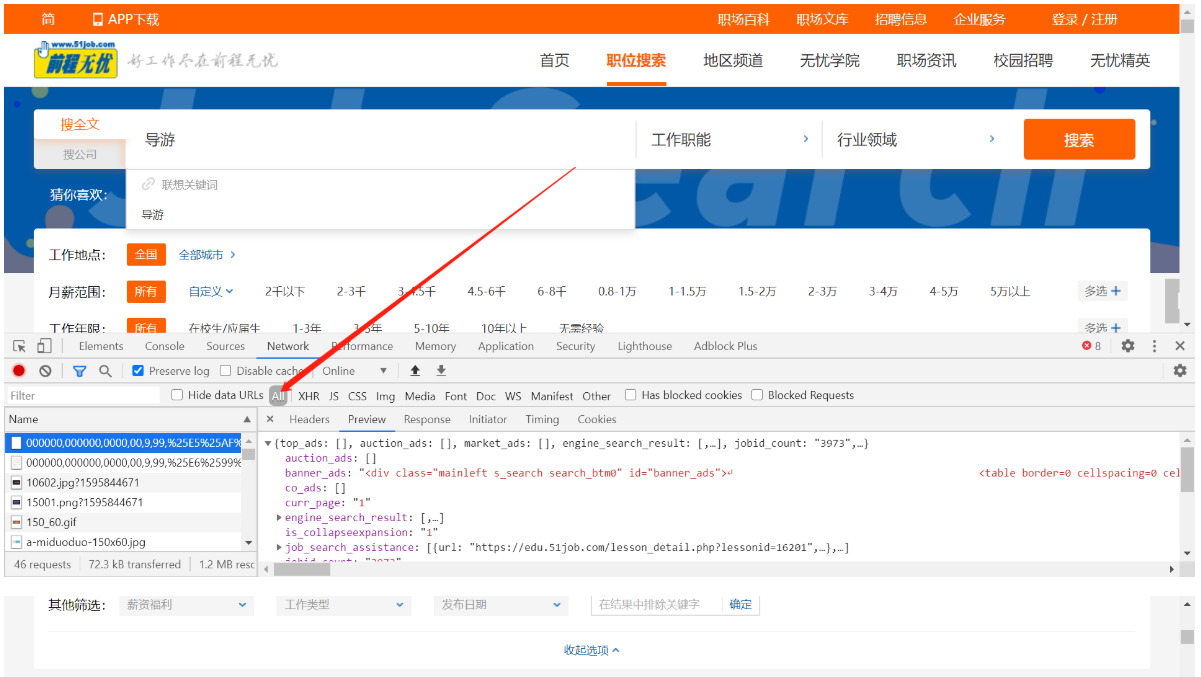
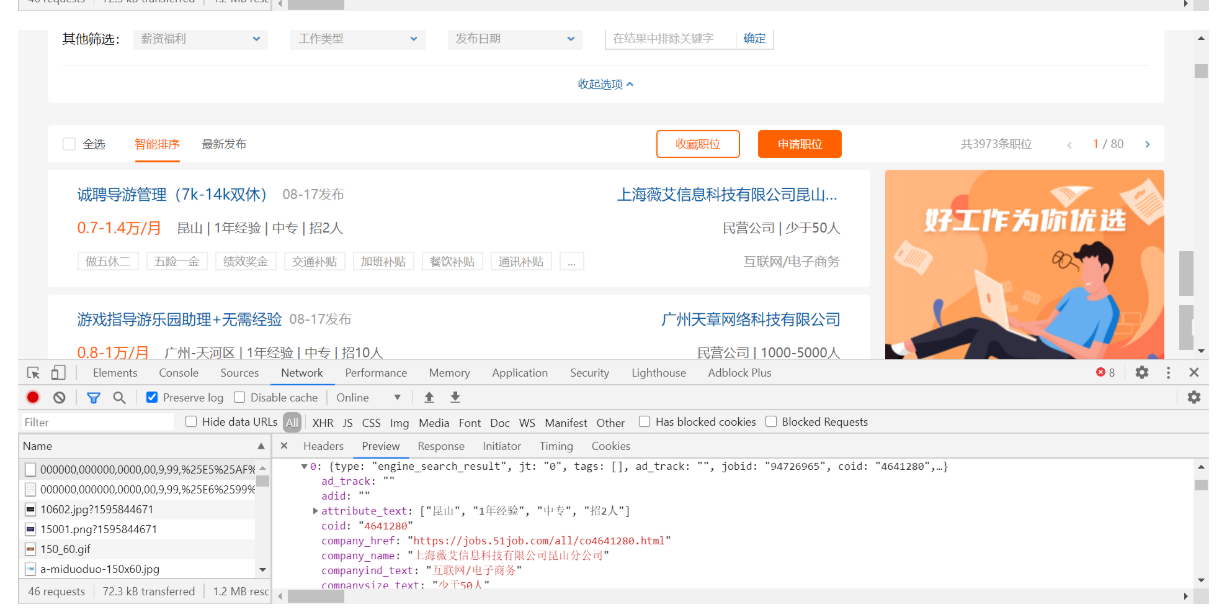
**旅游招聘数据分析之前程无忧**

