【参考答案】

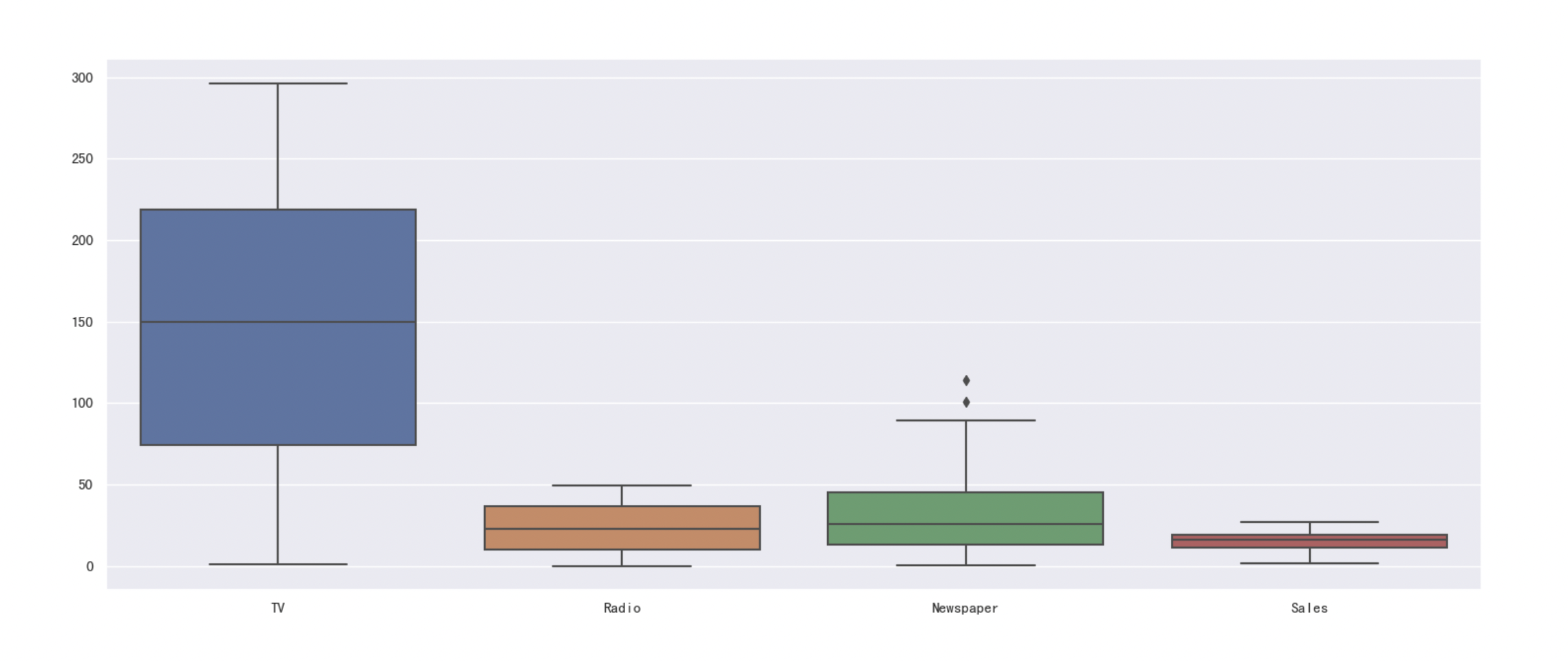
1. 连续型变量描述性统计（2分）



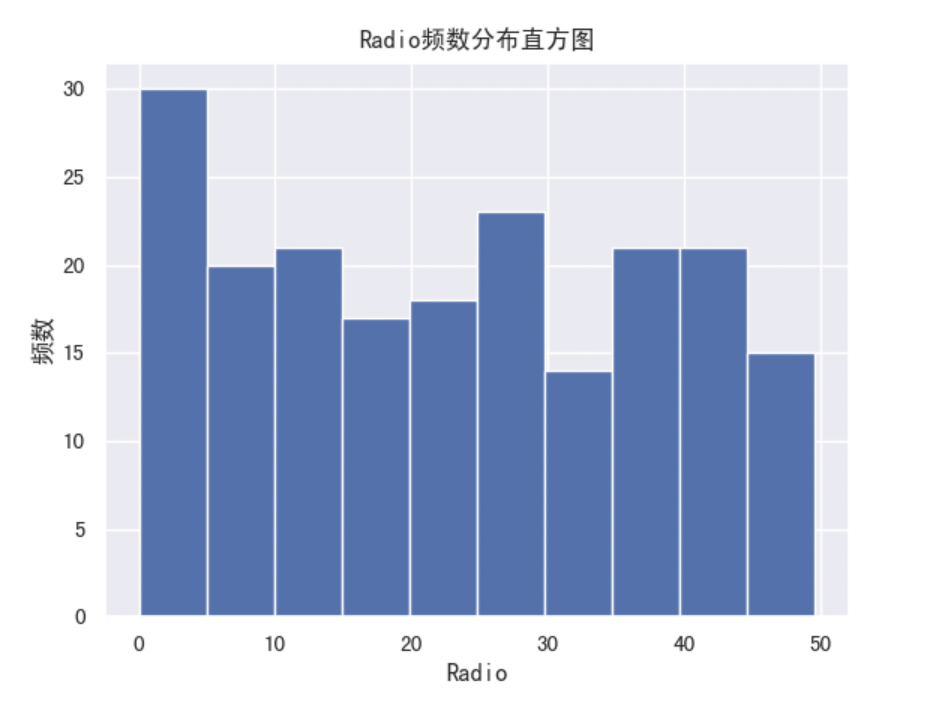
