Can you help reduce employee turnover?

📖 Background

Employee turnover describes the number or percentage of employees leaving an organisation during a specified time period, and who usually must be replaced. High rates of turnover are usually associated with:

- increased recruitment costs
- decreased productivity, and lower employee morale.

Employee turnover can occur for a number of different reasons: Some people switch careers, others move on due to toxic work environments, and still others move on because they receive a better offer elsewhere or due to changes in personal circumstances. That being said, most voluntary resignations occur due to management problems, lack of opportunities or burnout.

The goal of the following analysis was to use available internal employee data to better understand the turnover situation in the department, which types of employees were more likely to leave and why, and use those insights to present recommendations on how to tackle the issue.

The Board specifically requested answers to the following questions:

- Which department has the highest employee turnover? Which one has the lowest?
- Investigate which variables seem to be better predictors of employee departure.
- How could the organisation reduce employee turnover?

💾 The data

The department has assembled data on almost 10,000 employees. The team used information from exit interviews, performance reviews, and employee records.

- "department" - the department the employee belongs to.
- "promoted" - 1 if the employee was promoted in the previous 24 months, 0 otherwise.
- "review" - the composite score the employee received in their last evaluation.
- "projects" - how many projects the employee is involved in.
- "salary" - for confidentiality reasons, salary comes in three tiers: low, medium, high.
- "tenure" - how many years the employee has been at the company.
- "satisfaction" - a measure of employee satisfaction from surveys.
- "bonus" - 1 if the employee received a bonus in the previous 24 months, 0 otherwise.
- "avg_hrs_month" - the average hours the employee worked in a month.
- "left" - "yes" if the employee ended up leaving, "no" otherwise.

In [ ]:

```
```

library(tidyverse) df <- readr::read_csv('data/employee_churn_data.csv', show_col_types = FALSE) head(df)

Preparation data for analyzing

1. Data type

In [ ]:

```
summary(df)
```

```
 department          promoted           review          projects     
Length:9540        Min.   :0.00000   Min.   :0.3100   Min.   :2.000  
Class :character   1st Qu.:0.00000   1st Qu.:0.5929   1st Qu.:3.000  
Mode  :character   Median :0.00000   Median :0.6475   Median :3.000  
                   Mean   :0.03029   Mean   :0.6518   Mean   :3.275  
                   3rd Qu.:0.00000   3rd Qu.:0.7084   3rd Qu.:4.000  
                   Max.   :1.00000   Max.   :1.0000   Max.   :5.000  
   salary              tenure        satisfaction        bonus        
Length:9540        Min.   : 2.000   Min.   :0.0000   Min.   :0.0000  
Class :character   1st Qu.: 5.000   1st Qu.:0.3868   1st Qu.:0.0000  
Mode  :character   Median : 7.000   Median :0.5008   Median :0.0000  
                   Mean   : 6.556   Mean   :0.5046   Mean   :0.2121  
                   3rd Qu.: 8.000   3rd Qu.:0.6226   3rd Qu.:0.0000  
                   Max.   :12.000   Max.   :1.0000   Max.   :1.0000  
avg_hrs_month       left          
Min.   :171.4   Length:9540        
1st Qu.:181.5   Class :character   
Median :184.6   Mode  :character   
Mean   :184.7                      
3rd Qu.:187.7                      
Max.   :200.9                      
```

2. Missing values

There are no missing value for each vairables.

In [ ]:

```
data.frame(colSums(is.na(df)))
```

A data.frame: 10 × 1

	colSums.is.na.df..
	<dbl>
department	0
promoted	0
review	0
projects	0
salary	0
tenure	0
satisfaction	0

	colSums.is.na.df..
	<dbl>
bonus	0
avg_hrs_month	0
left	0

💻 Exploratory Data Analysis

Turnover Employee Rate by Deparment

In [ ]:
```
library(scales)
turnover<-df%>% group_by(department,left) %>% summarise(count=n())%>% filter(lef

## Summary statistic by department.
stat_sumary<-df%>% group_by(department) %>% summarise(total=n(),avg_satisfaction

inner_join(turnover,stat_sumary, on = c("departments")) %>% mutate(turnover_rate
```

```
`summarise()` has grouped output by 'department'. You can override using the
`.groups` argument.
Joining, by = "department"
```

