ASSIGNMENT COVER SHEET

PROGRAMME : Master's in Business Analytics (MsBA)

SUBJECT CODE AND TITLE : BAA5063 – Business Statistics Using R

ASSIGNMENT TITLE : Business Statistics Using R – Individual Assignment

LECTURER : Prof. Keshab Shrestha ASSIGNMENT DUE DATE: 16/12/2024

STUDENT'S DECLARATION

1. I hereby declare that this assignment is based on my own work except where acknowledgement of sources is made.
2. I also declare that this work has not been previously submitted or concurrently submitted for any other courses in Sunway University/College or other institutions.

[Submit "Turn-it-in" report (please tick √): Yes __√__ No _____]

NO.	NAME	STUDENT ID NO.	SIGNATURE	DATE
1.	Harresh A/L Ragunathan	19076090	*Harresh*	16/12/2024

E-mail Address / Addresses (according to the order of names above):

1. 19076090@imail.sunway.edu.my	

APPROVAL FOR LATE SUBMISSION OF ASSIGNMENT (If applicable)

IF extension is granted, what is the revised due date? ______________________

Signature of Lecturer: ______________________ Date: ______________________

Marker's Comments:

Marks and / or Grade Awarded: ______________________ Date: ______________________

ADDENDUM

USE OF ARTIFICAL INTELLIGENCE (A.I.) DECLARATION

Students are allowed to use AI to support completion of assessments. However, students are reminded to do so ethically and transparently. This is so that (a) submissions can be fairly and accurately marked; and (b) feedback can be provided on the content that reflects student ability, in order to help with future submissions. Students are also reminded that in accordance with the University's Academic Malpractice Policy, Item 4.11.2, "*… the representation of work: written, visual, practical or otherwise, of any other person, including another student or* ***anonymous web-based material*** *[emphasis added], or any institution, as the candidate's own*" is considered malpractice.

Declaration

[√] I / We used the following A.I. tools to produce content in this submission:

Tool	**Purpose**	**Prompts**	**Sections where AI output was used / Outcome(s) in the submission**
e.g. ChatGPT	*e.g. Generating points for the essay* *Structuring the essay*	*e.g. "Give me 5 key talking points for an essay on…"* *"Show me a structure for an essay on…"*	*e.g. The main point for Section 1.2 and 1.3 were generated by AI, but the discussion was not.* *The organization / structure of the essay was suggested by AI*
e.g. Grammarly	*e.g. Correcting grammar and spelling, improving sentence structure*	*N/A*	*e.g. Grammarly suggestions were used for all sections of the essay*

Note: Add additional rows if necessary.

OR

[ ] I / We did not use any A.I. tools to produce any of the content in this submission.

NO.	NAME	STUDENT ID NO.	SIGNATURE	DATE
1.	Harresh A/L Ragunathan	19076090	*Harresh*	16/12/2024

E-mail Address / Addresses (according to the order of names above):

1. 19076090@imail.sunway.edu.my	

Table of Contents

Introduction

Employee turnover, the rotation of workers around the labor market and between states of employment and unemployment (Ongori, 2007), is a challenge faced by many businesses. With employees leaving companies and being replaced by new recruits, it introduces more costs for the company as the hiring and training process is relatively expensive. Furthermore, these processes involve advertising the position, conducting interviews, and onboarding the new employees, which requires a substantial amount of time investment. In order for businesses to mitigate these costs, businesses may want to implement machine learning models, such as Logistic Regression, to not only predict employee turnover, but also identify the key aspects that influence employee turnover. By leveraging diverse and high-quality datasets, businesses can gain actionable insights regarding employee turnover, enabling them to implement various measures and strategies for employee retention.

Literature Review

Causes of Employee Turnover

There are various reasons why employee turnover occurs in business, both voluntarily and involuntarily. Voluntary turnover occurs when an employee decides to leave the company on their own accord (Ongori, 2007). Job satisfaction is a key factor that leads to high employee turnover. As highlighted by Masood (2024), enhancing job satisfaction and employee well-being are critical factors in reducing employee turnover. Additionally, Liu (2014) suggests that lack of growth opportunities is another factor that leads to high employee turnover. This lack of growth opportunities may include the lack of environment for success, lack of challenging responsibilities, and lack of recognition (Liu, 2014). On the other hand, involuntary turnover occurs when an organization terminates the services of an employee, resulting in the employee resigning unwillingly (Dwesini, 2019). As highlighted by Dwesini (2019), the reason for dismissal may be due to the employee's poor performance, layoffs, or downsizing.

Effects of Employee Turnover

Employee turnover affects businesses in various ways. Ongori (2007) highlights that employee turnover is expensive from the point of view of organizations. Due to high employee turnover, employees need to be replaced, which imposes a lot of costs. These replacement costs include searching of the external labor market for possible substitutes, selection between competing substitutes, and induction of the selected substitute (Sutherland, 2002). Furthermore, as Ongori (2007) suggests, in addition to the cost of replacement, a business' output will also be affected to a certain extent due to focusing on the recruitment and training process of these substitutes. In general, employee turnover results in negative impacts on businesses. However, in some cases, employee turnover can also result in positive effects. As suggested by Ampomah and Cudjor (2015), employee turnover could lead to the replacement of poor performing employees, introduce new skills and ideas to the organization, and aid in reducing redundancy in the organization

Logistic Regression

For this study, Logistic Regression will be used to predict employee turnover and identify the significant variables that affect employee turnover. Logistic Regression is a traditional classification algorithm that involves linear discriminants, which will output a probability that a given input belongs in a certain class (Zhao et al., 2019). It is generally utilized for two-class classification, and it is a measurable technique for predicting binary classes (Ponnuru et al., 2020). Additionally, Logistic Regression has many strengths such as that it is computationally efficient compared to other complex models. Additionally, it is simple and straightforward, allowing for the results to be easily interpreted. Lastly, it has the ability to handle both numeric and categorical predictors, allowing for a diverse range of predictors.

Exploratory Data Analysis

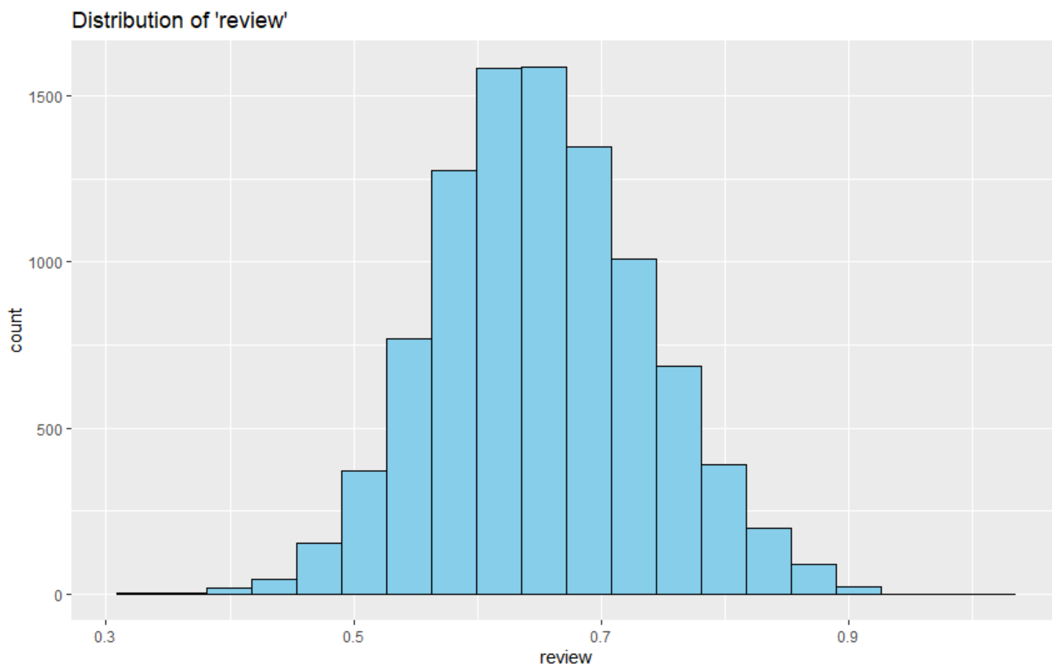
```
> head(df)
# A tibble: 6 × 10
  department promoted review projects salary tenure satisfaction bonus avg_hrs_month left
  <chr>         <dbl>  <dbl>    <dbl> <chr>   <dbl>        <dbl> <dbl>         <dbl> <chr>
1 operations        0  0.578        3 low         5        0.627     0          181. no
2 operations        0  0.752        3 medium      6        0.444     0          183. no
3 support           0  0.723        3 medium      6        0.447     0          184. no
4 logistics         0  0.675        4 high        8        0.440     0          189. no
5 sales             0  0.676        3 high        5        0.578     1          180. no
6 IT                0  0.683        2 medium      5        0.565     1          179. no
```

```
> str(df)
spc_tbl_ [9,540 × 10] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ department   : chr [1:9540] "operations" "operations" "support" "logistics" ...
 $ promoted     : num [1:9540] 0 0 0 0 0 0 0 0 0 0 ...
 $ review       : num [1:9540] 0.578 0.752 0.723 0.675 0.676 ...
 $ projects     : num [1:9540] 3 3 3 4 3 2 4 4 4 3 ...
 $ salary       : chr [1:9540] "low" "medium" "medium" "high" ...
 $ tenure       : num [1:9540] 5 6 6 8 5 5 5 7 6 6 ...
 $ satisfaction : num [1:9540] 0.627 0.444 0.447 0.44 0.578 ...
 $ bonus        : num [1:9540] 0 0 0 0 1 1 0 1 0 0 ...
 $ avg_hrs_month: num [1:9540] 181 183 184 189 180 ...
 $ left         : chr [1:9540] "no" "no" "no" "no" ...
```