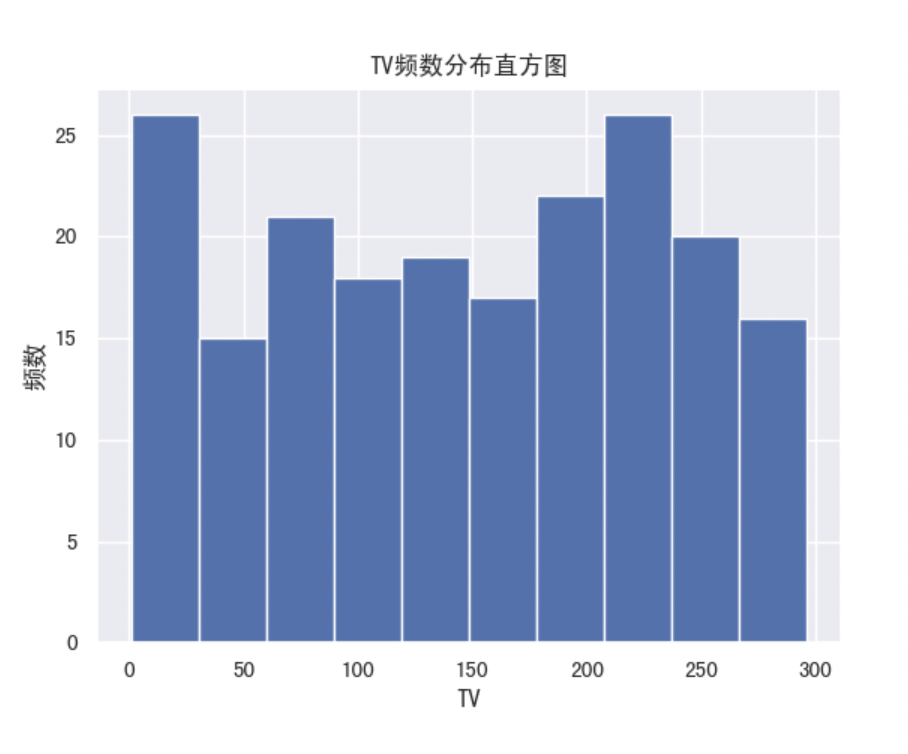
缺失值检测：销售额变量有缺失值，但由于缺失值占比较少，选择删除的方式来解决（缺失值分析2分）；采用数据加工里的全局删除（缺失值处理2分）。

