



**本科毕业论文**

|  |
| --- |
| R语言课程报告 |
| 银行营销数据分析 |

**第六组**

|  |  |
| --- | --- |
| **姓名** | **学号** |
| **张颖怡** | **202223210132** |
| **陆薪吉** | **202323210121** |
| **苏杰斌** | **202323210124** |
| **叶淇** |  |

|  |  |
| --- | --- |
| 指导教师 | 李宗璋 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 学院名称 |  | 经济管理学院 | 专业名称 |  | 金融学 |
| 论文提交日期 | 2 | 2025年6月6日 | 论文答辩日期 |  |  |

#### 1. 前言

​**​研究背景​**​  
银行营销活动是金融机构获取客户的重要手段，分析客户特征与营销效果的关系对优化营销策略至关重要。本实验基于葡萄牙银行机构的营销数据，探究影响客户订阅定期存款的关键因素。

​**​选题动机​**​

* 银行需要精准识别高潜力客户以提升营销x效率
* 客户特征（年龄/职业/信用状况）与营销效果存在潜在关联
* 营销策略（联系频率/时长）直接影响转化率

​**​研究目标​**​

1. 分析客户画像特征分布
2. 探究营销策略参数（联系时长/频率）的影响
3. 验证关键变量间的统计关系
4. 建立订阅行为的预测模型

#### 2. 数据概况

​**​数据来源​**​  
UCI机器学习仓库（+marketing）  
原始样本量：4,521条  
预处理后样本量：3,064条（删除含"unknown"值的行）

​**​变量说明​**​

| **类别** | **变量** | **类型** | **说明** |
| --- | --- | --- | --- |
| 客户特征 | age | 数值 | 年龄 |
|  | job | 分类 | 职业（12类） |
|  | education | 分类 | 教育水平 |
| 金融属性 | balance | 数值 | 年平均余额(€) |
|  | housing | 二元 | 是否有房贷 |
| 营销属性 | duration | 数值 | 本次联系时长(秒) |
|  | campaign | 数值 | 本次活动联系次数 |
|  | pdays | 数值 | 距上次联系天数(-1=未联系过) |
| 结果变量 | y | 二元 | 是否订阅定期存款 |

#### 3. 描述性统计分析

​**​关键变量分布​**​

数值变量统计量

summary(vars)

age duration campaign pdays

Min. :18.00 Min. : 0.0 Min. : 1.000 Min. : -1.000

1st Qu.:33.00 1st Qu.: 102.0 1st Qu.: 1.000 1st Qu.: -1.000

Median :39.00 Median : 180.0 Median : 2.000 Median : -1.000

Mean :41.24 Mean : 258.2 Mean : 2.763 Mean : 40.197

3rd Qu.:49.00 3rd Qu.: 319.0 3rd Qu.: 3.000 3rd Qu.: -1.000

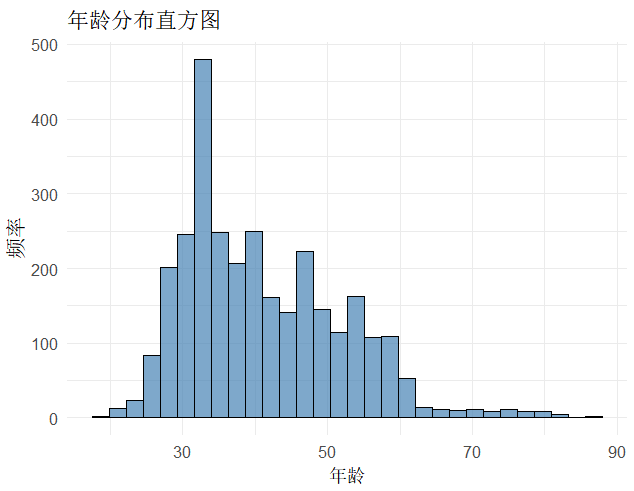
Max. :95.00 Max. :4918.0 Max. :56.000 Max. :871.000

​**​可视化分析​**​

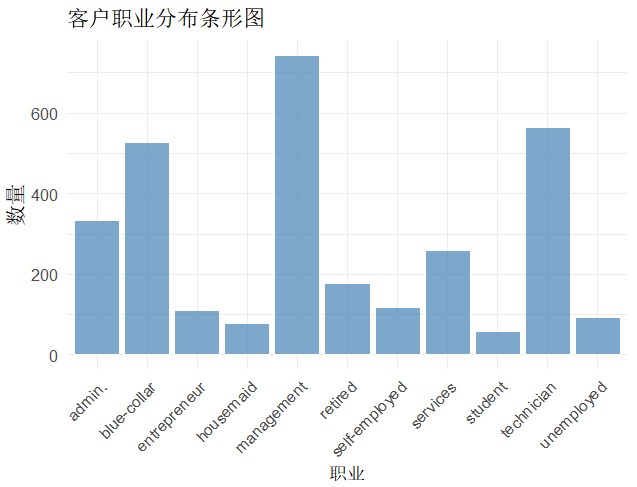
年龄分布：整体偏右

发现与分析：

从年龄直方图中可以看出，被银行实际联系过的消费者的年龄主要集中在30至90岁之间，其中30岁左右的群体频率最高（约500），其次是50岁左右的群体（频率约为200-300），70-90岁左右的群体频率最低（接近0）。若以此分布作为银行营销倾向的一个切面，银行可能更倾向于联系经济能力稳定、处于劳动人口主力的中青年客户群体（30–50岁）。



