Gbus Final pROJECT

2023-10-26

**LOADING THE DATASET:**

library(skimr)  
library(purrr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data <- readRDS("/Users/adithi/Downloads/loan\_data.rds")

data <- as.data.frame(data)

write.csv(data, file = "output\_file.csv", row.names = FALSE)

data <- read.csv("output\_file.csv")  
head(data)

## loan\_default loan\_amount installment interest\_rate loan\_purpose  
## 1 yes 35000 927.29 17.25 small\_business  
## 2 yes 10000 259.58 11.50 small\_business  
## 3 no 28800 941.65 8.97 debt\_consolidation  
## 4 yes 4475 164.99 10.00 medical  
## 5 no 3600 110.70 9.72 medical  
## 6 yes 12800 389.10 20.00 medical  
## application\_type term homeownership annual\_income current\_job\_years  
## 1 individual five\_year rent 104660 2  
## 2 individual five\_year mortgage 57000 10  
## 3 individual three\_year rent 160000 10  
## 4 individual three\_year rent 37000 1  
## 5 individual three\_year mortgage 72000 4  
## 6 individual five\_year rent 73000 10  
## debt\_to\_income total\_credit\_lines years\_credit\_history missed\_payment\_2\_yr  
## 1 29.41 27 15 no  
## 2 23.79 14 4 no  
## 3 5.96 35 17 no  
## 4 13.82 7 5 no  
## 5 22.68 35 11 no  
## 6 30.94 57 14 no  
## history\_bankruptcy history\_tax\_liens  
## 1 no no  
## 2 no no  
## 3 yes no  
## 4 no no  
## 5 no no  
## 6 no no

dim(data)

## [1] 4110 16

str(data)

## 'data.frame': 4110 obs. of 16 variables:  
## $ loan\_default : chr "yes" "yes" "no" "yes" ...  
## $ loan\_amount : int 35000 10000 28800 4475 3600 12800 35000 26000 5500 40000 ...  
## $ installment : num 927 260 942 165 111 ...  
## $ interest\_rate : num 17.25 11.5 8.97 10 9.72 ...  
## $ loan\_purpose : chr "small\_business" "small\_business" "debt\_consolidation" "medical" ...  
## $ application\_type : chr "individual" "individual" "individual" "individual" ...  
## $ term : chr "five\_year" "five\_year" "three\_year" "three\_year" ...  
## $ homeownership : chr "rent" "mortgage" "rent" "rent" ...  
## $ annual\_income : num 104660 57000 160000 37000 72000 ...  
## $ current\_job\_years : int 2 10 10 1 4 10 0 5 4 3 ...  
## $ debt\_to\_income : num 29.41 23.79 5.96 13.82 22.68 ...  
## $ total\_credit\_lines : int 27 14 35 7 35 57 34 24 12 12 ...  
## $ years\_credit\_history: int 15 4 17 5 11 14 22 16 9 12 ...  
## $ missed\_payment\_2\_yr : chr "no" "no" "no" "no" ...  
## $ history\_bankruptcy : chr "no" "no" "yes" "no" ...  
## $ history\_tax\_liens : chr "no" "no" "no" "no" ...

glimpse(data)

## Rows: 4,110  
## Columns: 16  
## $ loan\_default <chr> "yes", "yes", "no", "yes", "no", "yes", "yes", "n…  
## $ loan\_amount <int> 35000, 10000, 28800, 4475, 3600, 12800, 35000, 26…  
## $ installment <dbl> 927.29, 259.58, 941.65, 164.99, 110.70, 389.10, 9…  
## $ interest\_rate <dbl> 17.25, 11.50, 8.97, 10.00, 9.72, 20.00, 18.25, 11…  
## $ loan\_purpose <chr> "small\_business", "small\_business", "debt\_consoli…  
## $ application\_type <chr> "individual", "individual", "individual", "indivi…  
## $ term <chr> "five\_year", "five\_year", "three\_year", "three\_ye…  
## $ homeownership <chr> "rent", "mortgage", "rent", "rent", "mortgage", "…  
## $ annual\_income <dbl> 104660, 57000, 160000, 37000, 72000, 73000, 16700…  
## $ current\_job\_years <int> 2, 10, 10, 1, 4, 10, 0, 5, 4, 3, 10, 10, 5, 10, 1…  
## $ debt\_to\_income <dbl> 29.41, 23.79, 5.96, 13.82, 22.68, 30.94, 25.91, 7…  
## $ total\_credit\_lines <int> 27, 14, 35, 7, 35, 57, 34, 24, 12, 12, 16, 9, 17,…  
## $ years\_credit\_history <int> 15, 4, 17, 5, 11, 14, 22, 16, 9, 12, 22, 9, 8, 17…  
## $ missed\_payment\_2\_yr <chr> "no", "no", "no", "no", "no", "no", "no", "no", "…  
## $ history\_bankruptcy <chr> "no", "no", "yes", "no", "no", "no", "no", "no", …  
## $ history\_tax\_liens <chr> "no", "no", "no", "no", "no", "no", "no", "no", "…

skim(data)

