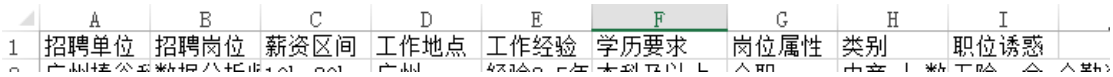
“数据分析”知多少，让数据分析告诉你

目的：通过对招聘网站“数据分析”职位信息进行提取和分析，让自己有一个更加全面和真实的了解，从而更好地判断自己是否能做好这个职位，是否真的喜欢这个职位，自己还有哪些差距和不足需要弥补，也从而能更好地指引后续的学习方向

数据：通过爬虫方式对拉勾网上的数据分析职位信息进行抓取，主要抓取的字段信息如下



	A	B	C	D	E	F	G	H	I
1	招聘单位	招聘岗位	薪资区间	工作地点	工作经验	学历要求	岗位属性	类别	职位诱惑

角度：主要从如下几个角度对数据分析职位做一些基本了解

1、 职位的地域分布情况（机会）

2、 职位的整体薪酬情况（待遇）

3、 不同地域的薪酬情况

4、 该职位所需的工作经验如何（普世标准）

5、 工作经验和薪酬的关系

6、 职位所属行业类别，需要具备的技能等（方向）

完整代码详见 github: https://github.com/zgjsjw/data_analysis

第一步：对抓取的原始数据进行清洗，去掉没有利用到的数据信息，通过粗略扫描统计由于空值占比较少，去掉后对整体分析结果影响不大，所以去掉空值，同时去掉重复的职位数据信息。

```
def clean_data(path_in, path_out):
    if not os.path.exists(path_out):
        os.mkdir(path_out)
    data_in = pd.read_csv(os.path.join(path_in, 'data.csv'))
    print("原始数据总共有{}行".format(len(data_in)))
    # 去掉一些不需要的字段：发布时间、职位描述、工作地址、爬取时间
    data_mid = data_in.drop(['发布时间', '职位描述', '工作地址', '爬取时间'], axis=1)
    #去掉空值
    data_no_na = data_mid.dropna()
    print("去掉空值后，还剩{}行".format(len(data_no_na)))
    #去掉重复值
    data_out = data_no_na.drop_duplicates()
    data_out.to_csv(os.path.join(path_out, 'result.csv'), index=False, encoding='utf_8_sig')
    print("去重后，有效数据{}行".format(len(data_out)))
    return data_out
```

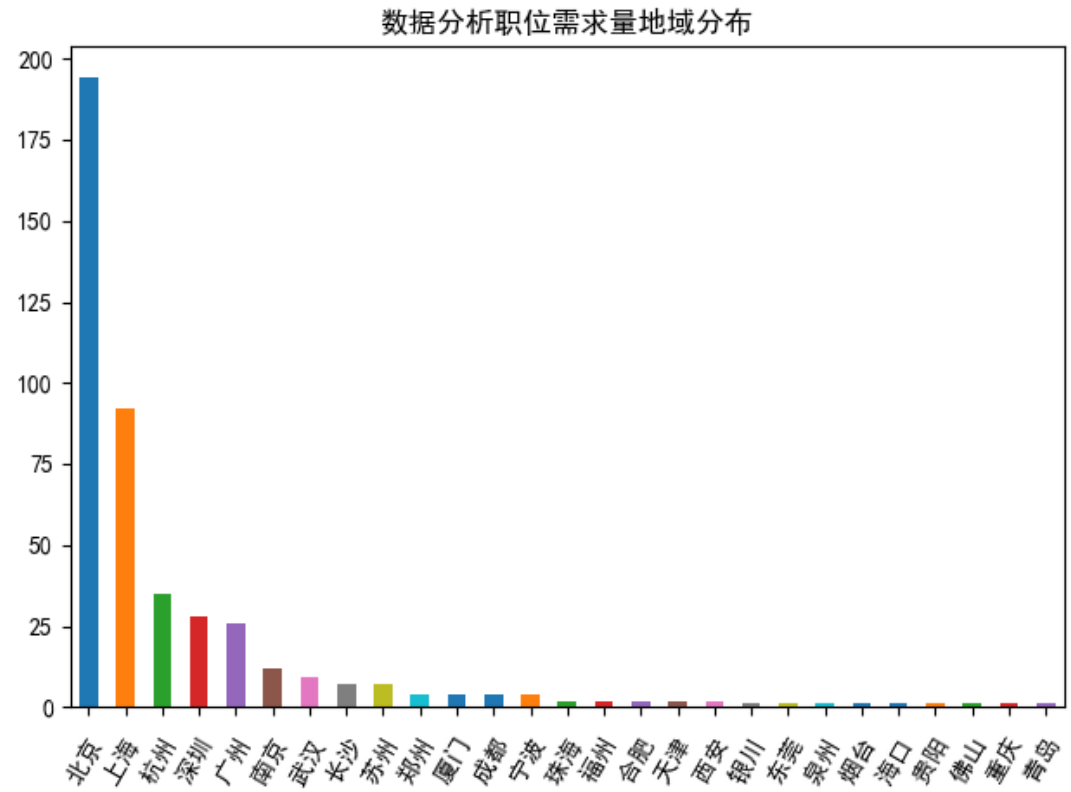
通过数据清洗可以发现，原始初始数据共有 450 条，其中空值数据 4 条，重复数据一条，异常数据占比较低，故作清除处理。

```
C:\Users\JIWei\Anaconda3\python.exe C:/Users/JIWei/Desktop/project_data_analysis/main.py
原始数据总共有450行
去掉空值后，还剩446行
去重后，有效数据445行
```