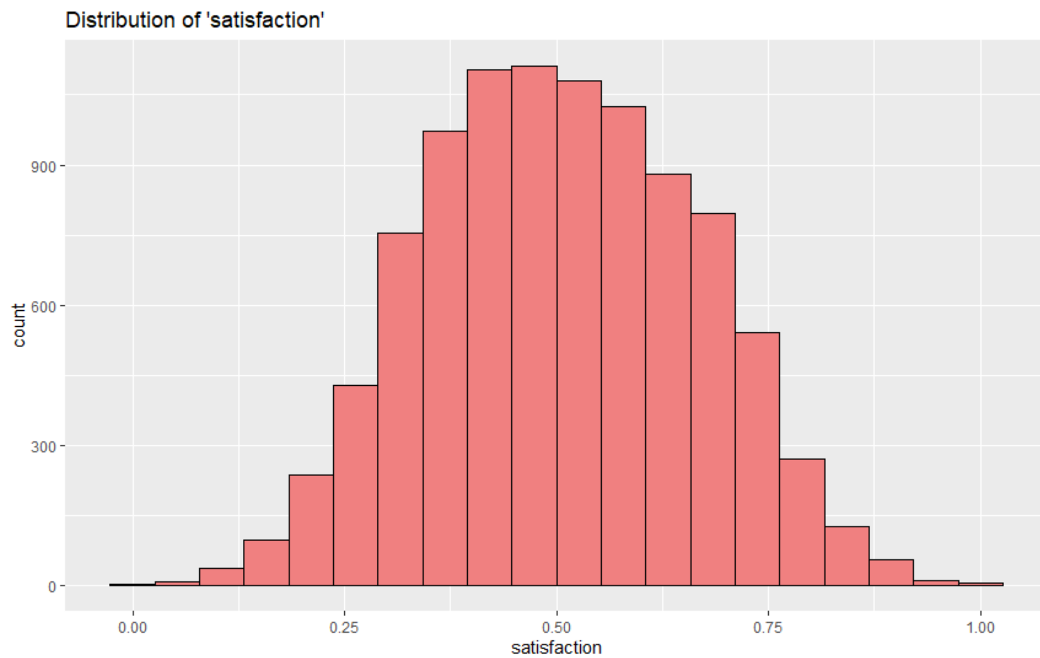
Upon initial inspection, it may be seen that the dataset has 9,540 rows and 10 columns, 3 categorical columns and 7 numerical columns. Although there are 7 numerical columns, it should be noted that the "promoted" and "bonus" columns contain binary values, therefore they are treated as categorical variables. Meanwhile, the "review", "tenure", "satisfaction", and "avg_hours_month" columns are continuous numerical variables, while the "projects" column is a discrete numerical variable.

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

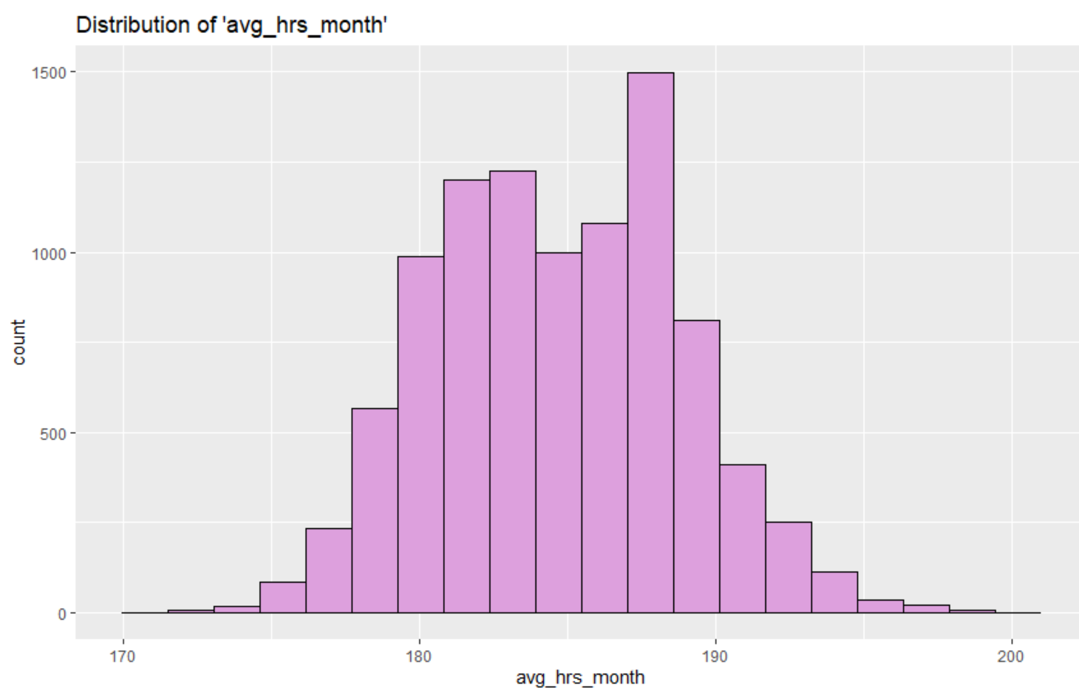
The above summary statistics highlight the key statistics for each column. Looking into "review", it may be seen that it has a mean of 0.6518 and a median of 0.6475, indicating a slight positive skew. This may imply that a handful of employees may have higher review scores compared to others. Observing the statistics for "projects", it may be noticed that it has a mean of 3.275 and a median of 3, indicating a slight positive skew. This gives the implication that some employees may work on more projects than others. Next, looking into the statistics for “tenure”, it has a mean of 6.556 and a median of 7, indicating a slight negative skew. This implies that a small number of employees have a shorter tenure compared to the majority of employees. Focusing on the statistics for “satisfaction”, it may be seen that it has a mean of 0.5046 and a median of 0.5008, which is close to symmetric, indicating a relatively normal distribution. Finally, observing the statistics for “avg_hours_month”, it has a mean of 184.7 and a median of 184.6, which also indicates a relatively normal distribution.