A grouped_df: 10 × 6

department	avg_satisfaction	avg_review	avg_projects	avg_tenure	turnover_rate
<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<chr>
IT	0.5158181	0.6477460	3.289326	6.609551	30.90%
logistics	0.4930845	0.6543736	3.275000	6.527778	30.83%
retail	0.5027688	0.6501520	3.266061	6.591175	30.56%
marketing	0.5024107	0.6576937	3.280549	6.503741	30.30%
support	0.5065793	0.6504666	3.268027	6.564626	28.84%
engineering	0.5049441	0.6506013	3.263852	6.558047	28.83%
operations	0.5046203	0.6533968	3.271353	6.608410	28.65%
sales	0.5045201	0.6516311	3.286245	6.535847	28.52%
admin	0.5194442	0.6470856	3.278960	6.498818	28.13%
finance	0.4971833	0.6549342	3.293532	6.440299	26.87%

Which department has the highest employee turnover? Which one has the lowest?

All departments share similar mean of satisfaction and reviews which the mean of satisfaction is about 0.5 and of reviews is about 0.65 for all the departments.Also the average tenure among deparment share similar which is close to the over mean tenure of all company, 6.6 year.

Although IT, logistic and retail has the highest rate employee turnover (about 31%) and the lowest rate is for finance (26.87%), the difference is not significant, which suggests that

reasons for leaving may be more systemic in nature.

Distribution of numeric variables

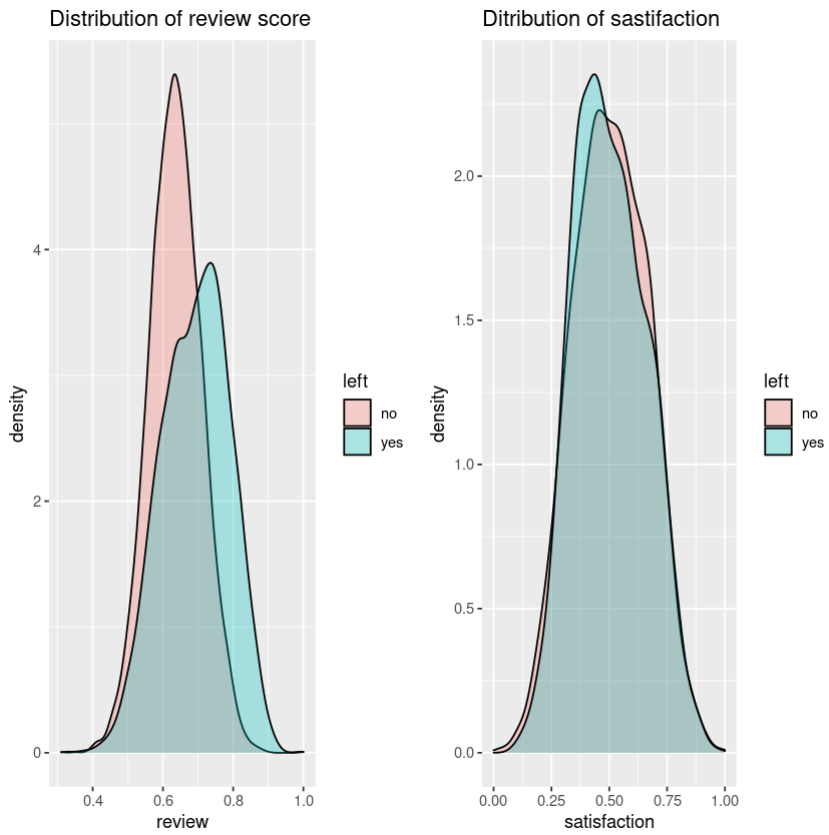
Staffs that left have higher performance reviews and lower satisfaction rates compared to those who stayed

