Week2　APP 用户及产品分析

1、用户的地域分布(分省按从大到小排列)是怎样的？

思路：从用户表中提取出用户所在城市的信息，对其进行统计分析得到相关分布信息

逻辑：

先观察数据，看是否需要进行处理，如去掉异常值或缺省值等。若需处理，处理完成后然后对 t_user_id_info 表根据 province 变量分组后使用 COUNT DISCTINCT 计算各省用户人数，再按降序排序，观察结果

SQL 代码：

①观察数据是否需要清洗

```
3  SELECT DISTINCT LENGTH(user_id) FROM t_user_id_info
4
```

信息　结果1　概况　状态

LENGTH(user_id)
32

```
1  # 观察数据是否需要清洗
2  SELECT DISTINCT province FROM t_user_id_info
3
4
5
6
```

信息　结果1　概况　状态

province
湖北省
广东省
辽宁省
安徽省
山东省
四川省
河北省
福建省
北京市
江西省
山西省
江苏省
浙江省
上海市
贵州省
河南省
重庆市
湖南省
广西壮族自治区
陕西省
黑龙江省

②按照所在地区分组计算人数，并按降序排列

查询创建工具 查询编辑器

```
4  #按照所在地区分组按降序排列
5  SELECT province,COUNT(DISTINCT user_id) as numbers FROM t_user_id_info
6  GROUP BY province
7  ORDER BY COUNT(DISTINCT user_id) DESC
8
9
```

信息 结果1 概况 状态

province	numbers
广东省	3143
江苏省	2810
山东省	2310
河南省	2064
四川省	2043
浙江省	1705
河北省	1363
安徽省	1319
广西壮族自治区	1203
湖南省	1120
湖北省	1013
北京市	963
上海市	907
辽宁省	899
重庆市	895
江西省	855
福建省	672
贵州省	618
黑龙江省	610
山西省	583
陕西省	576
云南省	570
吉林省	474

结论：由查询结果可得，用户数排名前 10 的省份分别是广东、江苏、山东、河南、四川、浙江、河北、安徽、广西壮族自治区和湖南，其中广东省拥有用户最多，用户数突破 3000 达到 3143，湖南省用户数也超过 1000.

2、用户的性别分布是怎样的?

思路：从用户表中提取出有关用户性别的信息，对其进行统计分析得到相关分布信息

逻辑：先观察数据，看是否需要进行处理，如去掉异常值或缺省值等。若需处理，处理完成后对 t_user_id_info 表根据 gender 变量分组后使用 COUNT DISCTINCT 计算各性别用户人数，再按降序排序，观察结果

SQL 代码：

①观察数据是否需要处理

```sql
#观察数据是否需要处理
SELECT DISTINCT gender,COUNT(DISTINCT user_id) as numbers
FROM t_user_id_info
GROUP BY gender
ORDER BY numbers DESC
```

信息 结果1 概况 状态

gender	numbers
男	20356
女	8179
未知	180

②按照所在性别分组观察分布

```sql
#按照所在地区分组按降序排列
SELECT gender,COUNT(DISTINCT user_id) as numbers FROM t_user_id_info
GROUP BY gender
ORDER BY numbers DESC
```

信息 结果1 概况 状态

gender	numbers
男	20356
女	8179
未知	180

结论：由查询结果可得，有 180 为用户不愿意告知性别，而在具体告知性别的用户中男性居多，有 20256 人，女性用户仅有 8179 人，不足男性用户的二分之一

3.付费用户与未付费用户的分布是怎样的？

思路：结合用户表和付费表，对其中有关付费用户和未付费用户的数据进行统计分析得到相关分布的信息

逻辑：使用 user_id 将用户表和付费表连接在一起，在 select 语句中用 CASE...WHEN 语句对 paid 与 unpaid 用户进行分组，然后使用 COUNT DISCTINCT 计算各组用户人数，按降序排序，观察结果

