In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

plt.rcParams["font.sans-serif"]=["simhei"]
```

In [2]:
```
df=pd.read_csv("lagou.csv")
```

In [3]:
```
df.head()
```

Out[3]:

	web-scraper-order	web-scraper-start-url	name	company	salary	demand	info
0	1609581844-342	https://www.lagou.com/beijing-zhaopin/shangyes...	外卖商业分析师	美团	20k-40k	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上
1	1609581844-332	https://www.lagou.com/beijing-zhaopin/shangyes...	高级商业分析师（策略）	美团	20k-40k	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上
2	1609581874-484	https://www.lagou.com/beijing-zhaopin/shangyes...	商业分析师	字节跳动	20k-40k	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上
3	1609581838-307	https://www.lagou.com/beijing-zhaopin/shangyes...	商业分析师-门票度假	美团	20k-30k	20k-30k\n经...	消费生活 / 上市公司 / 2000人以上
4	1609581874-483	https://www.lagou.com/beijing-zhaopin/shangyes...	商业分析	字节跳动	20k-40k	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上

1.数据处理

In [4]:
```
df=df.iloc[:,2:]
df
```

Out[4]:

	name	company	salary	demand	info
0	外卖商业分析师	美团	20k-40k	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上
1	高级商业分析师（策略）	美团	20k-40k	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上
2	商业分析师	字节跳动	20k-40k	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上

	name	company	salary	demand	info
3	商业分析师-门票度假	美团	20k-30k	20k-30k\n经...	消费生活 / 上市公司 / 2000人以上
4	商业分析	字节跳动	20k-40k	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上
...	...	...	...	...	...
257	商业数据分析师	滴滴	20k-40k	20k-40k\n经...	汽车丨出行 / 不需要融资 / 2000人以上
258	美团优选-商业分析师	美团	25k-50k	25k-50k\n经...	消费生活 / 上市公司 / 2000人以上
259	商业分析师-用户增长	猿辅导	15k-30k	15k-30k\n经...	移动互联网,教育 / D轮及以上 / 2000人以上
260	商业分析师-预算管理方向	美团	20k-40k	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上
261	商业分析高级经理(J15855)	美菜	25k-40k	25k-40k\n经...	电商 / D轮及以上 / 2000人以上

262 rows × 5 columns

In [5]:
```
df["salary"]=df["salary"].str.replace("k","")
```

In [6]:
```
df["salary"]
```

Out[6]:
```
0      20-40
1      20-40
2      20-40
3      20-30
4      20-40
       ...
257    20-40
258    25-50
259    15-30
260    20-40
261    25-40
Name: salary, Length: 262, dtype: object
```

In [7]:
```
df["salary_min"]=df["salary"].str.split("-").str[0].astype(int)
```

In [8]:
```
df["salary_max"]=df["salary"].str.split("-").str[1].astype(int)
```

In [9]:
```
df.head()
```

Out[9]:

	name	company	salary	demand	info	salary_min	salary_max
0	外卖商业分析师	美团	20-40	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上	20	40
1	高级商业分析师(策略)	美团	20-40	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上	20	40
2	商业分析师	字节跳动	20-40	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上	20	40

	name	company	salary	demand	info	salary_min	salary_max
3	商业分析师-门票度假	美团	20-30	20k-30k\n经...	消费生活 / 上市公司 / 2000人以上	20	30
4	商业分析	字节跳动	20-40	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上	20	40

In [10]:
```
df["salary_avg"]=(df["salary_min"]+df["salary_max"])/2
```

In [11]:
```
df.head()
```

Out[11]:

	name	company	salary	demand	info	salary_min	salary_max	salary_avg
0	外卖商业分析师	美团	20-40	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上	20	40	30.0
1	高级商业分析师（策略）	美团	20-40	20k-40k\n经...	消费生活 / 上市公司 / 2000人以上	20	40	30.0
2	商业分析师	字节跳动	20-40	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0
3	商业分析师-门票度假	美团	20-30	20k-30k\n经...	消费生活 / 上市公司 / 2000人以上	20	30	25.0
4	商业分析	字节跳动	20-40	20k-40k\n经...	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0

In [12]:
```
df["demand"]=df["demand"].str.replace(" ","").str.split("\n").str[1]
```

In [13]:
```
df["demand"]
```

Out[13]:
```
0       经验3-5年/本科
1       经验3-5年/本科
2       经验3-5年/本科
3       经验1-3年/本科
4       经验3-5年/本科
        ...
257     经验3-5年/本科
258    经验5-10年/本科
259     经验1-3年/本科
260     经验3-5年/本科
261    经验5-10年/本科
Name: demand, Length: 262, dtype: object
```

In [14]:
```
df["exp_demand"]=df["demand"].str.split("/").str[0]
```

In [15]:
```
df["edu_demand"]=df["demand"].str.split("/").str[1]
```

In [16]:
```
df.head()
```

Out[16]:

	name	company	salary	demand	info	salary_min	salary_max	salary_avg	exp_demand	edu_d
0	外卖商业分析师	美团	20-40	经验3-5年/本科	消费生活 / 上市公司 / 2000人以上	20	40	30.0	经验3-5年	
1	高级商业分析师（策略）	美团	20-40	经验3-5年/本科	消费生活 / 上市公司 / 2000人以上	20	40	30.0	经验3-5年	
2	商业分析师	字节跳动	20-40	经验3-5年/本科	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0	经验3-5年	
3	商业分析师-门票度假	美团	20-30	经验1-3年/本科	消费生活 / 上市公司 / 2000人以上	20	30	25.0	经验1-3年	
4	商业分析	字节跳动	20-40	经验3-5年/本科	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0	经验3-5年	

In [17]:
```
df["sort"]=df["info"].str.replace(" ","").str.split("/").str[0]
```

In [18]:
```
df["finance"]=df["info"].str.replace(" ","").str.split("/").str[1]
```

In [19]:
```
df["size"]=df["info"].str.replace(" ","").str.split("/").str[2]
```

In [20]:
```
df.head()
```

Out[20]:

	name	company	salary	demand	info	salary_min	salary_max	salary_avg	exp_demand	edu_d
0	外卖商业分析师	美团	20-40	经验3-5年/本科	消费生活 / 上市公司 / 2000人以上	20	40	30.0	经验3-5年	
1	高级商业分析师（策略）	美团	20-40	经验3-5年/本科	消费生活 / 上市公司 / 2000人以上	20	40	30.0	经验3-5年	
2	商业分析师	字节跳动	20-40	经验3-5年/本科	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0	经验3-5年	
3	商业分析师-门票度假	美团	20-30	经验1-3年/本科	消费生活 / 上市公司 / 2000人以上	20	30	25.0	经验1-3年	
4	商业分析	字节跳动	20-40	经验3-5年/本科	文娱丨内容 / D轮及以上 / 2000人以上	20	40	30.0	经验3-5年	

In [21]:
```
df.drop(["salary","demand","info"],axis=1,inplace=True)
```

In [22]:
```
df.head()
```

Out[22]:

	name	company	salary_min	salary_max	salary_avg	exp_demand	edu_demand	sort	finance
0	外卖商业分析师	美团	20	40	30.0	经验3-5年	本科	消费生活	上市公司

	name	company	salary_min	salary_max	salary_avg	exp_demand	edu_demand	sort	finance	
1	高级商业分析师(策略)	美团	20	40	30.0	经验3-5年	本科	消费生活	上市公司	2
2	商业分析师	字节跳动	20	40	30.0	经验3-5年	本科	文娱丨内容	D轮及以上	2
3	商业分析师-门票度假	美团	20	30	25.0	经验1-3年	本科	消费生活	上市公司	2
4	商业分析	字节跳动	20	40	30.0	经验3-5年	本科	文娱丨内容	D轮及以上	2

In [23]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 262 entries, 0 to 261
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype  
---  ------      --------------  -----  
 0   name        262 non-null    object 
 1   company     262 non-null    object 
 2   salary_min  262 non-null    int32  
 3   salary_max  262 non-null    int32  
 4   salary_avg  262 non-null    float64
 5   exp_demand  262 non-null    object 
 6   edu_demand  262 non-null    object 
 7   sort        262 non-null    object 
 8   finance     262 non-null    object 
 9   size        262 non-null    object 
dtypes: float64(1), int32(2), object(7)
memory usage: 18.5+ KB
```

