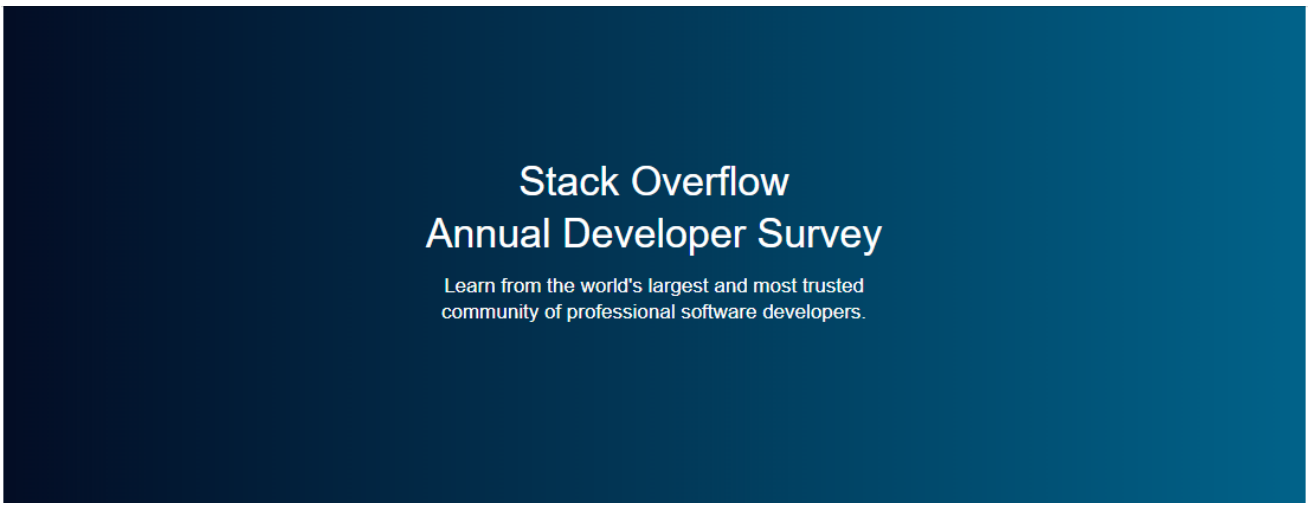
Stack Overflow 用户画像分析



项目介绍

本次项目的灵感来自于互联网，Stack Overflow 每年都会对其用户进行线上调研，从2011至2018，已持续了8年。

Stack Overflow2018年度开发者调查是针对183个国家和地区，获得了超过100,000个用户的回复，是对软件开发人员进行的最全面的调研，调研了开发人员从职业满意度、求职到教育、编码偏好各个方面的经验。

项目目的

本次项目主要利用Stack Overflow2018年近10万——98855个用户调研数据，以实现Stack Overflow的用户画像分析。Stack Overflow（简称Stack）想利用用户画像来吸引广告商的注意，以及投放广告以获得推广效果，目前需解决以下问题：

- **什么类型的广告适合投放在Stack？**
- **Stack在什么类型的平台发布广告可以使自己的推广效果达到最大呢？**

项目分析过程

**本项目主要用Python进行数据清洗，整合以及数据分析探索，使用Tableau进行数据可视化，使用Typora撰写Markdown数据分析报告。*

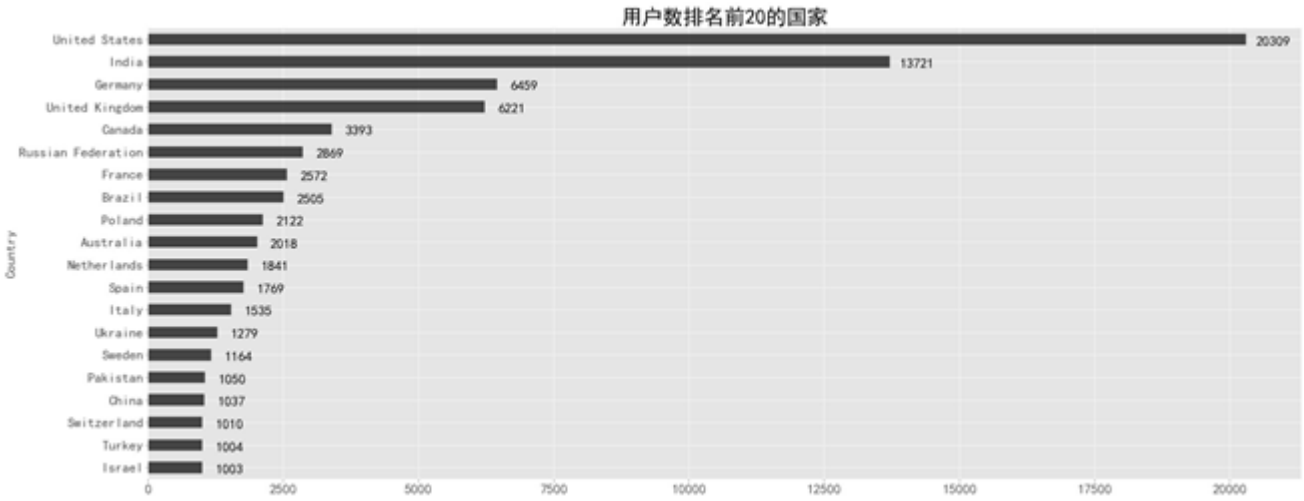
Stack此次的用户调研一共涉及128个问题，相当于128个特征，其中部分特征对项目的分析目的没有针对性，因此首先对特征进行筛选。

1.来自什么国家？

**共98443个受访者回答*

Stack的用户分布于全世界的183个国家。从世界范围来看，美国在Stack的用户数最多且占比超过20%，由于美国的计算机技术占领先地位，用户排名第一也是在合理的解释范围。其中印度的IT行业也是非常发达，其用户数在是stcak排名第二，且是第三的德国人数的2倍。中国排名17，如果没有网络安全的限制，中国的排名应该更高。

