**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？**

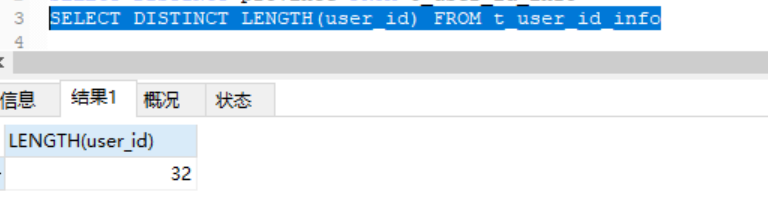
**思路：**从用户表中提取出用户所在城市的信息，对其进行统计分析得到相关分布信息

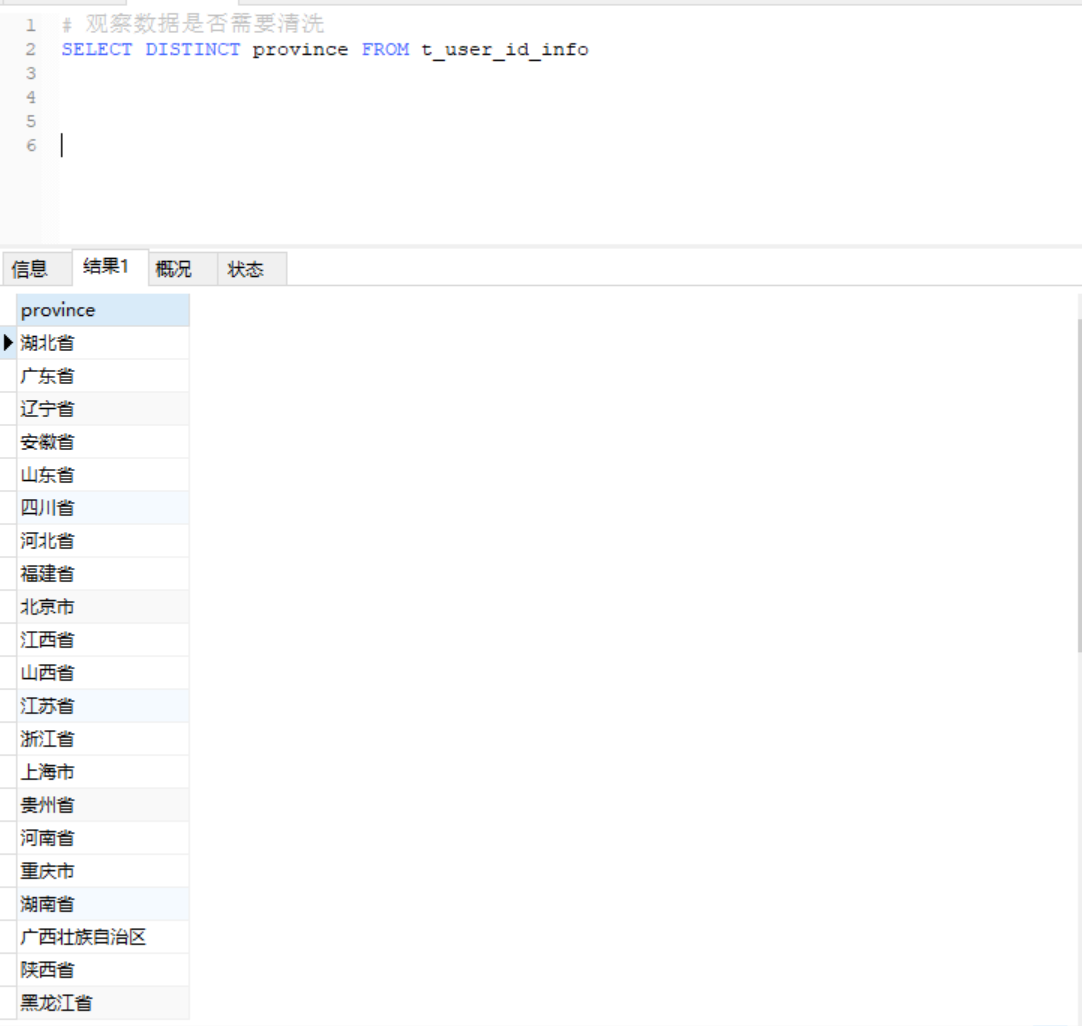