In [24]:
```
df.describe()
```

Out[24]:

	salary_min	salary_max	salary_avg
count	262.000000	262.000000	262.000000
mean	21.450382	38.423664	29.937023
std	7.883455	14.261108	10.855320
min	2.000000	3.000000	2.500000
25%	15.000000	30.000000	22.500000
50%	20.000000	40.000000	30.000000

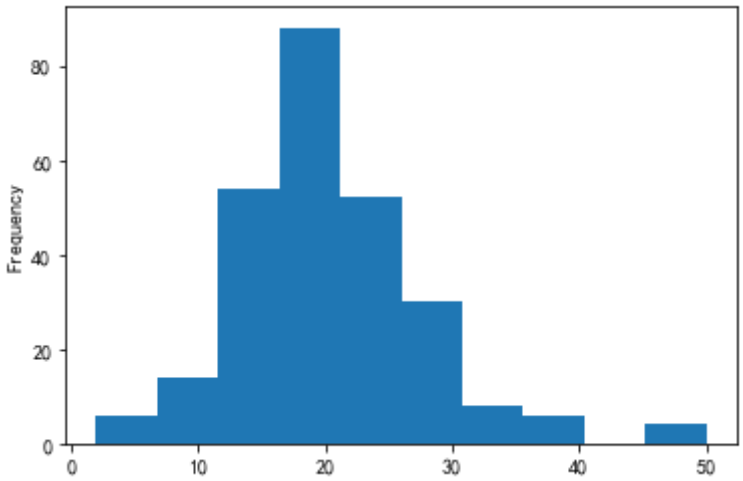
	salary_min	salary_max	salary_avg
75%	25.000000	45.000000	35.750000
max	50.000000	100.000000	75.000000