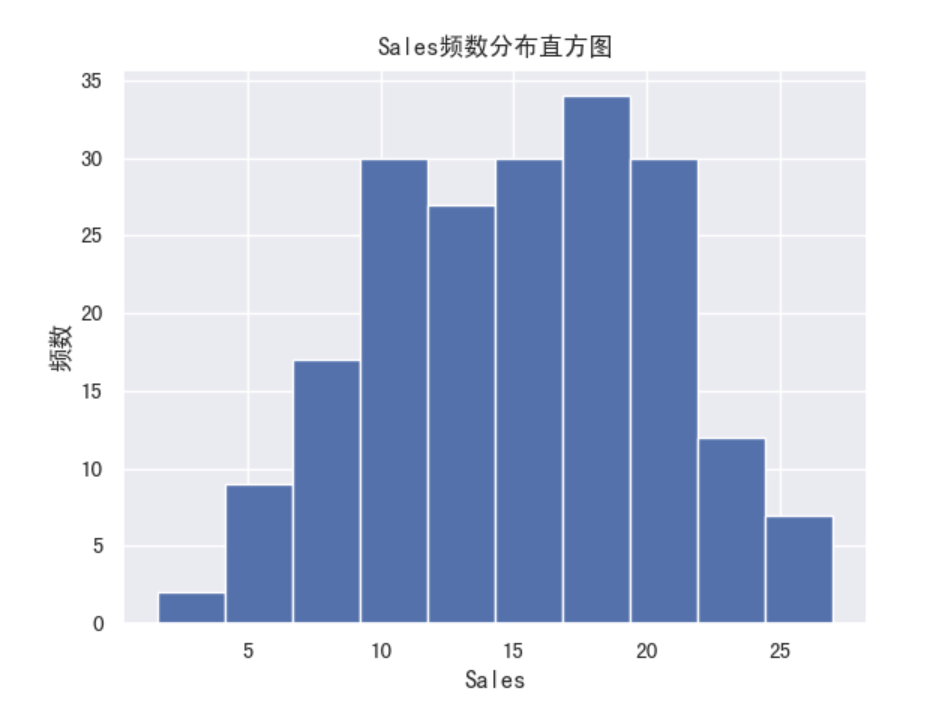
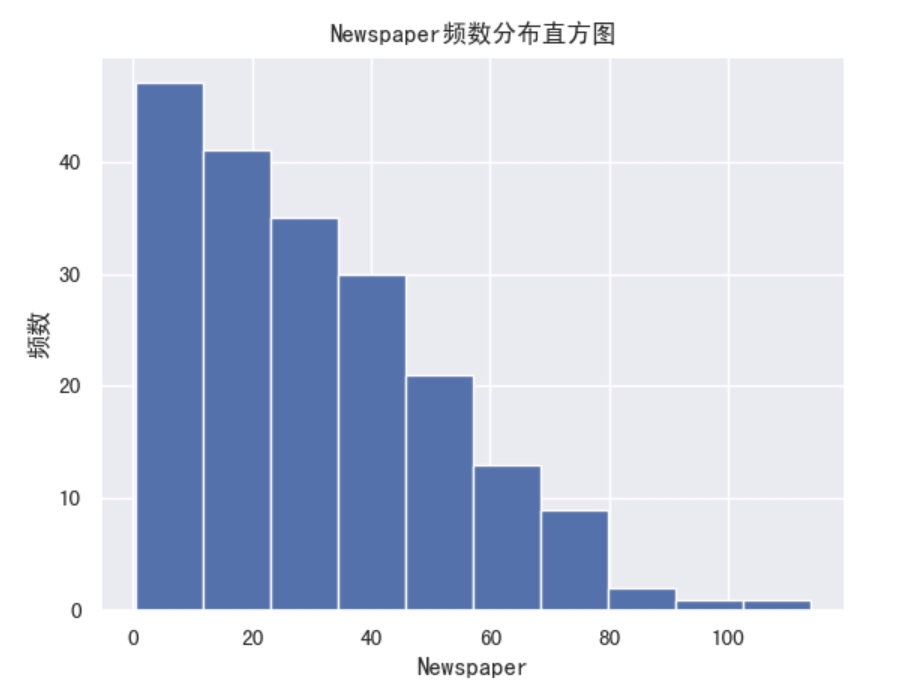


