## 项目背景

本周我们接到了某知名互联网公司的需求，他希望我们协助分析APP的用户属性和付费情况，以此来制定产品消息推送的策略。

对于一款APP来说，如何推送消息是一个非常重要的思考题。比如：

* 给哪些用户推送消息？
* 在什么样的时候推送消息？
* 推送怎样的消息可以带来用户的活跃与付费？

这些都是值得产品经理和业务人员去探索的问题。

因此，本周我们将会协助客户进一步分析APP用户数据以及产品消息推送数据。

## 界定问题

### What

* 用户的地域分布(分省按从大到小排列)是怎样的？
* 用户的性别分布是怎样的?
* 付费用户与未付费用户的分布是怎样的？
* 付费用户与未付费用户的地域.性别分布是怎样的?

### Why

* 你会优先向哪些地域和性别的用户进行推送，为什么？

### How

* 根据付费用户的地域以及性别分布，怎样设计对应的产品消息推送策略

## 数据搜集与评估

问题一，需要用户表的用户id和省份

问题二，需要用户表的性别和省份

问题三，需要获取付费表中付费用户数和用户表中未付费的用户数

问题四，按地域和性别分组，统计付费和未付费用户数

问题五，按地域和性别分组，统计推荐表中的可推荐量

## 数据的整理与清洗

用户表

首先确定，用户表的用户id位数，是否为32位

代码如下

SELECT DISTINCT length(user\_id )FROM t\_user\_id\_info

结果均为32位，无异常

之后确定，用户表的省份有没有格式问题

代码如下

SELECT DISTINCT(province) FROM t\_user\_id\_info

结果为无异常

此处要注意上海，北京，重庆属于直辖市和各个省份并列没问题。

然后检查用户表的性别有无异常

代码如下

SELECT DISTINCT(gender) FROM t\_user\_id\_info

结果无异常

付费表

首先，检查付费表用户id位数是否有异常

代码如下

SELECT DISTINCT length(user\_id )FROM paid

结果无异常，都是32位

然后，检查付费表用户付费标识是否有异常

代码如下

SELECT DISTINCT(is\_pay) FROM paid

结果无异常

推荐表

首先，检查推荐表中的省份有无异常

代码如下

SELECT DISTINCT(province) FROM push\_cnt

结果无异常

之后检查，推荐表中的性别是否有异常

代码如下

SELECT DISTINCT(gender) FROM push\_cnt

结果为无异常

最后检查app推荐启动量是否有异常

代码如下

SELECT DISTINCT(cnt) FROM push\_cnt

结果为无异常

## 数据的探索与可视化

问题一

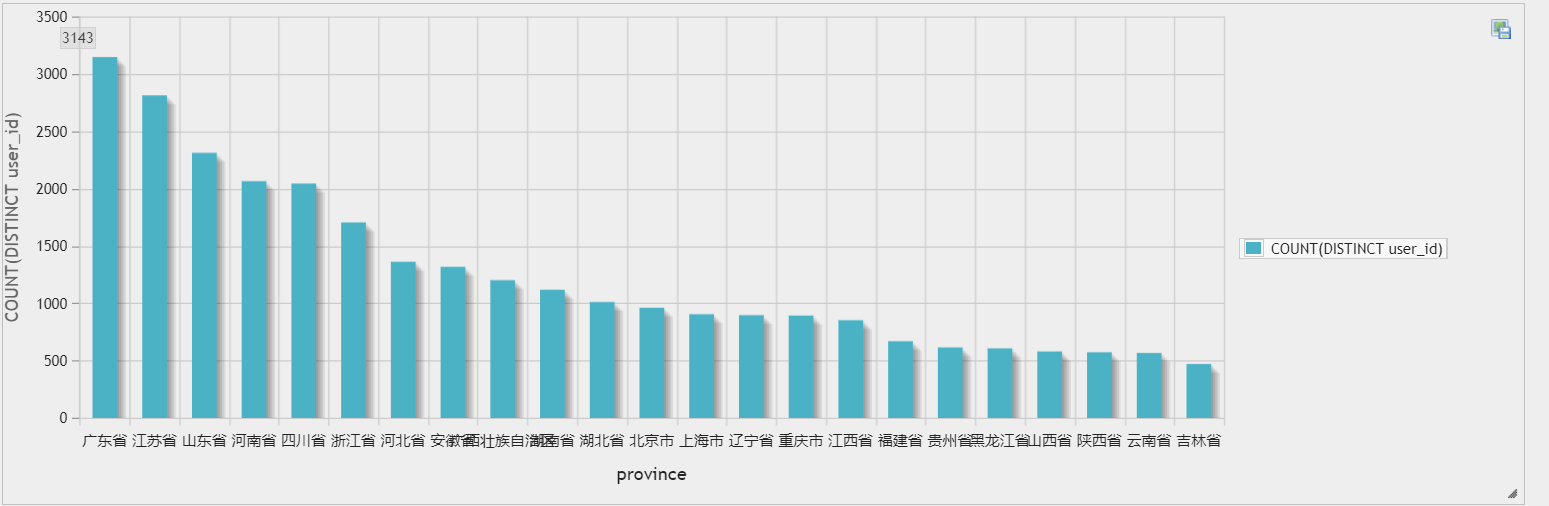
用户的地域分布(分省按从大到小排列)是怎样的？

代码如下

SELECT province ,COUNT(DISTINCT user\_id) FROM t\_user\_id\_info GROUP BY province ORDER BY COUNT(DISTINCT user\_id) desc