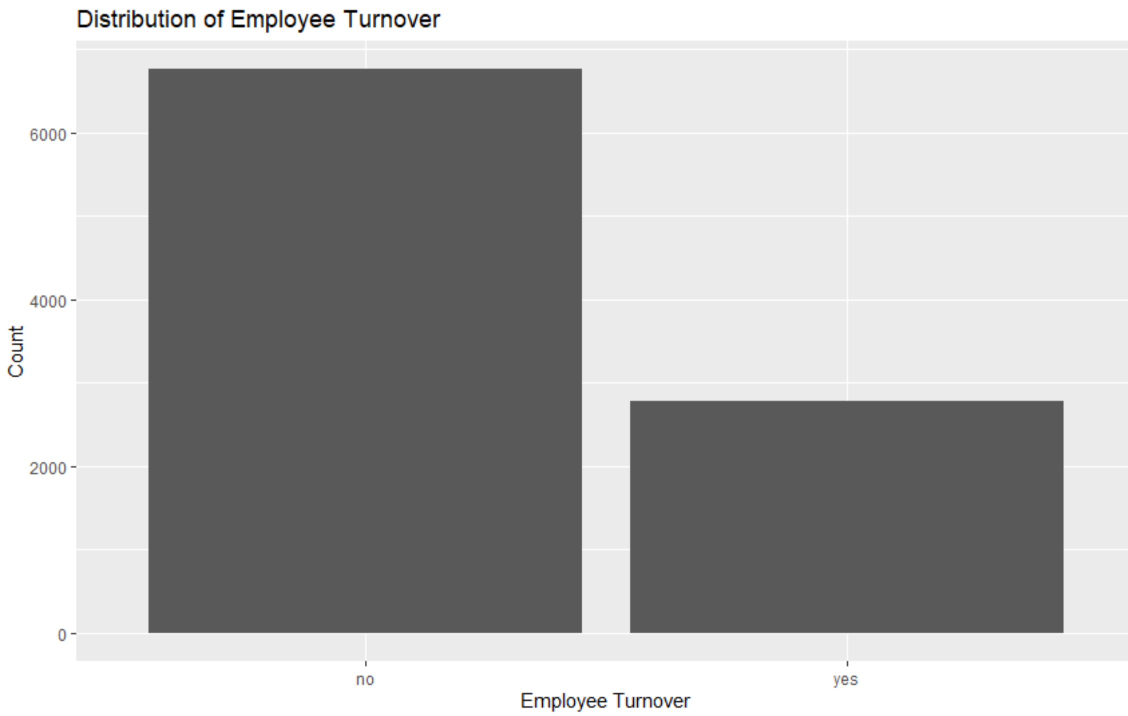
As highlighted by the summary statistics, the “review” variable possesses a slight positive skew as its mean is higher than its median, which is demonstrated by its histogram.



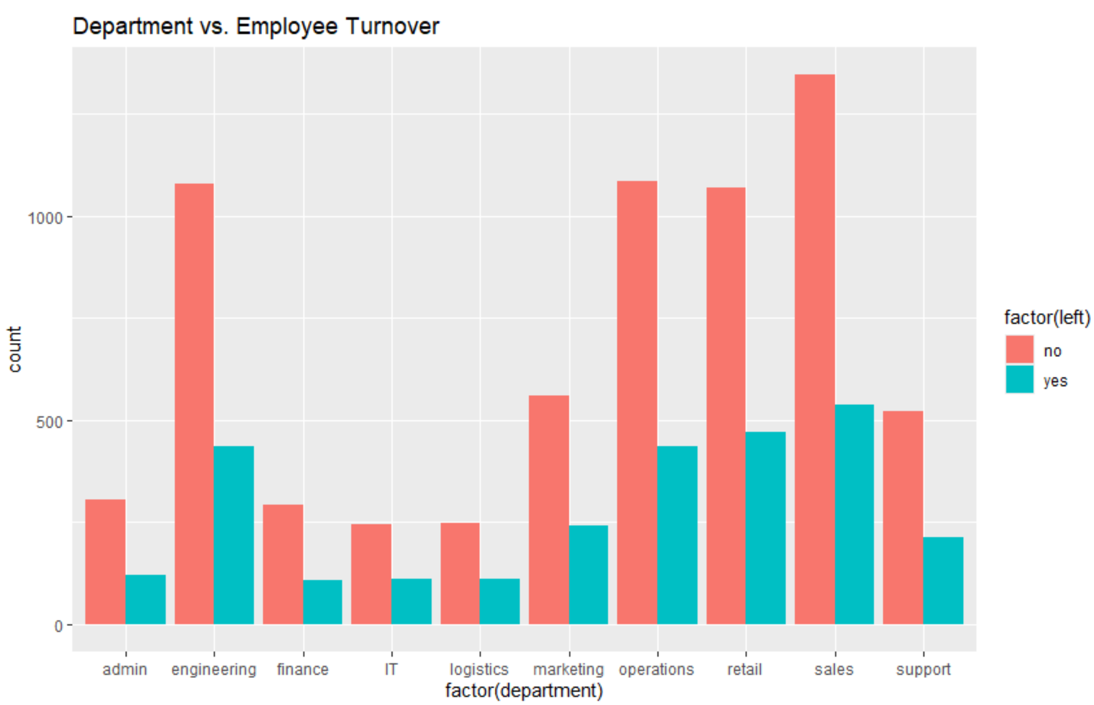
The histogram for “satisfaction” demonstrates a similar positive skew as its mean is also higher than its median.



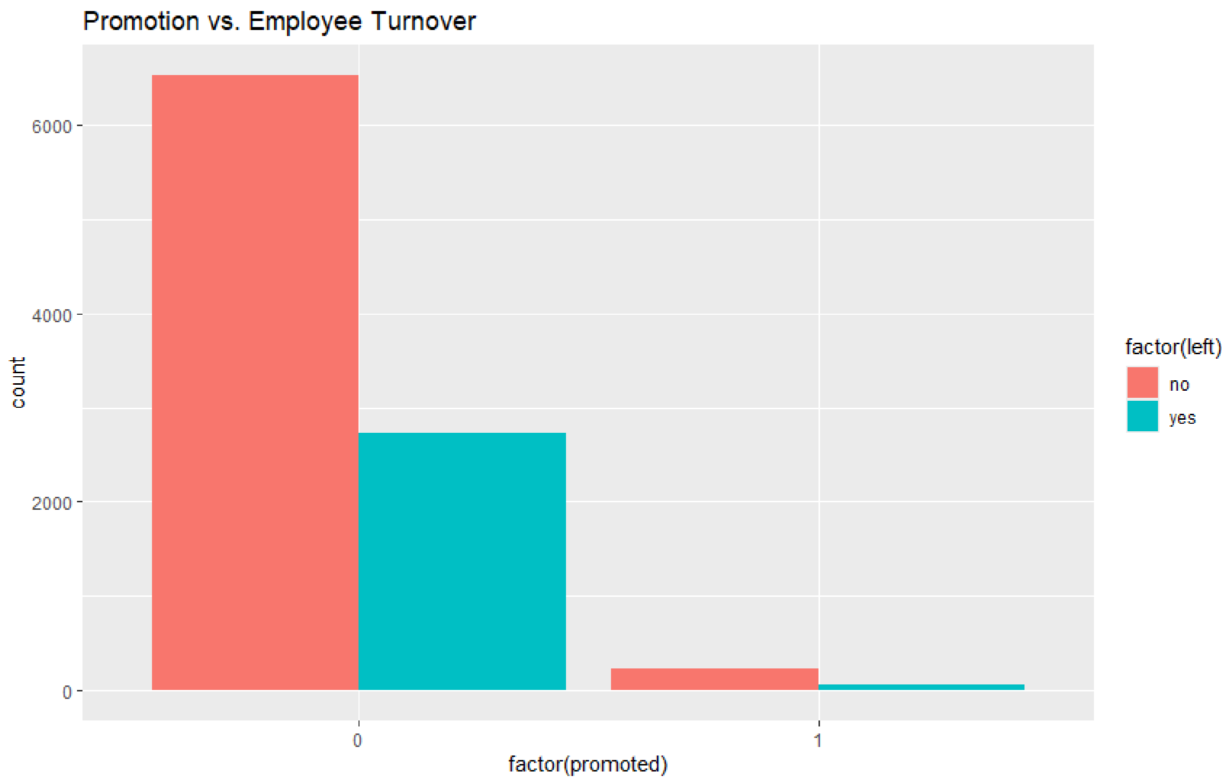
The histogram for “avg_hrs_month” demonstrates a relatively normal distribution, reflecting its summary statistics. However, it may be noticed that there is a dip around 170 average hours, followed by a spike.