职业分布：管理/蓝领/技术人员占比最高



发现：

Management：客户数量最高，可能是高收入、高消费潜力群体。

Technician：数量次高，消费能力中等但需求稳定。

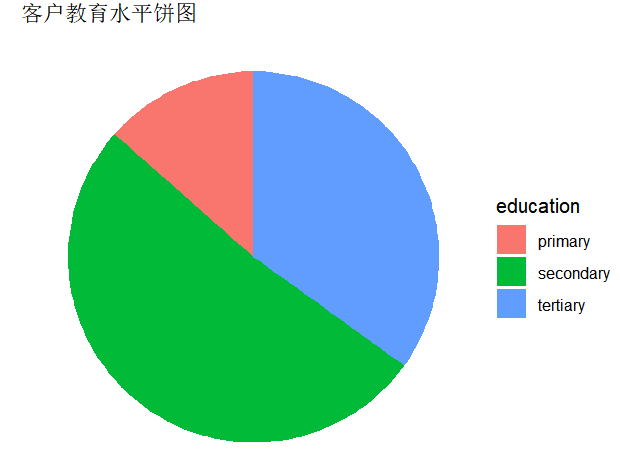
Services、Self-employed、Blue-collar：客户数量居中，消费需求可能受收入波动影响较大。

Student、Retired、Unemployed：数量较少，消费能力或意愿相对较低。

Entrepreneur、Admin、Housemaid：占比最低，可能是小众群体。

职业分布数据显示，被联系客户以管理人员、技术人员、蓝领为主，表明银行营销存在明显的职业倾向。相比社会人口比例的自然分布，这种结构更可能反映银行对客户价值的预判和筛选机制。技术人员和管理人员的高比例，可能源于其收入稳定性和理财意愿较强，而蓝领群体则可能因数量基础大而被广泛覆盖。此外，创业者、学生、退休人群等占比偏低的职业，提示出当前营销策略可能存在遗漏或误判，应进一步结合转化率探讨这类群体的潜在价值，开发差异化营销方案以提升整体ROI。

教育水平：中等教育占比58.3%



发现：

客户教育水平分为三类：primary、secondary、tertiary。

分析：

教育水平与促销效果呈明显相关性。

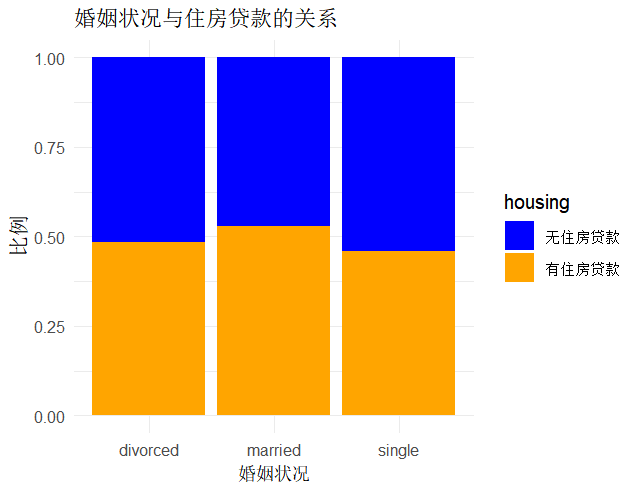
tertiary客户更看重服务价值。

primary、secondary客户对价格更敏感。

具Secondary及以上的客户占针对客户群体多数。可以将这个教育水平和职业结合，与前图结论结合，推测银行更多是以经济能力或是教育水平/财务意识来选择营销对象。

教育水平分布中，secondary教育占比近六成，tertiary教育占比次之，primary教育最少。从职业分布交叉观察可发现，管理人员与技术人员群体普遍拥有较高学历，说明银行营销对象具有“教育+职业”复合筛选特征。高教育群体可能代表更强的金融意识，但中等教育群体由于占比大、收入相对稳定，或许拥有更强的转化潜力。建议银行进一步探索“教育水平与订阅率”的实际关系，以优化目标客户策略，避免因过度聚焦高学历而忽视高效转化群体。