这个是我们要爬取的网站：[前程无忧](https://www.51job.com/)

前程无忧的网站是一个动态网站来的，单纯去抓取是无法抓取的，所以我们得找到他的接口，打开网站按f12

选择network选项卡





显然这个就是我们需要的找到的接口了

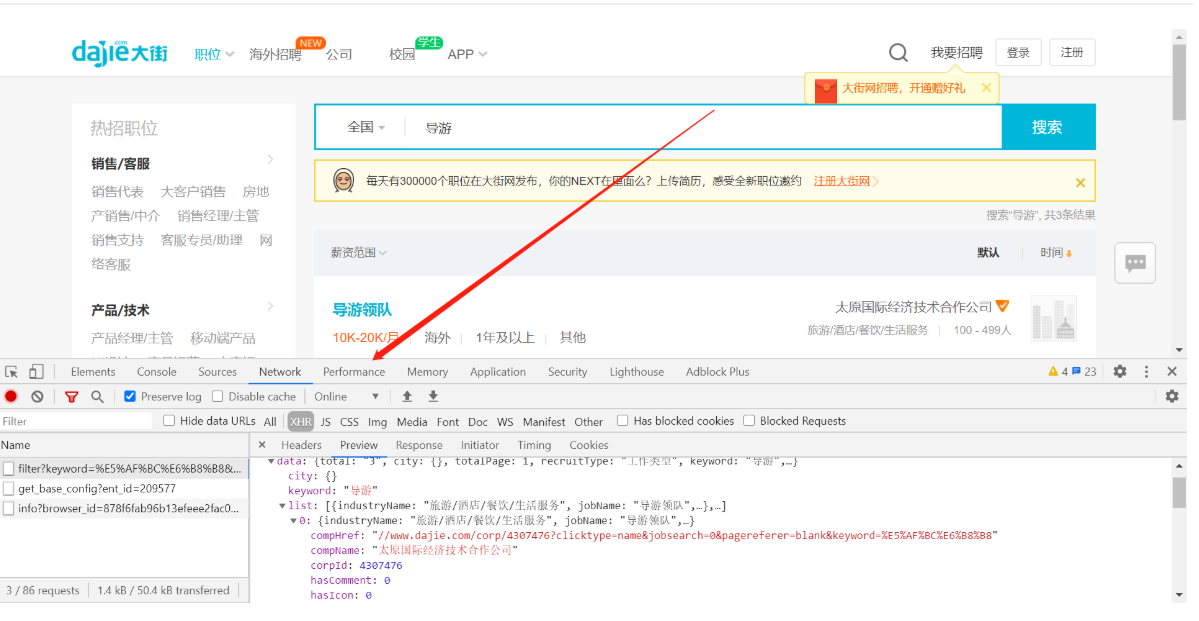
根据他给出来的URL接口<https://search.51job.com/list/000000,000000,0000,00,9,99,%25E5%25AF%25BC%25E6%25B8%25B8,2,1.html?lang=c&postchannel=0000&workyear=99&cotype=99&degreefrom=99&jobterm=99&companysize=99&ord_field=0&dibiaoid=0&line=&welfare=>

