1. 项目背景

- **Mission:**

数据集包含了一家公司员工的信息，包括他们的教育背景、工作经历、人口统计特征等。此数据集可广泛应用于人力资源（HR）和劳动力相关的各种分析。它能帮助我们深入了解员工留存情况、评估薪资结构等。请使用概率统计分析与建模技术帮助人力资源部门洞悉员工离职原因，并给出一些建议来减少员工离职行为。

- **Data:**

Canvas文件栏目：期末项目/employee.csv

1. Education：员工的教育背景，包括学位、毕业院校及所学专业。
2. Joining Year：员工加入公司的年份，反映其服务年限。
3. City：员工所在地或工作城市。
4. Payment Tier：将员工划分为不同的薪资级别。
5. Age：员工的年龄。
6. Gender：员工的性别认同。
7. Ever Benched：表明员工是否曾有过未被分配工作的临时状态。
8. Experience in Current Domain：员工在当前领域工作的年数。
9. Leave or Not：0未离职，1离职

2. 数据预处理

```
library(readr)
library(tidyverse)
library(ggplot2)
employee <- read_csv('F:/dataAnalysisModel/employee.csv')
# 检查是否有任何缺失值
any_missing <- any(is.na(employee))
cat("是否存在任何缺失值：", ifelse(any_missing, "是", "否"), "\n")

# 检查是否有空字符串
has_empty_strings <- sapply(employee, function(x) any(x == ""))
cat("是否存在任何空字符串：", ifelse(any(has_empty_strings), "是", "否"), "\n")
```

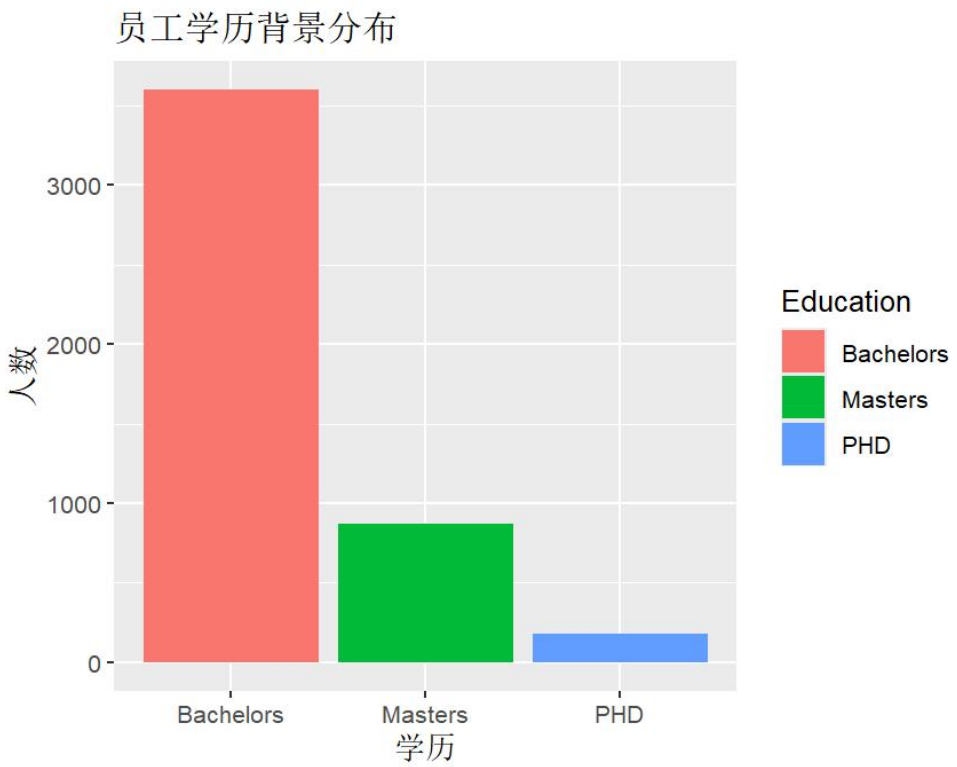
是否存在任何缺失值： 否
是否存在任何空字符串： 否

经分析得出本数据集不含缺失值与空值。

3. 数据分析

2.1 员工的学历背景分布

```
# 1. 员工的学历背景分布
education_distribution <- employee %>%
  group_by(Education) %>%
  summarise(Count = n())

ggplot(education_distribution, aes(x = Education, y = Count, fill = Education)) +
  geom_bar(stat = "identity") +
  ggtitle("员工学历背景分布") +
  xlab("学历") +
  ylab("人数")
```