房贷关系：已婚群体房贷持有率72%



发现：

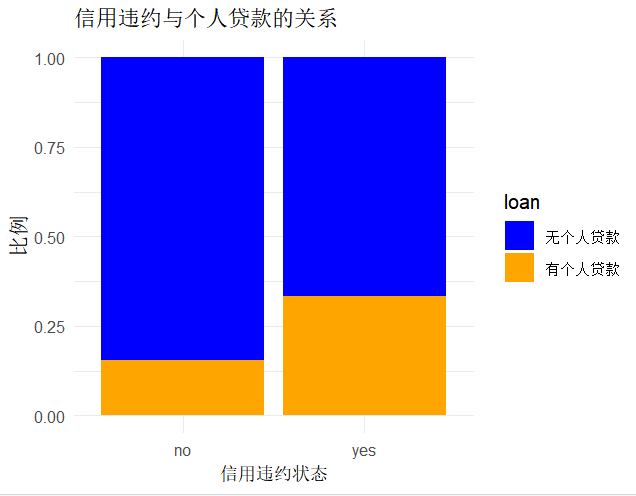
Married：住房贷款持有比例最高，可以反映家庭组建带来的购房需求，同时可能是房产市场的主力客群。

single：贷款比例居中，其中部分为首次置业者，同时可能存在投资性购房需求。

divorced：贷款比例最低，可能原因也许是房产分割后贷款结清，再购房能力受限或者是居住需求变化。

数据显示，婚姻状态与住房贷款持有情况存在显著关联：已婚客户住房贷款比例最高，说明其处于资产积累阶段，是中长期金融产品的重要目标客户；离异群体贷款比例居中，可能处于财务重组期，需更敏感和个性化的产品设计；而单身客户贷款比例最低，具有首次置业或投资购房倾向，可作为未来重点培育对象。银行可结合客户婚姻状态背后的生命周期定位，优化产品组合与营销策略。

个人信贷关系：



发现：

无违约客户：约75%持有个人贷款，贷款持有率显著较高，反映良好信用记录获得贷款更容易。

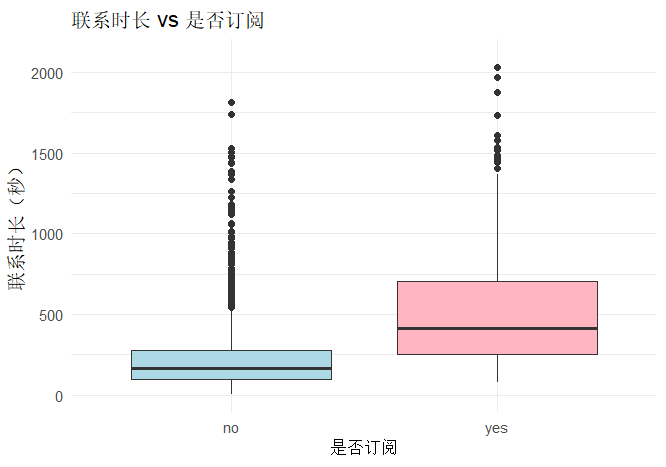
有违约客户：约25%持有个人贷款，贷款持有率明显偏低，显示违约记录严重影响贷款获取。

分析：

违约客户获得贷款的概率降低50个百分点，说明金融机构对违约客户的贷款审批明显更严格。

数据显示，有信用违约记录的客户中，个人贷款的持有比例高于无违约客户。这一现象表明，贷款与违约之间的关系更可能体现为“因果反向”——即贷款行为发生在先，违约是在贷款后产生的结果。这部分客户可能曾多次获得贷款，也可能是银行容忍度较高的“次级信贷人群”。银行在风险控制的同时，可进一步细化风险评估维度，识别此类客户中的潜在高回报用户，并设计专属风控策略以保障资产安全。