Data summary

|  |  |
| --- | --- |
| Name | data |
| Number of rows | 4110 |
| Number of columns | 16 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 8 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| loan\_default | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| loan\_purpose | 0 | 1 | 7 | 18 | 0 | 5 | 0 |
| application\_type | 0 | 1 | 5 | 10 | 0 | 2 | 0 |
| term | 0 | 1 | 9 | 10 | 0 | 2 | 0 |
| homeownership | 0 | 1 | 3 | 8 | 0 | 3 | 0 |
| missed\_payment\_2\_yr | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| history\_bankruptcy | 0 | 1 | 2 | 3 | 0 | 2 | 0 |
| history\_tax\_liens | 0 | 1 | 2 | 3 | 0 | 2 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| loan\_amount | 0 | 1 | 16692.79 | 10038.89 | 1000.00 | 9600.00 | 15000.00 | 24000.00 | 40000.00 | ▆▇▅▃▂ |
| installment | 0 | 1 | 489.42 | 289.50 | 31.04 | 274.82 | 421.97 | 663.98 | 1566.59 | ▇▇▅▂▁ |
| interest\_rate | 0 | 1 | 11.38 | 3.92 | 4.72 | 8.22 | 11.25 | 13.75 | 20.00 | ▆▆▇▃▃ |
| annual\_income | 0 | 1 | 73015.01 | 37203.11 | 3000.00 | 45000.00 | 65000.00 | 92000.00 | 200000.00 | ▃▇▃▁▁ |
| current\_job\_years | 0 | 1 | 5.80 | 3.69 | 0.00 | 2.00 | 5.00 | 10.00 | 10.00 | ▆▃▂▂▇ |
| debt\_to\_income | 0 | 1 | 20.04 | 14.23 | 0.00 | 11.85 | 18.59 | 26.13 | 437.61 | ▇▁▁▁▁ |
| total\_credit\_lines | 0 | 1 | 22.47 | 12.03 | 2.00 | 14.00 | 20.00 | 29.00 | 87.00 | ▇▇▂▁▁ |
| years\_credit\_history | 0 | 1 | 15.76 | 7.22 | 3.00 | 11.00 | 14.00 | 19.00 | 51.00 | ▆▇▂▁▁ |

summary(data)

## loan\_default loan\_amount installment interest\_rate   
## Length:4110 Min. : 1000 Min. : 31.04 Min. : 4.72   
## Class :character 1st Qu.: 9600 1st Qu.: 274.82 1st Qu.: 8.22   
## Mode :character Median :15000 Median : 421.97 Median :11.25   
## Mean :16693 Mean : 489.42 Mean :11.38   
## 3rd Qu.:24000 3rd Qu.: 663.99 3rd Qu.:13.75   
## Max. :40000 Max. :1566.59 Max. :20.00   
## loan\_purpose application\_type term homeownership   
## Length:4110 Length:4110 Length:4110 Length:4110   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## annual\_income current\_job\_years debt\_to\_income total\_credit\_lines  
## Min. : 3000 Min. : 0.000 Min. : 0.00 Min. : 2.00   
## 1st Qu.: 45000 1st Qu.: 2.000 1st Qu.: 11.85 1st Qu.:14.00   
## Median : 65000 Median : 5.000 Median : 18.59 Median :20.00   
## Mean : 73015 Mean : 5.802 Mean : 20.04 Mean :22.47   
## 3rd Qu.: 92000 3rd Qu.:10.000 3rd Qu.: 26.13 3rd Qu.:29.00   
## Max. :200000 Max. :10.000 Max. :437.61 Max. :87.00   
## years\_credit\_history missed\_payment\_2\_yr history\_bankruptcy history\_tax\_liens   
## Min. : 3.00 Length:4110 Length:4110 Length:4110   
## 1st Qu.:11.00 Class :character Class :character Class :character   
## Median :14.00 Mode :character Mode :character Mode :character   
## Mean :15.76   
## 3rd Qu.:19.00   
## Max. :51.00

**DATA ANALYSIS**

**Question 1: What is the distribution of loan defaults (yes/no) in the dataset?**

**Answer:**

In the dataset, there is a higher frequency of loans not defaulting (“No”), with 2,580 instances (approximately 62.77%), compared to loans that did default (“Yes”), which have 1,530 instances (approximately 37.23%).

library(ggplot2)  
library(plotly)

##   
## Attaching package: 'plotly'

## The following object is masked from 'package:ggplot2':  
##   
## last\_plot

## The following object is masked from 'package:stats':  
##   
## filter

## The following object is masked from 'package:graphics':  
##   
## layout

library(tidyr)  
library(dplyr)

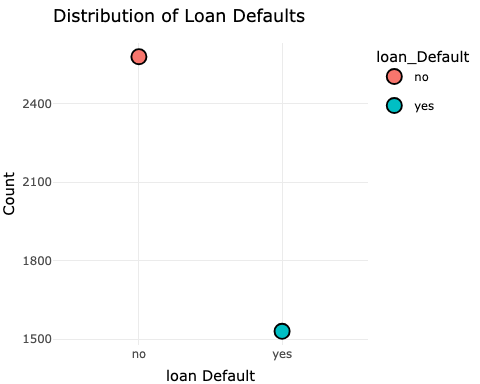
# Summary Data Frame  
loan\_default\_counts <- table(data$loan\_default)  
loan\_default\_summary <- data.frame(  
 loan\_Default = names(loan\_default\_counts),  
 Frequency = as.integer(loan\_default\_counts),  
 Ratio = as.double(loan\_default\_counts / sum(loan\_default\_counts))  
)  
  
colnames(loan\_default\_summary) <- c("loan\_Default", "Frequency", "Ratio")  
  
