**任务：**

1、将原始数据导入Mysql

2、Python进行数据读取，清洗，整理，分析和可视化

**一、从数据库读取数据**

1、连接数据库：create\_engine()

2、读取数据，得到DataFrame类型数据：read\_sql\_query()

**二、清洗数据**

1、删除positionId重复数据：drop\_duplicates()

2、薪资分列，计算最低和最高工资以及平均工资：find()

（1）以‘-’为分界点

（2）特殊情况，以‘K’为分界点

（3）数据类型转换为int型后计算平均工资

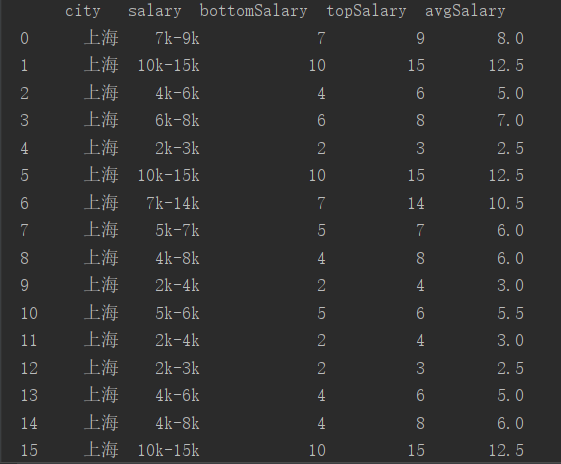


图1 薪资分列

**三、数据分析及可视化**

1、切取部分需要分析的数据字段

2、数据统计，分类数据：value\_counts()，数值数据：describe()

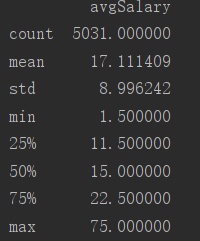
 

图2 数据统计

3、绘制avgSalary直方图：hist()

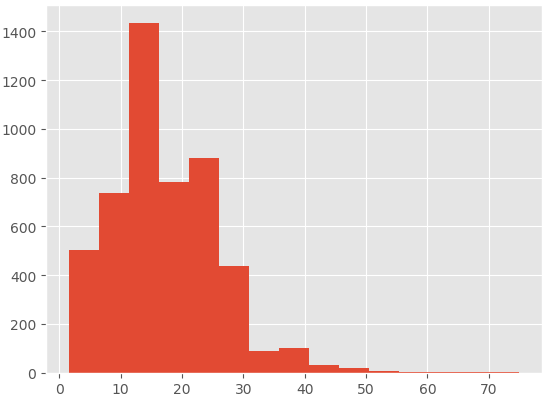


图3 avgSalary直方图

4、绘制箱线图不同城市、年限等单一变量对薪资影响：boxplot()，注：中文显示

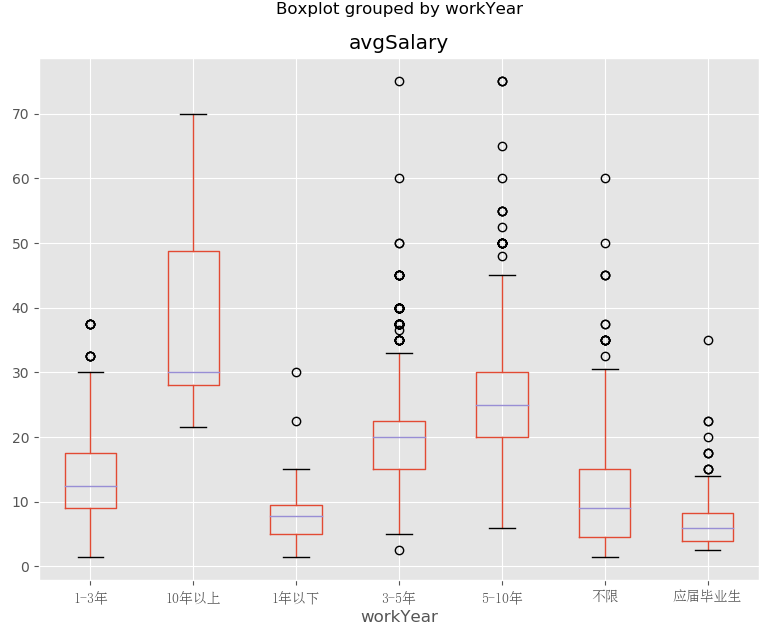
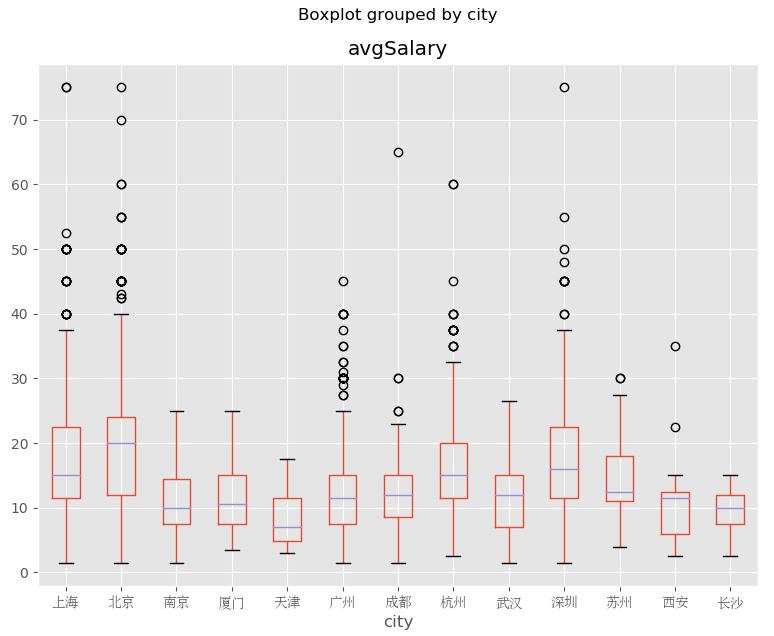


