H1B签证是国际学生完成大学和足够工作经验后所申请和持有的常见的签证类型。每年USCIS都收到了近两倍多的H1B签证申请，他们用一个数字标记每个申请表，并利用计算机来运行随机算法，所有的申请表均会被处理，因此H1B签证申请的成功与否很大程度上依赖于随机算法，我们无法根据数据做出分析。然而，劳工条件申请（LCA）处理过程成功认证是申请H1B签证的重要前提，并且劳工条件申请依赖于雇主提供的信息，**分析劳工条件申请过程的影响因素**有助于提升雇主的外籍劳工聘用率，进而增加有资格申请H1B签证的人数。

数据集h1b\_data.csv为美国外籍劳工认证数据中心披露的2017年前三季度关于LCA过程的审批数据，约50万条观测。数据集中每条观测为LCA过程的审批信息。该数据集包含35个变量。前15个变量为工作地点、申请状态、申请提交日期、申请审批时间、雇主名称、雇佣关系开始时间、雇佣关系结束时间、公司总人数、是否全职、现行工资、支付工资下限、支付工资上限、职业名称、公司所在经度和纬度。第16至21个变量以及第32、34个变量均以dcs开头，表示与申请审批时间有关的变量，分别为：星期、每月的第几周、年、月、哪年的第几天、哪年的哪月、是否为周末以及季度。第22至27个变量以及第33、35个变量均以sbm开头，与dcs相对应。第28至30个变量分别为受雇时长、审批时长以及薪资的分类变量。数据说明表如下：



**分析任务：**

1. 读入数据集，命名为h1b。（注：由于该数据集较大，可灵活采用针对大数据集的操作方式，例如，在R中可将数据转换为data.table格式）

**【描述性分析】**

2. 通过在美国地图上绘制各个工作地点的申请人数（单位：千人）热力图，从而通过整体视角了解H1B签证LCA过程在全美的分布情况。（提示：在R中可使用map\_data提取出美国的地图数据，根据WORKSITE、lon和lat这三个变量，汇总出各工作地点的申请人数（单位：千人），并通过ggplot绘制热力图）

3. 除了从地理热力图直观的感受数据，请将申请审批时间对应到日历上，来直观感受每一天的审批人数情况。（提示：请根据dcs\_year、dcs\_yearmonth、dcs\_monthweek、dcs\_weekday和dcs\_dayyear这五个变量分别汇总每天的申请人数（单位：千人），存入到dcs\_heat；在R中，根据dcs\_heat，利用ggplot函数绘每日申请人数的日历热力图）

4. 绘制画雇佣关系持续时间与申请状态的关系图。（提示：根据duration\_class和CASE\_STATUS两个变量汇总出各类受雇时长的申请状态，计算处不同受雇时长类别中申请状态的百分比，从而绘制对应的统计图）

5. 绘制各类数据岗位在美国的最低工资与申请状态关系图，其中，提取SOC\_NAME变量中MANAGEMENT ANALYSTS、OPERATIONS RESEARCH ANALYSTS、FINANCIAL ANALYSTS、MARKET RESEARCH ANALYSTS、STATISTICIANS和COMPUTER PROGRAMMERS这六类职业，将其英文名改为相应的中文名，分别为：管理分析师、运行分析师、金融分析师、市场分析师、统计学家和程序员。（提示：利用ggplot函数绘制出各职业不同申请状态下的最低工资（WAGE\_RATE\_OF\_PAY\_FROM）的箱线图）

**【建模分析】**

6. 以对申请状态为因变量，以TOTAL\_WORKERS、duration\_class、certi\_duration\_class、wage\_class、occupation、dcs\_weekend、sbm\_weekend、dcs\_quarter、sbm\_quarter和FULL\_TIME\_POSITION为自变量，建立逻辑回归，探求这些变量对于申请状态的影响，并利用summary函数将回归结果存入到mod.summary。（提示：在R中，利用glm函数建立模型）

7. 根据上述回归模型，计算混淆矩阵，并绘制ROC曲线，标注AUC值。（提示：在R中，使用pROC包中的roc函数绘制曲线）