数据分析报告

课程名称：程序设计基础（R）

|  |  |
| --- | --- |
| 开课学期： 2023 至 2024 学年 第一 学期  开课班级： 2021级大数据管理与应用1班 .  题目： 汽车mpg数据分析报告 .  学号： 202106140344 姓名： 邹元 .  完成时间： 2023 年 12 月 25 日 |  |

目 录

[一、背景与目标 3](#_Toc154439349)

[二、 数据说明 3](#_Toc154439350)

[（一） 数据来源 3](#_Toc154439351)

[（二） 变量说明 3](#_Toc154439353)

[1. 对变量进行说明 3](#_Toc154439354)

[2. 显示部分数据 4](#_Toc154439355)

[三、 数据探索 4](#_Toc154439356)

[（一） 读取数据 4](#_Toc154439357)

[（二） 查看数据整体信息 4](#_Toc154439358)

[（三） 缺失值探索 6](#_Toc154439359)

[1. 缺失值情况分析 6](#_Toc154439361)

[2.缺失值的处理以及数据类型的转换 6](#_Toc154439362)

[四、 数据可视化 7](#_Toc154439363)

[（一） 绘制条形图 7](#_Toc154439364)

[（二） 绘制点图 7](#_Toc154439365)

[1. 对气缸数和mpg绘制点图 8](#_Toc154439366)

[五、 相关性性分析 9](#_Toc154439367)

[六、 异常值处理 9](#_Toc154439368)

[七、 描述性分析](#_Toc154439369)

[八、 数据集构造 11](#_Toc154439370)

[（一）、重要变量选择 11](#_Toc154439371)

[九、 基于随机森林模型的汽车mpg预测模型构建及应用 11](#_Toc154439372)

[（一） 建立模型 11](#_Toc154439373)

[（二） 数据预测（这部分主要是呈现预测结果） 13](#_Toc154439374)

[十、 总结与建议 13](#_Toc154439375)

汽车油耗估算分析

1. 背景与目标

汽车油耗直接关系到车辆的运行成本，油耗越高，车辆的运行成本就越大。通过对汽车油耗的分析，可以找出油耗过高的原因，从而采取有效措施降低车辆的运行成本；汽车是城市空气污染的主要来源之一，而油耗过高会导致更多的二氧化碳和其他有害气体排放到大气中，加剧空气污染和气候变化问题。通过对汽车油耗的分析，可以降低汽车的碳排放量，减轻对环境的负面影响；汽车油耗与车辆性能密切相关。通过对汽车油耗的分析，可以发现车辆性能的不足之处，进而进行针对性的优化和改进，提高车辆的整体性能；汽车油耗过高可能是车辆出现故障的前兆。通过对汽车油耗的分析，可以及时发现潜在的故障问题，提前进行维修和保养，避免因车辆故障导致的意外事故和损失；随着能源供应日益紧张，石油等传统能源的价格不断上涨。通过对汽车油耗的分析，可以更好地管理和利用能源资源，保障国家能源安全。

汽车油耗分析预测的意义重大，不仅可以降低车辆运行成本、提高环保意识、优化车辆性能、预防故障发生，还可以保障国家能源安全。因此，车主应该重视汽车油耗分析预测，采取合理措施降低油耗，为可持续发展和绿色出行做出贡献。

1. 数据说明
2. 数据来源

该数据集来自卡内基梅隆大学维护的StatLib库。该数据集在1983年美国统计协会博览会上使用。

1. 变量说明
2. 对变量进行说明

表2-1对变量进行说明

|  |  |  |  |
| --- | --- | --- | --- |
| 字段名 | 数据类型 | 字段描述 | 取值说明 |
| mpg | 数值型 | 油耗，既汽车每消耗一加仑燃料能行驶的英里数 |  |
| cylinders | 数值型 | 汽车的发动机气缸数量 | 4  6  8 |
| displacement | 数值型 | 发动机的排量 |  |
| horsepower | 数值型 | 英制马力 |  |
| weight | 数值型 | 汽车的重量(kg) |  |
| acceleration | 数值型 | 一般指汽车静止加速100 km/h 或60 mph 所需要的时间(s) |  |
| Model year | 数值型 | 车型的发布时间 |  |
| carname | 字符型 | 车辆的名称 |  |