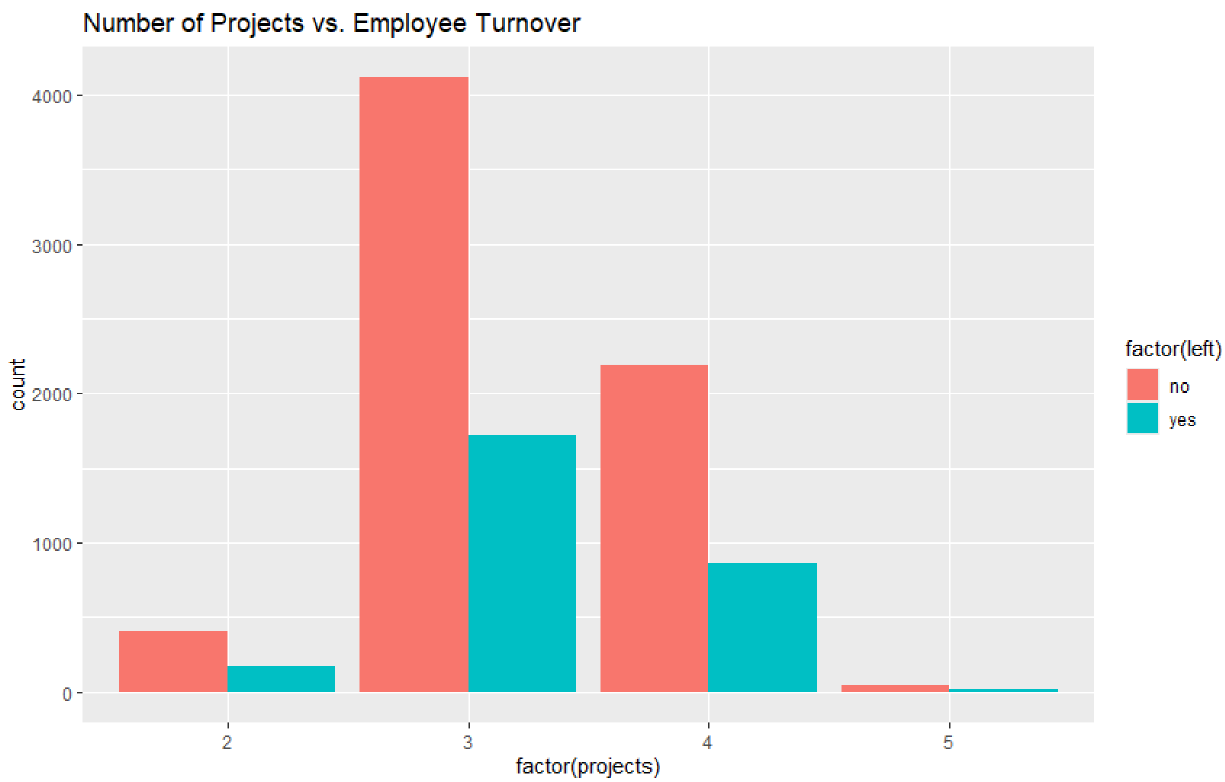
The bar chart above plots the number of employees that left and stayed. Based on the bar chart, it may be seen that there are significantly more employees that left the company compared to the employers who left.



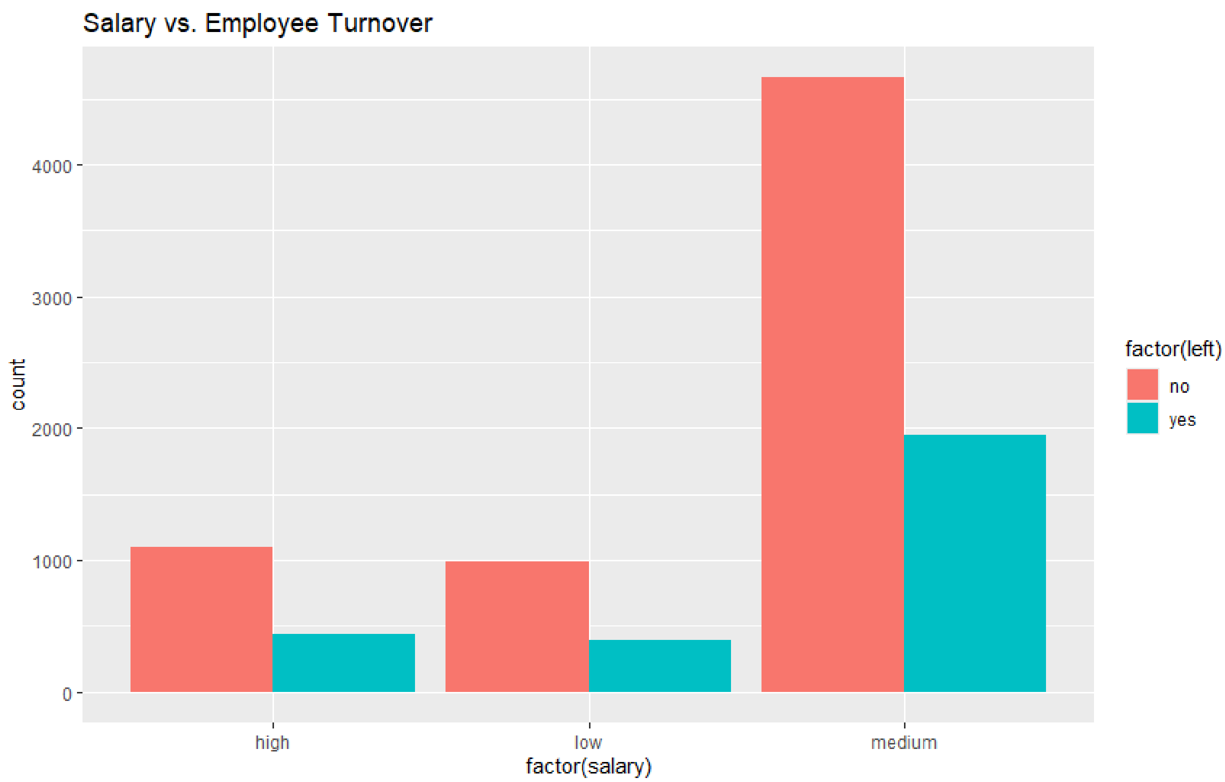
The bar chart above plots the number of employee turnover based on department. Based on the bar chart, it may be observed that the majority of employees who stayed are in the sales department, followed by the operations, engineering, and retail department. Meanwhile, the majority of employees who left are from the sales department, followed by the retail, engineering, and operations department.



The bar chart above plots the employee turnover based on whether they were promoted or not. Based on this bar chart, it may be seen that the majority of employees that left were not given a promotion. Meanwhile, the majority of employees who did not leave were also not given a promotion.