异常值检测：由箱线图可以看出，报纸广告投放费用有两个离群点，考虑真实情况，故不做异常值处理（异常值分析2分，做/不做处理，理由充分即可2分）。

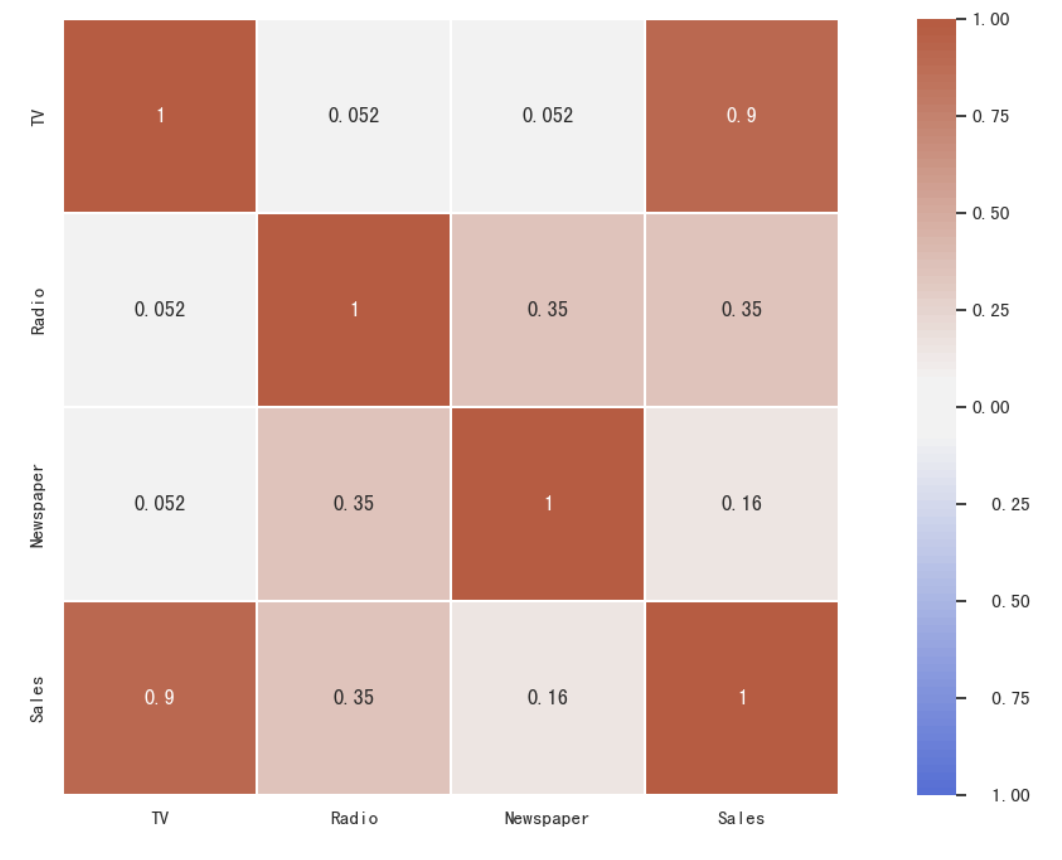


数据形态分析：通过直方图或偏度系数可知，报纸广告投放费用存在右偏，分析可知符合实际情况，为保留原始数据形态，在此不做boxcox变换（数据形态分析2分，做/不做处理，理由充分即可2分）。





（2）皮尔逊相关系数矩阵如下：

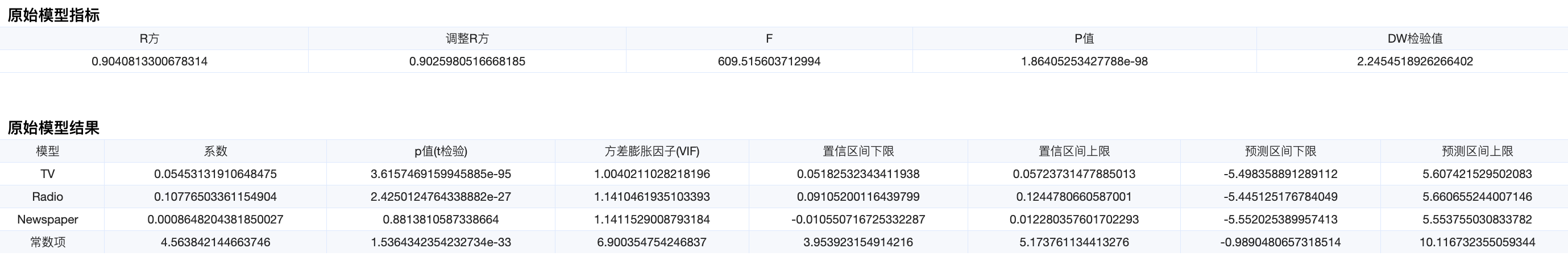


判定规则：相关系数的绝对值大于0.8时，则认为变量间存在高度相关关系；相关系数的绝对值均大于0.8且小于0.3，则认为变量间不存在中度相关关系；相关系数的绝对值小于0.3时，则认为变量间存在低度相关关系。