可以看出，Stack 的用户主要分布在北美，印度，欧洲等IT技术较发达的国家。

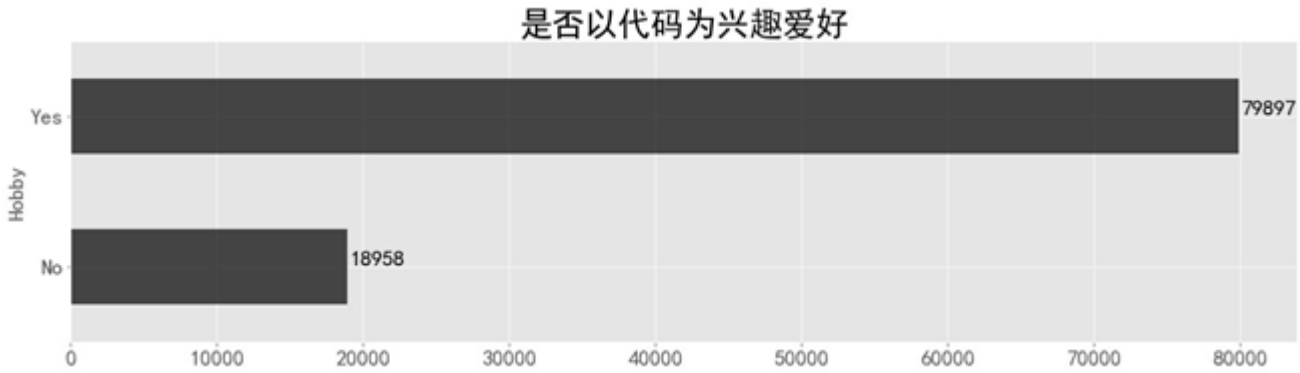


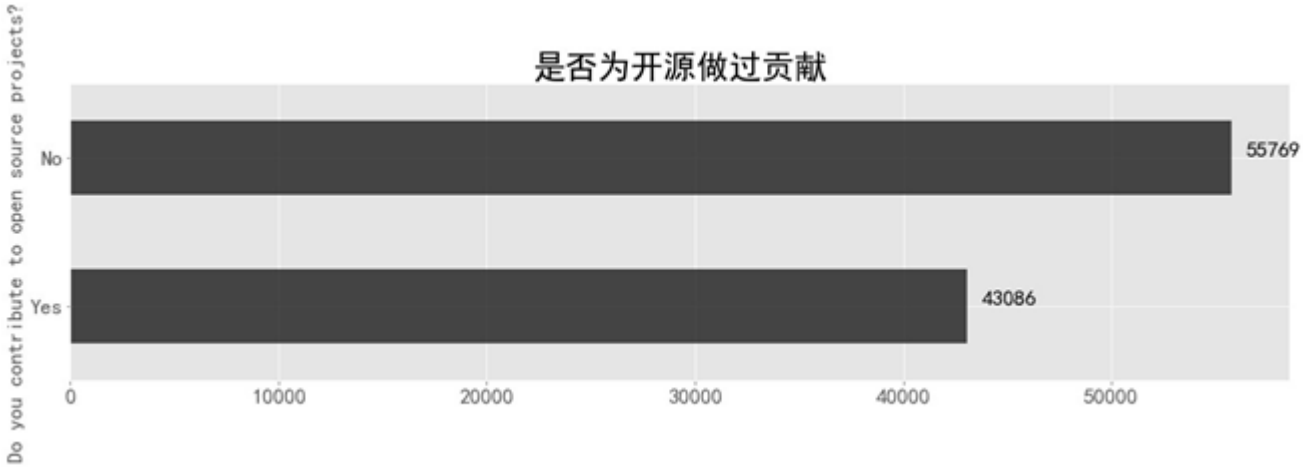
2.写代码是否是用户的兴趣？是否对开源项目做过贡献？

**均有98855个回答*

Stack的用户中大部分以写代码为兴趣，说明他们平时对IT领域的信息会比较关注。

Stack的用户中有43.6%的用户为代码开源做过贡献，这部分用户的技术能力相对较强，开源代码的门槛相对较高。

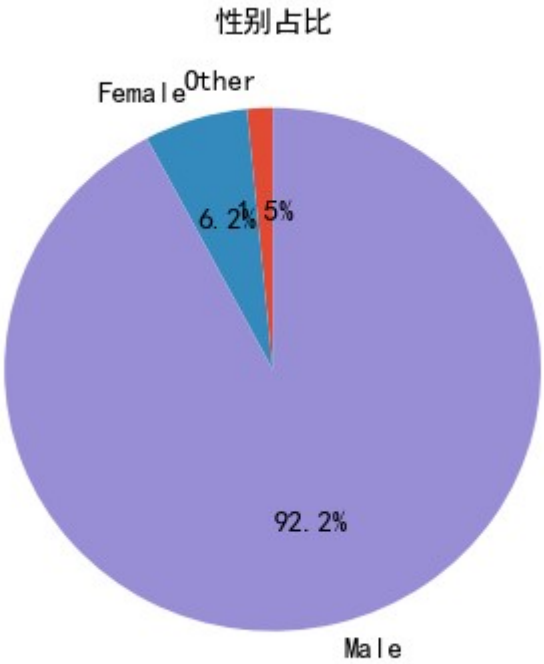