SQL 代码：

①观察数据是否需要处理

```sql
#观察数据是否需要处理
SELECT DISTINCT is_pay FROM paid
```

信息 结果1 概况 状态

is_pay
1

②连接用户表与付费表查看分布

```
SELECT (CASE WHEN is_pay=1 THEN '付费用户' ELSE '未付费用户' END) AS type,
COUNT(t_user_id_info.user_id) AS numbers
FROM t_user_id_info LEFT JOIN paid ON t_user_id_info.user_id=paid.user_id
GROUP BY type
```

信息 | 结果1 | 概况 | 状态

type	numbers
付费用户	19618
未付费用户	9097

结论：由查询结果可得，注册用户中付费用户居多，有 19618 位，而未付费用户有 9097 位

4.付费用户与未付费用户的地域，性别分布是怎样的?

思路：结合用户表和付费表，对其中有关付费用户和未付费用户的数据进行统计分析得到相关分布的信息

逻辑：

先对数据进行处理，为保护原始数据表不被破坏，将所需的数据筛选出来插入一个新表 paid_info，通过 user_id 将用户表和付费表连接在一起，使用 update 语句对 is_pay 字段进行更新，将 NULL 值更改为 0。最后对新表根据 paid、gender 和 province 变量分组后使用 COUNT DISCTINCT 计算各组用户人数，然后根据是否付费按人数降序排列，观察结果

SQL 代码：

①创建新表 paid_info

查询创建工具 | 查询编辑器

```
#创建新表paid_info
CREATE TABLE paid_info AS
SELECT a.user_id,gender,province,is_pay
FROM t_user_id_info AS a LEFT JOIN paid AS b ON a.user_id=b.user_id
```

信息 | 概况 | 状态

```
[SQL]CREATE TABLE paid_info AS
SELECT a.user_id,gender,province,is_pay
FROM t_user_id_info AS a LEFT JOIN paid AS b ON a.user_id=b.user_id

受影响的行: 28715
时间: 1.538s
```

```
6  #查看表paid_info
7  SELECT DISTINCT * FROM paid_info
8
```

信息 | 结果1 | 概况 | 状态

user_id	gender	province	is_pay
3e18d4b2d4ab941433bb	女	湖北省	1
f253f5b1a1f07075d21865	男	广东省	1
9ec676899fdffca77b5683	男	辽宁省	1
9f419cbf4212127588e6a4	男	安徽省	1
fce1332c01d9515b6cc508	男	湖北省	1
70fd3fb4a16a316557f616	男	四川省	1
4941b76df5cd8d6a88f1f0	女	河北省	1
02d349787279f831ace19c	男	福建省	1
669136617920500d3e3ff9	男	北京市	1
e99ee40d5aec445eeb575	男	江西省	1
728e0265ea6ebfe796389	男	安徽省	1
bfa2e37b5632a8dff74d9c	男	山西省	1
1642c5f72019565995161e	女	江苏省	1
d4c5943fe97eb5b2eb6cc	男	安徽省	1
0fa18d081ac69007754ab	男	四川省	1
4c2177bf6cb8bc743323d	男	安徽省	1
e7cacd2dba69d7d6ef9bb	男	山东省	1
7b382b578fd1a38b56fed	男	山西省	1
8889ed690cad4aef3f2008	男	北京市	1
77198db57e396141de91f	男	辽宁省	1
798861a33fb7243fe2701f	女	浙江省	1

SELECT DISTINCT * FROM paid_info | 只读 | 查询时间: 0.084s | 第 1 条记录 (共 28715 条)

②对 **is_pay** 字段进行更新

```
9  #is_pay字段进行更新
10 UPDATE paid_info SET is_pay='0'
11 WHERE is_pay IS NULL
12
13
```

信息 | 概况 | 状态

```
[SQL]UPDATE paid_info SET is_pay='0'
WHERE is_pay IS NULL

受影响的行: 9097
时间: 0.288s
```

```
13  #查看表paid_info
14  SELECT DISTINCT * FROM paid_info ORDER BY is_pay
```