1. 显示部分数据

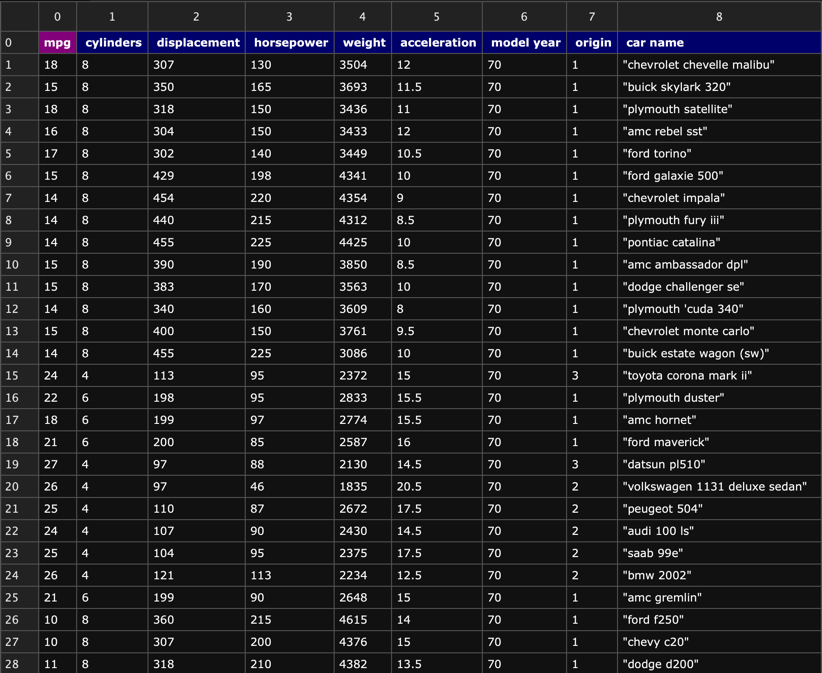


图2-1 数据集部分数据展示

1. 数据探索
2. 读取数据

# 设置工作目录

setwd('/Users/zouyuan/RStudyZouYuan')

#导入数据文件

data <- read.csv('./汽车油耗mpg估算/auto.mpg.csv',stringsAsFactors = F)