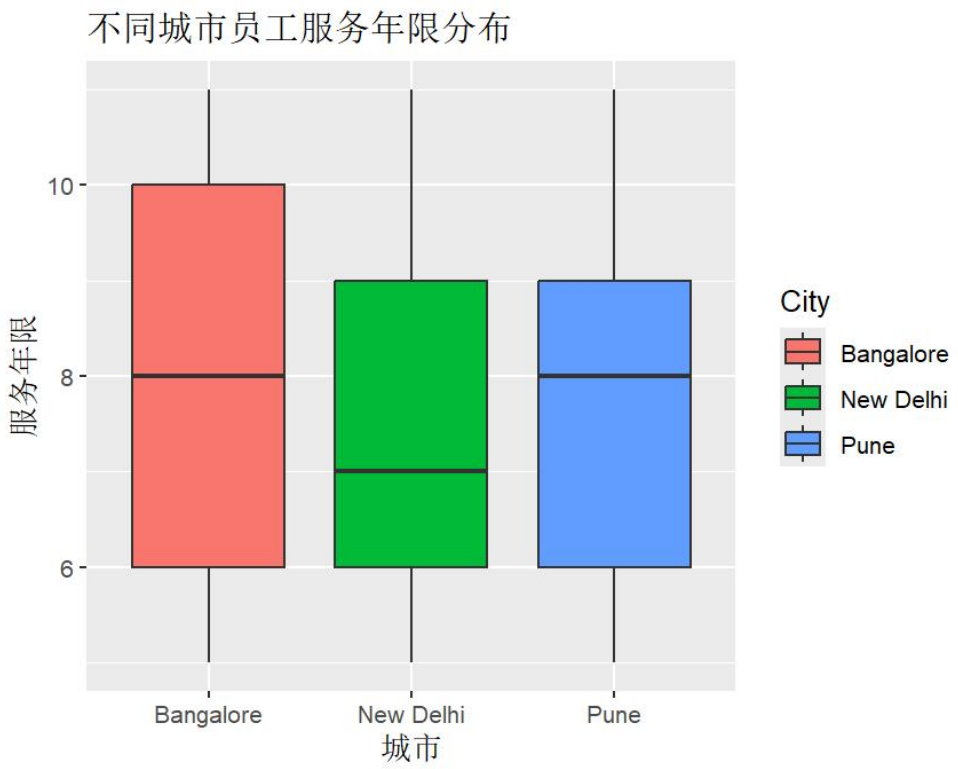


从图片可以看出，员工的学历背景呈现出明显的集中趋势。大多数员工的最高学历是学士学位，占据了绝大部分比例；而拥有硕士和博士学位的员工则相对较少。这反映了公司对人才招聘的学历要求更偏向于本科层次，本科学历已经能够满足大部分岗位的需求。

2.2 不同城市的员工服务年限有何差异？差异是否显著？

```
# 2. 不同城市员工服务年限差异
employee$ServiceYears <- 2023 - employee$JoiningYear
# 显著性检验：ANOVA
anova_result <- aov(ServiceYears ~ City, data = employee)
summary(anova_result)
# 可视化不同城市员工服务年限
ggplot(employee, aes(x = City, y = ServiceYears, fill = City)) +
  geom_boxplot() +
  ggtitle("不同城市员工服务年限分布") +
  xlab("城市") +
  ylab("服务年限")
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
City          2    341   170.4   50.12 <2e-16 ***
Residuals   4650  15812     3.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



从方差分析可以看出，不同城市的员工服务年限存在显著差异。New Delhi 员工的服务年限中位数明显低于其他城市，这表明该城市的员工流动性较高。而 Bangalore 和 Pune 的服务年限相对接近，呈现出较低的流动性，其工作机会更符合员工长期发展的需求。

2.3 薪资等级与当前领域经验之间是否存在某种关联？

```
# 3. 薪资等级与当前领域经验的关联性
cor_test <- cor.test(employee$PaymentTier, employee$ExperienceInCurrentDomain)
print(cor_test)
```

```
        Pearson's product-moment correlation

data:  employee$PaymentTier and employee$ExperienceInCurrentDomain
t = 1.2492, df = 4651, p-value = 0.2116
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.01042556  0.04702397
sample estimates:
       cor
0.01831432
```

相关性检验分显示薪资等级与当前领域经验之间的相关性极弱，且统计上不显著(p=0.2116)。这表明员工的薪资水平与他们在当前领域积累的工作经验没有直接的线性关系。

4. 通过构建贝叶斯网络预测员工是否会离职，评估模型质量

```r
library(e1071)
library(caret)

# 数据预处理：将分类变量转换为因子
employee$Education <- as.factor(employee$Education)
employee$City <- as.factor(employee$City)
employee$Gender <- as.factor(employee$Gender)
employee$EverBenched <- as.factor(employee$EverBenched)
employee$LeaveOrNot <- as.factor(employee$LeaveOrNot)