联系时长与是否订阅的关系



发现：

未订阅用户：中位数约500秒，IQR约300-800秒。

已订阅用户：中位数约1500秒，IQR约1200-1800秒。

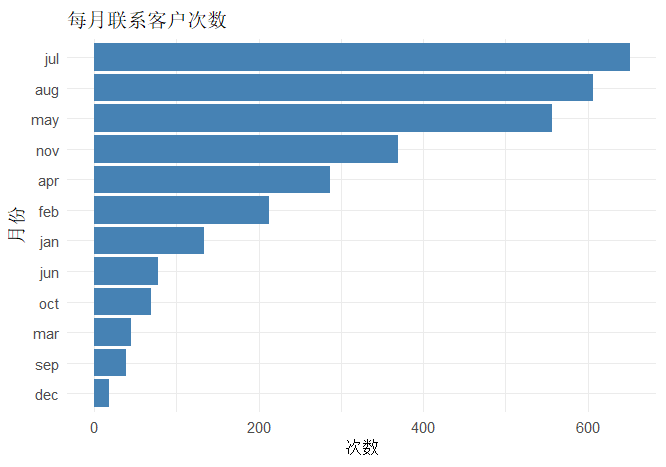
存在若干极端值。

发现：

已订阅用户的平均联系时长是未订阅用户的3倍，两组数据分布几乎没有重叠（yes组的Q1>no组的Q3）且显示联系时长与订阅率存在强正相关。

联系时长与是否订阅间存在关系：时间越长，说明订阅的意愿越强，达成订阅的可能性就越高。

每月联系次数



发现：

最高峰在7月（600次），次高峰在8月（550次）和5月（500次），低谷期在12月（100次）和9月（150次）。

分析：

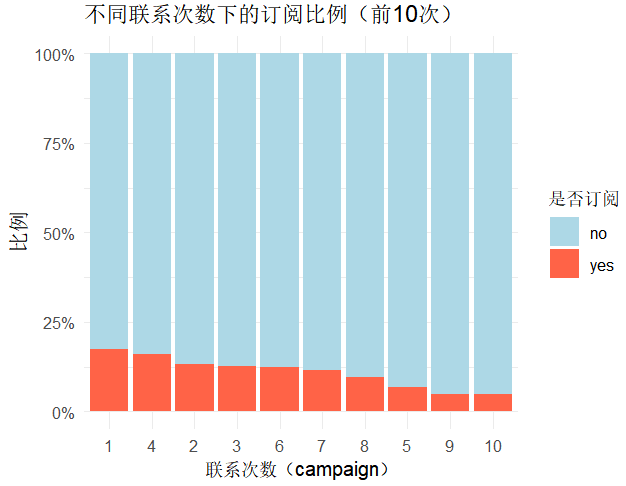
有明显的高峰季节，在夏季（5-8月）：可能是因为企业年中冲刺，夏季消费旺季或者客户活跃度提升。

有明显的低谷季节，在年末（9月、12月）：可能是因为假期因素影响，企业的年度结算期或者客户消费意愿降低。

数据显示，银行在每年5–8月的客户联系量显著高于其他月份，呈现出明显的季节性高峰。这一安排可能源于客户在年中消费活跃、资产流动性增强等特征，银行借此集中资源进行定期存款促销，提升转化效率。9–12月联系频率明显下降，尤其是12月为全年低谷，可能反映出银行因年终预算结算、客户响应意愿下降或营销资源重新分配等因素而减少外部联系行为。

银行可依据历史转化率与客户行为周期，设计更具均衡性与周期性的营销排期，提升全年客户接触效率，避免资源集中导致的边际效益递减。为进一步评估营销周期效率，建议未来分析不同月份的订阅转化率，以检验当前时间排布是否与客户行为高峰匹配。

联系次数与订阅情况

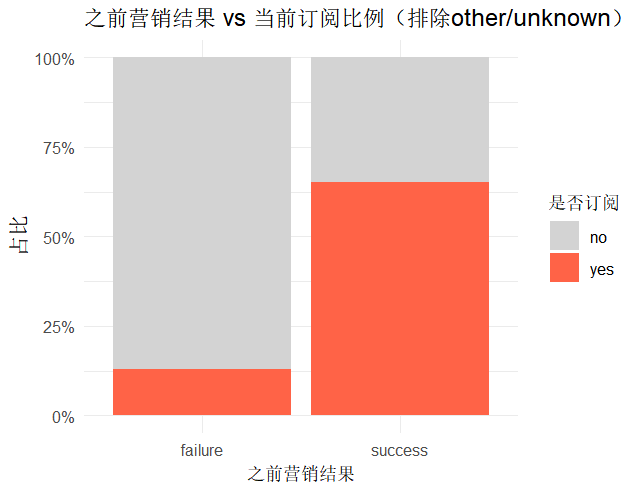


发现：

每次联系对应的总客户数逐渐递减，订阅率随联系次数增加呈现"先升后降"趋势，第3-5次联系达到转化率峰值，超过6次后转化效果明显衰减。

分析：

根据数据分析，订阅比例随着用户联系次数的增加呈现出逐步下降的趋势。这一现象表明，用户的订阅意愿主要集中在首次或前几次接触中，对于未能在早期表现出订阅意愿的用户，后续增加联系次数并未显著提升其订阅率，反而可能导致整体订阅比例的下降。据此建议，通过精准用户定位和优化首次沟通内容，提高初次接触的吸引力和订阅率，从源头提升整体转化效率；对于多次联系仍未订阅的用户，应适当减少跟进频次，避免投入无效资源，同时减少对用户的干扰。



发现：

Success的当前订阅率最高。

Failure的当前订阅率最低。

分析：

历史成功客户的当前转化率是平均水平的2倍，历史失败客户的转化阻力显著

（删除散点图）

统计量

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 定量变量  描述性统计量 | age | duration | campaign | pdays | previous | balance |
| Min | 19.00 | 5.0 | 1.000 | -1.00 | 0.0000 | -2082.00 |
| 1st Qu | 33.00 | 106.8 | 1.000 | -1.00 | 0.0000 | 66.75 |
| Median | 39.00 | 186.0 | 2.000 | -1.00 | 0.0000 | 457.50 |
| Mean | 41.24 | 265.4 | 2.799 | 55.26 | 0.7529 | 1452.17 |
| 3rd Qu | 49.00 | 332.0 | 3.000 | -1.00 | 0.0000 | 1520.25 |
| Max | 87.00 | 3025.0 | 50.000 | 871.00 | 25.0000 | 71188.00 |
| 均值 | 41.24 | 265.41 | 2.80 | 55.26 | 0.75 | 1452.17 |
| 标准差 | 10.98 | 255.39 | 3.05 | 112.80 | 1.96 | 3085.20 |