In [ ]:
```
library(gridExtra)

g1<-ggplot(df, aes(x=review, fill=left)) + geom_density(alpha=.3)+labs(title="Di

g2<-ggplot(df,aes(x=satisfaction, fill=left)) + geom_density(alpha=.3)+labs(titl

grid.arrange(g1, g2,  nrow = 1,ncol=2)
```



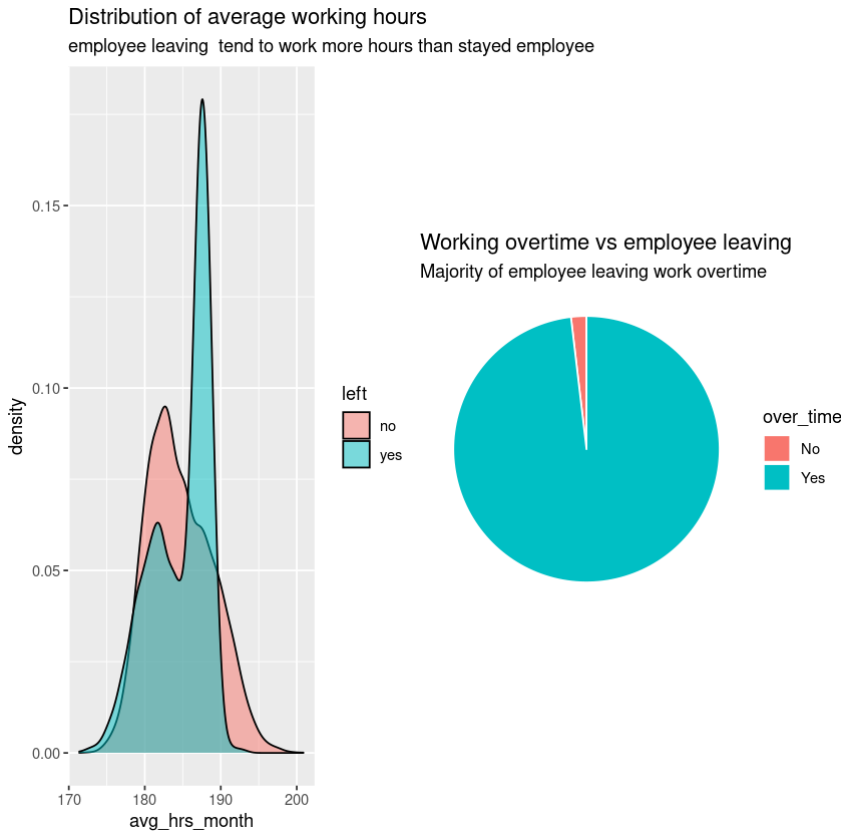
In general, employees who had left the organisation had received higher evaluation scores. Also, the slight right-skew in the distribution of satisfaction scores for those employees who had left the organisation indicated that a greater proportion were less satisfied relative to those that had not left, but it is not really significant.

Employee who left tend to work overtime compared to who stayed

In [ ]:
```
g1<-ggplot(df, aes(x=avg_hrs_month, fill=left)) +
    geom_density(alpha=.5)+labs(title="Distribution of average working hours",su


g2<-df %>% filter(left=="yes")%>%mutate(hours_per_day=avg_hrs_month/22,over_time
geom_bar(stat="identity", width=1, color="white") +
coord_polar("y", start=0)+labs(title="Working overtime vs employee leaving ",sub
```

```
grid.arrange(g1, g2,  nrow = 1,ncol=2)
```



We see the graph for employee leaving toward to left skew compared to employee staying, which indicate they tend to work more hours, and their hours focus on around 185 to 190 hours that why the graph is suddenly high at this interval.

Assuming that a month has 22 working days, for the group of employee that left, they tended to work more than 8 hours a day on average (assuming a 22 working days in a month), relative to those who stayed.