2.单变量分析

2.1数值变量

2.1.1 salary_min

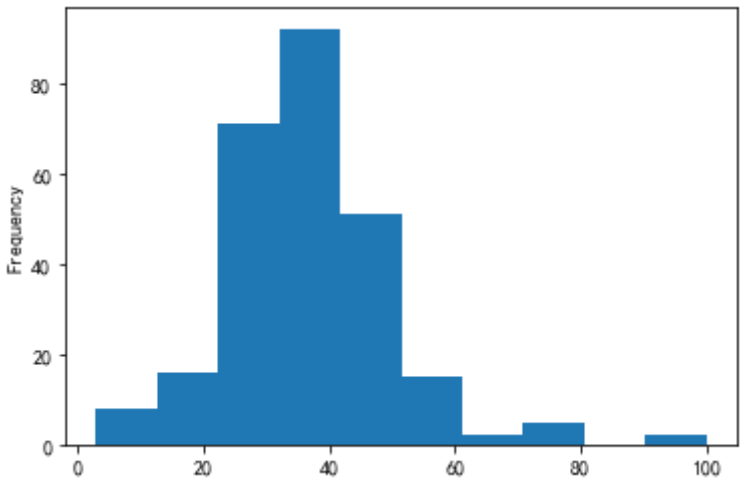
In [25]:
```
df["salary_min"].plot(kind="hist");
```



·最低工资集中在15k-30k之间
·有少数在45k以上

2.1.2 salary_max

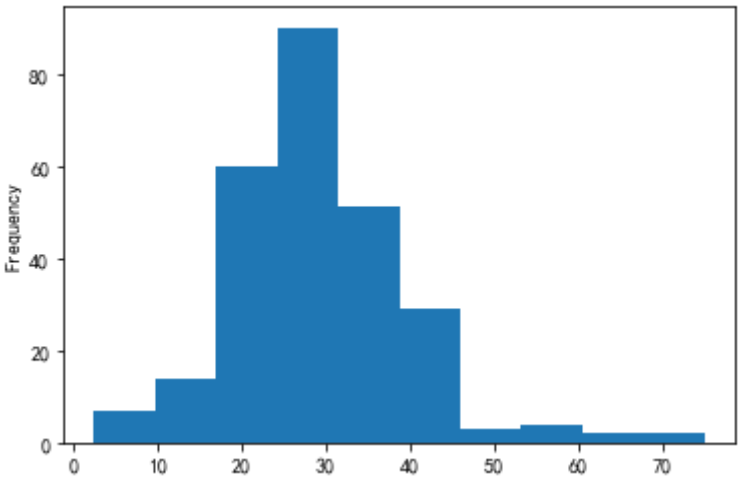
In [26]:
```
df["salary_max"].plot(kind="hist");
```



·最高工资集中分布在25k-50k之间
·有极少数高达90k以及以上

2.1.3 salary_avg

In [27]:
```
df["salary_avg"].plot(kind="hist");
```



·平均工资集中分布在20k-45k之间

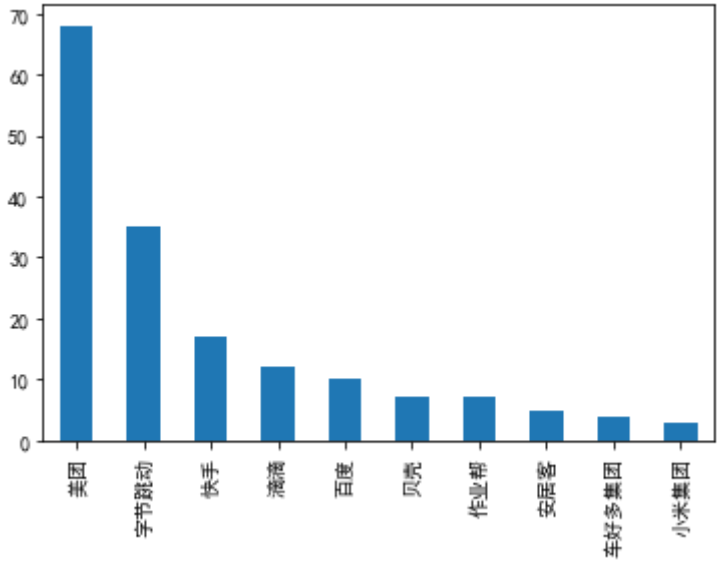
2.2 分类变量

2.2.1 company

In [28]:
```
df["company"].value_counts()
```

Out[28]:
```
美团          68
字节跳动        35
快手          17
滴滴          12
百度          10
            ..
慧科集团         1
联通大数据        1
Zenjoy      1
新片场          1
理想汽车         1
Name: company, Length: 83, dtype: int64
```

