A1 Submission

Takomborerwa Charles Mubaiwa(28080939)

2024

## Preliminary Steps

#Importing Data  
main\_data <- read.csv("A1\_data.csv", header = TRUE, stringsAsFactors = TRUE)

#Sampling Data  
set.seed(9)  
test\_index <- sample(1:1000, size = 900, replace = FALSE)  
test\_data <- main\_data[test\_index,]

# Assessment Questions

## Question 1

#Question 1.1  
print("Mean amount is: ")

## [1] "Mean amount is: "

mean(test\_data$AMOUNT)

## [1] 101719.4

#Question 1.2  
print("Median amount is: ")

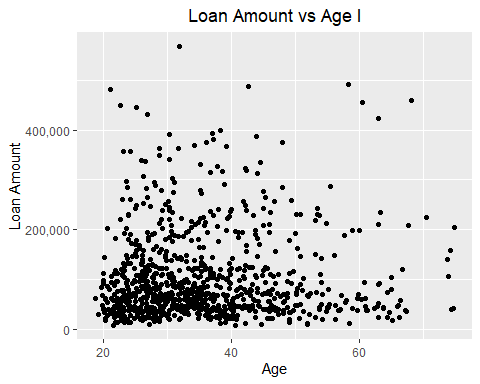
## [1] "Median amount is: "

median(test\_data$AMOUNT)

## [1] 71850

## Question 2

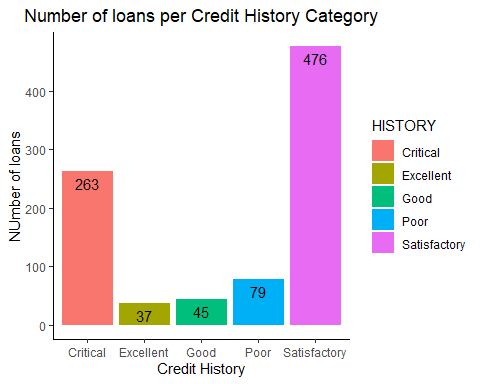
ggplot(test\_data) + geom\_point(mapping = aes(x = AGE, y = AMOUNT), position = "jitter") + scale\_y\_continuous(labels = comma) + labs(title = "Loan Amount vs Age I", y = "Loan Amount", x = "Age") + theme(plot.title = element\_text(hjust = 0.5))



There is a concentration of small loans given to clients between the age of 20 to 40, in the same range the loans above 200,000R are more spread with a high variability of larger loan and beyond age 40 loans are more evenly spread with a lower variability of larger loans.

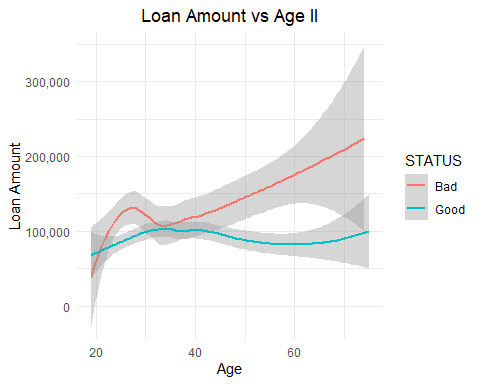
## Question 3

ggplot(test\_data, aes(x = HISTORY, fill = HISTORY)) + geom\_bar() + geom\_text(stat = "count", aes(label = after\_stat(count)), vjust = 1.5) + theme\_classic() + labs(title = "Number of loans per Credit History Category", y = "NUmber of loans", x = "Credit History") + theme(plot.title = element\_text(hjust = 0.5))



## Question 4

ggplot(test\_data, mapping = aes(x = AGE, y = AMOUNT, colour = STATUS)) + geom\_smooth(method = "loess", formula = 'y ~ x') + scale\_y\_continuous(labels = comma) + labs(title = "Loan Amount vs Age II", y = "Loan Amount", x = "Age") + theme\_minimal() + theme(plot.title = element\_text(hjust = 0.5))