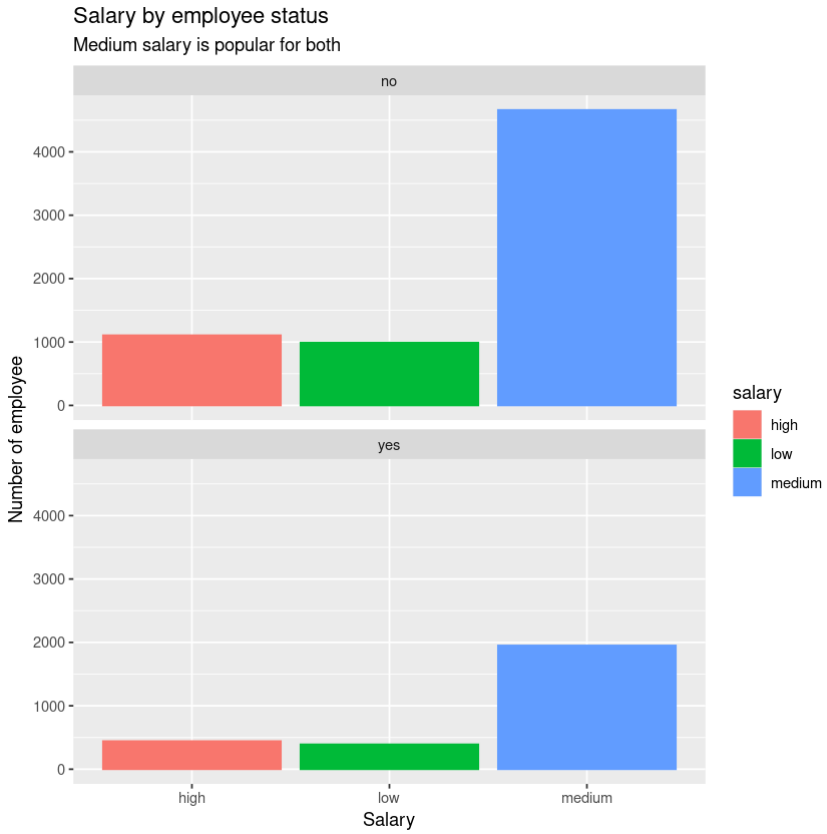
Treatments Vs Employee Turnover Rate

Which level of salary does have significant impact on turover rate?

In [ ]:

```
df%>% group_by(left,salary)%>% summarise(count=n())%>% ggplot( aes(x = salary,y=
geom_col() +
    facet_wrap(.~ left, ncol = 1)+
labs(title="Salary by employee status",y="Number of employee",x="Salary",subtitl
```

```
`summarise()` has grouped output by 'left'. You can override using the
`.groups` argument.
```



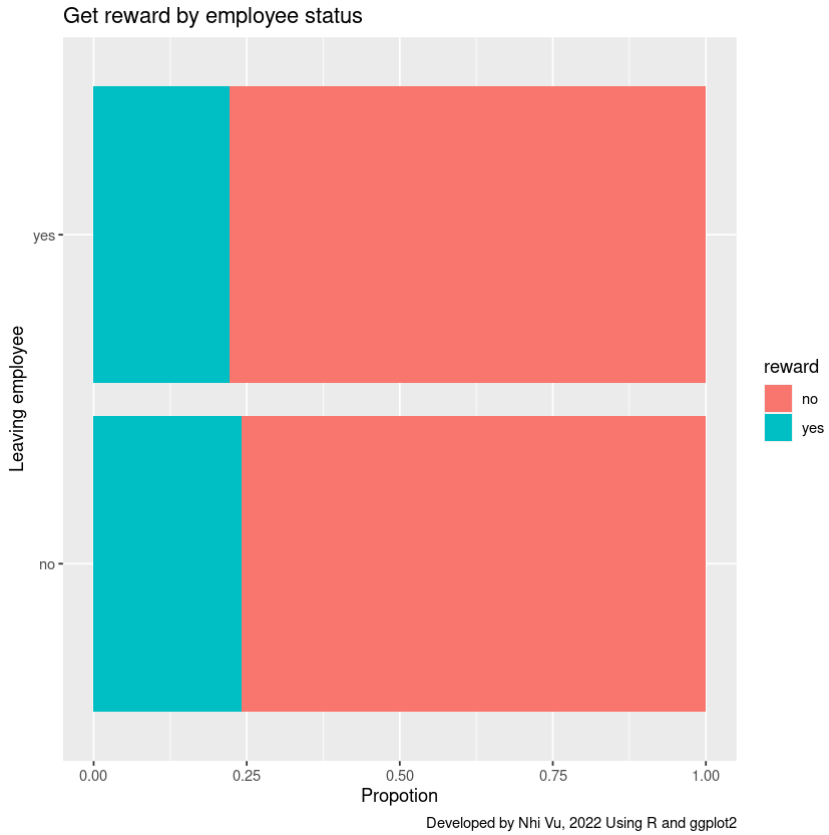
From the compared graph, we see no matter left or stayed employee, the distribution of salary looks similar for both statuses, which medium salary is the most common salary that emloyee in this organization receive.