In [29]:
```
df["company"].value_counts().head(10).plot(kind="bar");
```



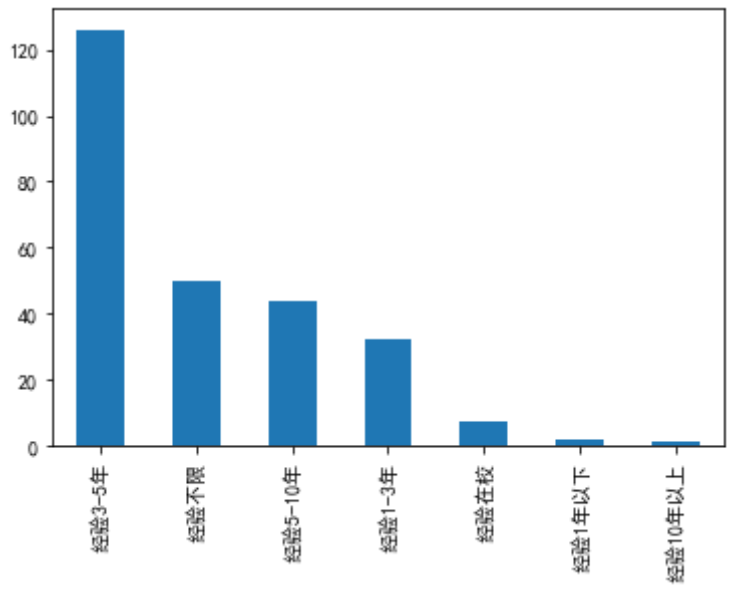
·招聘职位需求排名前10的公司如图所示
·其中美团需求最大，放出约70个商业分析相关的岗位，约是第二名的两倍
·其次是字节跳动、快手、滴滴以及百度等互联网公司
·第五名之后的企业放出岗位均为个位数，表明绝大多数公司放出的商业分析相关岗位极少

2.2.2 exp_demand

In [30]:
```
df["exp_demand"].value_counts()
```

Out[30]:
```
经验3-5年      126
经验不限        50
经验5-10年      44
经验1-3年       32
经验在校         7
经验1年以下       2
经验10年以上      1
Name: exp_demand, dtype: int64
```

In [31]:
```
df["exp_demand"].value_counts().plot(kind="bar");
```



·绝大多数要求有3-5年的工作经验

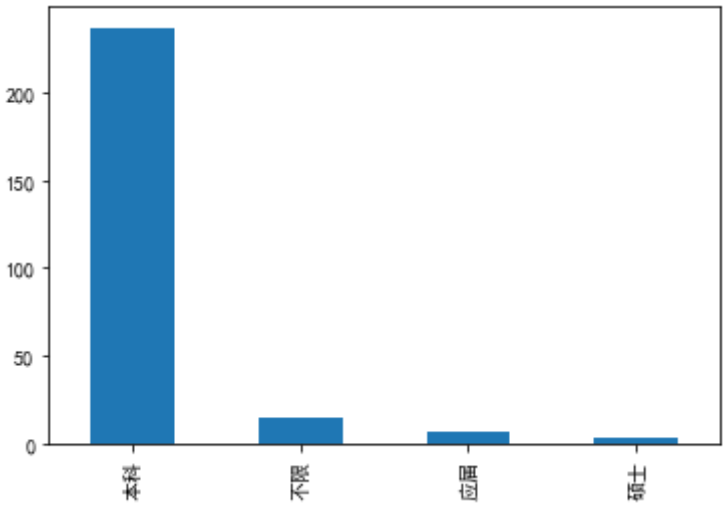
·不限经验的和要求有5-10年工作经验的差不多
·要求工作经验在1年一下和10年以上的几乎没有

2.2.3 edu_demand

In [32]:
```
df["edu_demand"].value_counts()
```

Out[32]:
```
本科     237
不限      15
应届       7
硕士       3
Name: edu_demand, dtype: int64
```

In [33]:
```
df["edu_demand"].value_counts().plot(kind="bar");
```



·绝大多数只要求拥有本科学历即可

2.2.3 sort

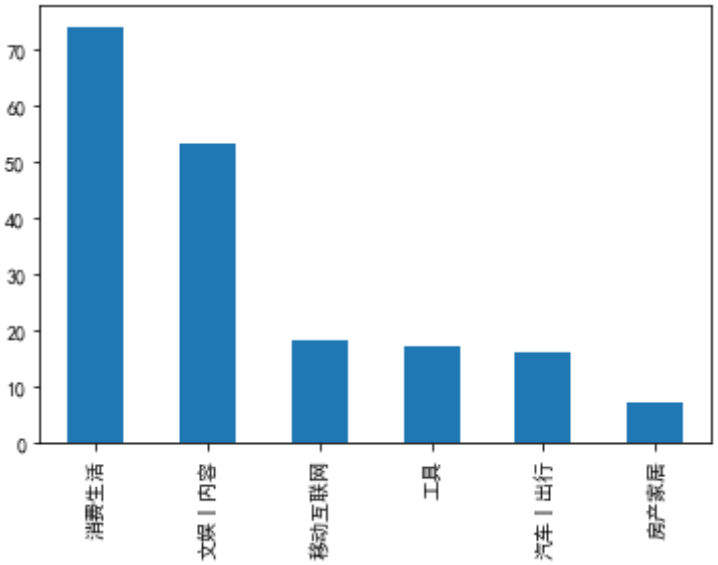
In [34]:
```
df["sort"].value_counts()
```

Out[34]:
```
消费生活                  74
文娱丨内容                 53
移动互联网                 18
工具                    17
汽车丨出行                 16
房产家居                   7
电商                      7
移动互联网,电商                6
移动互联网,教育                6
教育                      4
旅游                      4
社交                      4
数据服务                   4
移动互联网,企业服务             4
企业服务                   3
移动互联网,数据服务             3
硬件                      3
企业服务,数据服务              2
其他                      2
电商,消费生活                 2
广告营销                   2
移动互联网,广告营销             2
物流丨运输                  2
```

```
移动互联网,社交                          2
金融                                2
移动互联网,文娱丨内容                      1
移动互联网,医疗丨健康                      1
移动互联网,消费生活                        1
物联网,数据服务                          1
电商,移动互联网                          1
移动互联网,游戏                          1
硬件,数据服务                           1
电商,企业服务                           1
企业服务,信息安全,区块链,数据服务              1
企业服务,教育                           1
教育,移动互联网                          1
硬件,电商                             1
移动互联网,人工智能                        1
Name: sort, dtype: int64
```

