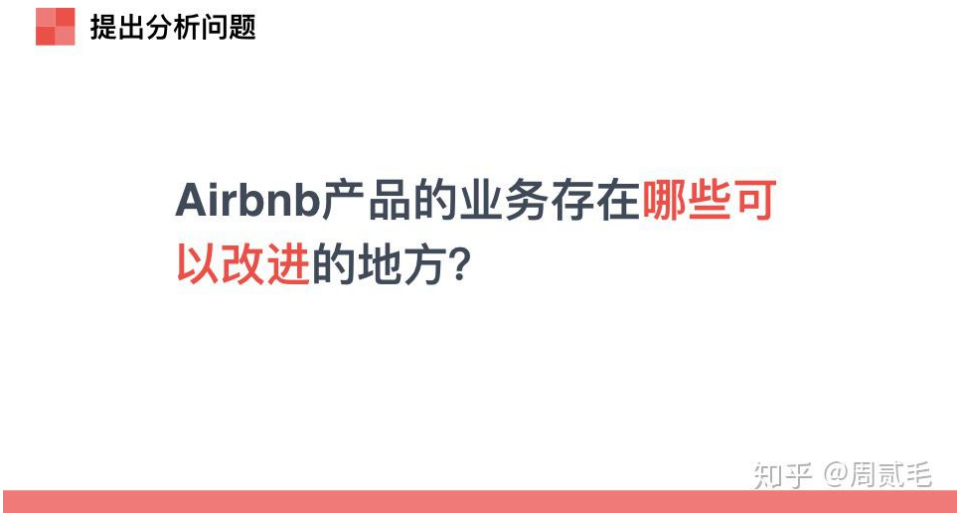
一、分析背景与目的

1.1、分析背景

Airbnb 成立于 2008 年，短短 9 年时间成为了短租民宿行业的巨头，并且强烈的冲击着传统酒店行业的格局。

目前 Airbnb 作为一款社区平台类产品，其业务遍布了 191 个国家，并且经常出现在商业分析的优秀案例中。Airbnb 在做好了产品体验、房源美感、民宿共享服务之后，这款产品和背后的业务是否存在可以改进的地方？

1.2、提出分析问题



一款产品的发展必然伴随着不断的迭代。在 AARRRR 模型中，第一个 A（用户获取）中，提高新用户获取的数量和质量是不断监测并优化的一个工作，哪些渠道的效果更好，企业就要及时调整和增加此渠道的投入，哪些渠道的效果很差，就要及时查找原因并给出解决。

另外转化漏斗分析也是数据分析环节的重要指标，可以从宏观角度了解整个产品的业务转化情况，企业针对流失率较高的漏斗环节进行改进，可以有效促进业务发展。

针对分析的目的，提出以下三个问题：

1. airbnb 的目标**用户群体**具有什么样的特征？
2. airbnb 当前的**推广渠道**有哪些是优质的、有哪些做的还不够好且需要改进？
3. 当前的**转化率和流失率**中哪里哪一个环节存在问题，或者有较大的改进空间？

二、分析维度

2.1、根据问题设立分析维度与分析指标

根据问题，提出来三个分析维度：

将着重从 airbnb 的***用户画像、推广渠道分析、转化漏斗分析三个***方面进行分析，去探索和分析 airbnb 在产品和业务上有哪些可以改进的地方，并给出实际性的建议，以提升和改进 airbnb 的渠道推广策略和产品设计。

1、用户画像分析

- 用户性别的分布特征；
- 用户年龄的分布特征；

- 用户地区的分布分布；
- 中国地区去国外预定的地区占比；

2、推广渠道分析

- 每月新增用户
- 不同用户端的注册量
- 不同推广渠道的注册量
- 不同营销内容的注册量
- 不同推广渠道的转化率
- 不同营销内容的转化率

3、转化漏斗分析

- 注册用户占比
- 活跃用户（非僵尸用户）占比
- 下单用户占比
- 实际支付用户占比
- 复购用户占比

三、数据清洗

3.1、数据集描述

数据集名称：Airbnb 顾客预订数据

数据集来源：
https://www.kaggle.com/c/airbnb-recruiting-new-user-bookings/data

数据集简介：此数据集是 kaggle 上的一个竞赛项目，主要用来制作目的地信息的预测模型。此数据聚集包含两张数据表，其中 train_user 表中为**用户数据**，sessions 表中为**行为数据**。

数据集量：21w * 15（train_user）、104w * 6（sessions）

3.1、列名称理解

原数据表的字段的列名非常规范，无需对列名称进行重命名，数据分析之前需要理解每个列名称的含义。

数据表一：数据表名称：train_users

id：用户 ID

date_account_created：帐户创建日期

date_first_booking：首次预订的日期

gender：性别

age：年龄

signup_method：注册方式

signup_flow：用户注册的页面

language：语言偏好

affiliate_channel：营销方式

affiliate_provider：营销来源，例如 google，craigslist，其他

first_affiliate_tracked：在注册之前，用户与之交互的第一个营销广告是什么

signup_app：注册来源

first_device_type：注册时设备的类型

first_browser：注册时使用的浏览器名称

country_destination：目的地国家

数据表二：数据表名称：sessions

user_id：与 users 表中的“id”列连接

action：埋点名称

action_type：操作事件的类型

action_detail：操作事件的描述

device_type：此次会话所使用的设备

3.2、重复值的处理

- train_users 为用户表中主键，所以每个用户只生成一条记录，所以如果 train_users 中 id 存在重复值，则需要处理。
- sessions 为用户会话记录表，存在一个用户多条记录

