## 1.项目目标

分析 APP 的用户属性和付费情况，以此制定产品消息推送的策略（给哪些用户推送消息，什么时候进行推送，推送怎样的消息可以带来活跃用户和付费）

## 2.项目任务

1. 用户的地域分布(分省按从大到小排列)是怎样的？
2. 用户的性别分布是怎样的?
3. 付费用户与未付费用户的分布是怎样的？
4. 付费用户与未付费用户的地域.性别分布是怎样的?
5. 根据付费用户的地域以及性别分布，设计对应的产品消息推送策略
    - 你会优先向哪些地域和性别的用户进行推送？ - 为什么？

## 3.数据清洗

在正式进行分析前需要对数据进行清洗
首先对数据进行检查，然后查找对应的数据问题，最后对数据进行清洗，一下为进行数据清洗所需代码
清洗用户表:

```
select distince length(user_id) from user_id_info #全部正常
select distinct(province) from user_id_info  #全部正常
select distinct(gender) from user_id_info
select user_id,gender from user_id_info where gender in ('男人','未知')
update user_id_info set gender='男' where gender='男人'
delete from user_id_info where gender='未知'
```

清洗付费表:

```
select distince length(user_id) from paid #全部正常
select distinct(is_pay) from paid  #全部正常,全都为 1，表明全部都为付费用户
```

## 4.问题解决

### 4.1 用户的地域分布(分省按从大到小排列)是怎样的？

1）构思所用的数据情况

```
select distinct province from user_id_info;
```

| 用户所在省份 | 用户人数 |
| --- | --- |
| 江苏省 | |
| 北京市 | |
| 河北省 | |
| 广东省 | |
| 上海市 | |

2）确定目标字段与目标数据表

目标字段：province，user_id，启动人数

目标数据表：用户表

3）确定要进行的数据操作和搭建代码框架

✧ 确定查询表为用户表

✧ 将数据按照省份进行聚合

✧ 将不同省份的数据按照用户数进行排序

✧ 查询人群，对应的人数

```
select
from
group by ……
order by …… desc
```

4)撰写代码

完善上述代码,最终代码为：

```
select province, count(distinct user_id) as number
from user_id_info
group by province
order by number desc
```

5)得出结论

基于运行结果：

| | |
| --- | --- |
| 广东省 | 1775 |
| 江苏省 | 1202 |
| 河北省 | 582 |
| 北京市 | 502 |
| 上海市 | 451 |

得出用户的地域分布情况为：从最终的图表结果可以看出，用户最多集中在广东省，其次是江苏省。在河北省，北京市和上海市该 APP 的用户较少。

**4.2 用户的性别分布是怎样的？**

1）构思所用的数据情况

```
select distinct gender from user_id_info;
```

| 性别 | 用户人数 |
| --- | --- |
| 男 | |
| 女 | |

2）确定目标字段与目标数据表
目标字段：gender，user_id，启动人数
目标数据表：用户表

3）确定要进行的数据操作和搭建代码框架
✧ 确定查询表为用户表
✧ 将数据按照性别进行聚合
✧ 将不同性别的数据按照用户数进行排序
✧ 查询不同性别的人群对应的人数

```
select
from
group by ……
order by …… desc
```

4)撰写代码
完善上述代码,最终代码为：

```
select gender, count(distinct user_id) as number
from user_id_info
group by gender
order by number desc
```

5)得出结论
基于运行结果：

| | |
| --- | --- |
| 男 | 2538 |
| 女 | 1974 |

基于运行结果可以发现，该 APP 的用户中以男性居多

**3.3 付费用户与未付费用户的分布是怎样的？**
1）构思所用的数据情况

| 用户付费情况 | 用户人数 |
| --- | --- |
| 付费用户 | |
| 未付费用户 | |

2）确定目标字段与目标数据表
目标字段： user_id，is_pay,启动人数
目标数据表：用户表，付费表
连接数据表的字段：user_id

3）确定要进行的数据操作和搭建代码框架

✧ 以用户表为主表进行左连接，连接付费表，连接字段为 user_id
✧ 将数据按照付费情况进行聚合
✧ 将不同付费情况的数据按照启动数进行排序
✧ 查询不同付费情况的人群对应的启动数

```
select
case
when then
else
end as
from
join
on
group by ……
order by …… desc
```