3.用户的性别，年龄以及编程年龄分布

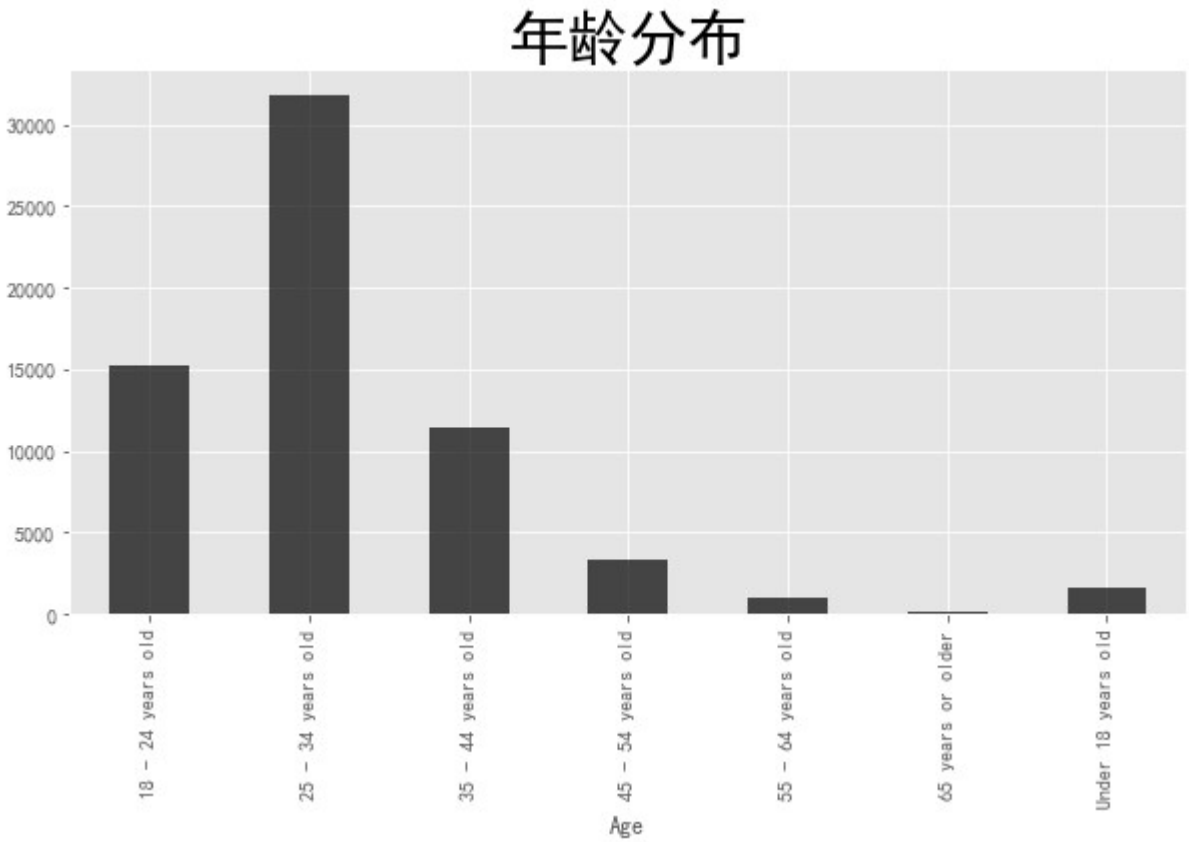
**共64469个回答*

Stack的用户男性占92.2%，女性仅占6.2%，说明对代码关心的男性远多于女性。



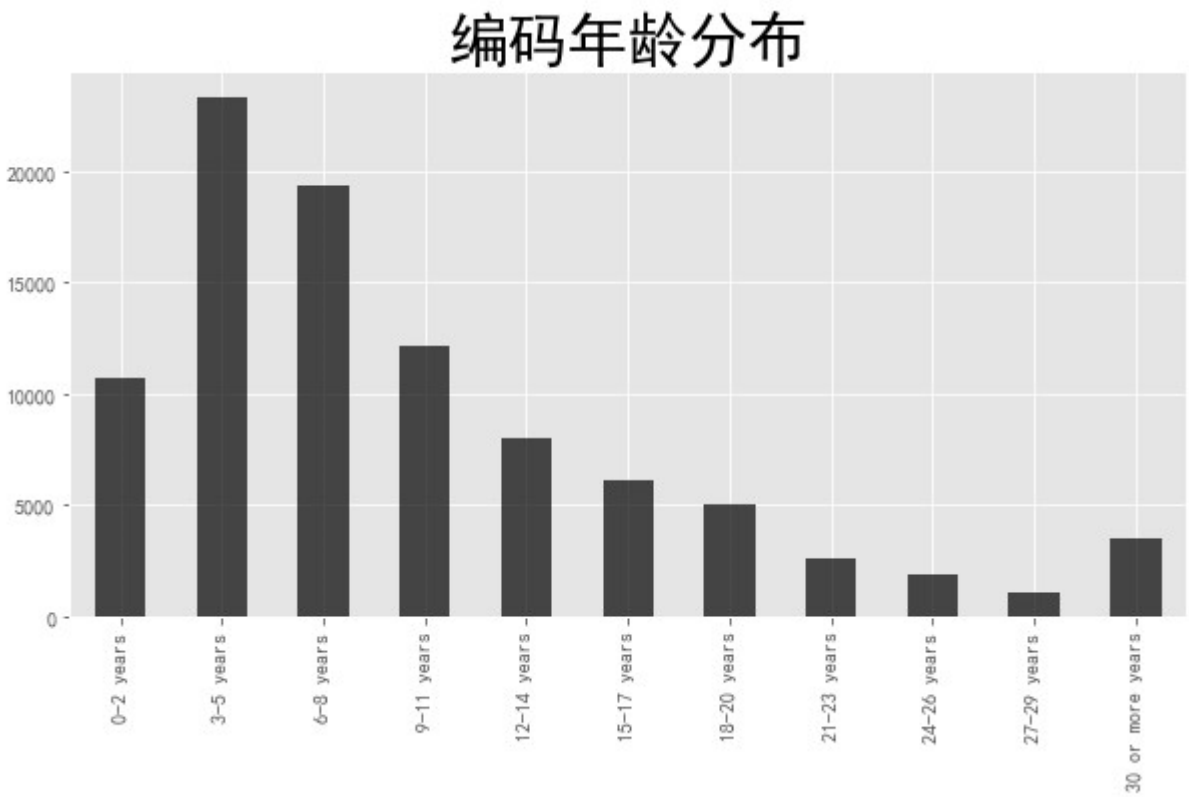
**共64574个回答*

用户年龄主要集中在25-30岁，以青年为主要力量。



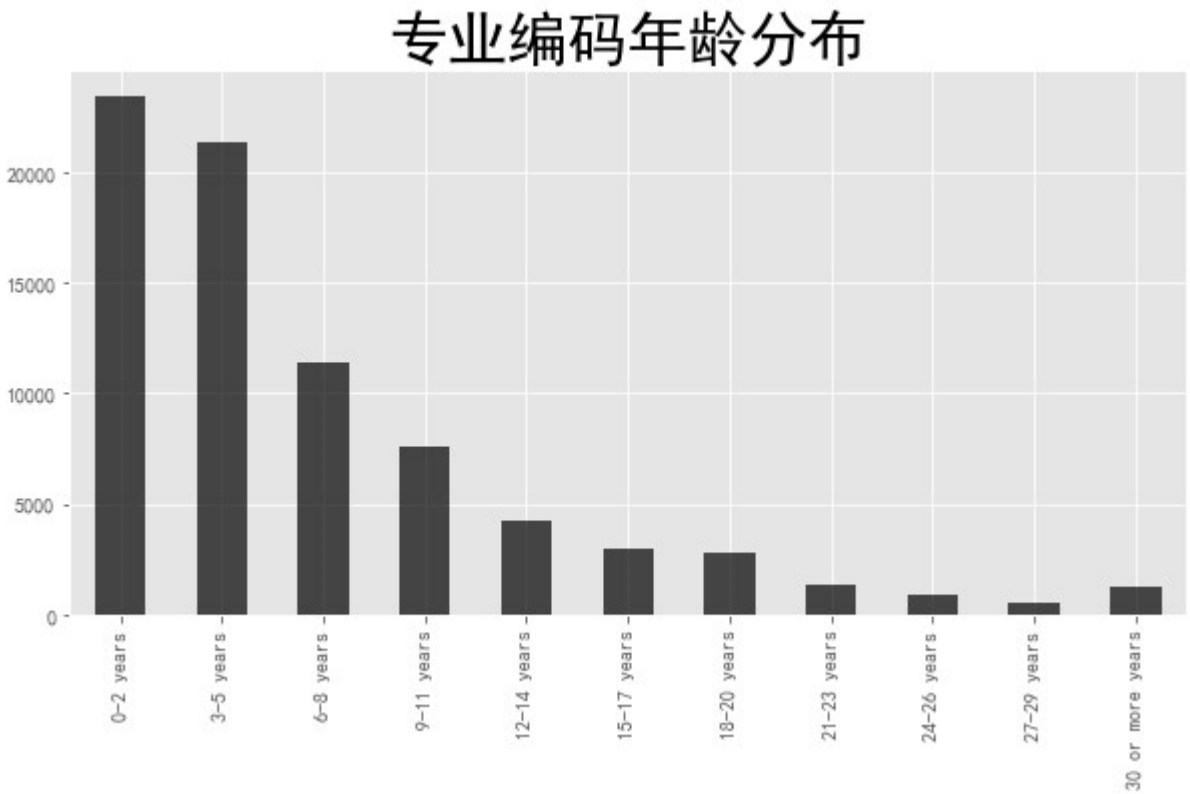
**共93835个回答*

用户编码年龄主要集中在3-5年，有一定的经验，仍需学习阶段。