结论：只需要*排查 train_users 中是否存在重复值*。

执行 SQL 后得出：count_id = 0。说明 train_users 数据表中不存在重复值。

3.3、缺失值处理

数据缺失数量较多，以下为**存在缺失值的列**：

- date_first_booking(首次预定时间)存在缺失值数量：124544 个。
- gender（性别）存在缺失值数量：95688 个。
- age（年龄）存在缺失值数量：87991 个。
- first_affiliate_tracked（用户通过那个营销广告注册）存在缺失值数量：6065 个。
- first_browser（注册时浏览器）存在缺失值数量：27266 个。
- action_type（埋点的操作类型）存在缺失值数量：1126204 个。
- action_detail（用户操作行为的描述）存在缺失值数量：1126204 个。

缺失原因推测及处理

- date_first_booking(首次预定时间)数据如果缺失,在业务上可以理解为此用户为“未预定用户”，也就是没有下单的用户。
- 性别、年龄由于客户端中这部分信息选填，空值为用户未填写。
- 其他四个数据是由于前端统计时数据没有统计到。

处理：实际分析中需要在 *where 条件排除掉空数据，再进行分析*。

3.4、异常值处理

以下为异常值的处理说明。

date_account_created 异常值处理：

date_account_created 中最小值为‘0000-00-00 00:00:00’；因为此异常数据只有 1 条，所以删除此条数据。

age（年龄）异常值处理：

age（年龄）的异常数据非常多；0～150 之间的数值都有，并且包含了 2014、2015 等数值。推测这些"脏数据"产生的原因是用户在客户端随意填写造成。

3.5、数据清洗中使用的 SQL

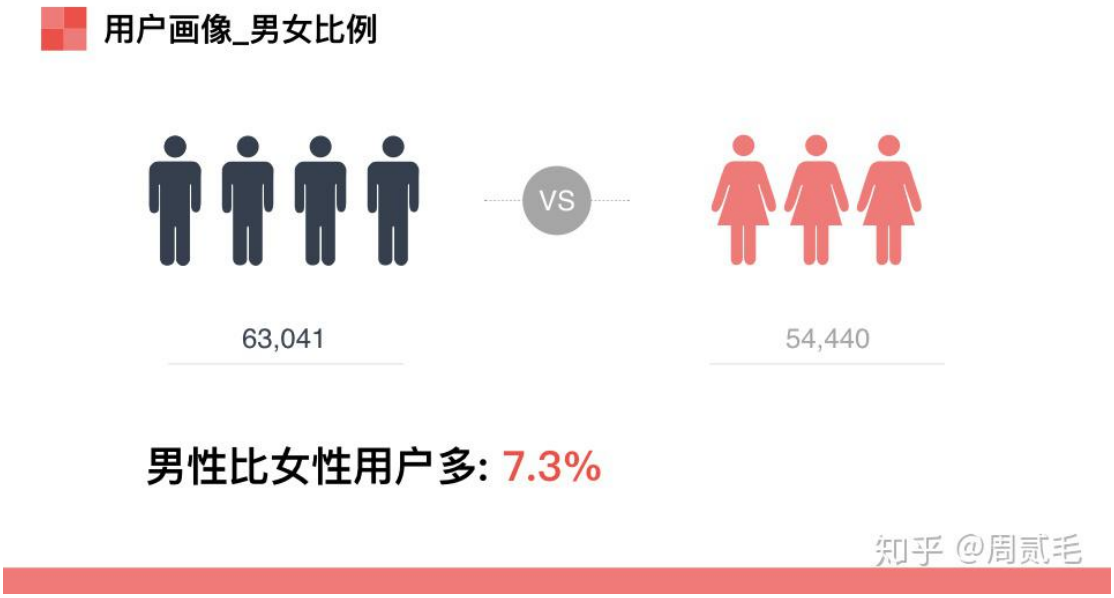
```
#检查数据中是否包含重复值
SELECT id, COUNT(id) AS count_id
FROM data.train_users
GROUP BY id
```

```
HAVING count_id > 1;

#通过以下 SQL 对每一列进行查询，通过替换 where 之后的条件，查询每一列包含的空值数量。
SELECT date_first_booking, COUNT(date_first_booking)
FROM data.train_users
WHERE date_first_booking = '0000-00-00 00:00:00';


#通过查看数据的极值（极大值、极小值）是否符合实际情况，来判断数据中是否存在异常值。
SELECT  min(age),max(age) FROM data.train_users;

#异常值处理：对于年龄不在 7～75 区间的数据删除（设置为 0-空值）
SET sql_safe_updates = 1;
UPDATE data.train_users
SET age = 0
WHERE id NOT IN (
                SELECT id
                FROM (
                        SELECT id
                        FROM data.train_users
                        WHERE age <= 75
                                AND age >= 7
                ) a
        );
```