1. 查看数据整体信息

View(Train) #查看训练集的整体信息

View(Test) #查看测试集的整体信息

1）样本量不大(400以内)，自变量较少（8个），待预测变量mpg。

2）自变量内容包括气缸数、排量、英制马力、重量、加速能力、设计日期、名称等非常详细的指标。

3）在horsepower变量中存在少量的缺失数据。

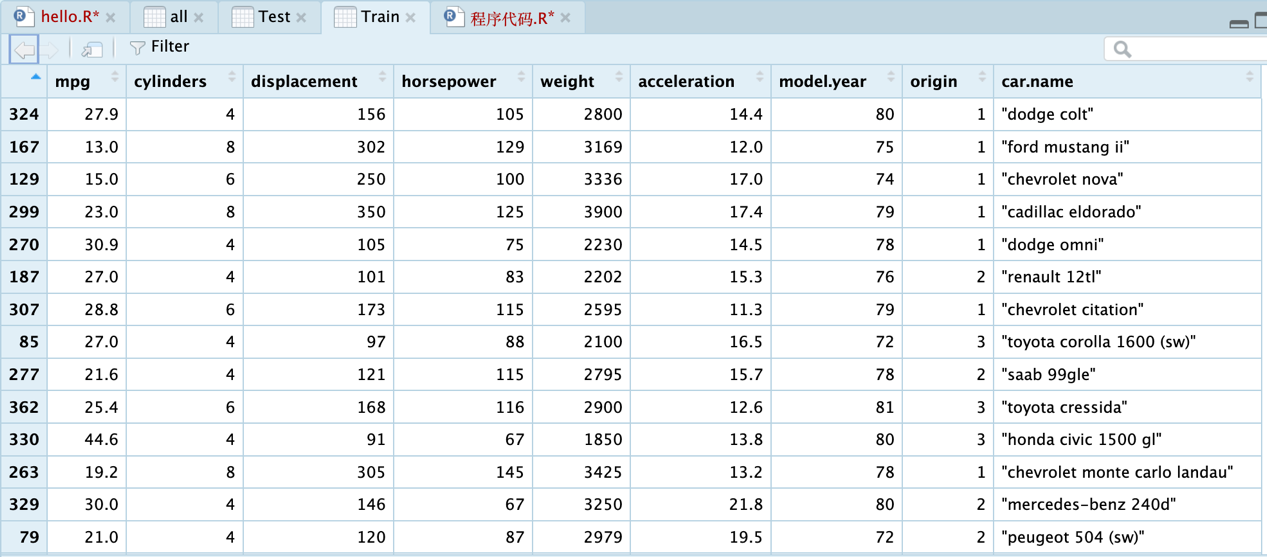


图3-1 查看训练集数据整体信息

表格

描述已自动生成

图3-2 查看测试集数据整体信息

# 合并训练集和测试集

all <- rbind(Train,Test) # 合并训练集和测试集

head(all) # 查看数据整体信息

表格

描述已自动生成

图3-3 查看数据集整体信息

1. 缺失值探索

数据分析之前，先对缺失的数据进行统计，分析有缺失值出现的特征对整个分析过程及结果是否有重要的影响。

1. 缺失值情况分析

*文本

描述已自动生成anyNA(all) #* *检测数据集是否存在缺失值*

图3-4 检测数据集是否存在缺失值

可知数据中存在缺失值。

*which(is.na(all),arr.ind = T) # 查看缺失值位置*

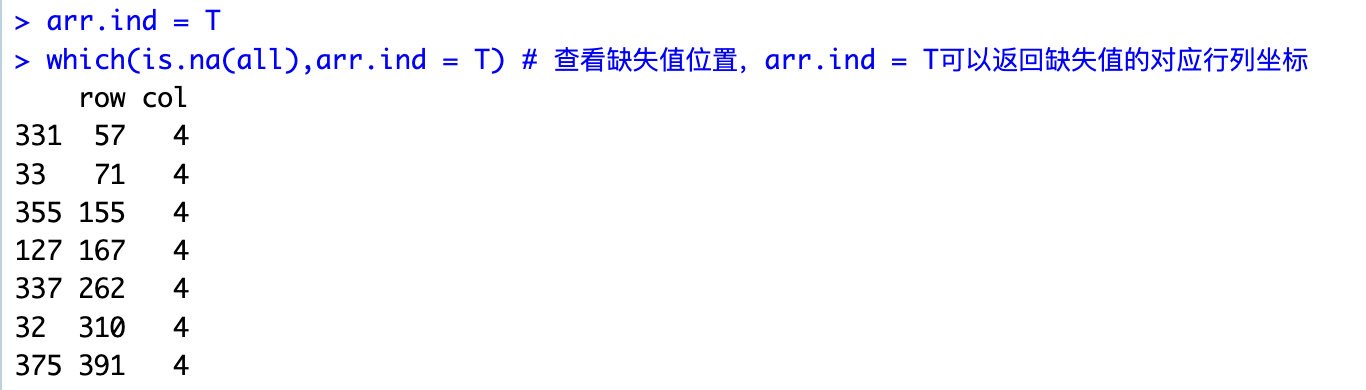


图3-5 查看缺失值的位置

从上图发现：数据集中存在的缺失中较少，通过观察变量horsepower的分布，得到完全随机的缺失值应该很少，可直接用众数进行填补。

1. .缺失值的处理以及数据类型的转换

*# 求取众数*

*tmp <- table(all$horsepower) #计算出horsepower列中每个值出现的次数*

*index <- which.max(tmp) # 找出最多次数的索引*

*tmp[index] # 输出对应的数据及次数*

*all$horsepower[is.na(all$horsepower)] <- all[150,4] # 用众数填补缺失值*

*anyNA(all) # 检测数据集是否存在缺失值*

*all$horsepower <- as.integer(all$horsepower)#转换数据类型*图形用户界面, 文本, 应用程序

描述已自动生成

图3-6 用众数填补缺失值

1. 数据可视化
2. 绘制条形图

*# 绘制条形图*

*counts <- table(all$cylinders) # 将mtcars的挡数制成一个频数统计表*

*counts # 查看counts的信息*

*barplot(counts, main="Car Cylinders", xlab="Number of Cylinders")*

*图表, 条形图

描述已自动生成*

图4-1 汽车气缸数统计条形图

从图中可以看出4缸车在数据集中占最大比例，其次是8缸车与6缸车，在后续的可视化、mpg预测以及总结结论的时候4、8、6缸车应该占有更高的衡量权重。

1. 绘制点图

由于数据集包含的数据较多，我们运用sample函数采用随机采样20条的方法对数据集进行处理。

*#随机选取20条数据绘制点图*

*all\_sample <- all[sample(nrow(all),20), ]*

*对排量和mpg绘制点图*

*# 提取mpg和displacement列*

*mpg\_displacement <- all\_sample[, c("mpg", "displacement")]*

*# 按照displacement进行排序*

*mpg\_displacement <- mpg\_displacement[order(mpg\_displacement$displacement), ]*