Based on the graph above customers of all ages who get loans below 100,000R pay back their loans and clients who take larger loans do not and this trend only decline

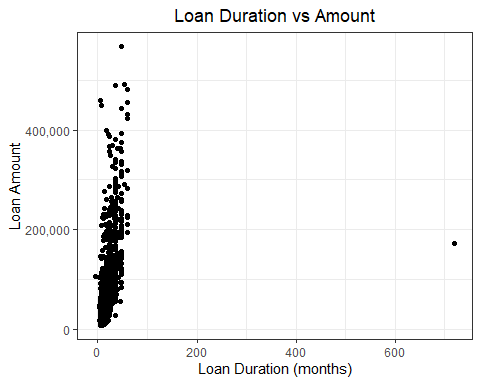
## Question 5.1

cor\_value <- cor(test\_data$DURATION, test\_data$AMOUNT)  
print(paste("The correlation between loan duration and amount is:", round(cor\_value, 3)))

## [1] "The correlation between loan duration and amount is: 0.309"

## Question 5.2

ggplot(test\_data) + geom\_point(mapping = aes(x = DURATION, y = AMOUNT), position = "jitter") + scale\_y\_continuous(label = comma) + theme\_bw() + labs(title = "Loan Duration vs Amount", x = "Loan Duration (months)", y = "Loan Amount") + theme(plot.title = element\_text(hjust = 0.5))



There are two unusual observations, one indicating a loan for 720 months and the other for a loan of negative 5 months. Loan duration cannot be negative, and the maximum loan out appears to be 60 months. The original correlation is weak due to the presence of these outliers in the data and it improves after removing them.

#Removing outliers  
new\_test\_data <- test\_data[!(test\_data$ID %in% c(431, 678)),]  
print(paste("The new correlation after removing outliers is:", round(cor(new\_test\_data$DURATION, new\_test\_data$AMOUNT), 3)))

## [1] "The new correlation after removing outliers is: 0.624"

## Question 6

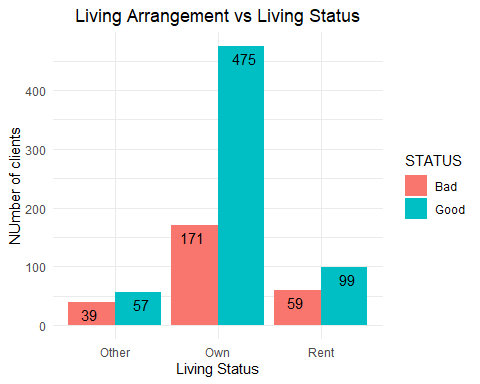
test\_with <- test\_data%>% mutate(COMB\_PURPOSE=PURPOSE)  
test\_with <- test\_with%>%mutate(COMB\_PURPOSE=if\_else(COMB\_PURPOSE == "Car - new", "Car", COMB\_PURPOSE))%>%mutate(COMB\_PURPOSE=if\_else(COMB\_PURPOSE == "Car - used", "Car", COMB\_PURPOSE))%>%mutate(COMB\_PURPOSE=if\_else(COMB\_PURPOSE == "furniture", "Household content", COMB\_PURPOSE))%>%mutate(COMB\_PURPOSE=if\_else(COMB\_PURPOSE == "TV", "Household content", COMB\_PURPOSE))%>%mutate(COMB\_PURPOSE=if\_else(COMB\_PURPOSE == "dom\_appliance", "Household content", COMB\_PURPOSE))  
purpose\_data <- test\_with%>%select(ID, AMOUNT, COMB\_PURPOSE)  
purpose\_data %>% group\_by(COMB\_PURPOSE) %>% summarize("Number of loans" = n(), "Average loan amount" = mean(AMOUNT))

## # A tibble: 6 × 3  
## COMB\_PURPOSE `Number of loans` `Average loan amount`  
## <chr> <int> <dbl>  
## 1 Car 313 117687.  
## 2 Household content 419 82816.  
## 3 business 88 127569.  
## 4 education 50 86316   
## 5 other 11 260782.  
## 6 repairs 19 84268.