# 划分数据集为训练集与测试集
set.seed(123)
train_index <- createDataPartition(employee$LeaveOrNot, p = 0.7, list = FALSE)
train_data <- employee[train_index, ]
test_data <- employee[-train_index, ]

naive_bayes_model <- naiveBayes(LeaveOrNot ~ ., data = train_data)
predictions <- predict(naive_bayes_model, test_data)

conf_matrix <- confusionMatrix(predictions, test_data$LeaveOrNot)
print(conf_matrix)

accuracy <- conf_matrix$overall['Accuracy']
print(paste("模型准确率为:", round(accuracy, 4)))
```

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 762 241
         1 153 239

               Accuracy : 0.7176
                 95% CI : (0.6931, 0.7411)
    No Information Rate : 0.6559
    P-Value [Acc > NIR] : 4.960e-07

                  Kappa : 0.3458

 Mcnemar's Test P-Value : 1.171e-05

            Sensitivity : 0.8328
            Specificity : 0.4979
         Pos Pred Value : 0.7597
         Neg Pred Value : 0.6097
             Prevalence : 0.6559
         Detection Rate : 0.5462
   Detection Prevalence : 0.7190
      Balanced Accuracy : 0.6654

       'Positive' Class : 0

"模型准确率为: 0.7176"
```

通过构建朴素贝叶斯模型，得到了较为满意的预测结果。模型的整体准确率为 71.76%，这表明大约 71.76% 的员工离职情况得到了正确预测。然而，从混淆矩阵中可以看出，模型在预测未离职员工方面表现较弱。具体而言，模型的灵敏度（83.28%）较高，说明它能够较好地识别离职员工；然而，特异性仅为 49.79%，即对于未离职员工的预测准确性不高。Kappa

值为 0.3458，属于中等水平，表明模型的预测能力一般，还有改进的空间。最后，McNemar 检验的 p 值为 1.171e-05，显示模型的预测结果显著优于随机猜测，具有统计学意义。

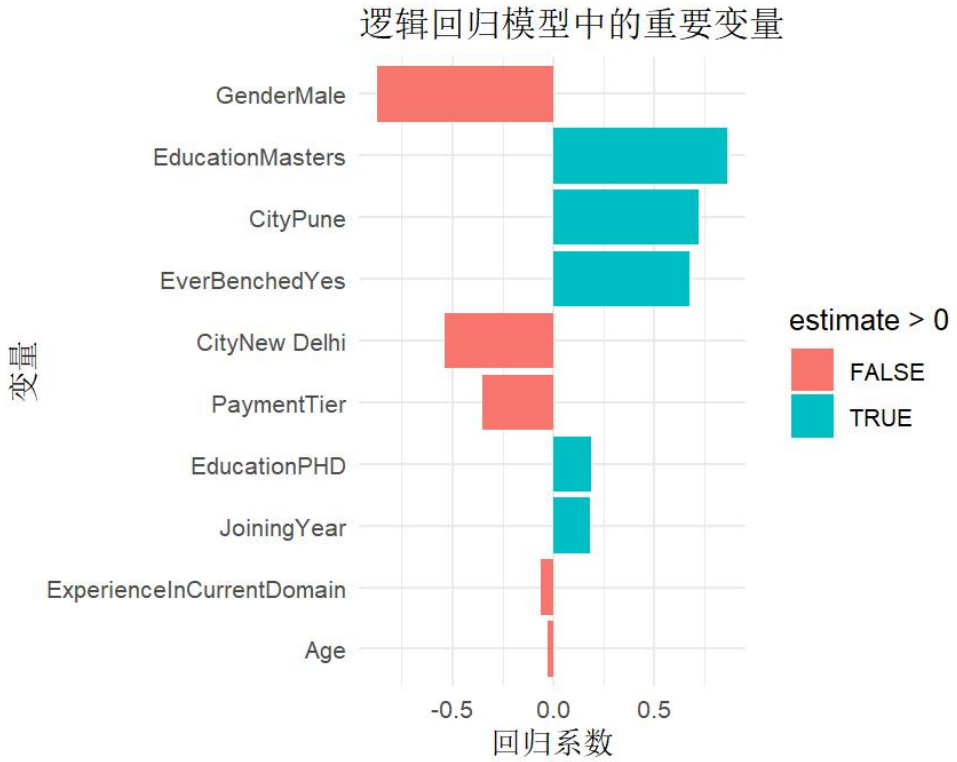
5. 构建逻辑回归模型对员工是否会离职进行预测，并通过模型解读员工离职的影响因素。

```
library(caret)
library(dplyr)
library(broom)
set.seed(123)  # 设置随机种子
train_index <- createDataPartition(employee$LeaveOrNot, p = 0.7, list = FALSE)
train_data <- employee[train_index, ]
test_data <- employee[-train_index, ]
# 构建逻辑回归模型
logit_model <- glm(LeaveOrNot ~ .-ServiceYears, data = train_data, family = binomial)
summary(logit_model)

# 预测测试集
test_data$pred_prob <- predict(logit_model, test_data, type = "response")
test_data$pred_class <- ifelse(test_data$pred_prob > 0.5, 1, 0)
test_data$pred_class <- as.factor(test_data$pred_class)

# 生成混淆矩阵
confusion_matrix <- table(Predicted = test_data$pred_class, Actual = test_data$LeaveOrNot)
print(confusion_matrix)
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("模型准确率:", round(accuracy, 4)))

# 可视化重要变量（回归系数）
coefficients <- tidy(logit_model)
coefficients <- coefficients %>%
  filter(term != "(Intercept)") %>%
  arrange(desc(abs(estimate)))

ggplot(coefficients, aes(x = reorder(term, abs(estimate)), y = estimate, fill = estimate > 0)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ggtitle("逻辑回归模型中的重要变量") +
  xlab("变量") +
  ylab("回归系数") +
  theme_minimal()
```

```
Call:
glm(formula = LeaveOrNot ~ . - ServiceYears, family = binomial,
    data = train_data)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)               -3.722e+02  4.515e+01  -8.243  < 2e-16 ***
EducationMasters           8.655e-01  1.132e-01   7.647 2.06e-14 ***
EducationPHD               1.908e-01  2.290e-01   0.833  0.40461
JoiningYear                1.854e-01  2.241e-02   8.273  < 2e-16 ***
CityNew Delhi             -5.355e-01  1.168e-01  -4.583 4.58e-06 ***
CityPune                   7.223e-01  9.853e-02   7.331 2.29e-13 ***
PaymentTier               -3.499e-01  7.429e-02  -4.710 2.47e-06 ***
Age                       -2.743e-02  8.515e-03  -3.222  0.00127 **
GenderMale                -8.744e-01  8.416e-02 -10.390  < 2e-16 ***
EverBenchedYes             6.781e-01  1.243e-01   5.455 4.88e-08 ***
ExperienceInCurrentDomain -6.034e-02  2.620e-02  -2.303  0.02130 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4193.1  on 3257  degrees of freedom
Residual deviance: 3690.2  on 3247  degrees of freedom
AIC: 3712.2

Number of Fisher Scoring iterations: 4
```

```
          Actual
Predicted   0   1
        0 820 279
        1  95 201
```

```
"模型准确率: 0.7319"
```