4)撰写代码
完善上述代码,最终代码为：

```
select count(a.user_id) as '启动量',
case
when b.user_id is not null then 'pay'
else 'not pay'
end as 'pay or not'
from user_id_info a
left join paid b
on a.user_id=b.user_id
group by b.is_pay
order by '启动量' DESC
```

5)得出结论
基于运行结果：

| pay or not | 启动量 |
|---|---|
| not pay | 4510 |
| pay | 2 |

基于运行结果可以发现，该 APP 的启动用户中绝大部分用户为非付费用户，付费用户所占比例极低。

### 3.4 付费用户与未付费用户的地域.性别分布是怎样的？

1）构思所用的数据情况

地域情况：

| 付费情况 | 地域 | 启动量 |
|---|---|---|
| 付费 | 广东省 | |
| 付费 | 北京市 | |
| …… | …… | |
| 非付费 | 广东省 | |

| | | |
|---|---|---|
| 非付费 | 北京市 | |
| …… | …… | |

性别分布情况：

| 付费情况 | 性别 | 启动量 |
|---|---|---|
| 付费 | 男 | |
| 付费 | 女 | |
| 非付费 | 男 | |
| 非付费 | 女 | |

2）确定目标字段与目标数据表
目标字段： user_id，is_pay, province, gender,启动人数
目标数据表：用户表，付费表
连接数据表的字段：user_id

3）确定要进行的数据操作和搭建代码框架
分为地域和性别两种情况分别进行
✧ 以用户表为主表进行左连接，连接付费表，连接字段为 user_id
✧ 将数据按照付费情况和省份或者按照付费情况和性别进行聚合
✧ 将不同付费情况的数据按照启动数进行排序
✧ 查询不同情况的人群对应的启动数

```
select
from
left join
on
group by ……
order by …… desc
```

4)撰写代码
完善上述代码,最终代码为：
地域部分：

```
select b.is_pay,a.province,count(a.user_id) as '启动量'
from user_id_info a
left join paid b
on a.user_id=b.user_id
group by b.is_pay,a.province
order by '启动量' DESC
```

性别部分：

```
select b.is_pay,a.gender,count(a.user_id) as '启动量'
from user_id_info a
left join paid b
on a.user_id=b.user_id
group by b.is_pay,a.gender
order by '启动量' DESC
```

5)得出结论

付费情况的地域分布：

| is_pay | province | 启动量 |
|---|---|---|
| *NULL* | 江苏省 | 1201 |
| *NULL* | 北京市 | 502 |
| *NULL* | 河北省 | 582 |
| *NULL* | 广东省 | 1774 |
| *NULL* | 上海市 | 451 |
| 1 | 广东省 | 1 |
| 1 | 江苏省 | 1 |

付费情况的性别分布：

| is_pay | gender | 启动量 |
|---|---|---|
| *NULL* | 女 | 1973 |
| *NULL* | 男 | 2537 |
| 1 | 男 | 1 |
| 1 | 女 | 1 |

**3.5 根据付费用户的地域以及性别分布，设计对应的产品消息推送策略？**

1.根据付费用户的地域分布情况，可以更多的向广东省和江苏省的用户进行推送。

```
select province,avg(cnt) as averagecnt from push_cnt group by province
order by averagecnt desc
```

同时根据推送表可以发现，广东省和江苏省的用户平均可接受的推荐量最高，因此综上，应该增多对这两个省的推送。

2.在性别方面，根据推送表中对男女时间平均可接受推荐量的比较，发现可以一定程度增加对男性的推送。

```
select gender,avg(cnt) as averagecnt from push_cnt group by gender
order by averagecnt desc
```

| gender | averagecnt_1 |
|---|---|
| 男 | 6282.6087 |
| 女 | 2556.5217 |