample)sam_tianchi_mum_baby.csv 包含了消费者在淘宝或天猫提供的900多个孩子的生日和性别。

(sample)sam_tianchi_mum_baby_trade_history.csv 该表包含淘宝会员的历史交易信息。

Out[3]:

	user_id	birthday	gender
0	2757	20130311	1
1	415971	20121111	0
2	1372572	20120130	1
3	10339332	20110910	0
4	10642245	20130213	0
...
948	2020957900	20140430	0
949	2080304899	20100713	0
950	2114469016	20140416	0
951	2186831536	20140519	1
952	2254611367	20111031	0

953 rows x 3 columns


Baby.csv

Out[2]:

	user_id	auction_id	cat_id	cat1	property	buy_mount	day
0	786295544	41098319944	50014866	50022520	21458:86755362;13023209:3593274;10984217:21985...	2	20140919
1	532110457	17916191097	50011993	28	21458:11399317;1628862:3251296;21475:137325;16...	1	20131011
2	249013725	21896936223	50012461	50014815	21458:30992;1628665:92012;1628665:3233938;1628...	1	20131011
3	917056007	12515996043	50018831	50014815	21458:15841995;21956:3494076;27000458:59723383...	2	20141023
4	444069173	20487688075	50013636	50008168	21458:30992;13658074:3323064;1628665:3233941;1...	1	20141103
...
29966	57747284	35169635909	50010549	50008168	21458:125202070;22019:3228688;22019:3248884;22...	1	20140109
29967	287541325	19778523000	50007011	50008168	21458:112788583;1633959:3523439;3130834:209537...	2	20140109
29968	82915321	12766532512	50011993	28	21475:137325;1628665:3233937;1628665:29798;162...	1	20131008
29969	78259523	18309305134	50013711	50008168	21458:30992;1628665:29778;1628665:29793;163395...	1	20131008
29970	758305789	20177445814	50018860	28	21458:3602856;1628665:29784;1628665:3233941;73...	1	20131008

29971 rows x 7 columns

Baby_trade_history.csv

The background of the slide is a light beige, textured surface. In the four corners, there are clusters of stylized green leaves. These leaves are layered and have white veins, giving them a paper-cut or origami-like appearance. The leaves in the top-left and bottom-left corners are a lighter green, while those in the top-right and bottom-right corners are a darker green. The overall aesthetic is clean and modern.