Employees who left tend to work more hours, whether they will get more rewards than people who stayed?

In [ ]:

```
## Create a new variable reward (get promoted or bonus)
df%>%mutate(reward=case_when(promoted==1|bonus==1 ~ "yes",TRUE~"no"))%>%group_by
labs(x = "Propotion", y = "Leaving employee",
title = "Get reward by employee status",
caption = "Developed by Nhi Vu, 2022 Using R and ggplot2")
```

```
`summarise()` has grouped output by 'left'. You can override using the
`.groups` argument.
```



Based on the graph above, we see the unfairness between both employee statuses. Despite working longer hours, those that left tend to receive proportionately fewer bonuses or promotions (22%) compared to those who stayed (25%).

Contribution and Turnover Rate

The majority of the staff that leave have been working for the company between 5 - 8 years

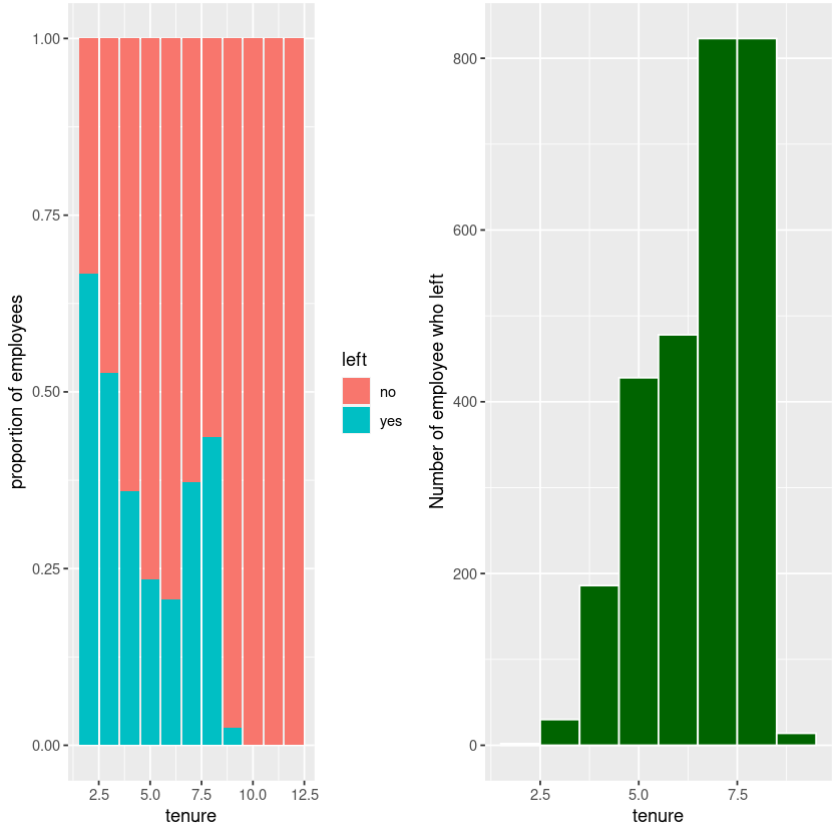
In [ ]:

```
g1<-df%>%group_by(left,tenure)%>%summarise(count=n())%>%ggplot(aes(x=tenure,y=co

g2<-df%>%filter(left=="yes")%>%ggplot( aes(x=tenure)) +
    geom_histogram(binwidth=1,colour="white", fill="dark green")+labs(y="Number

grid.arrange(g1, g2,  nrow = 1,ncol=2)
```

```
`summarise()` has grouped output by 'left'. You can override using the
`.groups` argument.
```