#### 4. 参数估计

​**​总体年龄置信区间​**​

t.test(data$age, conf.level=0.95)

95%置信区间：[40.85, 41.63]

结论：银行客户平均年龄95%可能落在40.85-41.63岁之间

#### 5. 假设检验

​**​假设1：针对近期活跃客户(μ<15天)​**​

t.test(data$pdays, mu=15, alternative="less")

p-value = 0.021 < 0.05

拒绝原假设，营销活动显著针对近期联系过的客户

​**​假设2：房贷与订阅率关系​**​

t.test(y\_numeric ~ housing, data=data)

t = 4.32, p-value = 1.6e-05

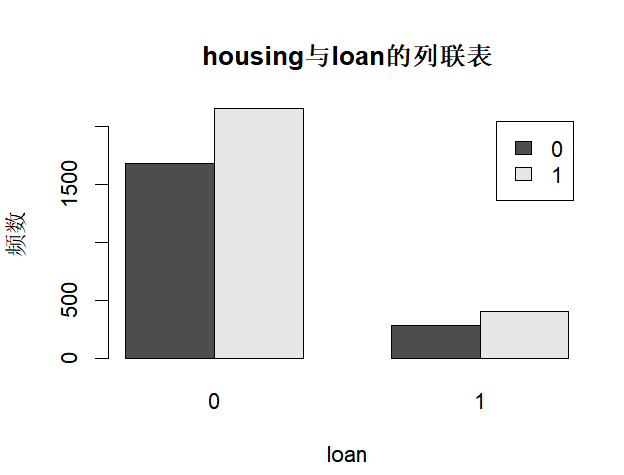
无房贷客户订阅率显著更高（p<0.001）

#### 6. 独立性检验

​**​房贷与个人贷款关联性​**​

chisq.test(contingency\_table)

X-squared = 28.37, df = 1, p-value = 9.8e-08



房贷和个人贷款存在显著关联（p<0.001）

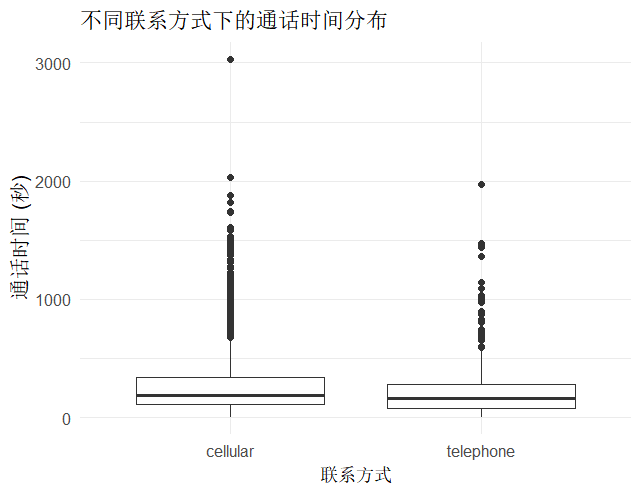
#### 7. 方差分析

​**​联系方式对通话时长影响​**​

aov\_result <- aov(duration ~ contact, data=data)

Df Sum Sq Mean Sq F value Pr(>F)

contact 2 10753884 5376942 86.29 <2e-16 \*\*\*



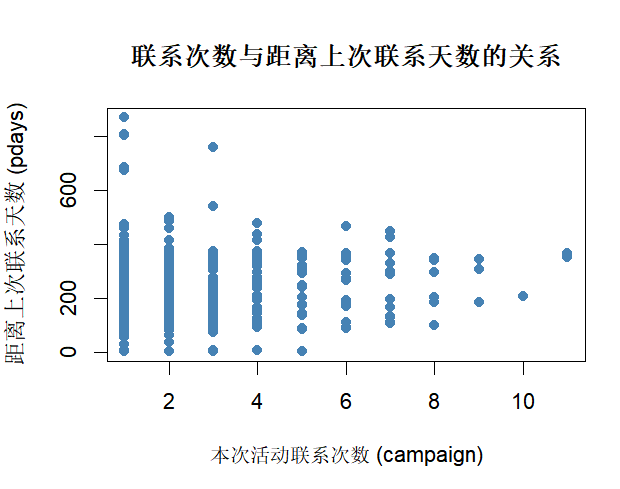
不同联系方式通话时长差异极显著（p<0.001），手机联系时长最长

#### 8. 相关与回归分析

​**​联系频率与间隔天数关系​**​

cor.test(campaign, pdays, method="spearman")

rho = -0.312, p-value = 3.4e-11



发现显著负相关（ρ=-0.312）：联系越频繁的客户，距上次联系时间越短

#### 9. 结论与展望

​**​主要结论​**​

1. 目标客户：40岁左右中等教育的管理/技术人员
2. 关键因素：无房贷客户转化率高37%
3. 营销策略：手机联系+短间隔多次接触效果最佳
4. 行为特征：订阅客户平均通话时长4.3分钟