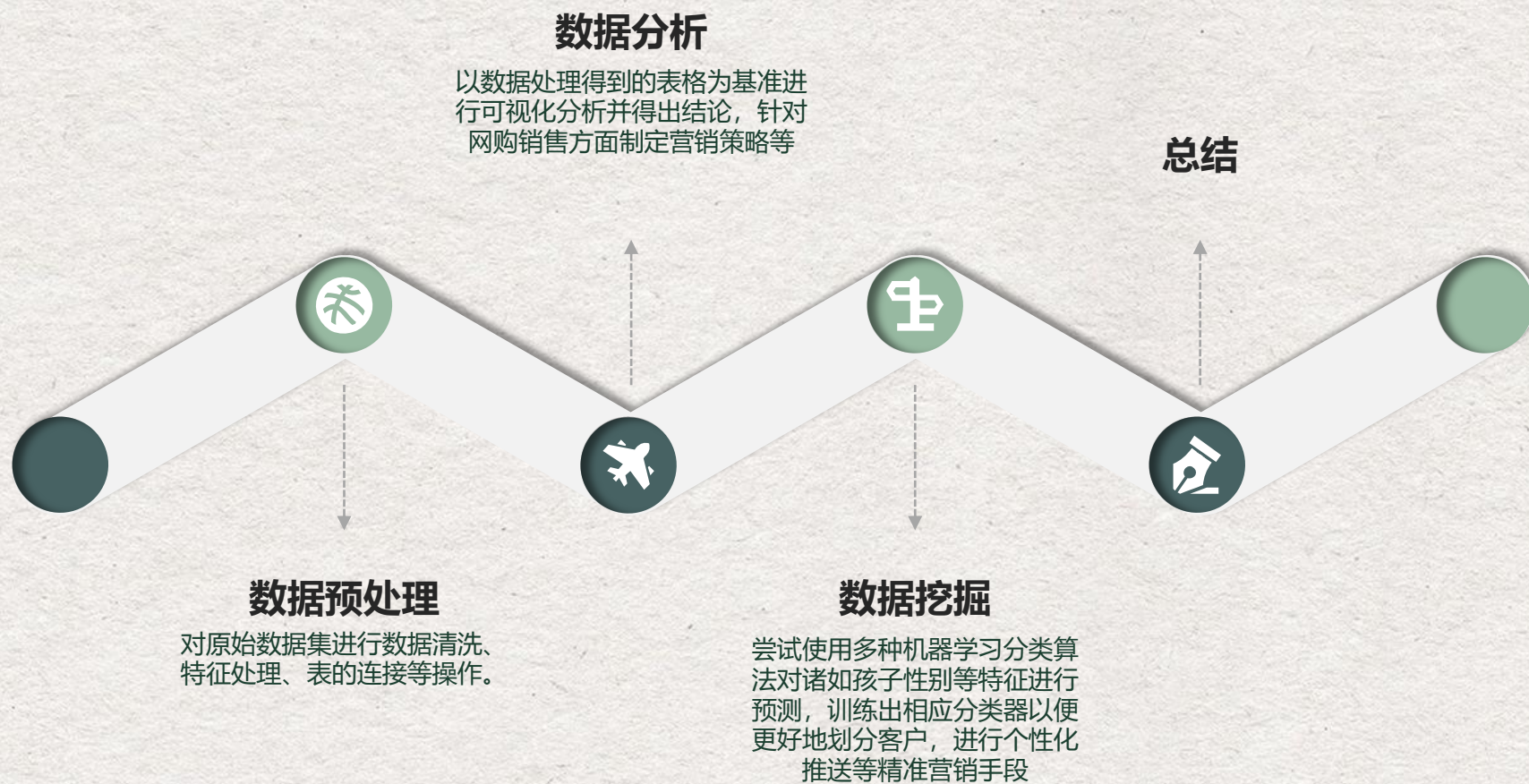
02

基本工作概述

THE PART TWO

03 数据分析成果展示

——数据分析



The background of the slide is a light beige, textured surface. In the corners, there are clusters of stylized green leaves. The leaves are layered and have white veins. Some leaves have small yellow dots on them. The leaves in the top right corner are more prominent and have a small yellow smiley face on one of them. The leaves in the bottom right corner are also layered and have small yellow dots. The leaves in the top left and bottom left corners are partially visible.

03

分析过程与 成果展示

THE PART THREE

03 过程与成果展示

——数据预处理

1

特征处理

- 更改列名 (英文→中文)
- 时间戳、出生日期、性别等转换格式

```
#数据读取
file_path1=r'(sample)sam_tianchi_mum_baby_trade_history.csv'
data1=pd.read_csv(file_path1)
# data1.head()
data1.columns = ['用户ID','订单编号','商品ID','根类别ID','特征','购买数量','时间戳']
data1=data1[['用户ID','订单编号','商品ID','根类别ID','购买数量','时间戳']]
#时间戳转换格式
data1['时间戳']=pd.to_datetime(data1['时间戳'].astype(str))
data1.head()
```

	用户ID	出生日期	性别
0	2757	2013-03-11	男
1	415971	2012-11-11	女
2	1372572	2012-01-30	男
3	10339332	2011-09-10	女
4	10642245	2013-02-13	女

Baby

	用户ID	订单编号	商品ID	根类别ID	购买数量	时间戳
0	786295544	41098319944	50014866	50022520	2	2014-09-19
1	532110457	17916191097	50011993	28	1	2013-10-11
2	249013725	21896936223	50012461	50014815	1	2013-10-11
3	917056007	12515996043	50018831	50014815	2	2014-10-23
4	444069173	20487688075	50013636	50008168	1	2014-11-03

Baby_trade_history

03 过程与成果展示

——数据预处理

2

数据清洗

- 检查处理空值
- 检查处理异常值
- 去除重复数据

```
# 1. 检查空值
# data1.info()
data1.isnull().sum()
data2.isnull().sum()
```

```
用户ID    0
出生日期  0
性别      0
dtype: int64
```

```
# 4. 去除重复数据
c_data1[c_data1.duplicated()]
c_data2[c_data2.duplicated()]
```

```
# 1) 检查时间戳是否存在异常值
print(data1[data1["时间戳"]>'2019-1-1'])

# 2) 检查购买数量是否存在异常值
print(data1["购买数量"].value_counts())
print(data1[data1["购买数量"]>1000])
# 1
#将购买数量超过1000的数据视作异常值
c_data1=data1[data1["购买数量"]<=1000]

# 3) 检查出生日期是否存在异常值
# print(data1["时间戳"].min())
#将出生日期早于2000. 07. 02与晚于2018. 12. 31的数据视为异常值
print(data2[(data2["出生日期"]<'20000702')|(data2["出生日期"]>'20181231')])
c_data2=data2[(data2["出生日期"]<='20181231')&(data2["出生日期"]>='20000702')]

# 3) 检查性别是否存在异常值
print(data2["性别"].value_counts())
# 未知性别暂不处理
```


03 过程与成果展示

——数据预处理

2

数据清洗

- 检查处理空值
- 检查处理异常值
- 去除重复数据

```
# 1. 检查空值  
# data1.info()  
data1.isnull().sum()  
data2.isnull().sum()
```

```
用户ID    0  
出生日期    0  
性别        0  
dtype: int64
```

```
# 4. 去除重复数据  
c_data1[c_data1.duplicated()]  
c_data2[c_data2.duplicated()]
```

	user_id	birthday	gender
0	2757	20130311	1
1	415971	20121111	0
2	1372572	20120130	1
3	10339332	20110910	0
4	10642245	20130213	0
...
948	2020957900	20140430	0
949	2080304899	20100713	0
950	2114469016	20140416	0
951	2186831536	20140519	1
952	2254611367	20111031	0

953 rows × 3 columns

清洗前

	用户ID	出生日期	性别
0	2757	2013-03-11	男
1	415971	2012-11-11	女
2	1372572	2012-01-30	男
3	10339332	2011-09-10	女
4	10642245	2013-02-13	女
...
948	2020957900	2014-04-30	女
949	2080304899	2010-07-13	女
950	2114469016	2014-04-16	女
951	2186831536	2014-05-19	男
952	2254611367	2011-10-31	女

