世纪佳缘用户画像

技术/设计/汇报：LucasX

NOTE

```
SELECT COUNT(*) AS '总人数' FROM candidates
```

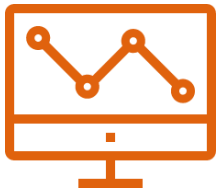
总人数
718395

通过Python爬虫抓取世纪佳缘婚恋网站71.8万条22-28岁年轻用户的信息进行分析
使用R、MySQL等进行统计、机器学习预测 eCharts3、R ggplot2、Excel2016 进行可视化
所有代码、数据、分析结果全部开源
仅作分析之用　不涉及任何商业用途以及可能被过分解读的地域歧视

Content

术语解释

技术解析



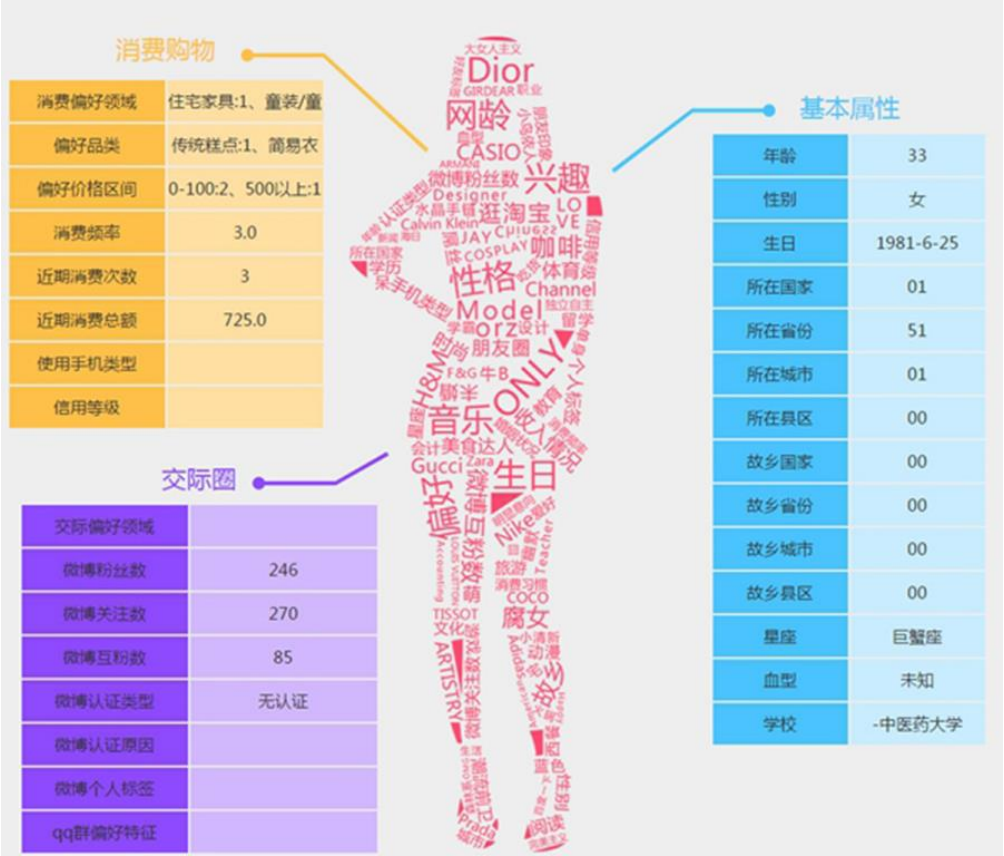
可视化

The Next

术语解释

用户画像

Persona是真实用户的虚拟代表,是建立在一系列真实数据(Marketing data, Usability data)之上的目标用户模型