四、用户画像分析

4.1、用户的性别分布特征

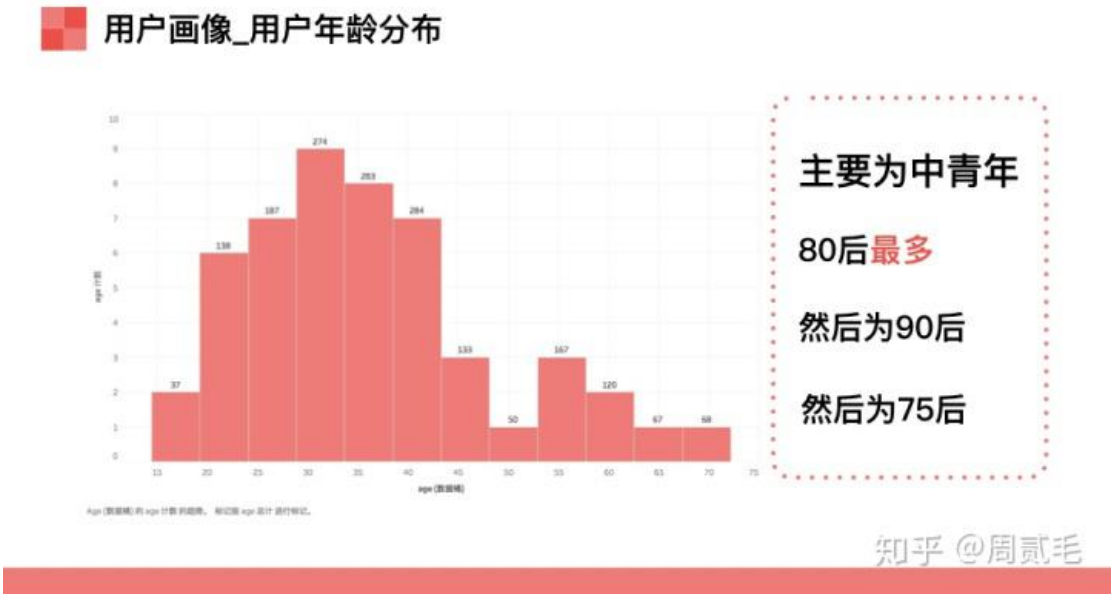
执行以上 SQL 得出：女性用户数量=54440；男性用户数量=63041



从以上可视化可以看出：

airbnb 的男女用户占比差别不大，其中男性用户多于女性用户。

4.2、用户的年龄分布；

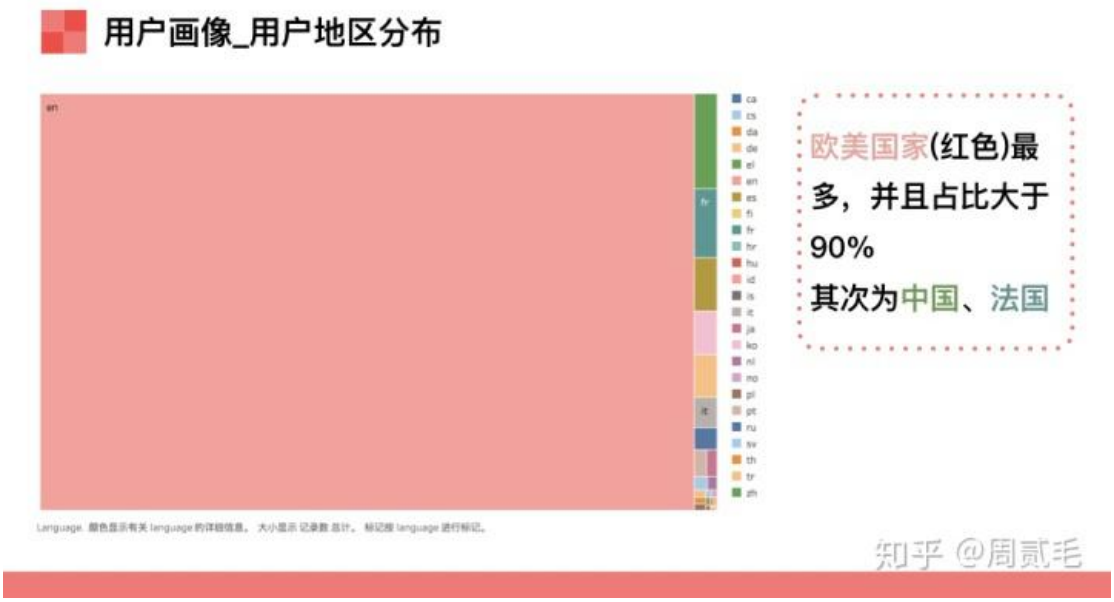


从直方图可以看出：

airbnb 的用户主要为“中青年群体”，其中用户数量最多的是 80 后～（29 岁～39 岁），其次为 90 后，然后为 75 后。

4.3、用户不同地区的分布

SQL 执行的结果可以看出：使用最多的语言排名前 5 的分别是英文、中文、法文、西班牙文、朝鲜文。

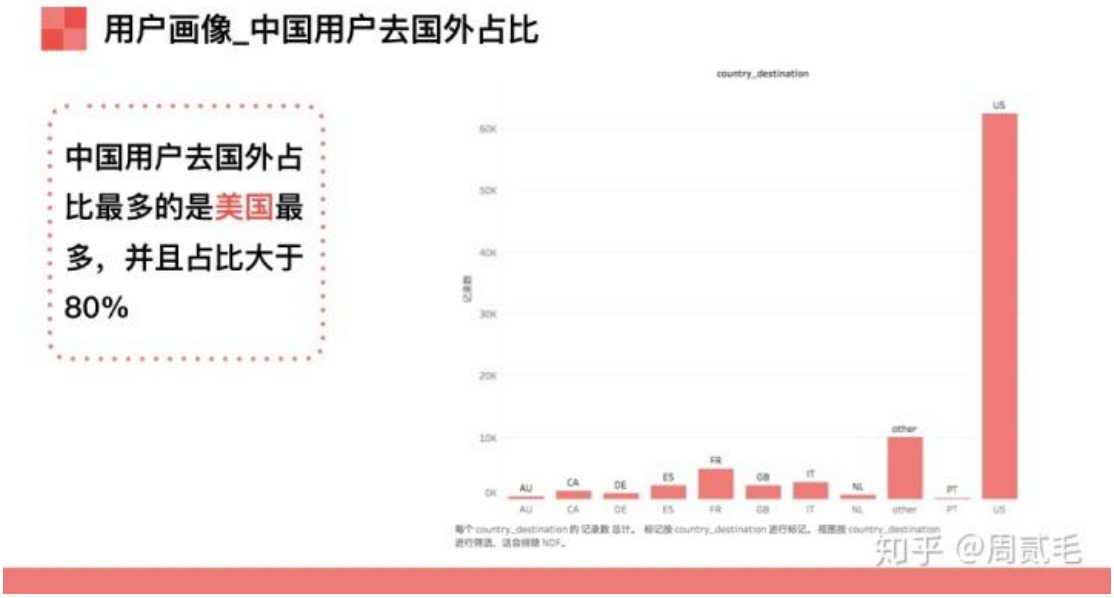


从可视化结果可以看出：

- airbnb 的产品真的很国际化，用户遍布多个地区。
- 有超过 90%的用户是英语国家（欧美）；airbnb 是 2013 年开始进入中国市场的（此数据集止于 2014 年），所以此时中文用户数量虽然排名第二，但是占比却非常小。

4.4、中国用户去国外预定的地区占比

从 SQL 执行的结果可以看出，中国人去国外预定民宿最主要集中在美国、然后是法国。除了这两个国家外，样本中其他国家的数量都少于 5。



从可视化结果可以看出：

- 中国用户去国外预定，占比最多的是美国。其余国家占比很小，总和不到20%。