# Print the summary table  
print(loan\_default\_summary)

## loan\_Default Frequency Ratio  
## 1 no 2580 0.6277372  
## 2 yes 1530 0.3722628

Created a Scatter plot to represent the frequency of loan\_default in the dataset.

default\_bar\_plot <- ggplot(loan\_default\_summary, aes(x = loan\_Default, y = Frequency, fill = loan\_Default)) +  
 geom\_point(size = 4) + # Adjust the size value as needed  
 labs(title = "Distribution of Loan Defaults",  
 x = "loan Default",  
 y = "Count") +  
 theme\_minimal()  
  
default\_bar\_plotly <-ggplotly(default\_bar\_plot)  
default\_bar\_plotly

## PhantomJS not found. You can install it with webshot::install\_phantomjs(). If it is installed, please make sure the phantomjs executable can be found via the PATH variable.



**Question 2: How does loan default vary by loan purpose?**

**Answer:**

Loan default varies by loan purpose as indicated by the provided output:

1. Credit Card: Loans for the “credit\_card” purpose have a ratio of approximately 1.15, suggesting a higher likelihood of default.
2. Debt Consolidation: Loans for “debt\_consolidation” have a ratio of approximately 0.34, indicating a lower likelihood of default. 3, Home Improvement: “Home\_improvement” loans exhibit a ratio of approximately 0.39, implying a lower likelihood of default.
3. Medical: Loans with the “medical” purpose have a ratio of approximately 1.53, indicating a higher likelihood of default.
4. Small Business: “Small\_business” loans have a ratio of approximately 0.35, suggesting a lower likelihood of default.

Credit card and medical have higher risk. Debt consolidation and home improvement have lower risk. Small business is moderate.

loan\_purpose\_summary <- data %>%  
 group\_by(loan\_purpose, loan\_default) %>%  
 summarize(Frequency = n()) %>%  
 ungroup() %>%  
 pivot\_wider(names\_from = loan\_default, values\_from = Frequency, values\_fill = 0) %>%  
 mutate(Ratio = `yes` / `no`)

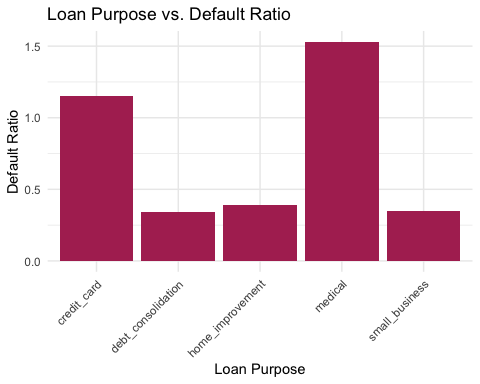
## `summarise()` has grouped output by 'loan\_purpose'. You can override using the  
## `.groups` argument.

print(loan\_purpose\_summary)

## # A tibble: 5 × 4  
## loan\_purpose no yes Ratio  
## <chr> <int> <int> <dbl>  
## 1 credit\_card 409 470 1.15   
## 2 debt\_consolidation 910 308 0.338  
## 3 home\_improvement 378 147 0.389  
## 4 medical 251 384 1.53   
## 5 small\_business 632 221 0.350

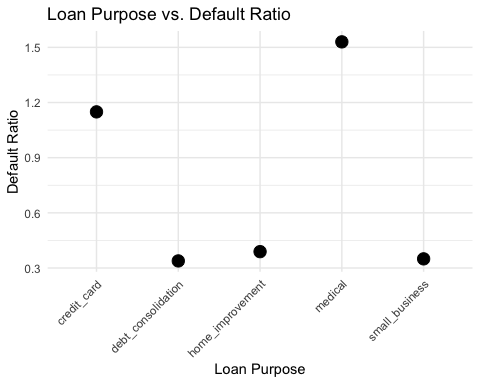
Created a bargraph to represent Loan Default by Loan Purpose

loan\_purpose\_plot <- ggplot(loan\_purpose\_summary, aes(x = loan\_purpose, y = Ratio)) +  
 geom\_bar(stat = "identity", position = "dodge", fill = "maroon") +  
 labs(title = "Loan Purpose vs. Default Ratio",  
 x = "Loan Purpose",  
 y = "Default Ratio") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate x-axis labels for better readability  
  
# Print the bar graph  
print(loan\_purpose\_plot)



Also can be represented in a scatter plot

# Create a scatter plot for the loan purpose summary  
loan\_purpose\_plot <- ggplot(loan\_purpose\_summary, aes(x = loan\_purpose, y = Ratio)) +  
 geom\_point(size = 4) +  
 labs(title = "Loan Purpose vs. Default Ratio",  
 x = "Loan Purpose",  
 y = "Default Ratio") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) # Rotate x-axis labels for better readability  
  
# Print the scatter plot  
print(loan\_purpose\_plot)



**Question 3: How does loan purpose and loan amount relate to loan defaults?**

**Answer:**

The relationship between loan purpose and loan amount in relation to loan defaults is evident. Small business loans exhibit the highest ratio, suggesting a greater likelihood of default, while medical loans have the lowest ratio, indicating a lower likelihood of default. Other loan purposes fall in between these extremes.