**逻辑：**

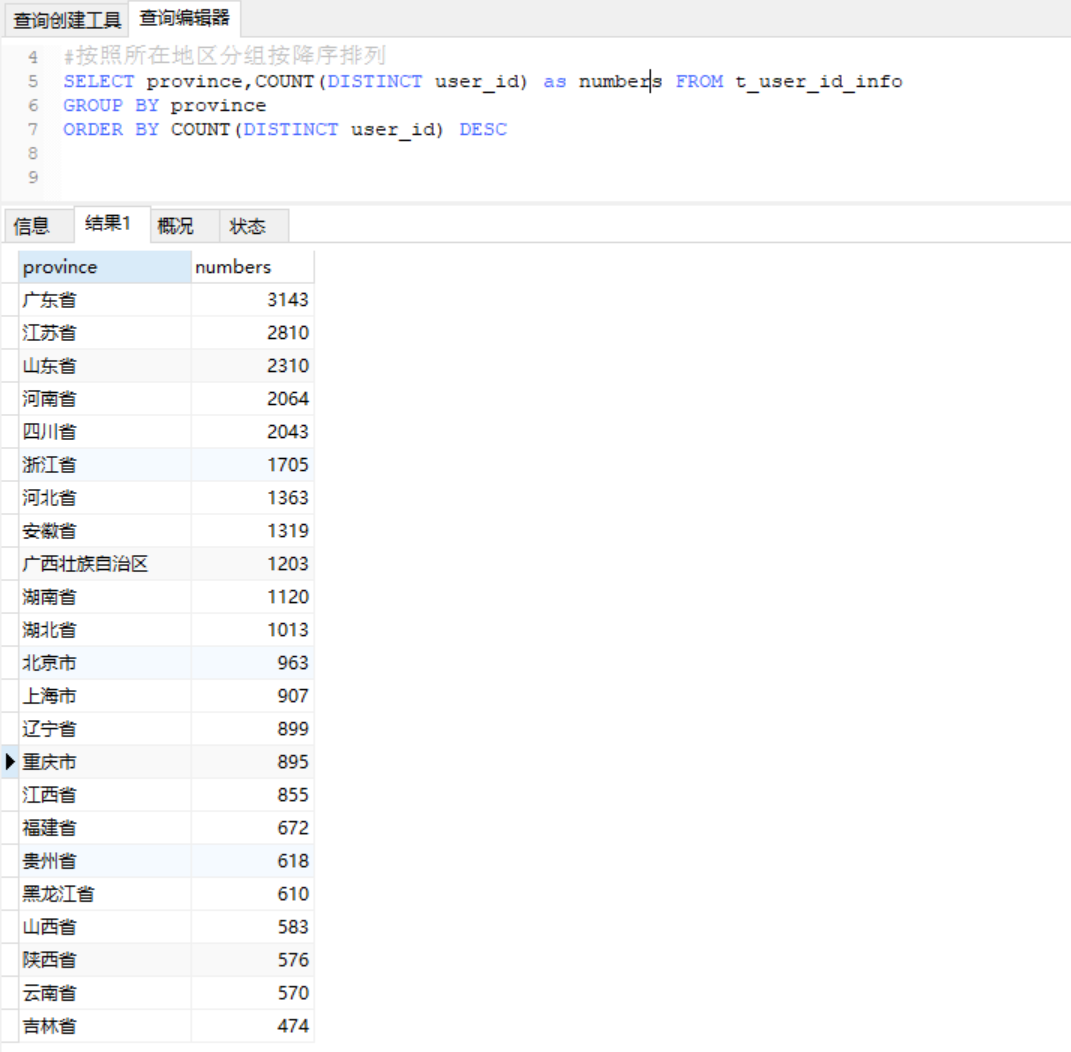
先观察数据，看是否需要进行处理，如去掉异常值或缺省值等。若需处理，处理完成后然后对t\_user\_id\_info表根据province变量分组后使用COUNT DISCTINCT计算各省用户人数，再按降序排序，观察结果