**共77903个回答*

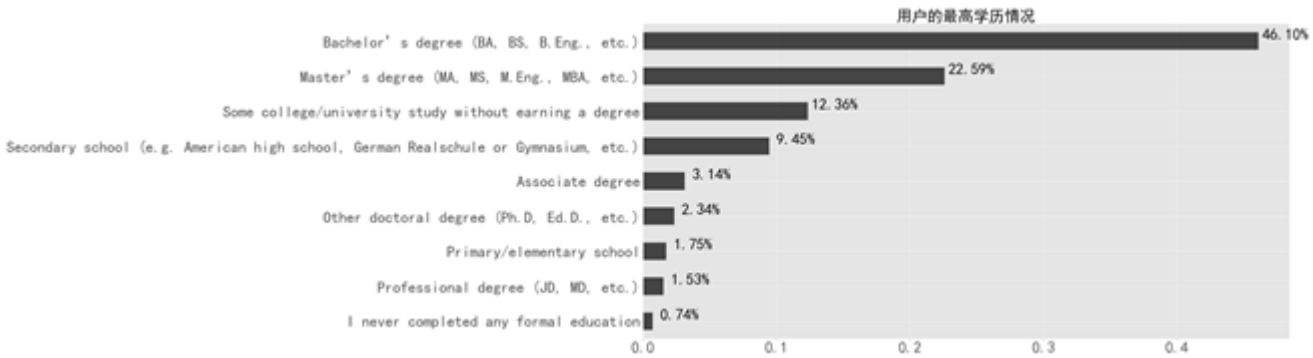
用户以编码为职业的时间集中在0-2年，3-5年这两个阶段，说明上Stack的用户大多都是刚入行不久，刷Stack的目的主要是提升个人技能。



4.用户的最高学历是怎么样的？

**共94703个回答*

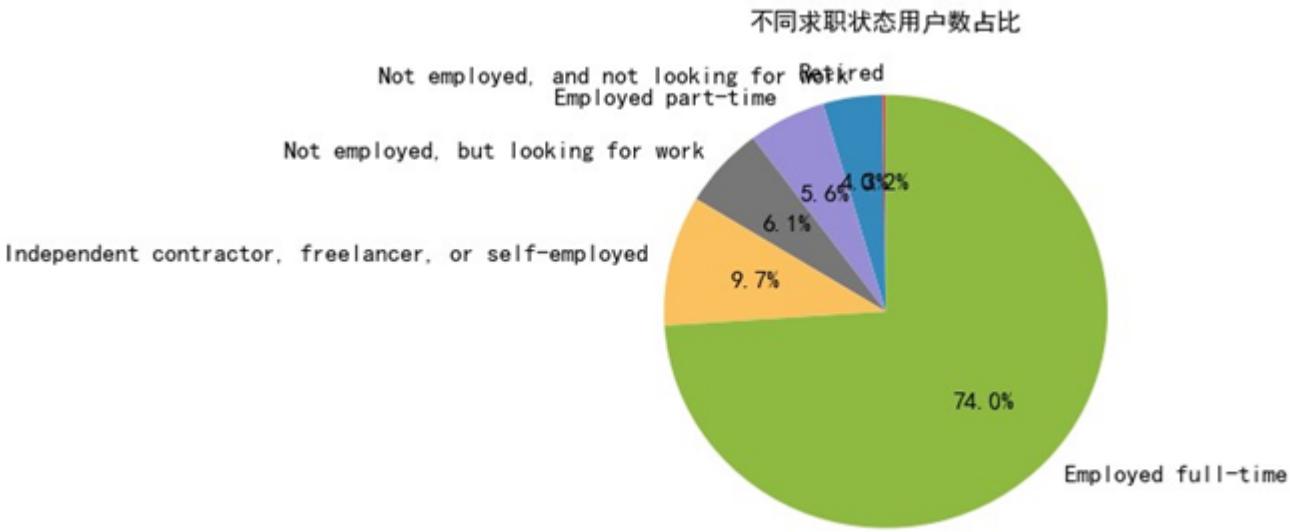
Stack用户拥有本科学历的占到46.1%，其次是硕士占到22.59%，而没有接触过正式教育的用户仅占0.74%。整体上看用户的学历普遍在本科及本科以上。