第二步：分析职位的地域分布情况

```
def data_area(data_in):
    data_count_city = data_in['工作地点'].value_counts()
    #print(data_count_city)
    #绘制地域分布条形图
    plt.figure()
    data_count_city.plot.bar()
    plt.xticks(rotation=60)
    plt.title("数据分析职位需求量地域分布")
    plt.tight_layout()
    plt.savefig(os.path.join(data_out_path, 'data_area.png'))
```

由于获取的数据信息中已直接包含‘工作地点’的信息，故直接取值进行分析即可：



从分析结果可以看出，不完全统计，大概有 20+地区有数据分析职位的需求信息，其中要有超出一半的职位分布在‘北上广深杭’，而且杭州的职位已经超出深圳和广州，可以初步判断出以阿里为巨头的互联网企业，在杭州已经形成了一种产业聚集效应。其中北京的职位数量遥遥领先，远超其他城市，几乎是第二需求量上海的双倍，分析前初步判断应该是上海具有大量的数据分析职位，毕竟上海作为全国金融中心，应该需要大量的数据分析人员，但从分析结果并结合数据来看，可能由于数据多属于互联网行业，而北京的互联网企业较多，所以该职位数量众多也是说得通的。

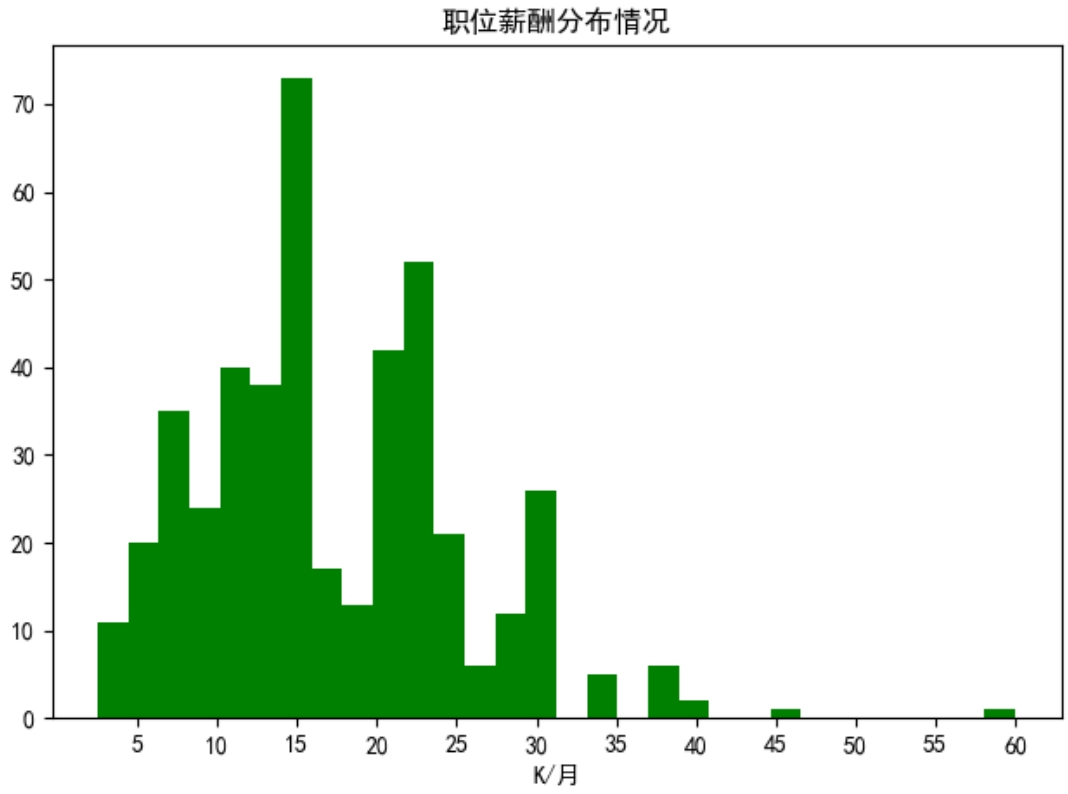
总的来看，在北上广深杭，数据分析有较多的职业机会，但同时也意味着会有大量的人才涌入到这几个地方，会造成较大的竞争，机会多，责任多，压力也会大。