**SQL代码：**

**①观察数据是否需要清洗**





**②按照所在地区分组计算人数，并按降序排列**

**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.

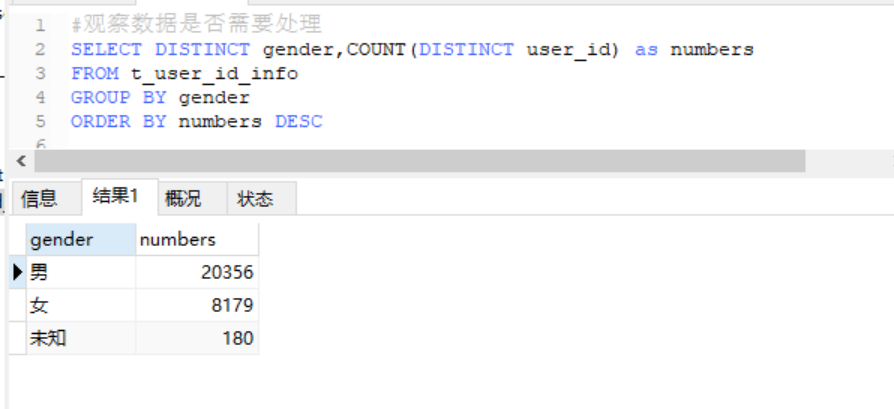
**2、用户的性别分布是怎样的?**

**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

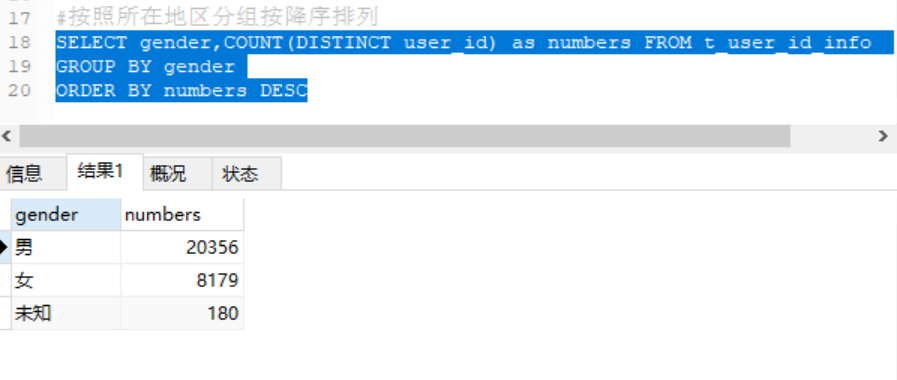
**逻辑：**先观察数据，看是否需要进行处理，如去掉异常值或缺省值等。若需处理，处理完成后对t\_user\_id\_info表根据gender变量分组后使用COUNT DISCTINCT计算各性别用户人数，再按降序排序，观察结果

**SQL代码：**

**①观察数据是否需要处理**



**②按照所在性别分组观察分布**



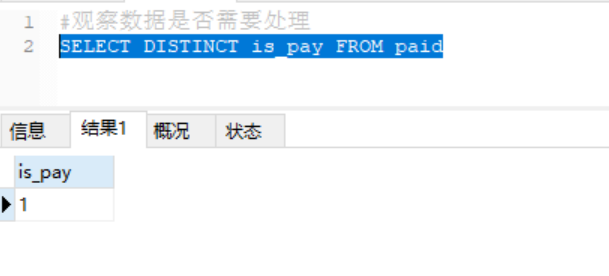
**结论：**由查询结果可得，有180为用户不愿意告知性别，而在具体告知性别的用户中男性居多，有20256人，女性用户仅有8179人，不足男性用户的二分之一

**3.付费用户与未付费用户的分布是怎样的？**

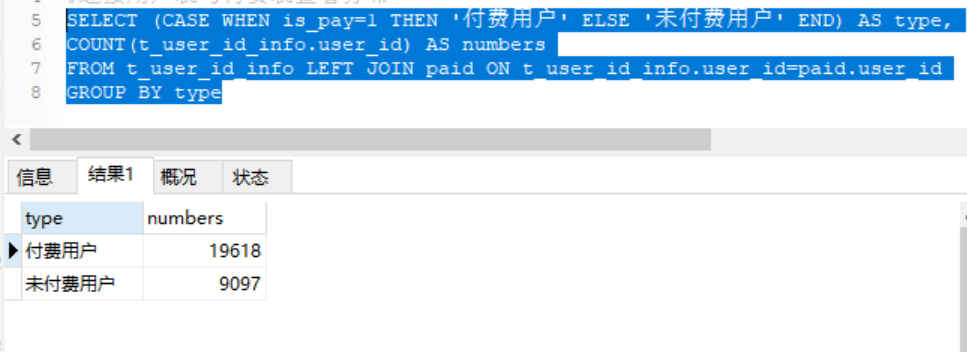
**思路：**结合用户表和付费表，对其中有关付费用户和未付费用户的数据进行统计分析得到相关分布的信息