5.目前的求职状态是什么？是否是学生？

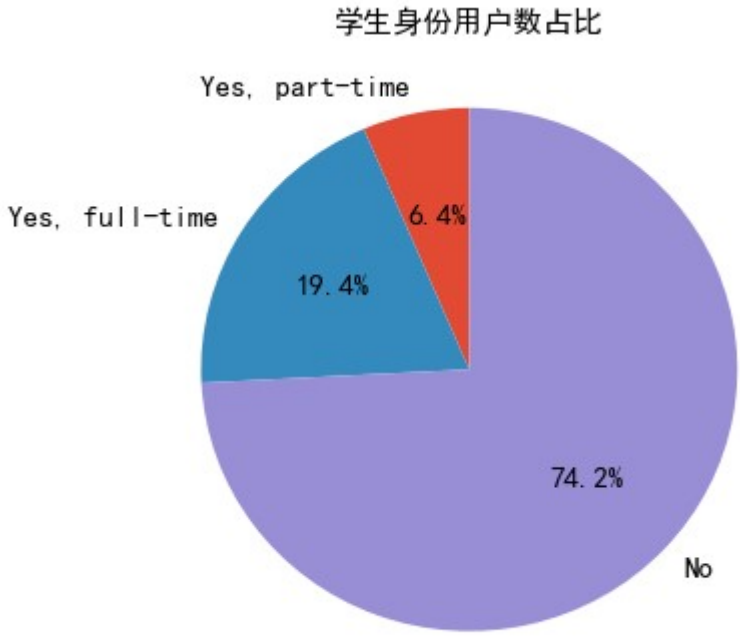
**共95321个回答*

Stack活跃的用户，74%都是在职员工，其次有9.7%的用户是自由职业者，5.6%的用户是拥有一份兼职，说明有近90%的用户都有一定经济能力。



**共9401个回答*

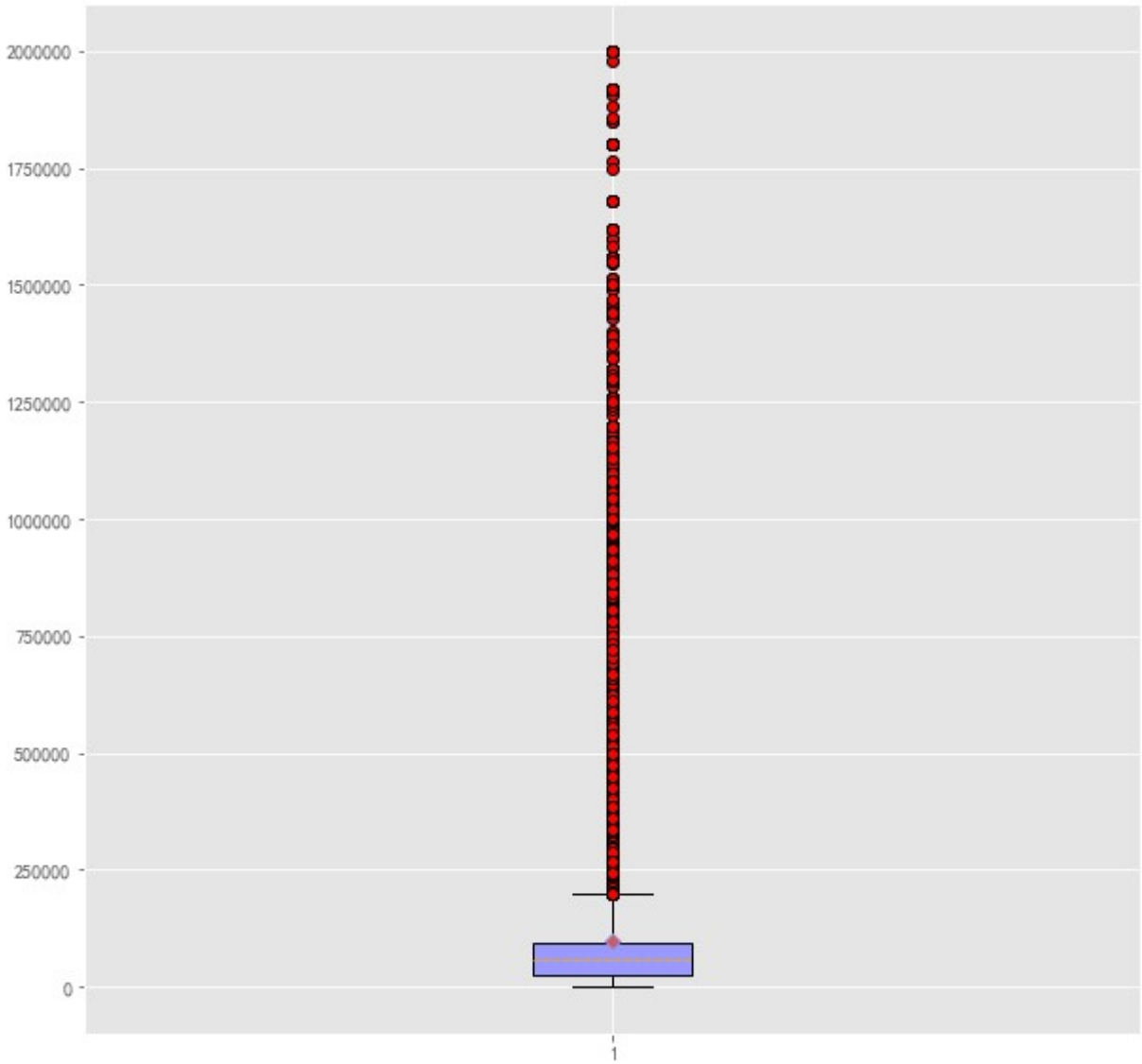
有74.2%的用户不是学生，结合求职状态数据可看出有全职工作的用户一般不是学生，但是全职学生或兼职学生的比例达到25.8%。虽然职业人士占Stack用户的大部分，但学生群体亦是一股不小的力量。