图4 不同城市、年限等单一变量对薪资影响

5、绘制箱线图不同城市及学历等多变量对薪资影响：isin()、boxplot()

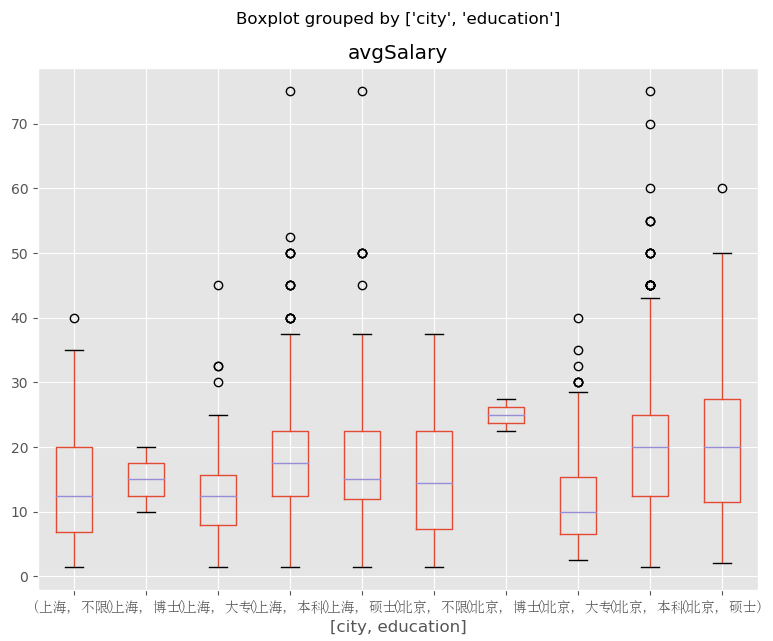


图5 不同城市及学历等多变量对薪资影响

6、多维度分析：groupby()

（1）统计不同城市不同学历的平均工资

（2）统计不同城市不同学历的招聘人数

（3）不同公司的招聘人数和平均工资统计

（4）不同公司最高薪资和最低薪资差值

（5）不同城市不同公司招聘人数前5

（6）不同城市不同招聘岗位前5

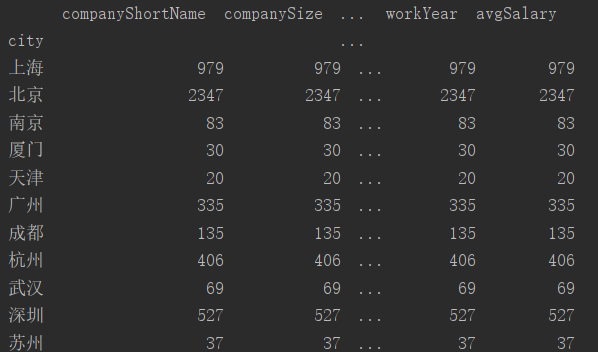
 

图6 多维度分析

7、多维度绘图：groupby()