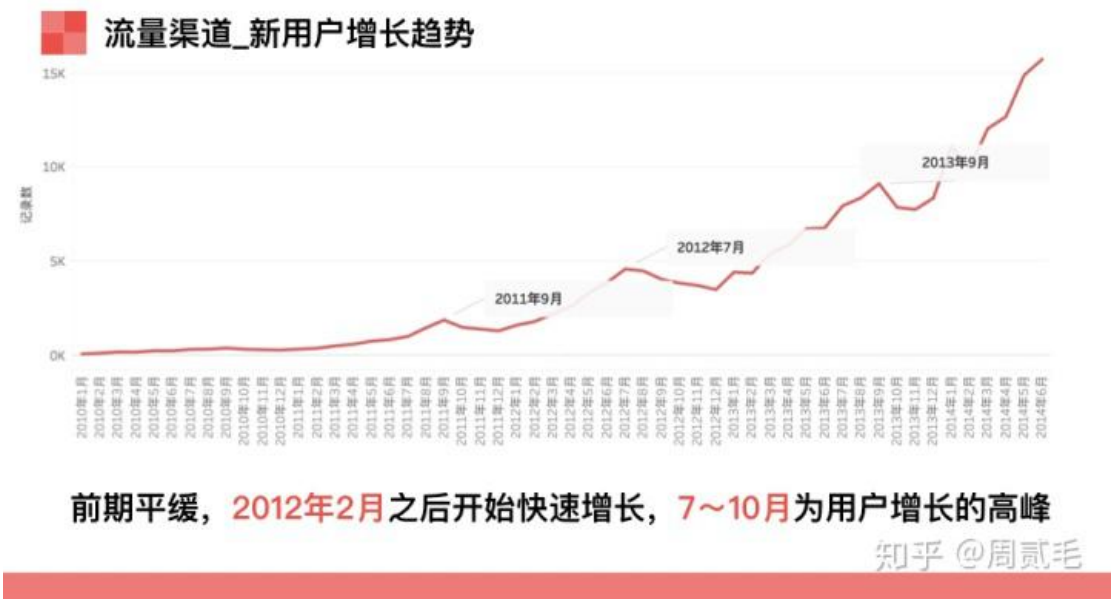
4.1、本章使用的 SQL 语句

```
#用户中女性用户的数量。
SELECT COUNT(id) AS '女性用户数量'
FROM data.train_users
WHERE gender = 'MALE';

#用户中男性用户的数量。
SELECT COUNT(id) AS '男性用户数量'
FROM data.train_users
WHERE gender = 'FEMALE';

#用户不同年龄的数量。
SELECT age, COUNT(id)
FROM data.train_users
GROUP BY age
HAVING age <> 0
ORDER BY age;

#用户不同语言的分布（通过手机系统语言数据）；
SELECT language, COUNT(language) AS lg_num
FROM data.train_users
GROUP BY language
ORDER BY lg_num;
```

```
#中国用户去国外预定的地区
SELECT language, country_destination, COUNT(country_destination) AS
cd_num
FROM data.train_users
GROUP BY language, country_destination
HAVING language = 'zh'
ORDER BY cd_num DESC;
```