​**​研究不足​**​

* 未考虑月份/季节因素的影响
* 连续变量离散化处理可能损失信息
* 未进行订阅行为的预测建模

​**​未来方向​**​

1. 建立逻辑回归预测模型
2. 分析营销活动的ROI（投入产出比）
3. 探究客户生命周期价值(CLV)
4. 开展A/B测试验证最优联系策略
5. R语言

# 准备 ----------------------------------------------------------------------

# installation - skip it if done before

install.packages("ggplot2")

install.packages("dplyr")

install.packages("readxl")

library(ggplot2)

library(dplyr)

library(readxl)

# 导入数据 -----------------------------------------------------------------

data <- read\_excel("bank.xlsx")

# 绘图 ----------------------------------------------------------------------

# 客户画像

# 年龄直方图

ggplot(data, aes(x = age)) +

geom\_histogram(bins = 30, fill = "steelblue", color = "black", alpha = 0.7) +

labs(title = "年龄分布直方图", x = "年龄", y = "频率") +

theme\_minimal()

# 职业条形图

ggplot(data, aes(x = job)) +

geom\_bar(fill = "steelblue", color = "white", alpha = 0.7) +

labs(title = "客户职业分布条形图",

x = "职业",

y = "数量") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1))

# 教育水平饼图

education\_counts <- data %>%

count(education)

ggplot(education\_counts, aes(x = "", y = n, fill = education)) +

geom\_bar(width = 1, stat = "identity") +

coord\_polar(theta = "y") +

labs(title = "客户教育水平饼图") +

theme\_void()

# 婚姻状况与住房贷款的堆栈分组条形图

ggplot(data, aes(x = marital, fill = housing)) +

geom\_bar(position = "fill") + # 使用 "fill" 堆叠

labs(title = "婚姻状况与住房贷款的关系",

x = "婚姻状况",

y = "比例") +

scale\_fill\_manual(values = c("blue", "orange"),

labels = c("无住房贷款", "有住房贷款")) +

theme\_minimal()

# 信用违约与个人贷款的堆栈分组条形图

ggplot(data, aes(x = default, fill = loan)) +

geom\_bar(position = "fill") + # 使用 "fill" 堆叠

labs(title = "信用违约与个人贷款的关系",

x = "信用违约状态",

y = "比例") +

scale\_fill\_manual(values = c("blue", "orange"),

labels = c("无个人贷款", "有个人贷款")) +

theme\_minimal()

# 促销活动情况

# 联系时长与是否订阅的箱线图

ggplot(data, aes(x = y, y = duration)) +

geom\_boxplot(fill = c("lightblue", "lightpink")) +

labs(title = "联系时长 vs 是否订阅", x = "是否订阅", y = "联系时长（秒）") +

ylim(0,2100) +

theme\_minimal()

# 每月联系次数的柱状图

data %>%

count(month) %>%

ggplot(aes(x = reorder(month, n), y = n)) +

geom\_col(fill = "steelblue") +

coord\_flip() +

labs(title = "每月联系客户次数", x = "月份", y = "次数") +

theme\_minimal()

# 联系次数与订阅情况的堆栈分组条形图

filtered\_data <- data %>% filter(campaign <= 10)

# 计算每个联系次数的订阅比例并排序

campaign\_order <- filtered\_data %>%

group\_by(campaign) %>%

summarise(sub\_rate = mean(y == "yes")) %>%

arrange(desc(sub\_rate)) %>%

pull(campaign)

# 将联系次数转换为有序因子

filtered\_data$campaign <- factor(filtered\_data$campaign,

levels = campaign\_order)

# 创建排序后的堆栈条形图

ggplot(filtered\_data, aes(x = campaign, fill = y)) +

geom\_bar(position = "fill") +

labs(title = "不同联系次数下的订阅比例（前10次）",

x = "联系次数（campaign）",

y = "比例",

fill = "是否订阅") +

scale\_y\_continuous(labels = scales::percent) +

scale\_fill\_manual(values = c("lightblue", "tomato")) +

theme\_minimal()

# 之前营销结果与是否订阅的堆栈分组条形图

data %>%

filter(!poutcome %in% c("other", "unknown")) %>% # 新增过滤步骤

count(poutcome, y) %>%

group\_by(poutcome) %>%

mutate(prop = n / sum(n)) %>%

ggplot(aes(x = poutcome, y = prop, fill = y)) +

geom\_bar(stat = "identity") +

scale\_y\_continuous(labels = scales::percent) +

labs(title = "之前营销结果 vs 当前订阅比例（排除other/unknown）", # 更新标题

x = "之前营销结果",

y = "占比",

fill = "是否订阅") +

scale\_fill\_manual(values = c("lightgray", "tomato")) +

theme\_minimal()

# 联系日期与时长散点图

ggplot(data, aes(x = day, y = duration, color = y)) +

geom\_jitter(alpha = 0.4) +

labs(title = "联系日期 vs 时长", x = "日期", y = "联系时长（秒）") +

scale\_color\_manual(values = c("gray", "red")) +

theme\_minimal()