（1）不同城市的平均工资柱形图

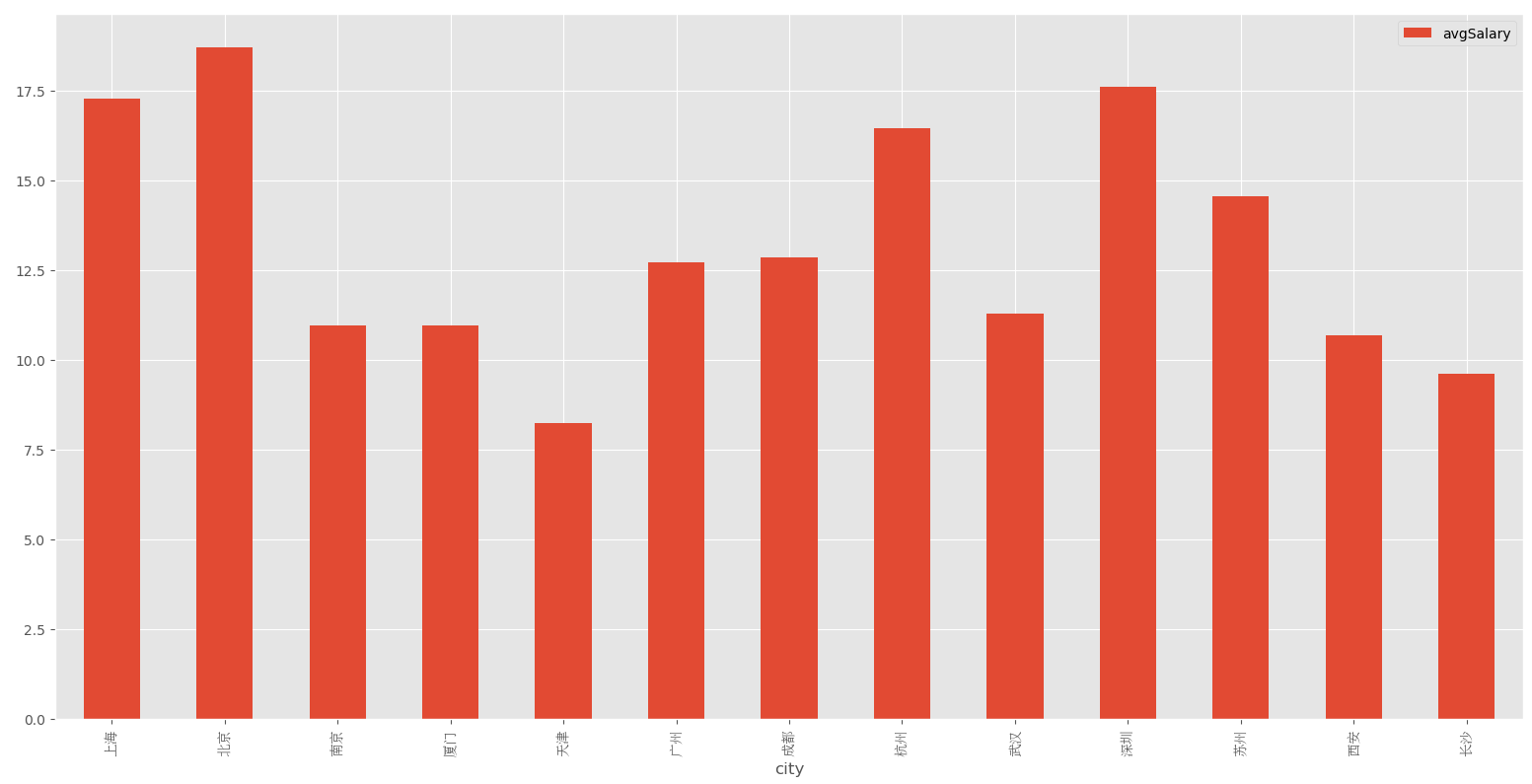


图7 不同城市的平均工资

（2）不同城市不同学历的平均工资柱形图

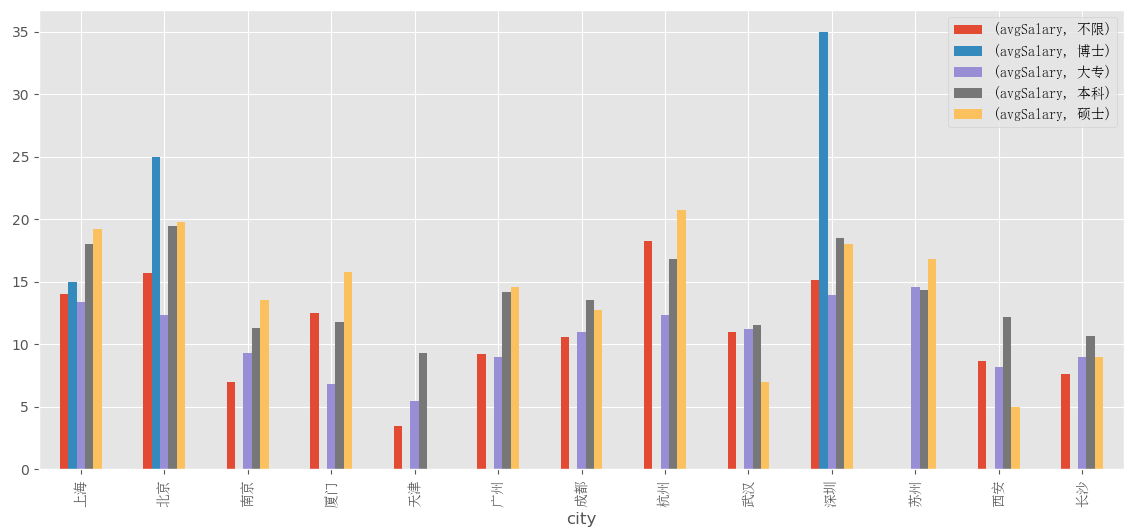


图8 不同城市不同学历的平均工资

8、数据分段并绘图：cut()

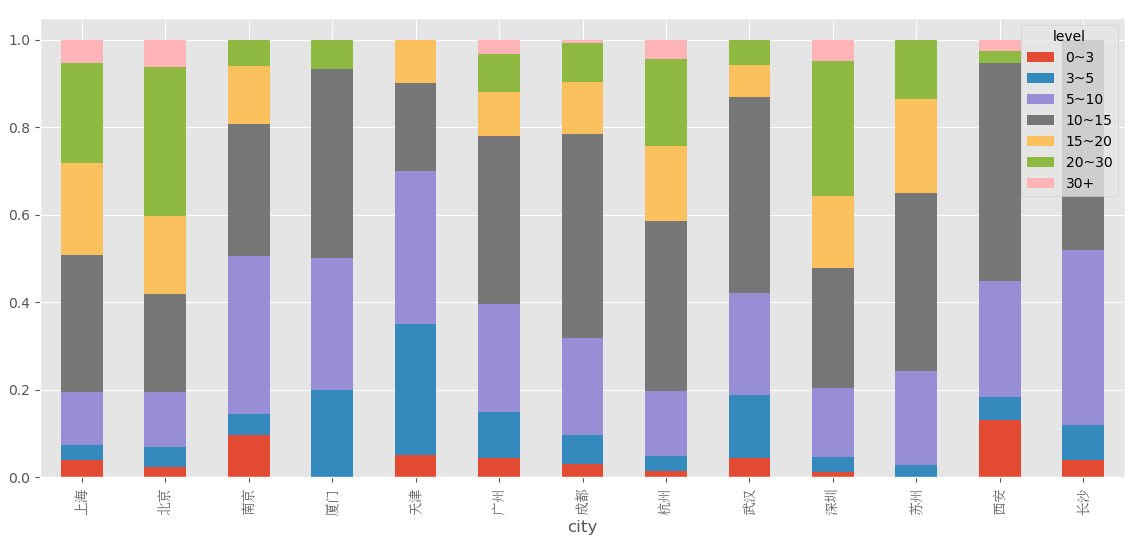


图9 数据分段

9、职位标签处理：WordCloud()

（1）去掉每组标签‘[]’：str()

（2）处理每组标签空格：replace()

（3）删掉空值：dropna()

（4）标签拆分：split()

（5）统计不同岗位标签数：apply(pandas.value\_counts)

（6）行列转换：unstack()

（7）删除空值，重置为DataFrame类型：dropna()、reset\_index()

（8）按标签分组，统计出现的次数：groupby()、count()

10、词云图片显示

（1）去掉标签引号：replace()

（2）绘图



图9 词云绘图