*# 绘制点图*

*dotchart(mpg\_displacement$mpg, labels = mpg\_displacement$displacement, cex = 0.7, main = "Gas Mileage for Car displacement", xlab = "Miles Per Gallon")*

*图表, 散点图

描述已自动生成*图4-2 汽车排量与油耗关系点图

从图中可以看出，汽车的油耗与排量大体呈负线性相关的关系。排量越大，汽车每行驶一英里所消耗的燃料越少；排量越小，汽车每行驶一英里所消耗的燃料越大。

1. 对气缸数和mpg绘制点图

*# 提取mpg和cylinders列*

*mpg\_cylinders <- all\_sample[, c("mpg", "cylinders")]*

*# 按照displacement进行排序*

*mpg\_cylinders <- mpg\_cylinders[order(mpg\_cylinders$cylinders), ]*

*# 绘制点图*

*dotchart(mpg\_cylinders$mpg, labels = mpg\_cylinders$cylinders,cex = 0.7, main = "Gas Mileage for Car cylinders", xlab = "Miles Per Gallon")*

*图表, 散点图

描述已自动生成*

图4-2 汽车气缸数与油耗关系点图

从图中可以看出，汽车的油耗与气缸数大体呈负线性相关的关系。气缸数越多，汽车每行驶一英里所消耗的燃料越少；气缸数越少，汽车每行驶一英里所消耗的燃料越多。

1. 相关性性分析

*# 相关性分析*

*states <- all[,1:5]*

*cor(states) #计算Pearson积差相关系数*

*cor(states,method = "spearman") #计算Spearman等级相关系数*

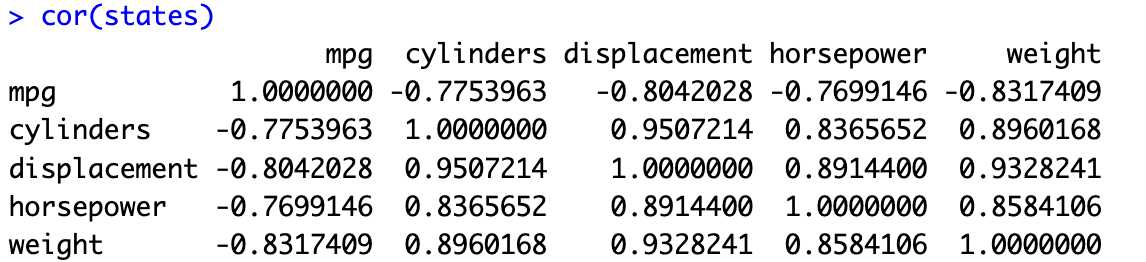
**

图5-1 计算Pearson积差相关系数结果图

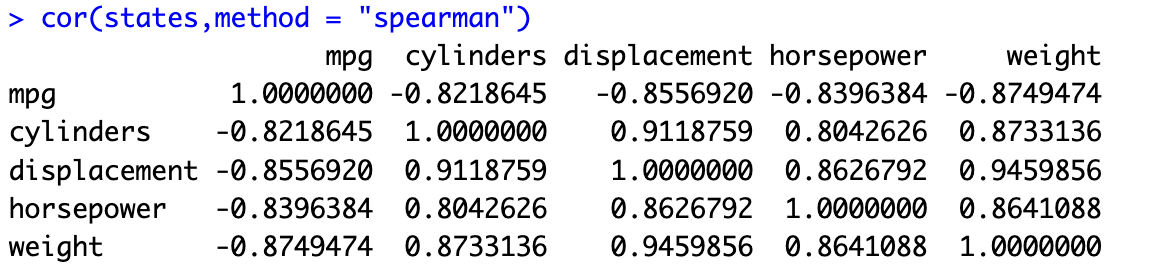
**

图5-2 计算Spearman等级相关系数结果图

从图中可以看出，油耗mpg与汽车的气缸数、排量、马力和重量都基本呈现负相关，且负相关的程度较大，说明选择的以上变量对于油耗mpg的影响较为显著。

1. 异常值处理

从数据可视化的密度图中可以看到，原始的数据并不是标准的正态分布，可能是受到了异常值的影响，我们可以通过绘制点图的方式来可视化数据，查看是否有必要要对数据进行异常值删除。