五、流量渠道分析

5.1、每月新增用户



从可视化结果可以看出：

- airbnb 的用户增长曲线健康，前期（2011 年之前）平缓，2012 年 2 月之后开始快速增长。
- 2012 年之后的增长速度很快。
- 此产品新用户的增加**存在季节性规律**：每年的 7～10 月，产品都会迎来用户增长的高峰，推测为夏季（北半球）是旅行的旺季，而短租产品本身就是旅行消费的一种。

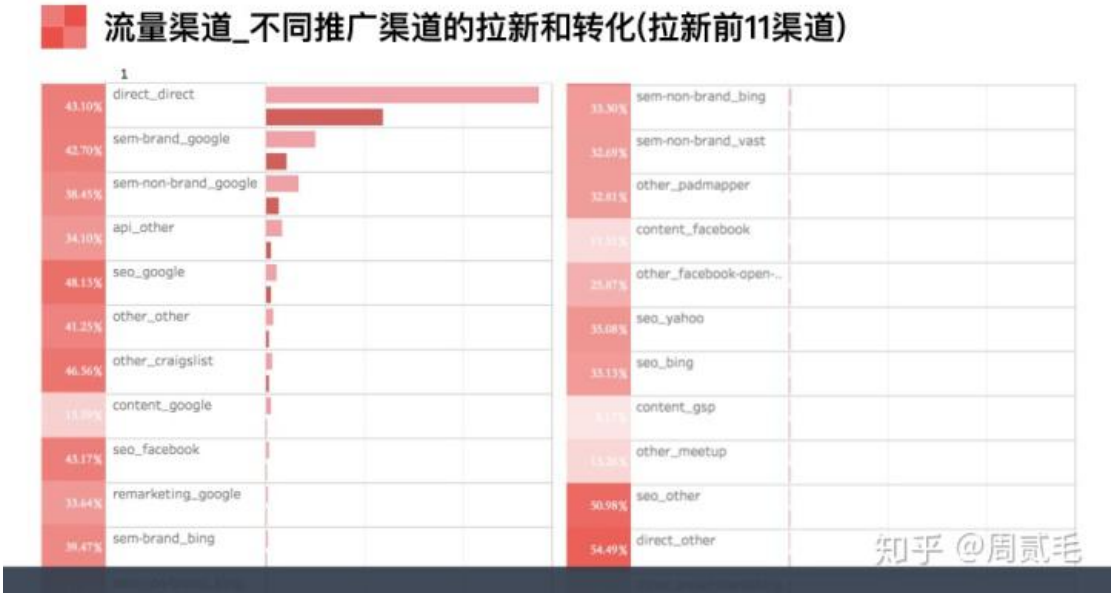
5.2、不同用户端的注册量

从可视化结果可以看出：

- 此数据为 2014 年之前的数据，当时智能手机还没有像现在一样普及，用户的注册设备 PC 大于移动设备。
- 苹果设备数量大于其他设备数量。

5.3、不同推广渠道的注册量

不同推广渠道的可视化：



备注：图表中排除了注册量在 150 以下的渠道

可视化图表中可以看出：

渠道注册量方面：

- airbnb 的整体渠道转化率表现很好，多数渠道的转化率都在 30%以上。
- 表现最好的为谷歌竞价（SEM）,其中品牌竞价注册量大于非品牌竞价的注册量。
- 渠道注册量符合二八定律，及前 7 个渠道（总共有 38 个渠道推广）的注册量已经占据了产品总的渠道来源的 90%以上。
- 另外从执行的 SQL 结果可以看出：drict（直接应用市场下载注册）的注册量最多，占总注册量的 64.38%。我们的分析目的是查看推广渠道的好坏，具体分析中排除了 drict。