第三步：分析该职位的整体薪酬情况

```
def data_salary(data_in):
    salary = data_in['薪资区间'].str.split('-').map(lambda t: (int(t[0][:-1]) + int(t[1][:-1]))/2)
    #print(salary)

    plt.figure()
    plt.title("职位薪酬分布情况")
    plt.hist(salary, bins=30, color='g')
    plt.xticks(range(5, 65, 5))
    plt.xlabel('K/月')
    plt.tight_layout()
    plt.savefig(os.path.join(data_out_path, 'data_salary.png'))
```

由于薪酬字段的信息不能直接进行运算显示，需要进行预处理，这里采用数据分析中常用的 labdba 函数，往往只需要很简要的一两行代码就可以很方便的进行数据处理，结果如下：



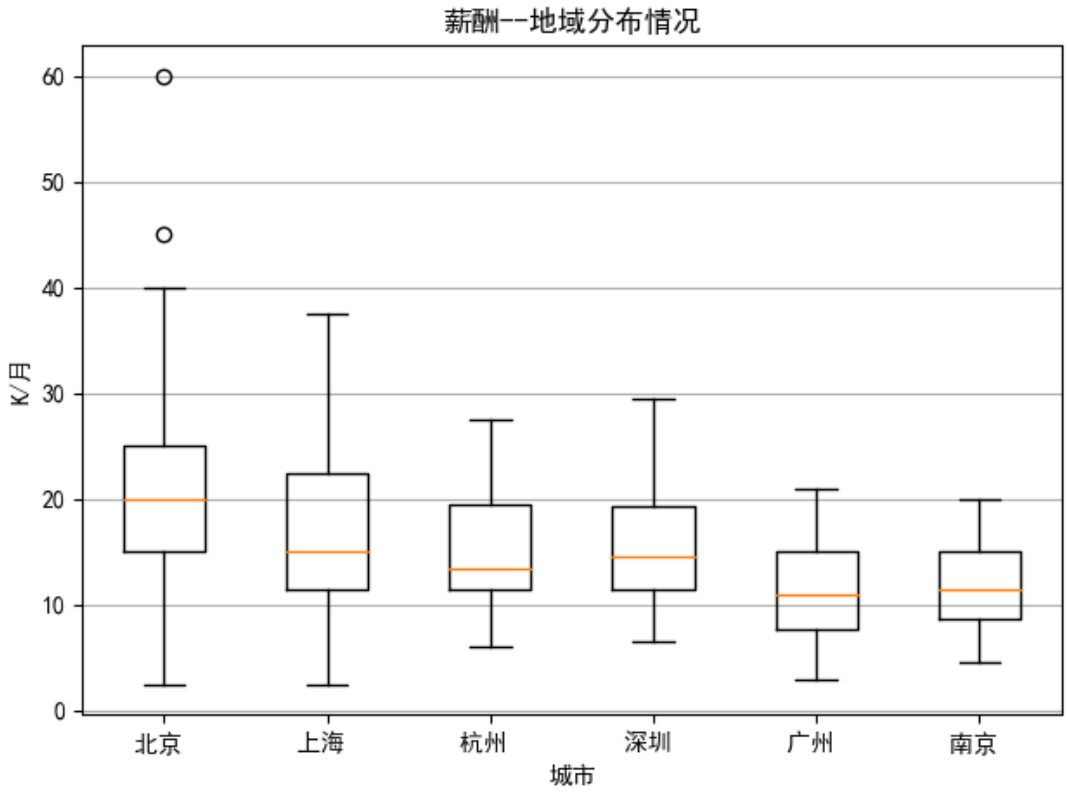
可以看出，数据分析职位的薪酬大多分布在 5K~25K 之间，还是挺令人期待的，同时也有 30K、40K 甚至 60K 的高薪情况，说明该职位后续的发展目前来看也是很令人鼓舞的，整体上，从薪酬来看，数据分析职位还是一个比较好的工作。