其中1是指页数，然后就根据他有多少页数就写一个for循环来进行叠加就好了

然后在获取到这个接口的信息之后，我们写一个正则表达式把我们需要的内容全部按顺序的爬下来，

最后把用pands来把我们爬下来的内容依次打印成CSV文件

**旅游招聘数据分析之大街网**



大街网的思路和上面一样，都是动态页面，需要我们耐心去找到他们的接口，然后根据接口给出来的URL：<https://so.dajie.com/job/ajax/search/filter?keyword=%E5%AF%BC%E6%B8%B8&order=0&city=&recruitType=&salary=&experience=&page=1&positionFunction=&_CSRFToken=&ajax=1>，来构建URL从而获取我们的页数

这里ajax=1指的是页数的意思，keyword是关键词，这里要用到parse.quteo(input("请输入关键词:"))的方法来获取加密后的字符串，从而构建我们的URL来获取我们需要的数据

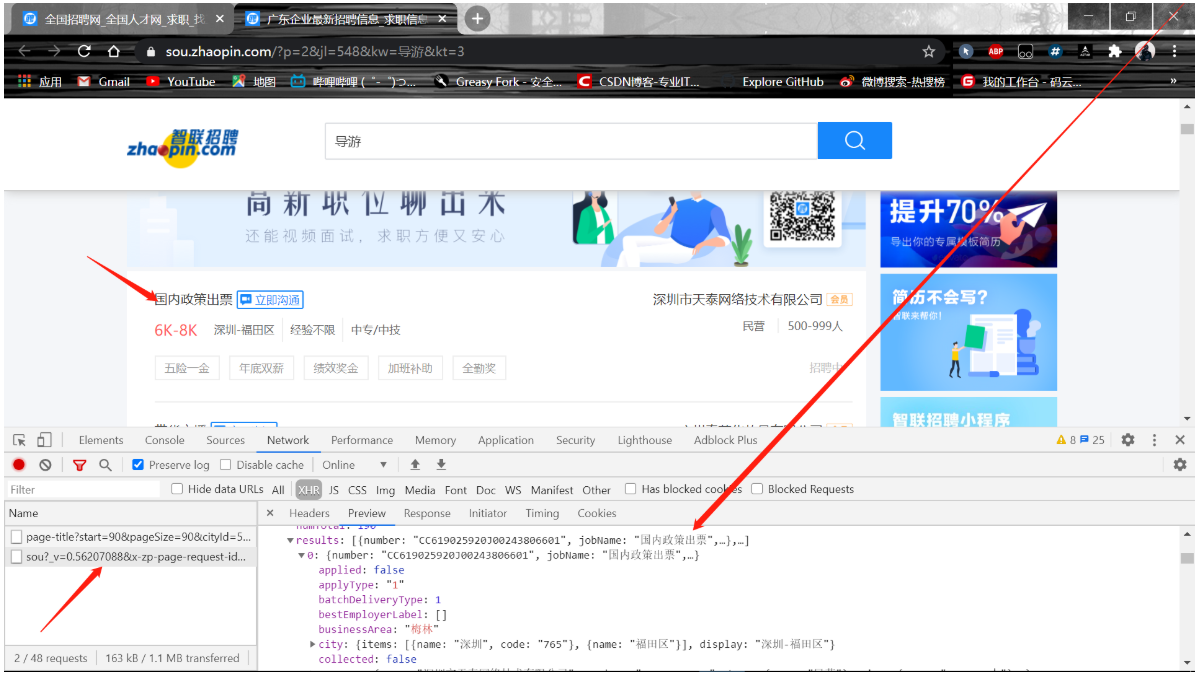
**旅游招聘数据分析之智联招聘**

智联招聘的具体步骤都是和上面类似，都是一个动态页面，需要我们去找到他们的接口才能获取数据，智联招聘的URL更为复杂：<https://fe-api.zhaopin.com/c/i/sou?_v=0.56207088&x-zp-page-request-id=9ad864f1dc6e497094eccb24778ef475-1597675999006-918393&x-zp-client-id=0452fae6-2162-4777-bc85-188030dac01f&MmEwMD=5jqng3nR1fCUImwlNiEahK1ZMeTdoCK3pcv6MJmDOAFpOPz4geS8inh5e6_R7Gs4dHUy3ezY5e2XIb1mBob_gPy7eHlCGCOE2RDajEOfEj69zR.TjRO0bomsNLZEKwJZaHRXPQRIXw0msOL5qzuYwZqBmZuJp_EGat_WDAST9fGGiG6D8tme5vWpVvIXMhfMBI0jBJQblppjqqaMg1C8BjMKFqpGFWORu0d2pNXVKOdGd0Eoz8LG5DZYTfLfWWA5Sg87HtsMVaBZL0FM3GrCP3W6_V51KL6858r2pOWyh6QqFNNQ_PCeJWlEv.GzaxcT3BAVd87QMw4oh5MfQC7Q8GiZMmd8fkQvyQpfCpiuoC9Z7eRV1QrnvzKExUhMtKG8yeDAvmgHt9aedv_FDt1VPkvVXyv_LQNslP9ukkiI7I.qLgsYCosDtH1T5YK9lB2KcOG3>

他是根据获取到页面的URL里面的params来操控的需要我们自己的定义一个Parmas参数，然后把这个参数加到我们的URL里面从而构成智联招聘的URL接口

params = {  
   '\_v': '0.56207088',  
   'x-zp-page-request-id': '9ad864f1dc6e497094eccb24778ef475-1597675999006-918393',  
   'x-zp-client-id': '0452fae6-2162-4777-bc85-188030dac01f',  
 'MmEwMD':'5jqng3nR1fCUImwlNiEahK1ZMeTdoCK3pcv6MJmDOAFpOPz4geS8inh5e6\_R7Gs4dHUy3ezY5e2XIb1mBob\_gPy7eHlCGCOE2RDajEOfEj69zR.TjRO0bomsNLZEKwJZaHRXPQRIXw0msOL5qzuYwZqBmZuJp\_EGat\_WDAST9fGGiG6D8tme5vWpVvIXMhfMBI0jBJQblppjqqaMg1C8BjMKFqpGFWORu0d2pNXVKOdGd0Eoz8LG5DZYTfLfWWA5Sg87HtsMVaBZL0FM3GrCP3W6\_V51KL6858r2pOWyh6QqFNNQ\_PCeJWlEv.GzaxcT3BAVd87QMw4oh5MfQC7Q8GiZMmd8fkQvyQpfCpiuoC9Z7eRV1QrnvzKExUhMtKG8yeDAvmgHt9aedv\_FDt1VPkvVXyv\_LQNslP9ukkiI7I.qLgsYCosDtH1T5YK9lB2KcOG3'  
}