渠道转化率存在的问题：

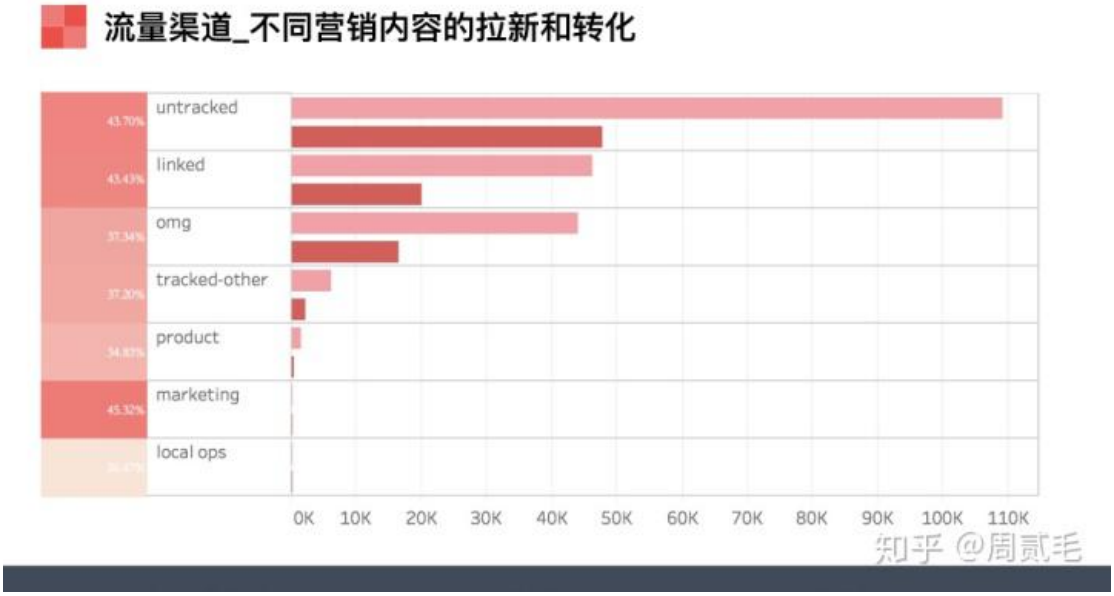
- 主要渠道（注册量在前 7 名的渠道）中，content_google 的转化率异常，明显低于转化率的均值。
- 主要渠道（注册量在前 7 名的渠道）中，api_other（其他产品的 API 对接）渠道的转化率虽然大于 30%，单数相较于其他渠道，转化率偏低。
- 发现 content（内容推广）这一种推广方式下各渠道的转化率都很低，其中 content_gsp 的转化率只有 8.2%。
- sem-non-brand_bing、sem-non-brand_vast 两种 SEM 渠道的转化率都偏低（35%以下）。

优质渠道（注册量和转化率都比较好）：

- seo_google 的注册量和转化率表现都很好。
- sem_brand_google 的注册量可转化率表现很好。

5.5、不同营销内容的注册量和转化率

不同营销内容的可视化：



从可视化图表中可以看出：

- 从上图可以看出注册量排名第一的营销内容为 untracked（未跟踪到），产品的数据跟踪异常。
- 从统计到的数据来看，linked 注册量位居第一。转化率方面，linked 和 product 两个营销内容的的转化率好，在 43%以上。local ops 的转化率非常低。

5.5、分析过程使用的 SQL

```
#查询每个月（date_account_created）新增注册用户的数量
SELECT date_format(date_account_created, '%Y-%M') AS year_moth,
COUNT(id)
FROM data.train_users
GROUP BY date_format(date_account_created, '%Y-%M')
ORDER BY year_moth;

#查询不同设备来源（first_device_type）注册的数量
SELECT first_device_type, COUNT(id) AS fdt_num
FROM data.train_users
GROUP BY first_device_type
ORDER BY fdt_num DESC;

#不同推广方式+渠道的注册数量
SELECT affiliate_channel, affiliate_provider, COUNT(id) AS ac_num
FROM data.train_users
GROUP BY affiliate_channel, affiliate_provider
ORDER BY ac_num DESC;
```

```
#不同推广方式+渠道的转化率
SELECT affiliate_channel, affiliate_provider,
SUM(CASE WHEN date_first_booking <> '0000-00-00 00:00:00' THEN 1 ELSE 0
END) / COUNT(id) AS ac_ratio
FROM data.train_users
GROUP BY affiliate_channel, affiliate_provider
ORDER BY ac_ratio DESC;

#不同营销广告内容的注册数量
SELECT first_affiliate_tracked, COUNT(id) AS fat_num
FROM data.train_users
GROUP BY first_affiliate_tracked
ORDER BY fat_num DESC;

#不同营销广告内容的转化率
SELECT first_affiliate_tracked,
SUM(CASE WHEN date_first_booking <> '0000-00-00 00:00:00' THEN 1 ELSE 0
END) / COUNT(id) AS fat_ratio
FROM data.train_users
GROUP BY first_affiliate_tracked
ORDER BY fat_ratio DESC;
```