*#异常值探索*

*par(mfrow = c(2,2))*

*plot(all[1:398,2],Train\_mpg,xlab='cylinders')*

*plot(all[1:398,3],Train\_mpg,xlab='displacement')*

*plot(all[1:398,4],Train\_mpg,xlab='horsepower')*

*plot(all[1:398,5],Train\_mpg,xlab='weight')*

*散点图

低可信度描述已自动生成*

图6-1 异常值分析点图

从图中可以看出，数据集中存在异常值，但是数量较少，对于数据集整体特征的影响较小，不做删除处理。

1. 描述性分析 Mpg是预测变量，我们可以作图看一下它的分布情况。 *#Mpg ggplot(data.frame(Train\_mpg),aes(Train\_mpg,color=I("white"),fill=I("darkred")))+geom\_histogram(bins=20)+theme(panel.background = element\_blank(),axis.line = element\_line())*

图表, 直方图

描述已自动生成

图7-1 mpg分布点图

从图上看，数据分布整体呈现右偏，这符合我们对油耗的常规认知，对大多数汽车，经济省油才是最主要的目标。

1. 数据集构造

（一）、重要变量选择

*#划分训练集和测试集*

*set.seed(1)*

*train <- sample(1:nrow(all),nrow(all)\*3/4) #取75%做训练集*

*Train <- all[train,]*

*Test <- all[-train,]*

*Test = Test[1:6] #构造测试集*

*Train = Train[1:6] #构造训练集*

*Train\_mpg <- Train$mpg # 把mpg从train数据集中拆分出来*

1. 基于随机森林模型的汽车mpg预测模型构建及应用
2. 建立模型

为了测试出多少课数时模型是最优的，我们先试探性地用800颗树试一下，然后画出错误率与决策树的数量的关系曲线。

*#基于随机森林模型的mpg预测模型构建*

*set.seed(1234)*

*# 验证性训练模型*

*rf\_ntree<-randomForest(x=all,y=all\_mpg,ntree=800,importance=TRUE,proximity=TRUE)*

*plot(rf\_ntree)*

*图表

描述已自动生成*

图9-1 随机森林模型训练错误率图

我们可以看出，在树的数量大于500的时候就基本稳定了，而且之后虽然还会有小幅度的下降，但error水平差不多，因此我们在保证效能的情况下，减少决策树的数量，减少运行时间，取ntree=500。

*# 正式训练模型*

*rf\_ntree<-randomForest(x=Train,y=Train\_mpg,ntree=500,importance=TRUE,proximity=TRUE)*

*rf\_ntree*

*文本, 信件

描述已自动生成*

图8-7 随机森林模型评估

模型在训练集上的Mean of squared residuals（残差平方均值）=1.97914 ；explained: 96.77%代表可以训练集中解释约96.77%的样本

1. 数据预测

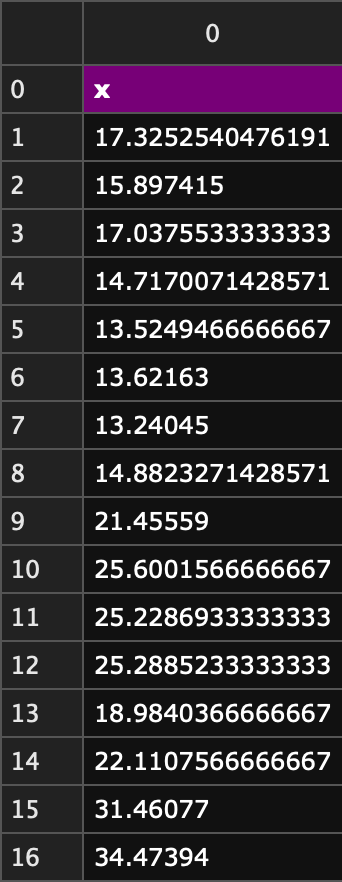


图9-3 随机森林预测mpg结果图

1. 总结与建议

通过对汽车mpg数据的一系列预处理、数据挖掘、建模，我们可以得出对汽车油耗影响最大的几个特征分别是Cylindrs、Displacement、Weight、Horsepower、Acceleration等，该项目能一定程度地为购车者提供有依据的购车建议。