**旅游招聘数据分析之数据清洗**



在获取完我们的数据之后，直接需要我们对数据进行清洗了，首先我们先创建两个文本文件，然后先对我们获取到的文件进行查重

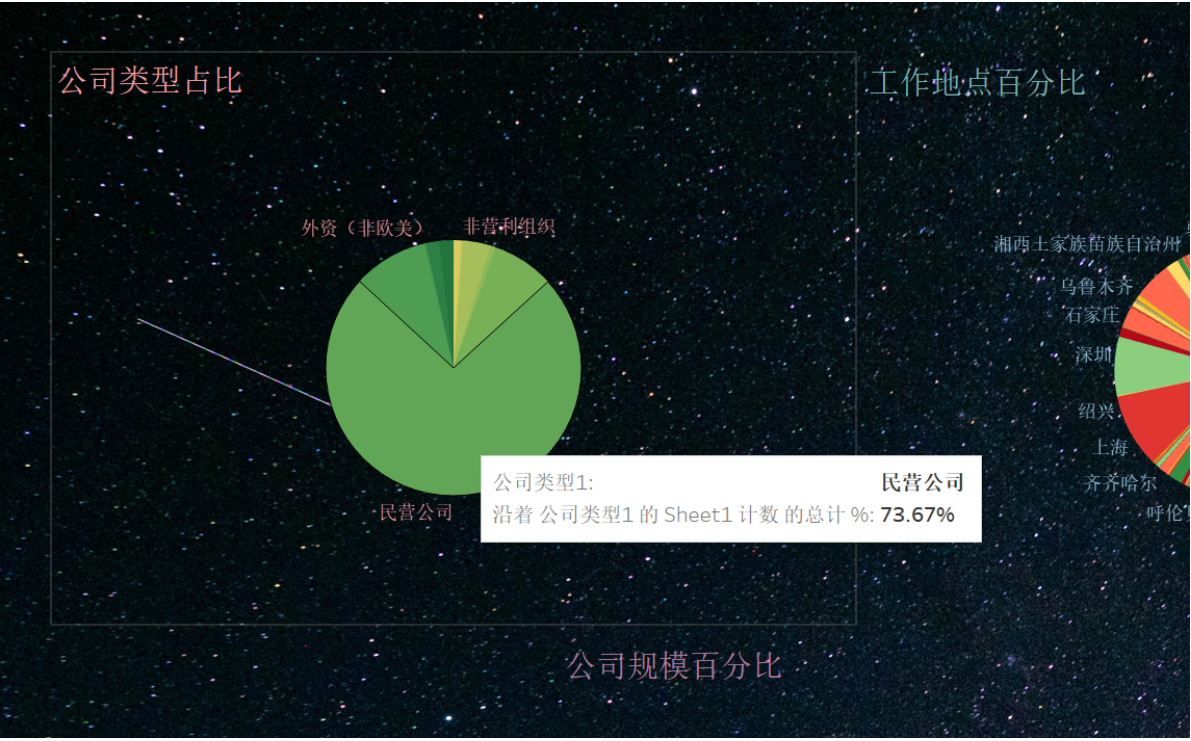
import pandas as pd  
data= pd.DataFrame(pd.read\_excel('数据大集成.xlsx','Sheet1'))  
no\_re\_row = data.drop\_duplicates()  
print(no\_re\_row)  
no\_re\_row.to\_excel("新(数据大集成).xls")

然后我们到查重好的文件里面，先将里面的部分内容进行复制，因为，数据量太大，文本文件大小有限制，超过这个范围，文本文件就容易出错，这个不是我们想要的结果，所以最好就是一步步复制，我这里是将Excel表进行一列列复制，然后再写一个程序，把里面不需要，多余的文字进行删除，清洗完之后就是我们的数据分析了