六、转化漏斗分析

执行 SQL 得出：用户总数量 = 135484。

6.1、活跃用户（非僵尸用户）占比

执行 SQL 得出：活跃用户总数量 = 114002。

计算可得：活跃用户占比 = 活跃用户总数量/用户总数量 = 84.144%

6.2、注册用户占比

执行 SQL 得出：注册用户总数量 = 73815。

计算可得：注册用户占比 = 注册用户总数量 / 用户总数量 = 54.482%

6.3、下单用户占比

执行 SQL 得出：下单用户总数量 = 10367。

计算可得：下单用户占比 =下单用户总数量 / 用户总数量 = 7.651%

6.4、实际支付用户占比

执行 SQL 得出：实际支付用户总数量 = 9019。

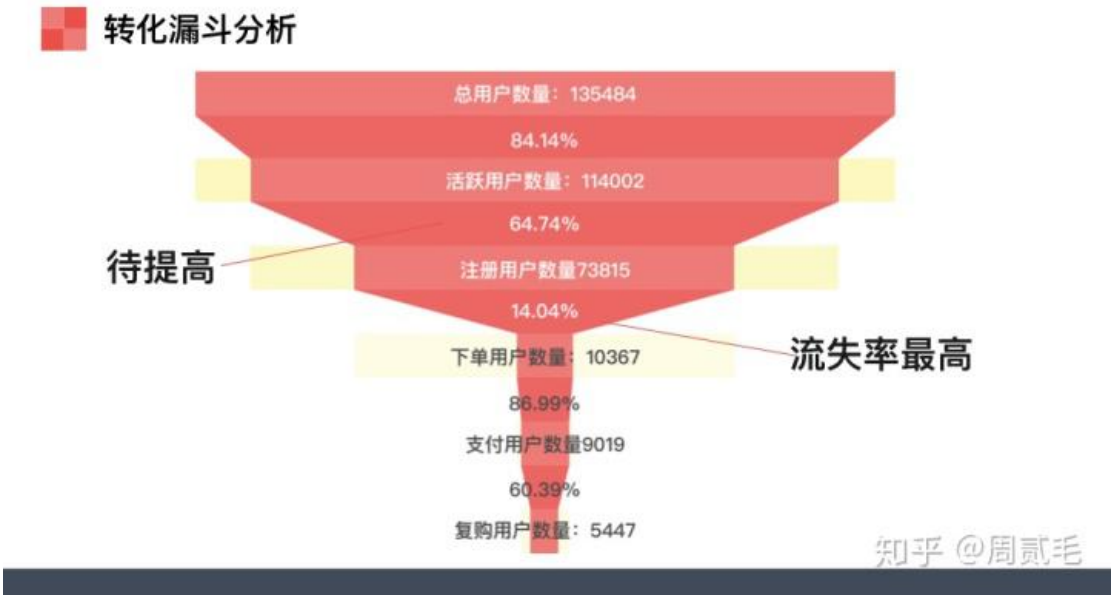
计算可得：付款用户占比 =实际支付用户总数量/用户总数量 = 6.656%

6.5、复购用户占比

执行 SQL 得出：复购用户总数量 = 5447。

计算可得：复购用户占比 = 复购用户总数量 / 用户总数量 = 4.0204%

airbnb 的漏斗模型：



从可视化图表中可以看出：

- 注册用户到下单用户是 airbnb 转化漏斗中流失率最高的一个环节。仅有 14%的注册用户下单、仅占全部用户的 7.651%。
- 活跃和复购环节表现的好，其中有 60%的下单用户复购，说明 airbnb 的产品和服务做的非常好。
- 下单用户中有大约 13%的用户没有最终支付，需要产品研发介入排查。

6.4、漏斗分析过程 SQL

```
#用户总数量：对 sessions 表中的 user_id 进行 group by，再统计数量，得出 sessions 表中所有的用户数量。
SELECT COUNT(*)  AS  '用户总数量'
FROM (
        SELECT user_id
```

```
        FROM data.sessions
        GROUP BY user_id
) new_sessions;


#活跃用户的定义：按照用户的操作总次数，如果用户操作产品大于等于10次，就可以说明用户为偏活跃的用户，另一方面说明此用户不是僵尸用户。
SELECT COUNT(*)  AS  '活跃用户总数量'
FROM (
        SELECT user_id
        FROM data.sessions
        GROUP BY user_id
        HAVING COUNT(user_id) >= 10
) active;

#注册用户：通过sessions表中的用户与注册用户表进行内关联，统计出sessions表中已注册用户数量
SELECT COUNT(*) AS  '注册用户总数量'
FROM (
        SELECT user_id
        FROM data.sessions
        GROUP BY user_id
) new_sessions
        INNER JOIN data.train_users tu ON new_sessions.user_id = tu.id;

#下单用户：用户行为中“reservations”为预定（下单）操作，通过统计进行了“reservations”的用户（group by去重），得出下单用户的数量
SELECT COUNT(*) AS  '下单用户总数量'
FROM (
        SELECT user_id
        FROM data.sessions
        WHERE action_detail = 'reservations'
        GROUP BY user_id
) booking;

#实际支付用户：用户行为中“payment_instruments”为支付操作，通过统计进行了“payment_instruments”的用户（group by去重），得出实际支付用户的数量
SELECT COUNT(*) AS  '实际支付用户总数量'
FROM (
        SELECT user_id
        FROM data.sessions
        WHERE action_detail = 'payment_instruments'
        GROUP BY user_id
```

```
) payed;

#复购用户:通过统计进行了“payment_instruments”操作次数大于 1 次的用户（group by 去重），得出实际复购用户的数量
SELECT COUNT(*) AS ‘复购支付用户总数量’
FROM (
        SELECT user_id
        FROM data.sessions
        WHERE action_detail = 'reservations'
        GROUP BY user_id
        HAVING COUNT(user_id) >= 2
) re_booking;
```