价值

精准营销

当给各个用户打上各种“标签”之后，广告主（店铺、商家）就可以通过标签圈定他们想要触达的用户，进行精准的广告投放

助力产品

用户画像能帮助产品经理透过用户行为表象看到用户深层的动机与心理

用户研究

人群的消费偏好趋势分析、高端用户青睐品牌分析、不同地域品类消费差异分析等等

技术解析

网络爬虫

爬虫是一种按照一定的规则，自动地抓取万维网信息的程序或者脚本，本例采用Python 3.5进行编写

R ggplot2

ggplot2是R语言高级绘图包

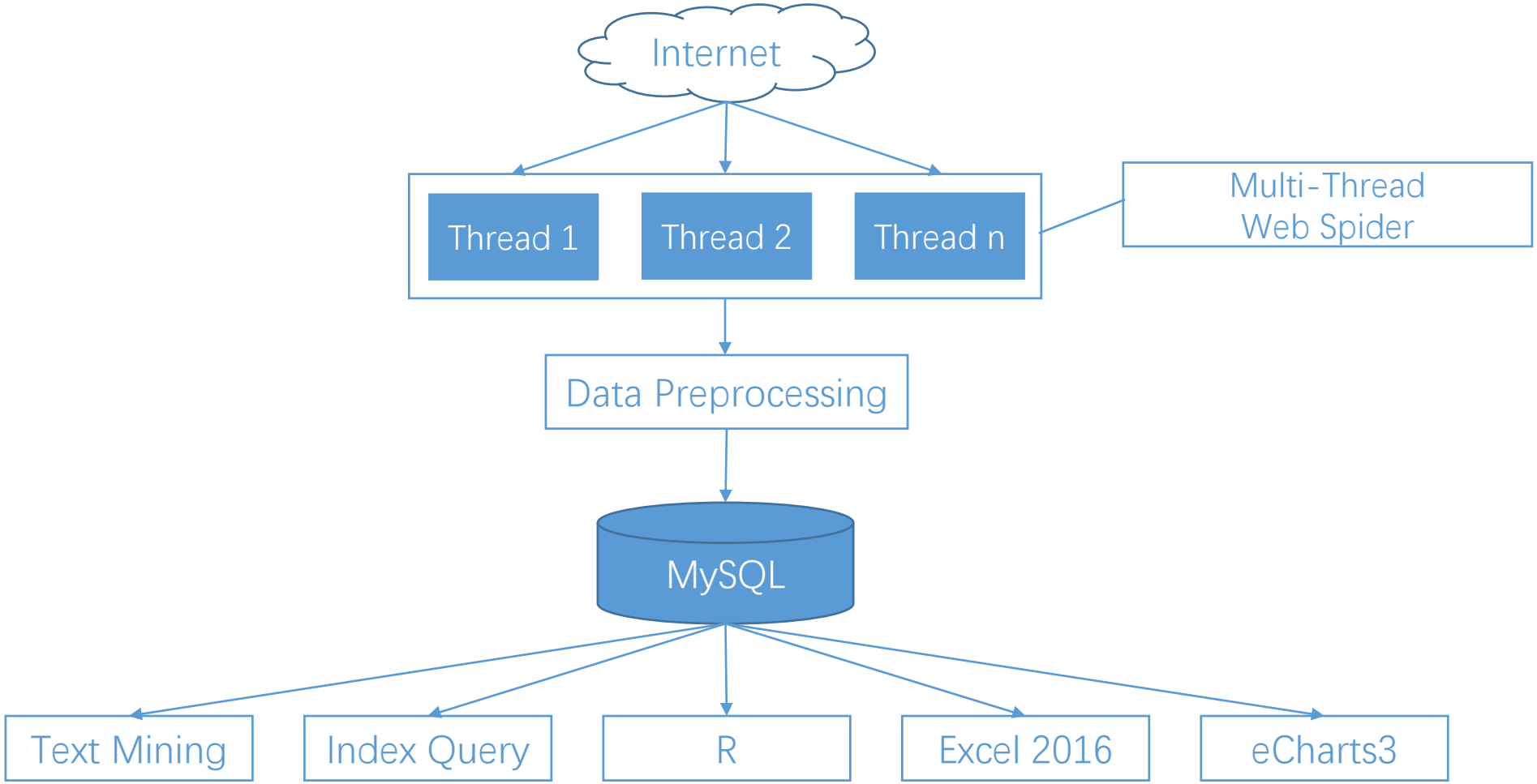
MySQL

MySQL是由Oracle公司开发的一套关系型数据库管理系统(RDBMS)，采用标准的SQL语法进行CRUD操作，目前被广泛应用在互联网信息存储中。本例部分采用Java myBatis进行数据处理