952 rows × 3 columns

清洗后

03 过程与成果展示

——数据预处理

3

表连接与特征构造

- 连接表数据清洗
- 构造孩子年龄属性（时间戳-出生日期）

```
# 表按用户ID连接
df = pd.merge(df_left, df_right)
df.columns=['用户ID', '订单编号', '商品ID', '根类别ID', '购买数量', '时间戳', '孩子出生日期', '孩子性别', '孩子年龄']
df['时间戳']=pd.to_datetime(df['时间戳'].astype(str))
df['孩子出生日期']=pd.to_datetime(df['孩子出生日期'].astype(str))
df['孩子年龄']=round((df['时间戳']-df['孩子出生日期']).dt.days/365)

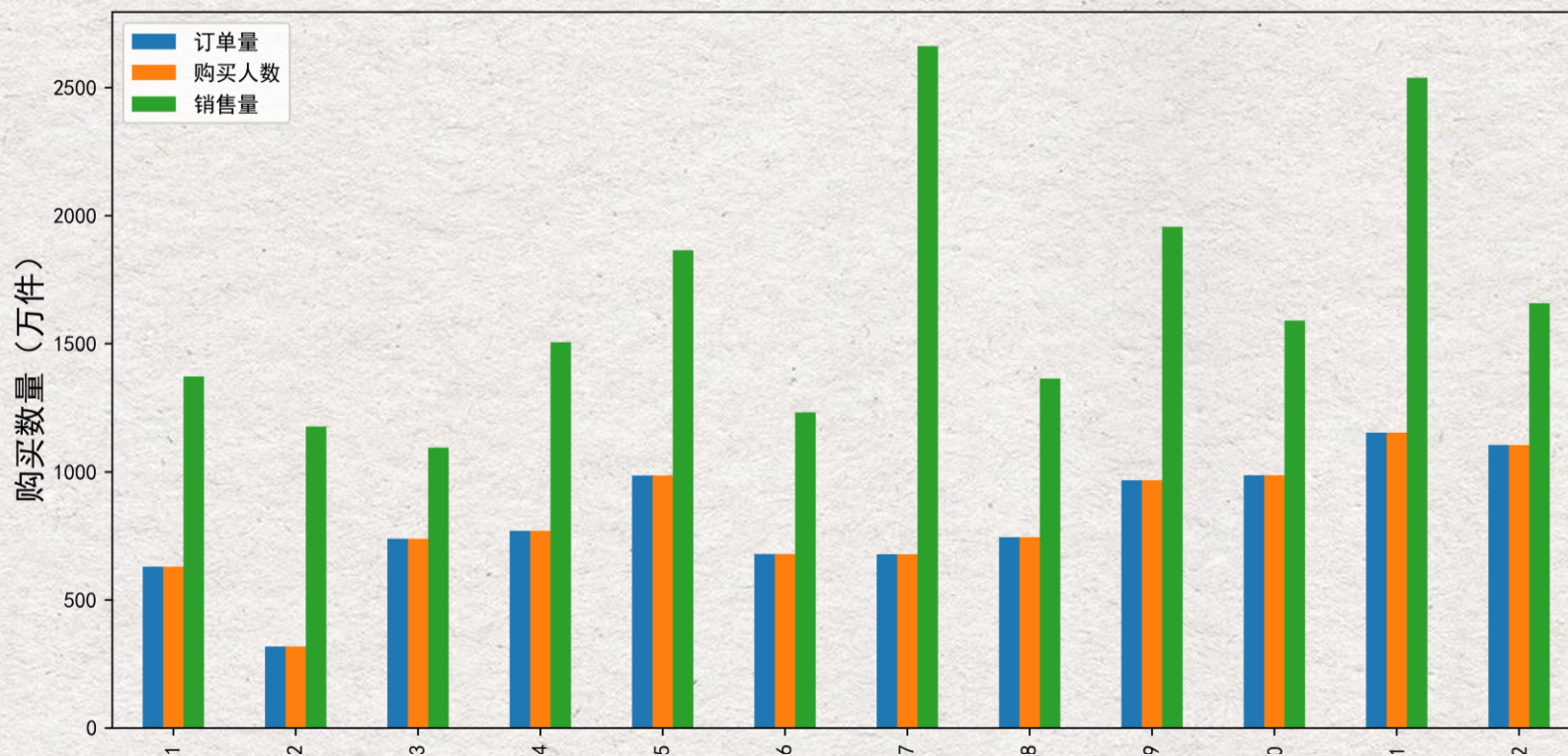
# 一、数据清洗
# 去掉未知孩子性别的数据
df=df[df['孩子性别']!='未知']
# 去掉孩子出生日期晚于购买时间的数据
df=df[df['孩子年龄']>=0]
```

	用户ID	订单编号	商品ID	根类别ID	购买数量	时间戳	孩子出生日期	孩子性别	孩子年龄
0	513441334	19909384116	50010557	50008168	1	2012-12-12	2011-01-05	男	1.94
1	377550424	15771663914	50015841	28	1	2012-11-23	2011-06-20	男	1.43
2	47342027	14066344263	50013636	50008168	1	2012-09-11	2010-10-08	男	1.93
3	119784861	20796936076	50140021	50008168	1	2012-11-29	2012-03-27	女	0.68
4	159129426	15198386301	50013711	50008168	2	2012-08-08	2010-08-25	女	1.96
...
950	685332320	12781785338	50018831	50014815	2	2013-06-01	2012-02-23	女	1.27
951	389326420	17164967407	50006820	28	1	2014-09-16	2013-07-17	女	1.17
952	359840716	17513925908	50013207	50008168	1	2013-03-18	2009-01-20	女	4.16
953	1372572	16915013171	50008845	28	1	2013-03-27	2012-01-30	男	1.16
954	54855720	39635136808	50018436	50014815	2	2014-09-13	2013-01-28	女	1.62