第四步：分析地域—薪酬的分布情况

```
def salary_area(data_in):
    data_in['average_salary'] = data_in['薪资区间'].str.split('-').map(lambda t: (int(t[0][:-1]) + int(t[1][:-1]))/2)
    #print(data_in['average_salary'])
    data = data_in.groupby("工作地点")['average_salary']
    city = data_in['工作地点'].value_counts().index[0:6]
    salary = []
```

地域主要选取的职位数量靠前的北上广深杭南，其他地方由于职位数量较少，这里不做分析讨论。

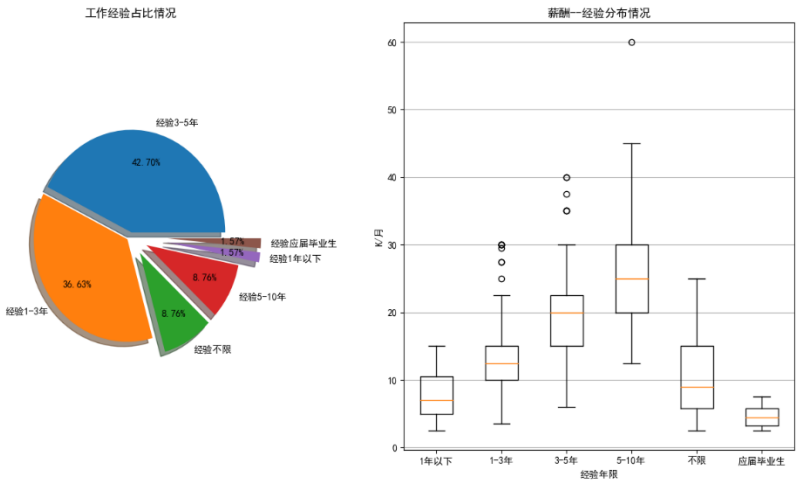
```
for key in city:
    salary.append(data.get_group(key).values)
#print(salary)
plt.figure()
ax = plt.subplot(1,1,1)
plt.title("薪酬—地域分布情况")
plt.boxplot(salary)
ax.set_xticklabels(city)
plt.xlabel("城市")
plt.ylabel('K/月')
plt.grid(True, axis='y')
plt.tight_layout()
plt.savefig(os.path.join(data_out_path, 'salary_area.png'))
```



从结果来看，北京的整体薪酬较高，其中位数薪酬 20K，居六个城市之首，上海和深圳大约 15K 左右，杭州、广州和南京相对而言稍微低一些（joking:也难怪，北上深的房价也是全国之首，所以压力大啊。。。），目前分析来看，数据分析师在北京是一个不错的选择，还有几个薪酬较高的异常值值得争取和期待（明天买机票飞北京!!!）。但上海的整体薪酬分布情况和北京差距不是很大，所以上海也是数据分析师一个比较不错的选择。

第五步：分析该职位工作经验和薪酬情况

```python
def experience(data_in):
    raw_data = data_in['工作经验'].value_counts()
    explode = 0.01 / raw_data.values * data_in['工作经验'].count() + [0,0,0,0,-0.4,-0.4]
    data_in['average_salary'] = data_in['薪资区间'].str.split('-').map(lambda t: (int(t[0][:-1]) + int(t[1][:-1])) / 2)
    # print(data_in['average_salary'])
    experience_salary = data_in.groupby("工作经验")['average_salary']
    dic = {'经验1年以下': 1, '经验1-3年': 2, '经验3-5年': 3, '经验5-10年': 4, '经验不限': 5, '经验应届毕业生': 6}
    #print(dic_new)
    salary = []
    x_labels = []
    for x in dic.keys():
        salary.append(experience_salary.get_group(x).values)
        x_labels.append(x[2:])
    plt.figure(figsize=(16,8))
    ax1 = plt.subplot(1,2,1)
    plt.title("工作经验占比情况")
    plt.pie(raw_data, labels=raw_data.index, autopct='%1.2f%%', shadow=True, radius=0.6, explode=explode,
            startangle=0, pctdistance=0.7)
    ax2 = plt.subplot(1,2,2)
    plt.title('薪酬—经验分布情况')
    plt.boxplot(salary)
    ax2.set_xticklabels(x_labels)
    plt.xlabel("经验年限")
    plt.ylabel('K/月')
    plt.grid(True, axis='y')
    plt.savefig(os.path.join(data_out_path, 'experience.png'))
```