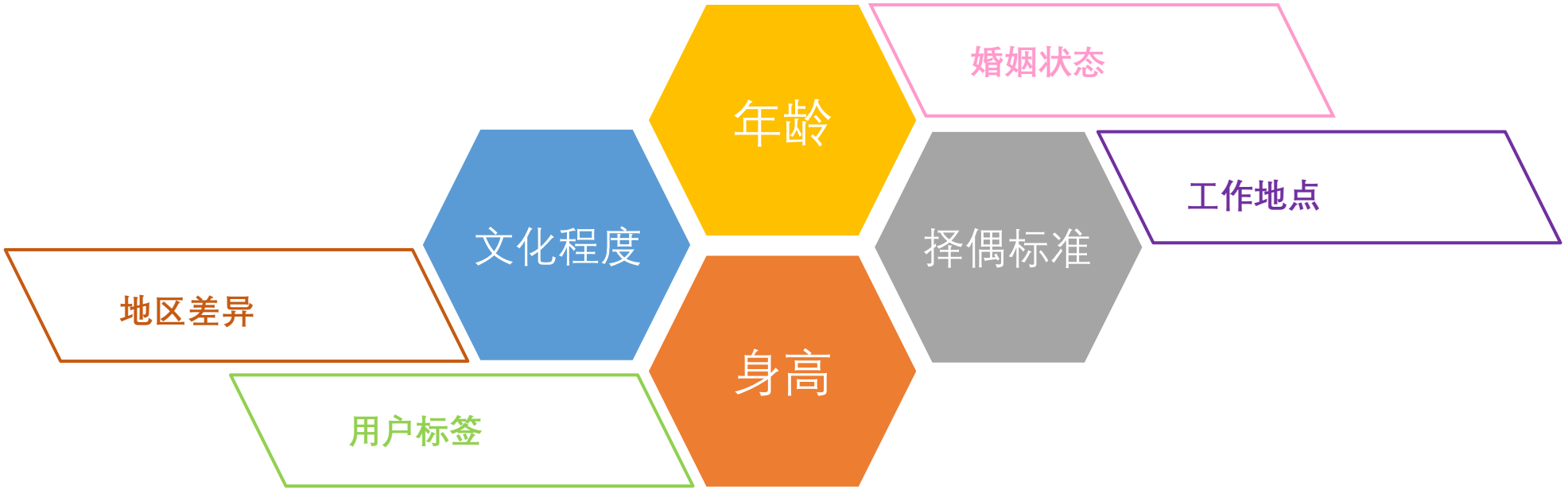
eCharts3

eCharts3是由百度前端团队打造的开源Web前端数据可视化组件，目前已成为BI和各大数据公司可视化的标准Web组件

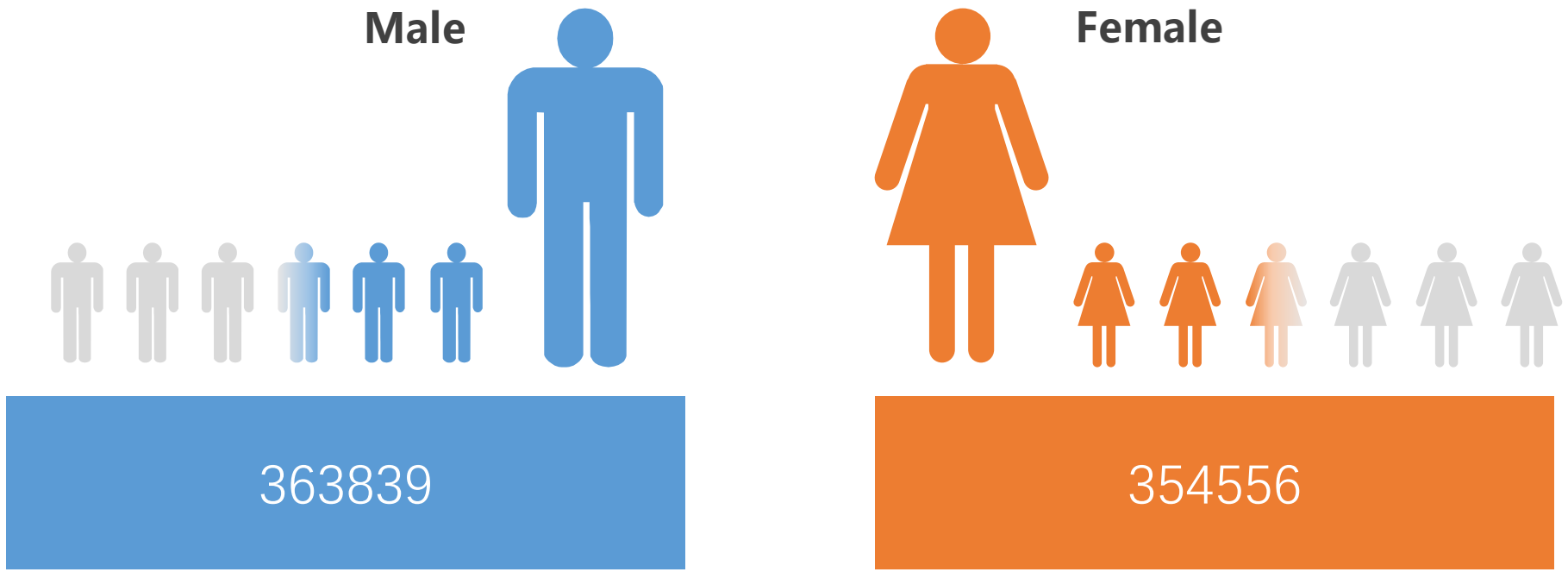
技术解析


Internet
Thread 1
Thread 2
Thread n
Multi-Thread
Web Spider
Data Preprocessing
MySQL
Text Mining
Index Query
R
Excel 2016
eCharts3