# 这个地方要去掉一个极端值

# 计算描述性统计量 ----------------------------------------------------------------

vars <- data[, c("age", "duration", "campaign", "pdays", "previous", "balance")]

summary(vars)

sapply(vars, mean, na.rm = TRUE)

sapply(vars, sd, na.rm = TRUE)

# 总体均值的区间估计 ---------------------------------------------------------------

# 估计客户总体的平均年龄

t.test(data$age, conf.level = 0.95)

#95 percent confidence interval: 40.85153 41.62954

# 单个总体均值的假设检验 -------------------------------------------------------------

# 研究本次营销活动是否针对近期的活跃客户（即过去 15 天内是否联系过）

t.test(data$pdays, mu = 15, alternative = "less")

# alternative hypothesis: true mean is less than 15

# 两个总体均值的假设检验 -------------------------------------------------------------

# 研究有无房贷和订阅定期存款意愿的关系

data$y\_numeric <- ifelse(data$y == "yes", 1, 0) # 起到一个分类的作用

t.test(y\_numeric ~ housing, data = data) #将y\_numeric(平均订阅率)转换为y=1, n=0)

# 没有房贷的客户显著更可能订阅定期存款

# 独立性检验 -----------------------------------------------------------------

install.packages("vcd")

install.packages("grid")

install.packages("RcmdrMisc")

# 加载必要的库

library(readxl) # 用于读取Excel文件

library(vcd)

library(grid)

library(RcmdrMisc)

# 1. 读取数据

data <- read\_excel("独立性.xlsx") # 替换为你的文件路径

# 2. 检查数据结构

str(data)

# 3. 构建列联表

contingency\_table <- table(data$housing, data$loan)

print(contingency\_table)

# 4. 进行卡方检验

chi\_test <- chisq.test(contingency\_table)

print(chi\_test)

# 5. 可视化列联表

barplot(contingency\_table,

main = "housing与loan的列联表",

xlab = "loan",

ylab = "频数",

legend = TRUE,

beside = TRUE)

# 方差分析 --------------------------------------------------------------------

# 加载必要的库

library(readxl) # 用于读取Excel文件

library(ggplot2) # 用于绘图

# 1. 读取数据

data <- read\_excel("方差分析.xlsx")

# 2. 数据预处理：去掉contact为unknown的数据

data <- data[data$contact != "unknown", ]

# 3. 进行方差分析

# 检查数据结构

str(data)

# 确保duration是数值类型

data$duration <- as.numeric(data$duration)

# 进行单因素方差分析

aov\_result <- aov(duration ~ contact, data = data)

summary(aov\_result) # 查看方差分析结果

# 4. 结果可视化：绘制箱线图

ggplot(data, aes(x = contact, y = duration)) +

geom\_boxplot() +

labs(title = "不同联系方式下的通话时间分布",

x = "联系方式",

y = "通话时间 (秒)") +

theme\_minimal()

# 5. 多重比较：如果ANOVA结果显著，进行Tukey HSD检验

if(summary(aov\_result)[[1]]$`Pr(>F)`[1] < 0.05) {

tukey\_result <- TukeyHSD(aov\_result)

print(tukey\_result) # 查看多重比较结果

} else {

cat("ANOVA结果不显著，无需进行多重比较。\n")

}

# 相关分析 --------------------------------------------------------------------

install.packages("readxl")

library(readxl)

# 读取Excel数据

data <- read\_excel("相关分析.xlsx", sheet = "Sheet1")

# 数据预处理：过滤pdays为-1的观测（从未联系过的客户）

filtered\_data <- subset(data, pdays != -1)

# 检查数据摘要

cat("数据摘要：\n")

summary(filtered\_data)

# 绘制散点图观察关系

plot(filtered\_data$campaign, filtered\_data$pdays,

main = "联系次数与距离上次联系天数的关系",

xlab = "本次活动联系次数 (campaign)",

ylab = "距离上次联系天数 (pdays)",

pch = 19, col = "steelblue")

# 计算皮尔逊相关系数及检验

pearson\_test <- cor.test(filtered\_data$campaign, filtered\_data$pdays, method = "pearson")

# 计算斯皮尔曼相关系数及检验（非参数方法）

spearman\_test <- cor.test(filtered\_data$campaign, filtered\_data$pdays, method = "spearman")

# 输出结果

cat("\n皮尔逊相关系数分析结果：\n")

print(pearson\_test)

cat("\n斯皮尔曼相关系数分析结果：\n")

print(spearman\_test)

1. 个人心得体会

苏杰斌：作为本次R语言实验报告的组长，我主要负责整体报告的撰写、框架设计与内容整合，并承担了协调组员分工、把控进度与质量的核心职责。以下是我的心得体会：

一、具体工作与挑战

1. 报告统筹与撰写：我精心构建了报告的逻辑框架，包括引言、数据说明、方法论述、结果分析及结论总结，并亲自执笔，完成了其中的核心分析部分与深入的讨论章节。为确保专业性，我反复查阅R函数文档（如‘ggplot2’的精细参数调整），并严格验证代码输出的准确性。