793 rows x 9 columns

03 过程与成果展示

——销售量分析

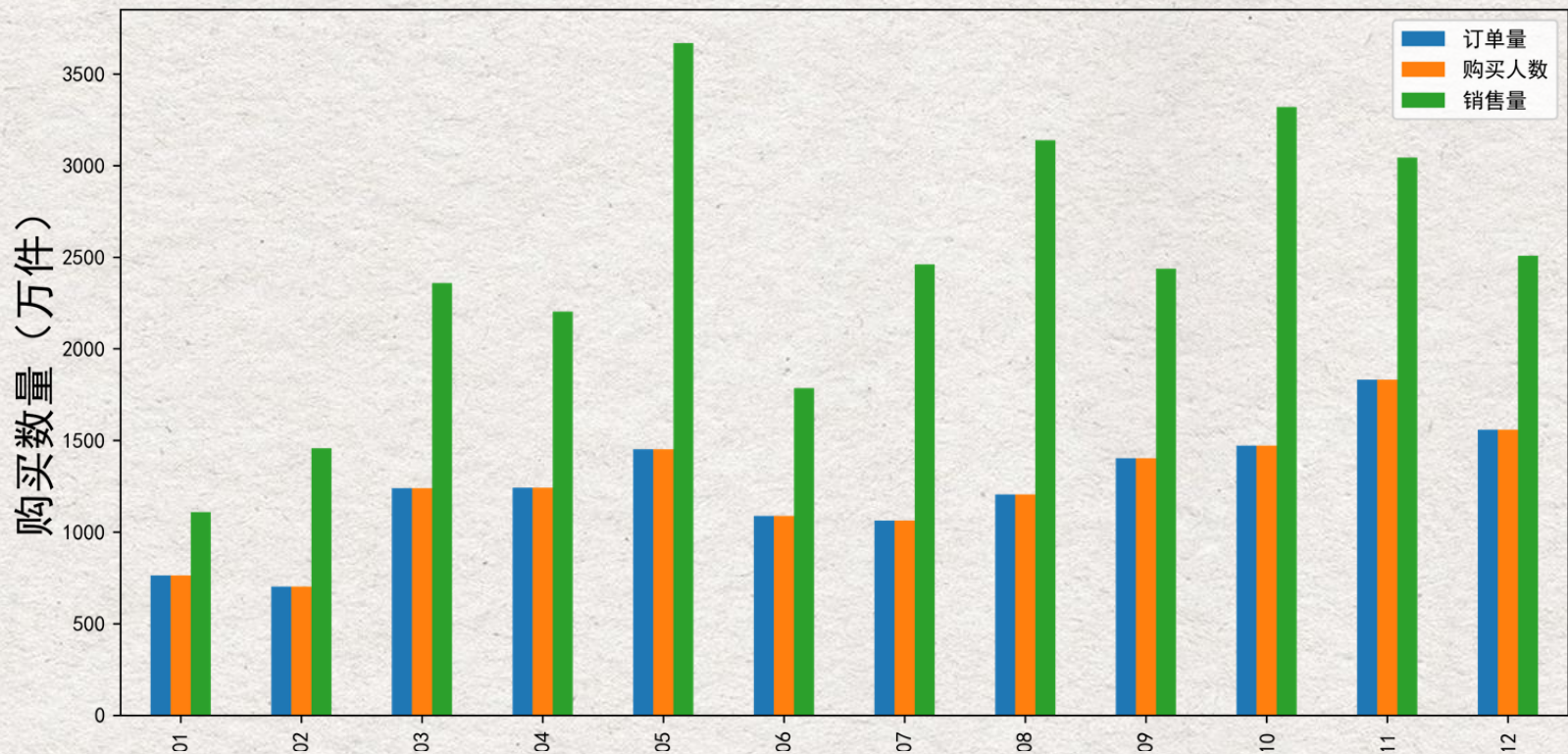


2013年销售数据图 (按月)

- 2013年月销售量数据在1000-2800范围内波动
- 3月达到最低值, 7月达到最高值, 5、7、9、12月份都有不错的销量表现
 - 就全年来看, 总体下半年销量好于上半年
- 双11导致11月订单量、购买人数、销售量均增高

03 过程与成果展示

——销售量分析

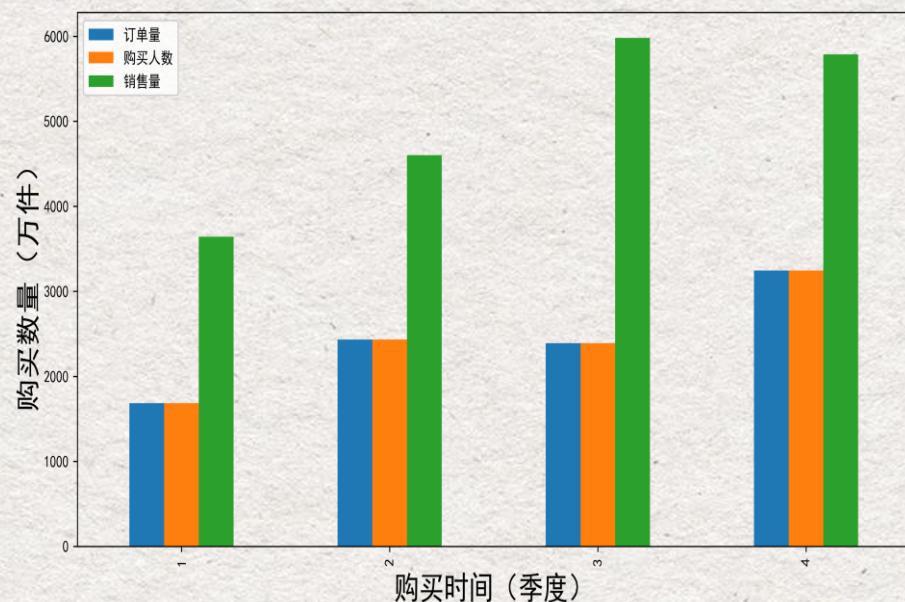


2014年销售数据图 (按月)