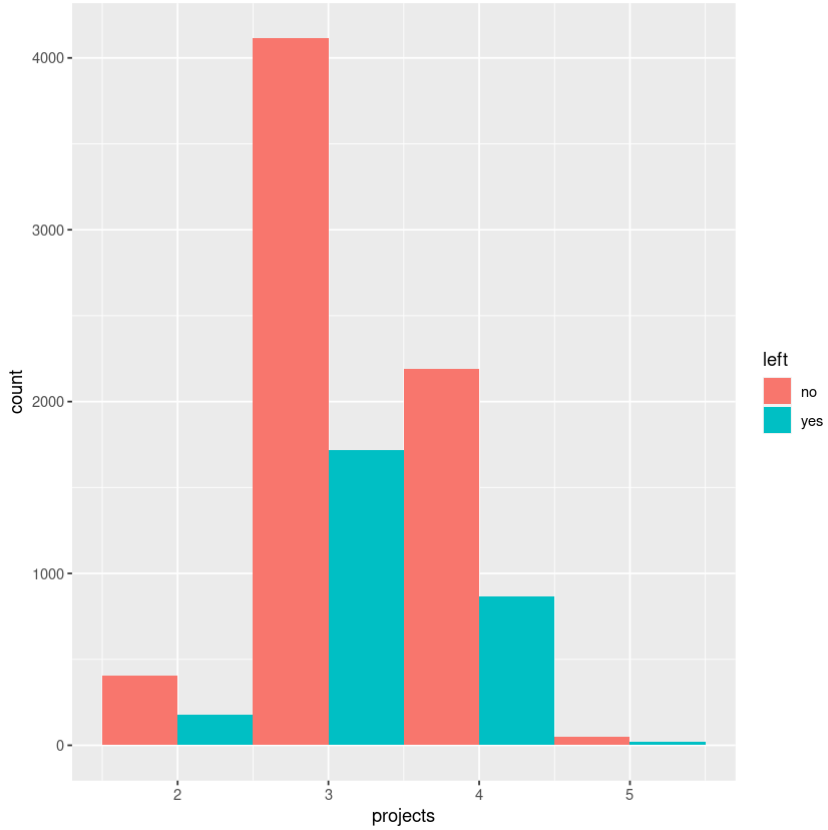


We found almost 91% employees that leave the company have been working for between 5-8 years, and at the first 2 years, we see more than 62.5% employees leaving company and employee who has been working more than 9 years tend to stay (which we find almost 0 % employee leaving at the tenure range more than 9 years.)

The majority of the staff that leave have been done for 3-4 projects

In [ ]:

```
ggplot(df, aes(x=projects, fill=left)) +
    geom_histogram(binwidth=1, position="dodge")
```



Based on tenure, and number of projects, we found that the staffs leaving are people have significant contribution to commpay,these employees are likely to be difficult and costly to replace, as they likely have deep institutional knowledge and skills that will be costly replace, and take time to retrain.

Significance Variables for Predicted model

Based on what we analyzed and visualization of each variable vs turnover rate, We can guess reviews, satisfaction and average working hours have significant impacts on the turnover rate. While department and salary show similar patterns in both employee status.

Finding significant coefficient effects in the logistic regression model predicts the probability of left

$H_0: \beta = 0$

$H_a: \beta \ne 0$

We use level significance $\alpha=0.05$. We reject H_0 when p_value of this coefficient effect less than α.