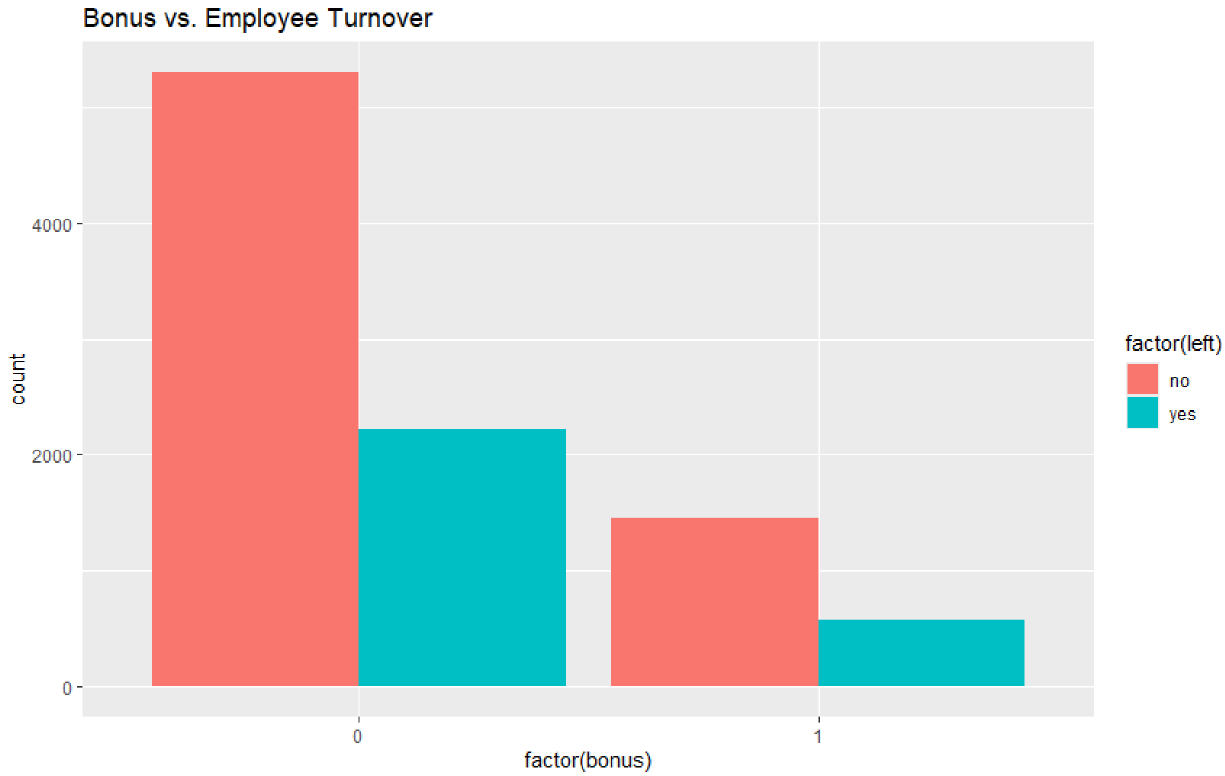


The bar chart above plots the employee turnover based on how many projects the employee has. It may be observed that the majority of employees who leave had 3 projects, while the majority who stayed also had 3 projects.



The bar chart above plots the employee turnover based on the employee’s salary range. It may be seen that the majority of employees who left had a medium-sized salary, while the majority of employees who stayed also had medium-sized salary.

The bar chart above plots the employee turnover based on the employee's tenure. Based on the bar chart, it can be seen that the majority of employees who leave had a tenure of 7 and 8 years. Meanwhile, the majority of employees who stayed had a tenure of 6 years.



Finally, the bar chart above plots the employee turnover based on whether the employee received a bonus or not. Based on the bar chart, the majority of employees who left did not receive a bonus. Meanwhile, the majority of employees who stayed also did not receive a bonus.

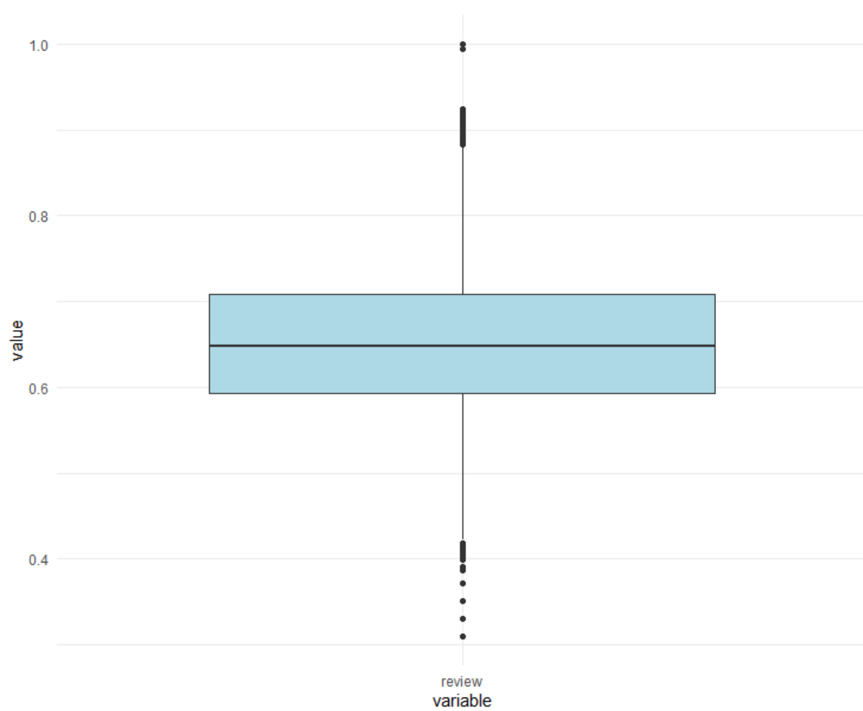
Based on the observations from the bar charts give a general understanding of how each variable affects employee turnover. However, it should be noted that these variables may not be directly related to the cause of employee turnover. For example, focusing on the salary, the medium category contains both the majority of employees who left and stayed. This does not directly imply that the majority of employees left because of their salary. However, it only implies] that the majority of the company's employees' salary is in this medium category. Therefore, statistically, this category will have the majority of employees who left. This also applies to other variables such as department, projects, and promotion.

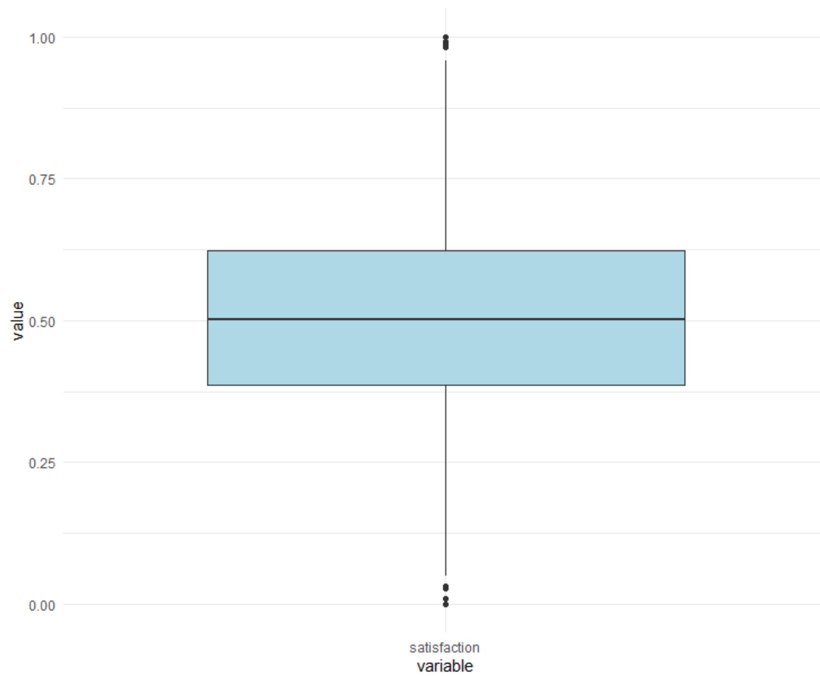
Data Pre-processing

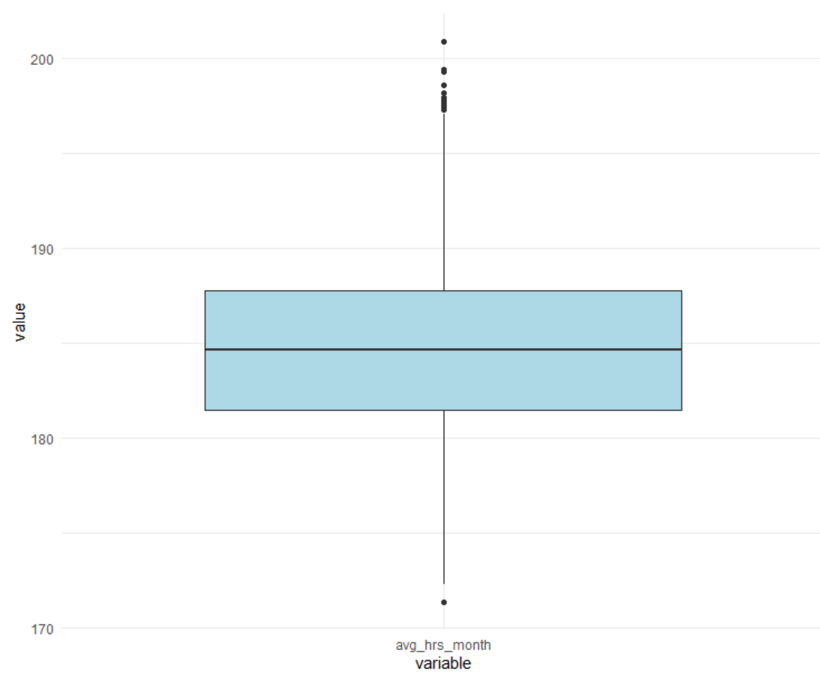
```
> # Check for missing values
> colSums(is.na(df))
  department     promoted       review     projects       salary       tenure  satisfaction        bonus avg_hrs_month
           0            0            0            0            0            0             0            0             0
        left
           0
```

```
> # Check for duplicates and handle them
> if (any(duplicated(df))) {
+   df <- df[!duplicated(df), ]
+   print("Duplicates have been removed.")
+ } else {
+   print("No duplicate rows found.")
+ }
[1] "No duplicate rows found."
```