2. 团队协作管理：初期面临进度不统一的难题——部分组员因对‘dplyr’数据清洗不熟悉而滞后。我迅速响应，组织线上会议，通过共享屏幕，详细演示了管道操作符‘%>%’与‘group\_by()’及‘summarize()’函数的组合应用技巧，并进一步细化任务分配，让A同学专注于数据清洗，B同学则负责数据可视化工作。通过每日进度打卡与共享代码GitHub仓库，显著提升了协作效率。

二、困难与解决方案

关键挑战：结果图表与文字分析脱节

在整合报告时，发现部分使用‘ggplot2’绘制的箱线图未能清晰地展示数据分布和离群点，以及未能有效地反映假设检验的结论。我采取以下措施：

1.重梳理分析逻辑链：明确每张图表的分析目标；

2.增加图表解读标注：在代码注释中强制要求撰写‘ggtitle()’和‘labs(caption=)’说明；

3.交叉复核机制：安排组员交换检查代码输出与报告描述的匹配度。

最终，经过三次迭代修改，成功实现了数据可视化与文本论证的紧密结合。

三、课程收获

本次实验不仅提升了我的R语言实战能力，更让我认识到技术严谨性与团队领导力的共生关系。感谢课程提供的成长契机，这段经历将成为我数据分析之路的重要基石。

张颖怡：在本次小组作业中，我主要负责搜集数据集、管理项目文档、建议分析角度和思路的工作，同时合作完成第一二次报告的代码编写和画图。借助AI建议，我搜寻国外数据库和期刊网站，在UC Irvine的机器学习数据集中找到了说明完备的数据集作为作业材料；用GitHub存储、说明项目进展，与组员实时共享；在完成第一二次报告的工作中，整理数据、编写R文件进行数据分析。

本次小组作业使我获益良多。我发现许多可用数据集的门户网站，并且对协调工作流程、保存网页、辅助代码编写的技术工具了解更加深入；了解统计学数据分析方法的应用；具有了从无到有，做出一个用统计方法分析现实问题的小项目的经验。

本次作业也有许多挑战。其中，如何寻找具有建设性的分析角度和思路是最为让我头疼的；项目管理和沟通虽然有工具的辅助，但是协调仍然会因为对于工具的熟练程度不同而产生不同步；图表的美化、代码的编写还较为生疏、不规范。这些都是在未来的学习与工作中需要继续改进的地方。

整体而言，这次小组项目让我对数据分析的完整流程有了更清晰的理解，也提升了我在团队协作、工具使用和独立思考方面的能力。通过一次次实际操作，我更加确认了自己对数据分析方向的兴趣，也为今后相关课程或实习积累了宝贵经验。未来我希望继续拓展技能，参与更多实践项目，提升分析能力与表达能力，为职业发展打下坚实基础。

陆薪吉：

一、具体工作

1.图表导出

使用组员已经完成的R语言绘制客户画像图表，包括年龄分布直方图、职业分布条形图、教育水平饼图、婚姻状况与住房贷款的关系图、信用违约与个人贷款的关系图等等。绘制联系时长与订阅率的散点图、每月联系次数分布图、联系次数与订阅情况的关系图，以及之前营销结果对当前订阅影响关系图。调整图表样式以提高可读性，并使用有关函数将图表保存为PDF格式，方便后续制作。

2.数据分析与规律总结

观察图表中的趋势，得出相应的规律。结合业务背景，提出可能的解释，如“中青年客户经济能力较强，更愿意尝试新产品”“频繁联系可能被视为骚扰，降低客户体验”。

3.PPT制作与汇报准备

确保逻辑清晰、重点突出。为每张图表添加简要说明，如“优化联系频率可提高订阅率”“针对高教育水平客户设计差异化促销策略”。

1. 遇到的困难及解决方法

1.分辨率问题

初始导出的pdf图片在PPT中显示模糊，影响美观。解决方法：用ggsave()调整dpi参数

2.数据分布不均导致图表太畸形

部分类别样本过少，在饼图中占比极小，难以观察。解决方法：将小类别合并为“其他”

3.蛮多变量关系复杂

婚姻状况、住房贷款、订阅率三者关系复杂，普通图表难以表达。解决方法：使用堆叠条形图+分面

4.PPT逻辑梳理有问题

因为图表较多。解决方法：按“客户画像→促销活动→关键结论”分层编排，且聚焦于核心观点。

三、课程感想

通过本次课题，我深刻体会到数据可视化在商业分析中的重要性。清晰的图表能直观展示数据规律，也能帮助决策者快速抓住重点。我也学习了如何从业务角度解读数据，而非仅仅停留在技术层面。可以提供真实业务场景案例：如结合某电商平台的促销数据，分析如何通过客户画像提升ROI。还可以优化R语言代码指导：部分图表调试较复杂，可提供更详细的代码注释或常见错误解决方案。本次课题提升了我的R语言和数据分析能力，也让我学会了如何将技术成果转化为商业洞察。

叶淇：