[TV与Sales]的相关系数值为0.9013，绝对值大于0.8，表明TV与Sales之间存在高度相关关系;[Radio与Newspaper]的相关系数值为0.3541，绝对值大于0.3且小于0.8，表明Radio与Newspaper之间存在中度相关关系;[Radio与Sales]的相关系数值为0.3489，绝对值大于0.3且小于0.8，表明Radio与Sales之间存在中度相关关系。

在上述判定规则基础上，由相关系数矩阵可以看出，所有自变量互相都呈低度相关关系，说明多重共线性问题不明显，综合考虑，选择除销售额外变量作为自变量，销售额为因变量，建立销售额的线性回归预测模型。（相关系数矩阵分析2分，多重共线性分析合理2分，变量选择合理2分）

采用原始进行回归，得到销售额的线性回归模型结果如下：（选择合适的方法建立模型原始/调优模型均可，5分）



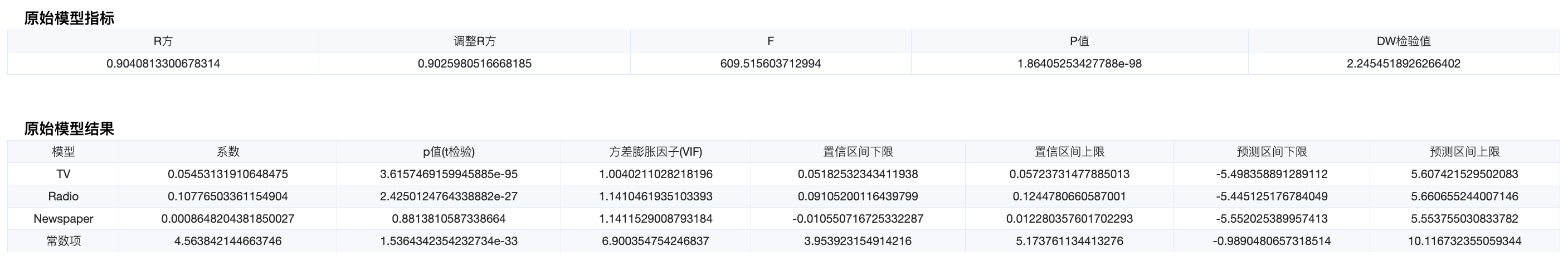
（3）模型调整R方为[0.9026]，R方>0.8表明模型拟合程度较好（调整R方2分，若写为R方1分）

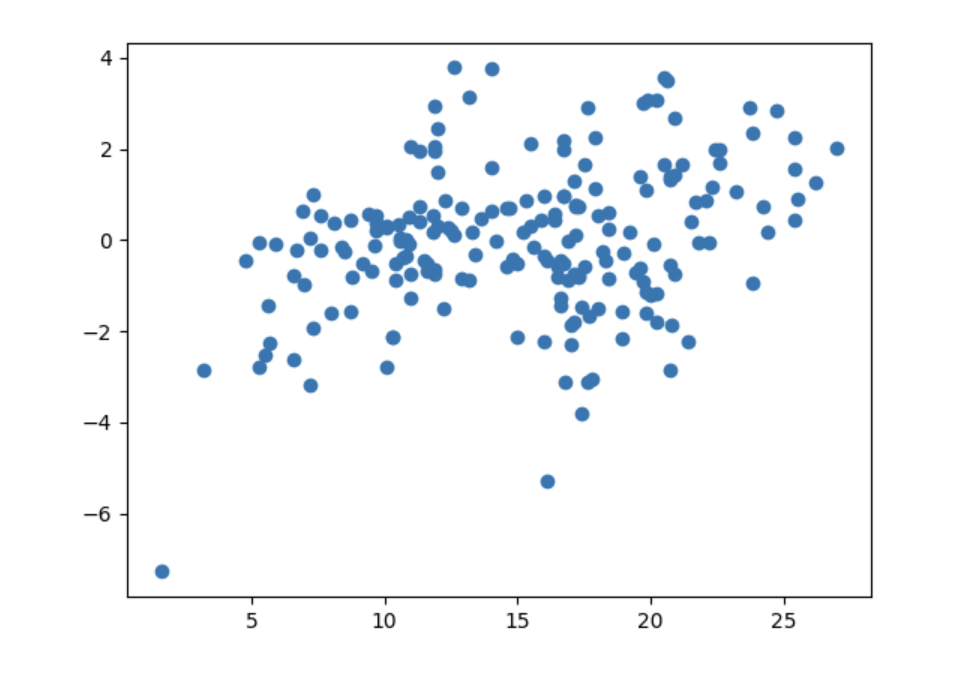
模型F检验P值为[1.8641e-98]，P值<0.05表明模型通过F检验，自变量（至少一个）和因变量间的线性关系是显著的，模型效果较好（2分）