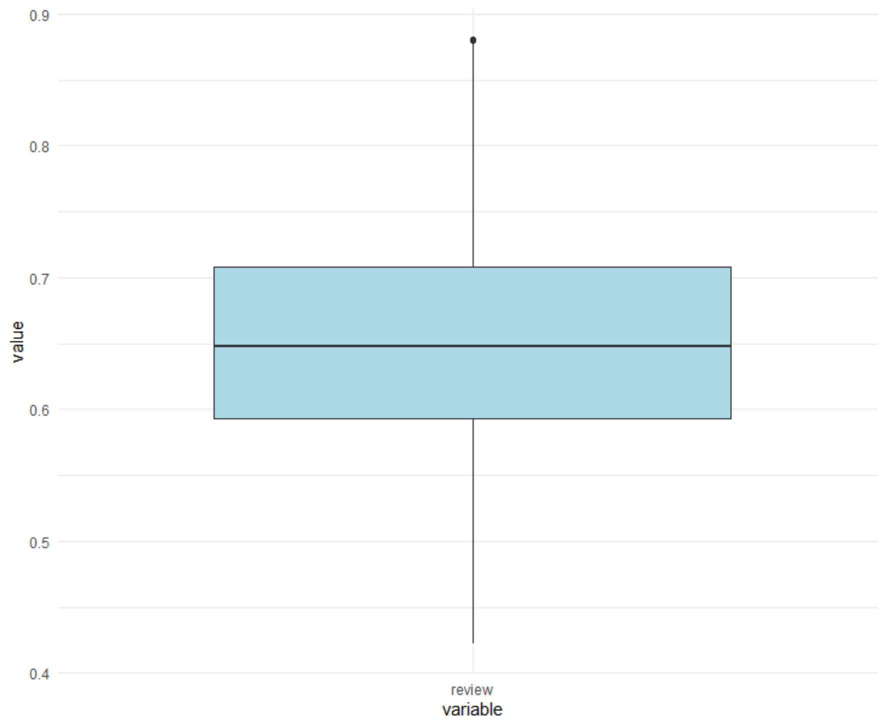
In order to begin the modelling process, the dataset must first go through pre-processing, where missing rows, duplicate rows, and outliers will be handled. As seen above, there are no missing values present in the dataset. Additionally, there are no duplicate rows found in the dataset.







The box plots have been plotted for the variables “review”, “satisfaction”, and “avg_hrs_month” as these are continuous variables. Based on these box plots, it may be noticed that the “review” column has a significant number of outliers, along with extreme values that deviate from its median. Observing the remaining box plots for “satisfaction” and “avg_hours_month”, it may be seen that there are outliers present. However, as opposed to the outliers in “review”, these outliers are reasonable as they are relatively within the range of the data. Therefore, the outliers for “review” will be removed, while the outliers for “satisfaction” and “avg_hours_month” will remain.



The figure above is an updated box plot for the “review” variable after the outliers have been removed. As seen in the box plot, the majority of outliers and extreme values have been removed.

It may be noticed that in the original dataset, the “department”, “salary”, and “left” columns are non-numerical categorical columns. Therefore, they must be converted to numerical columns. The “left” column contains the values “yes” and “no”, which is why these values in this column were converted to binary, where “0” represents “no” and “1” represents “yes”. In terms of converting the “department” and “salary” columns to numeric, one-hot encoding was performed as they both contain more than two categories. Furthermore, these columns are nominal categorical columns, as they do not have any intrinsic order or ranking. In addition to performing one-hot encoding, the reference categories for both “department” and “salary” were removed (“departmentsupport” and “salarymedium”). Finally, all binary columns were then converted to factors.

Methodology

Proposed Data

The data that will be used is from an anonymous large US company and retrieved from Kaggle. The company's HR department gathered information from about 10,000 of their employees who left the company between 2016-2020 and they used information from various events such as exit interviews, performance reviews, and employee records.

Data Dictionary

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

Data splitting

```
# Set a seed for reproducibility
set.seed(123)

# Split the data into training and testing sets (80% training, 20% testing)
trainIndex <- createDataPartition(df_cleaned$left, p = 0.8, list = FALSE)
train_data <- df_cleaned[trainIndex, ]
test_data <- df_cleaned[-trainIndex, ]
```

Once the data has been cleaned through pre-processing, the modelling process may begin. First and foremost, a seed will need to be set in order to reproduce the same results on repeat tests. Additionally, in order to evaluate the Logistic Regression model's performance, the data must first be split into training and testing sets, where 80% of the data will be used for training, while the remaining 20% will be used to evaluate the model.

Model testing

```
# Fit logistic regression model on training data
model1 <- glm(left ~ review + projects + tenure + satisfaction + bonus + avg_hrs_month +
                departmentadmin + departmentengineering + departmentfinance + departmentIT +
                departmentlogistics + departmentmarketing + departmentoperations + departmentretail +
                departmentsales + salaryhigh + salarylow,
              data = train_data,
              family = binomial)
```

After splitting the data, the Logistic Regression model can be trained, setting employee turnover ("left") as the target variable and the rest of the variables as the predictors. Once the model has been configured, a summary may be produced, which will output each coefficient's estimate, standard error, z-value, and p-value. Furthermore, the significant variables will be indicated.

In order to prevent any multicollinearity issues, the VIFs of each variable in the model will need to be inspected. If it is found that there is a multicollinearity problem present, one of the highly correlated variables will need to be removed from the model. After removing the variable, the summary may be produced once again, along with the VIFs of the updated model.

Performance metrics

```
# Predict probabilities on test data
test_data$predicted_prob <- predict(model2, newdata = test_data, type = "response")

# Convert probabilities to binary class predictions
test_data$predicted_class <- ifelse(test_data$predicted_prob > 0.5, 1, 0)

# View the confusion matrix for evaluation
confusion_matrix <- table(Predicted = test_data$predicted_class, Actual = test_data$left)
print(confusion_matrix)

# Extract values from the confusion matrix
true_positive <- confusion_matrix[2, 2] # Predicted 1, Actual 1
true_negative <- confusion_matrix[1, 1] # Predicted 0, Actual 0
false_positive <- confusion_matrix[2, 1] # Predicted 1, Actual 0
false_negative <- confusion_matrix[1, 2] # Predicted 0, Actual 1
```

Using the most updated model, the confusion matrix using the testing set may be produced. Producing a confusion matrix is a crucial step when evaluating the model's performance. The confusion matrix will display all the model's predictions, including the True Positives, True Negatives, False Positives, and False Negatives.

```
# Calculate accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
print(paste("Accuracy: ", round(accuracy, 2)))

# Calculate precision
precision <- true_positive / (true_positive + false_positive)
print(paste("Precision: ", round(precision, 2)))

# Calculate recall (same as sensitivity)
recall <- true_positive / (true_positive + false_negative)
print(paste("Recall (Sensitivity): ", round(recall, 2)))

# Calculate specificity
specificity <- true_negative / (true_negative + false_positive)
print(paste("Specificity: ", round(specificity, 2)))

# Calculate F1-score
f1_score <- 2 * ((precision * recall) / (precision + recall))
print(paste("F1-Score: ", round(f1_score, 2)))

# Plot ROC curve for model performance
roc_curve <- roc(test_data$left, test_data$predicted_prob)
plot(roc_curve, main = "ROC Curve", col = "blue")

# Calculate the AUC
auc_value <- auc(roc_curve)
print(paste("AUC: ", round(auc_value, 2)))
```

Using the values produced in the confusion matrix enables the computation of various performance metrics.