信息 | 结果1 | 概况 | 状态

user_id	gender	province	is_pay
0000e220fd999bf3490a7a	女	四川省	0
0002de862351668c6270d	女	广东省	0
001364b505e22d13d3da8	男	广东省	0
0025b77258b152ae81b84	男	北京市	0
0034bdd354e77a180e3fe	男	广西壮族自治区	0
0037bf7e679779fefe612b	女	江苏省	0
003c93df1bc86c7277507(	男	四川省	0
004736f80829ea686275d	女	湖南省	0
004b2afe7ed8eeec6ea94	女	山东省	0
004b661f7ba8f3b498945	女	浙江省	0
004f04e8764e8b206a2b5	男	上海市	0

③查看付费用户的分布

```
16  #查看付费用户的分布
17  SELECT (CASE WHEN is_pay='1' THEN '付费用户' END) AS type,
18  province,gender,COUNT(DISTINCT user_id) AS numbers
19  FROM paid_info
20  WHERE is_pay='1'
21  GROUP BY province,gender
22  ORDER BY numbers DESC
```

信息 | 结果1 | 概况 | 状态

type	province	gender	numbers
付费用户	广东省	男	1694
付费用户	江苏省	男	1424
付费用户	四川省	男	1155
付费用户	山东省	男	1053
付费用户	浙江省	男	929
付费用户	河南省	男	892
付费用户	安徽省	男	681
付费用户	河北省	男	602
付费用户	广西壮族自治区	男	595
付费用户	上海市	男	577
付费用户	湖南省	男	575
付费用户	北京市	男	528
付费用户	广东省	女	482
付费用户	湖北省	男	474
付费用户	江苏省	女	465
付费用户	重庆市	男	460

④查看未付费用户的分布

```
24  #查看微付费用户的分布
25  SELECT (CASE WHEN is_pay='0' THEN '未付费用户' END) AS type,
26  province,gender,COUNT(DISTINCT user_id) AS numbers
27  FROM paid_info
28  WHERE is_pay='0'
29  GROUP BY province,gender
30  ORDER BY numbers DESC
```

信息 | 结果1 | 概况 | 状态

type	province	gender	numbers
未付费用户	广东省	男	635
未付费用户	江苏省	男	552
未付费用户	山东省	男	504
未付费用户	河南省	男	435
未付费用户	四川省	男	377
未付费用户	江苏省	女	351
未付费用户	河南省	女	343
未付费用户	山东省	女	338
未付费用户	浙江省	男	322
未付费用户	广东省	女	315
未付费用户	安徽省	男	271
未付费用户	河北省	男	270
未付费用户	北京市	男	238
未付费用户	湖南省	男	237
未付费用户	广西壮族自治区	男	226
未付费用户	湖北省	男	215

结论：由查询结果可得，付费用户中，男性用户的人数显著较多，其中根据所在区域和性别分类后所得的用户数排名前十的地区（广东、江苏、四川、山东、浙江、河南、安徽、河北、广西壮族自治区和上海市）均为男性用户，广东省拥有高达 1694 的男性用户，而广东省也拥有最多的女性付费用户，有 482 人；而未付费用户中，广东省男性用户的用户数依旧是最多的，有 635 位，其次是江苏和山东省的男性用户，分别有 552 和 504 位。女性未付费用户在江苏省最多，有 351 人，在整个未付费分组中排名第 6。

5.你会优先向哪些地域和性别的用户进行推送？为什么？

我会选择优先向广东、江苏、四川和山东省的男性用户进行推送，因为根据 Q4 的查询结果，他们在付费与未付费分组中的人数都排名靠前，说明在他们之中，无论是已经开拓了的市场和潜在的市场都是巨大的，向他们进行推送可以在吸引有购买习惯的用户的同时吸引潜在的未付费用户前来消费。