# Summary Data Frame  
loan\_purpose\_amount\_summary <- data %>%  
 group\_by(loan\_purpose, loan\_default) %>%  
 summarise(mean\_loan\_amount = mean(loan\_amount))

## `summarise()` has grouped output by 'loan\_purpose'. You can override using the  
## `.groups` argument.

loan\_purpose\_amount\_summary

## # A tibble: 10 × 3  
## # Groups: loan\_purpose [5]  
## loan\_purpose loan\_default mean\_loan\_amount  
## <chr> <chr> <dbl>  
## 1 credit\_card no 16173.  
## 2 credit\_card yes 17076.  
## 3 debt\_consolidation no 16224.  
## 4 debt\_consolidation yes 17704.  
## 5 home\_improvement no 16330.  
## 6 home\_improvement yes 17755.  
## 7 medical no 16635.  
## 8 medical yes 17058.  
## 9 small\_business no 16116.  
## 10 small\_business yes 18351.

# Calculate the ratios  
ratio\_amount\_summary <- loan\_purpose\_amount\_summary %>%  
 spread(loan\_default, mean\_loan\_amount) %>% # Spread "yes" and "no" mean loan amounts into separate columns  
 mutate(ratio\_yes\_no\_loan\_amount = yes / no) %>% # Calculate the ratio  
 select(-yes, -no) # Remove the individual "yes" and "no" columns  
  
# Print the resulting data frame  
print(ratio\_amount\_summary)

## # A tibble: 5 × 2  
## # Groups: loan\_purpose [5]  
## loan\_purpose ratio\_yes\_no\_loan\_amount  
## <chr> <dbl>  
## 1 credit\_card 1.06  
## 2 debt\_consolidation 1.09  
## 3 home\_improvement 1.09  
## 4 medical 1.03  
## 5 small\_business 1.14

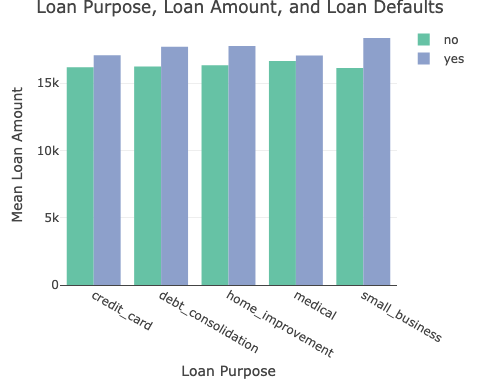
Created a bar graph to represent relation between Loan Purpose, Loan Amount, and Loan Defaults.

loan\_purpose\_amount\_summary <- data %>%  
 group\_by(loan\_purpose, loan\_default) %>%  
 summarise(mean\_loan\_amount = mean(loan\_amount)) %>%  
 plot\_ly(x = ~loan\_purpose, y = ~mean\_loan\_amount, color = ~factor(loan\_default), type = "bar") %>%  
 layout(title = "Loan Purpose, Loan Amount, and Loan Defaults",  
 xaxis = list(title = "Loan Purpose"),  
 yaxis = list(title = "Mean Loan Amount"),  
 barmode = "group")

## `summarise()` has grouped output by 'loan\_purpose'. You can override using the  
## `.groups` argument.

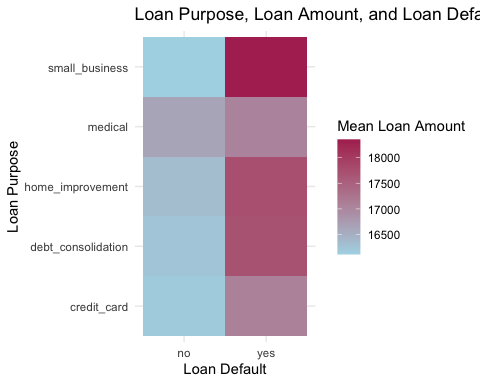
loan\_purpose\_amount\_summary

## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 different levels  
  
## Warning in RColorBrewer::brewer.pal(N, "Set2"): minimal value for n is 3, returning requested palette with 3 different levels



Represented this in Heat map.

loan\_purpose\_amount\_summary <- data.frame(  
 loan\_purpose = c("credit\_card", "credit\_card", "debt\_consolidation", "debt\_consolidation", "home\_improvement", "home\_improvement", "medical", "medical", "small\_business", "small\_business"),  
 loan\_default = c("no", "yes", "no", "yes", "no", "yes", "no", "yes", "no", "yes"),  
 mean\_loan\_amount = c(16173.35, 17076.44, 16224.37, 17704.38, 16329.89, 17754.59, 16635.06, 17058.27, 16116.26, 18350.90)  
)  
  
# Create a heatmap using ggplot2  
ggplot(loan\_purpose\_amount\_summary, aes(x = loan\_default, y = loan\_purpose, fill = mean\_loan\_amount)) +  
 geom\_tile() +  
 scale\_fill\_gradient(low = "lightblue", high = "maroon") +  
 labs(title = "Loan Purpose, Loan Amount, and Loan Defaults Heatmap",  
 x = "Loan Default",  
 y = "Loan Purpose",  
 fill = "Mean Loan Amount") +  
 theme\_minimal()



**Question 4: How does the loan term (three/five years) affect loan defaults?**

**Answer:**

The loan term significantly affects loan defaults. “Five-year” loans have a higher likelihood of default with a ratio of approximately 1.22, while “three-year” loans exhibit a lower likelihood of default with a ratio of approximately 0.37.