Accuracy measures the proportion of correct predictions (TP and TN) out of all predictions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision measures the proportion of correctly predicted positive instances out of all positive predictions.

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the proportion of correctly predicted positive instances out of all actual positive instances.

$$Recall = \frac{TP}{TP + FN}$$

Specificity measures the proportion of correctly predicted negative instances out of all actual negative instances.

$$Specificity = \frac{TN}{TN + FP}$$

F1-Score is the harmonic mean of precision and recall.

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

Alongside the confusion matrix, the Receiver Operating Characteristic (ROC) curve will also be produced, which graphically represents the trade-off between the model’s specificity and 1-specificity (False Positive Rate) at various threshold values. Its Area Under Curve (AUC) value quantifies the model’s performance in distinguishing between positive and negative classes. An ideal model will have an AUC value close to 1, indicating perfect distinguishing capabilities. Meanwhile, an AUC of 0.5 indicates a random classifier.

Analysis & Results

```
> summary(model1)

Call:
glm(formula = left ~ review + projects + tenure + satisfaction +
    bonus + avg_hrs_month + departmentadmin + departmentengineering +
    departmentfinance + departmentIT + departmentlogistics +
    departmentmarketing + departmentoperations + departmentretail +
    departmentsales + salaryhigh + salarylow, family = binomial,
    data = train_data)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -20.779179   5.318069  -3.907 9.33e-05 ***
review                  11.226230   0.411866  27.257  < 2e-16 ***
projects                -0.098395   0.046614  -2.111   0.0348 *  
tenure                   0.004887   0.092688   0.053   0.9580    
satisfaction             2.528408   0.210079  12.036  < 2e-16 ***
bonus1                  -0.088263   0.066801  -1.321   0.1864    
avg_hrs_month            0.062339   0.031772   1.962   0.0498 *  
departmentadmin1        -0.120659   0.159740  -0.755   0.4500    
departmentengineering1  -0.070567   0.117008  -0.603   0.5464    
departmentfinance1      -0.178109   0.164441  -1.083   0.2788    
departmentIT1            0.121019   0.167295   0.723   0.4694    
departmentlogistics1     0.125186   0.165701   0.755   0.4500    
departmentmarketing1    -0.039954   0.133136  -0.300   0.7641    
departmentoperations1   -0.094608   0.116929  -0.809   0.4185    
departmentretail1        0.043420   0.116507   0.373   0.7094    
departmentsales1        -0.050513   0.112926  -0.447   0.6547    
salaryhigh1             -0.030765   0.074247  -0.414   0.6786    
salarylow1              -0.064678   0.077804  -0.831   0.4058    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9137.3  on 7587  degrees of freedom
Residual deviance: 8218.7  on 7570  degrees of freedom
AIC: 8254.7

Number of Fisher Scoring iterations: 4
```

Model 1 utilizes all variables to predict employee turnover using Logistic Regression. The summary above displays the coefficient values for each predictor after training the Logistic Regression model utilizing all variables. As observed by the results, there are two significant variables that have a p-value of less than 0.001, which are “review” and “satisfaction”. Additionally, there are also two significant variables that have a p-value of less than 0.05, which are “projects” and “avg_hrs_month”. This implies that these variables contribute the most to the classification of employee turnover when utilizing all variables in the model.

```
> # VIF for multicollinearity
> vif(model1)
               review              projects                tenure          satisfaction                 bonus
             1.391331              1.002496             24.765913              1.342528              1.000941
         avg_hrs_month       departmentadmin departmentengineering     departmentfinance          departmentIT
            24.929399              1.484557              2.497068              1.445974              1.423614
  departmentlogistics   departmentmarketing  departmentoperations      departmentretail       departmentsales
             1.435627              1.877278              2.502404              2.525611              2.781328
           salaryhigh             salarylow
             1.035775              1.036797
```

After viewing the summary of model 1, the VIFs for each variable must be inspected. Based on the values above, it may be noticed that the VIFs for “tenure” and “avg_hrs_month” are relatively high compared to the rest of the variables, 24.76 and 24.92 respectively. This indicates that there is a multicollinearity issue when utilizing all variables, which is that the “tenure” and “avg_hrs_month” are highly correlated. This makes sense as both variables are related to time, indicating how long an employee works. In order to combat this multicollinearity issue, one of these variables must be removed from the model. In this case, the “tenure” variable was removed.

```
> summary(model2)

Call:
glm(formula = left ~ review + projects + satisfaction + bonus +
    avg_hrs_month + departmentadmin + departmentengineering +
    departmentfinance + departmentIT + departmentlogistics +
    departmentmarketing + departmentoperations + departmentretail +
    departmentsales + salaryhigh + salarylow, family = binomial,
    data = train_data)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)           -21.049682   1.399982 -15.036   <2e-16 ***
review                 11.226827   0.411711  27.269   <2e-16 ***
projects               -0.098374   0.046613  -2.110   0.0348 *
satisfaction            2.528222   0.210046  12.037   <2e-16 ***
bonus1                 -0.088272   0.066801  -1.321   0.1864
avg_hrs_month           0.063976   0.006743   9.488   <2e-16 ***
departmentadmin1       -0.120822   0.159709  -0.757   0.4493
departmentengineering1 -0.070670   0.116991  -0.604   0.5458
departmentfinance1     -0.178355   0.164377  -1.085   0.2779
departmentIT1           0.120905   0.167281   0.723   0.4698
departmentlogistics1    0.125114   0.165695   0.755   0.4502
departmentmarketing1   -0.040013   0.133131  -0.301   0.7638
departmentoperations1  -0.094690   0.116918  -0.810   0.4180
departmentretail1       0.043395   0.116505   0.372   0.7095
departmentsales1       -0.050644   0.112897  -0.449   0.6537
salaryhigh1            -0.030820   0.074239  -0.415   0.6780
salarylow1             -0.064685   0.077804  -0.831   0.4058
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9137.3  on 7587  degrees of freedom
Residual deviance: 8218.7  on 7571  degrees of freedom
AIC: 8252.7

Number of Fisher Scoring iterations: 4
```

Model 2 is similar to model 1, except “tenure” was removed as a predictor. The summary above displays the coefficient values after training model 2. Looking at the results, it may be noticed that after removing “tenure” as a predictor, there are now three significant variables that have a p-value of less than 0.001, which are “review”, “satisfaction”, and “avg_hrs_month”. Furthermore, there is only one significant variable that has a p-value of less than 0.05, which is “projects”. This implies that the most significant variables that contribute to model 2’s classification of employee turnover are “review”, “satisfaction”, and “avg_hrs_month”, with “projects” being a close contender.