import re  
from tqdm import tqdm  
list= []  
with open("stopword.txt","r",encoding="utf-8")as f:  
   for line in f.readlines():  
       line = line.strip('\n')  
       list.append(line)  
def clearBlankLine(i):  
   file1 = open('整理文件.txt', 'r',encoding="utf-8")  
   file2 = open('整理好的内容.txt', 'w', encoding='utf-8')  
   try:  
       for line in file1.readlines():  
           line = line.replace(i,"")  
           file2.write(line)  
           # line = line.replace("英语","").replace("薪聘","").replace("高","").replace("顺德区","").replace("旅游销售\\\\","") \  
           #     .replace("去哪儿", "").replace("网","").replace("旅游在线","").replace("客服","").replace("门店","").replace("全球","") \  
           #     .replace("包住宿", "").replace("月薪","").replace("携程","").replace("8500","").replace("8K","").replace("旅游产品专员\\\\","") \  
           #     .replace("急聘", "").replace("康养","").replace("7000+","").replace("酒店","").replace("薪","").replace("招聘","") \  
           #     .replace("同业", "").replace("无经验","").replace("+","").replace("轻松","").replace("工作","").replace("旅行海外","") \  
           #     .replace("同业", "").replace("无经验", "").replace("+", "").replace("轻松", "").replace("工作", "").replace(  
           #     "旅行海外", "") \  
           #     .replace("泰语", "").replace("客服", "").replace("主管", "").replace("门店", "").replace("扶贫", "").replace(  
           #     "旅行海外", "") \  
           #     .replace("高铁", "").replace("资深", "").replace("，", "").replace("轻松", "").replace("工作", "").replace(  
           #     "旅行海外", "") \  
           #     .replace("同业", "").replace("无经验", "").replace("+", "").replace("轻松", "").replace("工作", "").replace(  
           #     "旅行海外", "") \  
               # line = line.replace("0.1","1").replace("0.2","2").replace("0.3","3").replace("0.4","4") \  
           #     .replace("0.5", "5").replace("0.6","6").replace("0.7","7").replace("0.8","8").replace("0.9","9")  
           # line = line.replace("'","").replace("\\xa0","").replace("\*\*\*","").\  
           #     replace(",","").replace("★","").replace("◆","").replace("（","").\  
           #     replace("）","").replace("【","").replace("】","").replace("\\n","").\  
           #     replace('[','').replace(']',"").replace("...","").replace('\\\\',"/")  
           # line = line.strip(" ")  
           # line = line.replace("1.20k","12k").replace("1.40k","14k").replace\  
           #     ("1.50k","15k").replace("1.30k","13k").replace("10-150k/年","10-15k").\  
           #     replace("8-150k/年","8-15k").replace("15-200k/年","15-20k").\  
           #     replace("200元/天","6k").replace("1.5千以下","1.5k").\  
           #     replace("300元/天","9k").replace("100元/天","3k").replace("150元/天","4.5k").\  
           #     replace("7-120k/年","7-12k").replace("30-400k/年","3-4k").replace("20-300k/年","2-30k").replace("8-100k/年","8-10k").replace("8-200k/年","8-20k")  
           # line = line.replace("1-","1k-").replace("2-","2k-").replace("3-","3k-").\  
           #     replace("6-","6k-").replace("5-","5k-").replace("4-","4k-").replace("7-","7k-").replace("8-","8k-").replace("9-","9k-").replace("0-","0k-")  
           # line = line.replace("A轮","民营公司").replace("B轮","民营公司").\  
           #     replace("D轮及以上","国有企业").replace("天使轮","民营公司").\  
           #     replace("C轮","民营公司").replace("未融资","民营公司").replace("不需要融资","合资")  
           # line = line[0:2]  
           # line = line.replace("哈尔","哈尔滨").replace("大兴","大兴安岭").replace("防城","防城港").replace("呼和","呼和浩特").\  
           #     replace("呼伦","呼伦贝尔").replace("葫芦","葫芦岛").replace("红河","红河州").replace("景德","景德镇").replace("克拉","克拉玛依")\  