['TV', 'Radio']这2个自变量的t检验对应的P值是显著的，对因变量有显著影响；而['Newspaper']这1个自变量的t检验对应的P值是不显著的，对因变量没有显著影响，可以考虑剔除该变量，由于想要探究各渠道广告投放对销售额的影响，暂不处理（2分）

模型DW检验值为[2.2455]，在1.7-2.3之间，说明自变量和残差之间没有自相关性，模型构建良好（2分）

变量的VIF值均小于10，表明该数据不存在严重的多重共线性问题（2分）





该模型线性回归方程为该模型线性回归方程为Sales = 4.5638 + 0.0545\*TV + 0.1078\*Radio + 0.0009\*Newspaper（2分，合理即得分）

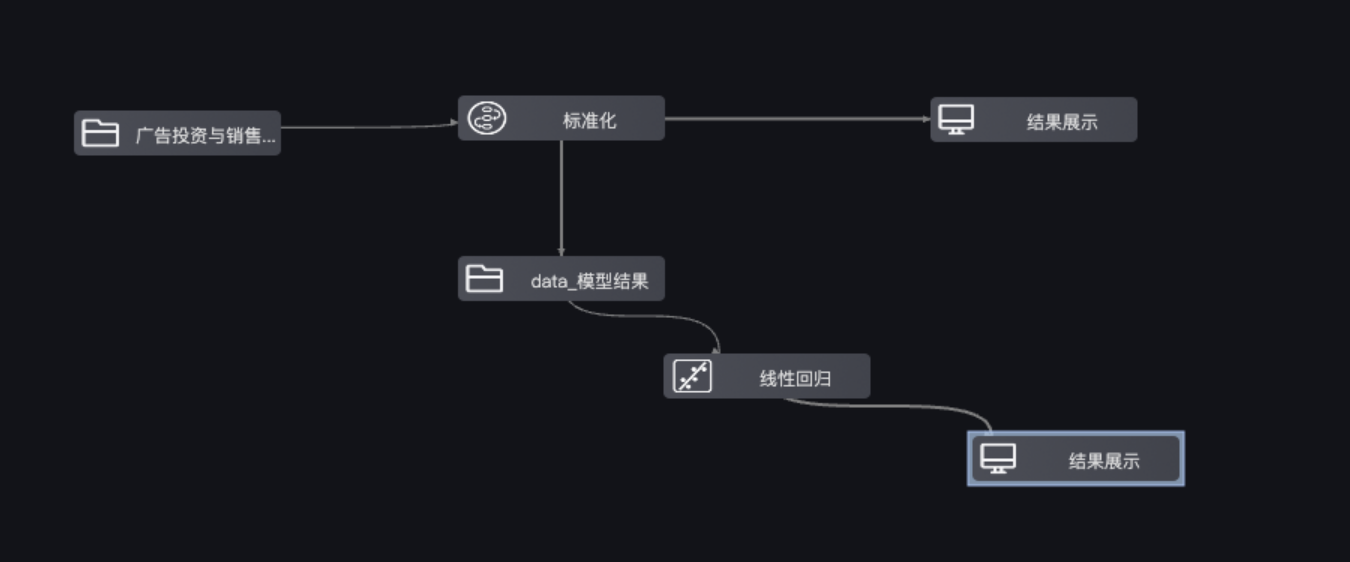
每个系数的解释如下：（3分，合理即得分）

0.0545 是TV的系数，它表示每增加一单位的电视广告投放费用，销售额将平均增加0.0545单位。

0.1082 是Radio的系数，它表示每增加一单位的广播广告投放费用，销售额将平均增加0.1078单位。

0.0009 是Newspaper的系数，它表示每增加一单位的报纸广告投放费用，销售额将平均增加0.0009单位。

（4）对预处理后的自变量进行最大最小规范化，构造线性回归模型（3分）





由线性回归结果可知，对销售额影响最大的是电视广告投放费用，其次是广播广告投放费用，最后是报纸广告投放费用。（3分，合理即得分）

建议该企业加大电视广告投放费用。（2分，合理即得分）