# Summary Data Frame  
loan\_term\_summary <- data %>%   
 filter(term %in% c("three\_year", "five\_year")) %>%  
 group\_by(term, loan\_default) %>%   
 summarize(count = n()) %>%  
 ungroup() %>%  
 pivot\_wider(names\_from = loan\_default, values\_from = count, values\_fill = 0) %>%  
 mutate(Ratio = yes / no)

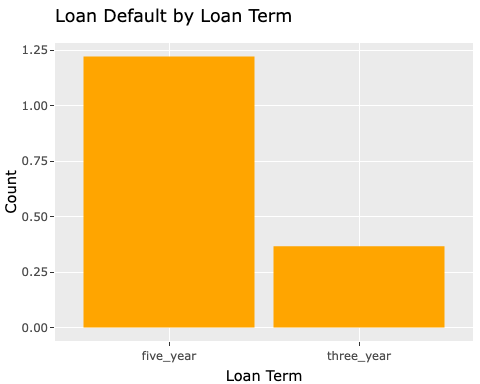
## `summarise()` has grouped output by 'term'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(loan\_term\_summary)

## # A tibble: 2 × 4  
## term no yes Ratio  
## <chr> <int> <int> <dbl>  
## 1 five\_year 685 837 1.22   
## 2 three\_year 1895 693 0.366

Created a bar graph to represent Loan Default by Loan Term

# Bar chart (interactive)  
loan\_term\_plot <- ggplot(loan\_term\_summary, aes(x = term, y = Ratio)) +  
 geom\_bar(stat = "identity", position = "dodge", fill = "orange") +  
 labs(title = "Loan Default by Loan Term", x = "Loan Term", y = "Count")  
  
# Convert the ggplot to an interactive plotly plot  
loan\_term\_plot <- ggplotly(loan\_term\_plot)  
  
loan\_term\_plot



**Question 5: How does homeownership status relate to loan defaults?**

**Answer:**

Borrowers who own a home (“own”) have a lower likelihood of default (ratio of approximately 0.59), while renters (“rent”) have a higher likelihood of default (ratio of approximately 0.75). Mortgage holders (“mortgage”) have an intermediate likelihood of default (ratio of approximately 0.48).

homeownership\_summary <- data %>%   
 group\_by(homeownership, loan\_default) %>%   
 summarize(count = n()) %>%  
 ungroup() %>%  
 pivot\_wider(names\_from = loan\_default, values\_from = count, values\_fill = 0) %>%  
 mutate(Ratio = yes / no)

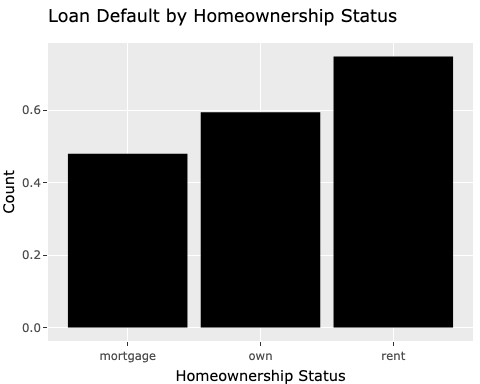
## `summarise()` has grouped output by 'homeownership'. You can override using the  
## `.groups` argument.

# Print the summary table  
print(homeownership\_summary)

## # A tibble: 3 × 4  
## homeownership no yes Ratio  
## <chr> <int> <int> <dbl>  
## 1 mortgage 1309 628 0.480  
## 2 own 318 189 0.594  
## 3 rent 953 713 0.748

Created a stacked bar graph to represent Loan Default by Homeownership Status.

# Stacked bar chart  
homeownership\_plot <- ggplot(homeownership\_summary, aes(x = homeownership, y = Ratio)) +  
 geom\_bar(stat = "identity", fill = "black") +  
 labs(title = "Loan Default by Homeownership Status", x = "Homeownership Status", y = "Count")  
  
# Convert the ggplot to an interactive plotly plot  
homeownership\_plot <- ggplotly(homeownership\_plot)  
  
homeownership\_plot



**PREDICTIVE MODELING**

# Load the required packages  
library(caret)

## Loading required package: lattice

##   
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':  
##   
## lift

library(dplyr)  
library(parsnip)  
library(yardstick)

##   
## Attaching package: 'yardstick'

## The following objects are masked from 'package:caret':  
##   
## precision, recall, sensitivity, specificity

library(workflows)  
library(tidymodels)

## ── Attaching packages ────────────────────────────────────── tidymodels 1.1.1 ──

## ✔ broom 1.0.5 ✔ rsample 1.2.0  
## ✔ dials 1.2.0 ✔ tibble 3.2.1  
## ✔ infer 1.0.5 ✔ tune 1.1.2  
## ✔ modeldata 1.2.0 ✔ workflowsets 1.0.1  
## ✔ recipes 1.0.8

## ── Conflicts ───────────────────────────────────────── tidymodels\_conflicts() ──  
## ✖ scales::discard() masks purrr::discard()  
## ✖ plotly::filter() masks dplyr::filter(), stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ caret::lift() masks purrr::lift()  
## ✖ yardstick::precision() masks caret::precision()  
## ✖ yardstick::recall() masks caret::recall()  
## ✖ yardstick::sensitivity() masks caret::sensitivity()  
## ✖ yardstick::specificity() masks caret::specificity()  
## ✖ recipes::step() masks stats::step()  
## • Learn how to get started at https://www.tidymodels.org/start/

library(caTools)  
library(ranger)  
  