- 2014年月销售量数据在1000-3600范围内波动，较2013年有明显增加
- 1月达到最低值，5月达到最高值，8、10、11月份都有不错的销量表现。
- 就全年来看，上半年仅5月销售量异常偏高，下半年整体销售量良好，11月仍为订单量、购买人数top

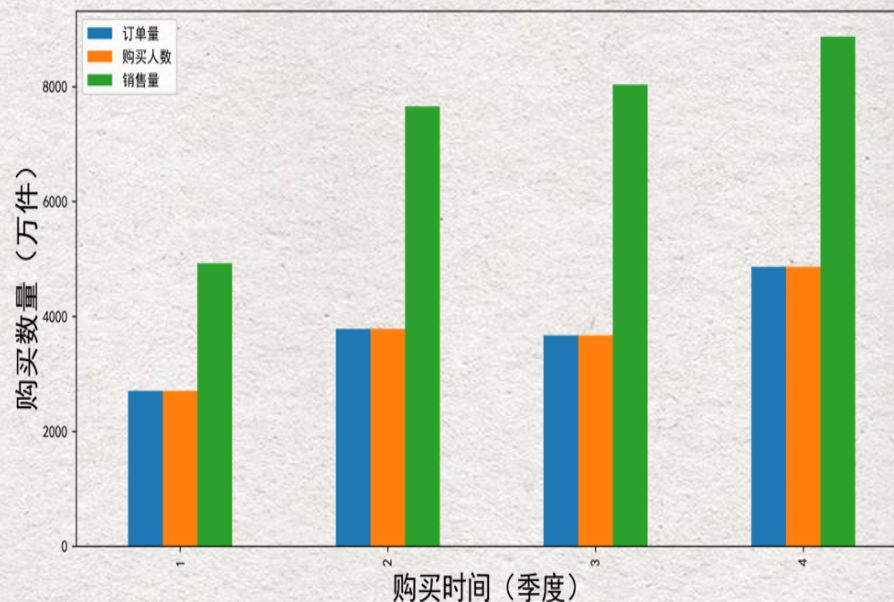
03 过程与成果展示

——销售量分析



2013年销售数据图（按季度）

- 季度销售量在3000-6000范围内，逐季增加，三季达到顶峰，四季稍有回落
- 三四季度的销售量明显优于一二季度

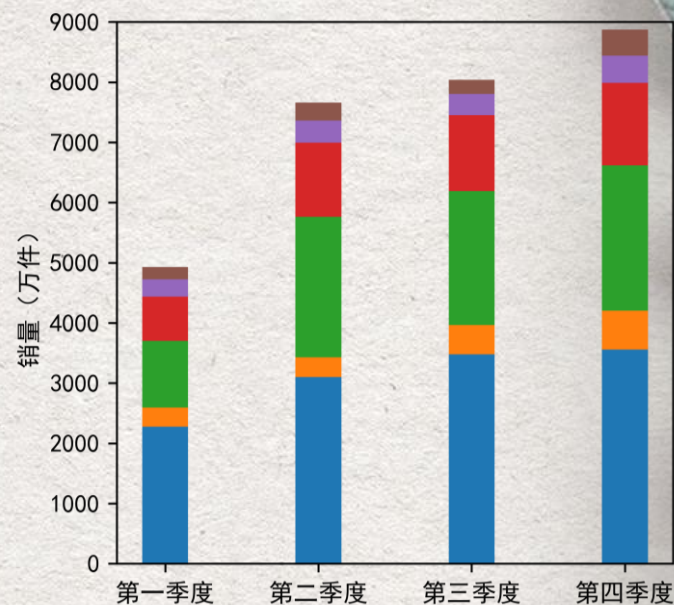
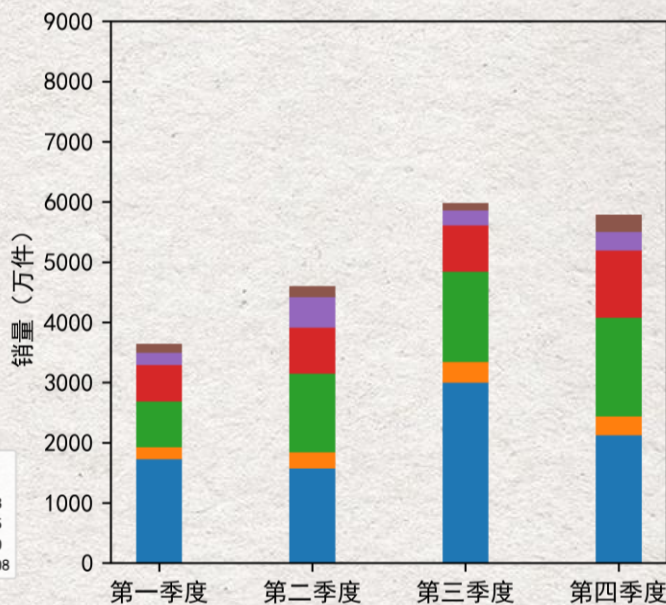
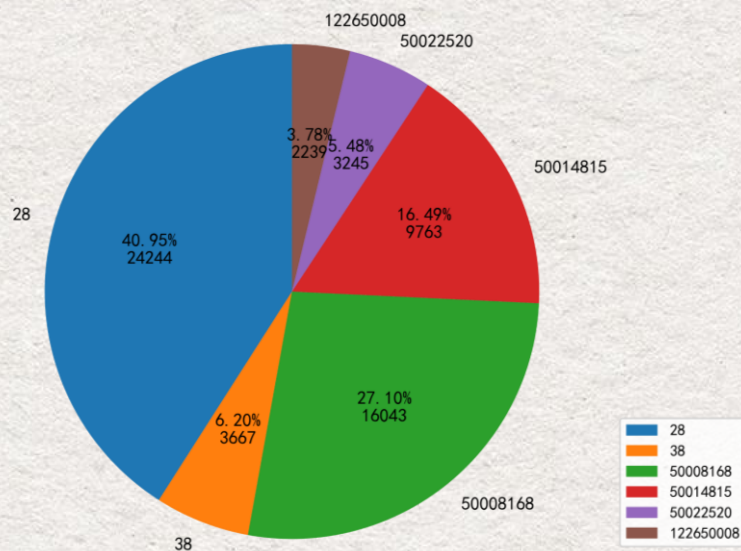