In [ ]:

```
#Need to convert Chr variables to factors
df$department <- as.factor(df$department)
df$salary <- as.factor(df$salary)
df$left <- as.factor(df$left)
df$promoted <- as.factor(df$promoted)
df$bonus <- as.factor(df$bonus)
```

```
## Fit all variables to predict the probability of left

fit<-glm(left~.,data=df,family="binomial")

## Find the significant variable which have p_value < 0.05

data.frame(summary(fit)$coefficients)%>%filter(Pr...z..<0.05)%>%arrange(Pr...z..
```

A data.frame: 5 × 4

	Estimate	Std..Error	z.value	Pr...z..
	<dbl>	<dbl>	<dbl>	<dbl>
review	11.16085772	0.36322317	30.727273	2.460718e-207
satisfaction	2.45789655	0.18655558	13.175144	1.219947e-39
(Intercept)	-21.47416376	4.74026443	-4.530162	5.893860e-06
promoted1	-0.55897190	0.15689187	-3.562784	3.669420e-04
avg_hrs_month	0.06604027	0.02829798	2.333745	1.960907e-02

As our prediction, we find review has the most impact on employee status, next is satisfaction, promoted and lastly avg_hrs_month. Also:

- Higher review scores increase the risk of staff leaving,
- Working longer hours increase the risk of staff leaving, as well as
- staff who revive promotion reduces the risk of staff leaving

🌳 Decision Tree to Model Leavers

In other hand, Decision Tree indicates the top factor influencing employee attrition is the avg_hrs_worked, next is satisfaction and following by reviews

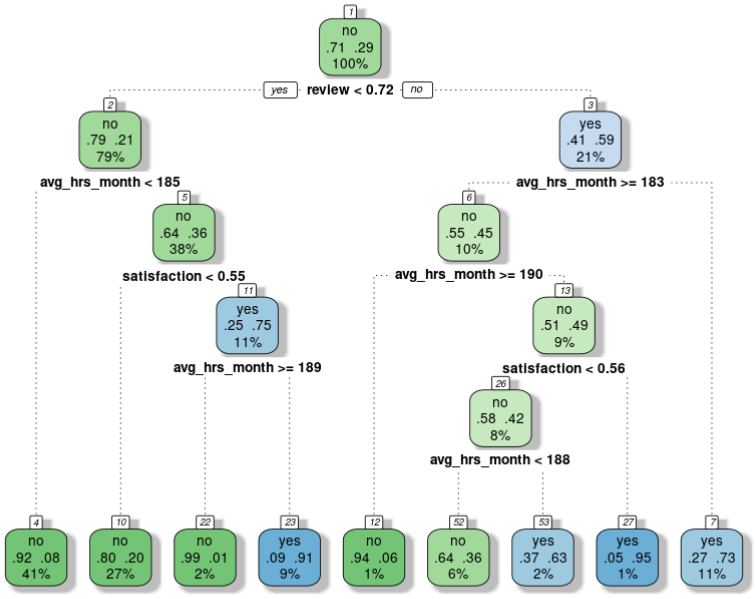
In [ ]:

```
list.of.packages <- c( "rattle", "rpart", "caret", "ggplot2")
new.packages <- list.of.packages[!(list.of.packages %in% installed.packages()[,
if(length(new.packages)) install.packages(new.packages)

library(rpart)
library(RColorBrewer)
library(rattle)
library(caret)

# Allow for randomness in the model and allow us re run it
set.seed(9999)
# Cross Validation - split the dataset into a training and testing set
train <- createDataPartition(df$left,p=0.5,list=FALSE)
df.trn <- df[train,]
df.tst <- df[-train,]

mtree <- rpart(left~., data = df.trn, method = "class", control = rpart.control(
# Code to plot the tree
fancyRpartPlot(mtree)
```



Rattle 2022-Apr-05 03:33:43 repl

From the tree we can see some further details of how these three main factors influence employee attrition. Employee get more than 0.72 review score is tend to leave, on the right side, if they get less than 0.72 but their working hours are greater 185 hours, they tend to leave, also we find that if their satisfaction is low, leading them to leave.

Conclusion and Recommendations

Important variables to predict the chance of leaving of staff are:

- Review score
- Hours working
- Satisfaction
- Reward

Because working more hours increase the risk of leaving, company should work on should be the hours being worked. Don't let them work too much. If employee work over time, to reduce the risk of leaving, this company should offer them reward.

Secondly, investing in gathering information from employees on what drives their satisfaction rating could further help break down what is creating lower satisfaction levels which is contributing to the attrition also.