           #     .replace("喀什","喀什地区").replace("马鞍","马鞍山").replace("牡丹","牡丹江").replace("秦皇","秦皇岛").replace("齐齐","齐齐哈尔").\  
           #     replace("七台","七台河").replace("黔东","黔东南").replace("石家","石家庄").replace("神农","神农架").replace("双鸭","双鸭山")\  
           #     .replace("石河","石河子").replace("图木","图木舒克").replace("五指","五指山").replace("乌鲁","乌鲁木齐").replace("西双","西双版纳")\  
           #     .replace("张家","张家界").replace("驻马","驻马店")  
           # line = line.replace("0k","0").replace("1k","1").replace("2k","2").replace("3k","3").replace("4k","4")\  
           #     .replace("5k","5").replace("6k","6").replace("7k","7").replace("8k","8").replace("9k","9")  
           # line = line.replace("1-1","1-1千/月").replace("1-2","1-2千/月").replace("1-3","-3千/月").replace("1-4","1-4千/月").\  
           #     replace("1-5","1-5千/月").replace("1-6","1-6千/月").replace("1-7","1-7千/月").replace("1-8","1-8千/月").\  
           #     replace("1-9","1-9千/月").replace("2-3","2-3千/月").replace("2-4","2-4千/月").\  
           #     replace("2-5","2-5千/月").replace("2-6","2-6千/月").replace("2-7","2-7千/月").replace("2-8","2-8千/月").\  
           #     replace("2-9","2-9千/月").replace("3-4","3-4千/月").\  
           #     replace("3-5","3-5千/月").replace("3-6","3-6千/月").replace("3-7","3-7千/月").replace("3-8","3-8千/月").\  
           #     replace("3-9","3-9千/月").\  
           #     replace("5-5","5-5千/月").replace("5-6","5-6千/月").replace("5-7","5-7千/月").replace("5-8","5-8千/月").\  
           #     replace("5-9","5-9千/月").replace("6-6","6-6千/月").replace("6-7","6-7千/月").replace("6-8","6-8千/月").\  
           #     replace("6-9","6-9千/月").replace("7-7","7-7千/月").replace("7-8","7-8千/月").\  
           #     replace("7-9","7-9千/月").replace("8-8","8-8千/月").\  
           #     replace("8-9","8-9千/月").\  
           #     replace("4-5","4-5千/月").replace("4-6","4-6千/月").replace("4-7","4-7千/月").replace("4-8","4-8千/月").\  
           #     replace("4-9","4-9千/月")  
           # line = line.replace("20-4万/月0","2-4万/月")\  
           #     .replace("-35","-3.5万/月").replace("-38","-3.8万/月").replace("-55","-5.5万/月").replace("-25","-2.5万/月").\  
           #     replace("-36","-3.6万/月").replace("-26","-2.6万/月").replace("-27","-2.7万/月").replace("-28","-2.8万/月").\  
           #     replace("-29","-2.9万/月").replace("-24","-2.4万/月").replace("-23","-2.3万/月").replace("-22","-2.2万/月").\  
           #     replace("-21","-2.1万/月").replace("-31","-3.1万/月").replace("-32","-3.2万/月").replace("-33","-3.3万/月").\  
           #     replace("-34","-3.4万/月").replace("-37","-3.7万/月").replace("-39","-3.9万/月").replace("-50","-5万/月").\  
           #     replace("-60","-6万/月")  
           # line = line.replace("-10","-1万/月").replace("-11","-1.1万/月").replace("-12","-1.2万/月").replace("-13","-1.3万/月").\  
           #     replace("-14","-1.4万/月").replace("-15","-1.5万/月").replace("-16","-1.6万/月").replace("-17","-1.7万/月").replace("-18","-1.8万/月").\  
           #     replace("-19","-1.9万/月").replace("-20","-2万/月").replace("-30","-3万/月").replace("-40","-4万/月")  
           # line = line.replace("2千/月.2万/月","2.2万/月").replace("千/月万/月","万/月").\  
           #     replace("1千/月.2万/月","1.2万/月").replace("1千/月.5万/月","1.5万/月").replace("3千/月.5万/月","3.5万/月")  
​  
​  
​  
   finally:  
       file1.close()  
       file2.close()  
​  
​  
if \_\_name\_\_ == '\_\_main\_\_':  
   for i in tqdm(list):  
       clearBlankLine(i)

**旅游招聘数据分析之数据清洗**

这里我采用的是tableau publish，我之所以选择它，是因为它简单易上手，而且弄出来的图，清晰可见，易懂，非常好用，



在这里我们直接鼠标移到饼图就可以看清楚，某个东西占的比重，方便我们做数据分析。