指标体系

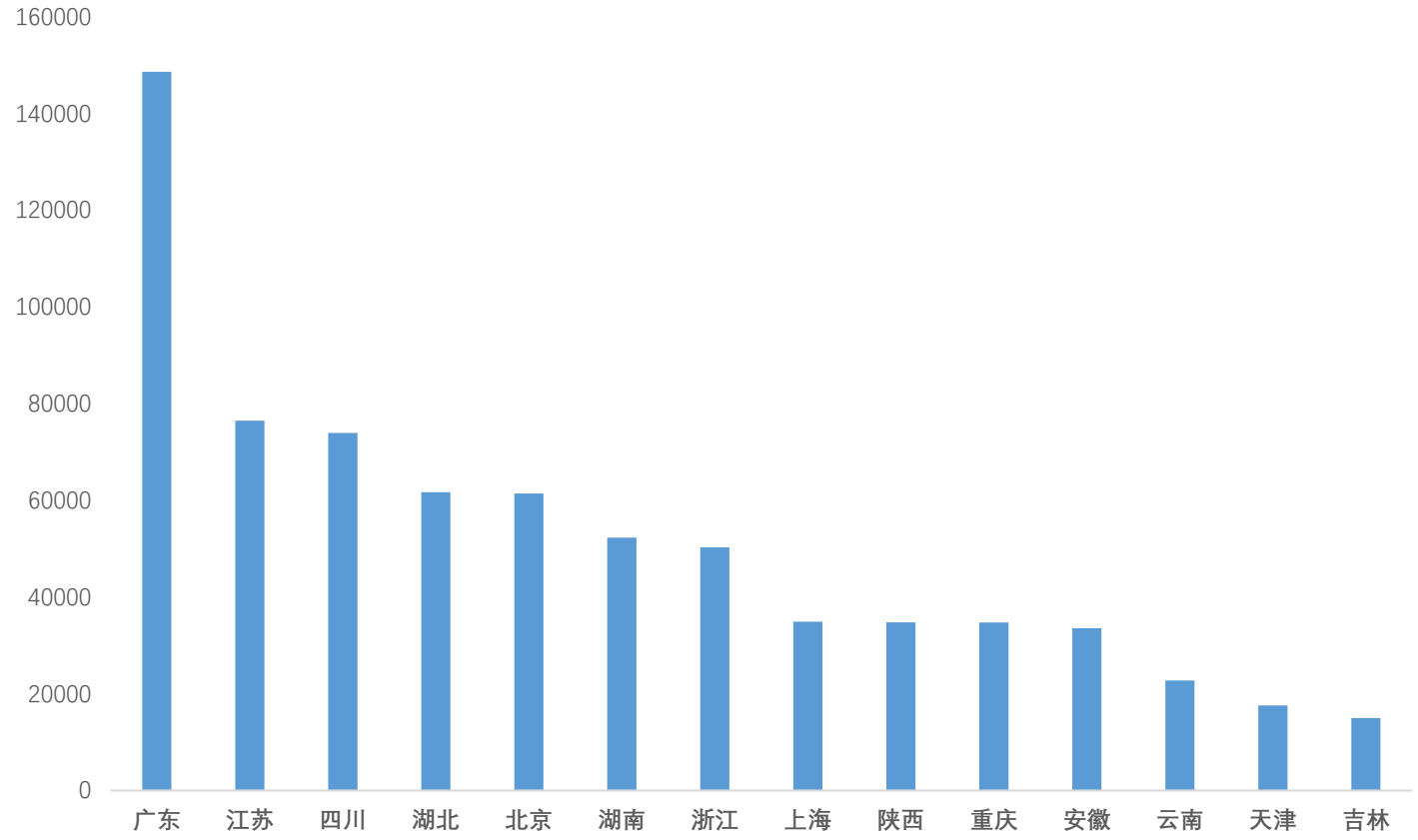


男女用户对比



男女用户比例基本接近1：1

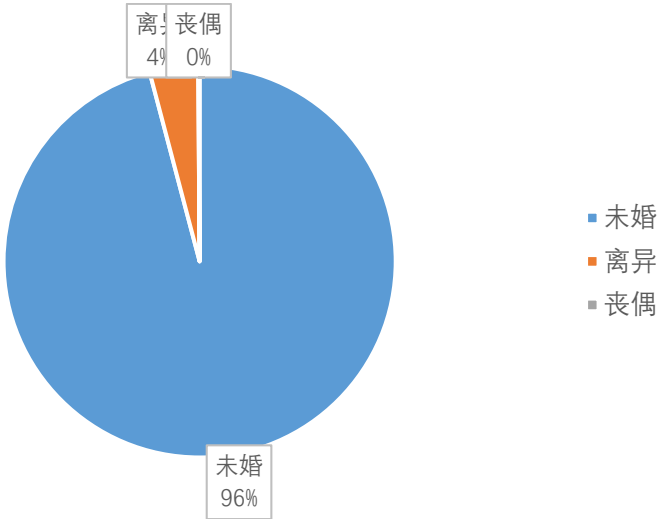
地区人数分析

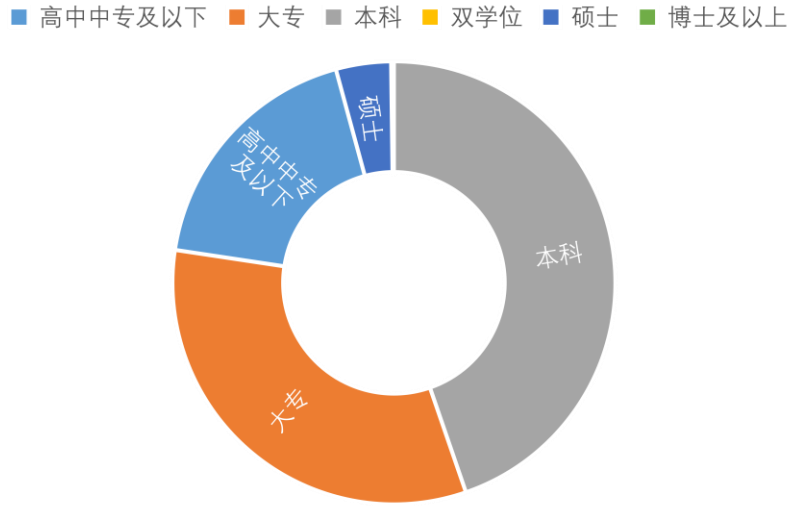


广东 江苏 四川相亲人数最多
吉林 天津 云南相亲人数最少

一线城市“不愁嫁娶”的天津

婚姻与学历

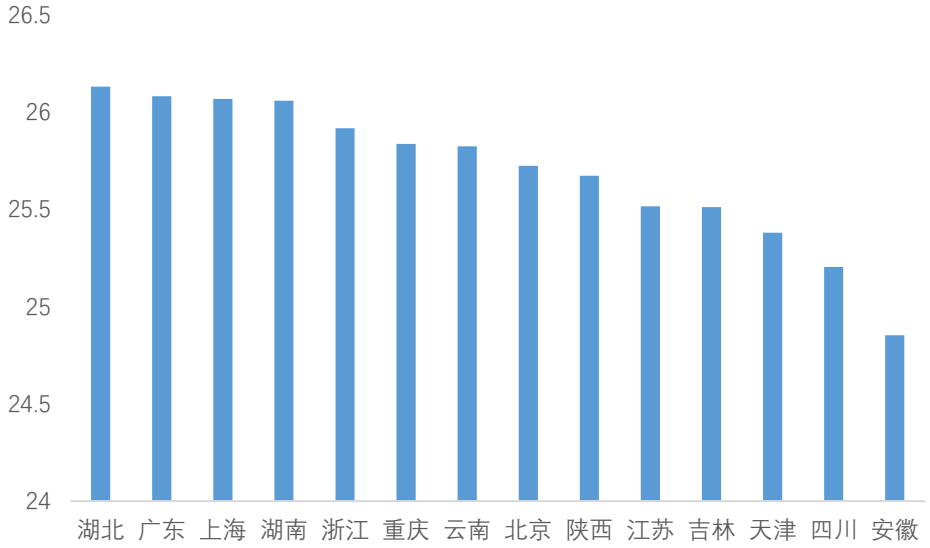




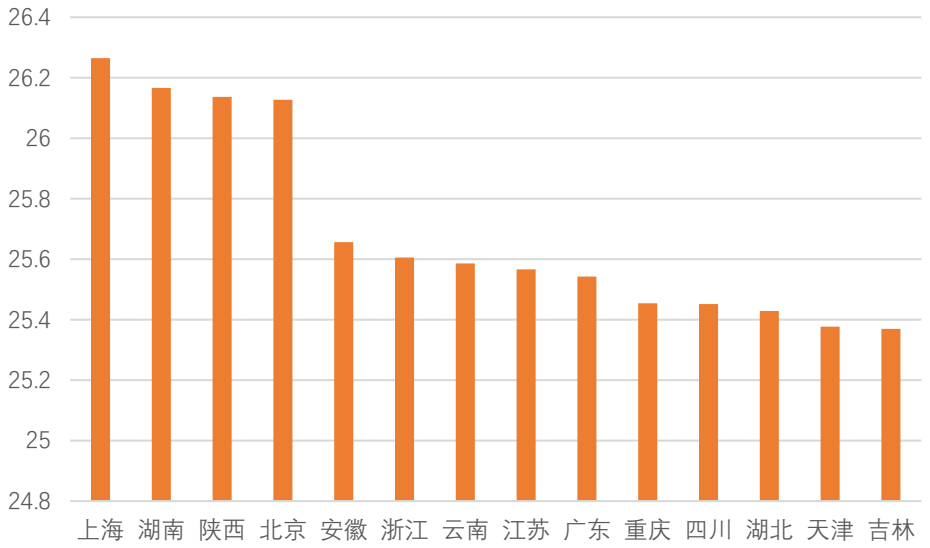
未婚群体占比96%
离异群体占比4%

接受过高等教育的群体占比约3/4
爬取的72W年轻用户文化程度较高

地区平均年龄分析

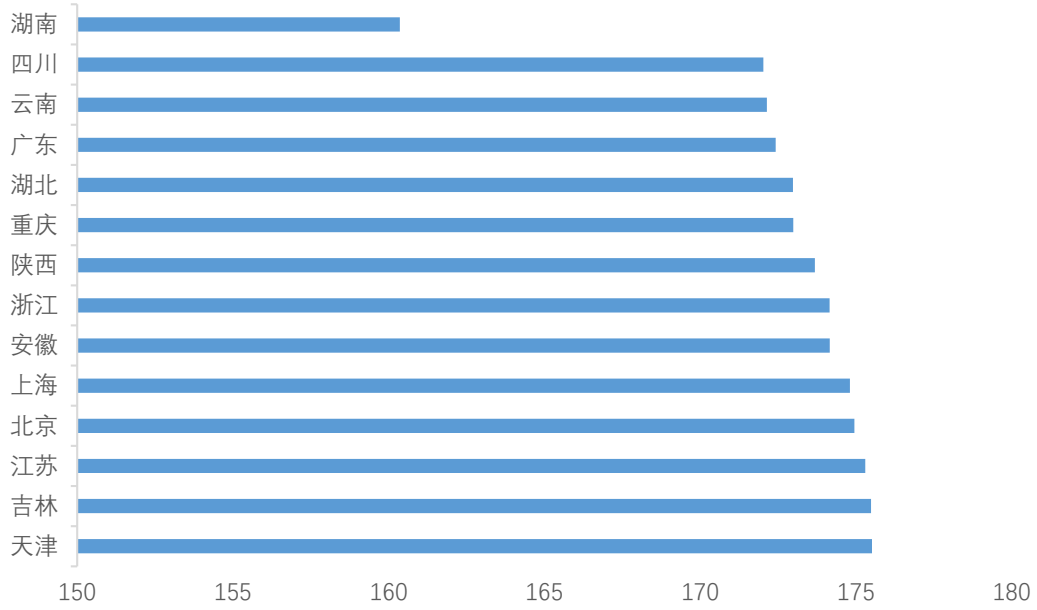


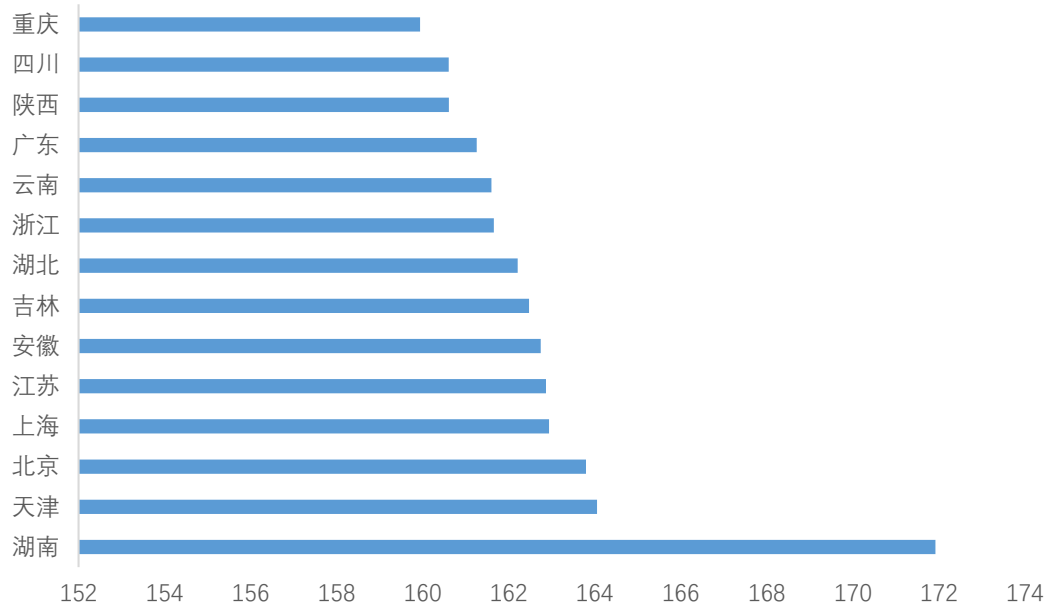