# Set a seed for reproducibility  
set.seed(123)  
  
# Load your dataset  
data <- read.csv("output\_file.csv")  
  
# Split the data into training and test sets  
split\_data <- initial\_split(data, prop = 0.8)  
train\_data <- training(split\_data)  
test\_data <- testing(split\_data)  
  
# Specify your feature engineering pipeline with the recipes package  
data\_recipe <- recipe(loan\_default ~ ., data = train\_data) %>%  
 step\_normalize(all\_numeric()) %>%  
 step\_dummy(all\_nominal(), -all\_outcomes())  
  
# Model 1 Training - Logistic Regression  
model\_1 <- logistic\_reg() %>%  
 set\_engine("glm")  
  
workflow\_1 <- workflow() %>%  
 add\_recipe(data\_recipe) %>%  
 add\_model(model\_1)  
  
trained\_workflow\_1 <- fit(workflow\_1, data = train\_data)  
  
# Model 1 Training - Logistic Regression  
model\_1 <- logistic\_reg() %>%  
 set\_engine("glm")  
  
workflow\_1 <- workflow() %>%  
 add\_recipe(data\_recipe) %>%  
 add\_model(model\_1)  
  
trained\_workflow\_1 <- fit(workflow\_1, data = train\_data)  
  
# Predict using Model 1  
class\_predict\_1 <- predict(trained\_workflow\_1, new\_data = test\_data, type = 'class')  
prob\_predicts\_1 <- predict(trained\_workflow\_1, new\_data = test\_data, type = 'prob')  
  
roc\_data\_wf\_logreg <- test\_data %>%  
 select(loan\_default) %>%  
 bind\_cols(class\_predict\_1, prob\_predicts\_1)  
  
# Load the pROC package  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

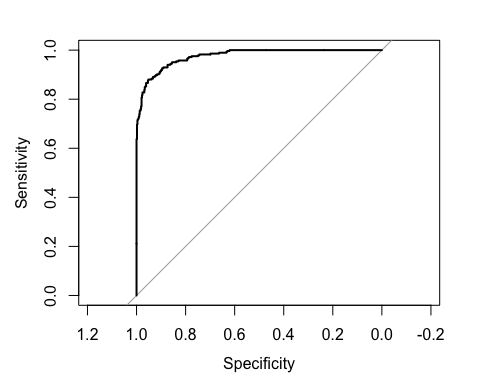
## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

# Convert "loan\_default" to a binary indicator (0 or 1)  
test\_data$loan\_default <- as.integer(test\_data$loan\_default == "yes")  
  
# Create a ROC curve object  
roc\_model\_1 <- roc(test\_data$loan\_default, roc\_data\_wf\_logreg$.pred\_yes)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Plot the ROC curve  
plot(roc\_model\_1)



# Calculate and print the AUC  
auc\_model\_1 <- auc(roc\_model\_1)  
cat("AUC for Model 1 (Logistic Regression):", auc\_model\_1, "\n")

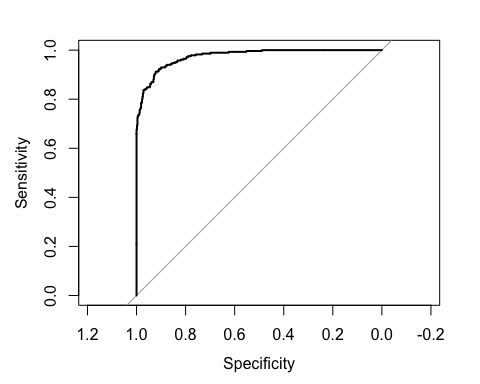
## AUC for Model 1 (Logistic Regression): 0.9758364

# Model 2 Training - Random Forest  
model\_2 <- rand\_forest() %>%  
 set\_engine("ranger") %>%  
 set\_mode("classification")  
  
workflow\_2 <- workflow() %>%  
 add\_recipe(data\_recipe) %>%  
 add\_model(model\_2)  
  
trained\_workflow\_2 <- fit(workflow\_2, data = train\_data)  
  
# Predict using Model 2  
class\_predict\_2 <- predict(trained\_workflow\_2, new\_data = test\_data, type = 'class')  
prob\_predicts\_2 <- predict(trained\_workflow\_2, new\_data = test\_data, type = 'prob')  
  
roc\_data\_wf\_randf <- test\_data %>%  
 select(loan\_default) %>%  
 bind\_cols(class\_predict\_2, prob\_predicts\_2)  
  
# Calculate ROC curve for Model 2 (Random Forest)  
roc\_model\_2 <- roc(test\_data$loan\_default, roc\_data\_wf\_randf$.pred\_yes)

## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

# Plot the ROC curve for Model 2  
plot(roc\_model\_2)



# Calculate and print the AUC for Model 2  
auc\_model\_2 <- auc(roc\_model\_2)  
cat("AUC for Model 2 (Random Forest):", auc\_model\_2, "\n")