·数据内部有重复
·但前6个占比较大，故可只观察前6个类型的公司分布

In [35]:
```
df["sort"].value_counts().head(6).plot(kind="bar");
```



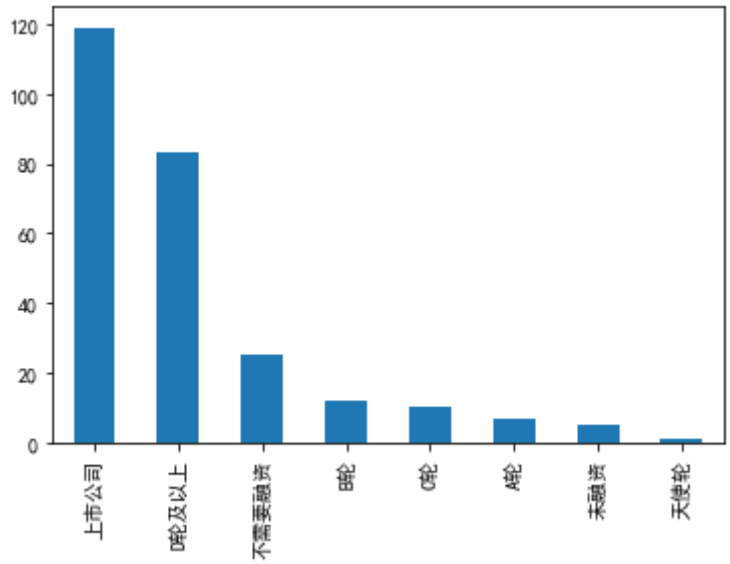
·消费生活类的公司放出的岗位最多，超过70个
·其次是文娱&内容，放出超过50个岗位
·移动互联网、工具和汽车&出行放出的岗位数量接近，约为20个

2.2.4 finance

In [36]:
```
df["finance"].value_counts()
```

Out[36]:
```
上市公司       119
D轮及以上       83
不需要融资       25
B轮           12
C轮           10
A轮            7
未融资            5
天使轮            1
Name: finance, dtype: int64
```

In [37]:
```
df["finance"].value_counts().plot(kind="bar");
```



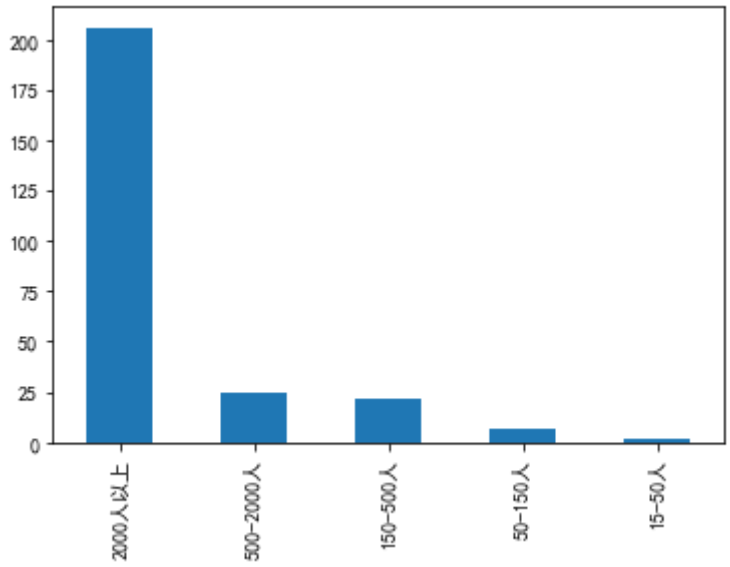
·其中最多的是上市公司和在D轮融资以上的公司，分别放出约120和80个岗位

2.2.5 size

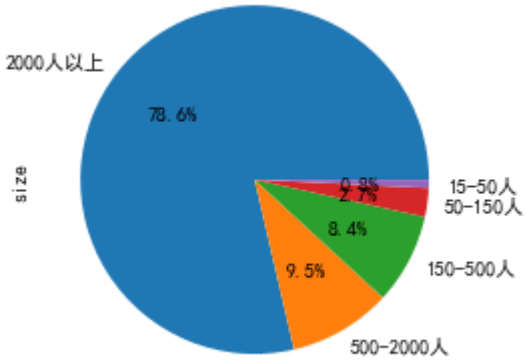
In [38]:
```
df["size"].value_counts()
```

Out[38]:
```
2000人以上      206
500-2000人      25
150-500人       22
50-150人         7
15-50人          2
Name: size, dtype: int64
```