**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**

②连接用户表与付费表查看分布



**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

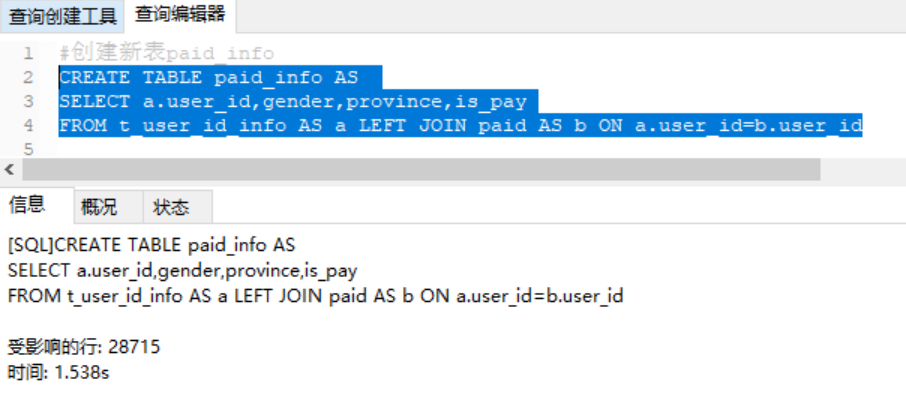
**4.付费用户与未付费用户的地域，性别分布是怎样的?**

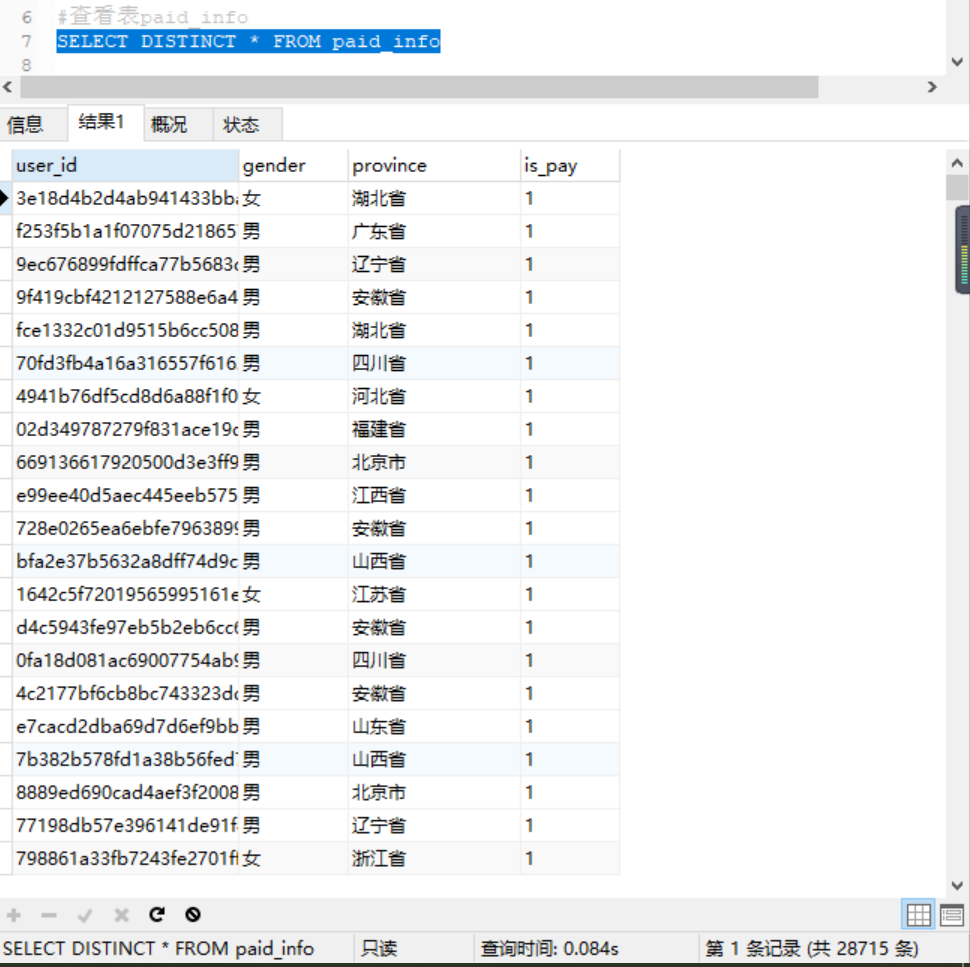
**思路：**结合用户表和付费表，对其中有关付费用户和未付费用户的数据进行统计分析得到相关分布的信息

