【参考答案】：

1. 描述性统计

缺失值：（3分）

该数据集所有变量都不存在缺失值，故不作处理。



离散型变量的描述性统计：（离散+连续描述性统计3分）



连续型变量的描述性统计：



 

均有一定程度右偏，但考虑样本的真实性，暂不处理

预测变量是否犯罪的样本比例约为1:4（192：:887），不太均衡，暂不处理，可以先查看结果再考虑要不要处理（均衡性3分）

异常值：（3分）

由箱线图可以看出，存在较多离群点，酌情考虑是否需要作异常值处理，这里考虑可能是不同类别人员导致的，暂不处理



标签编码：（3分）

将文本型字段进行标签编码处理，其中文化程度进行独热编码也给分，不独热编码也可以。

原假设：ED（文化程度）和是否犯罪（DEP）没有显著关联；

备择假设：ED（文化程度）和是否犯罪（DEP）有显著关联。

卡方检验分析结果：



P值为1.1341484727372575e-104，小于0.05，拒绝原假设。即ED（文化程度）和是否犯罪（DEP）有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）

原假设：HT0t6（近三个月入住时间在0点之后6点之前的次数）与是否犯罪（DEP）没有显著关联;

备择假设：HT0t6（近三个月入住时间在0点之后6点之前的次数）与是否犯罪（DEP）有显著关联。

单因素方差分析结果：



P值为8.466068202804786e-46，小于0.05，拒绝原假设。即HT0t6（近三个月入住时间在0点之后6点之前的次数）与是否犯罪（DEP）有显著关联。

（原假设备择假设合理4分，p值2分，结果分析合理并且结论正确2分）

1. 以是否犯罪作为因变量，其他变量作为自变量，构造分类模型。

逻辑回归结果：（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：

 

测试集：



逻辑回归结果既没有欠拟合也没有过拟合

决策树算法：深度为5，其他默认（选择一种分类算法2分+模型结果展示2分+模型结果分析4分）

训练集：







决策树结果既没有欠拟合也没有过拟合

综合来看，逻辑回归和决策树的结果都比较令人满意，针对1类犯罪的人员来说，决策树的预测效果更胜一筹，此外，两种方法在解释力上各有优势，逻辑回归通过回归系数进行解释，而决策树可以通过树的形状来解释分类规则。（比较2类不同模型的分析结论4分）

(4)

解释人员是否犯罪的影响因素：（4分，言之有理即可）

近三个月入住时间在0点之后6点之前的次数越多越倾向于犯罪

文化程度越高，越倾向于不犯罪

近三个月住宿次数越多，越倾向于不犯罪

近三个月上下网次数越多，越倾向于犯罪

建议：（3分，言之有理即可）

高度关注哪些近三个月入住时间在0点之后6点之前的次数较多、文化程度低、近三个月住宿次数较少，但是近三个月上下网次数较多的人员，或者针对预测模型对人员进行排查，多关注有犯罪倾向的人员。