通过对员工数据构建逻辑回归模型进行分析，识别出薪资等级、是否曾被闲置、学历、性别以及所在城市是影响员工离职的重要变量。

具体来看：性别方面，男性员工离职概率显著低于女性（回归系数 -0.87，$p < 0.01$），其离职可能性仅为女性员工的 41.5%（$e^{-0.8744}$）。此外，学历也对离职概率产生显著影响：与本科相比，硕士学历的员工离职倾向更高（回归系数 0.87，$p < 0.01$），其离职概率是基准水平的 2.38 倍（$e^{-0.8655}$）；而博士学历的影响不显著。城市差异同样明显，新德里 New Delhi 的员工离职概率较低，为基准城市（Bangalore 班加罗尔）的 58.3%（回归系数 -0.54，$p < 0.01$）；而浦那 Pune 的员工离职概率较高，是基准城市的 2.05 倍（回归系数 0.72，$p < 0.01$）。曾被闲置的员工，其离职概率是未被闲置员工的 1.97 倍（回归系数为 0.68，$p < 0.01$）。薪资等级的回归系数为 -0.35（$p < 0.01$），表明薪资越高，员工的离职概率越低；每提高一个薪资等级，离职的对数概率减少约 0.35。年龄和当前领域的工作经验也发挥了一定作用，年龄每增加一岁，离职概率略微降低（回归系数 -0.027，$p = 0.001$），而领域工作经验增加同样会减少离职可能性（回归系数 -0.060，$p = 0.02$）。

模型整体准确率达到 73.2%，其中正确预测未离职员工 820 人，正确预测离职员工 201 人，表现出较好的预测效果。

6. 模型可能存在的不足并给出优化建议

贝叶斯网络模型存在独立性假设不准确的问题，即变量之间的独立性假设可能不符合实际情况，并且自动学习的网络结构可能无法充分捕捉复杂的依赖关系。为此，可以通过改进

结构学习，采用更复杂的算法或结合领域知识来优化模型，同时尝试混合其他机器学习算法提升性能，并通过交叉验证来优化参数。

逻辑回归模型的问题主要在于特征选择不足，未考虑到一些潜在影响因素（如工作满意度、晋升机会），以及数据不平衡，导致对未离职员工的预测能力较差。优化方法包括增加更多的特征、去除冗余特征，并采用过采样或欠采样技术平衡数据，从而提升预测性能。