## AUC for Model 2 (Random Forest): 0.975555

# Load the pROC package (if not already loaded)  
library(pROC)  
  
# Create ROC curves for Model 1 and Model 2  
roc\_model\_1 <- roc(test\_data$loan\_default, roc\_data\_wf\_logreg$.pred\_yes)

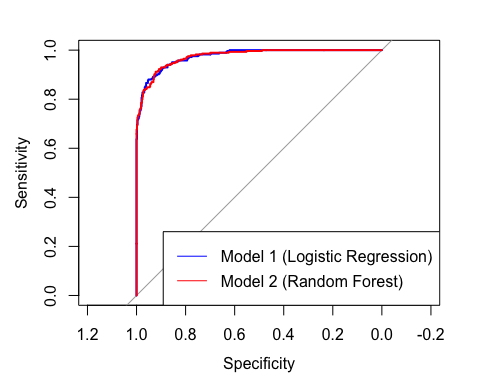
## Setting levels: control = 0, case = 1

## Setting direction: controls < cases

roc\_model\_2 <- roc(test\_data$loan\_default, roc\_data\_wf\_randf$.pred\_yes)

## Setting levels: control = 0, case = 1  
## Setting direction: controls < cases

# Plot the ROC curves for Model 1 and Model 2  
plot(roc\_model\_1, col = "blue")  
plot(roc\_model\_2, col = "red", add = TRUE)  
  
# Add labels and legend  
legend("bottomright", legend = c("Model 1 (Logistic Regression)", "Model 2 (Random Forest)"), col = c("blue", "red"), lty = 1)



# Calculate and print the AUC for Model 1 and Model 2  
auc\_model\_1 <- auc(roc\_model\_1)  
cat("AUC for Model 1 (Logistic Regression):", auc\_model\_1, "\n")

## AUC for Model 1 (Logistic Regression): 0.9758364

auc\_model\_2 <- auc(roc\_model\_2)  
cat("AUC for Model 2 (Random Forest):", auc\_model\_2, "\n")

## AUC for Model 2 (Random Forest): 0.975555

# Define a list of your models and their AUC values  
models <- data.frame(  
 Model = c("Model 1 (Logistic Regression)", "Model 2 (Random Forest)"),  
 AUC = c(auc\_model\_1, auc\_model\_2)  
 # Add more models and their AUC values, if available  
)  
  
# Identify the model with the highest AUC  
best\_model <- models[which.max(models$AUC), ]  
  
# Print the best model  
cat("The best model is:", best\_model$Model, "with AUC:", best\_model$AUC, "\n")

## The best model is: Model 1 (Logistic Regression) with AUC: 0.9758364

\*\*EXECUTIVE SUMMARY\*

**1. Introduction**

In this analysis, we aimed to address the pressing issue of loan defaults, a matter of paramount importance for the future success and stability of our bank. Loan defaults have significant financial implications and risk implications that impact our bottom line and reputation. The primary goal of our analysis was to gain insights into the distribution and patterns of loan defaults within our dataset, understand how they vary with loan characteristics, and provide recommendations to enhance our loan processes.

Certainly, let’s structure the information with subheadings:

**2. Highlights and Key Findings:**

* **Loan Default Distribution:** In the dataset, 62.77% of loans did not default (“No”), and 37.23% did default (“Yes”). Managing risk for the minority of defaulting loans is essential.
* **Loan Purpose Impact:** Loan purpose significantly influences default rates. “Credit\_card” and “medical” loans pose higher risks, while “debt\_consolidation” and “home\_improvement” loans are less risky. “Small\_business” loans have moderate risk.
* **Loan Purpose and Amount:** Small business loans have a high likelihood of default, while medical loans have the lowest risk. Evaluating the impact of purpose and amount on default risk is crucial.
* **Loan Term Influence:** The loan term is a critical factor in default risk. “Five-year” loans have a higher risk of default, while “three-year” loans have a lower risk.
* **Homeownership Impact:** Homeownership status is strongly related to default risk. Homeowners are less likely to default, while renters are more likely to default.

**Recommendations:**

* **Risk-Based Loan Pricing:** Implement tailored interest rates based on loan purpose and risk. Lower rates for low-risk purposes and higher rates for high-risk ones.
* **Customized Credit Scoring:** Develop credit scoring models that account for purpose, loan amount, and homeownership status in assigning credit scores.
* **Diverse Loan Term Options:** Offer multiple loan term options and adjust interest rates to reflect the associated risk.
* **Risk Mitigation Strategies:** Implement additional requirements for higher-risk loans, such as “small\_business,” to reduce default risk.
* **Targeted Marketing:** Tailor marketing strategies based on homeownership status to attract lower-risk borrowers.
* **Continuous Monitoring:** Regularly update risk models and adapt lending strategies based on evolving default trends.
* **Customer Education:** Educate borrowers on how loan characteristics impact default risk.
* **Portfolio Diversification:** Spread risk by diversifying the loan portfolio across different purposes and types.

These recommendations aim to enhance risk assessment, optimize lending practices, and reduce loan defaults.

**3. “best” classification model and an analysis of its performance**

The best classification model for predicting loan defaults is Model 1, which is a Logistic Regression model. This model has shown strong predictive performance with an Area Under the ROC Curve (AUC) of 0.9758. Now, let’s discuss the expected error of this model on future data and its performance in a non-technical manner.