There is a significant difference in loan amount based on purchase as average loans for large asset purchases such as a car and business are higher than that of relatively smaller purchases such as education or repairs.

## Question 7

ggplot(test\_data, mapping = aes(x = LIVING, fill = STATUS)) + geom\_bar(position = "dodge") + geom\_text(stat = "count", aes(label = after\_stat(count)), position = position\_dodge(width = 1.0), vjust = 1.5) + labs(title = "Living Arrangement vs Living Status", y = "NUmber of clients", x = "Living Status") + theme\_minimal() + theme(plot.title = element\_text(hjust = 0.5))



## Question 8

# Question 8.1  
print("Missing values in Current variable")

## [1] "Missing values in Current variable"

sum(test\_data$CURRENT == "")

## [1] 352

print("Missing values in Savings variable")

## [1] "Missing values in Savings variable"

sum(test\_data$SAVINGS == "")

## [1] 158

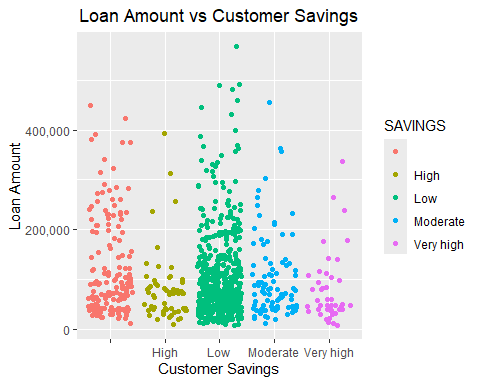
# Question 8.2  
MORE\_LOANS <- ifelse(test\_data$NUMPROD > 1, "Yes", "No")  
print("Observations where More loans is a yes")

## [1] "Observations where More loans is a yes"

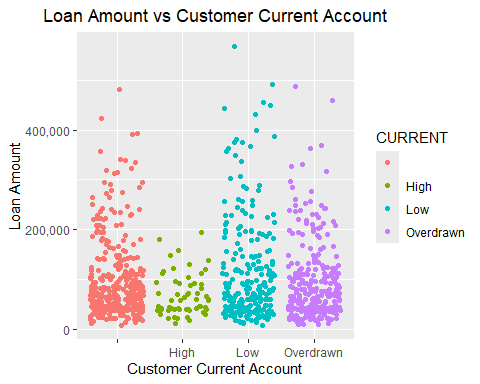
sum(MORE\_LOANS == "Yes")

## [1] 328

# Question 8.3  
ggplot(test\_data, mapping = aes(x = SAVINGS, y = AMOUNT, colour = SAVINGS))+ geom\_point(position = "jitter") + labs(title = "Loan Amount vs Customer Savings", y = "Loan Amount", x = "Customer Savings") + theme(plot.title = element\_text(hjust = 0.5)) + scale\_y\_continuous(label = comma)



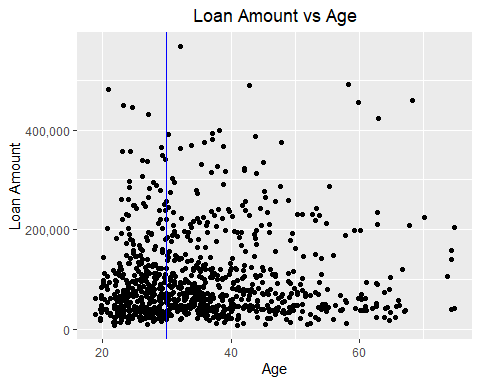
ggplot(test\_data, mapping = aes(x = CURRENT, y = AMOUNT, colour = CURRENT))+ geom\_point(position = "jitter") + labs(title = "Loan Amount vs Customer Current Account", y = "Loan Amount", x = "Customer Current Account") + theme(plot.title = element\_text(hjust = 0.5)) + scale\_y\_continuous(label = comma)



The Amount vs Savings indicates that those with low savings took a high frequency of small loans, although with respect to the specific loan amount no relationship is evident as all savings classes took loans of varying amounts. Due to missing data the Amount vs Current graph shows no clear relationship between the amount borrowed and current balance as missing values are proportional to the low and overdrawn balances.