2014年销售数据图（按季度）

- 季度销售量在4000-9000范围内，逐季增加，四季达到顶峰
- 与2013年相比，每季销售量都存在明显上升，总体上仍保持三四季度销量多于一二季度的趋势

03 数据分析成果展示

——商品类别分析



不同商品根类别市场占有率图

- 数据集中共存在六种商品类别，仅有编号信息
- 28、50008168、50014815三类商品占据主要市场份额，而38、122650008、50022520只占到了15.46%

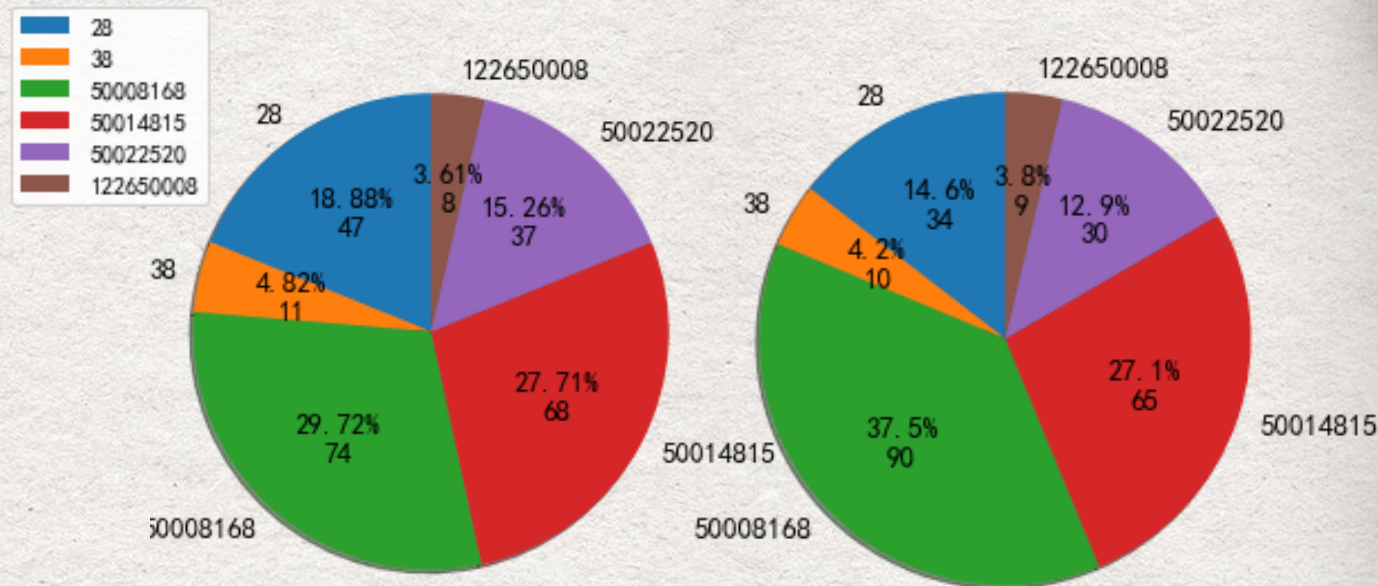
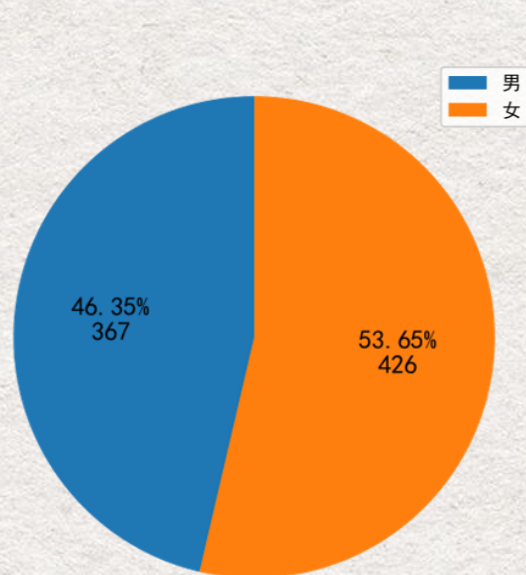