In [39]:
```
df["size"].value_counts().plot(kind="bar");
```



In [40]:
```
df["size"].value_counts().plot(kind="pie",autopct="%1.1f%%");
```

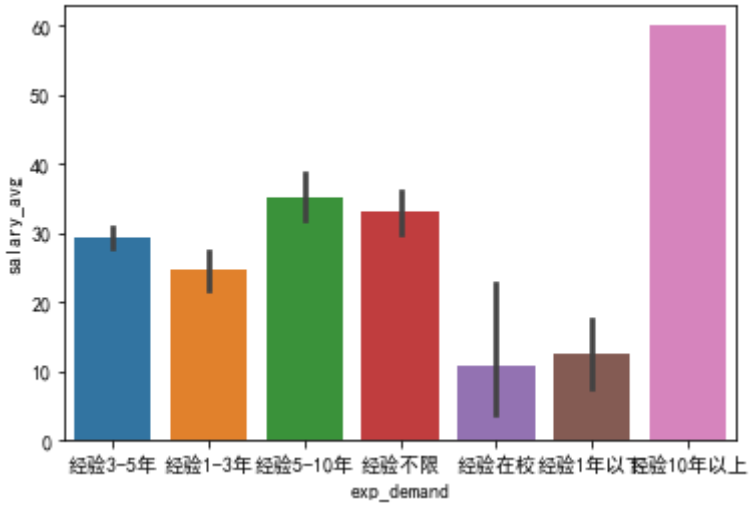


·2000人以上的大公司放出的岗位占比超过75%
·150-500和500-2000的中型公司的岗位数量接近

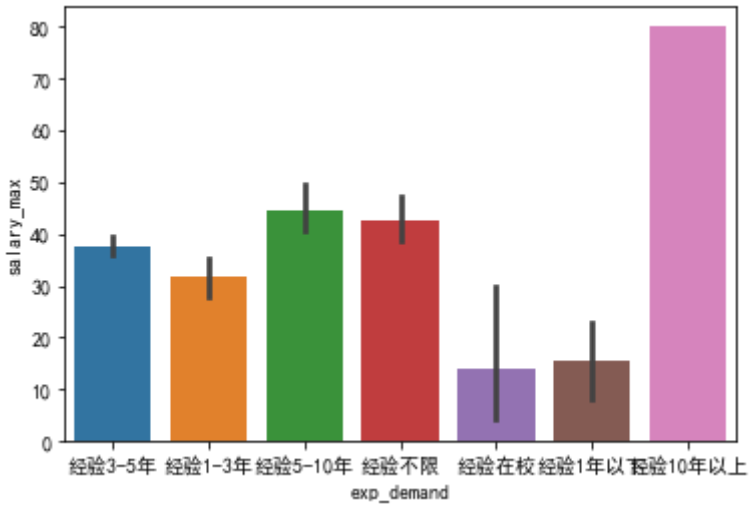
3.多变量分析

3.1 exp_demand

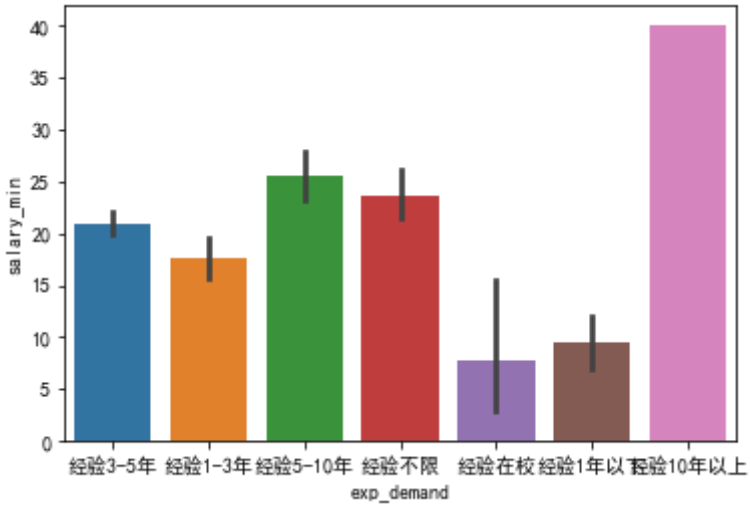
In [41]:
```
sns.barplot(x="exp_demand",y="salary_avg",data=df);
```



In [42]:
```
sns.barplot(x="exp_demand",y="salary_max",data=df);
```



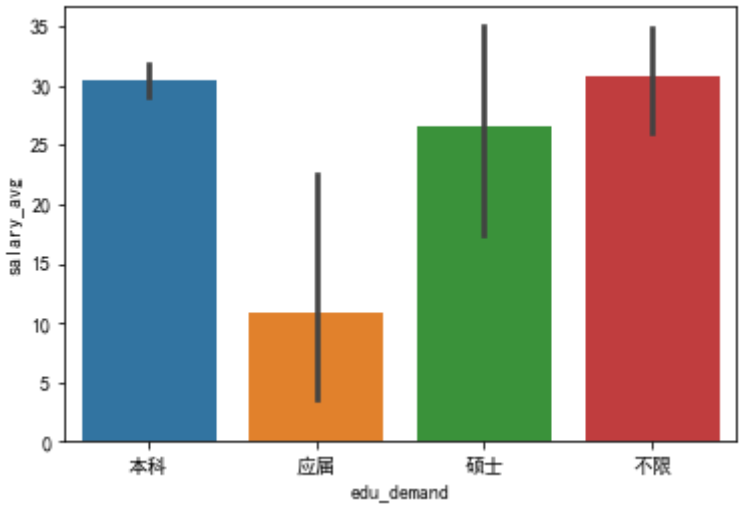
In [43]:
```
sns.barplot(x="exp_demand",y="salary_min",data=df);
```