## Question 9

ggplot(test\_data, mapping = aes(x = AGE, y = AMOUNT))+ geom\_point(position = "jitter") + labs(title = "Loan Amount vs Age", x = "Age", y = "Loan Amount") + theme(plot.title = element\_text(hjust = 0.5)) + scale\_y\_continuous(label = comma) + geom\_vline(xintercept = 30, color = "blue")



# Youth below 100000R  
youth\_below\_100k <- (test\_data%>%filter(AGE < 30 & AMOUNT < 100000)%>%summarise(youth = n())) / (test\_data%>%filter(AGE < 30)%>%summarise(youth = n()))  
# Youth below 80000R  
youth\_below\_80k <- test\_data%>%filter(AGE < 30 & AMOUNT < 80000)%>%summarise(youth = n()) / (test\_data%>%filter(AGE < 30)%>%summarise(youth = n()))  
# Youth below 20000R  
youth\_below\_20k <- test\_data%>%filter(AGE < 30 & AMOUNT < 50000)%>%summarise(youth = n()) / (test\_data%>%filter(AGE < 30)%>%summarise(youth = n()))  
# Older below 100000R  
older\_below\_100k <- test\_data%>%filter(AGE >= 30 & AMOUNT < 100000)%>%summarise(older = n()) / test\_data %>% filter(AGE >= 30)%>%summarise(older = n())  
# Older below 80000R  
older\_below\_80k <-test\_data%>%filter(AGE >= 30 & AMOUNT < 80000)%>%summarise(older = n()) / (test\_data%>%filter(AGE > 30)%>%summarise(older = n()))  
# Older below 20000R  
older\_below\_20k <-test\_data%>%filter(AGE >= 30 & AMOUNT < 50000)%>%summarise(older = n()) / (test\_data%>%filter(AGE > 30)%>%summarise(older = n()))  
  
print(paste("Proportion of youth with loans under 100,000R:", round(youth\_below\_100k$youth, 3)))

## [1] "Proportion of youth with loans under 100,000R: 0.668"

print(paste("Proportion of older clients with loans under 100,000R:", round(older\_below\_100k$older, 3)))

## [1] "Proportion of older clients with loans under 100,000R: 0.638"

print(paste("Proportion of youth with loans under 80,000R:", round(youth\_below\_80k$youth, 3)))

## [1] "Proportion of youth with loans under 80,000R: 0.556"

print(paste("Proportion of older clients with loans under 80,000R:", round(older\_below\_80k$older, 3)))

## [1] "Proportion of older clients with loans under 80,000R: 0.576"

print(paste("Proportion of youth with loans under 20,000R:", round(youth\_below\_20k$youth, 3)))

## [1] "Proportion of youth with loans under 20,000R: 0.366"

print(paste("Proportion of older clients with loans under 20,000R:", round(older\_below\_20k$older, 3)))

## [1] "Proportion of older clients with loans under 20,000R: 0.355"

Depending on the classification of a small loan (For Example 100,000R)the bank’s marketing manager’s claim is correct, however from the data as loans get smaller their claim is wrong as the proportion of small loans relative to the total number of youth or older customers is approximately equal.