2013 (左) 2014 (右) 年各季度商品类别分布图

- 与2013年相比，2014年各季度个商品的总销量均明显多于2013年。
- 38，50008168商品销量占各个季度销量的大多数，且38商品销量随季度变化改变较明显。
- 50022520、122650008商品销量各季度均明显保持稳定

03 数据分析成果展示

——性别分析



孩子性别分布图

Click here add your text Click here add your text
Click here add your text Click here add your text

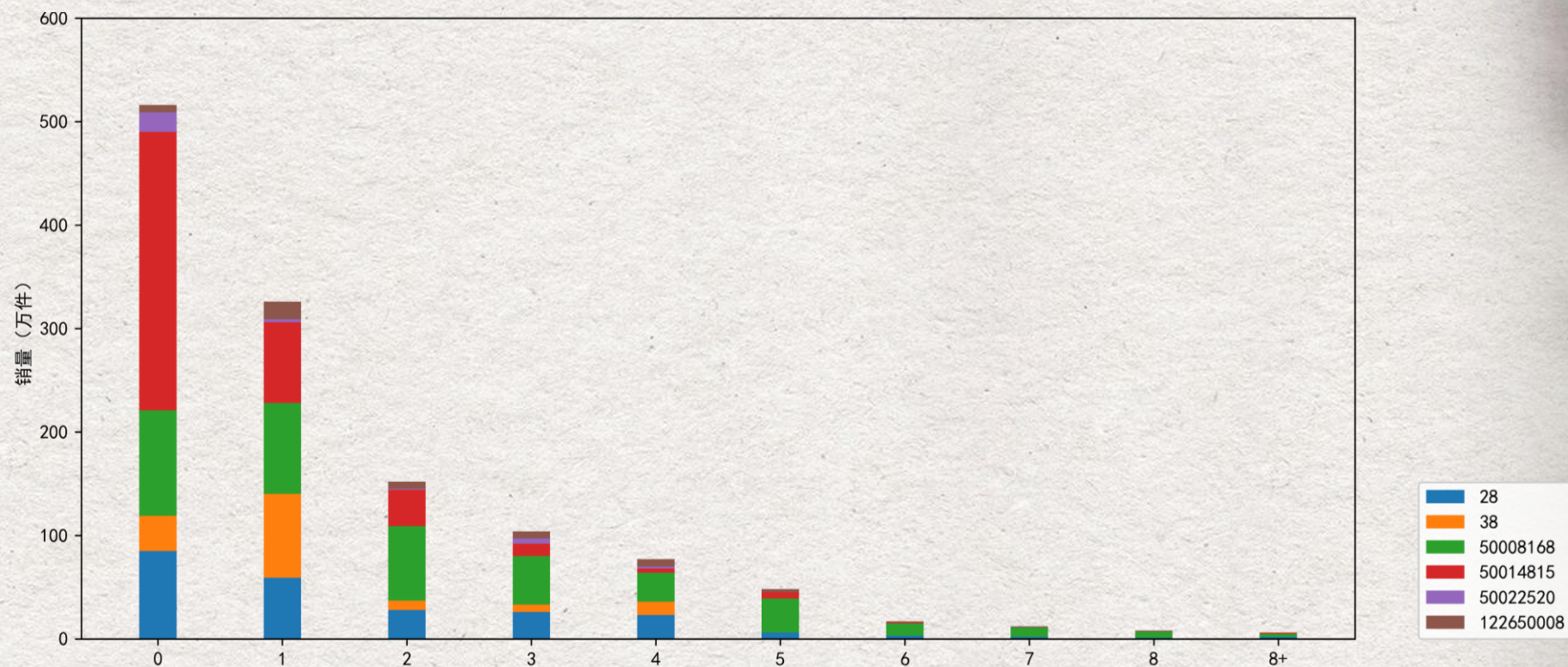


男性 (左) 女性 (右) 孩子商品类别分布图

- 不同性别孩子再商品类别分布上不存在明显差异
- 相比而言，男童28、50022520比例偏高，50008168、50022520偏低，其他相对持平

03 数据分析成果展示

——年龄分析



各年龄段各根类别销量构成图

- 在销售量上，0-3岁是整个市场的主要用户，其中0-365天的用户是母婴市场的主力军
- 0-1岁的孩子更偏向于50014815，尤其是0-365天的孩子，而在2岁+的孩子中，50008168占据了绝对优势
- 而且总体来说，年龄越大，人数越少的情况下，销量占比也越高。而28商品的主顾客群体为0-4岁的孩子

03 数据分析成果展示

——数据挖掘

1 查看数据类型及分布

```
round(train_df.describe(percentiles=[.45,.5, .6, .7, .75, .8, .9, .99]),2)
```