七、分析结论汇总

7.1、用户画像总结

- 用户性别中，男性用户多于女性用户，但是差别不大（7.3%的差距量）
- 用户年龄以中青年为主，用户数量最多的是 80 后～（29 岁～39 岁），其次为 90 后，然后为 75 后。
- 用户分布地区最多的为欧美地区，其次是中国，但欧美占比达到了 90%以上。（备注：截止 2014 年）
- 中国用户预订的最多的其他国家是美国，占比高达 90%以上。

7.2、流量渠道总结

- 前期（2011 年之前）平缓，之后（2012 年 1 月之后）开始快速增长，并且速度很快。
- 7～10 月是旅行旺季、此时也是 airbnb 用户增长的旺季。
- 苹果设备用户居多。
- drict（直接应用市场下载注册）的注册量最多，占总注册量的 64.38%。
- 注册量排名 7 的渠道，占据了产品全部渠道注册来源数量的 90%以上。

效果不好的渠道：

- 主要渠道中，content_google 的转化率异常，明显低于转化率的均值。
- api_other（其他产品的 API 对接）渠道的转化率也比较低。
- content（内容推广）这一种推广方式下各渠道的转化率都很低，其中 content_gsp 的转化率只有 8.2%。
- sem 中 sem-non-brand_bing、sem-non-brand_vast 两种 SEM 渠道的转化率都偏低

效果好的渠道：

- seo_google 的注册量和转化率表现都很好。
- sem_brand_google 的注册量可转化率表现很好，但是是付费渠道。

营销内容方面：

- 统计功能异常、数据追踪效果差。
- linked 和 product 两个营销内容的的转化率好。
- 相比较其他营销内容的转化率、local ops 的转化率非常低。

7.3、转化漏斗总结

- airbnb 转化漏斗中流失率最高的一个环节是“用户下单”，仅有 14%的注册用户下单。
- 下单用户中有大约 13%的用户最终没有支付成功。
- 注册率有待提高。
- 活跃和复购环节表现的好，说明 airbnb 的产品和服务做的非常好。

八、业务和产品上的建议

建议一：用户画像

- 根据年龄分布特征，建议 SEO 或者付费广告投放时，投放对象细化至年龄在 29～39 岁的男性。

建议二：关于推广渠道上的改进

- 7～10 月是业务的旺季，建议运营部门在每年的 7～10 月加大活动营销的力度，同时加大渠道广告的投放力度。
- 在主要渠道（注册量在前 7 名的渠道）中 content_google 非常低（只有 15%），建议运营部门计算此渠道的 ROI 和 ARPU（每客户平均收入），如果 ROI > ARPU，建议停止此渠道的投放。
- 在主要渠道（注册量在前 7 名的渠道）中 api_other（其他产品的 API 对接）渠道的转化率较低，建议产品设计部查找尾部排名的 API 对接产品，与对方产品沟通，从产品的流程设计、交互设计角度查找原因。
- SEO 推广下各渠道的拉新和转化都好，SEO 作为一种较低成本的获客方式（主要为人力成本），建议企业管理层日常要更加支持 SEO 相关的资源投入，甚至考虑扩大 SEO 的团队。
- 营销内容的埋点统计效果很差，常见两方面的原因。如果是研发导致的统计功能出错，需要立即修复。如果是运营人员不注重这一块的埋点统计，也就是运营同学没有提出埋点的任务，则需要行政介入，要求之后做好埋点工作。

- 不通营销内容的拉新和转化效果也不同，优秀营销内容（linked 和 product）和表现较差的营销内容（local ops），如果在活动过程建议对 local ops 及时优化内容甚至更换。如果是活动之后、建议运营部门针对不同质量的营销内容做对比分析，总结内容策划上的方法论，便于之后的实践。

建议三：转化漏斗方面

- 注册用户到下单用户是 airbnb 转化漏斗中流失率最高的一个环节，仅有 14%的注册用户真正下单，此环节作为企业营收的主要来源之一，建议围绕提高下单率做更多的工作。例如针对活跃用户的用户轨迹定期推送（产品 push+短信邮件）优质房源，此外提高下单转化率是一项长期工作、需要结合多种策略并行。
- 下单用户中有大约 13%的用户最终没有支付，需要排查具体原因（为什么已经下单啦，还是十分之一的用户没有结算），建议进行用户调研、或者在产品上统计用户未支付原因（是用户自身决策导致、还是产品流程的原因、还是支付类型不满足个别地区等）。

作者：周贰毛
链接：https://zhuanlan.zhihu.com/p/77558304
来源：知乎