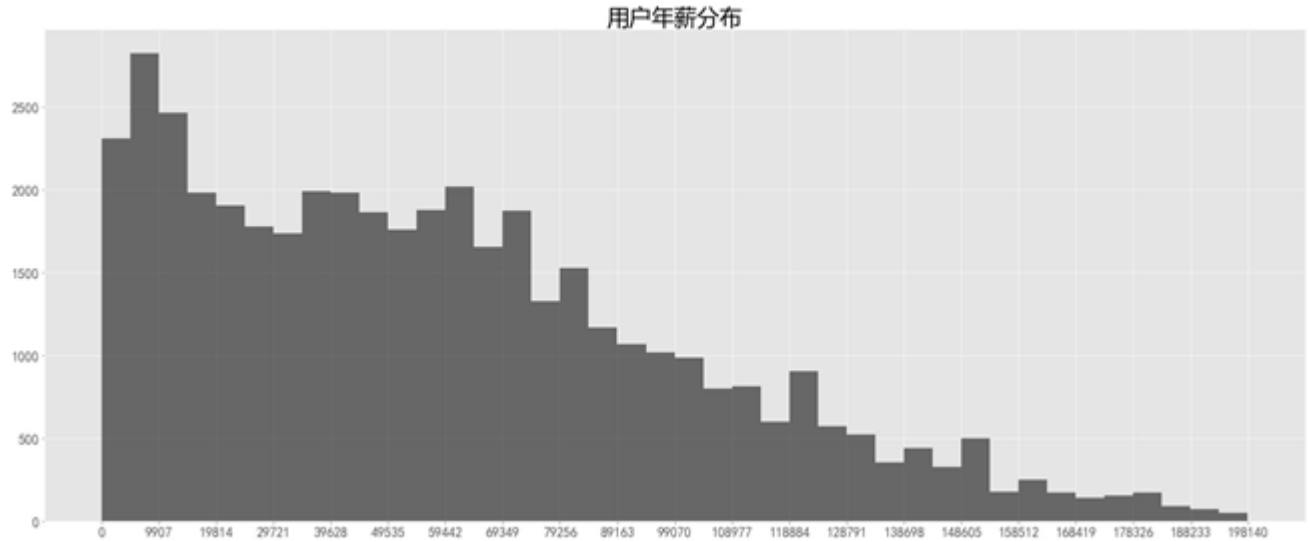
6.用户的收入分布

**共46860 个完整回答*

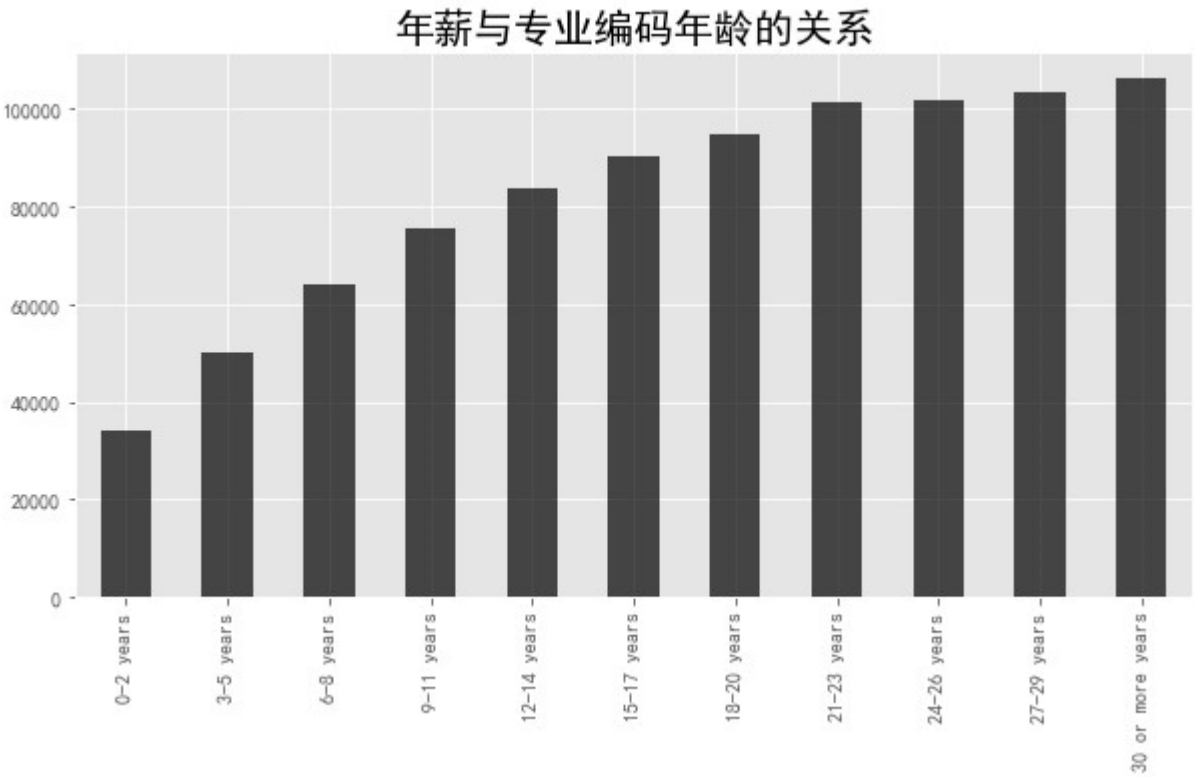
原始的薪资数据是以美元年薪换算（假设一年有50周，12个月），将薪资和薪资类型为NaN值或0值的数据剔除。在工资处理中发现有许多年薪的数据非常大，最大大到200w，通过箱型图看数据分布，发现异常值较多，上极限值为198139.5美元年薪，因此将异常值排除出去再来进行统计，这意味着我可能排除了类似facebook创始人马克·扎克伯格年薪这样的数据。



通过直方图可以看出，用户工资主要集中在3-5万年薪左右，这在美国的个人收入中算中等收入，且有很多年前超过十万的高收入人群，Stack用户有很大一部分工作中会用到编码，也可以推测出他们属于IT行业，而IT行业在美国算是高薪职业。其中0-9907美元年薪收入也占比较大的一部分，这可以用前面提到学生数据解释，有25.8%的用户是学生。



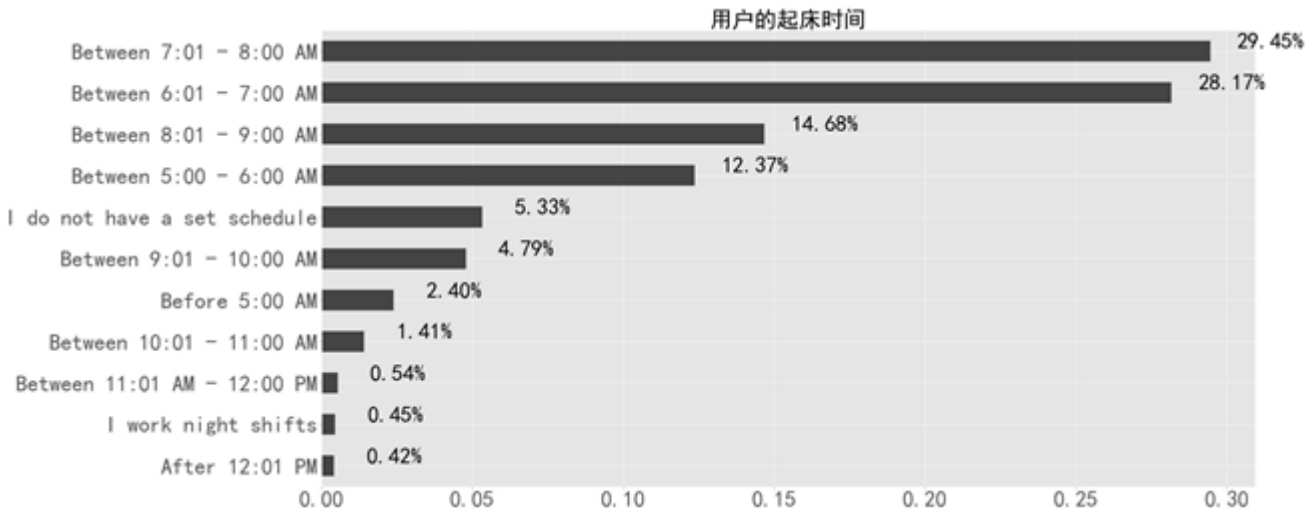
整体的平均年薪随编码年龄而增长，与编码相关的工作平均年薪在前20年的变化是比较快的，这也是职业的发展期，而超过20年后，平均年薪趋于稳定。