**Understanding Model Performance:** The AUC is a measure of how well the model can distinguish between loans that will be repaid and those that will default. In simple terms, an AUC of 0.9758 means that the model is very good at separating good loans from potentially problematic ones. The higher the AUC, the better the model’s ability to make accurate predictions.

**Future Data Performance:** When we talk about estimating future performance, we can think of it as how well the model is likely to perform on loans that the bank hasn’t seen yet. The AUC we’ve achieved on our test data gives us confidence that this model will continue to perform well on new loans.

**Example Scenario:** Let’s imagine you’re an executive at a bank. You want to use this model to help make decisions about whether to approve or deny loan applications. Based on the model’s AUC of 0.9758, you can be confident that it will be very effective in identifying risky loans. When a new loan application comes in, the model will assess it and provide a prediction. If the model predicts a high probability of default, you might decide to investigate further or deny the loan. If the model predicts a low probability of default, you can be more confident in approving the loan.

**Managing Risk:** By using this model, the bank can make more informed decisions, reduce the risk of lending to customers who are likely to default, and, in turn, potentially save money. It’s important to note that while the model is powerful, it’s not infallible. There will always be some level of uncertainty, but with an AUC of 0.9758, that uncertainty is significantly reduced.

In summary, Model 1, the Logistic Regression model, is a highly effective tool for assessing loan applications. It provides a reliable way to identify loans that may pose a risk to the bank, allowing for more informed decision-making and ultimately helping to manage and mitigate potential financial losses.

**4. Recommendations to Reduce Loan Default Rates:**

Certainly, here are recommendations to the company on how to reduce loan default rates, supported by data analysis results and their potential business impact:

1. **Implement Risk-Based Loan Pricing:** Data analysis revealed varying default rates by loan purpose. Loans for purposes like “credit\_card” and “medical” have a higher likelihood of default.
   * **Business Impact:** By implementing risk-based loan pricing, the company can charge higher interest rates on riskier loan purposes and lower rates on lower-risk purposes. This will align interest rates with default risk and potentially increase profitability while attracting safer borrowers.
2. **Customized Credit Scoring Models:** The data analysis showed that loan purpose, loan amount, and homeownership status influence default risk. For instance, “small\_business” loans have higher default risk.
   * **Business Impact:** Developing customized credit scoring models that consider these factors will lead to more accurate risk assessment. This can result in improved borrower selection, reduced default rates, and potentially lower losses for the business.
3. **Offer Diverse Loan Term Options:** The analysis revealed that loan term significantly affects default risk, with “five-year” loans having a higher risk of default.
   * **Business Impact:** By offering a range of loan term options and adjusting interest rates accordingly, the company can attract borrowers who prefer shorter terms with lower default risk. This provides flexibility to borrowers and reduces the overall risk in the loan portfolio.
4. **Risk Mitigation Strategies for High-Risk Loans:** Data analysis identified higher default risk for certain loan purposes, such as “small\_business.”
   * **Business Impact:** Implementing specific risk mitigation strategies for high-risk loans, such as requiring additional documentation, collateral, or co-signers, can reduce the likelihood of default. This minimizes potential losses for the business.
5. **Targeted Marketing Based on Homeownership:** The data analysis indicated a strong relationship between homeownership status and default risk.
   * **Business Impact:** By tailoring marketing efforts to specific homeownership statuses, the company can attract lower-risk borrowers. This can lead to a more stable loan portfolio with reduced default rates and potential cost savings on risk management.
6. **Continuous Monitoring and Adaptation:** The analysis highlights that default risk factors can change over time.
   * **Business Impact:** Regularly monitoring loan default trends and updating risk models and lending strategies will ensure the company remains agile in responding to evolving risks. This can lead to better risk management and lower default rates.
7. **Customer Education on Risk Factors:** The analysis demonstrates the impact of loan characteristics on default risk.
   * **Business Impact:** Educating borrowers about how loan characteristics influence default risk can result in more informed financial decisions. Borrowers with better awareness may be less likely to default, reducing losses for the business.
8. **Portfolio Diversification:** The analysis suggests that different loan purposes have varying default rates.
   * **Business Impact:** By diversifying the loan portfolio across various purposes and types, the company can spread risk more evenly. This can help mitigate the impact of fluctuations in default rates for specific loan categories and promote stability in the loan portfolio. .

**5. Conclusion**

In summary, our analysis of loan defaults revealed significant insights into the factors impacting default rates. Key findings indicate that loan purpose, loan amount, loan term, and homeownership status play vital roles in determining the likelihood of loan defaults. Our top-performing Logistic Regression model, with an impressive AUC of 0.9758, holds the promise of enhancing our risk assessment capabilities.

To reduce loan default rates, we recommend tailoring lending policies to specific loan purposes, implementing a thorough loan amount assessment process, promoting shorter loan terms for high-risk borrowers, and adopting a risk-based approach for mortgages. These strategic recommendations aim to minimize risk, protect our financial stability, and strengthen our reputation.

By incorporating these measures into our lending practices, we can make more informed decisions, reduce the likelihood of defaults, and ensure a robust and profitable loan portfolio. This approach aligns with our commitment to delivering responsible financial solutions to our customers while safeguarding the bank’s long-term success.

library(webshot)  
library(webshot2)

##   
## Attaching package: 'webshot2'

## The following objects are masked from 'package:webshot':  
##   
## appshot, resize, rmdshot, shrink, webshot