- 三种工资分析维度在用工作经历的要求分类下趋势相同
- 要求工作经验在10年以上的岗位的平均最高、最低和平均工资均最高
- 其次是要求工作经验5-10年的岗位
- 要求工作经验3-5年的反而没有不限经验的岗位高

3.2 edu_demand

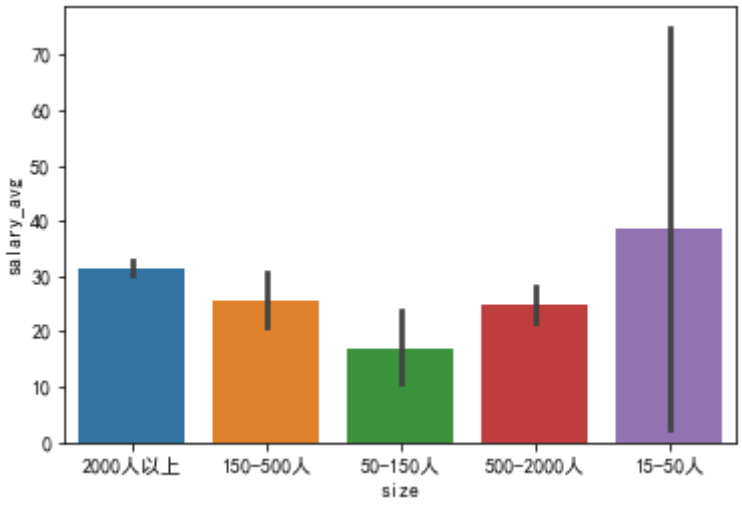
In [44]:
```
sns.barplot(x="edu_demand",y="salary_avg",data=df);
```



·各个对本科和不限学历的岗位平均工资相近
·要求硕士学历的反而低于要求本科和不限学历的岗位
·仅招收应届生的岗位平均工资最低

3.3 size

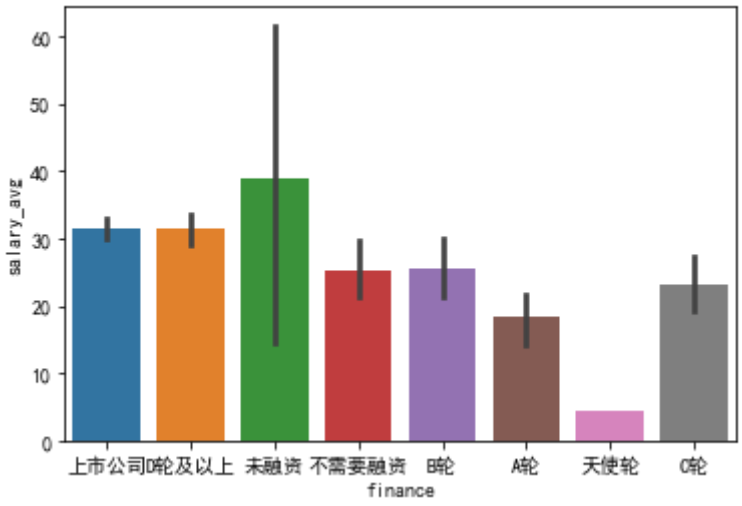
In [45]:
```
sns.barplot(x="size",y="salary_avg",data=df);
```



·2000人以上的大公司给出的平均薪资并不显著高
·15-50人的小公司类别内的平均工资相差极大，最高超过70k，最低仅有不到5k

3.4 finance

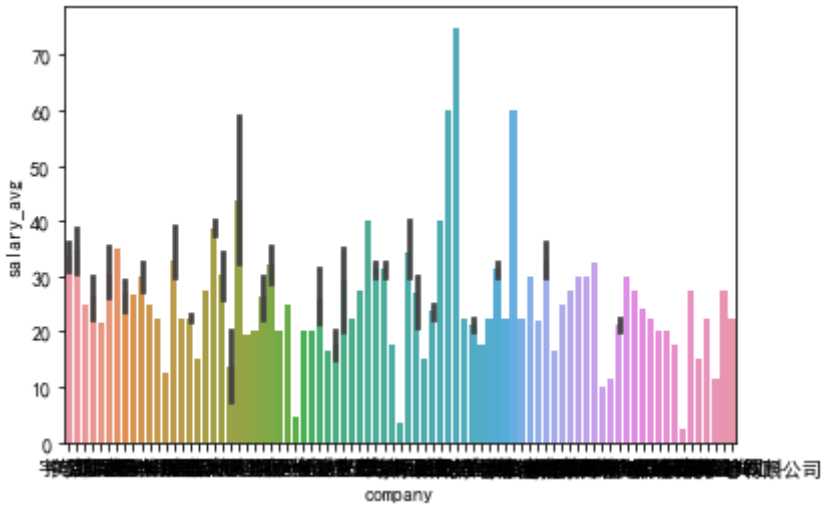
In [46]:
```
sns.barplot(x="finance",y="salary_avg",data=df);
```



·未融资的公司的岗位的平均薪资最高，但极差也最大
·上市公司和融资D轮以上公司的岗位的平均薪资接近，在30k左右
·融资B轮和不需要融资公司的岗位的平均薪资接近，在25k左右