	Column1	user_id	auction_id	cat_id	cat1	buy_mount	gender	age
count	650.00	6.500000e+02	6.500000e+02	6.500000e+02	650.00	650.00	650.00	650.00
mean	335.35	4.531264e+08	2.426991e+10	5.388015e+07	2.30	1.38	0.53	1.69
std	193.80	4.968262e+08	1.124221e+10	2.048415e+07	1.47	1.87	0.50	2.11
min	0.00	2.757000e+03	7.405612e+07	2.111220e+05	1.00	1.00	0.00	-1.00
45%	304.05	1.963321e+08	1.923268e+10	5.001245e+07	2.00	1.00	0.00	1.04
50%	336.50	2.769950e+08	2.011764e+10	5.001256e+07	2.00	1.00	1.00	1.19
60%	404.40	4.065938e+08	2.412212e+10	5.001364e+07	2.00	1.00	1.00	1.80
70%	469.30	6.591684e+08	3.560635e+10	5.001444e+07	3.00	1.00	1.00	2.40
75%	502.75	7.164043e+08	3.669493e+10	5.001671e+07	3.00	1.00	1.00	2.72
80%	536.20	7.684868e+08	3.755219e+10	5.001883e+07	3.00	1.00	1.00	3.14
90%	601.10	1.040472e+09	4.013649e+10	5.002646e+07	5.00	2.00	1.00	4.50
99%	662.51	2.109740e+09	4.315471e+10	1.214720e+08	6.00	10.00	1.00	8.84
max	670.00	2.298687e+09	4.367403e+10	1.224740e+08	6.00	30.00	1.00	11.80


03 数据分析成果展示

——数据挖掘



1 查看数据类型及分布

2 提出样本特征间关系的假设

- 婴幼儿性别与所购买的商品根类别特征相关
 - 婴幼儿性别与所购买的商品根类别特征相关
- 

03 数据分析成果展示

——数据挖掘

1 查看数据类型及分布

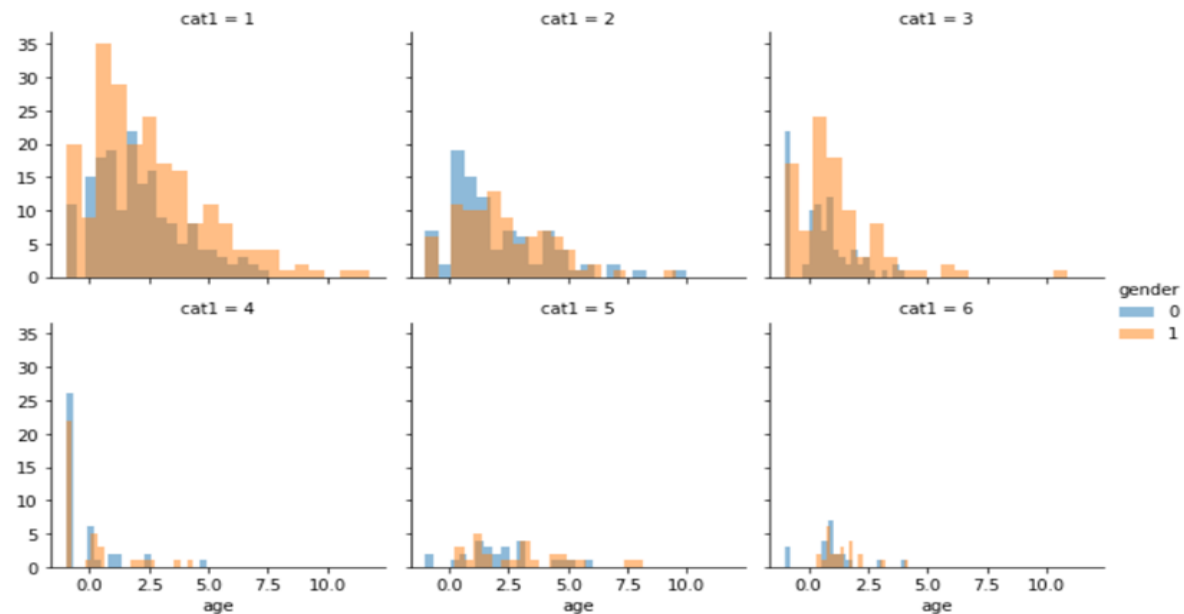
2 提出样本特征间关系的假设

3 进行特征分析

```
cols = train_df['cat1']
```

```
grid = sns.FacetGrid(train_df, col='cat1', hue='gender', col_wrap = 3, height = 3)  
grid.map(plt.hist, 'age', alpha=0.5, bins=20)  
grid.add_legend()
```

<seaborn.axisgrid.FacetGrid at 0x2c6004324c0>



03 数据分析成果展示

——数据挖掘

1 查看数据类型及分布

2 提出样本特征间关系的假设

3 进行特征分析


4 建立模型

Ada Boosting

```
# Ada Boosting
model_ADA = AdaBoostClassifier(n_estimators=100, learning_rate=1)
model_ADA.fit(X_train, Y_train)
print("训练准确率: ", end = ' ')
print(model_ADA.score(X_train, Y_train))
scores = cross_val_score(model_ADA, X_train, Y_train, cv=11)
'11折交叉验证: ' + str(scores.mean())
```

训练准确率: 0.6178686759956943

'11折交叉验证: 0.5188439011968424'

The background of the slide is a light beige, textured surface. In the four corners, there are clusters of stylized green leaves. These leaves are layered and have white veins, giving them a paper-cut or origami-like appearance. The leaves in the top-left and bottom-left corners are a lighter green, while those in the top-right and bottom-right corners are a darker green. The bottom-right cluster also features some small yellow dots and a small yellow smiley face on one of the leaves.

04

总结

THE PART THREE

04 总结

——收获与不足



遇到问题

- 数据预处理时对于异常值的判定与空缺值的处理
- 数据分析时的中文乱码问题
- 数据挖掘时准确度较低等
-

小组收获

- 掌握一定的数据分析、数据挖掘的代码编写能力
- 对数据科学的一般流程有了更深的了解
- 加强了沟通合作能力
-

项目不足

- 模块间耦合性较高
- 使用公开数据集
-

04 总结

——成员分工