结果为





问题二

用户的性别分布是怎样的?

代码如下

SELECT province ,gender,COUNT(DISTINCT user\_id) FROM t\_user\_id\_info WHERE gender='男' GROUP BY province ,gender ORDER BY COUNT(DISTINCT user\_id) DESC

SELECT province ,gender,COUNT(DISTINCT user\_id) FROM t\_user\_id\_info WHERE gender='女' GROUP BY province ,gender ORDER BY COUNT(DISTINCT user\_id) DESC

结果如下



问题三

付费用户与未付费用户的分布是怎样的？

代码如下

SELECT CASE WHEN

a.is\_pay IS NOT NULL THEN '付费' ELSE '未付费'

END AS c,

COUNT(b.user\_id)

FROM

paid a

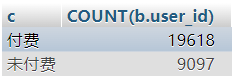
RIGHT JOIN t\_user\_id\_info b ON

a.user\_id = b.user\_id

GROUP BY

C

结果如下



问题四，按地域和性别分组，统计付费和未付费用户数

代码如下

SELECT

b.province,

b.gender,

CASE WHEN a.is\_pay IS NOT NULL THEN '付费' ELSE '未付费'

END AS '付费/未付费',

COUNT(b.user\_id)

FROM

paid a

RIGHT JOIN t\_user\_id\_info b ON

a.user\_id = b.user\_id

GROUP BY

'付费/未付费',

b.gender,

b.province

ORDER BY

COUNT(b.user\_id)

DESC

结果如下



问题五，按地域和性别分组，统计推荐表中的可推荐量

代码如下

SELECT

province,

gender,

cnt

FROM

`push\_cnt`

GROUP BY

gender,

province

ORDER BY

cnt

DESC

结果如下



## 11$KPG}X6HNVRZ[V1GIP@%H

## 数据的建模与商业决策

用户数排名前十的省份分别是，广东，苏州，山东，河南，四川，浙江，河北，安徽，广西湖南，湖北省。

男性用户数排名前五的省份是，广东，江苏，山东，四川，河南，女性用户数排名前五的省份是，江苏，广东，山东，河南，四川。

付费用户占用户数中的46%，付费用户排名前五的都是男性，且分布的省份为江苏，山东，四川，河南，浙江，付费女性分布省份前三有河南，河北，广西省。

可推荐量占前五的客户都是男性，其分布的省份为广东，江苏，山东，四川，河南省，可推荐量前三的女性分布的省份是山东，河南，广东。

### 决策

因此，我会优先向山东，江苏，四川，河南地域的男性用户和河南，河北，山东，广西的女性用户进行推送，因为这些地域的男客户购买力强，对产品的留存度高