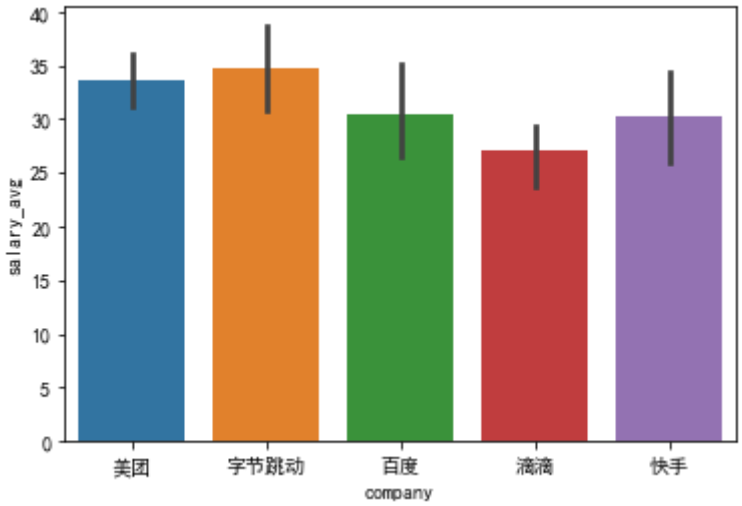
3.5 company

In [47]:
```
sns.barplot(x="company",y="salary_avg",data=df);
```



In [48]:
```
company_5=["美团","字节跳动","快手","滴滴","百度"]
dfcom5=df[df.company.isin(company_5)]
```

In [49]:
```
sns.barplot(x="company",y="salary_avg",data=dfcom5);
```

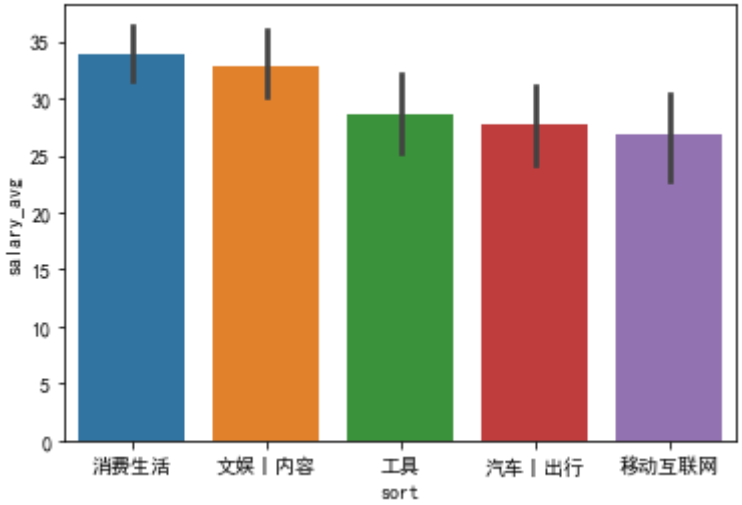


·5家公司给出的平均工资相差不大
·其中最高的是字节跳动，其次是美团和百度

3.6 sort

In [50]:
```
sort_5=["消费生活","文娱丨内容","移动互联网","工具","汽车丨出行"]
dfsort5=df[df.sort.isin(sort_5)]
```

In [51]:
```
sns.barplot(x="sort",y="salary_avg",data=dfsort5);
```



·5种类型的公司的平均工资也差距不大
·最高的是消费生活和文艺&内容
·工具、汽车&出行和移动互联网的平均工资接近

4.总结

综上所述，有：
·最低工资集中在15k-30k之间，最高工资集中分布在25k-50k之间，平均工资集中分布在20k-45k之间
·美团对商业分析师的需求最大，放出约70个商业分析相关的岗位，约是第二名的两倍，其次是字节跳动、快手、滴滴以及百度等互联网公司
·绝大多数岗位要求有3-5年的工作经验，不限经验的和要求有5-10年工作经验的差不

多，要求工作经验在1年一下和10年以上的几乎没有
·绝大多数岗位只要求拥有本科学历即可
·消费生活类的公司放出的岗位最多，超过70个，其次是文娱&内容，放出超过50个岗位，移动互联网、工具和汽车&出行放出的岗位数量接近，约为20个
·放出岗位最多的是上市公司和在D轮融资以上的公司，分别放出约120和80个岗位

·要求工作经验在10年以上的岗位的平均工资均最高，其次是要求工作经验5-10年的岗位
·而要求工作经验3-5年的工资反而没有不限经验的高
·各个对本科和不限学历的岗位平均工资相近，仅招收应届生的岗位平均工资最低
·大公司给出的平均薪资并不显著高，而小公司类别内的平均工资相差极大，最高超过70k，最低仅有不到5k
·未融资的公司的岗位的平均薪资最高，但极差也最大，上市公司和融资D轮以上公司的岗位的平均薪资接近，在30k左右，融资B轮和不需要融资公司的岗位的平均薪资接近，在25k左右
·放出岗位数量排名前5的公司给出的平均工资相差不大，其中最高的是字节跳动，其次是美团和百度
·对于公司类型，平均工资最高的是消费生活和文艺&内容类，工具、汽车&出行和移动互联网类的岗位的平均工资接近