第二次数据分析作业要求

这次作业的主要要求是根据Auto数据集建立一个逻辑回归模型来预测对于给定的一台车，其到底是耗油量高还是耗油量低。为此，我们需要完成以下任务：

1. 对数据集中衡量耗油量的mpg变量进行二值化处理，具体规则是若mpg取值大于其中位数则令其为1，否则为0。可将二值化后的新变量命名为mgp01，显然其为因变量即逻辑回归模型中的Y。

2. 选择合适的数据描述分析工具(如散点图与箱线图)来研究其他特征变量(不包括mpg)与mgp01间的关系，并回答问题：哪些变量可能会对预测mgp01有帮助？请详细陈述你的发现。

3. 将数据集随机分为训练集与测试集，两者比例分别为60%与40%，这里需合理使用R语言自带的sample()函数。

4. 根据第2步的分析确定逻辑回归模型中的因变量X，并用glm()函数在第3步得到的训练集上拟合逻辑回归模型。然后，将得到的模型用于预测测试集上数据的二值因变量值，并计算误差率。举例说明：若测试集上有100个样本，然后针对这些样本的自变量X预测得到的因变量值有10个与其本来的Y值不一致，则误差率10/100=10%，显然误差率越小表明模型预测效果越好。

5. 为避免单次计算的随机性，重复3-4步20次，从而会得到20个误差率，计算所有这20个误差率的平均值，此平均值可作为评估模型优劣的标准。

完成上述任务以后，请写一个报告详细陈述整个数据分析流程以及你们得到的数据分析结论。

Auto数据中的变量名说明：

mpg: miles per gallon

cylinders: Number of cylinders between 4 and 8

displacement: Engine displacement (cu. inches)

horsepower: Engine horsepower

weight: Vehicle weight (lbs.)

acceleration: Time to accelerate from 0 to 60 mph (sec.)

year: Model year (modulo 100)

origin: Origin of car (1. American, 2. European, 3. Japanese)

name: Vehicle name