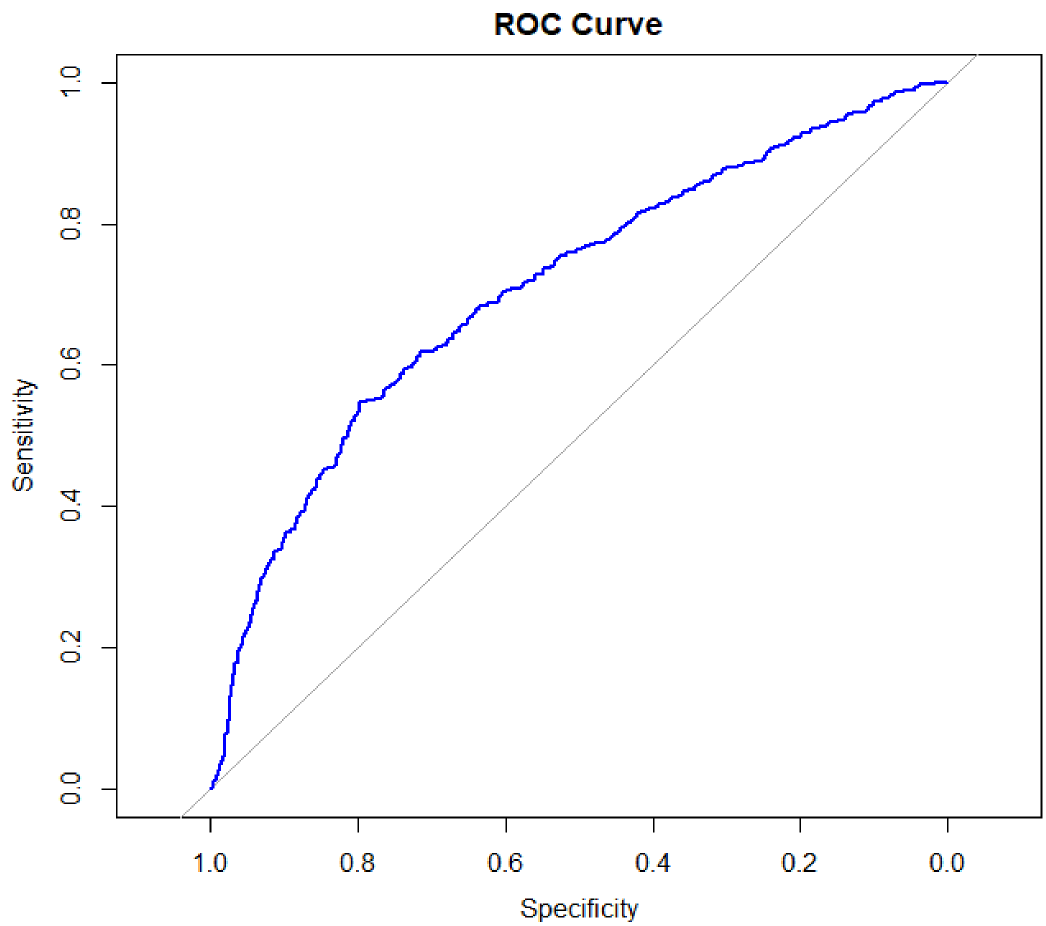
```
> # VIF for multicollinearity
> vif(model2)
               review              projects          satisfaction                 bonus         avg_hrs_month
             1.390287              1.002421              1.342125              1.000934              1.122853
      departmentadmin departmentengineering     departmentfinance          departmentIT   departmentlogistics
             1.483999              2.496361              1.444788              1.423366              1.435518
  departmentmarketing  departmentoperations      departmentretail       departmentsales            salaryhigh
             1.877120              2.501914              2.525554              2.779931              1.035569
            salarylow
             1.036794
```

After viewing the summary of model 2, the VIFs must be inspected once again to verify if there are any multicollinearity issues. As seen by the VIF values, there is no multicollinearity issue after removing the “tenure” variable, as all the VIFs are relatively low.

```
> print(confusion_matrix)
         Actual
Predicted    0    1
        0 1275  420
        1   72  129
```

The confusion matrix above displays the predictions made by the Logistic Regression model using the testing-set. It may be seen that its True Negative value is 1275, its True Positive value is 129, its False Negative value is 72, and its False Positive value is 420. Using these values, the performance of this model may be evaluated. This Logistic Regression model has an accuracy of 0.74, implying that the model correctly predicts about 74% of cases. Its precision is 0.64, meaning that out of all its positive predictions, about 64% of them are correct. It has a recall of 0.23, which means that the model is able to identify about 23% of actual positive cases. Its specificity is 0.95, meaning that the model is about to correctly identify about 95% of actual negative cases. Finally, its F1-Score is 0.34, indicating the balance between the model’s precision and recall. Based on these performance metrics, it can be concluded that the model performs moderately well due to its 74% accuracy. Its strongest point is its high specificity, implying that this model works exceptionally well in identifying negative classes. Where this model starts to fall off is its recall and F1-score. The model’s low recall indicates that it tends to miss class a significant number of actual positives. This low recall causes a large drop in its F1-score, which indicates that its balance between precision and recall is suboptimal.

-



By observing the ROC curve above, it can be seen that it is above the diagonal, indicating that the model performs better than random guessing. Furthermore, it has an AUC of 0.71, indicating that it has moderate discriminatory power. With an AUC of 0.71, it also implies that this model is able to distinguish between classes about 71% of the time. Overall, although its AUC suggests moderate performance, this model may struggle to distinguish between positive and negative classes.

Discussion

Based on the model testing, Logistic Regression is an appropriate model for predicting employee turnover due to its moderately acceptable performance metrics. Furthermore, using Logistic Regression, the significant variables were able to be identified. The most significant variables that affect the model's prediction for employee turnover, variables that have a p-value of less than 0.001, include “review”, “satisfaction”, and “avg_hrs_month”. Additionally, one variable that affects the model's prediction for employee turnover to a lesser degree, has a p-value less than 0.05, is “project”.

Based on the tested model, it may be implied that an employee's review score is the most important factor that affects the prediction of employee turnover due to its high positive coefficient value. Its high coefficient implies that as an employee's review score increases, the employee is likely to leave. Although a high review score may suggest good performance from that employee, it may also suggest that the high-performing employees are unhappy with their current experience, such as lack of opportunity, resulting in them searching for a better offer somewhere else. Since a high review score may result in higher employee turnover, businesses should implement various measures in order to retain these employees such as providing bonuses based on their performance or offering more opportunities in career development.

Satisfaction is another important factor that significantly affects the prediction of employee turnover. This implies that as an employee's satisfaction increases, the likelihood of them leaving increases. Generally speaking, an employee with a higher level of satisfaction is less likely to leave. However, satisfied employees may still want to leave due to being overworked or they seek better opportunities that the company may not be able to offer. In order to retain employees, businesses should improve their job roles in a sense where employees are able to grow and gain recognition.

Employees' average hours worked per month is a factor that significantly affects the prediction of employee turnover due to its positive, although relatively low, coefficient value. This implies that as an employee's average hours worked per month increases, the likelihood of the employee leaving is higher. In general, employees who work a high number of hours per month may feel burnt out, which will increase the likelihood of them leaving. Businesses should aim to reduce the employees' average working hours per month by improving their work-life balance by implementing flexi-hours or hybrid working. Additionally, businesses may want to offer employees various wellness programs to help them manage their stress.

Finally, the number of projects an employee has significantly affects the prediction of employee turnover. In this case, based on its negative coefficient value, an employee that is working on less projects is more likely to leave. This is because an employee who is not engaged with meaningful projects may feel undervalued, which may lead them into finding better opportunities elsewhere, where they may feel valued. Businesses should focus on assigning employees to more meaningful projects where their skill sets may be used to their full potential. In turn, these employees will feel valued, and their likelihood of leaving is decreased.

Conclusion

In conclusion, Logistic Regression is a suitable method in predicting employee turnover using various predictors. Furthermore, using Logistic Regression, the significant factors that determine whether an employee will leave or not were able to be identified, providing key insights on what aspects they should focus on to retain employees in the future. Although Logistic Regression produces acceptable performance metrics, there are areas that can be improved on, such as hyperparameter tuning to improve results. Furthermore, businesses should also explore other classification models such as Random Forest, XGBoost, or Gradient Boosting Machine to predict employee turnover, as these models are far more complex than Logistic Regression and may yield better results.

References

Ampomah, P. and Cudjor, S.K., 2015. The effect of employee turnover on organizations (case study of electricity company of Ghana, Cape Coast). Asian journal of Social Sciences and Management studies, 2(1), pp.21-24.

Dwesini, N.F., 2019. Causes and prevention of high employee turnover within the hospitality industry: A literature review. African Journal of Hospitality, Tourism and Leisure, 8(3), pp.1–15.

Masood, R.Z., 2024. Strategies for employee retention in high turnover sectors: An empirical investigation. International Journal of Research in Human Resource Management, 6(1), pp.33–41. doi:10.33545/26633213.2024.v6.i1a.167.

Ongori, H., 2007. A review of the literature on employee turnover.

Ponnuru, S.A., Merugumala, G., Padigala, S., Vanga, R., & Kantapalli, B., 2020. Employee attrition prediction using logistic regression. *International Journal of Research in Applied Sciences, Engineering and Technology*, 8(5), pp. 2871-2875.

Sutherland, J., 2002. Job-to-job turnover and job-to-non-employment movement. Personnel Review, 31(6), pp.710–721. doi:10.1108/00483480210445980.

Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., & Zhu, X., 2019. Employee turnover prediction with machine learning: A reliable approach. In: K. Arai, S. Kapoor, & R. Bhatia, eds. *Intelligent Systems and Applications*. IntelliSys 2018. Advances in Intelligent Systems and Computing, vol. 869. Cham: Springer, pp. 913-922. https://doi.org/10.1007/978-3-030-01057-7_56.