**逻辑：**

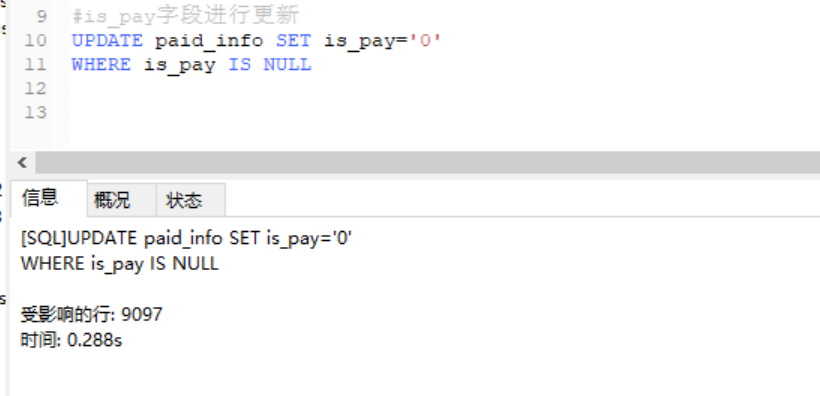
先对数据进行处理，为保护原始数据表不被破坏，将所需的数据筛选出来插入一个新表paid\_info，通过user\_id将用户表和付费表连接在一起，使用update语句对is\_pay字段进行更新，将NULL值更改为0。最后对新表根据paid、gender和province变量分组后使用COUNT DISCTINCT计算各组用户人数，然后根据是否付费按人数降序排列，观察结果

**SQL代码：**

**①创建新表paid\_info**



**②对is\_pay字段进行更新**

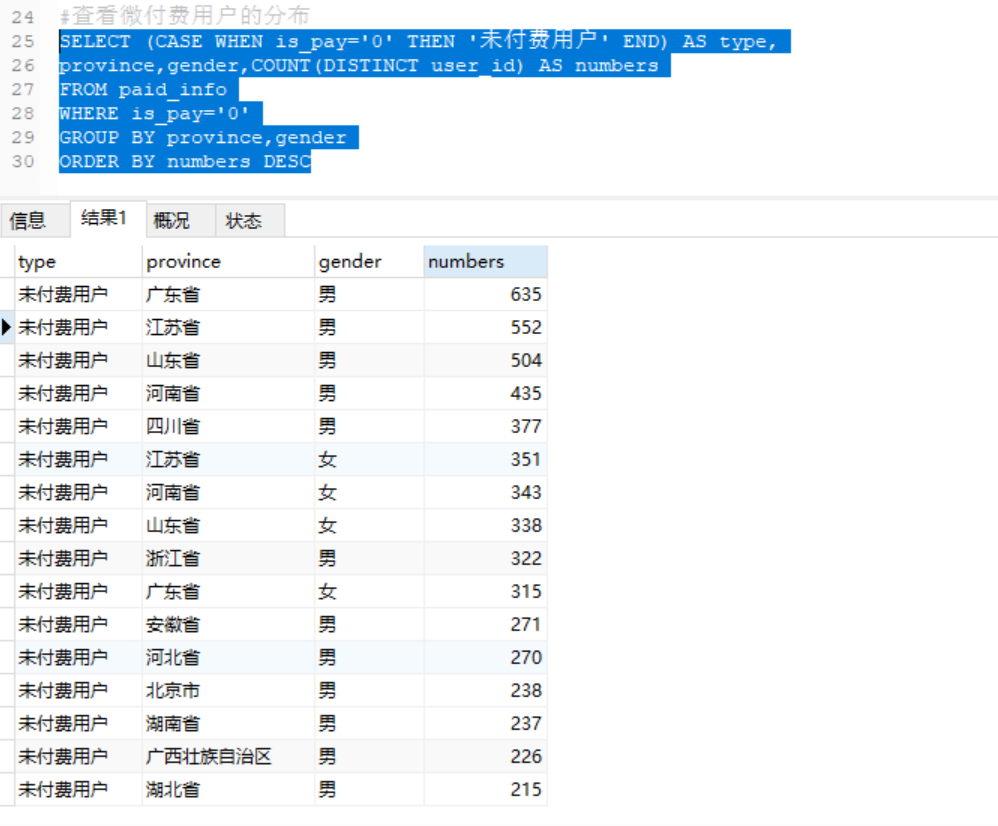




**③查看付费用户的分布**



**④查看未付费用户的分布**



**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。

**6. 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**