分析结果也反映了一定的市场行情，工作经验在 1-3 年和 3-5 年的需求量最大，1 年以下的需求量很少（所以对于转行来说，这是个无解的硬伤啊 T__T）。其次 5-10 年的经验需求量也不是很大，说明数据分析是一个偏向年轻化的职业，5-10 年的职位更多应该是具有数据分析经验的管理人员（所以转行要趁早，努力学，相信还是有机会的，鼓励下自己!!!）。

其次从经验—薪酬分布来看，整体的薪酬涨幅随着工作经验的增加还是挺可观的，有个一两年的工作经验，薪酬待遇可以翻一番，而且 1-5 年工作经验来看，还有个别机会争取异常的高薪。

第六步：分析数据分析职位所属的行业类别

由于拉勾网数据对行业类别字段信息划分不是很清晰，有些类别数据也可以当做职位技能要

求做参考，诸如 SQL、Hadoop、算法等。

```
def get_key_words(text):
    words = jieba.analyse.extract_tags(text, topK=20)
    return words
```

首先对‘类别’字段信息进行分词提取，存入到文本文件中，用于词云绘图使用。

```
def word(data):
    data['text'] = data['类别'].apply(get_key_words).map(lambda t:','.join(t))
    #print(data['text'])
    text_name = 'keywords.txt'
    text_path = os.path.join(data_out_path,text_name)
    #print(text_path)
    file = open(text_path, 'w')
    for word in data['text']:
        #print(word)
        file.writelines(word+',')

    file.close()
    Alice_mask = np.array(Image.open(pic))
    f = open(text_path).read().replace('数据分析','')\
        .replace('数据挖掘','').replace('数据','')
    #print(f)
    wordcloud = WordCloud(font_path=font_path, collocations=False, background_color="white",
                          mask=Alice_mask, scale=2,
                          prefer_horizontal=0.8).generate(f)
    #wordcloud = WordCloud(width=800, height=400, background_color="white")
    plt.figure()
    plt.imshow(wordcloud,interpolation='bilinear')
    plt.axis("off")
    plt.savefig(os.path.join(data_out_path, 'word.png'), dpi=500)
```

从词云分析结果来看，大多数公司为互联网和金融企业，同时他们也将自己较多的分类

为 BI 和 SPSS 这两类，同时 BI 工具和 SPSS 统计软件也是数据分析行业经常使用的两大商业化利器，说明往数据分析方向转，BI 和 SPSS 也是多少要了解一下的。其他的一些电商、教育、健康、风控、产品、广告等领域也是有较多的数据分析需求，说明数据分析职位分布的行业还是很广泛的，能够结合具体的数据分析技术和自己所处的行业业务知识。同时一些诸如算法、可视化、数据 ETL、NLP、SQL、Hadoop 等领域也需要很多数据分析师，而这些信息同时也是一些岗位技能专有名词，说明要做好一名数据分析师，需要掌握的技能还是不少的，其实具体的职位技能可以通过爬去网上具体技能要求进行更加有针对性的分析，总之，会的越多，机会越多，努力学习，天天向上。

总结结论：

1、 数据分析职位需求量主要集中分布在北上广深还有杭州；（选取地方很重要，机会多）

2、 整体薪酬还是不错的，大多分布在 5~25K 之间，也有寻求高薪的机会，只要够努力；（职位待遇不错，有努力有回报，良性循环）

3、 北京和上海是数据分析师不错的选择，杭州和深圳从待遇上来看也不错；

4、 数据分析行业大多数需要有行业经验的人才，且集中在 1~5 年之间；（市场标准，所以转行要努力趁早）

5、 行业经验愈多，待遇也愈多，且不同年限的待遇增长也较为可观；

6、 行业多集中在互联网和金融，同时还需要掌握大量的职业技能，诸如 BI、SPSS 工具，SQL、Hadoop 等数据 ETL 处理，算法、NLP、可视化等专业化技能，当然由于类别信息交叉集很多，各个行业的数据分析还有各自专业化的技能需求，具体行业方向和技能方向还需特定数据进行分析。（行业和技能方向）