7.用户的作息时间（WakeTime）

**共72146个回答*

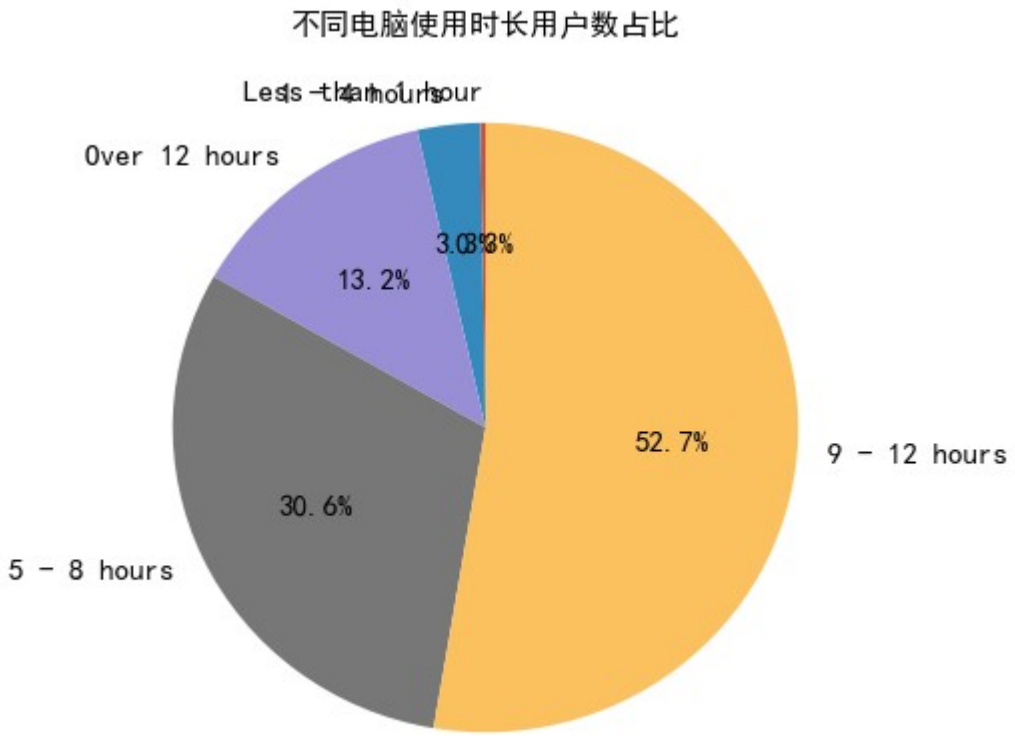
用户的作息大部分正常，占比最高的是7：01-8：00 AM 这一时间段，一般公司的上班时间为9：00。



8.用户一天花多少时间在电脑上

**共72133个回答*

如果按上班8小时计，那么上班工作用到电脑的时长在5-8小时，说明不用在非工作时间使用电脑，那么数据表明：只有30.6%的用户属于这一范围。程序员加班文化结合来看，大概有65.9%的程序员需要加班。