男性一般在25.8岁开始被催婚
湖北 广东 上海男性结婚较晚
安徽 四川 天津男性结婚较早



女性一般在25.7岁开始被催婚
上海 湖南 陕西女性偏晚
吉林 天津 湖北女性偏早

地区身高





天津 吉林 江苏等地男性较高
湖南 四川 云南等地男性较矮
男性平均身高为172.58cm

湖南 天津 北京女性较高
重庆 四川 陕西女性较矮
女性平均身高为162.57cm

技术解析

TF-IDF

TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF-IDF用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

$$TF = \frac{n_{ij}}{\sum_k n_{kj}} \qquad IDF = log \frac{|D|}{\{j: t_i \in d_j\}}$$

$$TF - IDF = TF*IDF$$

技术解析

Word Cloud

“词云”是对网络文本中出现频率较高的“关键词”予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使浏览网页者只要一眼扫过文本就可以领略文本的主旨。

常见的词云制作软件(包)有：Wordle Tagul R(wordcloud wordcloud2)
Baidu eChatrs d3.js etc.

择偶标准

对湖北、北京、四川的女性用户“择偶标准”进行分词，计算TF-IDF值，得到可视化字符云

对女人来说，“喜欢她”永远是第一位

简单 善良 孝顺 真心 逛街 幸福 生活是女人期待另一半拥有的品质

尽管身处物欲横流的社会

但大多数女性依然对纯真的爱情有着美好的憧憬

择偶标准

对湖北、北京、四川的男性用户“择偶标准”进行分词，计算TF-IDF值，得到可视化字符云

对男人来说

爱生活 喜欢他 真心 撒娇 孝顺 包容则是自己最关注的

在外打拼工作时，大多数男性希望另一半能作为港湾

自我评价

在提取35.2W条女性用户的自我简介进行分词之后

大多数女性都认为自己善良 简单 爱生活 孝顺

自我评价

在提取36.3W条男性用户的自我简介进行分词之后

大多数男性都希望能遇见倾心 爱生活 有责任心 有共同爱好的女性为伴

One more thing

除了基本信息及其分析、可视化
网络爬虫还抓取了**50万**头像图片数据集
利用**深度学习、计算机视觉**技术去训练该网络爬虫
使其具备“颜值偏好”
从而自动过滤掉不符合用户颜值喜好的用户
以提供更精准的推荐系统

One more thing

对分析报告或源代码中可能存在的错误
欢迎批评与指正!
对机器学习/深度学习/人工智能/计算机视觉/
NLP/推荐系统/搜索/数据挖掘/数据可视化
等领域感兴趣的童鞋一起交流

Github repository : https://github.com/lucasxlu/JiaYuan.git


Thanks
爬虫/分析/可视化/PPT制作/视觉设计：LucasX