## Question 10

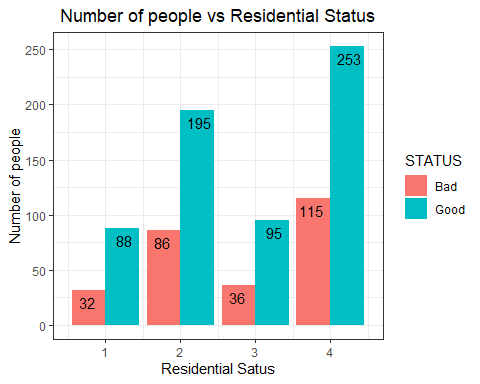
test\_data%>%group\_by(GENDER, MAR\_STAT)%>%tally()

## # A tibble: 4 × 3  
## # Groups: GENDER [3]  
## GENDER MAR\_STAT n  
## <fct> <fct> <int>  
## 1 "" "Single" 497  
## 2 "Female" "" 273  
## 3 "Male" "Divorced" 47  
## 4 "Male" "Married" 83

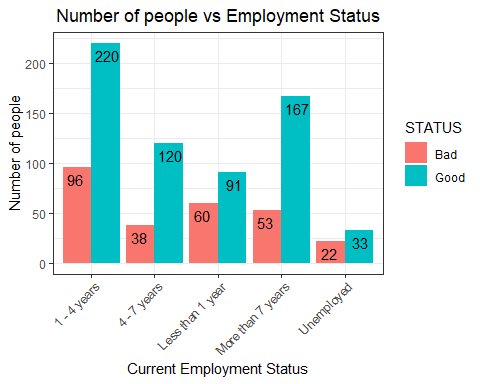
All single customers have been grouped with no gender and all female customers have no classification when it comes to marital status. These indicate missing data or a “mix-up” of categories as suggested. This highlights that the data has inconsistencies and is not very useful for certain analysis.

## Question 11

# Number of people vs Reside  
ggplot(test\_data, aes(x = RESIDE, fill = STATUS)) + geom\_bar(position = "dodge") + labs(title = "Number of people vs Residential Status", y = "Number of people", x = "Residential Satus") + theme\_bw() + theme(plot.title = element\_text(hjust = 0.5)) + geom\_text(stat = "count", aes(label = after\_stat(count)),position = position\_dodge(width = 1.0), vjust = 1.5)



#Number of people vs Employ  
ggplot(test\_data, aes(x = EMPLOY, fill = STATUS)) + geom\_bar(position = "dodge") + labs(title = "Number of people vs Employment Status", y = "Number of people", x = "Current Employment Status") + theme\_bw() + theme(plot.title = element\_text(hjust = 0.5)) + geom\_text(stat = "count", aes(label = after\_stat(count)), position = position\_dodge(width = 1.0), vjust = 1.5) + theme(axis.text.x = element\_text(angle = 45, hjust = 1))



After investigating the data set it is clear that the risk manager’s assessment is true. From the plots above, it clearly shows out of the proportion of people who take out loans in the various RESIDE and EMPLOY categories, those who have been employed the longest and stayed at the same residence the longest, (“i.e are more stable”) tend to pay their loan.

## Question 12

brief\_summary <- test\_data%>%group\_by(STATUS)%>%summarize("Mean Amount" = mean(AMOUNT, na.rm = TRUE), "Median Amount" = median(AMOUNT, na.rm = TRUE), "Mean Duration" = round(mean(DURATION, na.rm = TRUE), 2))  
  
print(brief\_summary)

## # A tibble: 2 × 4  
## STATUS `Mean Amount` `Median Amount` `Mean Duration`  
## <fct> <dbl> <int> <dbl>  
## 1 Bad 122902. 81000 28.0  
## 2 Good 92689. 70000 19.2

From the various analysis conducted through this assessment, it is clear that, the amount loaned, length of current residence as well living arrangement and length of current employment are the most important predictors of a customer’s credit status. For more insight to these findings, I refer to the graphs from previous analysis namely “Loan Amount vs Age II” graph, “Living Arrangement vs Living Status” graph, “Number of people vs Residential Status” graph, “Number of people vs Employment Status”. A brief summary from the table above stresses the point that the amount loaned is the most important predictor of a customer’s credit status.