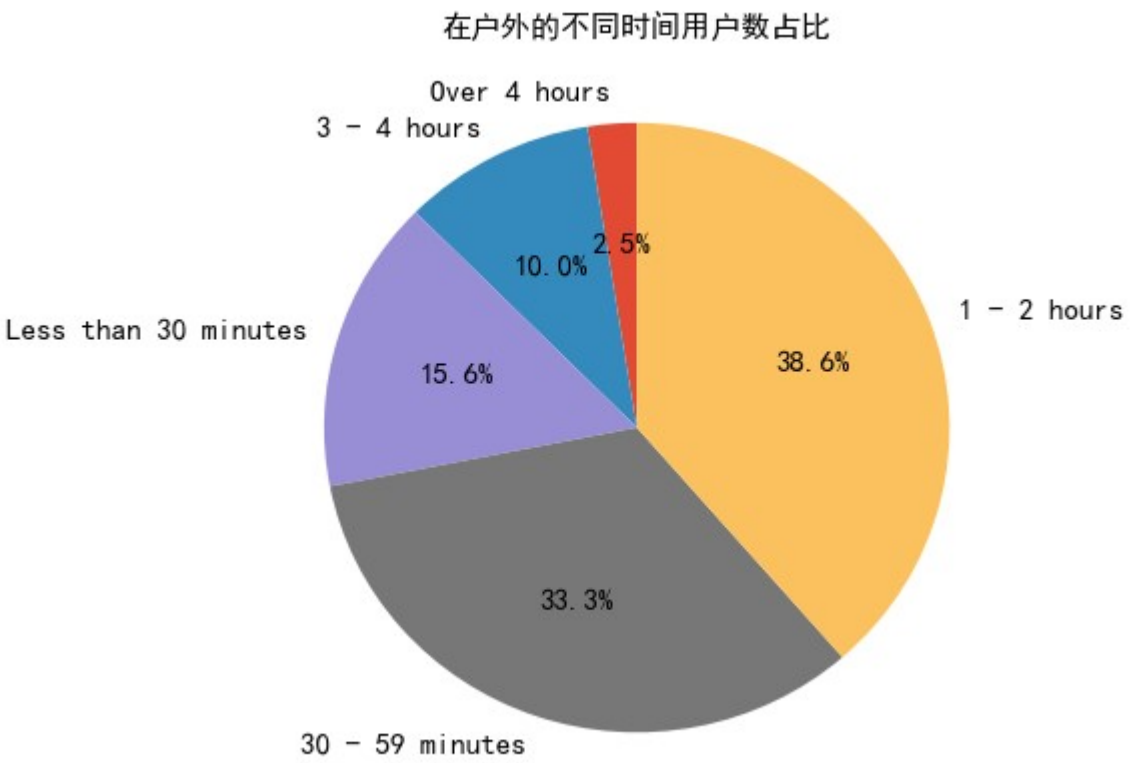


9.用户一天花多少时间在户外及锻炼的频率

**共72024个回答*

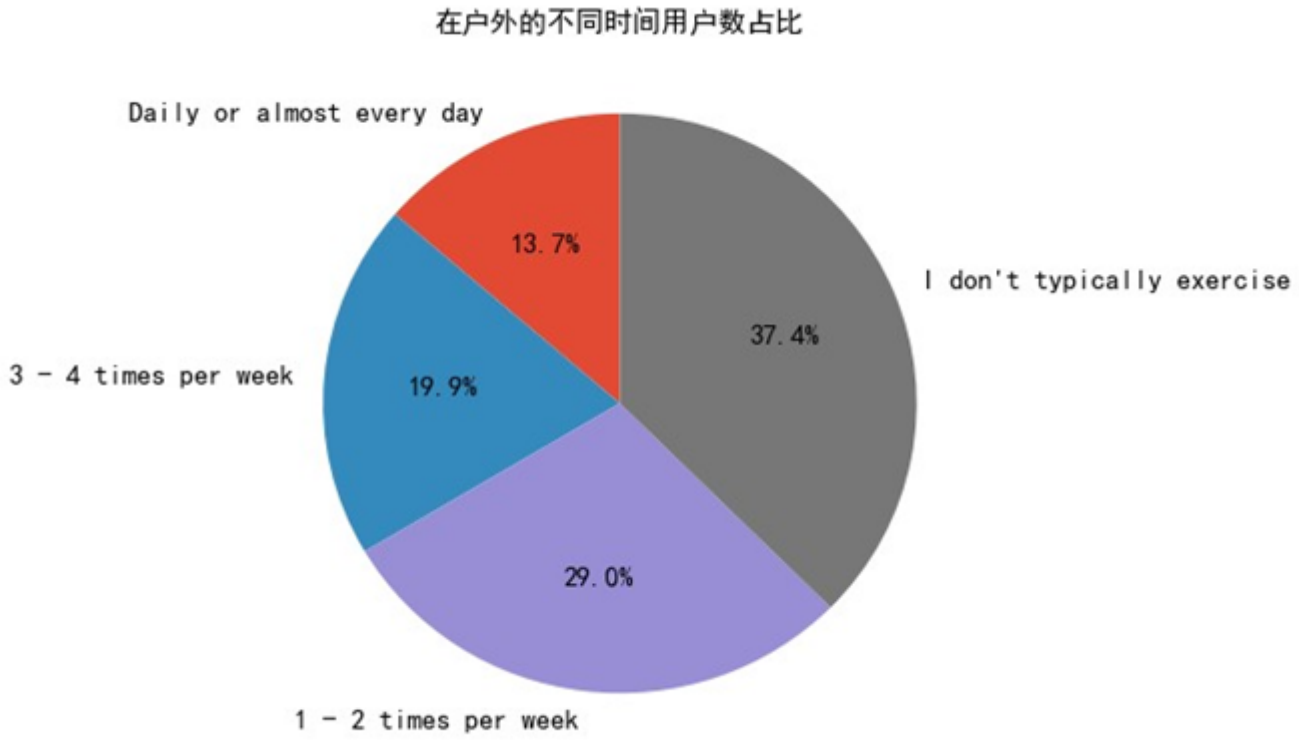
一天所花时间少于39分钟在户外的用户可以说是非常宅了，基本上没事不出门。

而有71.9%的用户在每天会在户外花30分钟-2小时，算上平时的通勤时间，在2小时内还是比较合理的。



**共72108个回答*

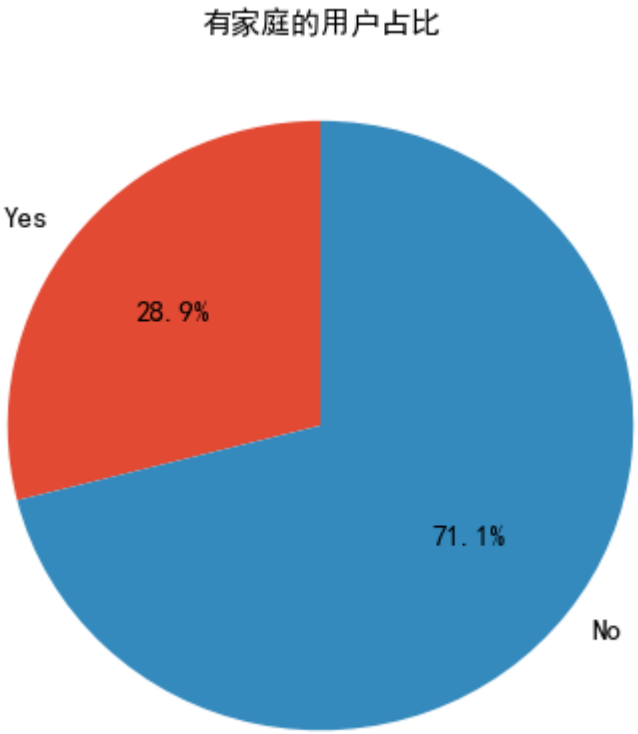
几乎不运动的用户占了很大一部分，在四种分类中占比也是最高。但是基本上satck用户还是挺注意健康的，62.6%的用户每周都至少有运动一次。



10.用户是否有家庭

**共62596个回答*

有家庭的用户占28.9%，而未婚/无家庭的用户占到71.1%，前面也有提到，Stack用户的年龄集中在22-34这一青年阶段，所以未婚/无家庭的占比大也有合理解释。



项目总结

我们从Stack的一些个人基本信息，如：国籍，性别，年龄，学历，收入，家庭，是否职业编程，求职状态，作息时间，使用电脑的时间，在户外的时间及锻炼频率，去描写了一个大体的用户画像。

由数据可以总结出：

- **Stack的用户来自世界范围，从而可以提现出satck是一个国际化平台；**
- **Stack的用户几乎都是年轻男性，很大一部分未婚； satck的用户学历高，收入客观，有一份相关编码的工作；**
- **satck的用户工作时间长，加班普遍，但也非常自律，坚持早起和运动。**

项目问题解决方向

- **什么类型的广告适合投放在 Stack?**

  从用户画像看出，针对男性的高品质，且不用花费太多时间去选择的产品，比较适合satck的用户，因为他们的收入不错，但时间主要花在工作上。例如可以投放最新的运动产品，数码产品等。

- **Stack在什么类型的平台发布广告可以使自己的推广效果达到最大呢？**

  可以有针对性地去程序员集中的地方投放推广广告，如程序员论坛，程序员博主等；

另一方面，大部分人把Stack当做一个提升自我技能及疑问解答的开源网站，那么可以和一些编程培训学校或网站合作，培养未来的中坚力量。

反思与致谢

- 本数据集一共有128个特征，本次报告只分析了有针对性的15个特征，还有很多特征值得去深挖了解，如可以通过年龄切片成不同的数据集，去多个方面对比，以发现每个年龄层不同的特征及偏好；
- 本分析的结果都只是统计学上的相关性，并不具备因果关系，因果性有待日后更严谨的大样本随机双盲试验得出结果。